Proceedings

of the

Conference on Directions for Mathematical Statistics

University of Alberta, Edmonton, Canada

12–16 August 1974

EDITOR: S. G. GHURYE

Special Supplement to
ADVANCES IN APPLIED PROBABILITY
September 1975

PARTICIPANTS

SESSION NO.	CHAIRMAN	SPEAKER
1	Harry E. Gunning President University of Alberta Edmonton, Alberta Canada T6G 2G1	Mark Kac Department of Mathematics Rockefeller University New York, N.Y. 10021 U.S.A.
2	G. S. Watson Department of Statistics Princeton University Princeton, N.J. 08540 U.S.A.	Meyer Dwass Department of Mathematics Northwestern University Evanston, Illinois 60201 U.S.A.
3	Emanuel Parzen Statistical Science Division State University of New York at Buffalo Amherst, N.Y. 14226 U.S.A.	Ronald Pyke Department of Mathematics University of Washington Seattle, Washington 98195 U.S.A.
4	Meyer Dwass Department of Mathematics Northwestern University Evanston, Illinois 60201 U.S.A.	Joseph Gani Division of Mathematics and Statistics C.S.I.R.O., P.O.Box 1965 Canberra City, A.C.T. 2601 Australia
5	Tore Dalenius Visiting Professor of Statistics Brown University Providence, R.I. 02912 U.S.A.	C. R. Rao Indian Statistical Institute 538 Yojana Bhavan, Parliament St. New Delhi 1 India
6	Bradley Efron Department of Statistics Stanford University Stanford, California 94305 U.S.A.	I. Richard Savage Department of Statistics Yale University New Haven, Connecticut 06520 U.S.A.
7	Ingram Olkin Department of Statistics Stanford University Stanford, California 94305 U.S.A.	John S. Chipman Department of Economics University of Minnesota Minneapolis, Minnesota 55455 U.S.A.

SESSION NO.	CHAIRMAN	SPEAKER
8	James V. Zidek Department of Mathematics University of British Columbia Vancouver 8, B.C. Canada	Peter Huber E.T.H. Clausiusstrasse 55 8006 Zurich Switzerland
9	Dean D. M. Ross Faculty of Science University of Alberta Edmonton, Alberta Canada T6G 2G1	L. Cavalli-Sforza Department of Genetics Medical School Stanford, California 94305 U.S.A.
10	Donald A. Dawson Department of Mathematics Carleton University Ottawa, Ontario Canada K1S 5B6	Wendell Fleming Department of Mathematics Brown University Providence, R.I. 02912 U.S.A.
11	Murray Rosenblatt Department of Mathematics University of California La Jolla, California 92037 U.S.A.	Dennis V. Lindley* Department of Statistics University College Gower Street London WC1E 6BT England
12	W. J. Hall Department of Statistics University of Rochester Rochester, N.Y. 14627 U.S.A.	Herbert Robbins Department of Mathematical Statistics Columbia University New York, N.Y. 10027 U.S.A.

* Professor Lindley spent the academic year 1974–75 at Iowa Testing Programs, 334 Lindquist Center, Iowa City, Iowa 52242, U.S.A.

CONTENTS

PREFACE

A reader who sees that this volume contains the proceedings of a conference is likely to exclaim, 'Oh no, not yet another conference!'. Such a reader would have a fair amount of justification for this reaction. This does seem to be the Age of Conferences, since every Tom, Dick, Harry, Jane and Mary seems to be organizing them. Usually these are conferences on specific research areas or topics and are designed for workers in those areas to come together to report on latest developments (although sometimes they seem to be designed just for workers in the areas to come together).

This volume essentially consists of talks given at a conference on Directions for Mathematical Statistics which I organized for a reason different from the usual one. The main aim of this conference was to take a little time out from proving more theorems in order to try to assess the state of the field—an attempt at what might be called rational algebraic meditation.

If we look at the history of any science, we find that, by and large, all scientific research is the eventual result of some question which was asked by someone in connection with some problem of the real world. Even many seemingly abstract origins can be traced back to some not so abstract problem which served as an inspiration. As the subject develops, it increases in sophistication and requires an increasing amount of concentration on it, leading to specialization. Eventually there comes a stage when people are working in this area without any reference to anything outside of that particular area. If they have the good fortune to be in a field which is naturally rich with potentialities and also has attracted first-class brains to it, this self-contained cooking can go for quite some time without loss of steam. On the other hand, if the Goddess of Fortune has not been so kind, the activity runs out of steam in a little while.

Looking at the field of statistics, it is obvious that the initial development came as a result of problems and observations in the natural sciences. The solution of these problems required both a philosophy, which is the philosophy of modern statistics, and mathematical techniques to transform this philosophy into concrete formulas. The need for the development of these mathematical techniques resulted in the creation of a field, still rather vaguely defined, which came to be known as mathematical statistics.

A careful study of the current literature in this area leaves some people with the idea that mathematical statistics perhaps is not of interest either mathematically or statistically. This raises a question which I feel needs to be examined:

Is mathematical statistics, as currently known, a dead-end street?

The conference was organized with a view to focusing attention on this

question. The program was intended to consist of talks by leading authorities dealing with various questions:

— what is pure research and applied research in our field?
— how much emphasis should be placed on applicability or 'relevance'?
— what are some areas of current interest in which statistical methodology will yield substantive dividends?

It was originally intended to have about sixteen speakers including several from areas like the social and biological sciences, which apparently still offer considerable opportunities for new and exciting statistical developments. A combination of various circumstances reduced the number of speakers to twelve, and resulted in representation from outside of statistics that was much less than originally planned. These twelve talks, in a somewhat revised form, are published here in the order in which they were given at the conference.

There were twelve sessions; the names of chairmen of sessions and speakers are listed on the next page. There were also official and unofficial discussants who made valuable contributions.Unfortunately the economics of the situation have dictated that the discussions be omitted from this report of the conference.

In closing, I wish to thank the speakers who agreed to contribute to the conference and the other participants who made it possible. I also wish to thank the members (teaching staff, secretaries and graduate students) of the Department of Mathematics of the University of Alberta without whose willing and enthusiastic services the physical arrangements necessary for the conference would not have been possible.

Finally, thanks are due to the National Research Council of Canada, the Canada Council and the University of Alberta for financial assistance which enabled the conference to be, to the University of Alberta for a publication grant which helped to cover some of the costs of publication and to Miss M. Hitchcock for her able handling of all matters concerned with publications. The conference participants, who expressed a desire to have a permanent record, owe a debt to the Applied Probability Trust for providing this vehicle for publication.

Edmonton, Alberta,
13 March 1975

S. G. GHURYE

Suppl. Adv. Appl. Prob. **7**, 3–4 (1975)
Printed in Israel

INTRODUCTION

HARRY E. GUNNING, *President, University of Alberta*

As President of this University, I am very happy to welcome the distinguished lecturers and other participants in this conference on Directions for Mathematical Statistics, and as a chemist, I feel singularly honoured to be given an opportunity to be included among you mathematicians.

I should like to use this opportunity, first of all, to dispel some misconceptions that people in more densely-populated parts of the world have about us. I know that Edmonton may seem to many of you to be on the northern fringe of civilization. Therefore, I should like to try to convince you that we, at this university, have many of the same academic ideals that you have at your various institutions.

I feel rather strongly that there has been far too much talk about teaching in universities, and that a university is not a proper place for teaching in the common meaning of the word. The main function of a university should be, in my view, to provide an environment in which students feel stimulated and challenged to bring out the best in themselves — an environment of constant intellectual ferment. Conferences such as this one fit in very closely with my concept of what a great university should be doing.

The theme of the conference is timely and of very broad appeal. As a research chemist, I have worked in both pure and applied research, and this constant conflict between the two is something that interests me a great deal. Now, all of us, I believe, would like to be doing something useful — something relevant, to use a much over-worked word. But for those of us working at the forefront of knowledge, it is an extremely difficult task to distinguish between the potentially useful and the dead-end street, to use Dr. Ghurye's pointed phrase. Unfortunately, new ideas do not identify themselves in this convenient fashion, and the complex sets of criteria for determining social utility are constantly undergoing change, and these changes are being brought about by the operation of forces which we do not, by any means, fully understand. I should like to cite an example from my own personal experience of the difficulty of determining whether a particular research endeavour has potential social utility.

Some twenty years ago, I was working on a problem in pure spectroscopy, dealing with the hyperfine structure of one of the resonance lines of mercury —

and certainly a more useless programme one could not possibly imagine. But it was an intriguing problem to me, for in it I saw the possibility that one might be able to take the individual hyperfine components of a resonance line which are due to the isotopes and their finite nuclear spin — such hyperfine components being separated from each other by only about 10 milli-Ångstroms — and excite single isotopes of mercury thereby. Well, we did a great deal of work in this field which we found to be fascinating indeed. However, the point I wish to make is that when we started looking into the photochemistry of the problem, we were amazed to discover that this turned out to be in fact the most effective method of separating isotopes yet discovered. Since that time, this technique has developed considerable potential industrial importance. This is but one of almost a countless number of examples of something that began as totally useless pure research and ended up, some years later, as very socially useful research indeed.

Now, whether we are working in mathematical statistics as you are or in physical chemistry as I am, there is one thing I am confident of; namely, that we share the common constraint that really good research ideas are extremely rare to come by. And perhaps what we need is not to change our conceptual mechanisms which have proven their functionality over a long period of time, but rather to be more sensitive in our minds to the needs of society, and see how we can expand the importance of our ideas by exploring their total significance to as many fields of endeavour as possible.

The number of different fields in which an idea is found to be useful is a good index of the basic worth or quality of that idea. For example, when a research contribution by a mathematician is found to be useful in chemistry (a field which in general is rather far from the core of mathematics), it is a clear indication that it has a basic content of substance and significance. This brings me to my pleasant task of introducing your first distinguished speaker. Professor Kac is an internationally famous mathematician who needs no introduction by me; I shall therefore content myself with pointing out that even I, a mere chemist, knew of some of the research contributions of Professor Kac before I was ever asked to introduce him to this audience. This came about because, for many years, I worked in the field of the statistical mechanics of chemically-reacting systems and I can tell you that chemists owe a great debt to Professor Kac for his important contributions to that field. We are very fortunate to have as our first speaker a distinguished and versatile mathematician who, out of his own experience, can explore for us the interconnections between pure and applied research.

Suppl. Adv. Appl. Prob. 7, 5–11 (1975)
Printed in Israel

SOME REFLECTIONS OF A MATHEMATICIAN ON THE NATURE AND THE ROLE OF STATISTICS

MARK KAC, *Rockefeller University, New York*

A principal purpose of this conference, as I am given to understand, is to try to look into the future of mathematical statistics. Now, there is a saying, attributed to Niels Bohr but apparently an old Danish proverb, that it is difficult to predict, especially the future. Therefore I will not engage in this extraordinarily difficult and involved task, but will restrict myself to illustrating some points which might be relevant to the theme of the conference.

First of all, it is easy to give numerous examples of research activity which started in one direction and happened to produce results in quite a different area. For example, when Sadi Carnot was working so hard on the steam engine, he was only interested in improving it. In actuality, he ended up discovering a fundamental law of nature — a discovery which has led to phenomenal consequences. At my own institution, which was then the Rockefeller Institute, O. T. Avery was working with his collaborators during the last world war on an extremely applied problem concerned with pneumonia. In the process of this investigation he discovered that the carriers of genetic information were not proteins but nucleic acids. As you all know, this discovery revolutionised the whole field of genetics, although the original problem which led to it was a specific applied medical problem. So, one never knows: It is not the problem, or the name attached to it that is pertinent. What matters is the special combination of the men, the problem, the environment — in fact, exactly those things which no one can possibly predict.

So now I would like to speak a little bit on what statistics, with or without the adjective 'mathematical', has meant to me. Perhaps I am not quite the right person to speak about it, because my connections with statistics are somewhat tangential; nevertheless, input from the outside can sometimes be useful.

My first exposure to statistics, although I did not realise then it was statistics, took place when I was 14 or 15 years of age. My class had at that time an extraordinary teacher of biology who gave us an outline of Darwin's theory and, in particular, explained how one of the claims of that theory, namely that individual characteristics are inheritable, was demolished by an experiment and a little bit of thinking. This was done by W. Johannsen, who published his results in 1909. There is a particularly vivid description of it in a book by

Professors Cavalli-Sforza and Bodmer (p. 523 of [1]). Johannsen took a large number of beans, weighed them and constructed a histogram; the smooth curve fitted to this histogram was what my teacher introduced to us as the Quetelet curve. That was my first encounter with the normal distribution and the name Quetelet.

At this time I would like to make a small digression which has its own point to make. Quetelet was an extraordinarily interesting man, who was a student of Laplace and was the first to introduce statistical methodology into social science (even though he was trained as an astronomer); he was also the author of some early books in the area (*Lettres sur la Théorie des Probabilités*, 1846; *Physique Sociale*, 1st ed. 1835, 2nd ed. 1869; *Anthropométrie*, 1870). It might interest some of you to know that Quetelet was private tutor to the two princes of Saxe-Coburg, one of whom, Prince Albert, later became Queen Victoria's Consort. He was the first major governmental figure to try to introduce some kind of rational thinking into the operations of the government, and thus may be considered as the forefather of Operations Research. Now, the point of this digression is that if you look back at this connection between Astronomy and Operations Research, via Laplace, Quetelet and Prince Albert, you will realise the significance of the Danish proverb I quoted at the beginning.

Anyway, coming back to Johannsen, he argued that if all individual characteristics are inheritable, then if we take the small beans and plant them, take the large ones and plant them, and plot separately the two histograms for the progeny of the small and the large beans, then we should again obtain Quetelet curves, one centred around the mean weight of the small beans used as progenitors and the other around that of the large ones. Now, he did carry out such an experiment and did draw those histograms, and discovered that the two curves were almost identical with the original one. Actually, there was a slight shift to the left for the small ones and to the right for the large ones, so that by repeated selection one could separate out the two populations. Of course, we now know that it is possible to distinguish between the genetic and the environmental factors, because the mean is controlled by the genetic factor while the variance is controlled by the environmental.

I have mentioned the above example because it illustrates a kind of unity of scientific thought: here was a basic problem in biology which was solved by a rather simple idea, which on closer analysis turns out to have an underlying mathematical foundation; and when one thinks some more about it, one is able to put a great deal of quantitative flesh on this extraordinarily interesting and impressive qualitative skeleton.

Another example of the same kind in which there is nothing quantitative or mathematical to begin with is connected with the great James Clerke Maxwell. I am sure all of you know that he invented a demon which has been named after

him; but it is relatively few people who know in what connection and for what purpose the demon was invented. You will find an excellent account of this in a magnificent article by Martin J. Klein, a distinguished physicist and a distinguished historian of physics (pp. 84–86 of [2]); but since this article might not be readily available to all of you, I take the liberty of repeating some of its contents.

As a matter of fact, Maxwell invented the demon in a reply to a letter from his friend Tait, who was writing a text-book on thermodynamics, and who always submitted to Maxwell for criticism whatever he wrote. Maxwell wrote to him: 'Any contributions I could make to that study are in the way of altering the point of view here and there for clearness or variety, and picking holes here and there to ensure strength and stability.' Maxwell then proceeded to pick the following hole (remember this was in 1867): He suggested a conceivable way in which 'if two things are in contact, the hotter *could* take heat from the colder without external agency', which would absolutely contradict the orthodox statement of the Second Law of Thermodynamics.

To quote from [2], 'Maxwell considered a gas in a vessel divided into two sections, A and B, by a fixed diaphragm. The gas in A was assumed to be hotter than the gas in B, and Maxwell looked at the implications of this assumption from the molecular point of view. A higher temperature meant a higher average value of the kinetic energy of the gas molecules in A compared to those in B, which is now well known to every student who has taken the elementary course in physics or physical chemistry. But as Maxwell had shown some years earlier, each sample of a gas would necessarily contain molecules having velocities of all possible magnitudes, distributed according to a law known afterwards as Maxwellian, which is, as a matter of fact, the same probability law as that described by a Quetelet curve. 'Now', wrote Maxwell, 'conceive of a finite being who knows the paths and velocities of all the molecules by simple inspection but who can do no work except open and close a hole in a diaphragm by means of a slide without mass.' This being is to be assigned to open the hole for an approaching molecule in A only if the molecule has a velocity less than the root mean square velocity of the molecules in B; it is to allow a molecule in B to pass through the hole into A only if its velocity exceeds the root mean square velocity of molecules in A. These two procedures are to be carried out alternately, so that the numbers of the molecules in A and B do not change. As a result of this procedure, however, 'the energy in A is increased and that in B diminished; that is, the hot system has got hotter and the cold colder, and yet no work has been done; only the intelligence of a very observant and neat-fingered being has been employed'. If we could only deal with the modecules directly and individually in

the manner of this supposed being, we could violate the Second Law. 'Only we can't', added Maxwell, 'not being clever enough'.

Some years later, in a letter to John William Strutt, better known as Lord Raleigh, he came even closer to what turned out to be one of the most signficant break-throughs in scientific thinking. 'For', he said, again referring to his demon, 'if there is any truth in the dynamical theory of gases, the different molecules in a gas of uniform temperature are moving with different velocities'. 'That was the essential thing', to quote again [2], and the demon only served to make its implications transparently clear. Maxwell even drew an explicit 'morale' from his discussion: 'The 2nd law of thermodynamics has the same degree of truth as the statement that if you throw a tumblerful of water into the sea, you cannot get the same tumblerful of water out again', and you see here the birth, the real birth — before Boltzmann and certainly before Gibbs — of the statistical approach to problems in physics, which proved to be of such enormous impact and usefulness.

Notice there was not a single formula in all these letters; there was only the drawing of conclusions from the variability of the velocities, which again can be attributed to the molecular structure of matter, and it immediately made possible further progress by again putting quantitative meat on the qualitative bones.

The two examples which I have given above were selected with a view to illustrating the basic strength of statistics: in both cases the analysis was based on the variability or random fluctuations that were present in each. Now random fluctuations exist everywhere and are often treated as a nuisance — 'the error term'. It is statistical methodology that is able to extract such information as is contained in this variability; and as such statistics is not a branch of this-or-that well-established science, but a discipline in its own right, — an important part of scientific methodology. If you look at it from this point of view, then what is important is not whether it is 'mathematical' statistics or 'demographic' statistics or 'applied' statistics, but how good it is as statistics.

Given that the subject of statistics cuts across inter-disciplinary boundaries, it is inevitable that it deals with problems which different research workers encounter in widely different areas. I want to give an example of this from my own experience. It goes back to the days of the last world war when one of the tasks which the Radiation Laboratory was charged with was the improvement of radar. Although most of the work in this connection was highly applied, there was also a theoretical group with which I was associated as a consultant. One problem that this group was interested in was to find out how the observer who is watching the radar scope reacts to what he sees on the scope. To simplify the problem, suppose that the observer on duty has been instructed to watch the scope diligently and whenever he sees a blip to ring a warning bell,

because there might be an enemy aeroplane. Now the problem is: When is a blip a blip? That is to say, when is a blip due to a signal and when is it due to (random!) noise?

The following experiment was performed: An observer was placed in front of the scope and told to watch a certain spot; he was told that with a 50:50 chance a signal would be put on or not put on, and he was to say 'yes' or 'no', according as he saw something or did not see anything. Now, of course, when the signal was strong, there was no question and the observer was right one hundred per cent of the time; but then the strength of the signal was slowly decreased until the signal-to-noise ratio became one or a little less, and the observer's probability of error underwent a corresponding increase.

Although the observer did not know in which trials there was a signal and in which there was not, the people in charge knew and were able to estimate from the records the observer's probability of a correct answer. It was then possible to compare this actual performance with a theoretical ideal. The theory of the ideal observer, which was developed in 1944 by a group of workers including Arnold J. F. Siegert, used a line of reasoning which has been familiar to statisticians for a long time, but was new to us. (For a detailed description see the book by Lawson and Uhlenbeck [3].)

The reasoning proceeds as follows: If there is no signal, the displacement, or deflection, on the scope is a random variable with density f_0; if there is a signal, the density is f_1. In actual practice, they can be taken to be two normal densities with the same variance and means proportional to the strength of the signal. Now, the 'ideal observer' knows f_0 and f_1 and is extraordinarily clever; and he sets out to construct a rule for when to say 'yes' and when to say 'no', the rule to be such as to minimise the probability of error. Now, we are on familiar ground, and the solution is obviously that given by a variant of the Neyman-Pearson theory, although the workers on the project did not know it as such.

Looking back at this experience, I wonder how many research workers there are in various fields all over the world, who are at this time struggling with problems whose solutions already exist in the statistical literature.

Anyway, there is a sequel to the above story. Many years later, I was consulting with an outfit involved with the space programme, in particular with automatisation of signal detection. The problem was to pre-teach an automaton so that it can learn to detect signals which it receives when floating around in outer space. In this connection it occurred to me to try to see whether an automaton exposed to the problem of signal detection could discover for itself the Neyman-Pearson theory empirically. To make a long story short, it turns out that in the very special case in which f_0 and f_1 are Gaussian, with the same variance it is possible to devise an automaton and a learning programme which turns the automaton into an ideal observer. Briefly, the learning programme is

as follows: The automaton sets itself an arbitrary threshold, and it is instructed that whenever the signal received exceeds the threshold, it must say 'yes', otherwise 'no'; if it answers correctly, the threshold is maintained at the same level, otherwise it is moved right or left depending on the kind of error; and so the process continues. It turns out that what we have here is a random walk with an attraction toward the true threshold; and then one can show that in a certain sense there is convergence in probability to the true threshold. (For a more detailed discussion see [4], [5] and especially the excellent book [6]. In general, one obtains a threshold criterion which is not necessarily that of the ideal observer, contrary to the erroneous claim made in [5].)

When this result was published, it came to the attention of psychologists interested in learning theory (notably Dr. D. Dorfman and his collaborators), who proceeded to make various experiments and modfications, and to publish papers in which I was referred to — with the result that I acquired an undeserved reputation among learning theory psychologists. (For a brief discussion of the psychological background see [6], pp. 25–26, where other references may be found.) More seriously, the interesting aspect is that a train of thought, which had started many years earler in an applied problem in one area, ended up later in theoretical investigations in a different direction because of the underlying statistical current. The other interesting aspect is the problem of constructing an automaton that performs as well as an ideal observer in more general situations. It is not an easy problem, and it should lead to interesting mathematics.

Finally, my last example is one in the opposite direction: a pure mathematical context in which statistical thought makes it possible to state new kinds of mathematical problems. Although I like the example immensely, I must restrict myself to a brief reference to it, since I have repeated it many times in print and otherwise. Now many of you know what is known as Descartes' rule of signs for estimating the number of positive real roots of a polynomial with real coefficients; namely, the number of positive real roots is never larger than the number of changes of sign in the sequence of coefficients. Now, we can ask a question: 'How good is Descartes' rule?'

This, of course, is a vague question, since there is no definition of what is meant by 'good' or 'not good' in this case. Now, there is nothing wrong with vague questions; it is the combination of vague question and vague answer that is bad. Many imprecisely stated questions have a tremendous amount good science in them. In our case, in order to measure how good Descartes' rule of signs is, we necessarily must consider an ensemble of polynomials; and once you have an ensemble, then necessarily you have to deal with it statistically. Now, every polynomial is identified uniquely by the vector of its coefficients; however, since the roots are unaffected if all the coefficients are multiplied by

the same non-zero number, therefore for our purpose, all polynomials are identified by points on the surface of the unit sphere. Assuming these to be uniformly distributed, we can then estimate the average value of the number of real roots for polynomials of a high degree. As I have stated earlier, I do not wish to repeat the details here, but merely draw attention to how, starting with a simple vague question about a pure mathematical problem, we have generated more mathematics by taking a statistical approach. (For some details see [7].)

In conclusion I would like to point out that these examples indicate the inter-play between ideas from different domains that goes on all the time, and is in fact extremely valuable for the development of human knowledge; they illustrate the fact that everything is connected to everything else, and it is impossible to separate completely any intellectual endeavor from any other. In particular, as regards mathematics, you cannot separate it from its applications to the external world, and you cannot separate statistics from mathematics, or mathematical statistics from applied statistics. Thus, in so far as I am at all able to look into the future and identify desirable directions for us to take, they point towards (a) unification, and (b) communication (amongst ourselves and with the outside world).

Let us be connected with as many reservoirs of inspiration and understanding as possible; let us ask questions, even vague questions, and try to answer them precisely; let us not worry about ' pure', about 'applied', about 'useful', about 'relevant', and so on. Let us, in brief, try to do our best, and what survives will be determined by Nature's Law of Survival of the Fittest; and what is fittest will be determined by the next generation in the light of what has survived. So be it!

I should like to thank Professor S. G. Ghurye for performing the nearly impossible task of reconstructing my remarks from a hopelessly garbled tape. I only hope that this archaeological feat will in some small measure be recompensed by the appearance of this article.

References

[1] CAVALLI-SFORZA, L. L. AND BODMER. W. F. (1971) *The Genetics of Human Populations.* W. H. Freeman & Co., San Francisco.

[2] KLEIN, M. J. (1970) Maxwell, his demon and the second law of thermodynamics, *Amer. Scientist* **58,** 84–97.

[3] LAWSON, J. L. AND UHLENBECK, G. E. (1950) *Threshold Signals.* Vol.24 of the Radiation Laboratory Series, McGraw-Hill, New York.

[4] KAC, M. (1962) A note on learning signal detection. *IRE Trans. on Information Theory* **8,** 126–128.

[5] KAC, M. (1969) Some mathematical models in science. *Science* **166,** 695–699.

[6] NORMAN, M. F. (1972) *Markov Processes and Learning Models.* Academic Press, New York.

[7] KAC, M. (1954) Signal and noise problems. *Amer. Math. Monthly* **61,** 23–26.

Suppl. Adv. Appl. Prob. 7, 12–16 (1975)
Printed in Israel

THE USE OF THE COMPUTER IN TEACHING STATISTICS

MEYER DWASS, *Northwestern University*

What I will talk about relates to pedagogy — the pedagogy of statistics in particular — though some of my comments may have relevance to the teaching of other subjects as well.

There is a growing involvement at universities these days in using the computer as an aid in teaching statistics. I want to describe our still limited experience at Northwestern University in these matters. I do not intend, nor am I able, to describe the 'state of the art' as a whole, but I will try to give some of my own philosophy, methods, guesses about costs and estimates of effectiveness and dangers.

Whatever we are doing at Northwestern has no doubt been done in part or whole with varying points of view elsewhere. But I do believe it is worth passing on our own approach because there must be many teachers who have not become aware of or have not seriously considered computer-oriented methods in teaching. Moreover, there is still a minimal exchange of ideas and information on the subject.

Up to this point, our experience at Northwestern has been limited to teaching several courses at the intermediate junior-senior level in statistics and a pre-calculus service course in statistics. In addition, we have taught some computer-oriented courses in finite mathematics for behavioral sciences, and several sections of calculus.

The system of computer-oriented instruction that I will describe emphasises the following:

(1) The availability of an adequate time-sharing computer facility.

(2) The involvement of the instructor in being able to use the computer in the classroom as an aid to teaching as easily as he or she uses chalk and blackboard, ditto machine and overhead projector.

(3) Student participation in learning to understand written computer programs, in being able to modify such programs and in writing programs on their own.

What do I mean by having the computer in the classroom? By this I mean having some device in the classroom which allows students to see the computer 'at work' and which allows the instructor or students to interact with the

computer. For a small class this can simply mean having a teletype in the classroom and passing printed output around for students to see. For a moderate or large class one can have wall-mounted video monitors driven by an instructor's keyboard terminal.

What does it cost to equip a classroom? My estimate is that one can set up a classroom with wall-mounted monitors sufficient for 60 students plus an instructor's console for about $5,000, with a marginal cost of about $700 for every additional 25 or 30 students. However, this is one cost area where, counter to current inflationary trends, prices may be going down. Also, technological innovations seem to come rapidly.

There are other display devices that are possible between the extremes of a single printing console and a set of video monitors. I have placed inexpensive transparent plastic paper in a teletype, ripped it off after it had been printed on and displayed the result on a conventional overhead projector. This has its disadvantages as the teletype is noisy and one must be careful not to smudge the printing on the plastic paper. There is a more sophisticated commercial version of this device in which an overhead projector is directly mounted on the teletype. By now there may be many variations on this theme involving overhead or opaque projectors.

Why does it benefit students to see the computer at work in the sense of seeing actual output coming off the computer? Suppose the material were prepared beforehand and made available in the form of notes or displayed by transparencies? Would not the net effect be the same? I think the answer is both yes and no. There is a danger that an instructor may be overwhelmed by the gadgetry and overemphasise it unnecessarily. There seems, however, to be a psychological value for students in seeing results produced on the spot. Also, just as with the traditional use of chalk and blackboard, much of what an instructor may do with the computer is spontaneous, evolving in response to the developing course of the lecture or queries by students.

What do I mean by student involvement? After all, one can use the computer in or out of the classroom, displaying all kinds of interesting computations, simulations and so forth, without students knowing a thing about how the programs work. Moreover, one can provide homework for students requiring only that existing programs be called in and some parameters or data entered. Very little need be known about interacting with a computer system and nothing need be known about programming for such uses. This approach seems to be the most common one in computer-oriented instruction using a time-sharing system; I do not mean to derogate the value of such an approach. We all know the value of SPSS (Statistical Packages for the Social Sciences), of various 'canned' interactive programs and rote-learning programs. But when students have learned enough elementary programming to modify existing

programs or to write programs of their own, such involvement with the program reinforces the theory they are learning. In elementary finite probability, for instance, understanding basic facts about sample spaces, combinatorics, random variables, frequency distributions often depends on understanding an algorithm or listing device for enumerating a large array of objects by brute force. In fact, practically, the student cannot enumerate such an array, but the process of describing to oneself how to do it often helps clarify the concepts. If the student succeeds in writing a program which does such a brute-force evaluation, he will have surely mastered the concept. I do not mean to suggest that one should forego the elementary mathematics which often makes brute-force counting unnecessary.

This makes the choice of a programming language rather important. I do not think there is a uniquely optimum choice, but I have developed a strong prejudice towards the use of BASIC as a student's first language. (I may be thinking of instructors without computer experience even more than I am of students.) In about three or four hours of out-of-class instruction in BASIC, students are able to learn a great deal of elementary programming. To achieve this, students should have a decent self-study manual and sufficient access to terminals. It helps a great deal if one provides a few lecture sessions on how to use the local system (which would not be discussed in the manual) and how to take the first steps in programming. It also helps if students are working in an area where there is an assistant available who helps them with the frustrating problems which most beginners meet. As to BASIC, I know there are strong adherents of various competitive languages. I would hope that what I am saying will not stand or fall on the basis of the particular language being used.

What kinds of pedagogic programs can be used for in-class and extra-class instruction? I think that the following non-mutually exclusive categories may account for most of the possibilities:

(1) Computation of numerical values of probabilities.

(2) Enumeration of sample spaces and evaluation of probabilities by brute-force counting.

(3) Computation of frequency distributions in one or several variables.

(4) Numerical integration to evaluate probabilities or moments.

(5) Simulation of chance experiments (using random-number generating devices) to obtain qualitative 'verification' of probabilities, moments, frequency distributions, power curves and so forth; also to obtain empirical information when the mathematical theory is not available, or when the course is not supposed to develop such mathematical theory.

Let me now outline one possible short course in elementary statistics. I will not describe the various examples of a traditional sort that are used and which

we are all familiar with; I will only refer to some of the more computer-oriented materials.

(1) Several lectures on how to use the local time-sharing computer system, plus an introduction to BASIC programming.

(2) Several lectures on axioms of finite probability, especially as they relate to sampling with and without replacement, independent trials and some dependent trial schemes.

(3) Classroom computer demonstrations showing listing of all possible sample points and calculation of their probabilities for various combinatoric problems and sampling schemes. Illustration of frequency theory of probability by large-sample simulations of elementary chance experiments and comparison of empirical frequencies with theoretical probabilities.

(4) A lecture giving definition and examples of random variables.

(5) Several lectures on frequency distributions, mean and standard deviation — both theoretical and empirical.

(6) Computer demonstrations in which empirical frequency distributions and moments are compared with the theoretical versions. Also some computer determinations of empirical distributions and moments where the calculation of the theoretical versions lies outside the scope of an elementary course.

(7) A lecture on independence and dependence.

(8) Several lectures on sample means and sample standard deviations and preciseness as a function of sample size. Computer demonstrations illustrating the law of large numbers and sampling distribution of various basic statistics.

(9) Several lectures on special distributions such as binomial, Poisson, hypergeometric. It is well to have 'canned' programs available to allow students to compute these distributions for parameters chosen arbitrarily. (This is actually easier than using tables.)

(10) Several lectures on chi-square goodness of fit tests and test of independence. (By this point, students will certainly be able to write their own programs computing the test statistic.) Computer demonstrations generating chi-square test values for randomly generated data. Large-scale simulation showing that the empirical distribution of chi-square values approximately corresponds to theoretical chi-square distribution.

(11) Several lectures on confidence intervals for binomial proportion, Poisson parameter. The stored computer programs for calculating the probabilitity distributions can be used to advantage in calculating the exact probability of telling the truth for various values of the parameters and comparing these with the advertised confidence level.

(12) A lecture on the normal distribution and the central limit theorem.

(13) Computer illustration of how the central limit theorem allows the generation of approximately normally distributed variables. A 'canned' numeri-

cal integration program can be made available to students for finding areas under the normal curve.

(14) Several lectures on confidence intervals for the mean of a normal population. Large-sample versions of same procedure for arbitrary populations based on the central limit theorem. Empirical computer 'verification' of probability of covering the true mean, using confidence intervals.

The above covers about one-quarter's worth of material and I hope it gives a rough idea of how the computer can be used to supplement a fairly traditional course. If there is a second quarter, many of us would want to continue with topics in regression, tests of hypotheses, power and non-parametric procedures.

In my own teaching, I have found such an elementary, computer-oriented course quite satisfactory. The computer demonstrations seem to provide students with insights which they would not get otherwise, since they are not ready for the underlying mathematics nor do they have any experience with real data. I have also tried this approach in a calculus-based course. Here one must be careful not to overemphasise the computer demonstrations, since students may prefer, and be better off with, not skimping on the mathematics. However in a more advanced course, one can give more sophisticated computer-tied outside assignments. I find that a weekly laboratory session using the computer is satisfactory.

Finally, I should point out what may be the main dangers in this computer approach to teaching statistics. (a) The computer tail may wag the whole statistical dog. That is, the course may turn into a course in computer programming with statistics achieving secondary importance. (b) Students may get the idea that there is no need to learn mathematics, since anything of interest can be 'approximately derived' by means of simulation or brute-force computation.

