Springer Series in Statistics

Perspectives in Statistics

Samuel Kotz
Norman L. Johnson
Editors

Breakthroughs in Statistics Volume I

Foundations and Basic Theory

Springer-Verlag

Samuel Kotz Norman L. Johnson
Editors

Breakthroughs in Statistics Volume I

Foundations and Basic Theory

Springer Science+Business Media, LLC

Samuel Kotz
College of Business
and Management
University of Maryland
at College Park
College Park, MD 20742
USA

Norman L. Johnson
Department of Statistics
Phillips Hall
The University of North Carolina
at Chapel Hill
Chapel Hill, NC 27599
USA

Library of Congress Cataloging-in-Publication Data
Breakthroughs in statistics / Samuel Kotz, Norman L. Johnson, editors.
p. cm. — (Springer series in statistics. Perspectives in statistics)
Includes bibliographical references and index.
Contents: v. 1. Foundations and basic theory — v. 2. Methodology and distribution.
ISBN 978-0-387-94037-3 ISBN 978-1-4612-0919-5 (eBook)
DOI 10.1007/978-1-4612-0919-5
1. Mathematical statistics. I. Kotz, Samuel. II. Johnson, Norman Lloyd. III. Series.
QA276.B68465 1993
519.5—dc20 93-3854

Printed on acid-free paper.

Typeset by Asco Trade Typesetting Ltd., Hong Kong.

9 8 7 6 5 4 3 2

Contents

Preface vii

Contributors xvii

Sources and Acknowledgments xix

Editorial Note xxiii

- Cox, G.M. (1957) "Statistical Frontiers" xxv

  Comments xxxvi

- *Introduction by S. Geisser* 1

  Fisher, R.A. (1922) On the Mathematical Foundations of Theoretical Statistics 11

- *Introduction by T.W. Anderson* 45

  Hotelling, H. (1931) The Generalization of Student's Ratio . . 54

- *Introduction by E.L. Lehmann* 67

  Neyman, J. and Pearson, E.S. (1933) On the Problem of the Most Efficient Tests of Statistical Hypotheses 73

- *Introduction by D.A.S. Fraser* 109

  Bartlett, M.S. (1937) Properties of Sufficiency and Statistical Tests 113

- *Introduction by R.E. Barlow* 127

  de Finetti, B. (1937) Foresight: It's Logical Laws, Its Subjective Sources 134

- *Introduction by M.R. Leadbetter* 175

  Cramér, H. (1942) On Harmonic Analysis in Certain Functional Spaces 179

- *Introduction by R.L. Smith* . . . 185
  Gnedenko, B.V. (1943) On the Limiting Distribution of the Maximum Term in a Random Series . . . 195
- *Introduction by P.K. Pathak* . . . 227
  Rao, C.R. (1945) Information and the Accuracy Attainable in the Estimation of Statistical Parameters . . . 235
- *Introduction by B.K. Ghosh* . . . 249
  Wald, A. (1945) Sequential Tests of Statistical Hypotheses . . 256
- *Introduction by P.K. Sen* . . . 299
  Hoeffding, W. (1948) A Class of Statistics with Asymptotically Normal Distribution . . . 308
- *Introduction by L. Weiss* . . . 335
  Wald, A. (1949) Statistical Decision Functions . . . 342
- *Introduction by D.V. Lindley* . . . 359
  Good, I.J. (1952) Rational Decisions . . . 365
- *Introduction by I.J. Good* . . . 379
  Robbins, H.E. (1955) An Empirical Bayes Approach to Statistics . . . 388
- *Introduction by H.P. Wynn* . . . 395
  Kiefer, J.C. (1959) Optimum Experimental Designs . . . 400
- *Introduction by B. Efron* . . . 437
  James, W. and Stein, C.M. (1961) Estimation with Quadratic Loss . . . 443
- *Introduction by J.F. Bjørnstad* . . . 461
  Birnbaum, A. (1962) On the Foundations of Statistical Inference . . . 478
- *Introduction by W.H. DuMouchel* . . . 519
  Edwards, W., Lindman, H., and Savage, L.J. (1963) Bayesian Statistical Inference for Psychological Research . . . 531
- *Introduction by N. Reid* . . . 579
  Fraser, D.A.S. (1966) Structural Probability and a Generalization . . . 587
- *Introduction by J. de Leeuw* . . . 599
  Akaike, H. (1973) Information Theory and an Extension of the Maximum Likelihood Principle . . . 610

Index . . . 625

Contents

Volume II: Methodology and Distribution

- *Introduction by G.A. Barnard*
  Pearson, K. (1900) On the Criterion that a Given System of Derivations from the Probable in the Case of a Correlated System of Variables is Such that it Can Reasonably be Supposed to Have Arisen from Random Sampling
- *Introduction by E.L. Lehmann*
  "Student" (1908) The Probable Error of a Mean
- *Introduction by S.C. Pearce*
  Fisher, R.A. (1925) Statistical Methods for Research Workers
- *Introduction by T.P. Speed*
  Fisher, R.A. (1926) The Arrangement of Field Experiments
- *Introduction by M.A. Stephens*
  Kolmogorov, A.N. (1933) On the Empirical Determination of a Distribution
- *Introduction by T. Dalenius*
  Neyman, J. (1934) On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling
- *Introduction by T.W. Anderson*
  Hotelling, H. (1936) Relations Between Two Sets of Variates
- *Introduction by G.E. Noether*
  Wilcoxon, F. (1945) Individual Comparisons by Ranking Methods