I do think that those of us who have made such beginning attempts to use the computer in teaching statistics have only scratched the surface and that there is a great deal to learn. I also believe that the prospects are optimistic about the usefulness of the computer as a pedagogic tool.

Suppl. Adv. Appl. Prob. 7, 17–37 (1975)
Printed in Israel

APPLIED PROBABILITY: AN EDITOR'S DILEMMA[1]

RONALD PYKE, *University of Washington*

Introduction

A light-hearted approach is probably best for a discussion of someone's problem. Since my problem may not be yours, the least I should do is make the hour go by as quickly and painlessly as possible. Although I will be somewhat facetious throughout, I hope you will detect the serious problem with which I am concerned: I will leave you with a serious challenge near the end of the talk. Indeed, the main purpose of a conference such as this is to challenge ourselves to think about serious problems related to our disciplines.

Professor Ghurye suggested the title of this talk; I reveal this in the hope of obtaining your sympathy. I would feel much more comfortable if I could say that when Professor Ghurye suggested this topic to me, he did so in the abbreviated form

'AP: An editor's dilemma'

for in this form I would have quickly and gladly accepted the invitation, being sure that AP could stand only for the *Annals of Probability*. (In correspondence between ourselves and our editorial boards, Professor Savage and I use AS and AP as standard abbreviations for the two *Annals* — and JAP for the *Junior Annals of Probability*? If this had been the case, many dilemmas could be easily identified; is a paper publishable or not? important or not? how to pry papers loose from referees? is a certain paper about statistics or does it concern probability?... However, this was *not* the case. I was not deceived. I walked into this challenge knowing full well that the object under discussion was the amorphous spectre of a discipline and not the concrete reality of a journal.

To insure that no confusion will arise during this hour, I will use script letters $\mathscr{AP}$ for the discipline of Applied Probability and Roman letters AP for the *Annals of Probability*. In particular J$\mathscr{AP}$ is used in place of the more usual JAP, and the title of this talk may then be written as

'$\mathscr{AP}$: An editor's dilemma'.

[1] This invited talk, presented at the Conference on Directions in Mathematical Statistics, University of Alberta, Edmonton, Canada, 12–16 August, 1974, was supported in part under National Science Foundation Grant No. GP-31361X2. Wherever possible, this paper agrees with the taped version of the talk, even though hindsight often lobbied for change.

I hope the audience will be able to distinguish between the 'script' inflection in my voice when I mention $\mathscr{AP}$ and my more normal 'Roman' rendering of AP.

(Before I continue, let me remark that I noticed this morning the prominence that mathematics, and probability and statistics in particular, enjoy here on the campus of the University of Alberta. Mathematics is elevated to the top two floors of its building; closest to the heavens. In the mathematics lounge, there are three tables of similar size. One had a Go board on it this morning, an indication that some applied research is carried on. The other two had journals on them. On one the *Annals of Statistics* was on top, while on the other the *Annals of Probability* was on top. Did our host visit the lounge early this morning to rearrange the journals to make us feel more at home?)

What is $\mathscr{AP}$? This in itself is an editor's basic dilemma. How does one define a discipline? Its definition would facilitate the determination of what portions of $\mathscr{AP}$ fall within the scope of AP, and perhaps what portions of $\mathscr{P}$ fall within the scope of J$\mathscr{AP}$. I was told just a moment ago that Professor Kac knew how to define $\mathscr{AP}$ and perhaps he will share this knowledge with us later during the discussion[2]. As for me, perhaps I should admit that I honestly do not really care how $\mathscr{AP}$ is defined, that I do not believe disciplines should even *be* defined, that I think walls and barriers between, and hence defining, disciplines should be broken down and that communication should be encouraged. With this admission, I should then sit down, thus freeing you from this hour's discourse. I could on the other hand, pretend that I know what $\mathscr{AP}$ is and attempt to convince you of my thoughts. Since I have suffered many hours on this question, I think it only fair that I elect the latter, thereby inviting you to suffer along with me. I encourage you to enter into this discussion, either during or following the talk. Hopefully by the end of the hour we will have some idea of $\mathscr{AP}$; what it is? where it is? how to foster it?... if to foster it?

Background

How many Applied Probabilists are here today? Would those of you who identity yourselves as Applied Probabilists raise your hand? — Notice how they cluster together!

How many Probabilists are here? — Notice that there are some Applied Probabilists who have not raised their hands. Let me ask the question again with the understanding that I mean 'Probabilist' with or without adjectives of any kind. — Now, that is a goodly number. And yet I still have the feeling that the Applied Probabilists are not quite sure if they are Probabilists.

[2] In the discussion which followed this talk, Professor Kac wondered if $\mathscr{AP}$ could be defined by the criterion, 'rejected by Pyke and accepted by Gani'. I wonder now if this means that $\mathscr{P}\backslash\mathscr{AP}$ relates to 'rejected by Gani and accepted by Pyke?

How many Statisticians, of any type whatsoever, are here today? — Look at that! You notice that they put their hands right up without hesitation!

How many Applied Statisticians are here? — Not too many.

Well, I think there is no need for me to define $\mathscr{AP}$, $\mathscr{AS}$, etc. Having answered these questions, it is clear that you know what they mean. Let me therefore leave $\mathscr{AP}$ undefined for the moment. Let me even postpone my two lemmas, if that is what a dilemma is. Rather, let me discuss first the editorial context in which this talk is given; it is, after all, entitled an *editor's* dilemma.

The history of probability and statistics goes back many centuries. According to one history of science, the earliest mention of probability in world literature is of a certain dice problem discussed by Benvenuti de'Rambaldi in 1390.[3] The use of lots in ancient civilization might even take the history back further. However, until the last hundred years, the disciplines of statistics and probability were very small indeed, being only slightly mathematical and without journals and editors. In fact, it was not until this century that the disciplines began to take on the form that we recognize today.

In the late 1920's, a need was felt for a journal which would focus on the mathematical advances in the theory of statistics. Much evidence had arisen to indicate the usefulness of a mathematical framework for statistics as well as probability. Many researchers had begun to formulate theoretical systems during the first three decades of this century. However, they had a dilemma. They found it difficult to publish their papers in the statistical journals of the day: they were too mathematical. They found it equally as difficult to publish in mathematical journals: they were too applied. Familiar? Out of this dilemma emerged a new journal, the *Annals of Mathematical Statistics*. It began in 1930 under the direction of Professor H. Carver. It was an individual's venture. (Possibly Gani = T^{34} (Carver), where by this I simply mean that translated by 34 years another journal was begun, for similar reasons and with a similar dependence upon an individual's energy and support.) Out of the depressions, academic and economic, of 1930, and involving Professor Carver's personal financial support, emerged a new journal. (Cf. a letter dated 14 April, 1972, from Professor H. Carver to Professor J. Hall which was reprinted in *Bulletin IMS* (1973) **2**, 11–14.) Four years later, in 1934, the Institute of Mathematical Statistics was founded as 'a society for encouraging the development, dissemination, and application of mathematical statistics'.

Today, new journals and new societies are still being formed. It undoubtedly will never be otherwise. In 1972, the *Annals of Mathematical Statistics* was replaced by two new journals, the *Annals of Statistics* and the *Annals of Probability*. Some of the events leading up to this are as follows.

---

[3] *History of Science*, G. Sarton, Harvard University Press, 1952.

There was a feeling in the late 60's on the part of some Probabilists, most of whom would call themselves Applied Probabilists, that the *Annals of Mathematical Statistics* and the Institute of Mathematical Statistics (IMS) were not representative of their interests. The coverage of statistics was also felt by some to be too narrow. In response to these criticisms, Professor Jack Kiefer, as President of the IMS, in 1969 requested the Committee on Operations to prepare a questionnaire to measure the interests of the IMS membership. About the same time, a committee on the restructuring of the *Annals* was set up, under the chairmanship of Professor Jack Hall. This committee's work straddled about three years, included an extensive sampling of members' views as requested by Professor Kiefer and culminated in the creation in 1973 of the two new *Annals*, AS and AP, as well as the *Bulletin* one year earlier.

The purpose of this split was in the first instance a practical one of halving an editor's burden. The 2500-page *Annals of Mathematical Statistics* was too much for one person. In addition however, the split would broaden the scope of the IMS by bringing in more probabilists, pure and applied, and encourage each of the new separate *Annals* to realize a healthier mix of pure and applied. As a discipline, probability needs its applications as well as its theory. The same is true for statistics.

Here then is one of my dilemmas. As a good statistician I predicted that, upon the split, the editor of AP would be deluged with papers on $\mathscr{AP}$ including queueing, reliability, dams and so on. It then would be easy to shape AP into a desirable blend of papers from all areas. Unfortunately, the prediction was incorrect. AP has *not* been deluged with papers on $\mathscr{AP}$. In fact, we have only received about a dozen papers out of more than 700 which I would call $\mathscr{AP}$. Of these we have accepted proportionately more than in other areas.

For me it is of interest to quote from the second Jeffery-Williams Lecture[4] which I gave, also in Canada, five years ago in 1969. The topic asssigned to me then was 'Whither Statistics and Probability?' (Note that the first word was spelled with an 'h', unlike the spelling used by Professor Robbins in the title of his Friday lecture.) In this lecture I mentioned that

> statistics and probability have relatively short histories as sciences go, ... As to its recent development, some factors to be considered include the demands, or lack thereof, by scientists or society for answers to specific problems, the calibre of researchers currently active in the field, the existing interactions with other disciplines and the social interest in the discipline as manifested perhaps by the degree of federal support. In mathematical sciences in particular a major factor determining development is the individualism of research; the freedom of each mathematician, statistician or probabilist to tackle any problem he wishes, guided in many cases only by his *own* estimate

[4] Published as 'Empirical Processes' in *Jeffery-Williams Lectures 1968–1972*, Canadian Mathematical Congress, Montreal 1972.

> of the problem's potential for publishable solutions. As problems yield solutions, these in turn generate papers, yield more problems, yield solutions, *ad infinitum*. A significant proportion of these papers are Ph.D. dissertations each of which signals the potential start of another branch of our discipline's tangled ivy plant of knowledge.

Many of our Ph.D. students were not in university more than a year or two before they were also giving out thesis topics to their students. These topics more than likely were simple outgrowths from their own theses, and would often lack a mature sense of direction. Our discipline grew very quickly; certain areas expanded far too quickly and some became rather stagnant.

If one looks at the history of probability theory, one is really looking at the history of limit theory. If one looks back over names like Bernoulli, De Moivre, Laplace and Poisson, limit theorems are seen to be the core of probability. If you look at publications in statistics or probability, whether in AP, J$\mathscr{A}\mathscr{P}$ or elsewhere, one sees that limit theory is well over half of the contents of our publications. Will the future be the same?

Unlike Professor Kac, I did in 1969 make some predictions as to the future directions of Mathematical Statistics and Probability. However, I did not publish them. Instead, I prepared the alternate lecture, 'Empirical Processes'. Consequently, only those with long memories can hold me to what I said. I did record, however, the underlying theme of my predictions, and that was that statistics would become more directly linked with applications. In general this has happened. My dilemma is: why in five years has it not happened in some of our journals? I believe the prediction is consistent with the historical development of probability and statistics. You can name as well as I the areas of application which have generated some of our discipline's more interesting theories. And at this time I believe it is very important that statistics and probability, whether two disciplines or one discipline or a sub-discipline of mathematics, should continue to improve its ties with areas of application. At this stage $\mathscr{P}$ and $\mathscr{S}$ are mature disciplines capable of standing on their own, but they should not isolate themselves from other disciplines. Particularly, I like to think of $\mathscr{P}$ and $\mathscr{S}$ as adopted children of mathematics. There were real problems in these subjects around the turn of the century and it turned out that mathematics had the type of deductive reasoning that was able to help immensely in the formulation of a framework in which solutions were possible. $\mathscr{P}$ and $\mathscr{S}$ remained as adopted children of mathematics up until, I would say, the last two decades. I think that today they are mature children, able to leave their mathematical home. They should not leave home, however, without leaving a communication channel open. They must not communicate only with mathematics, however. They must also communicate with other sciences. Applications bring in exciting new problems. You certainly will not think up a

new kind of stochastic process by meditating within an isolated ivy-covered tower. You must get out and find an interesting important problem; it alone could give you a new model, a new conjecture, new insight and even real satisfaction.

Where is $\mathscr{AP}$?

Let me briefly and quickly just indicate a coverage of $\mathscr{AP}$ by showing you something from some of our journals. (Last week there was a conference at York University on 'Stochastic Processes and their Applications'.[5] In the middle of the week, there was a $2\frac{1}{2}$ hour discussion on '$\mathscr{AP}$; its Nature and Scope'. It was very good. Perhaps the best way for me to have presented this talk would have been to have tape-recorded that session, and played it back for you today. Of course, in order to squeeze $2\frac{1}{2}$ hours into one hour I would have had to speed it up considerably. No matter, the resulting chipmunk-like voices would have conveyed some really interesting thoughts.) This first graph (Figure 1), given without Professor Gani's permission, is one of several interesting charts in the recent index and review of J.$\mathscr{AP}$ and A.$\mathscr{AP}$. It breaks down the papers in those journals into several topical categories. The top ones would probably in this context be called theoretical (branching processes, Markov chains,...). One notices that there is a fairly uniform coverage of these more theoretical topics. Down (Freudian slip?) in the more applied categories one notices a substantial covering of rather interesting areas, although Professor Gani in his talk will tell you of many other areas in which he would, I am sure, like to see more submissions.

Let me now look at the coverage of topics in AP. Remember, this represents only a year and a half with the data involving only nine issues. The data is based on the AMS Classification Numbers provided by the authors and so is very subjective. Authors give their papers a primary and a secondary classification. I simply gave weights of 2 and 1 to primary and secondary classifications, respectively. I think this gives a fairly good measure of content. However, I think that the key words if they were properly analysed and indexed would probably give an even better measure. In any event, consider the data summarized in Figure 2. (I never did take a course in histograms.) From this data, one sees that the heaviest coverage is in the area of weak limit theorems. There is a fairly uniform coverage of most types of stochastic processes. There are a few listings for 'Applications' in the AMS Classification Scheme. In our list, the word 'Applications' refers to the previous item in the listing. For instance, notice that Branching Processes scores 16, followed by Applications with a score of 5. As is clear from the graph, applications of all types have been

[5] The proceedings of this conference were published in *Adv. Appl. Prob.* **7**, 227–263.

TOPICAL SUMMARY

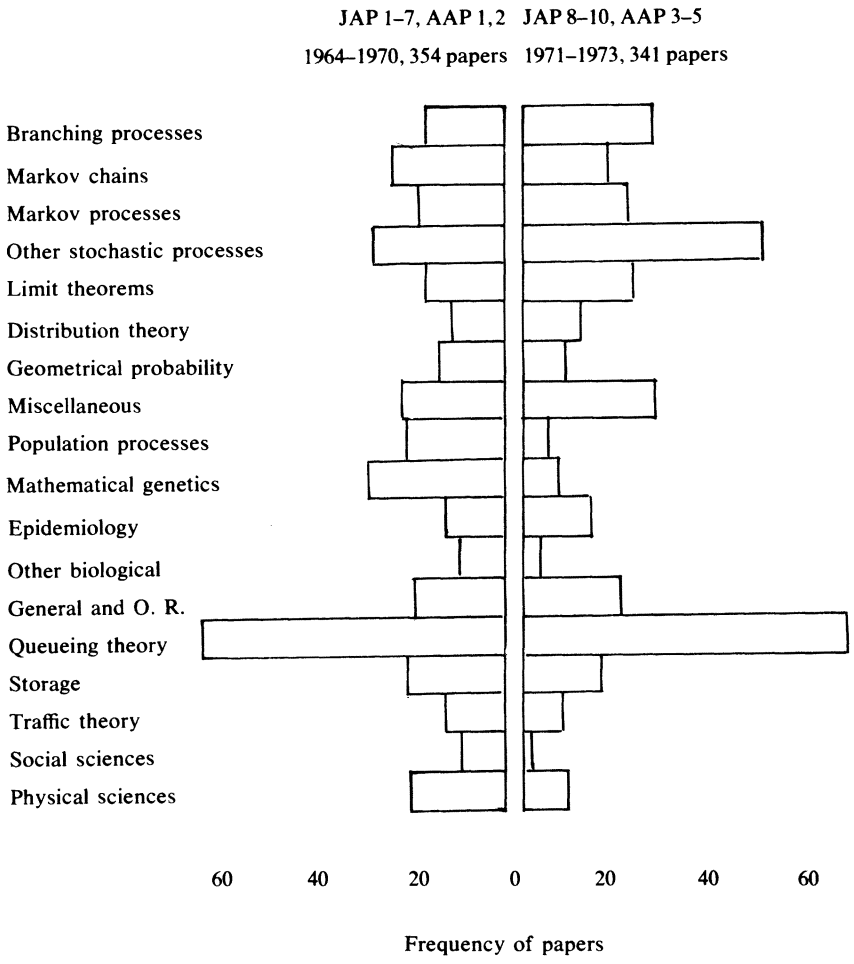


Figure 1

(Reprinted from *Complete Author and Subject Index No.* 2, Applied Probability Trust, 1974)

few. Branching Processes represents our most active area as far as applications are concerned.

There is a broad area of topics, such as optimal stopping, control theory, limit theorems, which are on the boundary between the coverages of AS and AP. There is very good communication between Professor Savage and myself, as there was between Professor Olkin and me before, concerning papers in these boundary areas. If one proves a limit theorem for a rank statistic for example, does this represent Statistics or Probability? This is difficult to determine, but it

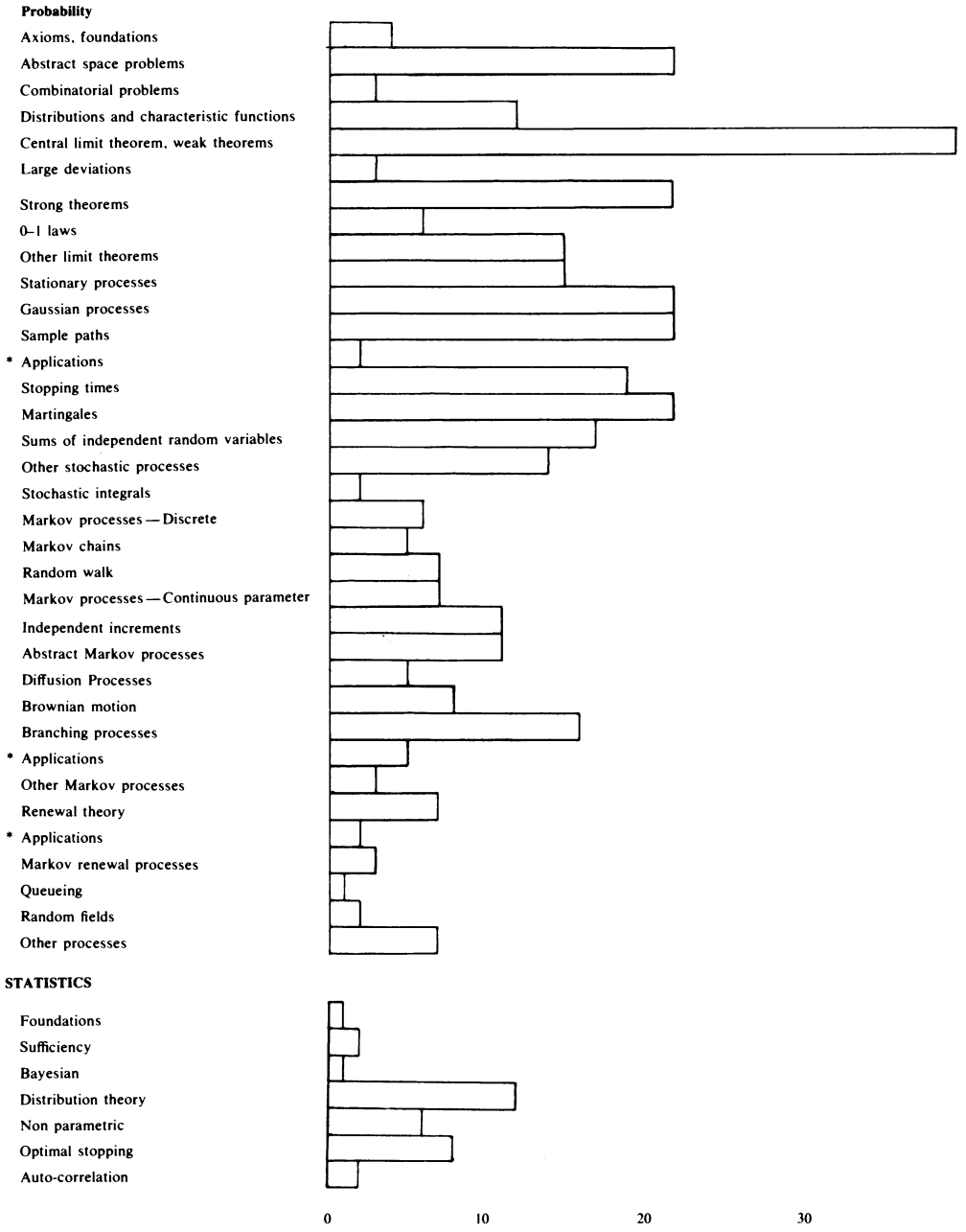


Figure 2

Topical breakdown. *Annals of Probability*, Volume 1, No. 1 (1973) through Volume 2, No. 3 (1974).

has presented no difficulty concerning which *Annals* should publish it, as both are possible.

If I consider Professor Gani's listing (Figure 1) and tabulate a similar summary for AP, I get the graph shown in Figure 3. You notice there that only five manuscripts were in areas that we would classify as really $\mathscr{AP}$, namely Genetics, Queueing and Operations Research. There is an upward (sic) bias in the coverage of AP rather than the uniformity of the spectrum covered by J$\mathscr{AP}$.

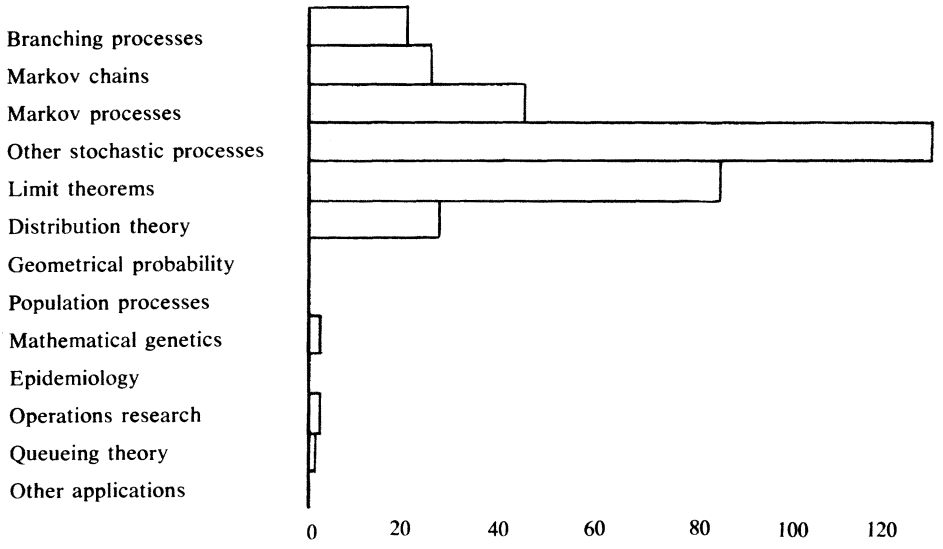


Figure 3

Summary of information presented in Figure 2

Question: Do you like this picture? Do I like this picture? This is a question we have to answer. By the way, the 'applied' papers in AP included:

—one on optimal filtering of stationary processes based on a problem arising in practice;

—one on shock and wear models in reliability theory, for which properties are obtained from various physically motivated models;

—one on an index of genealogical relatedness based on a particular genetic model;

—an application of random walks to a stock market model

plus several on branching processes and many-particle interactive models.

If you look at the coverage of $\mathscr{AP}$ in the general literature, rather than focus on these two journals, I think you see a much healthier outlook, and you find

some very impressive things. I just took, for example, the last issue (January-March 1974) of the *Science Citation Index*, (How many of you have used the *Science Citation Index*? — Good. I see they are being used.) Look up the word 'random' for example. You find five columns of references. Look up 'probability' and you find three columns. There are 170 listings in a column! The coverage is enormous, and you can go through other key words and see further evidence of the expansive coverage of probability throughout the scientific literature. However, what is more to the point is that if you look up 'random walks', say, and then run down the list of authors, I predict you will recognize very few names. By that I claim that we are too isolated. The people writing in these areas are to only a small extent, people 'of our type'. There are many, many areas of science where 'random walks' for example are studied and used.

If we now look at the ISI's *Statistical Theory and Method Abstracts* which is basically statistically oriented, what is the largest section? It is the red section of Stochastic Processes and Time Series. It had 101 pages in the latest issue I looked at (Volume 14, No. 4, 1973). This compares with sixty-five pages for the second-largest section (purple). The third-largest was in the pink section. Probability, with fifty-nine pages. This reflects the considerable emphasis that exists within the statistical literature of probabilistic topics.

What is $\mathscr{AP}$?

First of all let me remark that the adjective 'applied' appears very often throughout science, but is never defined. Even in Van Nostrand's *International Dictionary of Applied Mathematics*, there is no definition of 'applied', nor of 'applied mathematics'! I believe also that the word is often misused; properly speaking, the adjective should in most cases be 'applicable' rather than 'applied'. Let me quote from some 'applied journals' to illustrate some of the possible interpretations of 'applied'; all italics are mine.

J$\mathscr{AP}$ states that it 'contains research papers and notes *on applications* of Probability Theory to the biological, social and technological sciences'.

This is good, and implies 'applied' more than 'applicable'.

The Association of Applied Biologists 'exists to further the study of all aspects of biology and to *correlate* pure science with practice'.

Here, 'applied' is defined implicitly by the motivation, 'to correlate' theory with practice. This is excellent. It is the motivation that distinguishes the 'applied' from the 'pure'. (An aside: Professor Syski mentioned last week that the tools used by a probabilist might be a distinguishing feature between 'pure' and 'applied'. He mentioned that in the probability book by J. Neveu it is stated that only measures and not distribution functions are needed; whereas in the second book by W. Feller, it appears that the reverse is the case.)

The *Journal of Applied Physics* is 'devoted to general physics and its applications to other sciences, to engineering and to industry'... 'The editor welcomes manuscripts describing significant new experimental or theoretical results in applied physics or manuscripts *concerning* important new *applications* of physics to other branches of science and engineering'.

One can substitute 'probability' or 'statistics' for 'physics' in this definition and the resulting definition would not be too unreasonable.

Applied Microbiology is 'devoted to the advancement and dissemination of applied knowledge concerning microorganisms'.

This does not help. To me, a definition of a discipline, like $\mathscr{AP}$, should really be just an identification of an area of knowledge, and this is very difficult to put into words. It might best be done by a study of vocabularies. Take an individual's or a journal's vocabulary and correlate it with scientific vocabularies of all disciplines. It think that you will be able, by some suitable scaling method, to identify a region which we could identify as $\mathscr{AP}$.

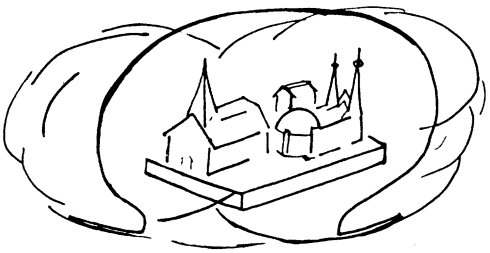
I now give my own definitions of $\mathscr{AP}$. This is done basically to entertain, but I hope that some serious points will be made. How you define your discipline depends on your outlook, your aim or your motivation. There are three main outlooks that I will emphasize today; that of a purist, a gentleman and a worker.

A. *The Purist's Definition.* The Purist's definition might go something like that given in Figure 4. His world is not ours, it is up in the ethereal heavens. I sort of detect an 'omega' surrounding its location; and it is clear that there is no crying nor suffering there caused by any 'down-to earth' considerations.

Definition A_1: $\mathscr{AP}$ refers to that part of abstract deductive reasoning generally referred to as (Pure) Mathematics. It is characterized by a language in which abstract concepts are couched in terms of real world objects (e.g., counters, dams, queues, epidemics, rumors, expectations, experiments, observations, models, etc.) in which non-determinism is present.

Although the vocabulary is borrowed from real objects, they have no more relationship to reality than do 'bundles', 'sheafs', 'categories', 'lines' or 'groups' in other provinces of this heavenly domain.

To stress the motivation behind each definition I think I should wear the right hat and address myself appropriately. Let me therefore (pretend to) don my bishop's mitre and sign my name as Bishop Pyke. You notice there is an ecclesiastical tone to this definition. The theology represented probably espouses that there is some good ($\mathscr{AP}$) in every probabilist. However, there exists an uncharitable Pharisaic outlook on the part of many pious purists toward those who tarnish themselves with reality.



A PURIST'S DEFINITION

Applied Probability refers to a part of abstract deductive reasoning generally referred to as (Pure) Mathematics. It is characterized by a language in which abstract concepts are couched often in terms of real world objects in which non-determinism is present.

BISHOP PYKE

Figure 4

After preparing this definition, I realized that there could of course be more than one bishop, and undoubtedly a Pope (or Archbishop) overall. I wonder, would it be Pope Hardy? I suspect that Pope Hardy might have frowned on my liberality, and given instead

Definition A_2. ap (note the lower case print) is a vocational area of technical expertise that involves the application of known probabilistic techniques to actual problems, in much the same way as glassblowing is applied chemistry.

I think we will ignore this and return to Definition A_1.

In any religion there is a broad spectrum of ministries. Sermons depend on the audience. On the one hand there are sermons given to farmers[6] in rural

[6] As a farmer myself, I do not by this comparison criticize our intelligence.

areas that are couched in familiar and practical terms. On the other hand, you have intellectual discourses expounded from cathedral pulpits. Although the Purist's $\mathscr{AP}$ is part of the pure realm of Mathematics, does the algebraic topologist think of us as living in a rural area without the same philosophical maturity as themselves?

With this definition, I would say we have ample $\mathscr{AP}$ in AP, but with my bishop's mitre on, I should be alert lest the number of down-to-earth papers taint the journal's purity. Possibly the purist views all probability as 'applicable' if not 'applied'.

While I am still wearing my bishop's mitre, may I quote ten commandments which are essential guides for preserving the purity of $\mathscr{AP}$.

Ten Commandments for $\mathscr{AP}$

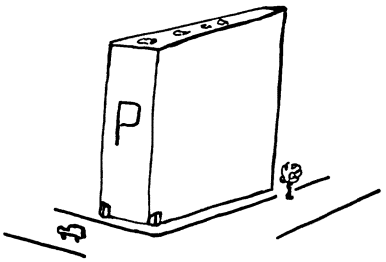
1. Love the Great Omega as your Ideal. Thou shalt have no other Ideals before this one.
2. Thou shalt not make nor serve any realistic models.
3. Thou shalt not take the name of $\mathscr{AP}$ in vain; use it only in the purity of abstraction.
4. Remember the sabbatical year to keep it holy, free from consultation on real problems.
5. Honor thy academic father by giving similar thesis problems to your academic children. Honor thy mother, Wisdom, that ye may be blest with long tenure.
6. Thou shalt not kill another's theory by exposing it to data.
7. Thou shalt not commit interdisciplinary adultery.
8. Thou shalt not plagiarize.
9. Thou shalt not bear false proofs.
10. Thou shalt not covet thy colleague's (probability) model no matter how beautiful or fertile.

These are unfortunately subscribed to by too many. How much better might be a paraphrase of the Great Commandment, 'Love thy neighbor sciences as your own!'

On the way to MacKenzie Hall, I noted a poem among the murals, all very nicely done, on the walls of the tunnels. The first four lines are:

Our life can grow in power
And happiness as it links
Itself productively to life
Other than our own.

B. *The Gentleman's Definition.* Let me put a different hat on, namely the gentleman's top-hat. The context here is of a discipline that lives in this world, in a Madison-Avenue type building. The building has no windows facing on reality, but receives its light through skylights in the roof. Whereas I quoted the ten commandments above, I should probably quote here from the *Wall Street Journal*: queueing theory—down two points; control theory—up one point; long-term prospects for growth in branching processes—excellent; and so on. The gentleman's definition, as given in Figure 5, is unfortunately the most common definition used by most us in this audience.



THE GENTLEMAN'S DEFINITION

Applied Probability is that part of mathematics which studies extensions, generalizations, simplifications, consequences of probabilistic results which have been used in some area of application during recent years.

R. PYKE, ESQ.

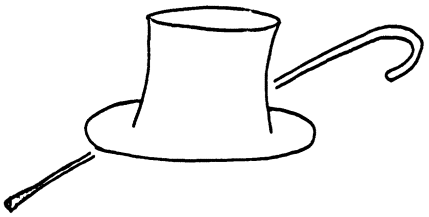


Figure 5

Definition B. $\mathscr{AP}$ is that part of mathematics which studies extensions, generalizations, simplifications and consequences of probabilistic results which have been used in some area of application during recent years.

Thus a gentleman looks upon $\mathscr{AP}$ as being concerned with problems that originated (albeit some time back!) with a real-life situation but he himself has

no direct contact with reality. Papers in this area are characterized by apologetic phrases like, 'These results have application to.∴' with a possible reference to the original paper that initiated the type of problem considered. These are materially, not idealistically, motivated which is even worse. Such definers of $\mathscr{AP}$ should get right with their ideal, and perhaps we will take time for confessions later.

There should probably be a pair of gloves in Figure 5 beside the top-hat, but I could not draw them very well. The gloves would indicate that gentleman $\mathscr{AP}$-ers are reluctant to get their hands dirty on real problems.

If you survey the program of the IMS meeting being held concurrently with this conference (The chairman of the Program Committee is here; I am sure he will forgive me) you find an unsatisfactory imbalance which I illustrate as follows. I tried to identify each paper on the program according to the categories $\mathscr{TS}$ (theoretical statistics), $\mathscr{AS}$, $\mathscr{TP}$ and $\mathscr{AP}$. I gave weights of 1, $\frac{1}{2}$ or 0 depending on how much I thought a paper pertained to each of these areas. I assigned the weight for units of fifteen minutes so that a forty-five-minute paper which is completely concerned with $\mathscr{AP}$ received a weight of 3 under that category and 0 under all others. I left out the Wald Lectures as that would bias the sample drastically. (It feels good to be a statistician!) The resulting weight are convicting:

$\mathscr{TS}$	$\mathscr{AS}$	$\mathscr{TP}$	$\mathscr{AP}$
49	1	38.5	2.5

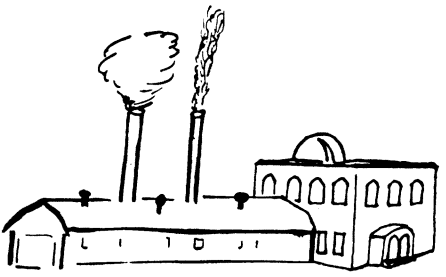
Here I have used the Gentleman's Definition since it would tend to increase the number of applied papers. (The Worker's Definition below would tend to make the case more deplorable.) One might note that the program at the recent Conference on Stochastic Processes and Applications contained also a very small number of applied papers.