- *Introduction by H.A. David*
  Mosteller, F. (1946) On Some Useful "Inefficient" Statistics
- *Introduction by M.L. King*
  Durbin, J. and Watson, G.S. (1950/1) Testing for Serial Correlation in Least Squares Regression, I/II
- *Introduction by N.R. Draper*
  Box, G.E.P. and Wilson, K.B. (1951) On the Experimental Attainment of Optimal Conditions
- *Introduction by N.E. Breslow*
  Kaplan, E.L. and Meier, P. (1958) Nonparametric Estimation from Incomplete Observations
- *Introduction by P.K. Sen*
  Chernoff, H. (1959) Sequential Design of Experiments
- *Introduction by E.J. Wegman*
  Box, G.E.P. and Jenkins, G.M. (1962) Some Statistical Aspects of Adaptive Optimization and Control
- *Introduction by L.V. Jones*
  Tukey, J.W. (1962) The Future of Data Analysis
- *Introduction by S.E. Fienberg*
  Birch, M.W. (1963) Maximum Likelihood in Three-Way Contingency Tables
- *Introduction by F.R. Hampel*
  Huber, P.J. (1964) Robust Estimation of a Location Parameter
- *Introduction by R.L. Prentice*
  Cox, D.R. (1972) Regression Models and Life Tables
- *Introduction by P. McCullagh*
  Nelder, J.A. and Wedderburn, R.W.M. (1972) Generalized Linear Models
- *Introduction by R.J. Beran*
  Efron, B. (1979) Bootstrap Methods: Another Look at the Jackknife

Contributors

GEISSER, S. School of Statistics, 270 Vincent Hall, University of Minnesota, 206 Church St., S.E., Minneapolis, MN 55455 USA.

ANDERSON, T.W. Department of Statistics, Sequoia Hall, Stanford University, Stanford, CA 94305 USA.

LEHMANN, E.L. Department of Statistics, University of California, Berkeley, CA 94720 USA.

FRASER, D.A.S. Department of Statistics, University of Toronto, Toronto, Canada M5S 1A1.

BARLOW, R.E. Department of Operations Research, 3115 Etcheverry Hall, University of California, Berkeley, CA 94720 USA.

LEADBETTER, M.R. Department of Statistics, University of North Carolina, Chapel Hill, NC 27599-3260 USA.

SMITH, R.L. Department of Statistics, University of North Carolina, Chapel Hill, NC 27599-3260 USA.

PATHAK, P.K. Department of Mathematics & Statistics, University of New Mexico, Albuquerque, NM 87131 USA.

GHOSH, B.K. Department of Mathematics, Christmas-Saucom Hall 14, Lehigh University, Bethlehem, PA 18015 USA.

SEN, P.K. Department of Statistics, University of North Carolina, Chapel Hill, NC 27599-3260 USA.

WEISS, L. College of Engineering, School of Operations Research and Industrial Engineering, Upson Hall, Cornell University, Ithaca, NY 14853-7501 USA.

LINDLEY, D.V. 2 Periton Lane, Minehead, Somerset TA 24 8AQ, United Kingdom.

GOOD, I.J. Department of Statistics, Virginia Polytechnic and State Univ., Blacksburg, VA 24061-0439 USA.

WYNN, H.P. School of Mathematics, Actuarial Science and Statistics, City University, Northampton Square, London EC1V OHB, United Kingdom.

EFRON, B. Department of Statistics, Sequoia Hall, Stanford University, Stanford, CA 94305 USA.

BJØRNSTAD, J.F. Department of Mathematics and Statistics, College of Arts and Sciences, University of Trondheim, N-7055 Dragvoll, Norway.

du MOUCHEL, W.H. BBN Software Corporation, 10 Fawcett Street, Cambridge, MA 02138 USA.

REID, N. Department of Statistics, University of Toronto, Toronto Canada M5S 1A1.

de LEEUW, J. Social Statistics Program, Depts. of Psychology and Mathematics, University of California, 405 Hilgard Avenue, Los Angeles, CA 90024-1484 USA.

Sources and Acknowledgments

Cox Gertrude M. (1957) Statistical frontiers. *J. Amer. Statist. Assoc.*, **52**, 1–10. Reproduced by the kind permission of the American Statistical Society.

Fisher R.A. (1922) On the mathematical foundations of theoretical statistics. *Philos. Trans. R. Soc. London, Ser. A.*, **222A**, 309–368. Reproduced by the kind permission of the Royal Society.

Hotelling H. (1931) The generalization of Student's ratio. *Ann. Math. Statist.*, **2**, 368–378. Reproduced by the kind permission of the Institute of Mathematical Statistics.

Neyman J. and Pearson E.S. (1933) On the problem of the most efficient test of statistical hypotheses. *Philos. Trans. R. Soc. London, Ser. A.*, **231**, 289–337. Reproduced by the kind permission of the Royal Society.