C. *The Worker's Definition.* I now come to what I believe is the important definition of $\mathscr{AP}$. According to this definition, $\mathscr{AP}$ is located in working proximity to the problem mill but with telescopic access to the theoretical heavens. (Of course, the smoke in Figure 6 is clean, filtered and harmless.)

Definition C. $\mathscr{AP}$ is that part of the full coverage of man's knowledge that uses both deductive and inductive reasoning to describe or explain real phenomena in which randomness plays a basic role.

The hat I must wear during this discussion is of course the hard-hat. (A hard head is an acceptable alternative.) This definition conveys what I think is truly $\mathscr{AP}$. This is real science, working with colleagues from many disciplines on real

problems. This is where the action should be. If of course, a new aspect of theoretical probability develops from such a collaboration, so much the better. I claim this is the only way important new types of stochastic processes will arise.



THE WORKER'S DEFINITION

APPLIED PROBABILITY IS THAT PART OF THE FULL COVERAGE OF MAN'S KNOWLEDGE THAT USES BOTH DEDUCTIVE AND INDUCTIVE REASONING TO DESCRIBE OR EXPLAIN REAL PHENOMENA IN WHICH RANDOMNESS PLAYS A BASIC ROLE.

PYKE—12280 UW

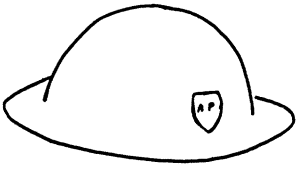


Figure 6

Let me give an example from my own career. As a fresh Ph.D., I was part of the Statistics Department at Stanford University. At that time there was an ONR grant on applied mathematics under which Professor Chernoff was responsible for consultation with the U.S. Naval Radiological Laboratories at Hunter's Point in San Francisco. We would on occasion drive to Hunter's Point for discussion with workers there about real problems. Professor Chernoff was my interpreter as he was uniquely capable of understanding a client's problem and translating it into the narrower vocabulary that someone like myself possessed. (Surely a major reason why more true $\mathscr{AP}$ is not in evidence is the difficulty that most of us have in communicating with persons in other disciplines.) Out of one particular problem at Hunter's Point came Markov renewal (or semi-Markov) Processes. The problem was one of identifying radioactive sources electronically, rather than chemically, by means of a multiple channel analyser. This instrument would record not only the total

number of radioactive impulses but would record the number of impulses according to their amplitudes. There were in effect 256 Geiger counters, each recording those impulses with amplitudes in a specified range. By comparing the resulting histogram with a library of theoretical frequency curves for each possible radioactive element, one could estimate the proportions of each element present in the sampled material. Out of this problem came my theory of and interest in Markov renewal processes. Following the original paper I, and others, went on to develop the theory further, possibly publishing along the way some papers of the type I have previously criticized.

Have these papers been used in other applied areas? I looked in the most recent Citation Index again under the listing 'Pyke', (honestly, this is the first time I have ever done this,) and I saw that there were a few papers listed. My 1961 papers on Markov renewal processes had ten citations. One was in *Eng. Cyber. Rev.*, one was in the *SIAM J.A.M.A.*, two were in *Biometrics*, one was in *Z. Wahrscheinlichkeitstheorie* and five were in J$\mathscr{AP}$. I found it rather rewarding to find here several references to papers appearing in applied journals. Isn't this the type of $\mathscr{AP}$ we should encourage?

At this time, if you are not yet asleep, may I ask you some questions? (That's the first one!) Let me read the titles and brief portions from some recently published papers. As I describe these papers would you decide for me whether they represent $\mathscr{AP}$, or not, and whether they could have been published in AP or not. By these examples I also wish to emphasize again the enormous scope of $\mathscr{AP}$ and to suggest that AP and even J$\mathscr{AP}$ cover only an extremely small portion of it. (At this point in the talk I quoted from the papers displayed in Figure 7 and briefly discussed their probabilistic contributions.) It is very easy to find large numbers of probabilistic papers like these in journals which are rarely referenced by most of us. The notation and vocabulary may differ from what we are used to, but the problems and results are often very interesting. The papers selected for this talk were chosen fairly randomly from the current periodicals shelf of the Science Library at York University last week. I encourage you to make regular checks through your own libraries or through lists of tables of contents.

Directions for research in $\mathscr{AP}$

What do you feel are the important directions for *research* in $\mathscr{AP}$? I think you could answer this as well as I. However, there is a simple answer. Find where quantiative science, social, natural or physical, is going today and one will find directions for $\mathscr{AP}$.

Medical research could be mentioned as only one of several important directions. Medical research has grown dramatically in the past decade, with

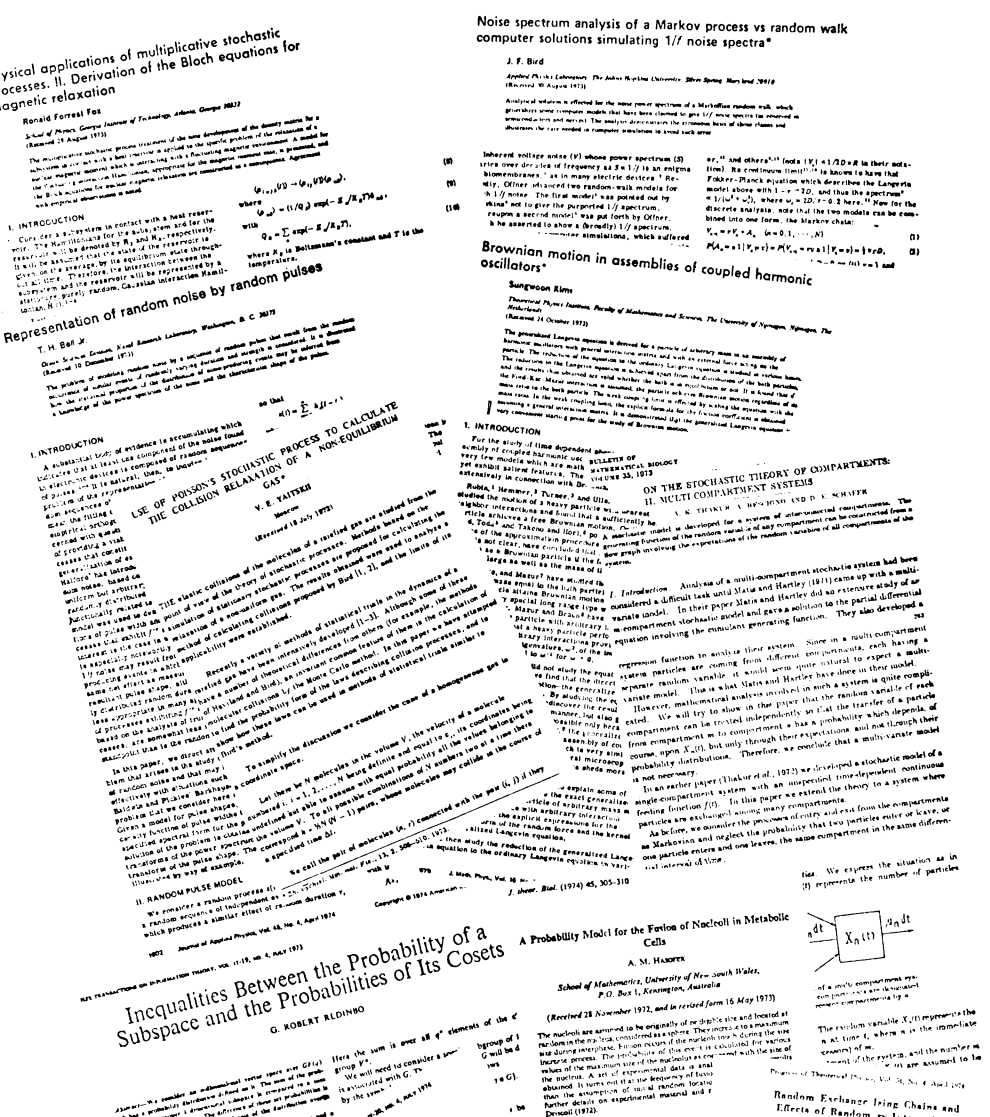


Figure 7

promise of continued growth ahead. Probabilistic models are needed in many basic areas of genetics, virology, epidemiology, microbiology, physiology,... More and more data for testing models are becoming available; for example, in clinical trials and from national surveys of comparative merits of surgical techniques and drugs.

I think some of the other speakers will give more specific examples, so rather than continue with suggested areas of applications and directions, let me return in conclusion to my basic dilemma:

What is $\mathscr{AP}$? Where is $\mathscr{AP}$? How to foster $\mathscr{AP}$?

I have already addressed the first two questions. I now suggest some steps for resolving the third.

How to foster $\mathscr{AP}$?

The following are but just a few of the possible ways of fostering $\mathscr{AP}$.

A. Break down walls — departmental walls, societal walls, journal walls, prejudicial walls. All such walls encourage narrowness and isolation; they certainly cannot foster the $\mathscr{AP}$ of Definition C. Work with colleagues in other fields; new theory will come from applications. Encourage students to do the same. (At York University I mentioned something I have been doing for several years though it is not original with me. When a student comes up for his general Ph.D. examination, I pose a general question of the type, 'Tell me three important breakthroughs in science that have occurred in the past ten years.' The answer most often is silence. Are we encouraging our students, let alone ourselves, to be as broad as possible in their general knowledge of science? After all, the demand for Ph.D.'s has fallen in many areas. We can now approach our production of Ph.D.'s more leisurely. Let us make sure that the ones we produce are genuinely excited about science as a whole.)

B. Test the fit of your models. If you are wearing that gentleman's top-hat, at least go out and get some data and use that sub-area of $\mathscr{AP}$ called $\mathscr{AS}$.

C. Study harder problems: Robustness of models, transient behavior of processes, asymptotic expansions and rates of convergence. Make use of the computer and algorithms in place of 'double generating functions' or such. Go after practical solutions to harder questions, not just explicit answers to simplified models.

D. Publish only useful (in a broad sense) papers. Use the 'Abstracts' section of the *Bulletin of the IMS* for statements of many results; if the proof has nothing new in it, the statement is all that is needed and you probably do not need an extra paper in your bibliography. The reader of the abstract will supply the proof if he doubts the results, and write to the author if he has difficulty.

E. In journals, AP included, editors should

i) consider publishing reviews, abstracts, titles of interesting papers from other disciplines:

ii) raise standards concerning the applicability of published papers;

iii) work for breadth.

F. In societies, IMS included, be open, encourage other areas and sponsor interdisciplinary activities. If there is a group of people desiring a symposium on a special topic, get behind and push it. Small groups meeting on special topics within $\mathscr{AP}$ can be of great use and lasting value, particularly if Definition C is understood.

G. In classrooms; this is where there is a most important need for change. Check your curricula. Do not hesitate to change courses, texts and approaches. (By the way, if you want a very poor definition of $\mathscr{AP}$ consider the tables of contents of many of the recent books which have $\mathscr{AP}$ in their titles.) Use computers for data and for sample 'omegas' of stochastic processes. Augment and motivate lectures with examples of their applicability. Do not compromise our students' need for strength in mathematics and in theoretical $\mathscr{P}$ and $\mathscr{S}$. However they need much more. Ph.D.'s may take longer, as a certain amount of apprenticeship as well as theory is involved. However, it should be well worth it.

In closing, let me break one of the earlier commandments and plagiarize from someone not present.

> $\mathscr{AP}$ or not $\mathscr{AP}$, that is the question
> Whether 'tis nobler in the mind to suffer
> The fame and tenure from outrageous models,
> Or to take arms against a sea of problems
> And by solving, end them. To try $\mathscr{AP}$ —
> Ah more, and by $\mathscr{AP}$ to say we solve
> For man's sake the thousand natural shocks
> Of life and air too; 'tis a consummation
> Devoutly to be wished to try $\mathscr{AP}$!
> $\mathscr{AP}$, perchance a dream, ay there's the rub,
> For does $\mathscr{AP}$, a breath of dreams become,
> When we have shuffled off this abstract soil?
> Trust in the cause — there's the respect.
> $\mathscr{P}$ makes calamity if not applied.
> For who would bear the whips and scorns of time?
> The Professor's wrong, who would problems spare,
> That grunt and sweat in real life form.

Or that give the thread to answers, useful links,
An undiscovered theory, from whose home
No worker returns without the will
To seek an answer for those ills we have,
To apply to others that we know not of.
This conscience does make cowards of us all.
Let not the native hue of resolution,
Be sicklied o'er with the pale cast of doubt,
So that enterprises of great pitch and moment,
With disregard for current needs and pleas,
Lose the name of action.

If the author of the original version of the above were here, I fear he might say, on behalf of the pure probabilist:

Get thee to a nunnery:
Why would'st thou be a breeder of sinners?
I am myself indifferent honest.

One feature of an introspective discussion like this should be the resultant spirit of confession. I noticed such a spirit at the Conference last week at York University, where several contributors 'confessed' to having published a paper of such a type (e.g., an unrealistic generalization of a generalization of ...) and pledged not to do it again. I am therefore tempted to conclude with an old-fashioned invitation: While smoothly filtered pure white noise plays in the background, may I challenge you with the words:

Just as I am, without $\mathscr{AP}$,
Now that a flood of needs I see.
For theory with utility;
Not words born in futility
I come ...

Benediction.[7] In the name of (the Ω, the $\mathscr{A}$, and the $\mathscr{P}$): AP.

[7] Suggested by R. J. Griego following the talk.

Suppl. Adv. Appl. Prob. 7, 38–49 (1975)
Printed in Israel

THEORY AND PRACTICE IN APPLIED PROBABILITY

J. GANI,* *University of Sheffield*

1. Introduction

Professors Marc Kac and Ron Pyke have very ably and entertainingly presented you with several points which I had also intended to underline in this paper. I propose to reinforce their statements by discussing some examples of problems in applied probability, and putting forward a few conclusions based on my personal experience.

From this outline of my views, two basic themes will emerge. The first and most important is the unity of theory and applications in probability; you will undoubtedly note both the close interrelation and the delicate counterpoint of these two facets of our subject. The second subsidiary theme concerns communication between theoretical and applied probabilists, and the dissemination of information within our field.

Let me begin by remarking that most phenomena in real life, and in the biological, physical, social or technological sciences, have a large random component. It would thus seem natural to use probabilistic methods to solve a number of the problems arising in these fields. My understanding of applied probability is that it consists of the application of the theory and calculus of probability to the solution of problems in these, as well as other areas. I can think of no better way to substantiate this point than by providing you with four examples of such problems.

1.1 *A problem in virology: the random covering of a sphere*

I first came across this problem at the Australian National University in 1961. There was then a strong team of virologists, among them John Cairns and Stephen Fazekas de St Groth, who were working on the influenza virus (type A) at the School of Medical Research. This roughly spherical virus, of radius 40 mμ, attaches itself to the surface of healthy blood cells, which are assumed to be flat, and infects them. One of the methods suggested for preventing infection was to increase the number of antibodies in the bloodstream; these cylindrical

*Now at CSIRO Division of Mathematics and Statistics, Canberra.

cigar-shaped bodies, 27 mμ in length, by attaching themselves normally to the virus surface, would protect a spherical cap from contact with any cell (see Figure 1).

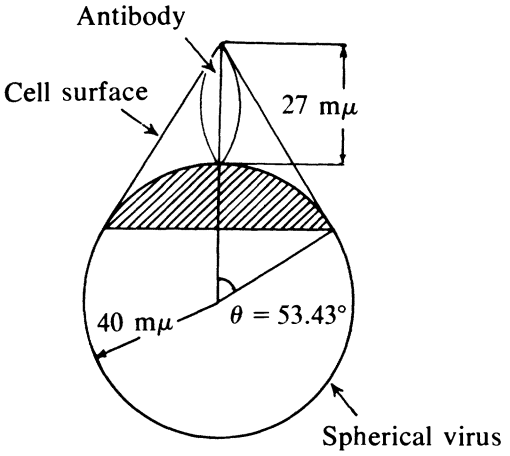


Figure 1
Influenza virus and antibody

Looked at concretely, the problem was to ensure that a sufficient number i of antibodies, randomly distributed over the sphere, would protect it entirely from contact with a cell. In probabilistic terms, given i antibodies distributed at random on a sphere, what was the chance that the i spherical caps subtended by them would cover the sphere entirely? This is a very complicated problem of geometrical probability; for the angle $\theta = 90°$, the exact probability $P(i)$ of coverage by i antibodies is

$$P(i) = 1 - \frac{i^2 - i + 2}{2^i} \qquad (i = 1, 2, \cdots)$$

(see Gilbert (1965)). When $\theta < 90°$, as in the case of the virus, for which $\theta = 53.43°$, Moran and Fazekas (1962) were able to show that for large i

$$P(i) \sim \exp\left[-\tfrac{1}{2}\pi^2 \left(\left\{ 2[\tfrac{1}{2}(1+\cos\theta)]^i \left[1 + \frac{i^2}{\pi^2} \tan^2 \tfrac{1}{2}\theta \right] \right\}^{-1} - 1 \right)^{-1} \right];$$

they went on to verify this formula by a simple Monte Carlo experiment.

Under certain simplifying assumptions, the approximate probability $P(n_0, n_1, \cdots, n_s; t)$ that x_0 antibodies distribute themselves over n_i ($i = 0, 1, \cdots, s$; $\Sigma_{i=0}^{s} n_i = N$) virus particles having i attachments, where s is the maximum number of possible attachment sites, has been obtained (see Gani (1971) for a review). Morgan (1971) has shown how, in some cases, exact values can be found for this probability. Assuming independence of the virus particles, one can then derive the useful probability

$$Q(t) = \sum_{n_i} P(n_0, n_1, \cdots, n_s; t) P(1)^{n_1} \cdots P(s)^{n_s}$$

of non-infectivity, that is the probability that the virus does not infect healthy cells.

1.2 *A problem of directional navigation*

In a paper recently read to the Royal Statistical Society, David Kendall (1974) has investigated some interesting models of bird navigation. Ornithologists have long known that birds have a homing propensity, and have often wondered how it operates. It is now thought that birds may navigate by reference to the sun and the stars in some instinctive fashion; Kendall has proceeded to construct two mathematical models which would simulate bird navigation fairly realistically.

The first is referred to as the Manx model, after the Manx Shearwater, which flies across the Atlantic to its breeding grounds in Europe. The second Bessel model has a similar formulation but with steps of random rather than fixed size. The bird is assumed to set off from a distance of several hundreds (or thousands) of miles, a not uncommonly long journey. It tries to point in the direction it believes to be home, but many commit an error when setting off: this may be an error of navigation, or it may be a deflection due to head winds or other such causes.

This directional error may follow either of the two circular distributions, the von Mises VM(k) or the wrapped normal WN(k) with parameter k; for all practical purposes, these are almost indistinguishable. It may be relevant to point out here that Hartman and Watson (1974) have proved that for each $k > 0$, there exists a probability measure μ_k on the Borel sets of $[0, \infty)$ such that

$$\mathrm{VM}(k) = \int_0^\infty \mathrm{WN}(\lambda) \mu_k(d\lambda).$$

This is an attractive theoretical result relating two distributions, whose consideration has been dictated by very practical reasons.

Manx motion can be very simply described: a bird flies fixed laps of approximately 20 miles each, at a speed of roughly 40 m.p.h. At the end of each

lap, it redirects itself towards its goal, but commits an angular error having one of the above circular distributions. When the bird finally arrives to within a radius of approximately 10 miles of its home, it recognizes its destination and heads directly for it. In Bessel motion, the lap length, as well as the deflection from the correct direction, is assumed to be random. Under appropriate conditions, both converge to motions of the Browian type.

Kendall's paper contains a good deal of data, subtle mathematics, diagrams of simulations of both Manx and Bessel motions, and some interesting probabilistic results. Not satisfied with these, Kendall remarks, 'As usual in applied mathematics, all the mathematician can do is to guide the practical man towards the point at which the really difficult thinking and experimentation has to begin'.

This modest comment may help to explain why journal editors have such difficulty in attracting similar papers of high excellence in applied probability. Experimental data takes time to collect, models require careful development and verification, and all this only to act as guides to the practical scientists; how much more simple to write the largely self-contained mathematical essays with which we are so familiar.

1.3 *The analysis of type counts in literary texts*

Applications of probability in the social sciences are rarely found in the periodicals read by applied probabilists; social scientists usually publish their probabilistic models in specialized journals of their own. There is, however, the occasional exception; two years ago, a paper by Brainerd (1972) appeared in the *Journal of Applied Probability* on type and token counts in literary texts. One of its main interests was the size of Shakespeare's working vocabulary; this was derived from the relationship between the count of previously unused words (types) in a given count of all words (tokens) in his plays. The problem can be formulated as one of random sampling with replacement from a word pool, in this case, Shakespeare's vocabulary.

Brainerd postulated a non-homogeneous Markov chain process for the type count X_n in a given token count n, such that

$$\Pr\{X_{n+1} = i + 1 \mid X_n = i\} = f(n, i) \qquad (i \leqq n)$$

where $\Pr\{X_1 = 1\} = 1$. When $f(n, i)$ was simplified to the form $g(n)$, he was able to obtain the p.g.f.

$$G_n(s) = \sum_{i=1}^{n} \Pr\{X_n = i\} s^i = \prod_{j=0}^{n-1} \{1 + (s-1)g(j)\} \qquad (|s| \leqq 1);$$

from this the mean and variance of the type count for n tokens were derived.

Brainerd initially estimated Shakespeare's vocabulary on the assumption that $g(n) = e^{-\alpha n}$. But why Shakespeare, you may ask; the reason is that his towering position in English literature has led to the compilation of several concordances of his works, beginning with Mrs Cowden Clarke's published in 1845. It is known from this that Shakespeare's plays total about 310,000 words, of which approximately 30,000 are different; assuming that Shakespeare did not use his entire vocabulary in his plays, its total size could be expected to be somewhat larger than this figure. Brainerd refined his model by subdividing vocabulary into the larger Group I of substantive words, and the small Group II of auxiliary words numbering approximately 500. But the fit of his model to the data remained somewhat unsatisfactory.

In a subsequent paper McNeil (1973) proposed an alternative model in which $f(n, i) = \alpha(M - i)$, M being the total vocabulary and α some suitable parameter ($\alpha \leqq 1/M$). In this, the probability of a new word was proportional to the number of words as yet unused. I have recently become interested, together with I. Saunders (1975), in refinements of this model in which the vocabulary is broken up into Group I and Group II words. The refined model fits Shakespeare's data very well, particularly when the parameters are varied slightly for the comedies, tragedies and histories. However excellent the fit, we must accept the model with scepticism, since it is clear that no author selects words from his vocabulary pool at random; syntax and the direction of his thoughts dictate his choice. We conjecture that this structuring of language does not seriously affect the results of our modelling.

1.4 *Problems of water storage*

The probabilistic theory of water storage was developed by Moran in 1952–3 and published in 1954 as a result of problems which had arisen in the construction of dams in the Australian Snowy Mountains Hydroelectric Scheme. Very roughly, when constructing a dam, an engineer will wish to know, given the distribution of inputs and the required annual water release, what dam content will result in the water's running dry less than once in a hundred years.

In more technical terms, if K is the capacity of the dam, $\{X_t\}$ the set of annual imputs during the years $(t, t+1)$, M the annual release occurring at the end of each year, and $\{Z_t\}$ the set of dam contents after this release, then

$$Z_{t+1} = \min\{Z_t + X_t, K\} - \min\{Z_t + X_t, M\} \qquad (t = 0, 1, \cdots).$$

It is readily seen that if the $\{X_t\}$ are assumed independent, $\{Z_t\}$ forms a Markov chain for which the transition probability matrix can be very simply written.

The stationary probability distribution of $\{Z_t\}$ can be found, and from it the probability P_0 of the dam's running dry; this will depend on K.

On altering the value of K in the model, one can obtain by a sequence of approximations that particular one which will make $P_0 \leqq 0.01$, as required. The general theory of storage for discrete inputs is very similar to that for queueing, but it becomes quite distinct when continuous inputs are considered. The methods of the Moran model have come into common usage by engineers, particularly since the WRA Reservoir Yield Symposium held at Oxford in 1965. Since inputs are serially correlated, Lloyd (1963) suggested a refinement in which the $\{X_t\}$ form a Markov chain; in this case $\{Z_t, X_t\}$ define a bivariate Markov chain, and methods similar to those for the Moran model continue to apply. In subsequent reviews Gani (1969) and Lloyd (1974) have described the further development of storage theory and its applications.

An elegant result in this theory, which holds for the case of independent $\{X_t\}$, is originally due to Kendall (1957). It concerns the probability of first emptiness of the dam, and states that:

$$\Pr\{Z_t = 0 \text{ for first time at } t = T \mid Z_0 = u\} = \frac{u}{T} p^{(T)}_{T-u}$$

where $p^{(T)}_{T-u} = \Pr\{X_1 + \cdots + X_T = T - u\}$. A similar result also holds when the $\{X_t\}$ form a Markov chain.

I have now given a very brief outline of four probability models in each of the biological, physical, social and technological fields. In every case, the relationship between theory and application has been close, each suggesting further developments in the other area. The counterpoint of theory and practice is complex, not easily categorized, but very much in evidence. There is neither contradiction nor competition between these two. Sometimes particular aspects of an application will lead to a rewarding theory; not infrequently a theoretical result will find useful application in a very concrete problem. Each area aids and advances the other; neither is dominant, and neither can be forgotten in an objective appraisal of our field. Let me now turn to my secondary theme: communication in the field of applied probability.

2. Criteria for papers in applied probability

One might expect that work in applied probability would proceed from the collection of data to inference, estimation and the building of probability models from which practical conclusions could be drawn. These might then be compared with the data in order to verify the validity of the model. Regrettably few of the papers written in applied probability are of this type; one always responds with pleasure to an author who tackles a real problem containing hard

data and builds a theoretical model to explain it. The majority of contributions are essentially concerned with the construction of models which are never validated by reference to real data.

But instead of commenting negatively on these, let me state in a more positive vein that David Kendall's previously quoted paper on bird navigation models fulfils most of the requirements which I would expect of a paper in applied probability. He has clearly made himself thoroughly conversant with the biological literature in bird navigation, having discussed it with, among others, his friend, Dr. D. L. Lack, the ornithologist, to whose memory the work is dedicated.

Kendall is not averse to explaining in detail the background of the problem, nor is he slow in seeking and analysing data on which to base his hypotheses. He proceeds to test these by repeated simulations which lead to graphical representations of Manx flight. Having obtained plausible results, he returns to the mathematics of the problem, compares Manx and Bessel motions theoretically and numerically, finally proceeding to some diffusion approximations. He concludes with a lengthly study of the hitting times to the circumference of the homing target, and with further simulations relating to Manx motion.

Such a thorough investigation of a scientific problem is rare. A far commoner approach consists of probabilistic model-building almost for its own sake. A well-established model may at one time have been developed in a given area, say population studies, to answer a genuine scientific question. A subsequent author, sensing that the modification of certain conditions will lead to a neat mathematical solution, will write a paper on the modified model. He is then followed by others who make a cottage industry of such minor modifications.

It would be foolish to suggest that some variants of the original model are not legitimate or valuable, but there comes a point where these are reduced to classroom exercises lacking any scientific content. It is these we should try to discourage, these whose publication an editor will endeavour to avoid. How delightful it would be if papers in applied probability dealt with real scientific problems, and went some way towards their solution.

To avoid any false impressions, let me hasten to add that I greatly admire those probabilists whose main concern is not the solution of applied problems, but rather the development of elegant and deep theories in the field of probability. What I am trying to isolate are the rarely defined qualities of spuriousness, imitativeness and artificial model-making, which lead to the sad waste of so much technical ability. If I could be distinctly more encouraging, might I suggest to research workers with high technical skills that they devote their talents to the solution of problems worthy of their mettle.

But let me leave here this difficult discussion of scientific taste, and continue with the history of applied probability.

3. The Applied Probability journals: their origin and development

The term 'Applied Probability' was coined on the occasion of a symposium held by the American Mathematical Society in 1955; its proceedings were published in a volume of that title. For my own part, I discovered the term through the Methuen Monographs in Applied Probability and Statistics edited since 1959 by Professor M. S. Bartlett. He has expressed the opinion that neither field could exist without the other, and I tend to agree with him.

In the late 1950's I had just completed my postgraduate training with Professor P. A. P. Moran in Canberra, and was being initiated into the mysteries of writing research papers, and of corresponding about these with editors of learned journals. The field of applied probability had not, as yet, been clearly defined and research workers would not infrequently find that a paper sent to a statistical journal would be returned with a polite note stating 'We regret that we are a statistical journal; your paper would seem to be more appropriate for a mathematical periodical'. An established mathematical journal would later return the paper with an equally polite note saying 'We regret that we are a mathematical journal; the referee was uncertain as to whether your paper contained any original mathematical developments'.

There were some journals, such as the *Proceedings of the Cambridge Philosophical Society*, whose wide mathematical coverage and catholic tastes allowed them to accept original papers in applied probability, as in other fields. But such periodicals were few; there seemed to be a problem in search of a solution. In 1962, I went on sabbatical leave from the Australian National University, and visited Britain and the USA. My colleagues in both countries encouraged me in the belief that an avenue of publication for applications of probability theory was desirable; with their cooperation, I decided to act.

An editorial board was formed; with the financial assistance of the London Mathematical Society, and later the University of Sheffield, the Trustees of the Applied Probability Trust launched their first journal in 1964. The first volume of the *Journal of Applied Probability* had 396 pages and about as many subscribers, but it has never looked back; *Advances in Applied Probability* followed in 1969 and *Mathematical Spectrum*, a student magazine, in 1968–9. These have been self-sustaining for many years, and are most efficiently run from the Sheffield office of the Trust by our Executive Editor, Miss Mavis Hitchcock.

I would hardly wish the titles of the Applied Probability journals to canonize the name of the field; quite the contrary. The name was selected because it seemed to define what a sizeable group of probabilists was writing about at the time; but one would not wish to freeze the situation permanently, or imply that there was mystical applied probability approach to problems. Applied probabil-

ity must be recognized as the small subfield of mathematics which it is. In practice, the journals are concerned with papers on the building of probabilistic models in such areas as population processes, mathematical genetics, epidemiology and other biological processes, operational research, queueing theory, storage and traffic theory. There are fewer contributions forthcoming on problems in the social sciences, and only occasional ones in the physical sciences, although statistical mechanics and probabilistic quantum mechanics remain active fields of research. Some papers are concerned with the play of ideas and the construction of models in their own right; of these some fall short of real science, and detract from the contribution which applied probability could make to it.

While this is the present situation, the aim of the Applied Probability journals goes even further: it is to bring together accounts of probabilistic methods used in the solution of problems in every possible field of science. But I should admit that we have difficulty in attracting papers from physicists, social scientists, engineers, psychologists, medical scientists, all of whom tend to publish in their own more specialised periodicals. While we have tried to persuade them to write the occasional review paper in order to keep our readers in touch with the realities of practical problems in these important fields of research, we have not so far proved successful in doing so.

4. The spread of the Applied Probability journals

The Editorial Office of the Applied Probability journals is developing the practice of publishing Journal Index volumes regularly; two have been produced over the past ten years. The first *Complete Author and Subject Index*, which appeared in 1971, covered Volumes 1–7 of the *Journal* (JAP); the second *Index*, published in 1974, analysed Volumes 8–10 of the *Journal* and Volumes 1–5 of *Advances* (AAP). The opportunity has been taken in each *Index* to survey the field of applied probability, list the numbers of papers in each of its subsections, and consider significant trends in the research interests of authors.

The histograms in Figure 2 will give some idea of the number of papers published in 18 subsections of our field during 1964–70 and 1971–73 respectively. You will note that roughly as many papers appeared in the last three years of the decade as in the first seven, so that the volume of papers published (let alone submitted) had approximately doubled by the end of this period.

When χ^2 tests of the contingency table type were carried out on all the data, the research interests appeared to have been significantly redistributed. But when the relatively small subsections in biological, social and physical applications were removed, χ^2 no longer proved significant. The most important change was recorded in the biological subsection, where it seems likely

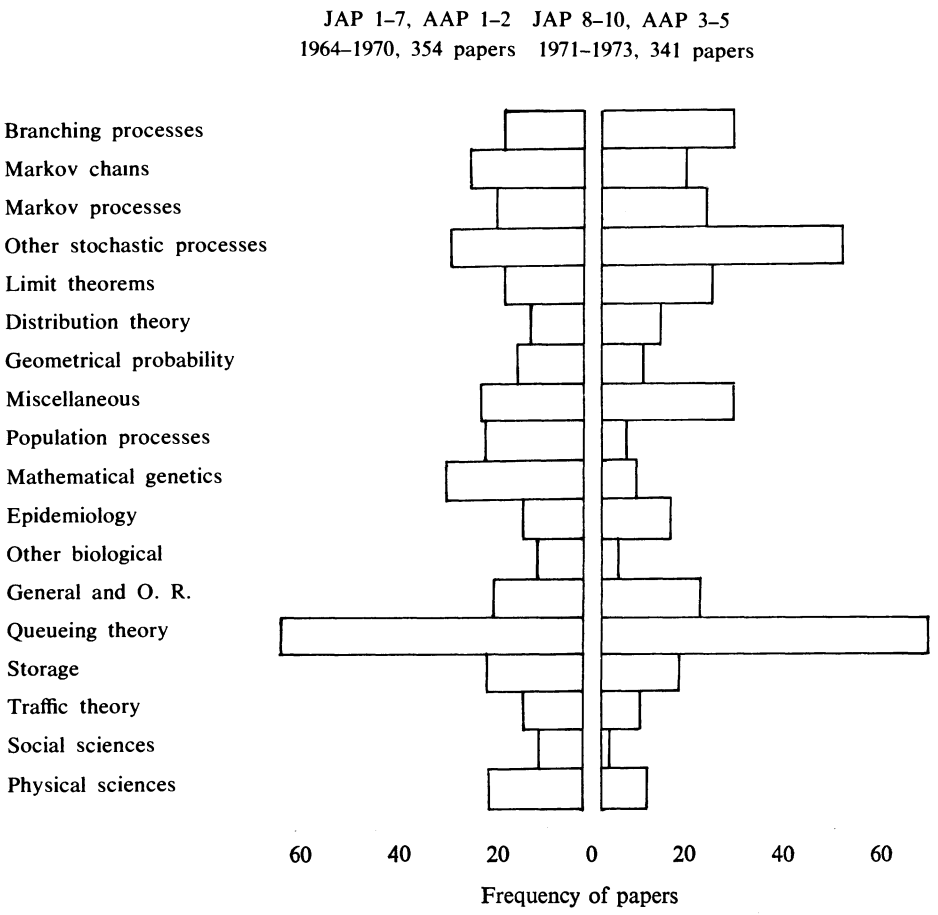


Figure 2

Histograms for numbers of papers published in the Applied Probability journals, 1964–1973

that new journals such as *Theoretical Population Biology* initially published in 1970, and *Mathematical Biosciences* from 1967 have drawn off potential submissions in population processes and mathematical genetics. While welcoming the important contributions made by these journals of high standing, I regret that we are not able to attract more research papers in the biological area, as I believe that important probabilistic developments are likely to take place in it.