Bartlett M.S. (1937) Properties of sufficiency and statistical tests. *Proc. R. Soc. London, Ser. A*, **168**, 268–282. Reproduced by the kind permission of the Royal Society.

de Finetti B. (1937) Foresight: Its logical laws, its subjective sources. *Ann. Inst. H. Poincaré*, **7**, 1–68, (english translation by Henry E. Kyberg, Jr.). Reproduced by the kind permission of Robert E. Krieger Publishing Company.

Cramér H. (1942) On harmonic analysis in certain functional spaces. *Ark. Mat. Astr. Fys.*, **28B**(12). Reproduced by the kind permission of the Royal Swedish Academy of Sciences.

Gnedenko B.V. (1943) On the limiting distribution of the maximum term in a random series. *Ann. Math.*, **44**, 423–453 (English translation).

Rao C.R. (1945) Information and the accuracy attainable in the estimation of statistical parameters. *Bull. Calcutta Math. Soc.*, **37**, 81–91. Reproduced by the kind permission of the Calcutta Mathematical Society.

Wald A. (1945) Sequential tests of statistical hypotheses. *Ann. Math. Statist.*, **16**, 117–196. Reproduced by the kind permission of the Institute of Mathematical Statistics.

Hoeffding W. (1948) A class of statistics with asymptotically normal distributions. *Ann. Math. Statist.*, **19**, 293–325. Reproduced by the kind permission of the Institute of Mathematical Statistics.

Wald A. (1949) Statistical decision functions. *Ann. Math. Statist.*, **29**, 165–205. Reproduced by the kind permission of the Institute of Mathematical Statistics.

Good I.J. (1952) Rational decisions. *J. R. Statist. Soc. Ser. B.*, **14**, 107–114. Reproduced by the kind permission of the Royal Statistical Society and Basil Blackwell, Publishers.

Robbins H.E. (1955) An empirical Bayes approach to statistics. *Proc. 3rd Berkeley Symp. Math. Statist. Prob.* **1**, 157–163. Reproduced by the kind permission of the Regents of the University of California and the University of California Press.

Kiefer J.C. (1959) Optimum experimental designs. *J. R. Statist. Soc. Ser. B.*, **21**, 272–304. Reproduced by the kind permission of the Royal Statistical Society and Basil Blackwell, Publishers.

James W. and Stein C.M. (1961) Estimation with quadratic loss. *Proc. 4th Berkeley Symp. Math. Statist. Prob.*, **1**, 311–319. Reproduced by the kind permission of the Regents of the University of California and the University of California Press.

Birnbaum A.W. (1962) On the foundations of statistical inference. *J. Amer. Statist. Assoc.*, **57**, 269–306. Reproduced by the kind permission of the American Statistical Association.

Edwards W., Lindman H., and Savage L.J. (1963) Bayesian statistical inference for psychological research. *Psychol. Rev.*, **70**, 193–242. Reproduced by the kind permission of the American Psychological Association.

Fraser D.A.S. (1966) Structural probability and a generalization. *Biometrika*, **53**, 1–9. Reproduced by the kind permission of the Biometrika Trustees.

Akaike H. (1973) Information theory and an extension of the maximum likelihood principle. *2nd Intern. Symp. Inf. Theory*, (B.N. Petrov and F. Csàki, eds.) Akad. Kiàdo, Budapest, 267–281.

Editorial Note

To illustrate the enormous strides in Statistical Sciences during the last three and a half decades and to exhibit the direction of these developments the Editors decided to reproduce the well-known American Statistical Association Presidential Address by Gertrude Cox *Statistical Frontiers* delivered on September 9, 1956, at the 116th Annual Meeting of the ASA in Detroit and printed in the March 1957 issue of the *Journal of American Statistical Association.*

Gertrude Cox (1900–1978), an illustrious representative of the classical school of modern statistics in the K. Pearson - R.A. Fisher tradition, delivered her address on the state of statistical sciences just before the major impact and eventual dominating position of computer technology in statistical methodology and practice, and the expansion of appreciation of statistical methodology to various new branches of medical engineering and behavioral sciences. Although the comparison between the state of statistical sciences in the fall of 1956 and in the fall of 1990 (when these lines are being written) is self-evident for readers of these volumes, we thought that it would be expedient to solicit comments on this subject. Each person was requested to provide a 200–400 word commentary on *Statistical Universe in* 1990 *versus* 1956. Respondents' comments are printed, with minor editorial alterations, following G. Cox's address.

Statistical Frontiers*

Gertrude M. Cox
Institute of Statistics,
University of North Carolina

1. Introduction

I am going to ask you to look forward as we try to discern, as best we can, what the future holds for statisticians. If ten years ago we had predicted some of the things we are doing today, we would have been ridiculed. Now, my concern is that we may become too conservative in our thinking.

Civilization is not threatened by atomic or hydrogen bombs; it is threatened by ourselves. We are surrounded with ever widening horizons of thought, which demand that we find better ways of analytical thinking. We must recognize that the observer is part of what he observes and that the thinker is part of what he thinks. We cannot passively observe the statistical universe as outsiders, for we are all in it.

The statistical horizon looks bright. Exciting experiences lie ahead for those young statisticians whose minds are equipped with knowledge and who have the capacity to think constructively, imaginatively, and accurately.

Will you, with me, look upon the statistical universe as containing three major continents: (1) descriptive methods, (2) design of experiments and investigations, and (3) analysis and theory. As we tour these continents, we shall visit briefly a few selected well developed countries, where statisticians have spent considerable time. As tourists, we shall have to stop sometimes to comment on the scenery, culture, politics, or the difficulties encountered in securing a visa. With our scientific backgrounds, we should spend most of our time, seeking out the new, the underdeveloped, the unexplored or even the dangerous areas.

* Presidential address, at the 116th Annual Meeting of the American Statistical Association, Detroit, Michigan, September 9, 1956.

It is one of the challenges of the statistical universe that, as new regions are discovered and developed, the horizon moves further away. We cannot visit all the frontiers for they are too numerous. I believe that we should try to visualize the challenges of the future by looking at typical types of unsolved problems. I hope you will find the trip so interesting that you will revisit some of these statistical frontiers not as tourists but as explorers.

You know how many folders and guide books one can accumulate while traveling. I am not going even to list the ones used. This will leave you guessing whether I am quoting or using original ideas. Many people in this audience will recognize their statements used with no indication that they are quotations.

2. Descriptive Methods Continent

In planning our tour, I decided to take you first to the descriptive methods continent, for it is the oldest and has the densest settlement. The lay conception of descriptive methods ordinarily includes these countries: (1) collection of data; (2) summarization of data including such states as tabulation, measures of central tendency and dispersion, index numbers and the description of time series; and (3) the presentation of data in textual, tabular, and graphic form.

The collection of data is the largest country on this descriptive methods continent. This country is of common interest and concern to the whole statistical universe and is by far the oldest country. Official statistics existed in the classic and medieval world. In fact, in 1500 B.C. in Judea the population is given as 100,000 souls. Practical necessity forced the earliest rulers to have some count of the number of people in their kingdom.

The collection of official statistics has increased in importance over the years as evidenced by the large units of our Federal Government such as Census, Agriculture, and Labor, organized to collect all kinds of useful data.

Before going into the frontier area to collect more data, one should check carefully the sources of data in the settled areas to be sure that he is not about to perform needless duplication. The decision will have to be made whether to take a census, or to take a sample from the population. Here, as we stand on a ridge, we look over into the sampling country which we shall visit later.

Between the collection and the summarization of data countries, there is this border area, where the police (editors) check our schedule to make sure the blanks are filled and that no absurd or highly improbable entries have been made. As we continue our tour, our papers and passports will be checked frequently.

Our first stop in the summarization country is at the state called tabulation. Here the data on all items from the individual schedules are tabulated and cross-tabulated. A visit here is prerequisite to all further study of the data by statistical methods.

I shall have to ask you to pass up a visit to the well-known array, ranking, and frequency tables states. There still exists disputed area around the frequency table, such as the choice of the beginning and extent of class intervals. These historic frontiers and the political devices such as ratios, proportions and percentages are visited by many tourists.

Let us proceed to two other states, where the calculations of measures of central tendency and dispersion are made. The central tendency state has several clans. In one, the arithmetic mean is the rule. A second group has a median rule, and a third group prefers the mode rule.

Near the mainland, there are islands between it and the analysis and theory continent. Even on these islands mathematical definitions are required for the rules used for measuring central tendencies such as the geometric and harmonic means.

As we go on into the dispersion state you will note that the topography is becoming less familiar. Yet variation of individuals in a measurable characteristic is a basic condition for statistical analysis and theory. If uniformity prevailed, there would be no need for statistical methods, though descriptive methods might be desired.

This variation state also has several clans. One advocates the range as the simplest measure to describe the dispersion of a distribution. Another prefers the use of the mean deviation, while the most densely populated clan advocates the standard deviation. Nearby is a frontier area where dwell less familiar and relatively uninteresting groups such as the quartile deviation and the 10–90 percentile range.

In this descriptive methods continent, placed in the summarization of data country are other states settled by special purpose groups. Let us now visit two, the index number and the description of time series states, to look at some of their unsettled and disputed frontier problems.

The index number state, consisting of one or a set of measures for one or a group of units, evaluates indirectly the incidence of a characteristic that is not directly measurable. We do not have time to visit the single factor index area, but proceed directly to the wide open frontiers of multi-factor indexes. For example, the price and level-of-living indexes are well known and of vital interest. On this frontier: (1) Which multi-factor index formula is the best? (2) What items should be included? (3) What is the proper weighting of items? (4) Is the fixed base or chain method best? (5) How frequently should the base be changed? (6) When and how can you remove obsolete commodities and add new ones into the index? and (7) If the index number has no counterpart in reality, should it be discarded? To settle these frontiers, developments are needed on the borders with the theory continent.

In the description of time series state, we find measures recorded on some characteristic of a unit (or a group of units) for different periods or points of time. There are several method groups governing this state such as inspection, semi-averages, moving averages and least squares. Of course, there are disputes about which method is best. One of the frontier problems is how to handle nonlinear trends. One group of statisticians exploring in this state deals with time series accounting for secular trend, cyclical, periodic, and irregular movements.