Research interests have remained centred on three main areas: the general theory of stochastic processes, population processes and mathematical genetics, and queueing theory and allied fields. Between 1964 and 1970, approximately 41% of the papers in the Applied Probability journals were concerned with probability theory, but between 1971 and 1973 this proportion rose to 54%.

I must admit that my original aim had been to devote two-thirds of the journals to applications, and only one-third to theory having a basis in applied problems; in this I have been unsuccessful. But I should at least wish to achieve a balance of 50% applied and 50% theoretical material; if this were not so, the journals would deny both their origin and their function.

It is, of course, entirely consistent with the current trends in our field that an increasing amount of effort should be expended on what one might call the theory of applied probability. This deals with the mathematical and probabilistic implications of stochastic models. Research has been carried out in branching processes, Markov chains and processes, as well as more general stochastic processes (point processes, Brownian motion, time series and prediction). There is also a fair amount of work in limit theorems, distribution theory and geometrical probability. Much of this effort is valuable, but, regrettably, some consists of marginal improvements on previous results and the re-proving of theorems under less stringent conditions.

I have now commented sufficiently on various aspects of our field; let me try to give a brief summary of my personal viewpoint.

5. A personal view of applied probability

The first point I should like to stress is the breadth of our subject: it encompasses many real life problems and every scientific field including the biological, physical and social sciences, engineering and technology. Applied probability feeds on practical problems, but requires a high level of theoretical competence in probability. It would be difficult, if not impossible, for any journal or set of journals to collect together the diverse strands of the subject, but the effort is nevertheless worth making.

The second point is one of method. In solving applied probability problems, all approaches are useful: eclecticism is a positive asset. Classical mathematical analysis, numerical methods, statistical calculations, probabilistic limit theorems, simulation and every other branch of mathematics are legitimate weapons in the search for a solution. Applied probability is a small branch of mathematics, and must not hesitate to draw on the resources of its parent tree. Nor must it tie itself to any school or tradition, whether it be British empiricism, French abstractionism or North American theoreticism; its strength lies in the universality of its traditions, and the versatility of its mathematical methods.

The third point concerns the delicate interrelation of theory and practice in applied probability. There is a subtle counterpoint between the two, and no easy formula can be prescribed for their correct balance. But they are closely interlocked, the two facets supporting rather than competing with each other.

Without practice, applied probability is trivial; without theory it becomes shallow.

My final point is that close contact with experiment and reality is essential to the healthy development of the subject: the collection and analysis of data cannot be avoided. In attacking each problem, there should be a complete cycle from the examination of data to the development of a theoretical model; this should be followed by the statistical verification of the model, and its subsequent refinement in the light of its goodness of fit.

My personal view, and I must stress that it is purely personal, is that too much of our effort may have been diverted into the game of model building for its own sake, as well as in following through the mathematical implications of new models. I believe that we can only achieve depth in applied probability by considering real life problems and by validating our models on the basis of real data. I speak with some experience of the complexities of biological problems on which I have worked recently. My feeling is that only by paying close attention to practical data, and considering genuine scientific problems, can applied probability achieve its full stature.

Though perhaps a little subjective, I hope that these remarks will have proved of some interest to you as workers in applied probability. If so, I shall feel that my experience may have helped the growth of our very lively field of research.

References

BRAINERD, B. (1972) On the relation between types and tokens in literary text. *J. Appl. Prob.* **9,** 507–518.

CLARKE, MARY COWDEN (1845) *Complete Concordance to Shakespeare.* Knight & Co., London.

GANI, J. (1969) Recent advances in storage and flooding theory. *Adv. Appl. Prob.* **1,** 90–110.

GANI, J. (1971) Some attachment models arising in virus populations. *Statistical Ecology,* (Ed. G.P. Patil, E.C. Pielou and W.E. Waters) Vol. 2, pp. 49–86. Pennsylvania State University Press.

GANI, J. AND SAUNDERS, I. (1975) Some vocabulary studies of literary texts. *Proc. Mahalanobis Memorial Meeting,* Calcutta, 1974.

GILBERT, E. N. (1965) The probability of covering a sphere with N circular caps. *Biometrika* **52,** 323–330.

HARTMAN, P. AND WATSON, G. S. (1974) "Normal" distribution functions on spheres and the modified Bessel function $I_\nu(x)$ (to appear).

KENDALL, D. G. (1957) Some problems in the theory of dams. *J. R. Statist. Soc.* B **19,** 207–212.

KENDALL, D. G. (1974) Pole-seeking Brownian motion and bird navigation. *J. R. Statist. Soc.* B **36,** 365–402.

LLOYD, E. H. (1963) Reservoirs with serially correlated inflows. *Technometrics* **5,** 85–93.

LLOYD, E. H. (1974) Wet and dry water: Remarks on pure and applied mathematics as illustrated by the recent history of stochastic reservoir theory. *Bull. I.M.A.* **10,** 348–353.

MCNEIL, D. R. (1973) Estimating an author's vocabulary. *J. Amer. Statist. Assoc.* **68,** 92–96.

MORAN, P. A. P. (1954) A probability theory of dams and storage systems. *Aust. J. Appl. Sci.* **5,** 116–124.

MORAN, P. A. P. AND FAZEKAS DE ST GROTH, S. (1962) Random circles on a sphere. *Biometrika* **49,** 384–396.

MORGAN, B. J. T. (1971) On the solution of differential equations arising in some attachment models of virology. *J. Appl. Prob.* **8,** 215–221.

Suppl. Adv. Appl. Prob. 7, 50–61 (1975)
Printed in Israel

SOME PROBLEMS OF SAMPLE SURVEYS

C. RADHAKRISHNA RAO, *Indian Statistical Institute*

1. Introduction

India is one of the developing countries which adopted and uses sample surveys on a large scale and in a continuing way for collection of statistics for administrative purposes. These surveys are carried out by a department of the Government called the N.S.S. (National Sample Survey) which employs about 1000 field investigators and 1200 technical and managerial staff. The N.S.S. engages itself in an enormous task of collecting data on a wide variety of subjects in several rounds during a year, each round covering a period of three to four months, and tabulating data as desired by the users. The tabulation department of the N.S.S. produces about four estimates every minute round the clock on some aspect or other of the socio-economic and demographic situation in the country (the only comparable activity in India seems to be sixteen or so babies born every minute). In my talk today, I shall mention some of the practical problems encountered in conducting large-scale sample surveys like the N.S.S. I would like to thank Professor S. G. Ghurye for giving me this opportunity.

Before doing so, I shall briefly touch upon some theoretical aspects of sample surveys, partly because of their intrinsic interest in the context of statistical inference about finite populations and partly because of the current controversies sparked off by some issues raised by Godambe ((1955), (1965), (1974)), which seem to remain unresolved till today. The issues involved belong to the domain of inductive inference by which new knowledge is created, and in such processes controversies are inevitable. However, many of the controversies could be avoided if the basic issues are clearly stated and statisticians do not insist on a monolithic structure for all problems of statistical inference. Much damage has been done by fashions and slogans in statistics introduced by theoretical statisticians who have no experience of handling live data and extracting information from them.

2. Statistical setup of sample survey problem

2.1 *The frame*

We consider a space F of identifiable *units* (elements), called the population to be surveyed. F may be a discrete space such as a list of towns, households or

individuals or a continuous space such as area under a particular crop, etc. It is useful to define a σ-field of Borel sets B_F in F and also consider a natural measure P_F defined on B_F. We call the triplet

$$(F, B_F, P_F) \tag{2.1}$$

the *frame* or the *label space.*

Associated with each unit $f \in F$, there is a *known* real-valued vector $c(f)$ called vector of concomitant measurements and an *unknown* real valued vector $\theta(f)$ called vector of parameters (or variate values). The function $\theta(\cdot)$ is assumed to be B_F-measurable. Also associated with each f there is a real vector-valued random variable $V(f)$ whose distribution is specified by $\theta(f)$. It is possible to take observations on $V(f)$ for values of $f \in s$, any chosen subset of F called a *sample of units.* The observed data including the concomitant measurements may be represented by

$$s^* = (s, \{v(f): f \in s\}, \{c(f): f \in F\}) \tag{2.2}$$

where $v(f)$ denotes all the observations taken on $V(f)$. Based on s^* we are required to draw inference on the unknown function $\theta(\cdot)$ defined on F. More specifically, the types of problems encountered in practice may be classified as follows:

(a) *Mapping*, i.e., providing an estimate of $\theta(f)$ for each $f \in F$.

(b) *Estimating* certain characteristics such as
   (i) $\int \theta \, dP_F$, i.e., the mean value of θ, and
   (ii) $\max_{f \in F} \theta(f)$, i.e., locating the maximum where θ is a scalar function

(c) *Testing* hypotheses such as

$$\int \theta_i dP_F = \int \theta_j dP_F$$

where θ_i and θ_j are the ith and jth components of θ.

Note 1. The concept of a frame is implicit in all sample survey problems and the term frame, I believe, was first proposed by R. A. Fisher. The need for a frame to clearly define parameters under estimation is also emphasised in a recent paper by Särndal (1974)

Note 2. Mapping is important in estimating, for instance, underground reserves of coal, minerals, etc.. The subject of sampling for mapping purposes is not very well developed.

Note 3. Problem (c) mentioned above occurs in the design of experiments where $\theta_i(f)$ and $\theta_j(f)$ may represent the yields of two treatments i and j applied on unit f.

2.2 *Sample Survey Design (SSD)*

In Section 2.1, we considered a chosen set $s \in F$ called a sample of units. We shall also associate with each $f \in s$ an integer n_{fs}, which represents the number of observations to be taken on the random variable $V(f)$. Without complicating the notation, let us represent by s not only the units in it but also the numbers n_{fs} attached to the units. Let $S = \{s\}$ be a specified set (or space) of samples (of units), B_s be a σ-field of Borel sets in S, and P_s be a probability measure defined on B_s. The measure P_s provides the probability or probability density for an element $s \in S$. We shall call the triplet

$$(S, B_s, P_s) \tag{2.3}$$

a Sample Survey Design (SSD), which is fundamental in all sample surveys and which has been specially emphasised by Godambe ((1955), (1965)).

Note 4. Some authors of sample surveys consider a sample of units not as a subset of F, but as an ordered set of elements of F allowing repetitions. If a single observation is taken on each unit in such a sample, there is no *essential difference* if we consider the sample as a set of distinct units and take a number of observations on each unit equal to the number of times the unit is repeated in the original sample. We shall adopt the latter convention which is already considered in defining an SSD.

2.3 *Kolmogorov setup*

A sample consists of an element $s \in S$ and the values $v(f)$ observed on each $f \in s$ as indicated in Section 2.2,

$$s^* = (s, \{v(f): f \in s\}, \{c(f): f \in F\}) \tag{2.4}$$

where the range of each component of $v(f)$ is taken to be the entire real line. Thus s^* is the product of s and an Euclidean space of suitable dimensions. Let S^* be the space of all samples s^* and B^* be a σ-field of Borel sets in S^*. Given an SSD and a specified function $\theta(\cdot)$, we can determine the probability measure on B^*, which may be represented by P_θ. Then we have the usual Kolmogorov setup

$$(S^*, B^*, P_\theta) \tag{2.5}$$

and the statistical problem is one of determining $\theta(\cdot)$ given an element of S^*. The triplet (2.5) in the context of sampling from finite populations allowing B^* to be the set of all possible subsets is considered by Basu (1969).

Keeping the Borel sets in (2.1), (2.3) and (2.5) in the background, the essential elements in a sample survey setup may be written as a triplet

$$(F, S^*, P_\theta) \tag{2.6}$$

where F is the frame, S^* is the space of sample units and variate values, and P_θ is the probability measure which, for given θ, provides the probability or probability density of a given element $s^* \in S^*$.

A simple example of sample surveys which has generated considerable controversy is as follows. F is a list of N labelled units $f_1, \cdots, f_N$. A sample s is a selection of a subset of units. S has a finite number of elements and the probability of $s \in S$ is p_s such that

$$p_s > 0 \quad \text{and} \quad \sum_{s \in S} p_s = 1. \tag{2.7}$$

Further, $V(f)$ is a degenerate scalar random variable which takes the value $\theta(f)$ with probability 1, if $\theta(f)$ is the true value of the function $\theta(\cdot)$ at f.

The degeneracy of $V(f)$ has created some unnecessary controversy between Basu (1969) and Godambe (1974) about the definition of sample space. Godambe considers a sample space depending on the parameter unlike S^*. Both of them give the same expressions for probabilities of observed events under given values of $\theta(f)$ and, therefore, they have the same likelihood and distributions of statistics they wish to consider for purposes of inference on unknown parameters. Thus, there is no room for controversy if what is relevant in statistical inference is the specification of probabilities of observed events under different hypotheses. However, I have chosen a realistic situation by considering $V(f)$ to be a non-degenerate random variable (with non-zero density everywhere if necessary) to avoid measure-theoretic problems such as *undominatedness, zero probabilities* for impossible events and consideration of *non-Borel* sets, which seem to obscure real issues. I am sure, Godambe will agree with my setup[†] which is in line with Basu's (in fact the traditional one) in the situation I am considering. We shall now review the controversies about the role of the likelihood in the usual Kolmogorov's setup.

3. Problem of inference

3.1 *Likelihood approach*

Let us consider the set up (F, S^*, P_θ) and an observed sample s^*. Then

$$p(s^* | \theta) = p(s)\, p(\{v(f): f \in s\} | \theta) \tag{3.1}$$

$$= p(s)\, p(\{v(f): f \in s\} | \{\theta(f): f \in s\}) \tag{3.2}$$

where the p's involved are either density functions or probabilities depending on the nature of S, S^* and $V(f)$. The expressions (3.1) and (3.2) show that the likelihood of θ given s^* depends only on $\{\theta(f): f \in s\}$, so that it provides no

[†] The same setup was considered in some detail in Rao (1971).

information on $\{\theta(f), f \in \bar{s}\}$, where $\bar{s}$ is the complement of s in F. Thus, strict adherence to the likelihood principle tells us only about $\{\theta(f): f \in s\}$ and nothing about the function θ on unobserved units as noted by Godambe (1966) and others. Thus the uninformative nature of the likelihood cannot be attributed to some measure-theoretic difficulties arising out of degenerate distributions etc., but is inherent in the problem itself of first choosing some units and making observations on them.

But does it mean that no inference is possible on θ, or its mean value or its variance over F? The answer is no, since the marginal likelihood based on $\{v(f)\}$, summing over all possible s, provides discrimination between alternatives of $\theta(\cdot)$ and thus enables estimation of $\theta(\cdot)$ or its mean value, or its variance. The problem is not new, and the whole subject of estimation of variance components comes under the setup we are considering. Statisticians are used to *random effects* linear models in analysis of variance, which is more complex than the sample survey model (see Rao (1972)).

Faced with the situation where likelihood alone is unable to provide an answer, Basu (1969) advocates the use of prior information on the unknowns and application of Bayesian techniques. Other writers ignore the labels on observed units, after determining the distinct units, and draw inference based on variate values only using standard techniques, such as maximum likelihood. See for instance Hartley and Rao ((1968), (1969)) and the references cited in these articles. The author (Rao (1971)) indicated the possibility of ignoring labels only in subsets of observed units. We shall briefly examine these procedures.

Basu (1969) places too much faith on the universality of the likelihood principle in statistical inference instead of considering it as one of possible techniques. In doing so he is inevitably led to accept Bayesian techniques and some arbitrariness in the choice of prior distributions. Basu is silent on the types of priors he would recommend although the notion of exchangeable priors has found favour with many (see Erickson (1969))[†]. Suppose we want to estimate the number of mosquitoes in a region by selecting some villages and ascertaining the number in each. It is probably known that the number of mosquitoes is highly variable from village to village but it is *not known* which of the villages are more infested than the others. Does the notion of exchange-

[†] There seems to be another logical difficulty. If a customer says that he has observed certain units and knows the true values on these units and wants an estimate of the total for all units, the statistician should look for a prior on the unknowns only (i.e., values on unobserved units). He may choose a prior depending on observed values, in which case he will have a wider choice than blindly choosing an exchangeable prior on all units and deriving a posterior on unobserved units. Bayesians seemed to have missed the type of procedure I am suggesting which is implicit in the method of estimation I have indicated in Rao (1971), by using post stratification.

ability reflect our ignorance of the nature of ***relationship*** between ***number*** of mosquitoes and *label* of a village? What is the best way of utilising the knowledge we may acquire about the relationship after a sample is observed? I suggest that one would be better off by not entertaining any false notion of exchangeability but attempting estimation by post survey stratification when the observed variate values are found to be heterogeneous, in which case labels play an important role. Observe that stratification implies recognition of dependence of variate values on labels, *viz.*, that the variate values associated with labels in one stratum are in general larger or smaller than in another. In an earlier paper (Rao (1971)), I have given an extreme example where, after observing a sample, *one of the observed units* could be considered as one stratum if its variate value is found to be much different from those on others in the sample. The rest of the units in the population could be considered as a second stratum in which case the rest of the units in the original sample would constitute a valid sample from the second stratum. Such an approach may be considered unorthodox but is likely to provide a better estimator than taking a simple average of the observed values.

It is not clear in what sense and in what situations one considers labels as uninformative. Random assignment of labels to units can only alter the relationship between variate values and labels. Why should one deny oneself an opportunity of looking for a possible relationship between variate values and labels from observed data by discarding labels and postulating that they are uninformative? I have already mentioned post survey stratification, and there may be other ways of analysing data where labels play a role, specially when new concomitant variables, even of a qualitative nature, become available. Of course, in many situations, we may not be able to see the relevance of labels. What is the appropriate analysis in such cases? Then the procedures of Hartley and Rao and that of the author (Rao (1971)) seem to be reasonable, although it would have been more satisfactory if these prescriptions could be brought under a unified theory which specifies at what stage and on what subsets labels should be ignored.

3.2 *An alternative approach*

Let us consider the simple situation where F is a list of units $f_1, \cdots, f_N$ with unknown associated variate values $\theta_1, \cdots, \theta_N$. A sample s^* is a set of n units and variate values attached to these units

$$s^* = (f_{s1}, \cdots, f_{sn}\,;\, \theta_{s1}, \cdots, \theta_{sn}). \tag{3.3}$$

Let S be the set of all possible samples of n units and p_s be the probability of $s = (f_{s1}, \cdots, f_{sn})$, a sample of units. The parameter to be estimated is the population total

(3.4) $$T = \theta_1 + \cdots + \theta_N$$

or a linear function $\lambda_1\theta_1 + \cdots + \lambda_N\theta_N$ where λ_i are specified. (The latter may be regarded as a population total with $\lambda_i\theta_i$ taking the place of θ_i as the variate value on the ith unit.) We shall assume that associated with each unit i, there are known concomitant measurements represented by vector $x_i, i = 1, \cdots, N$.

3.2.1 *Minimum variance unbiased estimator*

Let s^* be an observed sample, which provides us with the variate values on n units, whose total is denoted by T_1. Further let the total of variate values on unobserved units be T_2, the population total being $T = T_1 + T_2$. We may estimate T by $T_1 + g(s^*)$ where $g(s^*)$ is a *good predictor* of T_2 on the basis of s^*. (The problem is posed as one of predicting the total of unobserved values having observed the values on some members of the population.) The function $g(s^*)$ may be determined by the usual conditions in prediction problems:

(3.5) $$E[T_2 - g(s^*)] = 0$$

(3.6) $$E[T_2 - g(s^*)]^2 \text{ is a minimum.}$$

We may rewrite the conditions (3.5), (3.6) in the form

(3.7) $$E[T - h(s^*)] = 0$$

(3.8) $$E[T - h(s^*)]^2 \text{ is a minimum}$$

where $h(s^*) = T_1 + g(s^*)$. Then $h(s^*)$ is an unbiased minimum variance estimator of the parameter T. It is well known that no such estimator exists if we insist on minimum variance uniformly for all values of the parameter $(\theta_1, \cdots, \theta_N)$.

However, slight modifications of the conditions might yield unique estimators. One such modification is to minimise the variance (3.8) averaged over a super population model for $(\theta_1, \cdots, \theta_N)$. I have suggested (Rao (1971)) as a natural model the set of all permutations of $(\theta_1/\pi_1, \cdots, \theta_N/\pi_N)$ with equal probability for each, for fixed values of the ratios θ_i/π_i, where π_i is the probability of inclusion of the ith unit in a sample. In such a case, it is shown that the Hurvitz-Thompson (HT) estimator is the best. In the paper cited (Rao (1971)), I have assumed linearity of the estimator which was shown to be not necessary by Thompson in the discussion of the paper. A simple proof in the general case of Thompson is indicated in my reply to the comments on the paper.

Equally one could place restrictions on the function $h(s^*)$ and minimise the variance for a fixed set of parameters. For instance if $h(s^*)$ is symmetric in $(\theta_{s1}/\pi_{s1}, \cdots, \theta_{sn}/\pi_{sn})$, then again the HT estimator is the best.

3.2.2 *Best-fitting function*

In Section 3.2.1, we saw how unbiasedness and minimum variance arise in a natural way when we try to predict the unobserved values using the sample, which is made possible by *probability selection* of samples. Let us look at the problem in a slightly different way.

Having observed s^*, we know the values $\theta(f_{s1}), \cdots, \theta(f_{sn})$ of the function θ at the units $f_{s1}, \cdots, f_{sn}$ and the problem is to estimate the values at $f_{\bar{s}1}, \cdots, f_{\bar{s},N-n}$, which are the labels in the set $\bar{s}$ of unobserved units. The nature of the function $\theta(f)$ is unknown; otherwise, the problem might have a simple answer. Let us attempt to construct a function

$$r[\theta(f), x] \tag{3.9}$$

where f is the label of a unit and x is the associated concomitant vector variable, which may be considered as more homogeneous with respect to the units than $\theta(f)$ itself. A sample s^* provides us with n values $r_{s1}, \cdots, r_{sn}$, of r on the basis of which we shall try to predict each of the values *of* $r_{\bar{s}1}, \cdots, r_{\bar{s},N-n}$, by the same function k of $r_{s1}, \cdots, r_{sn}$, using a criterion such as minimising

$$E_{s\in S}\left[\sum_{i=1}^{N-n} \pi_{\bar{s}i}(r_{\bar{s}i} - k)^2 \Big/ \sum_{i=1}^{N-n} \pi_{\bar{s}i}\right] \tag{3.10}$$

subject to

$$E_{s\in S}\left[\sum_{1}^{N-n} \pi_{\bar{s}i}(r_{\bar{s}i} - k)\right] = 0. \tag{3.11}$$

The problem, as stated, is difficult to solve. Let us replace (3.10) and (3.11) by

$$E_{s\in S}\left[\sum_{1}^{n} \pi_{si}(r_{si} - k)^2 + \sum_{1}^{N-n} \pi_{\bar{s}i}(r_{\bar{s}i} - k)^2\right], \tag{3.12}$$

$$E_{s\in S}\left[\sum_{1}^{n} \pi_{si}(r_{si} - k) + \sum_{1}^{N-n} \pi_{\bar{s}i}(r_{\bar{s}i} - k)\right] = 0. \tag{3.13}$$

Now if we demand that k should be asymmetric in $r_{s1}, \cdots, r_{sn}$ (which is natural, since no recognisable relationship may exist between r and f), then it easily follows that

$$k = (r_{s1} + \cdots + r_{sn})/n. \tag{3.14}$$

Having found k, we may estimate $\theta(f)$ for an unobserved unit by solving the equation

$$r(\theta(f_i), x_i) = k, \quad i = \bar{s}1, \cdots, (\bar{s}, N - n). \tag{3.15}$$

Denoting the solution by $\hat{\theta}_{\bar{s}i}$, an estimate of the population total may be obtained as

$$\theta_{s1} + \cdots + \theta_{sn} + \hat{\theta}_{\bar{s}1} + \cdots + \hat{\theta}_{\bar{s},N-n} \tag{3.16}$$

The estimator (3.16) may not always be unbiased.

For example, the choice

$$r[\theta(f), x] = \frac{\theta(f)}{\pi_f} \tag{3.17}$$

leads to the HT estimator. Other choices may be tried depending on one's knowledge of the likely relationship between θ and (f, x). Further, there is also the problem of specifying a suitable restriction on function k. I have assumed symmetry in arriving at the solution. Other conditions are worth exploring.

I have considered the problem of estimation of the population total as one of prediction of values on unobserved units on the basis of an observed sample and suggested some methods of prediction. The SSD plays a significant role in this approach. No doubt other methods should be tried. For instance, instead of predicting values of individual units, one may predict some function of variate values on blocks of units. The procedure I am proposing is implicit in the method of ratio estimation, familiar to survey statisticians.

4. Some practical problems

Having discussed the problem of statistical inference in sample surveys, I would like to mention some practical problems for which no satisfactory solutions exist and which require careful study.

4.1 *A problem of non-response*

An important problem in sample surveys is that of non-response or response errors, i.e., situations where the variate values on certain units included in the sample are unascertainable or subject to large errors, resulting in bias in estimators. While response errors of certain types can be detected and corrected by built-in checks in the questionnaire, the problem of non-response cannot be easily tackled. I shall describe a recent study made of a somewhat unusual problem at the Indian Statistical Institute (Sengupta (1966) and Rao and Sengupta (1966)) to explain the difficulties involved in non-response situations.

The problem was to estimate the mean direction of flow of an extinct river of geological times in a given region. A simple approach would be to choose points at random or at equal intervals along the river bed and measure the direction of flow at each point. The average of directions at selected points would then provide an estimate of the mean direction of flow. As the river is extinct, observation on direction of flow cannot be made at any desired point, leading to a non-response situation. Measurements can be made only at certain points

where there is rock formation, known as 'outcrops'. The frequency of outcrops is not uniform along the river bed and the average of directions observed at all the outcrops may, therefore, be biased as an estimator of the mean direction of flow. The following method of sampling was adopted to avoid serious bias.

A topographical map of the river bed was obtained and a square lattice was superimposed on the map. In each square grid, measurements of directions were obtained at some of the outcrops selected at random. Averages were computed separately for each grid and then a simple average of the grid averages is taken as an estimate of the mean direction of flow. Such a procedure may minimise bias to a certain extent depending on the size of the grid. Other methods of reducing bias in the above problem and in similar problems of non-response may be explored.

This study points out the need for a re-examination of data on directions of rock magnetism collected by geologists and analysed by Fisher (1953). If the outcrops, at which measurements of direction are possible, are not uniformly distributed over space, the observed data from a random sample of outcrops may be biased for estimating the true mean direction. A well-designed survey for collection of data is essential in such cases to minimise bias due to non-response.

4.2 *Use of concomitant variables*

The use of concomitant variables in estimation of population parameters is stressed in Section 3.2.2. Unfortunately not much work is done about the proper utilisation of concomitant variables in designing sample surveys and in estimating unknown parameters. Most of the available techniques like pps sampling, ratio estimators, etc. refer to the use of a single concomitant variable only. In a recent survey for estimation of area under wheat in different states of India, stratified simple random sampling was used with villages in a stratum as basic units. Ratio estimates using separately geographical area of a village and area under wheat in the previous year as concomitants gave widely differing results. Later verification with complete enumeration figures showed that the use of geographical area gave better estimates in some states, and the area under wheat in the previous year in the other states. In such cases it is worth exploring whether estimators could be improved by using both the concomitant variables in a suitable way. The method of constructing an optimum linear combination of ratio estimators based on individual concomitants, known in the literature, may not be the best procedure.

4.3 *Adjustment of estimates*

Suppose that, in a given year, independent sample surveys for yield of cereals have been conducted in different states of a country and an unbiased estimate together with standard error is obtained for each state. Let $X_1, \cdots, X_k$

be the estimates and let all of them have nearly the same standard error. Should one consider the problem as one of simultaneous estimation of k parameters (compound decision) and report adjusted estimates for the states using the method of James and Stein (1961)? It is well known that James-Stein procedure would grossly underestimate in the case of states with high yields and grossly overestimate in the case of states with low yields. This is clearly not desirable, as individual estimates giving a true picture of disparities between states is essential for policy purposes.

A similar suggestion has been made for adjusting estimates of yield for the whole country over a number of years including the current year. The following table gives unbiased and adjusted estimates for the years 1966–73. The adjusted value is obtained by the formula, (1/5) (overall average) + (4/5) (unbiased estimate).

Estimated yield of cereals in millions of tons during 1966–73

Estimate	1966	1967	1968	1969	1970	1971	1972	1973
Unbiased	52	61	62	66	68	64	67	73
Adjusted	53.4	61.6	62.4	65.6	67.2	64.0	66.4	71.2

As a result of adjustment the figure for current production (for the latest year) is reduced by nearly two million tons! This may be largely due to serious bias introduced by James-Stein adjustment in the estimation of parameters with highest values. Further, if one is studying the trend of production of cereals over time, the average increase per year as indicated by the unbiased estimates is about 4 million tons whereas the corresponding figure for adjusted estimates is 3.2 million tons. The latter estimate of trend is seriously biased downwards (see Rao (1974) for further remarks). It appears that in situations such as the above, adjustment by James-Stein procedure is not desirable.

References

BASU, D. (1969) Role of the sufficiency and likelihood principles in sample survey theory. *Sankhya* A **31,** 441–454.

ERICKSON, W. A. (1969) Subjective Bayesian models in sampling finite populations. *J.R. Statist. Soc.* B **31,** 195–233.

FISHER, R. A. (1953) Dispersion on a sphere. *Proc. Roy. Soc. Lond.* A **217,** 295–305.

GODAMBE, V. P. (1955) A unified theory of sampling from finite populations. *J. Statist. Soc.* B **28,** 310–328.

——— (1965) Contributions to the unified theory of sampling from finite populations. *Int. Statist. Rev.* **33,** 242–258.

——— (1966) A new approach to sampling from finite populations, I: Sufficiency and linear estimation. *J. Statist. Soc.* B **28,** 310–319.

——— (1974) A reply to my critics. Paper presented at the International Symposium on Recent Trends of Research in Statistics, Calcutta, December 1974.

HARTLEY, H. O. AND RAO, J. N. K. (1968) A new theory of sample surveys. *Biometrika* **55,** 547–557.

———— (1969) A new estimation theory for sample surveys II. *New Developments in Survey Sampling*, Wiley Interscience, New York. 147–169.

JAMES, W. AND STEIN, C. (1961) Estimation with quadratic loss. *Proc. Fourth Berkeley Symp Math. Statist. Prob.* 361–379.

RAO, C. R. (1971) Some aspects of statistical inference in problems of sampling from finite populations. *Foundations of Statistical Inference*, Ed V. P. Godambe and D. A. Sprott. Holt, Rinehart and Winston, 177–202.

——— (1972) Estimation of variance and covariance components in linear models. *J. Amer. Statist. Assoc.* **67,** 112–115.

——— (1974) Some thoughts on regression and prediction—Part I. *Gujarat Statistical Review* **1,** 7–32.

RAO, J. S. AND SENGUPTA, S. (1966) A statistical analysis of cross-bedding azimuths from the Kamthi formation around Bheemaram, Pranahita-Godavari valley. *Sankhya* B **28,** 165–174.

SÄRNDAL, E. C. (1974) Continuous survey sampling models. (unpublished paper).

SENGUPTA, S. (1966) Studies on orientation and imbrication of pebbles with respect to cross-stratification, *J. Sed. Petrology* **36,** 362–369.

Suppl. Adv. Appl. Prob. 7, 62–71 (1975)
Printed in Israel

COST-BENEFIT ANALYSIS OF DEMOGRAPHIC DATA

I. RICHARD SAVAGE, *Yale University*

0. Introduction

Statistical theory and demography have had limited interaction for many years. There was an International Statistics Institute satellite conference in the summer of 1971 to help remedy the situation but the proceedings do not seem to be forthcoming. Actually there is some mistrust of statistical ideas by the official demographers. Thus demographers give projections (extrapolations under deterministic models) rather than predictions (means with standard deviations) of future populations, see Keyfitz (1972) and Hoem (1973). There is a tendency to present demographic results without a probabilistic assessment of the associated uncertainty. These assessments are difficult to make but they are very helpful to the well-prepared user.

The areas to be discussed involve theoretical and applied knowledge from a variety of disciplines, such as economics, political science, sociology, and statistics. Although I believe the academic statistician is one of the few socially acceptable dilettantes, my faith is shaken when confronted with such a serious problem.

Finally, not all of demography is considered. Attention is focussed on major data series collected either by census or very large survey.

Uses of demographic data in governmental activities are considered in contrast to the scientific study of populations.

1. Cost-benefit (economic)

Census data are collected for a great variety of purposes (see Research Publications Inc. (1974)) and, while we are waiting for the next census, estimates are needed. This great apparent demand for data has not built up a technology for justifying the Census budget[1]. Census data is a public good and the economic value of such goods is not easily determined, see Arrow (1965). An effort to do this for crop data is found in Hayami and Peterson (1972), but much additional effort is required to develop technique.

This manuscript was also the basis for a lecture at the Bayes Methods Conference. Yale University, Spring 1974. The research was supported by the National Science Foundation Grant No. GS–41617.

The impression is that much of the U.S. statistical program has the technology to optimize in terms of costs. That is, given a budget, an agency will come close to getting the most data possible. This impression is correct in reference to sampling errors in surveys: given a data requirement, in particular size of sampling error, a nearly optimally designed survey will be used. On the other hand, minimization of total (expected) risks due to costs of data collection and terminal actions is not common. A real problem covers many related issues in which the statistician should participate,

(1) Should new data be collected or can analytic methods obtain legally and economically necessary data from existing sources?

(2) Should current definitions and procedures be maintained or modified[2]?

(3) What level of accuracy — sampling variation, biases, inadequate definitions, non-sampling errors — is needed?