Note that most of the folks in this area are economists. The public health and industrial scientists are beginning to explore here. They have such problems as fatigue testing, incubation period of a disease, and the life time of radioactive substances.

This is rather an exhausting tour, so much to be seen in so short a time. However, before you leave the descriptive methods continent, I want you to visit the presentation of results country. The availability and usefulness of whatever contribution to scientific knowledge the project has yielded are dependent upon the successful performance in this country.

As we enter the presentation of results country, you will be asked to swear allegiance to logical organization, preciseness, and ease of comprehension. In this country, certain conventions in structure and style of the form of presentation have developed and are generally accepted.

The methods of presentation of results divide into several states: textual, tabular, and graphic. The textual state gives only statements of findings and interpretation of results. The tabular state has two types of tables, the general and the special purpose tables, according to their functions. In the graphic state, presentation of quantitative data is represented by geometric designs. It is obvious that the tourist naive in mathematics will enjoy this state. Some of the named community settlements are: the bar diagram, area diagram, coordinate chart, statistical map, and pictorial statistics.

3. Design of Experiments and Investigations Continent

Later in discovery and development was the analytical statistics hemisphere where the tools and techniques for research workers are provided and used. The northern continent, called Design of Experiments and Investigations, is divided into two major sections, the design of experiments and the design of investigations.

My own random walks have taken me into the design of experiment section of this continent more frequently and extensively than into any other area we shall visit.

This section is divided into four major countries: (1) completely randomized, (2) randomized block, (3) latin square, and (4) incomplete block designs. The first three countries are the oldest and are well developed. However, in the latin square country, let us visit a newly explored state, where the latin square is adjusted so as to measure residual effects which may be present when the treatments are applied in sequence.

We might inquire about the uprisings in the latin square country when nonrandom treatments are assigned to the rows and columns. This takes you over into the incomplete block design country. It is hoped that this area will be placed in the incomplete block design country without further trouble.

The selection of the treatment combinations to go into these countries takes us into another dimension of this statistical universe. We have single factor and multiple factor treatment combinations. Small factorial groups fit nicely into our design countries. If several factors are involved, we may need to introduce confounding. This requires settlement in the incomplete block design country, where there are more blocks than replications. Some confounded areas are settled, such as those where confounding on a main effect, the split-plot design country. Here you find political parties with platforms ranging from randomization to stripping the second factor. This latter complicates its trade relations with the analysis countries.

Let us continue in the incomplete block design country and cross the state where confounding on high-order interactions is practiced. Right near, and often overlapping, is a new state using confounding on degree effects. These two states are being settled, with good roads already constructed, but the border has not been defined or peacefully occupied.

A rather new and progressive group of settlers are the fractional replication folks. Their chief platform is that five or more factors can be included simultaneously in an experiment of a practicable size so that the investigator can discover quickly which factors have an important effect on their product. In this area the hazard of misinterpretation is especially dangerous when one is not sure of the aliases. The penalties may be trivial. However, it seems wise not to join this group unless you know enough about the nature of the factor interactions.

The balanced and partially balanced incomplete block states are being settled very rapidly. So far as experimental operations are concerned, the incomplete block design country is no more difficult to settle than the complete block design country. It will take some extra planning and analysis to live in the incomplete block country and you will have to have adjusted means. The weights to use to adjust the means are still in a frontier status.

There are numerous frontier areas in this incomplete block country where roads and communications have been established. There are 376 partially balanced incomplete block design lots with $k > 2$ and 92 lots with $k = 2$ from which to choose. These lots have two associate classes.

We should look at some of the newer settlements as (1) the chain block and the generalized chain block design states; (2) the doubly-balanced incomplete block design state where account can be taken of the correlation between experimental units; and (3) the paired comparison design areas for testing concordance between judges, together with the appropriate agreements with the analysis continent. Beyond the latin square country dikes have been built to provide new land. There are latin squares with a row and column added or omitted, or with a column added and a row omitted. Further work covering more general situations will give this design continent more areas for expansion.

Let us go now to another large new country which, after negotiations, has been established by taking sections of the design and analysis continents. The process has raised some political issues and questions of international control. The development came about because, in the design continent, there is a two-party system with data measured (1) on a continuous scale (quantitative variable) or (2) on a discontinuous scale (qualitative variable). These party members have settled side by side in the design continent for single-factor groups.

If we have factorial groups, we have to consider both whether the measures are continuous or discontinuous and whether the factors are independent or not. To handle these problems, some of the continuous scale statisticians have established a response surface country. To prepare for the peaceful settlement of this response surface country a portion of the regression analysis state has been transferred. Whether this separation of portions of countries to make up a new country will hold, only time will tell.

Here in this rather new response surface country, observe that major interest lies in quantitative variables, measured on a continuous scale. In this situation, it is often natural to think of response as related to the levels of the factors by some mathematical function. The new methods are applicable when the function can be approximated, within the limits of the experimental region, by a polynomial.

In this tropical and exciting response surface country, the central composite and non-central composite states have been settled for some time. Some of the other borders are not firmly fixed, as would be expected in a new country. New states identified as first, second, third, and higher-order designs are seeking admittance to this country. They overlap with some of the older countries. We can stand over here on this mountain top and see many frontiers as the very special central composite rotatable design area, which has been named and partially settled with some roads constructed. Over there is the evaluation frontier where the relative efficiency of these designs and methods needs to be determined.