Many, including myself, have presumed that one or possibly several major programs could more than justify the cost of important Federal statistical activities. Revenue sharing — five billion dollars per year — is a potential candidate to show the value of Federal statistics[3]. The presumption is that believed errors in the data base would suggest major inequities in the allocation of funds. Once the inequities are located then special interest groups (individual cities, states, racial groups, etc.) would be willing to remove them by paying or lobbying for better data.

An important data series for allocating in revenue sharing is the 1970 state populations. It is well known that about 2 per cent of whites and 7 per cent of blacks are not enumerated in the census[4]. Thus a state with relatively few blacks (Minnesota) would appear to have a substantial advantage over a state with relatively many blacks (Mississippi). The Urban League (see Hill and Steffes (1973)) and Savage and Windham (1974) have independently followed this line of thought and come to different conclusions. The Urban League found major inequities, but apparently used the incorrect formula. Savage and Windham found relatively minor inequities even when they assumed urban areas had much higher undercount rates than the other areas. Currently, the Stanford Research Institute is doing a detailed analysis of this problem.

There are a few economists interested in the value of information quality of statistical data. To rationalize the Federal statistical program, the work of the economists must be developed to the extent that budgets can be at least partially explained in terms of the data being created.

2. Cost-benefit (political)

Neither the Constitutional requirement of a census nor the Supreme Court doctrine of one-man-one-vote yields a simple mandate for the quality of data. It is doubtful that current data are satisfactory from philosophical, legal, or

political viewpoints. Federal data are used for many decision-making activities, such as location of new facilities, level of the Federal Reserve discount rate, or drafting new legislation. A good statistical system might be used as indicative of a solid government. Also, many private political decisions are based on census and other public data.

For the activities just described, the cost-benefit relations are particularly awkward. The cost of collecting data is mostly expressed in dollars, but the benefits are rather vaguely expressed in political good[5]. The finding of a unit to compare cost and benefits (political) as well as to combine the political and economic benefits is a problem for political scientists and economists, see Seltzer (1973).

At this point in history, it is particularly embarrassing to talk about the value of a seat in the U.S. House or New York Assembly. Savage and Windham (1974) suggest that it is likely that Oklahoma should have received one of Connecticut's seats in the U.S. House[6]. How should the political force of Oklahoma protect itself from such injustice in the future? Apparently similar studies have not been done for the state legislatures.

I recall almost no discussion of what constitutes a 'fair' census[7]. As remarked at the beginning of this section, the system appears unfair. Specifically, the U.S. Census is believed (by many, including the Census) to have different undercount rates for specified segments of the population. Inequities, economic and political, occur mainly at the state and smaller geographical levels. At least for race these undercount rates are believed to be known with high precision, but at the state level the (current) precision is low, and little is being done to improve it. If the state undercount rates were known with high precision, two kinds of action could be taken. If the believed rates resulted in inequities, administrative and legislative action could be used to improve the situation. A more flexible possibility is that the aggrieved parties could apply political pressure to their constituents to be counted and they could insist on better counting procedures.

It is interesting that those procedures which (I think) would most efficiently improve the population data are politically expensive, i.e., unpopular. Better internal migration data appears to be a real bargain. Use of administrative data — Social Security and Internal Revenue Service — is useful but would be much better with greater use of matching of data files which is opposed by Congress. Registration cards with continuous reporting gives good data (and undesired side effects). The common use of the Social Security number is the apparent beginning of a possibly undesired registration system, see Secretary's Advisory Committee (1973).

What is a fair system? It is unsatisfactory to say fairness occurs when no group is willing to pay more for bias reduction.

3. Cost-benefit (social)

The process of obtaining information requires specialists in many areas. For demographic data the sociologist could play a central role. The report of the Advisory Committee on Problems of Census Enumeration (1972) contains much sociological probing of why it is hard to count people. The immediate pay-off of that effort is limited, but I suspect that it is worth making other sociological efforts to understand the process of obtaining data from people. (Anthropologists and psychologists are also needed.)

A new need for data is in the social indicator movement[8]. Development of reasonable accuracy standards for these data has not begun. There will be interesting statistical problems since the desired form for the data is in time series. At the moment the movement is using descriptive statistics. Eventually, however, formal models and social accounting will be developed, see Clark (1973). Then the demand on data quality will be much increased.

Social accounts dealing with the quality of life form a slippery subject. Analytical demography dealing with the quality of life is a highly developed subject[9]. My reading in analytic demography has shown me many clever devices for interrelating various data series. At the same time practically nothing is said about the error structure of the output or the actual ways of using the output.

4. Statistics

What can statistical theory do for the work discussed in Sections 1–3? On the non-analytic side it can locate problems. It warns against a deterministic interpretation of an uncertain world. It highlights the uselessness of projections when trying to relate to the future. It might encourage finding operational meaning to catch-phrases such as 'one-man-one-vote'.

On the analytic side statistics should encourage workers in the data system to form problems. (A good problem is a good guide.) As a working hypothesis, I think it is appropriate to assume that the collection of the major statistical series and some of their important uses can be formulated as a statistical decision problem[10]. The solutions of such problems would give strong arguments for how much data to collect and what kind of terminal losses to anticipate.

The use of the working hypothesis would require cooperation, skill and judgement. The total Federal statistical activity is gigantic and its uses are many. Thus, a first task is to settle on a problem (or problems) of reasonable size. Then start filling in the pieces of the decision model. Just seeing those pieces will be interesting and useful.

Apparently very little is known about the consequences (economic, political

or social) of allocations and decisions based on imperfect data. Even for distribution of such things as seats in the U.S. House or dollars to states in revenue sharing the data required are substantial and the formulas are involved. Analytic models might show the consequences of changes in the data base, but it is likely to be simpler and just as enlightening to evaluate the consequences of a few plausible sets of data or do a Monte Carlo study. Thus a large-scale computer is useful to explore consequences. These studies would take on a pointed meaning if the consequences were given an economic interpretation.

The consequence side of a problem will not be followed here. As a closing effort I will try to say something about the analytic structure of a demographic problem — one of the least involved problems. This should show the origin of the error structure. Of course the following is schematic, incomplete, and tentative.

To make life simple assume the census occurs every year, people are born, immigrate and die just before the census. Then use the notation:

n_{ij} = number of people in the country for ith census of age j.
b_i = number of people born just before the ith census.
m_{ij} = number of people (net) immigrating into the country just before the ith census of age j.
d_{ij} = number of people who die just before the ith census of age j.

Then,

$$n_{i0} = b_i,$$

and

$$n_{ij} = n_{i-1,j-1} + m_{ij} - d_{ij}, \quad j \geqq 1.$$

Even the working definitions of these quantities are non-trivial. Also these quantities are not observed in a large society. Rather, the data system generates much related data.

One might desire the joint distribution for the terms $\{n_{ij}, b_i, m_{ij}, d_{ij}\}$. For simplicity, consider

$$n_0 = \sum_{j=0} n_{0j},$$

the total population in year 0. With obvious notation

$$n_i = n_{i-1} + b_i + m_i - d_i$$

and our knowledge regarding n_0 could be developed by combining our

knowledge about the right-hand-side quantities. Of course other strategies exist involving more detail on the right-hand-side, such as

$$n_i = n_{i-2} + b_{i-1} + b_i + m_{i-1} + m_i - d_{i-1} - d_i.$$

A greater number of lags can be used and demographic detail included, such as age, race, and locale. With primary interest in n_0 there is no formal statistical reasoning to tell us which variables to use. The choices involve the following considerations:

(1) If too many variables are used, then it is too difficult to construct a model.

(2) The amount of data and computation increases very rapidly as the number of variables increases.

(3) The quality and quantity of data is not the same for all variables.

(4) If the number of variables used is large, then the analysis will have available more internal checks[11].

The least involved datum regarding n_0 — total population — is N_0, the census total[12]. But what does N_0 do regarding our distribution for the value of n_0? Knowledge of previous censuses and the census prodecures would yield for the 1970 U.S. Census $\Pr(N_0 \geqq n_0 | N_0) \leqq \varepsilon$ where ε is very small. That is, it is very unlikely to have a larger census count than population in the United States.

Likewise, $\Pr(1.05 N_0 \leqq n_0 | N_0) \leqq \varepsilon$. A statement like $\Pr(1.025 N_0 \leqq n_0 |$ 1970 Census$) \sim 1/2$ would make use of much more detail of the Census than N_0. Techniques for assessments of probabilities appear to have been developed without explicit regard for the problem of which information should be used at what cost: see L.J. Savage (1971). Perhaps the assessment of probabilities regarding n_0 can be brought into sharp focus by (inappropriately) reducing the alternative sets of useful data. The construction of a probability distribution for a complex quantity, n_0, from a wealth of more or less relevant data is a major challenge.

In considering n_0, one notices that the stochastic natures of relevant data are different for the several variables. Thus in thinking about N_0 as evidence for n_0, it is known that:

(1) N_0 comes from a stochastic process in the sense that it contains random errors of a clerical nature and biases as a result of misunderstandings of instructions;

(2) N_0 contains errors because of deceptions, deliberate omissions, age heaping, etc.;

(3) N_0 fails to include data from some individuals because their household was missed.

Our experience makes it clear that variation from (2) and (3) (particularly (3)) is more than the variation from (1). How to assess this variation is unclear. The obvious strategy is to start considering components of the population.

To assess b_0, the most important datum is B_0, the number of registered births. The stochastic relation between B_0 and b_0 is relatively simple compared to N_0 and n_0[13]. The proportion of births in hospitals or under the supervision of licensed individuals is increasing and near 1. At the least on a local basis, it is relatively easy to compare b and B. Clerical errors in evaluating B are almost negligible. Although I do not know a good way to assess my belief about b_0 it seems like a simpler task than working with n_0.

Definitely, there are problems of probability assessment. The role of sampling distributions is small—clerical errors can have simple sampling distribution models, but those errors are usually relatively small. The trick is to avoid pure guesswork as a substitute for assessment. The strategy is to work with the better-understood components of the population.

5. Education statistics

Successful rationalization of portions of the Federal statistical system will not come easily, see J. Duncan (1974). The preceding sections say that the theory might not yet be available, and certainly no one has much experience in the task.

A suggestion of the President's Commission on Federal Statistics (1971) is to audit various statistical programs within the Government bureaus. The idea is to check that the work is being done well, to bring new techniques to old problems, to remove deficiencies, and to praise success. These audits have not yet been done. However, Health, Education, and Welfare has asked the Committee on National Statistics to review their programs in educational statistics, see Martin (1974). Part of the program is to rationalize the education statistics program. Miracles are not expected. The side benefits of problem formulation can justify the review.

* * *

In summary, the Federal data series are sources for assessing our beliefs about important quantities. The complexities of the loss functions and the data (error) structure take these topics out of neat standard packages. Solutions appear to require some new theory, learning to work with large unwieldy structures, and judicious simplification of messy problems.

Notes

[1]The following quotation supports the contention that the technology to justify a census budget is not developed. The quotation also makes a suggestion for the improvement of the technology.

> It is clear that a quinquennial census of population and housing would improve the quality of small area sample surveys carried out in the last half of the decade. The Commission is unwilling to say that such a census would provide the most effective statistics for small areas that can be had for the cost. But the Commission has not seen evidence even that the statistics thus provided would be effective at all, let alone the most effective possible.
>
> Evaluation of the case for the quinquennial census might proceed as follows: derive rates of change in various social and economic activities and use these rates to estimate the cost of misallocation that has resulted in the last one-half of a decade from dependence on decennial data. If this cost of misallocation exceeds the cost of the census, taking the census would be worthwhile.
>
> This eminently logical procedure, while exceedingly difficult to follow, would improve our understanding of the importance of intercensal updating. It may be prohibitively difficult because the rates of change to be identified are so many and so varied, and interact in such a complex manner, that the mathematical equations of the models often are not solvable, thus precluding thorough analysis. Even more important, so little is known about the sensitivity of decisions themselves, and about alternative means of developing data, that a whole new field of cost-effectiveness analysis would have to be developed. (From page 125 and 126 of the President's Commission on Federal Statistics (1971).)

[2]There is strong feeling that a slight change in definition or of procedure can make major public differences. Thus in the computation of unemployment rates the previous procedure was to count people where they work, and the new, 1974, is to count people where they live. This change has been explained by the Commissioner of Labor Statistics to the members of Congress by the use of a 13-page memorandum, see also Wetzel (1974). The immediate call for this and similar changes is the strong reliance on unemployment data of the Comprehensive Employment Act (1973). It would be interesting to know if these changes make an appreciable change in the consequences of the Act, see Johnston and Wetzel (1969). See *New York Times*, 28 April 1974, letter from Ewan Clague.

[3]Congress apparently gave limited thought to the implementation of revenue sharing. The process is first to divide the money between the states and then to divide it between local governments. The rules for division at the state level are complicated and require extensive data resources which are not readily available even after a census, see Savage and Windham (1974).

The division within states will require extensive new data sources.

When Congress planned the legislation they were most successful in that the states received what the Congress thought they would receive. Poor data on 'tax effort' may have caused Alabama a moderately large loss.

[4]Siegel's papers (1974a and b) contain the most serious effort to evaluate the undercount for the 1970 census. His analysis is official. It does not contain any material on the undercount rates for geographical regions below the national level.

[5]There are important non-monetary costs associated with data collection. The time of the respondent is easily converted to money, but his annoyance or pleasure — a political cost — is more difficult to measure. The Statistical Policy Division has a major problem of keeping respondents cooperative in Federal data collection.

[6]The Savage and Windham (1974) analysis of the U.S. House allocation was not consistent with the U.S. Census analysis since the two groups used different 'state populations'. In 1967 it was anticipated that the Savage-Windham and Census methods would yield the same results, see footnote 8 of Subcommittee on Census and Statistics (1970).

[7]There are problems of human values associated with the census. Discussion has included: (1) confidentiality, (2) privacy, (3) duty to be counted versus genocide.

[8]It is not clear what the social indicator movement is. The guiding thought appears in O. Duncan (1969). The first United States official social indicator volume is Statistical Policy Division (1974). The Russell Sage Foundation and Social Science Research Council place much importance on this subject. A key activity is to develop time series which describe the quality of life.

[9]Shryock and Siegel (1973) presumably give a good view of what official demographers do. In particular they include substantial material on analytic demography.

[10]The decision theoretic Bayesian framework is used in this paper. It is the most demanding statistical framework. It requires the formulation of a complete problem. In application, one would work at parts of the problem and possibly never bring everything into operation. But this Bayesian view keeps clear what needs to be done. Raiffa and Schlaifer (1961) or Savage (1968) are textbook versions of the viewpoint. Guttentag (1973) expresses this viewpoint in an applied social science setting.

[11]For the U.S. population there are certain established facts which can be used to check the quality of data. For example, the number of 20-year-old black males in 1960 must be more than the number of 30-year-old black males in 1970. The number of male live births must exceed the number of female live births. The U.S. Census shows an undercount when checked against such facts.

[12]No figure reported by the Census is simple. The total count, N_0, is the product of a very complicated process. And N_0 is not immediately fixed, since the Bureau of the Census corrects and adjusts several times before the final value is obtained. Yet, the relation between N_0 and n_0 is simple and direct compared to some other data useful in assessing our belief about n_0.

[13]Although the number of registered births is now close to the number of births, this was not the case several decades ago. A similar remark applies to deaths. Apparently international migration data is never very good.

References

Advisory Committee on Problems of Census Enumeration, Division of Behavorial Sciences — National Research Council (1972) *America's Uncounted People*, National Academy of Sciences, Washington, D.C.

AGNEW, R. A. (1972) Crude confidence interval estimates for future U.S. population levels. *Amer. Statistician* **26** (1), 47–48.

ARROW, K. J. (1965) Criteria for social investment. *Water Resources Research* **1**, 1–8.

CLARK, T. N. (1973) Community social indicators: From analytical models to policy applications. *Urban Affairs Quarterly* **9**, 3–36.

DUNCAN, J. W. (1974) Developing better long range plans for federal statistics. *Statistical Reporter* **75** (4), 49–54.

DUNCAN, O. D. (1969) *Toward Social Reporting: Next Steps*. Russell Sage Foundation, New York.

GUTTENTAG, M. (1973) Evaluation of social intervention programs. *Ann. N.Y. Acad. Sci.* **218**, 1–13.

HAYAMI, Y. AND PETERSON, W. (1972) Social returns to public information services: Statistical report of U.S. farm commodities. *Amer. Econ. Rev.* **62**, 119–130.

HILDRETH, C. (1963) Bayesian statisticians and remote clients. *Econometrica* **31**, 422–438.

HILL, R. B. AND STEFFES, R. B. (1973) *Estimating the* 1970 *Census Undercount for State and Local Areas*. National Urban League, Washington, D.C.

HOEM, J. M. (1973) *Levels of Error in Population Forecasts*. Artikler fra Statistisk Sentralbyrå 61, Oslo, Norway.

JOHNSTON, D. F. AND WETZEL, J. R. (1969) Effect of the census undercount on labor force estimates. *Monthly Labor Review* **92** (3), 3–13.

KEYFITZ, N. (1972) On future population. *J. Amer. Statist. Assoc.* **67**, 347–363.

MARSCHAK, J. (1966) Economic planning and the cost of thinking. *Social Research* **33**, 151–159.

MARTIN, M. E. (1974) The work of the Committee on National Statistics. *Amer. Statistician* **28** (3), 104–107.

MUHSAM, H. V. (1956) The utilization of alternative population forecasts in planning. *Bull. Res. Coun. Israel* **5c**, 133–146.

PRESIDENT'S COMMISSION ON FEDERAL STATISTICS (1971) *Federal Statistics.* (2 vols.), Stock no. 4000–0269, U.S. Government Printing Office, Washington D.C.

RAIFFA, H. AND SCHLAIFER R. (1961) *Applied Statistical Decision Theory.* Graduate School of Business Administration, Harvard University.

RESEARCH PUBLICATIONS, INC. (1974) *United States Decennial Census Publications* 1790–1960. New Haven, Connecticut.

SAVAGE, I. R. (1968) *Statistics: Uncertainty and Behavior.* Houghton Mifflin Company, Boston.

SAVAGE, I. R. AND WINDHAM, B. (1974) *Effects of Bias Removal in Official Use of United States Census Counts.* Unpublished.

SAVAGE, L. J. (1971) Elicitation of personal probabilities and expectations. *J. Amer. Statist. Assoc.* **66**, 783–801.

Secretary's Advisory Committee on Automated Personal Data Systems (1973) *Records, Computers and the Rights of Citizens.* Stock no. 1700–00116, U.S. Government Printing Office, Washington D.C.

SELTZER, W. (1973) *Demographic Data Collection: A Summary of Experience.* An occasional paper of the Population Council distributed by Key Book Service, Inc., Bridgeport, Conn.

SHRYOCK, H. S., SIEGEL, J. S., LARMON, E. A., (editorial assistant) *et al.* (1973) *The Methods and Materials of Demography.* (2 vols.), Stock no. 0301–2289. U.S. Government Printing Office, Washington, D.C.

SIEGEL, J. S. (1974a) Estimates of coverage of the population by sex, race, and age in the 1970 census. *Demography* **11**, 1–23.

SIEGEL, J. S. (1974b) *Estimates of Coverage of Population by Sex, Race, and Age: Demographic Analysis.* Evaluation and Research Program PHC(E)–4, U.S. Bureau of the Census, U.S. Government Printing Office, Washington, D.C.

Statistical Policy Division of Office of Management and Budget (1974) *Social Indicators* 1973. Stock no. 0324–00256, U.S. Government Printing Office, Washington, D.C.

Subcommittee on Census and Statistics (1970) *The Decennial Population Census and Congressional Apportionment.* House Report No. 91—1314, U.S. Government Printing Office, Washington, D.C.

U.S. Bureau of the Census (1974) *Standards for Discussion and Presentation of Errors in Data.* Technical Paper No. 32, U.S. Government Printing Office, Washington, D.C.

WETZEL, J. R. (1974) New procedures for estimating unemployment in state and local areas. *Statistical Reporter* **74** (11), 181–184.

Suppl. Adv. Appl. Prob. 7, 72–83 (1975)
Printed in Israel

THE AGGREGATION PROBLEM IN ECONOMETRICS

JOHN S. CHIPMAN, *University of Minnesota*

1. Introduction

On looking over the list of participants in this conference, I find that I am apparently the only one who was not formally trained as either a mathematician or statistician. You will understand, then, why, during the last few days, I have felt a little bit like an intruder at a Quaker meeting. I have also been interested to hear my field described as an 'application', or an 'applied field'. I think I should hasten to point out that economics runs the gamut from the purest of pure theory to the most grubby empirical work. The term 'substantive field' might perhaps be a more apt description, because we ourselves engage in the same types of discussions concerning the proper balance between theory and applications, and the problem of bridging the gap between the two.

This problem is particularly great, I think, in a non-experimental field, such as economics. In experimental fields, it is taken for granted, I think, that the empirical data on a variable, say x, correspond to what x is supposed to mean in theory. Of course, it is admitted that there are errors of measurement and rounding errors and so forth, so that what is actually observed is x plus a random error; nevertheless, by and large it is considered that there is a correspondence between the variables of the model and the data to be collected. Now, in economics the situation is that this is hardly ever the case. If you look at a model of the economy — one that we might think of as a 'true model' — it would depict a huge system with millions of individuals and firms, each one of which has a demand or supply function for various commodities, which will be a function of literally thousands of prices of commodities defined in very detailed manners. We rarely have such detailed and comprehensive data on prices, and we hardly ever have data on individual transactions; and even if we did, we do not have the means to handle such large amounts of data simultaneously. If you consult the consumer price index for data on prices, you will find them all grouped. Whether we like it or not, that is the way they are published; and even if you should get data on, say, the price of beef in August of 1974, this itself is an aggregate — over different qualities of beef, over different localities and over different days of the month. So we never really have a situation in which we have a true correspondence between the variables of the model and the data we actually collect. Therefore, we cannot really hope

to apply the models that we have, directly. Even if we should entertain a forlorn hope that a century hence we could persuade governments to collect data on all these variables, we simply cannot wait that long. In the meantime, what do we do? We construct approximative models or aggregative models. And we hope that these aggregative models approximate the true models. This, then, is the general background of the aggregation problem.

2. Modelling aspects

Let me now introduce a bit of structure. My general framework will be that of a multivariate multiple regression model; however, I will start out with the modelling aspects before getting into the statistical aspects. The situation is depicted in Figure 1 (cf. Malinvaud [8], Chipman [2]). We have a set, or space, $\mathscr{X}$.

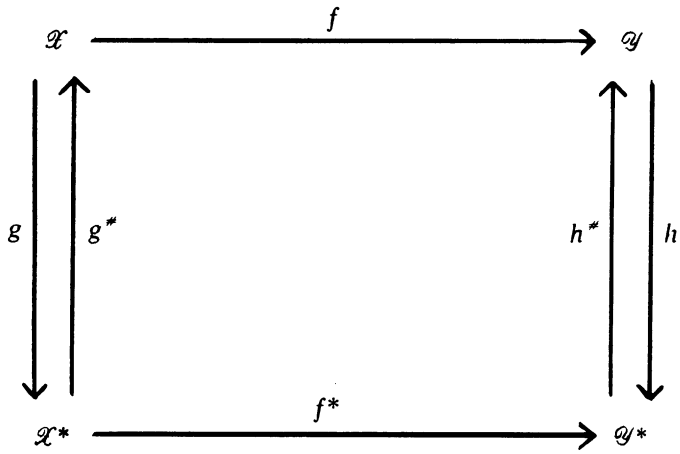


Figure 1

Its elements, x, are k-tuples consisting of the magnitudes of k 'exogenous', or 'independent', variables. $\mathscr{Y}$ is another space whose elements, y, are m-tuples consisting of the magnitudes of m 'endogenous', or 'dependent', variables. The 'true model' (put in quotation marks, because nobody ever believes any model to be really true), consists of a mapping $f: \mathscr{X} \to \mathscr{Y}$. For example, $\mathscr{X}$ might consist of the possible income levels of k different individuals, and $\mathscr{Y}$ of the possible consumption expenditures of those same individuals (here, $m = k$); $f: \mathscr{X} \to \mathscr{Y}$ would then be a k-tuple of functions describing the relationship $y_i = f_i(x_i)$ between consumption and income for each individual in the

economy ($i = 1, 2, \cdots, k$). Now, we do not have data on incomes and consumption expenditures for each and every individual. So we have to aggregate. We represent the aggregation process as a mapping $g: \mathscr{X} \to \mathscr{X}^*$ where $\mathscr{X}^*$ is a space of lower dimension $k^* < k$. We do have data on *national* income; recourse to such data will correspond to a very extreme case of aggregation. As our grouping function g we may take $g(x) = \Sigma_{i=1}^{k} x_i$, i.e., g simply sums the incomes of the k individuals. In this case, $\mathscr{X}^*$ is a space of just one dimension ($k^* = 1$). Likewise, we can aggregate consumer expenditures according to a mapping $h: \mathscr{Y} \to \mathscr{Y}^*$, where $\mathscr{Y}^*$ is a space of dimension $m^* < m$. In this case we can let $\mathscr{Y}^*$ be a one-dimensional space of aggregate consumer expenditures, and take $h(y) = \Sigma_{i=1}^{m} y_i$, where $m = k$ and $m^* = k^* = 1$. Our problem now is to find a function or mapping $f^*: \mathscr{X}^* \to \mathscr{Y}^*$ which in some sense approximates or mirrors the mapping f.

There are some theoretical cases of interest in which what we may call *perfect aggregation* is possible. Following Theil [11] we may distinguish two such cases: (a) $h(f(x)) = f^*(g(x))$ for all $x \in \mathscr{X}$ and some function f^*; (b) the observed data x are restricted to a k^*-dimensional subspace $\mathscr{X}^0 \subset \mathscr{X}$, and $h(f(x)) = f^*(g(x))$ for all $x \in \mathscr{X}^0$ and some f^*. Referring to Figure 1, Case (a) requires $h \circ f = f^* \circ g$ for some f^*, i.e., that there exist f^* such that the diagram commutes. It can be shown (cf. Chipman [2]) that if, in the above example, $\mathscr{X} = \mathscr{Y}$ is the non-negative orthant of k-dimensional space, then so long as at least one f_i is continuous at a point, a necessary and sufficient condition for the existence of a solution f^* to the equation $f^* \circ g = h \circ f$ is that f_i be of the form $f_i(x_i) = a_i + bx_i$ for each i, whence the solution is given by the affine function $f^*(x^*) = a + bx^*$. This is the famous Keynesian 'consumption function' which for many years formed the basis for predictions of national income and employment. Case (b) will be fulfilled if $x_i = \lambda_i x^*$ for all i, since then f^* may be defined by $f^*(x^*) = \Sigma_{i=1}^{k} f_i(\lambda_i x^*)$. These two theoretical cases show that either *similarity of tastes* or *proportionality of incomes* can allow for the possibility of perfect aggregation. In general, *structural uniformity* and *multicollinearity* allow perfect aggregation to hold; accordingly, the belief that such uniformities and multicollinearities are approximately fulfilled in the real world lies at the basis of the faith that economists place in aggregative models.

Another illustration may be taken from the field of international trade. One is interested, for instance, in studying the impact of changes in world prices (e.g., oil) on domestic prices within a country (e.g., gasoline prices, food prices, wage rates, profits etc.). In this case, $\mathscr{X}$ would be a space of k-tuples of prices of internationally traded commodities, and $\mathscr{Y}$ a space of m-tuples of prices of commodities on the domestic market. Now, monthly U.S. statistics [12] of export, import and wholesale prices, even though they involve considerable aggregation, are so detailed and microscopic that they are almost impossible to

handle both computationally and conceptually. About 2400 commodity groups are listed, ranging from a category as fine as, e.g., microcrystalline wax, to a heterogeneous aggregate such as, e.g., 'cod, cusk, haddock, hake, pollock, smoked or kippered, not otherwise prepared or preserved and not canned.' Prices for each category, obtained by dividing monthly values by monthly quantities, are an aggregate over the month and over ports of entry or exit. These prices tend to fall into groups which move up and down together, so that the observations x tend to lie very close to a lower-dimensional subspace $\mathscr{X}^0 \subset \mathscr{X}$. Thus, one way or another, we are going to have to use a model of reduced rank. We may distinguish two problems: (1) Assuming that we already have methods of aggregating groups of prices into price indices, how can we best choose a pseudo model relating the domestic price indices to the international ones? (2) How should we group the prices and form price indices to begin with? The second of these problems belongs to the class of problems that Richard Savage was talking to us about yesterday.

Let us now formulate this more precisely, starting with problem (1). I shall assume that the so-called independent variables x possess means, variances and covariances, and hence a moment matrix $\mathscr{E}xx' = M$. (Variables x, y, etc. will be considered as column vectors and a prime denotes transposition; $\mathscr{E}$ is the expectation operator.) What shall be the criterion for optimal choice of the simplified model f^*? Minimisation of forecast error seems a reasonable criterion, and squared forecast error is, of course, a very natural measure to consider. An ideal forecast of the aggregated dependent variable y^*, conditional on values of the unaggregated independent variable x, would be given by $y^* = (h \circ f)(x)$; on the other hand, an investigator employing an aggregative model would use as his forecast $\hat{y}^* = (f^* \circ g)(x)$. We need a definition of the distance between these two composed functions, $h \circ f$ and $f^* \circ g$. Suppose, then, that the functions f, g and h are such that the joint distribution of $g(x)$ and $(h \circ f)(x)$ has finite first and second moments, and that f^* is required to be such that $(f^* \circ g)(x)$ has finite first and second moments. We shall take as our measure of the distance the non-negative definite matrix

$$
\begin{aligned}
d(f^* \circ g, h \circ f) &= \mathscr{E}\,[(f^* \circ g)(x) - (h \circ f)(x)]\,[(f^* \circ g)(x) - (h \circ f)(x)]' \\
&= \mathscr{E}\,[f^*(x^*) - y^*]\,[f^*(x^*) - y^*]', \qquad (1)
\end{aligned}
$$

which will be called the matrix of *aggregation bias*. One such matrix will be defined as greater than or equal to another if the difference between them is non-negative definite; we then seek an aggregative model f^* which minimises (1) in terms of this partial ordering. Now (cf. Doob [3], pp. 271–272; see also Chipman [2]), the function f^* which minimises (1) is precisely the conditional expectation of y^* given x^*, $\mathscr{E}(y^*|x^*)$.

In the special case in which the functions f, g, h are all affine, i.e., sums of homogeneous linear functions and constant functions, this conditional expectation becomes simply

$$(2) \qquad \mathscr{E}(y^*|x^*) = (h \circ f \circ g^\#)(x^*), \quad \text{where} \quad g^\#(x^*) = \mathscr{E}(x|x^*).$$

The problem of finding an optimal choice of f^* boils down in this case to that of finding the conditional expectation of x given x^*, which defines the mapping $g^\#: \mathscr{X}^* \to \mathscr{X}$ (see Figure 1).

Since $\mathscr{E}(x|x^*)$ itself is in general not an affine function, even if f, g and h are, the problem may be further simplified by considering the case in which the aggregative model f^* is required to be affine. Then we may replace the mappings f, g, h, f^* by the linear transformations, or matrices, F, G, H, F^*, where we adopt the convention that the first components of x and y are, respectively, dummy 'variables' taking on the constant value 1, the remaining components corresponding to the magnitudes of the substantive variables; accordingly, G and H may be considered as block diagonal matrices, and F and F^* as block lower-triangular matrices, each of whose upper left block is equal to 1. (The remaining diagonal blocks of G and H will themselves typically have block diagonal structure, the diagonal blocks being rows of weights which, for example, form price indices of groups of commodities out of the individual prices in the respective groups.) We may add a random error term e to the 'true model', to obtain

$$(3) \qquad y = Fx + e, \qquad \mathscr{E}(e|x) = 0, \qquad \mathscr{E}(ee'|x) = \Sigma.$$

Given the grouping transformations $x^* = Gx$ and $y^* = Hy$, our problem is then to find an aggregative model,

$$(4) \qquad y^{**} = F^*x^* + e^*, \qquad \mathscr{E}^*(e^*|x^*) = 0, \qquad \mathscr{E}^*(e^{*\prime}|x) = \Sigma^*,$$

which best approximates (3) in the sense of minimising the mean squared forecast error

$$(5) \quad \mathscr{E}(F^*x^* - y^*)(F^*x^* - y^*)' = (F^*G - HF)\,M(F^*G - HF)' + H\Sigma H',$$

where $y^* = Hy$. The second term on the right in (5) is a constant, and the first term is the aggregation bias, which is minimised if and only if (cf. Chipman [2])

$$(6) \qquad F^* = HFG^\# + Z^*(I - GG^\#),$$

where Z^* is arbitrary and $G^\#$ is any matrix satisfying

$$(7) \qquad \text{(i)} \quad GG^\#GM = GM, \qquad \text{(iv)} \quad G^\#GM = (G^\#GM)'.$$

I call such a matrix $G^\#$ a *generalised quasi-inverse* of G. In the case in which M

is positive definite, properties (i) and (iv) of (7) are equivalent to the corresponding properties of the oblique version (cf. Chipman [1]) of the Moore-Penrose generalised inverse of G (cf. Penrose [9]) — the latter corresponding to the special case in which $M = I$. The transformation (6) is, by virtue of (7), equivalent with probability 1 to $F^* = HFG^{\#}$, and F^*x^* may be called the *best homogeneous linear predictor* of y^* given x^*. Given our convention that the first components of x and x^* are identically 1, it also coincides with what Doob ([3], p. 77) designates as the *wide sense conditional expectation* of y^* given x^*, denoted $\hat{\mathscr{E}}(y^*|x^*) = F^*x^*$; the transformation $G^{\#}: \mathscr{X}^* \to \mathscr{X}$ is, in turn, the wide sense conditional expectation of x given x^*, $\hat{\mathscr{E}}(x|x^*) = G^{\#}x^*$.

Substituting (6) in (5) and making use of (7) we see that the minimum aggregation bias is equal to

$$(8) \qquad HF(I - G^{\#}G)M(I - G^{\#}G)'F'H' = HF(I - G^{\#}G)MF'H',$$

and this reduces to zero if and only if $HF(I - G^{\#}G)M = 0$, in which case we may say that aggregation is 'almost perfect'. In Case (a) we assume that M has full rank; perfect aggregation then holds whenever F satisfies the bilinear restriction $HF(I - G^{\#}G) = 0$. Since in this case $G^{\#}$ is a generalised inverse of G, this bilinear restriction is precisely Penrose's necessary and sufficient condition for the existence of a solution F^* to the equation $F^*G = HF$ (cf. Penrose [9]). In the case of our international trade example, this condition defines certain structural similarities among industries whose prices are grouped together to form price indices. In Case (b) we assume that observations on x are constrained to lie in the k^*-dimensional subspace $\mathscr{X}^0 = G^{\#}G\mathscr{X} \subset \mathscr{X}$, with probability one, i.e., $(I - G^{\#}G)M = 0$. In our international trade example, this would correspond to a case in which the k prices could be partitioned into k^* collinear groups.

Let us now take stock of the situation. I have characterised the aggregation problem as a two-fold one: (1) Given certain modes of aggregation $g: \mathscr{X} \to \mathscr{X}^*$, $h: \mathscr{Y} \to \mathscr{Y}^*$, such as summing and forming weighted indices, what is the best choice of an aggregative model $f^*: \mathscr{X}^* \to \mathscr{Y}^*$? (2) Given a best choice of f^* for each choice of g and h, what is the best choice of the modes of aggregation g and h? We have obtained a solution of problem (1), which is conceptually a very simple one: f^* should be the conditional expectation of $y^* = h(f(x))$ given $x^* = g(x)$. In the case in which f, g, and h are affine transformations, and f^* is also required to be an affine transformation, the solution is computationally simple as well: it consists in calculating a generalised quasi-inverse $G^{\#}$ from G and data on M, and then computing $F^* = HFG^{\#}$. It is very unsatisfactory to leave the problem there, however, since there is no guarantee that the given choices of modes of aggregation g and h will exploit to the fullest the structural similarities and multicollinearities that may be present.

To accomplish this we must proceed to problem (2), which may be formulated in the following manner: Let there be given a set $\mathcal{G}$ of pairs of grouping mappings (g, h), where $g: \mathcal{X} \to \mathcal{X}^*$ and $h: \mathcal{Y} \to \mathcal{Y}^*$. For each such pair (g, h) we may in principle determine the function f^* which minimises (1), namely $f^*(x^*) = \mathcal{E}(y^* | x^*)$, and the corresponding minimising value of the bias matrix (1), which is a function of g and h. Finally, we may select that pair $(g, h) \in \mathcal{G}$ which minimises the norm of (1), i.e., the square root of its trace.

The above method of solving problem (2) is simple conceptually but intractable computationally, since the dependence of $\mathcal{E}(y^* | x^*)$ on g and h is not in general of any simple kind. If we require f, g, h, and f^* to be affine, the problem becomes a good deal simpler. Defining the M-norm of $F^*G - HF$ by

$$\|F^*G - HF\|_M = \sqrt{(\text{trace}\,(F^*G - HF)\,M\,(F^*G - HF)')} \tag{9}$$

and substituting (6) in (9) we obtain, as in (8),

$$\inf_{F^*} \|F^*G - HF\|_M = \sqrt{(\text{trace}\,HF\,(I - G^{\#}G)\,MF'H')}. \tag{10}$$

Defining a 'grouping matrix' G as a matrix with non-negative elements and at most one positive element in each column, we may consider the set $\mathcal{G}$ of pairs (G, H) of grouping matrices of given orders, and select that pair which minimises (10). This is a problem of the quadratic programming type. An economist would probably further want to restrict $\mathcal{G}$ so that the indices Gx and Hy were economically interpretable, which would mean limiting the modes of aggregation to summing and forming price indices with natural quantity weights; the problem would then take on aspects of integer programming problems. Considerable work on such types of problems has been carried out by W. D. Fisher [5], in a related formulation, but much work remains to be done to obtain an efficient computational algorithm. One would also wish to relax the assumption that the spaces $\mathcal{X}^*$ and $\mathcal{Y}^*$ have fixed dimensionality, and allow this to be one of the questions to be determined.

Before proceeding to discuss the statistical estimation problem, I would like to say something about the *disaggregation problem*. As has been suggested by W. D. Fisher [5], problems of aggregation and disaggregation naturally arise in a setting in which economic investigations are carried out in a decentralised fashion by different agencies. We can think of there being two types of investigators: researchers in universities and advisers to policy makers in government. The latter would like to have a model they can use for making forecasts of the effects of changes in taxes and government expenditures, the effects of devaluation and of changes in world oil prices, etc. And they want to make such calculations speedily on the basis of fairly simplified models. In this scheme of things, we can think of the role of the university researchers as being

that of estimating the model (the aggregative model, that is), and the role of the government advisers as being that of estimating the parameters of this model and making forecasts on the basis of it. But now, even though the government advisers like to work with these very crude models, nevertheless in many cases they do have to make detailed forecasts. It is not enough just to make a forecast that the general price level, or even the general price of food, will be such and such; you would want them to be able to make predictions about the price of meat, or even about the prices of beef and pork and lamb. It is not enough to make predictions about the general level of employment; you would want to have predictions of the levels of employment in different industries and regions. This means that it is not enough to furnish them with an aggregative model f^*: $\mathscr{X}^* \to \mathscr{Y}^*$; they also need to be supplied with a disaggregation rule $h^\#$: $\mathscr{Y}^* \to \mathscr{Y}$ (see Figure 1). Defining the *disaggregation bias* by the matrix

$$
\begin{aligned}
d(h^\# \circ f^* \circ g,\ f) &= \mathscr{E}[(h^\# \circ f^* \circ g)(x) - f(x)][h^\# \circ f^* \circ g)(x) - f(x)]' \\
&= \mathscr{E}[h^\#(\hat{y}^*) - y][h^\#(\hat{y}^*) - y]', \qquad (11)
\end{aligned}
$$

where $\hat{y}^* = f^*(x^*) = \mathscr{E}(y^* \mid x^*)$, (11) is minimised with respect to $h^\#$ when $h^\#(\hat{y}^*) = \mathscr{E}(y \mid \hat{y}^*)$. In the case in which all the mappings are assumed to be linear transformations, the optimal 'degrouping mapping' becomes $h^\#(\hat{y}^*) = \hat{\mathscr{E}}(y \mid \hat{y}^*) = \hat{\mathscr{E}}(\tilde{y} \mid \hat{y}^*)$, where $\tilde{y} = (f \circ g^\# \circ g)(x), g^\#(x^*) = \hat{\mathscr{E}}(x \mid x^*)$, and $\hat{y}^* = h(\tilde{y})$; the corresponding 'degrouping matrix' $H^\#$ is then a generalised quasi-inverse of H with respect to

$$W = \mathscr{E}\tilde{y}\tilde{y}' = FG^\#GMG'G^{\#\prime}F' = FG^\#GMF'$$

(satisfying (7) where the symbols G and M are replaced by H and W). In effect, the model F is then 'estimated' by $\hat{F} = H^\#HFG^\#G$. The procedure might be described as 'non-Archimedean Bayesian,' since first $G^\#$ is chosen so as to minimise aggregation bias, and then, given this choice, $H^\#$ is chosen so as to minimise disaggregation bias. In a third stage, one could then select $(G, H) \in \mathscr{G}$ so as to minimise $\|F - \hat{F}\|_M$.

3. Statistical aspects

The discussion up to this point cannot be said to come under the heading of 'statistics' in the sense in which this term has generally been understood since the time of R. A. Fisher, although it surely conforms to the broad definition supplied by Fisher himself, namely the 'reduction of data' [4]. The treatment is, nevertheless, still incomplete, since nothing has been said so far about how one should go about obtaining numerical estimates of F^* and $\hat{F}$ from empirical data. To this subject I now finally turn.

According to the customary notation used in regression analysis, in which the order of matrix multiplication is reversed, we may write the multivariate multiple regression model corresponding to (3) in the form

$$(12)\quad Y = XB + E, \qquad \mathscr{E}(E \mid X) = 0, \qquad \mathscr{E}\{(\text{row } E)'(\text{row } E) \mid X\} = V \otimes \Sigma,$$

where Y is an $n \times m$ matrix of n random observations on the m jointly dependent variables, X is an $n \times k$ matrix of n observations on the k independent variables, and B (taking the place of F') is the $k \times m$ matrix of regression coefficients; 'row E' stands for the row vector of rows of the $n \times m$ matrix E of error terms, Σ is the $m \times m$ 'contemporaneous covariance matrix', and V is the $n \times n$ sample covariance matrix — assumed to have positive but not necessarily full rank; $\otimes$ is the symbol for Kronecker multiplication. It will be assumed that X is a random variable having the same autocovariance structure as $E \mid X$; this may be expressed by specifying it as the dependent term in the multivariate multiple regression model

$$(13)\quad X = \iota\mu + U, \qquad \mathscr{E}U = 0, \qquad \mathscr{E}\{(\text{row } U)'(\text{row } U)\} = V \otimes \Theta,$$

where ι is the column of n ones, μ a row of k means, and Θ the $k \times k$ contemporaneous covariance matrix of the independent variables in (12). Given the grouping transformations

$$(14)\quad X^* = XG, \qquad Y^* = YH$$

where G and H are $k \times k^*$ and $m \times m^*$ column-wise grouping matrices, we define

$$(15)\quad M = \mu'\mu + \Theta, \qquad G^{\#} = (G'MG)^{-}G'M.$$

We now seek a consolidated multivariate multiple regression model

$$(16)\quad Y^{**} = X^*B^* + E^*, \qquad \mathscr{E}^*(E^* \mid X^*) = 0,$$
$$\mathscr{E}^*\{(\text{row } E^*)'(\text{row } E^*) \mid X^*\} = V \otimes \Sigma^*$$

which best approximates (12). Moreover, we seek a suitable estimator of B^*.

Instead of discussing optimal estimation procedures, I will confine myself to a much more limited question: Under what circumstances, and according to what criteria, can the generalised least squares (Gauss-Markoff) procedure, as applied to the model (16), be justified or rationalised when the true model is assumed to be (12)?

A bi-affine function $\tilde{B} = AYK + C$ will be said to be a *Gauss-Markoff estimator* of B with respect to the model (12) if, for all estimable bilinear functions $\psi B\phi$ (i.e., functions for which there exists $\tilde{B}$ such that $\mathscr{E}(\psi\tilde{B}\phi \mid X) =$

$\psi B\phi$ for all B), $\psi\tilde{B}\phi$ has minimum variance (conditional on X) in the class of bi-affine functions of Y. We can call $\psi\tilde{B}\phi$ the *best bilinear unbiased estimator* of $\psi B\phi$. It can be shown that the class of Gauss-Markoff estimators of B is given by $\tilde{B} = X^{\ddagger}Y + (I - X^{\ddagger}X)Z$, where Z is arbitrary and $X^{\ddagger}$ is any matrix (which always exists) satisfying $XX^{\ddagger}X = X$ and $XX^{\ddagger}V = (XX^{\ddagger}V)'$. (In fact, it can be shown that col $\tilde{B}$ — the column vector of columns of $\tilde{B}$ — is Gauss-Markoff within the class of affine functions of col Y; cf. [2].) Now, let us assume that an investigator employs an estimator $\tilde{B}^*$ of B^* which is Gauss-Markoff *relative to* the model (16), i.e., which would be Gauss-Markoff if (16) were the true model; in particular, we may choose

$$\tilde{B}^* = (X^{*\prime}V^{+}X^*)^{-}X^{*\prime}V^{+}Y^* + [I - (X^{*\prime}V^{+}X^*)^{-}X^{*\prime}V^{+}X^*]Z^*, \tag{17}$$

where A^{-} denotes a generalised inverse of the matrix A, i.e., any matrix satisfying $AA^{-}A = A$ (cf. Rao [10]), and where V^{+} is a matrix satisfying (a) $VV^{+}V = V$, (b) $VV^{+\prime}X = X\Gamma$ for some Γ, and (c) rank $X'V^{+}X$ = rank X (cf. Zyskind and Martin [14]). It will also be assumed (as is always possible) that V^{+} is symmetric. Let us define

$$\tilde{M} = (\iota'V^{+}\iota)^{-1}X'V^{+}X,\ \tilde{G}^{\#} = (G'\tilde{M}G)^{-}G'\tilde{M} = (X^{*\prime}V^{+}X^*)^{-}X^{*\prime}V^{+}X. \tag{18}$$

Then we can readily see that the estimator (17) can be expressed, with probability 1 (conditional on X), as

$$\tilde{B}^* = \tilde{G}^{\#}\tilde{B}H. \tag{19}$$

On the other hand it can be shown (cf. Chipman [2]) that if the first column of X (and of X^*) is a column of ones, and the remaining columns of X (and hence of X^*) are contained in the column space of V (in which case any V^{+} satisfying (a) above also satisfies (b) and (c)), then

$$\hat{\mathscr{E}}(Y^*|X^*) = X^*G^{\#}BH \qquad \text{with probability 1.} \tag{20}$$

Thus, aggregation bias $\mathscr{E}$ [row $X(GB^* - BH)$]$'$[row $X(GB^* - BH)$] is minimised when B^* in (16) is chosen to be $B^* = G^{\#}BH$. A Gauss-Markoff estimator of this B^* (relative to the model (12)) is then given by $G^{\#}\tilde{B}H$. Accordingly, to justify (19) as an estimator of $G^{\#}BH$ we need to justify $\tilde{G}^{\#}$ as an estimator of $G^{\#}$, or equivalently, $\tilde{M}$ as an estimator of M.

Defining $\iota^{\ddagger} = (\iota'V^{+}\iota)^{-1}\iota'V^{+}$, and the estimators

$$\tilde{\mu} = \iota^{\ddagger}X, \qquad \tilde{\Theta} = (\iota'V^{+}\iota)^{-1}\bar{X}'V^{+}\bar{X}, \qquad \text{where} \quad \bar{X} = (I - \iota\iota^{\ddagger})X, \tag{21}$$

we see that $\tilde{\mu}$ is the Gauss-Markoff estimator of μ in (13) and that

$$\tilde{M} = \tilde{\mu}'\tilde{\mu} + \tilde{\Theta}. \tag{22}$$

Under certain conditions, $\tilde{\Theta}$ is a consistent estimator of Θ; then, $\tilde{M}$ can be justified as an estimator of M, and therefore (17) of $B^* = G^{\#}BH$.

The special cases of perfect aggregation may be briefly considered. In Case (a), if X and Σ have full column rank and moreover rank XG = rank $G = k^*$ and rank ΣH = rank $H = m^*$, and if G^- is any left inverse of G, then it can be shown that (17) is the best bilinear conditionally unbiased estimator of G^-BH subject to the bilinear restriction $(I - GG^-)BH = 0$. Note that in this case, $G^{\#}BH = G^{\#}GG^-BH = G^-BH$, and likewise $\tilde{G}^{\#}BH = G^-BH$, so no question of estimating M arises. In Case (b), if it is assumed that for some diagonal matrix D we have rank XG = rank $G'DG$ = rank $G = k^*$, and that $X = XGG^{\ddagger}$ where $G^{\ddagger} = (G'DG)^{-1}G'D$, then it follows that $\tilde{G}^{\#} = G^{\ddagger}$ hence, in view of (19), (17) is the best bilinear unbiased estimator of $G^{\ddagger}BH$. Here, one must have recourse to the above argument to justify $\tilde{B}^*$ as an estimator of $G^{\#}BH$.

It is often the case in applications that a bilinear restriction such as $(I - GG^{\#})BH = 0$ is unlikely to be fulfilled — whether exactly or approximately — except in conjunction with additional restrictions on B. In such a case, Gauss-Markoff estimation is not fully efficient. This fact was first noticed, and exploited, by Zellner [13], and has long been recognised in the context of simultaneous equations models, of which (12) is the 'reduced form' (cf. Koopmans *et al.* [7]). Adding to this the fact that in large systems such as (12) considerable improvement over Gauss-Markoff estimators (in terms of reduction of mean square error) appears to be possible (cf. James and Stein [6]), we may conclude that, in the aggregation field, much work remains to be done.

Acknowledgment

The research reported here was supported in part by Ford Foundation Grant 750–0114. I wish to thank Sudhish Ghurye and Dit Sang Ho for calling my attention to errors and ambiguities in an earlier draft. They are not responsible, however, for any that remain.

References

[1] CHIPMAN, J. S. (1964) On least squares with insufficient observations. *J. Amer. Statist. Assoc.* **59**, 1078–1111.

[2] CHIPMAN, JOHN S., Estimation and aggregation in econometrics, in Nashed, M. Z. (ed.), *Generalized Inverses and Applications*, Academic Press, New York (to appear.).

[3] DOOB, J. L. (1953) *Stochastic Processes*. Wiley, New York.

[4] FISHER, R. A. (1922) On the mathematical foundations of theoretical statistics. *Phil. Trans. Roy. Soc.* A, **222**, 309–368.

[5] FISHER, W. D. (1969) *Clustering and Aggregation in Economics*. Johns Hopkins Press, Baltimore.

[6] JAMES, W., AND STEIN, C. (1961) Estimation with quadratic loss. *Proc. Fourth Berkeley Symp. Math. Statist. Prob.* **I**, 361–379.

[7] KOOPMANS, T. C., RUBIN, H. AND LEIPNIK, R. B. (1950) Measuring the equation systems of dynamic economics. In Koppmans, T. C. (ed.), *Statistical Inference in Dynamic Economic Models*, Wiley, New York pp. 53–237.

[8] MALINVAUD, E. (1956) L'agrégation dans les modèles économiques. *Cahiers du Séminaire d'Econométrie*, No. 4, 69–146.

[9] PENROSE, R. (1955) A generalized inverse for matrices. *Proc. Camb. Phil. Soc.* **51**, 406–413.

[10] RAO, C. R. (1966) Generalized inverse for matrices and its applications to statistics. In David, F. N. (ed.), *Festchrift for J. Neyman: Research Papers in Statistics*, Wiley, New York pp. 263–279.

[11] THEIL, H. (1954) *Linear Aggregation of Economic Relations*. North-Holland Publishing Company, Amsterdam.

[12] U.S. Bureau of the Census, *U.S. Imports — General and Consumption, Schedule A Commodity and Country*, Report FT 135, and *U.S. Exports — Schedule B Commodity by Country*, Report FT 410, January 1974. U.S. Government Printing Office, Washington D.C.

[13] ZELLNER, A. (1962) An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *J. Amer. Statist. Assoc.* **57**, 348–368.

[14] ZYSKIND, G. AND MARTIN, F. B. (1969) On best linear estimation and a general Gauss-Markov theorem in linear models with arbitrary nonnegative covariance structure. *SIAM J. Appl. Maths* **17**, 1190–1202.

Suppl. Adv. Appl. Prob. 7, 84–89 (1975)
Printed in Israel

APPLICATIONS VS. ABSTRACTION: THE SELLING OUT OF MATHEMATICAL STATISTICS?

PETER J. HUBER, *Swiss Federal Institute of Technology, Zürich*

Seven years ago there was a conference on 'The Future of Statistics' [1]. Curiously, only one speaker addressed himself directly to that topic, and afterwards a reviewer noted that a conference with such a title implicitly questioned whether statistics had a future at all. Reassuringly, the field has survived so far, but we now attend another conference motivated by a concern about the future. Also among the present speakers few have stuck their neck out and have addressed themselves to the stated topic 'Directions for Mathematical Statistics'.

This preoccupation with the future, combined with a reluctance to talk about it, surely suggests that something is wrong with today's mathematical statistics. The organiser of the present meeting even offers a precise diagnosis: '... the pursuit of mathematical abstraction in statistics for its own sake is a dead-end street; [the conference] is aimed at emphasising the need for relevance to real problems.' At yesterday's lunch somebody suggested the Chinese solution: that every individual devote some percentage of his time to manual labor, that is, to applied problems.

I do not think that this diagnosis is correct. In my opinion it would be a grave mistake to seek the cause of the difficulties in the pursuit of abstraction *per se*. But the diagnosis contains enough truth that it might conceivably become quite dangerous to the field of mathematical statistics, namely by suggesting inappropriate corrective measures. After yesterday's talks and discussions I was afraid we were quite close to a great cultural revolution with purges and public confessions.

To begin with, I would like to rise to the defence of honest-to-goodness pursuit of mathematical abstraction. It is certainly not a dead-end street: at worst it leads straight into pure mathematics, and I hope we are never going to close that border!

In mathematical statistics one needs mathematical abstraction to retain intellectual control of the developments. For instance, we need rigorous optimality results (in some idealised, but typical cases) so that we can reasonably decide when it is no longer worthwhile to push for a better procedure in a dirty practical situation. The theoretically optimal procedure, if

there is any, might be worse than useless (perhaps it neglects a side issue like the cost of computation), but maybe we have a feasible alternative which comes within an epsilon of the performance of the ideal one.

A personal experience from the Princeton robustness study [2] may serve to illustrate another point: without the help of some highly abstract notions like Volterra derivatives of functionals, weak convergence and the like, we would have stood helpless in front of a bewildering plethora of estimates. As usual, insight comes only with thinking in models.

I shall now try to elaborate my own diagnosis of the situation. Clearly, I have been much influenced by Thomas Kuhn's ideas on the structure of scientific revolutions [3]. In statistics as well as in any other field of applied mathematics (taken in a wide sense), one can usually distinguish (at least) three phases in the development of a problem. In Phase One, there is a vague awareness of an area of open problems, one develops *ad hoc* solutions to poorly posed questions, and one gropes for the proper concepts. In Phase Two, the 'right' concepts are found, and a viable and convincing theoretical (and therefore mathematical) treatment is put together. In Phase Three, the theory begins to have a life of its own, its consequences are developed further and further, and its boundaries of validity are explored by leading it *ad absurdum*; in short, it is squeezed dry.

All three stages are indispensable. Many of the best statistical procedures are invented already in the groping phase, but their merits are accurately understood only later, perhaps only after considerable squeezing. The middle phase is transient and in extreme cases may consist of a single paper or book. Without a fair amount of Phase Three one would never be able to identify the limitations and shortcomings of a certain theory.

Unfortunately, the 'gropers' and the 'squeezers' tend to have an almost infinite disdain for each other. The squeezers will point out, quite justly, that in statistics groping contains little (and often poor) mathematics, and that there is no place for it in a journal like the *Annals of Mathematical Statistics*. After all the Institute of Mathematical Statistics and the *Annals of Mathematical Statistics* were founded to create an organisation and an outlet for the mathematically minded statisticians. On the other hand, squeezing very soon will give results no longer directly relevant to applications, and it will therefore be rejected by a majority of the gropers.

In view of this antagonism it is difficult to keep a reasonable balance. It appears that over the years unnecessarily many papers of the squeezing type have passed the editorial gates of the *Annals*, while even first-rate papers of the groping type would not even have been submitted. I think the recent change in the title of the *Annals* was a deliberate step in the right direction for several reasons, among them: these double names always suggest something inferior to each of the two parts. (Once I mentioned such a double name when talking to

a colleague in our chemistry department, and he quickly countered: Is this another cow-horse—an animal that runs like a cow and yields milk like a horse?) But I do not think that too much abstract mathematical work is going on in mathematical statistics. On the contrary, there is perhaps too much work which is neither good mathematics nor applicable statistics. And anyhow, it is tempting to go on with squeezing even after the last drop of juice has been pressed out.

Against this background, deeper reasons for anxiety about the future of mathematical statistics become apparent: too many of the activities belong to the later stages of Phase Three. But this is only a symptom, not the cause of the trouble; the cause is of course that most areas of today's mathematical statistics have passed their Phase Two quite some time ago. Therefore, and perhaps somewhat paradoxically, I expect salvation not from cutting back on abstraction, but on the contrary, from new areas reaching Phase Two, the principal phase of abstraction!

Now, what are the directions into which mathematical statistics might or should develop in the next few years? It is risky to make predictions, but it should at least be possible to identify underdeveloped areas, which have not yet reached Phase Two. Whether these potential growth areas will ever get there, is of course a different question.

It has been said that statistics is the art of collecting and interpreting data. I shall therefore subdivide statistics into the following four broad areas:

(1) Design of experiments
(2) Data acquisition
(3) Data analysis
(4) Inference: estimates, tests, Bayesian inference...

I omit stochastic modelling on purpose, despite its importance — it belongs so much to the particular field of application that it is difficult to discuss in a broad and general framework.

For historical reasons, statisticians have been preoccupied mainly with the first and the last of these areas, design and inference. I feel that the first by now is rather overdeveloped. For instance, the sociologists of my acquaintance tend to use such sophisticated stratified designs that it is difficult to draw inferences — they have lost the advantages of randomisation offered by a simple random sample. Also from the point of view of robustness the 'naive' designs (e.g., uniform allocation of observations) appear to be preferable to the more sophisticated ones [4]. The fourth area, inference, is in a fairly good shape, too. There has been considerable activity here, but I suspect — and I hope that Professor Lindley will not shoot me — that the recent trends like Bayesian inference or robustness have peaked by now; we can expect a considerable number of elaborations, but I would be surprised if there were any radical new

departures from those ideas which are now in the pipeline. For instance in robustness, the experience gained with simple estimates of location — both from rather abstract theory and from Monte Carlo experiments interpreted with its aid — now at several places has led to definite proposals for general regression calculations which are currently being implemented and tested in complicated real life situations. Covariance analysis would be next in line.

The impact of the computer age on this area (inference) has been subtle but critical: there is more data to be processed, this can only be done by computer, and this now means that all the pattern recognition and all the semi- and subconscious checking which an intelligent human calculator would have performed almost automatically, have to be made formal and explicit, or else they will go down the drain. In other words, we must 'robustify' our procedures, and informal data analysis must get formal recognition.

The areas of data acquisition and data analysis have in fact been step-children of professional statistics — in principle, they had been left to the experimenter. Data acquisition has made tremendous advances during the past decade, in particular in the natural sciences, thanks to the emergence of minicomputers as automatic collection devices, but also in the social sciences (with the availability of various data banks).

There have been some significant innovations in data analysis. I would certainly name numerical spectrum analysis (which belongs rather here than in estimation) and non-metric multidimensional scaling [5]. Then there has been a frantic, but perhaps somewhat disorganised and bewildering, activity in cluster analysis in the past few years. As a common feature, many of the newer data analytic methods tend to use fairly sophisticated transformations of all kinds of data into visual patterns, to be inspected by the human eye.

But, as a whole, data analysis has not kept up with the advances of the other areas. I have already mentioned that the informal and semi-conscious scrutiny of the data by the human calculator is lost as soon as the data is processed by machine. To counterbalance this, it has become increasingly popular to offer computer-drawn standard graphs, like 2-dimensional scatter plots, histograms and normal plots, sometimes on the line-printer, and more conveniently through some interactive graphics terminal with hardcopy facility. But the approach has remained rather pedestrian in character: the computer merely did what the statistician could easily have done with pencil and graph paper. Perhaps also here the philosophy of the 'low-cost terminal' has slowed down more imaginative approaches, cf. [6], pages 399f.

There have been only a few isolated attempts to combine the highly developed faculty of human beings to recognise spatio-temporal patterns with the superior data-handling capacity of the electronic computer (the most exciting among these perhaps are J. W. Tukey's pioneering experiments with

the interactive analysis of up to 9–dimensional scatter plots). Pattern recognition still is much too difficult a task to be left entirely to a machine, and it will certainly remain so for decades to come. It is not so that pattern recognition would not be amenable to a mathematical treatment; there is even a perfectly general and straightforward definition (in terms of the shortest program which produces the given pattern, compare [7]), for which one easily proves that no machine can do it! The problem is to find a more restricted formulation which would allow a feasible solution.

Evidently this whole field of data analysis is very much in the groping phase; several of its aspects should be susceptible to a unifying and clarifying systematic approach, well fitting into mathematical statistics.

The reader may have wondered that I have not mentioned the many interesting actual and potential applications of mathematical statistics to various scientific fields (apart from a passing reference to stochastic modelling). There are two reasons. The main one is that in my opinion these applications should not be claimed for statistics, but should be counted with the particular field of application and should in principle also be published there. It is clear also to me that mathematical statistics and mathematical statisticians need the challenge of ever new applied problems as their driving force. But I think that at this time mathematical statistics as a field can give more to the various applied fields than it can receive from them: there is more need for tailor-made models and procedures in these fields now than for systematic unification. But this is an enormously difficult task; the really deep and innovative applications can only be done by somebody who is fully competent in both fields, in statistics and in the particular field of application, and who is motivated by the latter. Thus, these are really directions for individual mathematical statisticians, not for the field of mathematical statistics (and this was my second reason for not mentioning the applications before). Institutionally, one might facilitate this mutual interaction and formation of bi-specialists for instance by allowing that a doctoral student can have a joint affiliation with statistics and with another field, where he does his research work: whether his degree is ultimately awarded as one in statistics or in that other field is immaterial, it is only important that it is good research work.

We must avoid, by all means, the creation of a new sub-branch of statistics — applied mathematical statistics — which would combine the worst features of all those components of its triple name.

References

[1] WATTS, D. G. (ed.) (1968) *The Future of Statistics.* Academic Press, New York and London.

[2] ANDREWS, D. F., BICKEL, P. J., HAMPEL, F. R., HUBER, P.J., ROGERS, W. H. AND TUKEY, J. W. (1972) *Robust Estimates of Location.* Princeton University Press.

[3] KUHN, T. S. (1962) *The Structure of Scientific Revolutions*. University of Chicago Press.

[4] HUBER, P. J. (1974) Robustness and designs. In *A Survey of Statistical Design and Linear Models*, ed. J. N. Srivastava. North-Holland Publishing Co., Amsterdam.

[5] KRUSKAL, J. B. (1964) Multidimensional scaling. *Psychometrika* **29**, 1–27 and 28–42.

[6] NEWMAN, W. M. AND SPROULL, R. F. (1973) *Principles of Interactive Computer Graphics* McGraw-Hill, New York.

[7] MARTIN-LÖF, P. (1966) The definition of random sequences. *Information and Control* **9**, 602–619.

Supp. Adv. Appl. Prob. 7, 90–99 (1975)
Printed in Israel

CULTURAL AND BIOLOGICAL EVOLUTION: A THEORETICAL INQUIRY

L. L. CAVALLI-SFORZA, *Stanford University*

The bases of a mathematical theory for biological evolution were created around the 1920's, mostly by the independent work of R. A. Fisher, J. B. S. Haldane and S. Wright. The depth and breadth of developments that have taken place since that heroic phase is witnessed by the number of papers in existence today. J. Felsenstein has recently collected the relevant bibliography. J. Crow and M. Kimura have expounded in a lucid book the major mathematical developments. L. Cavalli-Sforza and W. Bodmer have summarised the applications to Man.

One should not underestimate the difficulties of applying a mathematical theory to biological data. Observations are almost always limited in number, and there are many sources of error. Only in a few organisms can one hope, by present techniques, to estimate with some accuracy basic quantities like, for instance, mutation rates. Intensities of selection have been evaluated only in a few cases. An assumption frequently made — that selection coefficients are constant — is of doubtful validity. It is only seldom that population sizes, migration and, in general, population structure can be studied with some detail. In many respects Man is an organism of choice for study; but again, there are serious limitations. Thus, data on population structure are fairly easily obtained been living human populations, but inference on earlier situations is very uncertain. Even so, it has been possible to apply the theory in a number of circumstances, and it has benn reassuring to find that theoretical predictions were verified, within limits of error. Moreover, the mathematical theory of biological evolution has provided a conceptual framework which has helped considerably the understanding of evolutionary phenomena.

Some of the major problems are still largely unsolved. Thus, the roles of chance and of natural selection in determining molecular evolution (basically, the substitutions of amino-acids in proteins) are unclear and a source of debate.

Nobody would doubt that adaptive phenomena are of importance, but the observations leave room for some non-adaptive variation due to evolution for 'neutral' genes.

Factors of biological and of cultural evolution

All evolutionary processes are basically similar, whichever the objects that evolve. Biological evolution concerns living organisms. Natural *selection* is the consequence of the individual variation among organisms of a species in the capacity to survive to adulthood and to leave progeny. Those individuals who are more successful in these respects contribute proportionately more of their genes to the future generations. In the process of *transmission* of genes from one generation to the next there happen rare, transmissible changes by a random process called *mutation*; it is these changes which form individual, transmissible variation, on which natural selection operates when determining evolution. In addition, in every population of finite size there are sampling effects which generate stochastic variation, and thus add randomness (random genetic *drift*) to the process. Finally the movement, active or passive, of individuals (migration) is a factor which contributes in an important way to growth, differentiation of populations and the exchange between them.

Evolving 'objects' in cultural evolution are ideas, skills, social customs and in general behaviour. They also change by a *mutational* process, which is sometimes simply passive copy error and at other times an active effort (an innovation). Unlike the genes, 'culture' (loosely defined as above) is not only *transmitted vertically* from parent to offspring but also *horizontally*, between any two or more individuals. In the past history of the human species, much teaching and imitation learning may have taken place from parent to offspring; some still does now, but there has been a shift to more complex types of transmission. Relatives other than parents, and members of the social group outside the family, from age peers to professional teachers and social leaders, have taken a greater and greater part in the formation of our cultural activity. Cultural transmission tends to take place through complex networks formed and maintained by social interactions, e.g., political and religious organisations. Mechanisms of cultural transmission and change still leave room for a role of chance; a *stochastic* variation is inevitable also in the cultural processes. In fact it may be even more pronounced there than in biological evolution, where the effective population size is usually large and determines the magnitude of the chance fluctuations; in fact, in cultural processes the equivalent number may be much smaller. In the extreme case, one social leader may determine the behaviour of a large number of subjects.

There is no difficulty in visualising the importance of *migration* in cultural evolution, which is very similar to that in the biological counterpart. There is also no difficulty in understanding that an equivalent of *natural selection* must play an important role in cultural evolution, but it is difficult to foresee ways of measuring it as convincing as those which are in use for biological evolution.

There, the darwinian 'fitness' is defined in terms of the capacity to reproduce and survive, and this is the natural measurement of adaptive value. In cultural evolution many, and perhps most, changes represent successful adaptations, but a measurement of adaptive values is more difficult to achieve because of the complex nature of cultural transmission. Consider an innovation which spreads successfully in a society and should therefore have, superficially at least, adaptive value. The spread of the innovation may actually reflect, in part at least, conformity or obedience in a well organised social network. Acceptance rates of innovations, or, in general, cultural traits, can be measured but they do not necessarily depend only on their adaptive values. They seem to be a property of the social system rather than of the advantage of the innovation itself.

Examples of models useful both in cultural and in biological evolution

Some simple theoretical developments can be used in both kinds of evolutionary processes with little or no change. A little-known example is that of the spread of an advantageous mutation. In this model, first suggested by R. A. Fisher (1937), a single advantageous mutation giving a reproductive advantage to its carriers measured by s, spreads under the pressure of natural selection, in a population whose individuals have a migration pattern comparable to ordinary diffusion (with diffusion coefficient m) at a constant radial rate of advance which is proportional to the geometric mean of s and m. A very similar model was used by D. G. Kendall (1948) for the spread of a rumour, and by J. G. Skellam (1951) for the spread of a population. An example given by Skellam is the spread of the muskrat in Central Europe. A few individuals which escaped from a breeding place were successful in establishing themselves out of captivity and in rapidly reproducing. The presence of the animals was recorded farther and farther away from their place of origin. The square root of the area occupied by muskrats increased linearly with time, as postulated by the theory.