Progress has been made on strategies to be used for determining the optimum combination of factor levels. In addition to locating the maximum of y, it is often desirable to know something about how y varies when the factor levels are changed from their optimum values. The efficient location of an optimum combination of factor levels often requires a planned sequential series of experiments.

Most experimentation is sequential, since the treatments are applied to the experimental units in some definite time sequence. To explore in this area, the process of measurement must be rapid so that the response on any unit is known before the experimenter treats the next unit. A method of sequential analysis gives rules that determine, after any number of observations, whether to stop or continue the experiment.

The full sequential approach is often not practical, thus the two or multiple stage sequential plan with groups of units handled at one time takes us into the frontiers of this region. So far, the matter of testing hypotheses has been given major attention, but now sequential methods hold promise of increasing the efficiency of both testing and estimation procedures.

Are you ready now to visit the investigations (more popularly known as sampling) section of this design continent? Since this section borders on the descriptive methods continent, both continents find that it is essential to maintain trade relationships.

In all fields of experimentation and in most collections of descriptive data only a sample from the population can be considered. How to do this efficiently presents an extensive horizon.

I hope you did not forget to get a visa permit to travel into the sample design territory. We shall quickly cross the settled simple random sampling country. Here is the method of sampling in which the members of the sample are drawn independently with equal probabilities. This is a satisfactory place to settle if the population is not highly variable. On the frontier between this country and the other countries of this area, there are two problems: (1) How could the present sampling procedures be improved if the observations followed a standard distribution form? (2) What are the effects of nonrandomness? The inhabitants of these frontiers invade the settled areas frequently, and frontier battles result.

Next, we must cross the systematic sampling country. It is very difficult to secure permission from a statistician to enter this country. However, it is densely settled mostly by older people who have lived here all their lives. We frequently hear about uprisings and renewed efforts of this group to acquire all the advantages of the simple random sampling country.

It appears that settlement in the systematic sampling country can safely be recommended if one of the following conditions exists, (1) the order of the population is essentially random, or (2) several strata are to be used, with an independent systematic sample drawn from each stratum. There may be populations for which systematic sampling gives extremely precise estimates of the mean but never gives reliable estimates of the variance of the mean.

Perhaps the most popular section of the sampling area is the stratified random sampling country. The population is divided into parts called strata, then a sample is drawn independently in each part. One popular political party selects the number of units per stratum by optimum allocation. The second party advocates selection of a proportionate number of units per stratum. Some recently explored frontier areas are: (1) the determination of optimum allocation in multivariate studies, (2) the improvement of criteria for the construction of strata, and (3) the selection of the optimum number of strata.

If you are interested in large sample surveys, you will want to visit the multi-stage sampling country. Here the first stage units may be selected with probability proportional to size, the second stage units with equal probability. An adjacent area has been explored where first stage units are selected with arbitrary probability.

In newer areas of the multi-stage sampling country more than one first stage unit per stratum is drawn in order to permit internal assessment of the sampling errors of estimates. Even here many of these large surveys have been relegated to the archives without securing the sampling errors of estimates. This is done perhaps because of the complexity of the estimating formulas. Electronic computing machines are helping to settle this difficulty. In fact, the machines may open up even wider frontiers for settlement in the sample design countries.

In all the sampling territory, there are many internal political and economic frontiers to be cleared. These sampling countries now have fair control over sampling errors but relatively little over non-sampling errors. They realize the need to find an economic balance between investment on sample and investment on measurement technique. To these developing frontiers, we can add others such as: (1) What are the relative efficiencies of the various sampling plans? (2) What is the effect of nonresponse? and (3) What is an efficient method to sample for scarce items? Efforts are being made to clear out the underbrush and to settle some of this frontier area around the sampling territory.

4. Statistical Inference; Analysis and Theory Continent

In the analytical statistics hemisphere, we have visited the northern design of experiments and investigations continent. Let us start our tour of the southern statistical inference or the analysis and theory continent. The broad problem of statistical inference is to provide measures of the uncertainty of conclusions drawn from experimental data. All this territory, in the statistical universe, has been discovered and settled by a process of generalizing from particular results.

Let us visit several analytical technique countries, keeping in mind that the level of civilization in each of these countries is determined largely by the status of its theoretical development.

First, here is the beautiful and popular *t*-test country, where testing of hypotheses and setting up of confidence intervals for univariate populations are performed. This area is a tourist photographic paradise, but we cannot tarry. I know you will return.

Hurriedly, the way some tourists travel, we shall cross another univariate country, analysis of variance. Almost all statisticians, except maybe a few theorists, have enjoyed the beautiful lakes and mountains in this country. Among the attractive features to explore are the orthogonal sets of single degrees of freedom, the separation of simple effects when interaction exists, the use of both continuous and discontinuous variables and even the fitting of regression models for the fixed continuous variable. This latter region is being urged to establish an alliance with the response surface country.

We have time to observe only a few frontier problems: (1) What is the power of analysis of variance to detect the winner? (2) How do you analyze data which involve both a quantal and a graded response? (3) How do you attach confidence limits to proportions? (4) What about nonhomogeneity of variance when making tests of significance? and (5) Should we enter these countries with nonnormal data? I may just mention a subversive area, at least it is considered so by some, that is, the region where effects suggested by the data are tested.

Are you ready now to visit the correlation country? Bivariate populations are often interesting because of the relationship between measurements. First, let us visit the well developed product moment correlation section, where the cultural level is high due to theoretical verifications. Around here are several unincorporated areas, quite heavily populated by special groups, but not too well supported by theory. You should be careful if you visit the method of rank difference, ρ (rho), the non-linear, η (eta), the biserial or the tetrachoric coefficients of correlation districts.

While we travel across to the regression country, I might mention that its constitution has several articles like the constitution of the correlation country. The two are confused by some users of statistics and even by statisticians.

We had better check to see if you have your visa before we enter the regression country. Some of the acceptable reasons for granting visas are: (1) to see if Y depends on X and if so, how much, (2) to predict Y from X, (3) to determine the shape of the regression line, (4) to find the error involved in experiments after effect of related factor is discounted or (5) to seek cause and effect.

Some near frontier areas are being settled, such as those where there are errors in both the X and the Y variables. Other frontiers include the test of the heterogeneity of two or more regressions. How do we average similar ones? What about the nonlinear regression lines?

As we leave the bivariate countries of the analysis and theory continent and enter the multivariate countries, we find that life becomes more complicated. All kinds of mechanical, electrical and electronic statistical tools have come into use. These countries have been developed from, but are not independent of, the univariate and bivariate areas by a process of successive generalizations. For example, people were taken from the t-test country and by generalization they developed the statistics T country. This T group does all the things done by the t group for any number of variates simultaneously, be they mutually correlated or independent.

In this multivariate area, new territory related to the analysis of variance has been explored and is called the multivariate analysis of variance. Here are theoretical frontiers to be explored. Some are (1) What are the values of the roots of a determinantal equation and what particular combination of them should be used for a particular purpose? (2) What are the limitations and usefulness of the multivariate analysis of variance country? and (3) What are the confidence bounds on parametric functions connected with multivariate normal populations?

The next time you come this way, I wish you would stop to explore the areas where the discriminant function and factor analysis methods are used. There may be some danger that the latter will not be able to withstand the attacks being made by those who advocate replacing factor analysis by other statistical methods. I personally believe the factor analysis area will resist its attackers and will remain in the statistical universe as a powerful country.

The simple correlation ideas were generalized into two new countries, the multiple correlation country and the less well known canonical correlation country, which has two sets of variates.

Crossing the multiple regression country, we look at the frontiers. There are situations where it is desirable to combine scored, ranked, and continuous data into a multiple regression or factor analysis. How can this be done legally? What about the normal distribution assumptions?

I cannot resist having you visit the analysis of covariance country for it accomplishes some of the same purposes as do the design countries. Covariance helps to increase accuracy of estimates of means and variances. However, dangerous mountains exist in this country. The explorers may need to develop added theory to enable the applied statistician to reach the top of such cliffs as the one where the X variable is affected by the treatments. If the treatments do affect X, a covariance analysis may add information about the way in which the treatments produce their effects. The interpretation of the results when covariance is used requires care, since an extrapolation danger may be involved. Now that I have acknowledged that we are in a dangerous area, I might state that the dangers of extrapolation exist in all regression and related areas, and especially back in the response surface country.

We are ready to enter the variance component country, where separate sources of variability are identified. Estimates of these variance components are desired. These estimates are used to plan future experiments, to make tests of significance, and to set confidence limits.

This country is relatively new, so that adequate statistical theory has not been developed, thus leaving rugged frontiers: (1) The assumption of additivity needs to be explored in detail, (2) A clear statement is needed of how to decide whether the interaction in a two-way classification is nonrandom or random, (3) More exact methods of assigning confidence limits for the variance components need to be developed, (4) How does one handle the mixed model? (5) How can one detect correlated errors? (6) What can be done to simplify the analysis of data with unequal variances? (7) What are the effects of various types of nonnormality on the consistency and efficiency of estimates? and (8) Some study needs to be made of the proper allocation of samples in a nested sampling problem when resources are limited and good estimates of all components are desired.

Another section of the variance component country is called components of error. The problem of choosing the correct error term in the analysis of two or more factors depends upon whether the factors are random or nonrandom or upon the question you ask. Do you want the mean difference between treatments averaged over these particular areas with narrow confidence limits, or do you want mean differences averaged over a population of areas of which these areas are a sample with broad confidence limits?

So far, we have visited almost exclusively the parametric inference countries. Let us take a glimpse at the frontier in the nonparametric inference territory. When the experimenter does not know the form of his population distribution, or knows that it is not normal, then he may either transform his data or use methods of analysis called distribution free or nonparametric methods. This territory is being settled. The area dealing with the efficiency of certain tests for two by two tables has been partially settled and some general theorems on the asymptotic efficiency of tests have been proved.

Some of the frontiers are: (1) What is the general theory of power functions for distribution free tests? (2) What is the efficiency of nonparametric tests? (3) Can sequential methods be applied to nonparametric problems, and (4) How can two nonnormal populations be compared?