Ammerman and I have used this model to measure the expansion of 'early farming' from the area of origin in the Near East across Europe. The rate of advance was about 1 km per year, and much of the evidence indicates that this was a spread of farming people (demic diffusion) rather than that of an innovation (which is also called 'stimulus' or cultural diffusion). In the case of spread of early farming, the basic cultural trait followed in the analysis were cereals and in particular wheat, and radiocarbon was found appropriate for dating its spread. Cereals were first cultivated in the Near East and their domestication was accompanied by a complex cultural adaptation. In some cases the adoption of the cultivation of cereals may have been due to cultural diffusion. But early farming offered the potential for a geographic expansion of

farmers. Farming allowed a considerable increase in the carrying capacity of the land. Some of the bars to reproduction of hunting-gathering cultures were at least potentially removed. Migration was favoured by the necessity of shifting agriculture at regular intervals. Much of the expansion of farmers probably took place in a near vacuum: hunter-gatherers who lived in the same regions were either rare or tended to concentrate in somewhat different ecological niches. Demic expansion of early farming is thus very likely, even if not yet entirely proved (Ammerman and Cavalli-Sforza (1971a) (1971b)).

Other traits diffused most probably culturally, that is with little or no displacement of people. But the kinetics of their spread can be given by the same model. Thus, the use of iron, originating about 1500 B.C. in Anatolia, spread at an approximately constant radial rate of 3 km per year, apparently in agreement with the Fisher-Kendall-Skellam model. Some other cultural diffusions show a more complex pattern. Pottery may have travelled even faster, so much so that the error of radiocarbon dating may be too high for following the diffusion of pottery over ranges of 1000 km or so. Pottery spread rapidly from an unknown source and was quickly adopted by the pre-ceramic farmers of the Near East and Turkey, reaching soon the western outposts of agricultural expansion; it then followed the advance of farmers, and spread through Europe with them. Unlike pottery, or even iron, copper and obsidian depended more heavily on the local availability of scarce raw material. They therefore show a more complex diffusion pattern. In general, the spread of people or of cultural traits may follow the simple model of a constant radial of expansion, especially when there are no serious complications from geographical barriers or from cultural heterogeneity that prevent free diffusion. It seems reasonable that cultural diffusion is found to be on an average faster than demic diffusion, even in early times when transportation and communications were much slower than today.

Another simple model which was introduced independently, both in biological and cultural evolution, is that of the rate of decay of evolutionary units: amino-acids in proteins, words in languages. It was shown by Zuckerkandl and Pauling that the substitution of amino-acids in proteins can be described by a negative exponential. In other words, the probability of substitution of a given amino-acid is constant over time. A substitution at a given amino-acid position takes on an average about one billion years. A complication that is not fully understood is that different proteins have different substitution rates. Some proteins are extremely conservative, like histones; at the other extreme of the range are some highly variable ones, like fibrinopeptides. Even within a protein, there may be some variation in the rate at different positions. Amino-acids of key physiological significance appear much more stable and are often totally conserved.

Entirely similar developments took place, independently, in linguistics. The rate of substitution of words in languages was at first believed to be constant. On this hypothesis, it was expected that the proportion y of words not substituted by unrelated ones over a given time t follows, as in the case of molecular evolution, a negative exponential: $y = e^{-ct}$ where c is a constant such that it takes of the order of 1000 years, on an average, for substitution of a word. It was suggested that the proportion of related words ('cognates') between two languages can be used to estimate the time of separation between the languages (Swadesh: '*glottochronology*').

It was later shown, however, that the substitution rate differs for different words. This variation is sufficiently large that, on closer inspection, the overall rate of substitution can hardly be described by a negative exponential (Kruskal *et al.*). A good fit of decay curves is obtained by the function $y = (1 + kt)^{-n}$, where k and n are constants (Sgaramella-Zonta and Cavalli-Sforza). This formula is generated by the assumption of a specific distribution of the rates of change of individual words (the 'gamma' distribution).

It is interesting that use of the relative frequency of amino-acid substitutions in molecular evolution and that of word substitutions in linguistics generate similar types of problems. In both cases there are sufficient instances of locally altered evolutionary rates that the prediction of evolutionary time on the basis of observed substitutions suffers from fairly large errors. Glottochronology in particular has limited usefulness; the evolution of a language may happen to be especially slow (e.g. in extreme isolation) or especially fast (e.g. with the onset of foreign political control). Conditions that may alter local rates of molecular evolution are not necessarily the same as those of linguistic evolution. Thus it is not surprising that in some cases the genetic similarity between human groups is not at all related to linguistic similarity.

A third model of greater sophistication than the two above may be cited, which is potentially useful in both types of evolution. It was generated (by Karlin and McGregor) to account for the distribution of mutant genes in a population; it has been applied quite successfully, but so far only in a cultural context. Karlin and McGregor's distribution assumes that a gene can mutate to a new allele with a constant probability μ. In a population of constant size N, individuals die according to a Poisson process, and are substituted by births of individuals which carry the same allele with probability $(1 - \mu)$ or have a new random one with probability μ. The total number of alleles that are possible may be finite or infinite. The distribution predicted is that at equilibrium, for the number of alleles expected to be found in $1, 2, \cdots, N$ individuals. Usually, in genetic models mutation and migration are interchangeable and thus μ may indicate both mutation and migration.

This theory was used to predict the distribution of surnames in a population

(Yasuda *et al.* (1974)). Surnames are clearly transmitted culturally, but with rules that parallel closely biological ones (in most cultures they are transmitted from fathers to children). The Poisson process, in which an individual has a constant probability of dying during all intervals of any given length, is not clearly ideal to represent the human life cycle, but the approximation thus introduced does not seem serious. The fit of the Karlin-McGregor distribution to actual surname data was extremely good. This makes it possible, given the population size N and the number of different surnames present in it, to compute the parameter μ, which is the same as mutation and immigration per generation. With surnames, mutation is very small and thus μ refers essentially to immigration (of males). Estimates of immigration, applying the theory to the observed distribution of surnames, were found to be in good agreement with those obtained independently fron a direct study of immigration data.

In general, theories for biological and cultural evolution may be more easily interchanged when transmission rules can be taken to be formally the same in the two processes. This is not always the case, and a quantitative study of the rules of cultural transmission seems appropriate.

A theory of cultural transmission

A characteristic of cultural transmission is that the cultural experience of an individual is influenced by that of a wide variety of other individuals, alive or dead, living nearby or far away. Cultural traits (including skills and behaviour-patterns) which are learned early in life are more likely to be learnt from a biological parent or parents (or an individual of the family playing the parental role). For such traits there will be considerable confounding of biological and cultural transmission. Traits learned later in life may show a lesser role of parental influence. Age peers, teachers, heroes and, today, the mass media may play an overwhelming role in the determination of such traits. A measurable trait in individual i may then have a value x which will depend on the value of the trait, X_i in a variety of individuals. The final outcome may be expressed as follows:

$$X_i = W_1X_1 + W_2X_2 + \cdots + W_jX_j + \cdots + W_NX_N = \Sigma W_jX_j \tag{1}$$

where N is the number of individuals forming the social group to which the individual belongs. The W_j are *weights* which for some purposes may be standardised so that their sum is one, although this is not always necessary. The weights indicate the relative importance that each of the 'teachers' has had in forming the trait of individual i. Many W_j values may be zero indicating no influence of the corresponding individual j. The individual whose trait X_i is being studied will almost always introduce something of its own, either because

of involuntary 'error' or because of voluntary change (which may sometimes be considered as true innovation). These contributions may by expressed by ε_i giving:

$$X_i = \Sigma W_j X_j + \varepsilon_i \tag{2}$$

For some purposes, ε_i can be considered as a random variable, and then predictions can be made of the evolution of a group and differentiation between groups. An interesting consequence is that the variation between individuals belonging to the same group (variance within a group) will soon stabilise to a small value dictated mostly by the variation of ε_i. General formulas have been given by Cavalli-Sforza and Feldman (1973).

This conclusion is of some general importance. It has not often been widely realised that the variation for many cultural traits between individuals must be low for social interaction to be possible. This is immediately clear for language. Unless individual variation is small, there is no intelligibility. Individual variation in language is adjusted through a long learning process to a minimum, compatible with full (or almost full) mutual understanding. A relatively low variation is also necessary in individual moral values, or else social life would be impossible. Only for certain skills is it useful, and tolerated or encouraged, that differentiation between individuals (or sexes, etc.) takes place. Apart from these exceptions, there is a contrast between the great amount of genetic variation among individuals of the same species or subpopulation of it, and the homogeneity for cultural traits of individuals of the same social group.

It is of some interest that the model of transmission given above (Equations (1) and (2)) can be simplified to coincide with an earlier model:

$$X_i = \tfrac{1}{2}X_{j1} + \tfrac{1}{2}X_{j2} + \varepsilon_i \tag{3}$$

by which R. A. Fisher represented the 'blending' model of inheritance. This was used in the last century by F. Galton and K. Pearson for predicting the transmission of continuous biological traits like stature and correlations between relatives, until R. A. Fisher (1918) showed that Mendelian inheritance provided a better model for the same purposes. A polygenic (multifactorial) trait can be represented by the same expression (3), with the addition of a term on the right-hand side which symbolises Mendelian segregation. Other things being equal, the Mendelian polygenic model has a higher variance between individuals than the model of blending inheritance because of σ. This provides a formal explanation of the higher variance expected for a trait transmitted biologically than for one transmitted culturally.

Expressions such as (1) indicate that the analysis of cultural transmission can gain from a matrix representation. Matrices of cultural transmission that can be thus constructed have a relationship with those used to describe social

networks. Special matrices can be given to represent transmission through (1) a parent, (2) a leader or teacher, (3) a social hierarchy and so on (Feldman and Cavalli-Sforza).

Interaction of cultural and biological evolution

Among anthropologists, use of the word 'cultural' is restricted to Man. But undoubtedly much of what has been said applies also to animals other than man, as is well summarised in the book *L'Animale Culturale* by D. Mainardi.

Still, it is unquestionable that the cultural life of Man is richer, on an average, than that of animals, mostly thanks to a well-articulated language and a highly-developed skill in making tools.

Both in Man and in animals there have been obvious cross-influences of biological and cultural evolution. A cultural activity of a given kind requires an adequate biological substratum, as the comparison of different species shows unequivocally. One can teach the meaning of a number of words to a chimpanzee or a gorilla, but only by making recourse to some form of sign language because the capacity of these Primates to vocalise is not adequate for use of our language. The variation between individuals of one species may also in part be genetic, as extreme cases of genetically-determined mental deficiencies show. But in less extreme cases the distinction of sociocultural and of biological sources of variation is much more difficult to obtain. Whenever parents themselves, or more generally the family environment, are of importance in determining a behavioural trait, correlations generated between relatives mimic closely those due to chromosomal inheritance. Only the study of adoptive relationships can help to separate biological and cultural transmission, and even then the expectations are complex (see Cavalli-Sforza and Feldman (1973)).

It is not only at the level of transmission that the joint action of genes and culture is of importance. Undoubtedly, there have been many reciprocal influences also at the evolutionary level. They are frequently so numerous that the distinction of cause and effect becomes impossible. The use of tools or of language, and the structure and nervous control of the hand or tongue, etc., are obvious examples. In a few situations causal relations may be simpler. The direct use of milk in the adult diet has apparently determined natural selection in favour of the capacity to digest milk as an adult. This capacity seems to be determined by a gene frequent in Europeans, especially in the North of Europe, and in some N. African tribes, but practically absent elsewhere. Here it seems possible that a new social custom, developed with the domestication of cattle, has started natural selection in favour of a rare gene in some populations that adopted the custom.

The reverse may also be true and the adoption of a social custom may have been facilitated by a pre-existing genetic difference between populations. Examples can be given especially for remedial activities (wearing glasses, using hearing aids or sign language, and so on). Among these, one may be cited which affects all humans, not just a handicapped segment of mankind. In a very successful book *The Naked Ape*, D. Morris lists a number of explanations that have been given for the fact that humans lost their fur. Incidentally, 'hairlessness' is a more appropriate word than 'nakedness', although clearly a less sensational one. But none of the reasons given, including one strongly favoured by Morris, carries much strength. Assume instead that the reason was cultural; some hairless mutants learnt to protect themselves from the cold by putting on some kind of cloth, most probably an animal fur. A cultural adaptation of this kind may transform an otherwise lethal mutant into one which may be advantageous in a climate with strong seasonal fluctuations, and also give to a species a better chance to adapt to almost any climatic condition.

The transition to hairlessness may have occurred in the last hundred thousand years or much earlier and we really have no idea how it happened. This suggestion is not a scientific example, but just a parable to show the kinds of interactions which there can exist between genetic and cultural adaptation.

It is difficult to predict at this stage whether the attempt to quantify the study of cultural evolution can be as successful as that of biological evolution. However, it does seem to me to be a problem which should be exciting and interesting for people working in the area of probability and statistics.

Acknowledgments

The material contained in this paper is similar in substance to that published in an article appearing in *Ateneo Parmense* and is reprinted with permission. It was also used in part for the sixth R.A. Fisher Memorial Lecture, London, 1974. Dr. S. Ghurye's help in revising the manuscript is gratefully acknowledged. Research done under grants NIH GM 8043326, GM 20467 and GM 10452.

References

AMMERMAN A. AND CAVALLI-SFORZA L. L. (1971 a) Measuring the rate of spread of early farming in Europe. *Man* **6**, 674–688.

AMMERMAN A. AND CAVALLI-SFORZA L. L. (1971 b) A population model for the diffusion of early farming in Europe. *Institute of Archaeology Research Seminar in Archaeology and Related Subjects*, Sheffield 1971, ed. Colin Renfrew. Duckworth, London (1973).

AMMERMAN A., CAVALLI-SFORZA L. L. AND WAGENER D. K. (1973) Towards the estimation of population growth in Old World prehistory. School of American Research Seminars, Sante Fe, New Mexico, January, 1973.

BODMER W. F. AND CAVALLI-SFORZA L. L. (1971) Variation in the fitness and molecular evolution. Vol. 5. Darwinian, New Darwinian and Non-Darwinian Evolution. *Proc. 6th Berkeley Symp. Math. Statist. Prob., Conference on Evolution*, 155–175.

CAVALLI-SFORZA L. L. (1971) Similarities and dissimilarities of sociocultural and biological evolution. In *Mathematics in the Archaeological and Historical Sciences*, eds. F. R. Hodson, D. G. Kendall and P. Tautu. Edinburgh University Press, Edinburgh, 535–541.

CAVALLI-SFORZA L. L. (1972) Origin and differentiation of human races (Huxley Memorial Lecture). *Man* **7**, 15–25.

CAVALLI-SFORZA L. L. (1974) The role of plasticity in biological and cultural evolution. *Ann. N.Y. Acad. Sci.* **231**, 43–59.

CAVALLI-SFORZA L. L. AND FELDMAN, M. W. (1973) Models for cultural inheritance. I. Group mean and within group variation. *Theoret. Pop. Biol.* **4**, 42–55.

CAVALLI-SFORZA L. L. AND FELDMAN, M. W. (1973) Cultural versus biological inheritance: Phenotypic transmission from parent to children (A theory of the effect of parental phenotypes on children's phenotype). *Amer. J. Hum. Genet.* **25**, 618–637.

CROW, J. AND KIMURA, M. (1970) *An Introduction to Population Genetics Theory*. Harper and Row, New York.

FELDMAN M. AND CAVALLI-SFORZA, L. L. Models for cultural inheritance: a general linear model. *J. Math. Psychol.* (submitted for publication).

FELSENSTEIN, J. AND TAYLOR, B., EDS. (1974) *A Bibliography of Theoretical Population Genetics*. National Technical Information Service, U. S. Department of Commerce. Report No. RLO–2225–5–18.

FISHER, R. A. (1918) The correlation between relatives on the supposition of Mendelian inheritance. *Trans. Roy. Soc. Edin.* **52**, 399–433.

FISHER, R. A. (1930) *The Genetic Theory of Natural Selection*. Clarendon Press, Oxford.

FISHER, R. A. (1937) The wave of advance of advantageous genes. *Ann. Eug.* **7**, 355–369.

GALTON, F. (1889) *Natural Inheritance*. Macmillan, London.

HALDANE, J. B. S. (1932) *The Causes of Evolution*. Harper and Row, New York.

KARLIN, S. (1967) The number of mutant forms maintained in a population. *Proc. Fifth Berkeley Symp. Math. Statist. Prob.* **4**, 415–438.

KARLIN, S. AND MCGREGOR, J. (1964) On some stochastic models in genetics. In *Stochastic Models in Medicine and Biology*, ed. J. Gurland. University of Wisconsin Press, Madison, Wisc.

KENDALL, D. G. (1948) A form of wave propagation associated with the equation of heat conduction. *Proc. Camb. Phil. Soc.* **44**, 591–594.

KRUSKALL, J. B., DYEN, I. AND BLACK, P. (1971) The vocabulary method of reconstructing language trees: Innovations and large-scale applications. In *Mathematics in the Archaeological and Historical Sciences*, eds. F. R. Hodson, D. G. Kendall, and P. Tautu. Edinburgh University Press, Edinburgh, 361–380.

MAINARDI, D., (1974) *L'Animale Culturale*. Rizzoli Editore, Milan.

MORRIS, D. (1967) *The Naked Ape*. Cape, London.

PAULING, L. AND ZUCKERKANDL, E. (1965) Evolutionary divergence and convergence in proteins. *Evolving Genes and Proteins*, eds. V. Bryson and H. J. Vogel. Academic Press, New York.

PEARSON, K. (1894–96) Contributions to the mathematical theory of evolution. *Phil. Trans. Roy. Soc.* A **185**, 71–110; **186**, 343–414; **187**, 253–318.

PEARSON, K. (1930) *The Life, Letters and Labours of Francis Galton*, Vol. III, pp. 1–137. Cambridge University Press.

SGARAMELLA-ZONTA L. AND CAVALLI-SFORZA L. L. (1972) A method of the detection of a demic cline. *Proceedings Workshop on Population Structure, Hawaii* (in press).

SGARAMELLA-ZONTA L. AND CAVALLI-SFORZA L. L. (1975) Variation in the rate of language evolution. To appear.

SKELLAM, J. G. (1951) Random disposal in theoretical populations. *Biometrika* **38**, 196–218.

SWADESH, M. (1952) Lexico-statistic dating of prehistoric ethnic contacts. *Proc. Amer. Phil. Soc.* **96**, 452–463.

WRIGHT, S. (1931) Evolution in Mendelian populations. *Genetics* **16**, 97–159.

YASUDA N., CAVALLI-SFORZA L. L., SKOLNICK M. AND MORONI A. (1974) The evolution of surnames. An analysis of their distribution and extinction. *Theoret. Pop. Biol.* **5**, 123–142.

Supp. Adv. Appl. Prob. 7, 100–105 (1975)
Printed in Israel

DIFFUSION PROCESSES IN POPULATION BIOLOGY

WENDELL H. FLEMING, *Brown University, Providence, R.I.*

1. Introduction

Mathematical models in population biology deal with the dynamics of natural populations, i.e., with temporal changes in the numbers of individuals and with the composition of populations containing several types of individuals. In population genetics theory, one is concerned with individuals of different genetic types. In ecological models several species interact in some way, for instance by competing for the same resource. A third example is a population in which individuals are categorized by age.

We are concerned here with stochastic population models, in particular models involving Markov diffusion processes. The random effects may arise in several distinct ways. One is from random fluctuations in an environmental parameter, for instance the population growth rate (Section 2). A second is from chance fluctuations in population numbers, or in the frequencies of different types, as individuals die and new ones are born (Section 3). In models of such phenomena, diffusion processes appear as the limits of certain Markov chains (e.g., multitype branching processes or birth-death processes) after a suitable rescaling. In Section 4 geographically structured populations are considered. In some models of geographically structured populations, the movement of individuals within the habitat where the population lives is also treated as random. In such models this introduces yet another stochastic effect.

Before proceeding further, two comments should be made. First, while this paper is part of a conference on 'Directions of Mathematical Statistics', its topic belongs to applied probability. Population biology has been a source of inspiration in statistics since the early days of Pearson and Fisher, but we do not treat statistical aspects here. Among recent work one should mention Ewens's sampling theory of selectively neutral alleles [6].

The second comment is that stochastic population models are regarded by some as a frill, those holding this view believing that progress toward understanding natural evolution comes instead by studying deterministic

Research supported by the National Science Foundation under GP-28931X2.

models. In population genetics this issue is related to the controversy whether neutral mutations play a significant role in evolution. See, for instance [2].

In neutral gene theory, genetic types may persist because of the chance effects of random sampling, and not because they convey selective advantages to individuals who have them.

2. Random environmental fluctuations

Let us consider a population having only one type of individual; and let $N(t)$ denote the number of individuals at time t. A very simple model for the change in population size is:

$$\frac{dN}{dt} = f(N). \tag{2.1}$$

If $f(N) = rN$, then the population growth is malthusian (i.e., exponential). If $f(N) = rN - \beta N^2$, then the growth is logistic.

If the environment is subject to irregular fluctuations, which are taken as random, then $N(t)$ is a stochastic process. For an introduction to this topic see [16]. In particular, in the logistic model one can suppose that either $r = r(t)$ or $\beta = \beta(t)$ is a stochastic process. Then (2.1) becomes a stochastic differential equation for $N(t)$. The probability distribution of $N(t)$ can often be found explicitly, by a suitable change of variables. (See [12], III [19] and references cited there.) For instance, suppose that $\beta = 0$ and $r(t) = \bar{r} + \sigma\, dw/dt$, for constants $\bar{r}$, $\sigma > 0$, and $w(t)$ a Brownian motion (i.e., Wiener process). Then one gets an explicit solution by introducing $Y = \log N$. The formal derivative dw/dt is called a white noise. In calculating $d(\log N)$ one must specify whether the Itô or Stratonovich stochastic differential calculus is being used. In the Itô calculus, $d(\log N)$ is computed from the Itô stochastic differential rule ([11], Chap. 8) in the Stratonovich calculus the usual formula $d(\log N) = N^{-1}dN$ holds [18].

In this example, $N(t)$ is a 1-dimensional Markov diffusion process. The generator is

$$\mathcal{G} = \frac{\sigma^2 N^2}{2}\frac{d^2}{dN^2} + \bar{r}N\frac{d}{dN} \tag{2.2}$$

if the Itô calculus is used. With the Stratonovich interpretation

$$\mathcal{G} = \frac{\sigma^2 N^2}{2}\frac{d^2}{dN^2} + \left(\bar{r}N + \frac{\sigma^2 N}{2}\right)\frac{d}{dN}. \tag{2.3}$$

3. Diffusion approximations

This method has been extensively applied in population genetics theory. It involves the introduction of a Markov diffusion process as an approximation to a rescaled Markov chain with many states. By using the limiting diffusion various quantities can often be computed, such as the equilibrium probability distribution or mean exit time from a given interval. Similar ideas have been applied to many other kinds of problems. Examples are the so-called invariance principles, in which a discrete time process consisting of sums of independent random variables is approximated, after rescaling, by a Brownian motion. Other applications include optimal stopping problems and queues. For results about convergence to the limiting diffusion process see [11], Chap. 9, [14].

The earliest application of the diffusion approximation technique in population biology was made by Feller [7]. In this paper, Feller obtained a 1-dimensional diffusion as an approximation to rescaled branching processes. The generator has the form

$$\mathcal{G} = \frac{c^2 p}{2}\frac{d^2}{dp^2} + bp\frac{d}{dp} \tag{3.1}$$

for suitable constants b, c.

Feller also considered the Wright model in population genetics. Let $p(t)$ now denote the frequency of one of two possible gene types in a population of fixed total size N. Then $0 \leqq p(t) \leqq 1$. For the somewhat more general model considered in [3], Chap. 8, the generator of the approximating diffusion has the form

$$\mathcal{G} = \frac{p(1-p)}{4}\frac{d^2}{dp^2} + g(p)\frac{d}{dp} \tag{3.2}$$

if time is measured in units of N. Here $g(p)$ is a polynomial of degree $\leqq 3$ arising from deterministic effects of natural selection and mutation. In the selectively neutral case, $g(p) = \alpha p - \beta$ is of degree 1.

It is often convenient to describe diffusion processes as the solutions to stochastic differential equations (Itô sense). For the stochastic differential equation

$$\frac{dp}{dt} = bp + cp^{\frac{1}{2}}\frac{dw}{dt}, \tag{3.3}$$

with w a Brownian motion, the generator is (3.1). For the equation

$$\frac{dp}{dt} = g(p) + \left[\frac{p(1-p)}{2}\right]^{1/2}\frac{dw}{dt} \tag{3.4}$$

the generator is (3.2). In Section 4 we shall consider corresponding stochastic partial differential equations, when the population is geographically structured.

Diffusion processes in more than one dimension are obtained by considering several types at a given gene locus, or several gene loci. For many of the known results neutrality is assumed. Sometimes the quantities of interest are certain moments of the gene frequency process, which turn out to obey a system of linear differential equations. See for instance [13], Chap. 7, for the problem of linkage disequilibrium. A truncation procedure discussed in [10] could be used to extend the method of linear differential equations to the case of near neutrality, at the expense of increasing dimensionality.

4. Geographically structured populations

In nature significant fluctuations are often observed in the composition of a population from place to place within its habitat. The composition may also fluctuate with time, for instance as a new type introduced at one location disperses throughout the habitat. In complete isolation (no dispersal) populations at each place evolve independently of each other. At the other extreme rapid dispersal mixes a population fast enough that it acts as a unit. Models of geographically structured populations are concerned with intermediate situations, which neither of these extremes fits.

One way to formulate a geographically structured model is to divide the population into discrete colonies (or niches), with certain rates of migration between colonies. This is the basis of the stepping stone model in population genetics. (See [3], Chap. 9.9, [13], Chap. 8.) Bailey [1] considered a discrete colony population model, in which a birth-death process is going on independently in each colony besides an exchange of individuals between adjacent colonies.

One can also consider populations distributed over a continuous habitat R, contained in r-dimensional space ($r = 1, 2$ or 3). In a series of papers, including [15], Malécot considered the problem of identity of genes, in a continuous habitat-discrete time model. We shall now mention some recent results about population models in which both temporal and spatial parameters are continuous. We believe that such models deserve further study, despite the technical difficulties innvolved.

One kind of model is formulated in terms of stochastically driven partial differential equations. Dawson [4], p. 313 considered the equation

$$\frac{\partial p}{\partial t} = \frac{\partial^2 p}{\partial x^2} + cp^{1/2}\frac{\partial w}{\partial t}, \tag{4.1}$$

where x denotes a point of a 1-dimensional habitat and $\partial w/\partial t$ is a space-time

white noise. One can regard $p(t,x)$ as a space-time diffusion approximation to Bailey's model with equal birth and death rates after rescaling. If (4.1) is considered foor x in a finite interval R, with either $p=0$ or $\partial p/\partial x=0$ as boundary conditions, then existence of a solution is not known. However, the means and covariances of a solution can be calculated, if the solution exists. For a two-dimensional habitat R, with $\partial^2/\partial x^2$ replaced by the Laplace operator, there is no solution $p(t,x_1,x_2)$ as a process with $E\|p(t,\cdot,\cdot)\|^2<\infty$ where $\|\ \ \|$ is the norm in the Hilbert space $L^2(R)$. In [5], Dawson circumvented this difficulty by introducing the concept of measure diffusion process. His measure-valued solution is not absolutely continuous with respect to Lebesgue measure on R. If it were, the Randon-Nikodym derivative would be the non-existent solution $p(t,x)$ of (4.1).

The state space of the measure diffusion is the space $\mathcal{M}$ of finite non-negative measures on r-dimensional space. The generator is formally the sum of two operators A_S, A_T, where A_S is determined by random dispersal alone and A_T by random births and deaths alone. Dawson's method is to first study the two corresponding semigroups $\{S_t\}$, $\{T_t\}$, defined on a Banach space of w^* continuous functions on $\mathcal{M}$. Then a theorem of Trotter is used to obtain the semigroup associated with the desired measure-valued process.

A space-time diffusion approximation to the stepping stone model in population genetics can also be made. Now $p(t,x)$ denotes the frequency of a gene type at time t, place x. The corresponding stochastic partial differential equation is then, assuming a 1-dimensional habitat and selective neutrality.

$$\frac{\partial p}{\partial t}=\frac{\partial^2 p}{\partial x^2}+\alpha p-\beta+\left[\frac{p(1-p)}{2}\right]^{1/2}\frac{\partial w}{\partial t}, \tag{4.2}$$

where α,β are positive constants. In [9] the mean and covariance of a solution were found in equilibrium, when (4.2) is considered for x in a finite interval with boundary conditions $\partial p/\partial x=0$. However, existence of a solution is unknown. For a two-dimensional habitat there is no solution. This difficulty disappears if instead of (4.2) one considers the following equation, which can also be reasonably regarded as an approximation to the stepping stone model ([8], §6):

$$\frac{\partial p}{\partial t}=\frac{\partial^2 p}{\partial x^2}+\alpha p-\beta+\left[\frac{p(1-p)_+}{2}\right]^{1/2}\frac{\partial W}{\partial t}, \tag{4.3}$$

where $a_+=\max(a,0)$ and correlations of spatial increments of the process W are allowed. If W is a Wiener process in $L^2(R)$ with covariance operator of finite trace, then Viot [20] has shown that a solution of (4.3) exists for bounded R of any dimension with zero Neumann boundary conditions. Moreover, the solution is unique and satisfies $0\leqq p(t,x)\leqq 1$.

The theory of branching diffusion processes provides another kind of continuous parameter model for geographically structured populations. Sawyer [17] used this method to study a population of 'rare' mutant genes in a uniform population of 'normal' genes. The branching diffusion formulation has the advantage that formulas can be obtained directly for various quantities of interest. Among such quantities are the mean number of mutants in a region A, the covariance of numbers in regions A, B, the probability density of a mutant at place x given that there is one at y, and the probability of identity by descent.

References

[1] BAILEY, N. T. J. (1968) Stochastic birth, death and migration processes for spatially distributed populations. *Biometrika* **55,** 189–198.

[2] CROW, J. F. (1972) Darwinian and non-darwinian evolution. *Proc. Sixth Berkeley Symp. Math. Statist. Prob.* **5,** 1–22, University of California Press, Berkeley.

[3] CROW, J. F. AND KIMURA, M. (1970) *An Introduction to Population Genetics Theory.* Harper and Row, New York.

(4) DAWSON, D. A. (1972) Stochastic evolution equations. *Math. Biosci.* **15,** 287–316.

[5] DAWSON, D. A. (1975) Stochastic evolution equations and related measure processes. *J. Multivariate Anal.* **5,** 1–52.

[6] EWENS, W. J. (1972) The sampling theory of selectively neutral alleles. *Theor. Popn Biol.* **3,** 87–112.

[7] FELLER, W. (1951) Diffusion processes in genetics. *Proc. Second Berkeley Symp. Math. Statist. Prob.* 227–246, University of California Press, Berkeley.

[8] FLEMING, W. H. (1975) Distributed parameter stochastic systems in population biology. *Proc. IRIA Symposium on Control Theory, Numerical Methods and Computer Systems Mod.*, Springer Lecture Notes in Economics and Mathematical Systems, No. 107.

[9] FLEMING, W. H. AND SU, C. H. (1974) Some one dimensional migration models in population genetics theory. *Theor. Popn Biol.* **5,** 431–449.

[10] FLEMING, W. H. AND TSAI, C. P. (1975) Some stochastic systems depending on small parameters. Proceedings of the International Symposium on Dynamical Systems, Brown University, Providence, Rhode Island, Academic Press, New York.

[11] GIKHMAN, I. I. AND SKOROKHOD, A. V. (1969) *Introduction to the Theory of Random Processes.* Saunders, Philadelphia.

[12] KIESTER, A. R. AND BARAKAT, R. (1974) Exact solutions to certain stochastic differential equation models of population growth. *Theor. Popn. Biol.* **6,** 199–216.

[13] KIMURA, M. AND OHTA T. (1971) *Theoretical Aspects of Population Genetics.* Monographs in Population **4,** Princeton University Press.

[14] KUSHNER, H. J. (1974) On the weak convergence of interpolated Markov chains to a diffusion. *Ann. Prob.* **2,** 40–50.

[15] MALÉCOT, G. (1967) Identical loci and relationship. *Proc. Fifth Berkeley Symp. Math. Statist. Prob.* **4,** 317–332, University of California Press, Berkeley.

[16] MAY, R. M. (1973) *Stability and Complexity in Model Ecosystems.* Monographs in Population Biology **6,** Princeton University Press.

[17] SAWYER, S. Branching diffusion processes in population genetics. Yeshiva University preprint.

[18] STRATONOVICH, R. L. (1966) A new representation for stochastic integrals and equations. *SIAM J. Control* **4,** 362–371.

[19] TUCKWELL, H. C. (1974) A study of some diffusion models of population growth. *Theor. Popn Biol.* **6,** 345–357.

[20] VIOT, M. (1975) Brown University Lefschetz Center for Dynamical Systems Technical Report. 75–3.

Supp. Adv. Appl. Prob. 7, 106–115 (1975)
Printed in Israel

THE FUTURE OF STATISTICS — A BAYESIAN 21ST CENTURY

D. V. LINDLEY, *University College London and University of Iowa*

The thesis behind this talk is very simple: the only good statistics is Bayesian statistics. Bayesian statistics is not just another technique to be added to our repertoire alongside, for example, multivariate analysis; it is the only method that can produce sound inferences and decisions in multivariate, or any other branch of, statistics. It is not just another chapter to add to that elementary text you are writing; it is that text. It follows that the unique direction for mathematical statistics must be along the Bayesian road.

The talk is divided into three sections. In the first I shall state the Bayesian position and explain how it differs from that which is currently popular. I had hoped that it would not be necessary to include this section, but others have persuaded me that I should. In the short time available, a complete statement is not possible; but the literature contains many better and fuller statements than can be given here.* In the second section the central thesis will be justified; and in the third I shall undertake what I see as the real point of the lecture, namely a study of future directions for statistics. It had originally been my intention to follow Orwell and and use 1984 in the title, but de Finetti (1974) suggests 2020; hence the longer time span.

1. Bayesian Statistics

The distinguishing feature of Bayesian statistics is that *all* unknown quantities are random variables: not just the data, but other variables, like parameters, are, before they are observed, random. The act of observation changes the status of the quantity from a random variable to a number. Here are two examples.

Example 1. Consider n Bernoulli trials in which an event occurs on just r of them. The usual random mechanism is governed by a parameter, the chance of the event occurring in a single trial, usually denoted by p, but here by θ. You will be used to r being a random variable having a density, $p(r)$, or more

* Two references are Lindley (1971) and De Groot (1970). The best is de Finetti's two-volume work, of which Volume 1 has just appeared (1974) in an English translation.

accurately, $p(r|\theta)$, since it depends on θ. The Bayesian position is that θ is also a random variable with its density, $p(\theta)$ say. After we have conducted the n trials and seen the event occur r times, r ceases to be a random variable (the notation often reflects this in a change from R to r). However, θ, not being observed, retains its random variable status, its density changing from $p(\theta)$ to $p(\theta|r)$ in consequence of the observation. The change here is governed by Bayes' theorem: $p(\theta|r) \propto \theta^r(1-\theta)^{n-r}p(\theta)$. Bayesian analysis is concerned* with the distributions of θ and how they are changed by observations: sampling-theory statistics is concerned with the only distribution it has, $p(r|\theta)$, a distribution, which the Bayesian claims, is irrelevant after R has been observed to be r.

Example 2. There are many problems which are concerned with the means of several normal distributions: for example, the common two-way classification (rows and columns, say) using the analysis of variance and the concepts of main effects and interactions. In the Bayesian position the cell means are themselves random variables whose distributions, as in Example 1, are affected by observations. The distinction between Model I and Model II analyses therefore disappears, though the parameters in the latter model, for example, the variance component for rows, are random variables in the Bayesian treatment, though not in the orthodox one. We return to these two examples later in the talk.

Although all unobserved quantities are, in the Bayesian view, random, the concept of probability thereby implied is not based on frequency considerations. Probability is a relationship between 'you' and the external world, expressing your views about that external world. In particular, the Bernoulli 'probability', θ in Example 1, is not a probability in this sense, because it describes a property of the external world. We refer to it as the *propensity* of the event to occur. The important point here is not the names as such, but the appreciation of the difference between, on the one hand, a relationship between you and the sequence, and, on the other, a property of the sequence. The function of names is to distinguish things: the same name is given to things which are alike; different names to things which are dissimilar. A rose by any name would smell as sweet but it would be confusing if the alternative name was daffodil.

Other concepts enter into the Bayesian approach, in particular that of utility and the combination of it with probability in the notion of expected utility. The

* Though not exclusively. Sometimes it is useful to talk about the unconditional distribution of r, $p(r)$, not $p(r|\theta)$, as when we contemplate the possible results of the n trials. Such distributions are not available in sampling-theory statistics.

final maximisation of this quantity solves any decision problem. But the utility notion is itself probabilistic, so that essentially everything follows from the basic remark that unobserved quantities are random: that is, have a probability structure. All the calculations in the system are within the probability calculus (which is why Jeffreys (1967) uses probability in the title of his great book on statistics). In particular problems of point estimation disappear: the 'estimate' is the probability distribution and any single value is nothing more than a convenient partial description of this distribution.

There is a useful distinction to be made between 'inference' and 'decision'. The Bayesian view is that the only purpose of an inference is its potential use in a decision problem. To achieve this potentiality it is only necessary to provide the probability distribution conditional on the data. This provision is inference. Decision-making adds the utility ingredient, calculates an expectation, using the inferential probability, and performs the maximisation. The distinction occurs outside statistics: law and medicine are mentioned below.

2. Justification

The first complete justification for this viewpoint known to me was given by Ramsey (1964) in 1926. His work lay unappreciated for almost thirty years and modern work begins with Savage's (1954) important book. The best up-to-date treatment in a textbook is probably De Groot's (1970). An alternative approach is due to de Finetti (1964) in 1937. Ramsey's argument is essentially along the following lines. In considering the way in which people would themselves wish to act in the face of uncertainty, the statistician is led to state certain axioms that they would not wish to violate. An example of these is the one Savage so charmingly called the 'sure-thing' principle. It says that if A is preferred to B when C obtains, and also when C does not obtain, then A is preferred to B when one is uncertain about C. From these axioms it is possible to develop a mathematical system that we call Bayesian statistics. In particular, it is possible to *prove* that uncertain quantities have a probability structure; the property that we took as basic to the system. I know of no objection to these axioms that has persisted, and it is a pity that many critics of the approach do not pay more attention to them instead of misrepresenting the position and so making it look ridiculous.

We should, at this point, take note of a great advantage the Bayesian position has over all other approaches to statistics: namely, in the way just described, it is a formal system with axioms and theorems. We all know and appreciate the great impetus given to probability theory by Kolmogorov's (1950) 1933 axiomatisation of that field. A more striking example is provided by Newton's statement of the laws of mechanics. Only when a system has a formal structure

can we be quite sure what it is we are talking about, and can we teach it to all intelligent enough and willing to listen. Fisherians have condemned Bayesian statistics as a 'monolithic structure'. Would they term Newtonian mechanics monolithic? Critics often refer to Bayes as a Messiah; would they grant the same status to Newton? I find this messianic attitude particularly curious when uttered by Fisherians who appear to regard the collected works, Fisher (1950), and his last book, Fisher (1956), as the old and new testaments respectively.

An important theorem within the formal system is that which says that inferences should follow the likelihood principle. Now it so happens that almost all statistical techniques violate this principle and therefore do not fit into the system. As a result all these techniques must be capable of producing nonsense. And this indeed is so. In Lindley (1971) I have given a list of counter-examples to demonstrate how ridiculous every statistical technique can be. Thus in Example 1 above suppose it is required to test the hypothesis $\theta = 1/2$, by a standard significance test. Then a vast range of significance levels can be produced by varying the sample space, or equivalently changing the stopping rule. Careful reflection shows that this is not exactly sensible. Or consider Kendall and Stuart's (1970) optimum estimate of θ^2 in Example 1, namely $r(r-1)/n(n-1)$, when $r = 1$: to estimate a chance as zero when the event has occurred is incredible.

The above justification for Bayesian statistics is at a theoretical level, though its practical implications are immense. But an important alternative justification rests on the pragmatic fact that it works. Bayesian statistics satisfies the two basic requirements of science in resting on sound principles and working in practice. Let me demonstrate this using the two examples above.

Example 1. Consider n_1 trials with r_1 successes observed, and contemplate n_2 further trials and ask what are the chances of r_2 additional successes. First let us note that this is a practical problem. The physician who treated n_1 patients with a drug and had r_1 respond successfully, could legitimately ask what might happen if the treatment were used on n_2 further patients. Indeed Pearson (1920) went so far as to describe it as one of the fundamental problems of practical statistics. Although it rarely occurs in quite the simple form here presented, a solution to it is essential before more complicated and realistic problems are discussed. But then notice that sampling-theory statistics has no simple way of answering the question. For within that subject it is not possible to talk of $p(r_2|n_2; r_1, n_1)$: only probablities conditional on θ are admitted. The difficulty is circumvented by either making statements about θ — to which the doctor's response is that he is treating *this* patient, not a long-run frequency of patients — or, rarely, to resort to the complexities of tolerance intervals. So immediately we see that Bayesian statistics has one practical advantage over the standard approach. But let us go further and consider the Bayesian answer. For

simplicity take the case $n_2 = r_2 = 1$: the chance of success on one further trial. Under certain assumptions* the probability is $(r + a)/(n + a + b)$ — omitting the suffixes — where a and b refer to the initial (prior) views of the sequence. Compare this with r/n, the usual point estimate of θ. The most obvious difference between the two is the occurrence of a and b in the former but not the latter. But doesn't this make good, practical sense? The usual estimate says that it does not matter whether it is a sequence of patients, transistors, drawing-pins or coins, the estimate is always the same. The Bayesian argument says it is necessary to think about whether it is patients, transistors, drawing-pins or coins that are being discussed, for which it is could affect the choice of a and b. For example, with drawing-pins I would take $a = b = 2$, but with coins $a = b = 100$, say. The resulting Bayesian answers for modest values of n are very different: isn't that right? Wouldn't your reaction to drawing-pins (about whose tossing propensities you probably know very little) be different from those with coins (which are well-known to have propensities near 1/2)?

Example 2. The techniques available for studying the two-way layout are extensive and one faces an embarrassment of choices which the textbooks do not resolve. One can perform an analysis of variance with its associated significance tests. But if, for example, the main effect of rows and the interaction are significant at 1 percent, but not the column effect, how is one supposed to estimate a cell mean? What multiple comparisons are to be applied? The Bayesian approach is quite clear. first you have to think about those rows and columns: are they important factors or are they nuisance factors that good experimental design has suggested be included? What do you know about the factors — is one a control? And so on, thinking about the real problem in order to assess an initial distribution. Having done this, Bayes theorem is applied to provide answers to all questions in the form of a probability distribution. Under certain assumptions the expectation of the parameter describing the cell in the ith row and the jth column is a linear function of four quantities, the overall mean $x_{..}$, the row effect $x_{i.} - x_{..}$, the column effect $x_{.j} - x_{..}$ and the interaction $x_{ij} - x_{i.} - x_{.j} + x_{..}$, the weights depending on the appropriate variance components. The estimates avoid all multiple comparison difficulties and any ambiguities over the meaning of significance tests: see Lindley (1975).

(A further point arises here: it was not mentioned in the original lecture but occurs in Rao's paper and was prominent in the discussion. It is now well-known that the usual estimate of a multivariate normal mean is unsatisfactory and that the Stein (1956) estimate is preferable. Unfortunately this

* The basic assumption is that the trials are exchangeable. This is weaker than the assumption of a Bernoulli sequence.

estimate, and analogous estimates provided by empirical Bayes methods, are ambiguous in that they do not declare what multivariate distribution is to be used. If studying hogs in Montana, why not add data on butterflies in Brazil and increase the dimensionality? Curiously, the estimates for the hogs will change. The difficulty can be resolved by recognizing that the Bayes distribution will reflect the difference between hogs and butterflies and will only produce the Stein estimate when certain exchangeability assumptions are valid. Hogs are not exchangeable with butterflies!)

3. Directions for Statistics

As I hinted in the introduction, in my view it would have been better not to have included the above material in the talk, since it is already available in the literature, but instead to have concentrated on future directions for our subject. This was, as I understood it, the purpose of the conference, and is a topic not too well covered in the literature; a notable exception is Watts (1968). In the time remaining to me I can only provide a cursory guide into the next century.

Bayesian statistics rests on the all-embracing notion of probability as describing your belief about the state of the world. Once it is admitted that such beliefs, obeying the calculus of probabilities, exist, we have an important measurement problem: how to assess them? According to the thesis, your beliefs can be described numerically: how are these numbers to be found? Associated with this idea there is the concept of utility, describing numerically your valuation of the worth of an outcome: how are these to be evaluated? One method is to relate the beliefs to gambles, but this is, for obvious reasons, not entirely satisfactory. A modified form of this is to consider a *scoring rule.* A subject, asked to assess the probability of some event, gives the value p. If the event occurs he is awarded a prize $\phi(p)$; if not, he obtains $\phi(1-p)$. It is easy to see that only some functions ϕ will qualify, in the sense that in order to maximise his expected score the subject will declare his correct probability. The simplest qualifying function is $\phi(p) = (1-p)^2$. This has been used by de Finetti, but in meteorology is called the Brier scoring rule. Can we train people to be good probability assessors using the Brier, or a similar, rule? Clearly this must be a subject for much research if the Bayesian ideas are to be implemented.

One of the most important papers in this field is that of Savage (1971). His work is theoretical and needs to be supplemented by experimental studies. To perform these we will need the help of psychologists. At the moment many psychologists waste their time trying to find out how people make decisions in practice. It turns out that they aren't natural Bayesians: so then, the psychologists ask, what rules do people apply? Now, why do this, why not teach people

how to make decisions sensibly: that is to maximise expected utility. So let us persuade these psychologists and market researchers away from the problem of why a housewife buys this type of detergent rather than that — a problem that in any case will disappear with the capitalist system — and get them on to real problems.

The assessment of utilities presents similar difficulties but can often be solved directly in terms of gambles since utilities themselves involve gambles. For example, consider three states of health, here referred to loosely as good, bad and intermediate. Assigning utilities of 1 and 0 to the first two respectively, the utility of the intermediate state can be assessed by considering someone in that state, contemlating an operation that may restore his good health but has a chance p of reducing it to bad. What is the maximum p for which the operation will be adopted? The intermediate utility is then $(1 - p)$. In my experience such ideas are acceptable because subjects can appreciate the problem. The Bayesian rules enable several such assessments to be combined, so as to handle more complicated and realistic decisions.

In pursuing a sound path it is important to rescue those who are trying to cross the mountains by bad routes. What shall we do with sampling theory statistics, with significance tests, with confidence intervals; with all those methods that violate the likelihood principle? The answer is, let them die. Their role has been to provide valuable stepping-stones to the future and our appreciation of the originators of these ideas should not be diminished by this remark: for it is largely by the pursuit of the notions that we have reached the understanding that we have today. Each of these techniques has its Bayesian equivalent, which makes better practical sense, and I see no excuse for wasting our time on them except in a course on the history of our subject.* It is, I think, generally acknowledged that sampling-theory statistics is in trouble. Hence this conference, and hence the emergence of new ideas like data analysis. Notice that data analysis is the antithesis of Bayesian statistics, for it is an informal, unstructured field in which there are no rules. It is the negation of scientific method. It is a field in which bright ideas of a few clever men abound, but these ideas are, because of the informality of the subject, difficult, if not impossible, to convey to the average statistical practitioner. Contrast this with a formal system with theorems stated under precise conditions and its comparatively simple method of communication to all who are interested. I do not wish to be thought to be decrying informality as such: far from it. Messing about with the data, making plots of it and such aids to thought are an essential ingredient of

* At University College London we are working towards an integrated programme on Bayesian statistics, with one course on sampling-theory ideas for reasons of history and communication.

any good statistics. But let it be allied to a good formal framework and regarded as an approximation to a full Bayesian treatment. Newtonian mechanics provides a good analogue again. Many problems are impossible to solve strictly within that framework and much ingenuity is devoted to finding workable approximations that produce valuable answers. Do your data analysis, but remember, to make sense, you must never forget the rules of coherent behaviour, any more than an engineer can forget Newton's laws.

Having cleared some dead wood from the path, let us go forward in a more constructive vein. Statistics has had its greatest successes in those fields of science where the long-run frequency view of probability is appropriate — for example, in agriculture, where experiments may be repeated but nevertheless the variation is sufficienly large for naive techniques to be inappropriate. But with the widening of the notion of probablity to embrace non-repeatable situations the potential scope of statistics is enormously increased. We can now enter into fields that were previously denied to us, without any loss in the traditional ones, where propensity and exchangeability replace long-run frequencies and randomisation. The future of statistics looks very bright to me and perhaps the most important thing I have to say to you today is to ask you to recognise this enormous widening of our subject. For if we do not recognise this, others will take over. Let us not repeat the split between OR and statistics. Only statisticians know how to process evidence: only statisticians know how to make decisions. (The obvious adjective must be added in two places.)

An an illustration of this widening of the range of applications of statistics, consider the situation in law. In a court of law, one of the problems is, in probability language, to assess $p(G|E)$, the probability that the defendant is guilty, G, given the evidence, E. The judge and jury would clearly wish this assessment to be done using Bayes theorem; assuming, that is, they do not themselves wish to stand accused of violating the axioms, such as the sure-thing principle. At the moment it is unrealistic to be able to do this except in special cases. One such case is forensic medicine, where the evidence is precisely stated and certain probabilities are obtainable from scientific evaluations outside the court — such as the chance that two hairs, one from the suspect, one found at the scene of the crime, have come from the same head. Again notice, as with the Bayesian solution of Pearson's problem, that such probabilities do not arise naturally in the usual treatment of this problem.

Utility considerations also enter into legal matters. The jury, in some situations, is not called upon to pass sentence, that is the prerogative of the judge. He has a decision problem to solve and will require utility assessments, either imposed by statute, or by himself, preferably the former. One thing seems clear: fines should be in utiles. A wealthy man should pay more for a parking offence than an impecunious student. An interesting example of the

way in which general theorems could influence legal practice is to be found in the result which says that the expected value of sample information is non-negative. This goes against the concept of non-admissibility of certain types of evidence. My personal view is that the reason for some things being legally inadmissible is that their use as evidence is difficult, not that they are not evidence. But Bayes theorem could again oblige. (I similarly have little sympathy with those who argue for privacy of certain types of information, for example, salaries. The difficulty lies in how we *use* the information — now solved in principle — not with the facts per se.)

Another field where statistics could make a significant impact is that of diagnosis and management in medicine. The problem here is to calculate $p(D|S)$, the probability that the patient has the disease, D, given the symptoms, S: and then the use of this probability, combined with utility considerations, to determine the best management of the patient. Indeed, there is scarcely a field of human endeavour that cannot be assisted by some statistical considerations. The future is bright — but can we take advantage of it?

It has been mentioned above that certain ideas, like confidence intervals, should be allowed to die. In some branches of statistics the interment cannot be completed until a Bayesian form has been born. An example of such a topic is multivariate analysis. This is a most peculiar subject in some ways. The literature on it is vast and yet it contains substantial contradictions and difficulties that most practitioners in the field ignore. We have only recently discovered how to estimate the mean of a multivariate normal distribution: we still do not know how to estimate the dispersion matrix. And yet elaborate multivariate techniques, and their associated computer packages, have been developed and extensively used. The need for sound statistical analyses of many variables is an urgent practical necessity. The problems arise acutely in the medical diagnosis situation where many signs and symptoms are typically available. The extensive literature on multidimensional contingency tables scarcely comes to grip with this problem. Least squares is similarly unsound, at least in high dimensions, but the replacement there is simpler, because it is often fairly easy to impose a reasonable probability structure on the parameters to obtain reasonable posterior judgments.

The mention of multivariate ideas naturally leads us to consider the role of the computer. The broad line of the development is clear. Bayesian statistics is within the calculus of probabilities and the only calculations are those implied by this calculus. The computer is needed for the more complex probability manipulations, for evaluation of expected utilities and the subsequent maximisation. Multidimensional integration is extensively involved in the elimination of nuisance parameters. Man thinks, the computer calculates: that is the basic rule. A Bayesian data package will require thoughtful specification of the

model; thoughtful assessment of the initial distribution (and utility, if decision is involved) followed by calculation according to the laws of probability. It will not be as easy to use as today's packages because the user will have to think whether it is data on hogs or butterflies that he is analysing.

The future of statistics is bright. We can expand greatly: but where are the recruits to come from? We need to attract able young people into the field: people who have the mathematical experience, and exposure to scientific ideas, to make good statisticians. My hope is that by teaching Bayesian ideas we shall succeed in this. The formal system will make it easier to teach, and will appeal to the mathematical mind. The fact that it works will bring in the interested scientist.

I have spoken of the 21st century. I wish the change could come sooner. How about a moratorium on research for two years? In the first of these we will all read de Finetti's first volume: the next year will do for the second. It would do you, and our subject, a lot of good.

References

DE FINETTI, B. (1964) Foresight: its logical laws, its subjective sources. *Studies in Subjective Probability*, ed. Henry E. Kyburg, Jr. and Howard E. Smokler, pp. 93–158, Wiley, New York, (Translation of La prévision: ses lois logiques, ses sources subjectives, *Ann. Inst. H. Poincaré*, **7** (1937), 1–68.)

DE FINETTI, B. (1974) *Theory of Probability: a critical introductory treatment.* Volume 1 (Volume 2 to appear) Wiley, New York. (Translation of *Teoria delle probabilità, sintesi introduttiva con appendice critica* (1970) Giulio Einaudi, Torino.)

DE GROOT, M. H. (1970) *Optimal Statistical Decisions.* McGraw-Hill, New York.

FISHER, R. A. (1950) *Contributions to Mathematical Statistics.* Wiley, New York.

FISHER, R. A. (1956) *Statistical Methods and Scientific Inference.* Oliver and Boyd, Edinburgh.

JEFFREYS, H. (1967) *Theory of Probability.* 3rd edition (corrected). Clarendon Press, Oxford.

KENDALL, M. G. AND STUART, A. (1970) *The Advanced Theory of Statistics*, Volume 2. Griffin, London.

KOLMOGOROV, A. N. (1950) *Foundations of the Theory of Probability.* Chelsea, New York. (Translation of *Grundbegriffe der Wahrscheinlichkeitsrechnung* (1933), Springer, Berlin.)

LINDLEY, D. V. (1971) *Bayesian Statistics, a Review.* SIAM, Philadelphia.

LINDLEY. D. V. (1975) A Bayesian Solution solution for two-way analysis of variance. *Proc. 1972 Meeting of Statisticians, Budapest.* (To appear).

PEARSON, K. (1920) The fundamental problem of practical statistics. *Biometrika* **13**, 1–16.

RAMSEY, F. P. (1964) Truth and Probability. *Studies in Subjective Probability*, ed. Henry E. Kyburg, Jr. and Howard E. Smokler, pp. 61–92, Wiley, New York, (Reprinted from *The Foundations of Mathematics and Other Essays.* (1931), 156–198, Kegan, Paul, Trench, Trubner & Co. Ltd., London.

SAVAGE, L. J. (1954) *The Foundations of Statistics.* Wiley, New York.

SAVAGE, L. J. (1971) Elicitation of personal probabilities and expectations. *J. Amer. Statist. Assoc.* **66**, 783–801.

STEIN, C. M. (1956) Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proc. Third Berkeley Symp. Math. Statist. Prob.* **1**, 197–206. University of California Press, Berkeley.

WATTS, D. G. (1968) *Conference on the Future of Statistics.* Academic Press, New York.

Suppl. Adv. Appl. Prob. 7, 116–121 (1975)
Printed in Israel

WITHER MATHEMATICAL STATISTICS?

HERBERT ROBBINS, *Columbia University*

Despite the pun, I do regard the question as a serious one, and one of manageable size compared to the other well-known Whither problems that confront us as human beings. Let me begin by saying that of course I disagree with the position expressed by Professor Lindley, an abstract of whose paper I have had the privilege of seeing. The attitude that says 'Burn all the books, for if what they say is in the Koran they are unnecessary, and if not they are blasphemous' is a familiar one, though not often stated explicitly by practicing scientists. If indeed the future of statistics is to include an exclusively Bayesian 21st century, then the question mark can be removed from my title. However, I don't think Professor Lindley's prediction has a high probability except in his personal sense, so I think we should feel free to consider some alternatives to the Dark Age that he envisions.

This branch of mathematical science is a relatively new one. Two sources that I have found most informative concerning its 'early' history are Karl Pearson's three-volume work on the life, letters and labours of Sir Francis Galton, and the British statistical journals during the period 1920–40 that contain the famous controversies involving R.A. Fisher, J. Neyman and E.S. Pearson, and their sometimes bewildered contemporaries of lesser rank. It would be a most useful thing, especially during times when nothing really new and important seems to be going on, for students and professors to acquaint themselves with at least this much of the historical background of their subject. An intense preoccupation with the latest technical minutiae, and indifference to the social and intellectual forces of tradition and revolutionary change, combine to produce the Mandarinism that some would now say already characterises academic statistical theory and is most likely to describe its immediate future.

People are not born to be mathematical statisticians, and for this discussion we may ignore those who become so because it is a profession that is somewhat more pleasant or better suited to their talents than any other. What, then, motivates those who will become the innovators on which the future of our discipline depends?

One view regards statistics, like mathematics in general, as the handmaiden of the sciences, physical, biological and social. At a lower intellectual level,

statistical theory is asserted to be of great importance in processing evidence and making decisions in business, government, community health care and so on. There is much to be said for this utilitarian concept of statistics as a tool in the service of something or other, and cerainly as long as the ruling institutions of society find statisticians useful, they will be encouraged and rewarded according to their utility. Moreover, the statisticians of the past came into the subject from other fields — astronomy, pure mathematics, genetics, agronomy, economics etc. — and created their statistical methodology with a background of training in a specific scientific discipline and a feeling for its current needs. Hence, the question arises whether statistics should be studied and taught as an autonomous general discipline like mathematics, or only as the statistical part of one or another of the sciences.

The autonomy and unity of statistics, somewhat questionable from the utilitarian point of view, is championed most explicitly by the school of what I shall somewhat facetiously call pure statistics, in analogy with pure mathematics. Pure statistics on a large scale is a development of the last forty years, and is usually described by referring to the current issue of the *Annals of (Mathematical) Statistics*. I do not mean to imply that this journal is totally unconcerned with applications, but only that immediate practical applications are cerainly not the main interest of its authors, who are involved in erecting a new edifice of which even G.H. Hardy might have approved, complete with existence theorems, non-existence theorems, asymptotic theorems, postulate sets and all the familiar apparatus of pure mathematics, so engrossing to the initiate and impenetrable to the outsider.

I myself came to statistics thirty years ago from the field of pure mathematics, and have found the experience to be rewarding but sometimes disconcerting. Let me mention two instances. A friend who is an applied statistician and for whose work I have the greatest admiration said to me recently, 'I have no idea what you have been doing since you went into statistics, but people tell me that some of your work gets pretty close to the borderline of being potentially useful'. In the second case, a few years ago I was talking to one of the truly eminent mathematical statisticians of our time, a man considerably older than myself, and I thought to profit from the occasion by asking in what direction he thought the greatest advances in statistical theory might be made. His answer, given with some emotion, was, 'The greatest thing you or anyone else can do in statistics is to convince mathematicians that statistics is actually a part of pure mathematics'!

There is no doubt that mathematical statistics is destined to become increasingly important during the next century, if civilization itself endures. And it is worth speculating, at such a conference as this, on the form that future statistics will take, whether Bayesian, Neyman-Pearsonian, data-analytical or

whatever variety as yet undreamed of. Such speculations have had little predictive value in the past, as when all the fundamental problems of physics seemed to have been solved around 1900 and all that remained was to apply the basic rules that had served so well since Kepler, Galileo and Newton. Nevertheless, we are responsible not only for our own work but to some extent for that of our present and future students, and this responsibility demands that each of us should have some view of what is desirable and what should be avoided in planning for the future.

Mathematical statistics as a formal structure may be expected to flourish most in an intellectually free society, where all hypotheses are to be tested and all estimates are to be at least consistent, if not unbiased. The experience of theoretical physics, however, must never be forgotten. The work of Planck, Einstein, Dirac, Heisenberg, Fermi and others during the greatest outburst of scientific creativity of modern times produced as a direct consequence nuclear weapons that are ready and able to kill us all when the signal is given by the small group of persons who control them. And even if the threat of immediate annihilation were to be removed, the future of this technological society seems desperate indeed, for reasons with which you are all familiar to the point of boredom. Thus, the future of statistics as the handmaiden of the sciences may seem destined to be a degraded one, while as an autonomous discipline it is largely irrelevant to the problem of the survival of our species.

I believe, however, in the best Victorian tradition, that such depressing thoughts should be kept firmly in check. I cannot prove that this is desirable for everyone, but it accords with my nature to do so. So for the future I recommend that we work on interesting problems, avoid dogmatism, contribute to general mathematical theory or concrete practical applications according to our abilities and interests and, most important, formulate for ourselves a canon of humanistic values that will inspire and justify our work on a higher level than that of the well-trained and useful technician. It was L.J. Henderson, I think, who said that it was not until recently that a sick person would have been well-advised to consult a doctor. I am not sure whether someone with a statistical problem would be well-advised even today to consult a statistician. So I would add to the preceding injunctions, that if you are consulted about a practical problem and aren't sure that you can supply the correct answer, at least try to follow the Hippocratic Oath and do no harm. (For example, *never* state that a chi-squared test of independence or homogeneity in a contingency table rejects, or does not reject, the null hypothesis at the .05 significance level.)

The 'interesting problems' to which I referred above change, of course, from one year to the next. At one time, when I was trying to acquire some skill in probability theory, I found it very interesting to try to relax the condition for a

certain type of convergence from that of the finiteness of the fourth moment to that of the second. At other times I have enjoyed working on more grandiose general theories like stochastic approximation, empirical Bayes and compound decision theory, optimal stopping and confidence sequences, without much regard for obtaining the best possible results. Right now I am working on a problem that any non-statistician would probably think has long been solved: how to decide which of two binomial p's is the greater.

To be specific, suppose that observations $x_1, x_2, \ldots$ and $y_1, y_2, \ldots$ are coming in from some laboratories or hospitals and represent Bernoulli trials with unknown parameters p_1 and p_2 respectively of being 1's rather than 0's. We are asked to decide whether $p_1 > p_2$ or $p_2 > p_1$ (the option of saying that we do not know and that it is not worth finding out is assumed to be excluded). A sequential procedure due to A. Wald is to choose some positive integer B and stop with $N =$ first n such that either

$$(1) \qquad (x_1 + \cdots + x_n) - (y_1 + \cdots + y_n) = B,$$

or

$$(2) \qquad (x_1 + \cdots + x_n) - (y_1 + \cdots + y_n) = -B;$$

in case (1) we assert that $p_1 > p_2$, while in case (2) that $p_2 > p_1$. For this procedure it can be shown that for $p_1 > p_2$, say,

$$(3) \qquad P_{p_1, p_2}(\text{error}) = \frac{1}{1+\lambda^B}, \qquad E_{p_1, p_2}(N) = \frac{B(\lambda^B - 1)}{(p_1 - p_2)(\lambda^B + 1)}$$

where $q_i = 1 - p_i$ and λ is the odds ratio $p_1 q_2 / p_2 q_1$. The properties (3) of this procedure make it very suitable in some practical applications.

However, we have in (1) and (2) tacitly assumed that the observations are coming in *pairwise*: $(x_1, y_1), (x_2, y_2), \cdots$. When this is not the case, so that at any given stage of experimentation we may be confronted with values $x_1, \cdots, x_m$ and $y_1, \cdots, y_n$ for which $m \neq n$, then Wald's procedure cannot be applied, except by discarding some of the information. For such cases, let us consider the following procedure: stop with $(M, N) =$ first pair (m, n) such that setting $u_m = x_1 + \cdots + x_m$, $v_n = y_1 + \cdots + y_n$, either

$$(4) \qquad \frac{2nu_m - 2mv_n}{m+n} \geq B,$$

or

$$(5) \qquad \frac{2nu_m - 2mv_n}{m+n} \leq -B;$$

in case (4) assert that $p_1 > p_2$, while in case (5) that $p_2 > p_1$.

If the observations are in fact coming in pairwise, then this procedure reduces to Wald's. In the non-pairwise case the following properties seem to hold (again for $p_1 > p_2$, say):

$$(6) \qquad P_{p_1,p_2}(\text{error}) \leq \frac{1}{1+\lambda^B}, \qquad E_{p_1,p_2}\left(\frac{2MN}{M+N}\right) \geq \frac{B(\lambda^B - 1)}{(p_1 - p_2)(\lambda^B + 1)},$$

with approximate equalities when p_1 and p_2 are close together.

In addition to being applicable to cases where the rates at which the observations are coming in are not under our control, this non-pairwise procedure becomes especially interesting when the allocation of the observations is under our control, and we wish to minimize not the total sample size $M + N$ but, e.g., the sample size from the population with the smaller p value, or the total number of 0's observed, where 1 denotes 'cure' and 0 'non-cure' for some human disability. At this point mathematics, probability theory, computing science and statistical inference combine to make contact with the fate of actual human beings, both the experimental subjects and those to whom the asserted conclusion will be applied, and it seems fitting to conclude my remarks here and open this last session to general discussion. I regret very much that I have been unable to be with you in person, and I hope that we may all join in saying, 'Mathematical Statistics — may it never wither!'

Addendum: A heuristic derivation of the relations (6)

For any $m, n \geq 1$ the joint likelihood function of $x_1, \cdots, x_m$ and $y_1, \cdots, y_n$ is $f(p_1, p_2) = p_1^{u_m} q_1^{n-u_m} p_2^{v_n} q_2^{n-v_n}$. Consider, now, the likelihood ratio

$$\frac{f(p_2, p_1)}{f(p_1, p_2)} = \lambda^{v_n - u_m} \cdot \left(\frac{q_2}{q_1}\right)^{m-n},$$

and suppose that

$$p_1 = p + \varepsilon, \quad p_2 = p - \varepsilon,$$

where $\varepsilon > 0$ and ε/p and ε/q are both small. Then

$$\lambda = \frac{p_1 q_2}{p_2 q_1} = \frac{(1+\varepsilon/p)(1+\varepsilon/q)}{(1-\varepsilon/p)(1-\varepsilon/q)} \cong e^{2\varepsilon/pq}, \quad \frac{q_2}{q_1} = \frac{1+\varepsilon/q}{1-\varepsilon/q} \cong e^{2\varepsilon/q},$$

so that

$$\frac{f(p_2, p_1)}{f(p_1, p_2)} \cong e^{2\varepsilon/pq}\,[v_n - u_m + p(m-n)] \cong \lambda^{v_n - u_m + p(m-n)}.$$

If $m + n$ is large then with high probability

$$\frac{u_m + v_n}{m+n} \cong p + \frac{(m-n)\varepsilon}{m+n},$$

so the exponent of λ becomes

$$v_n - u_m + (m-n)\left[\frac{u_m + v_n - (m-n)\varepsilon}{m+n}\right] = -z_{mn} - \frac{(m-n)^2\varepsilon}{m+n} = -c_{m,n},$$

where we have put

$$z_{m,n} = \frac{2(nu_m - mv_n)}{m+n}, \quad c_{m,n} = z_{m,n} + \frac{(m-n)^2\varepsilon}{m+n}.$$

Hence

$$\frac{f(p_2,p_1)}{f(p_1,p_2)} \cong \lambda^{-c_{m,n}}.$$

To the extent that this approximation holds it follows that

$$P_{p_1,p_2}(\text{assert } p_2 > p_1) = \sum_{m,n=1}^{\infty} \sum \lambda^{c_{m,n}} f(p_2,p_1)$$

where the second sum is over those outcomes for which $(M,N) = (m,n)$ and $z_{m,n} \geq B$. Since $c_{m,n} \geq z_{m,n}$, it follows that

$$P_{p_1,p_2}(\text{assert } p_1 > p_2) \geq \lambda^B P_{p_2,p_1}(\text{assert } p_1 > p_2) = \lambda^B[1 - P_{p_1,p_2}(\text{assert } p_1 > p_2)],$$

(assuming that the sampling rule is symmetric in the x's and y's) so that

$$P_{p_1,p_2}(\text{assert } p_1 > p_2) \geq \frac{\lambda^B}{1+\lambda^B}$$

and hence the first part of (6) holds.

A heuristic derivation of the second part of (6) is based on the fact that

$$E_{p_1,p_2}(z_{M,N}) \geq \frac{B\lambda^B - B}{1+\lambda^B},$$

while presumably

$$E_{p_1,p_2}(z_{M,N}) \cong (p_1 - p_2)E_{p_1,p_2}\left(\frac{2MN}{M+N}\right).$$

When $m = n$ the preceding argument simplifies to give a proof of the exact relations (3). A detailed treatment of the simpler case of two normal populations is given in *J. Amer. Statist. Assoc.* (1974) **69**, 132–139.