There are three more general frontiers I wish to mention. (1) How far are we justified in using statistical methods based on probability theory for the analysis of nonexperimental data? Much of the data used in the descriptive methods continent are observational or nonexperimental records. (2) What are the effects of nonnormality, heterogeneity, nonrandomness and nonindependence of observations to which standard statistical methods are applied? And (3) How can we deal with truncated populations in relationship problems?

As we complete our tour of the three continents, I wish to emphasize the fact that there are many important problems of design and statistical inference which remain unexplored.

5. Training Frontier

Our travels took us to only a part of the statistical universe, but we managed to observe many frontier areas. I hope one thing impressed you: that is, the extent of the need for statisticians to explore these areas. In recent years, there have been advances in statistical theory and technology, but the prompt application of these to our biological, social, physical, industrial, and national defense needs has created an unprecedented demand for intelligent and highly trained statisticians. Research workers in many fields are requesting the statistician to help both in planning experiments or surveys and in drawing conclusions from the data. Administrators are facing the quantitative aspects of problems, such as optimum inventories, production schedules, sales efforts, pricing policies and business expansion, which call for new mathematical methods for solving problems concerned with decision making.

Comments on Cox (1957) Statistical Frontiers

G.A. Barnard

Perhaps the most obvious omission from her survey is any mention of computers, which might be thought of as large mechanised tractors which are in course of ploughing all the land she traversed, and bringing about in every area *new* and luxuriant growth. When the University of London took delivery of its Mercury Computer in the mid fifties I recall saying that at last we could really draw likelihood curves and contour plots. Yet it was not until the late seventies that a set of likelihood plots appeared in the "Annals", in a paper by Morris De Groot. Of course it would not have been possible in the mid fifties to foresee all the graphics we now can have on desk top computers, nor the computer-intensive possibilities of projection pursuit and the bootstrap. The statistical world has yet to adjust to the full possibilities opened up by computers.

A feature that has developed since Gertrude wrote—and which is to a large extent a consequence of initiatives which she promoted—is the direct involvement of statisticians in applications of their ideas. The Kettering award to David Cox is an instance of what I mean—the award normally goes to a piece of medical research. Related to this sort of development are the powerful pressures exerted both in the US and in the UK, as well as in other countries, for guaranteeing public access to accurate statistics as an important bulwark of democracy.

I.J. Good

In 1977, Gertrude Cox presented a colloquium at Virginia Tech with the title "A Consulting Statistician: Facing a Real World." She encouraged students to become consultants and see the world, and she received a standing ovation. Here's one anecdote from this colloquium. Gertrude had been invited to do some statistical consulting for a mining company and she insisted that she should be allowed to go down the mine. She was one of the rare women at that time (Eleanor Roosevelt was another) to do so. This anecdote reveals her determination and her hands-on down-into-the-earth approach to consulting. Her love of traveling, which was clear from the colloquium, would help to explain the "geographical" structure of her presidential address in 1957. Perhaps she had specific continents and countries in mind and used them to organize her address.

Her address contained about fifty suggestions for research projects, many of which are still topical. For example, she mentioned the problem of combining discrete and continuous data, an area of considerable current interest for medical diagnosis. She said she'd let us guess which ideas were original to her, but I think her main aim in this address was to be useful rather than original.

Ideas in one continent often affect others, and can even affect another world. For example, one of the earliest ideas in Gertrude's continent, Yates's adding-and-subtracting algorithm for obtaining the interactions in a 2^n factorial experiment (Yates, 1933, pp. 15 and 29; Cochran & Cox, 1957, §5.24a) led to an influential Fast Fourier Transform. It was probably anticipated by Gauss.

Gertrude's address had little to say about Computerica (two sentences on page 7), nothing on multivariate categorical data, and, apart from the two words "decision making" on page 10, she didn't mention Bayesiana. Fisher had pushed that continent away but by 1957 it was already drifting back fast.

The prediction on page 11 that "statisticians are destined for a larger role" was correct and probably influential. It was anticipated by Wilks (1950) who acknowledged the prophet H.G. Wells but without a citation. In fact Wells (1932, pp. 372 and 391) said "The science of statistics is still in its infancy—a vigorous infancy", and on page 391 "... the movement of the last hundred years is all in favour of the statistician."

References

Cochran, W.G. and Cox, G.M. (1957). *Experimental Designs.* 2nd ed. New York: Wiley.

Wells, H.G. (1932). *The Work, Wealth and Happiness of Mankind.* London: Heinemann. American printing (1931), two volumes, Doubleday, Doran & Co., Garden City.

Wilks, S.S. (1950). Undergraduate statistical education, *Journal of the American Statistical Association* **46**, 1–18.

Yates, F. (1937). *The Design and Analysis of Factorial Experiments.* Harpenden, England: Imperial Bureau of Soil Science.

D.V. Lindley

Guide books do not ordinarily concern themselves with the politics or philosophy of the countries they are describing and tourists, save in exceptional cases, ignore the manner of government. In this respect, Dr. Cox really is a tourist, not mentioning the philosophy of statistics. In 1956, this was reasonable, since Savage, the revolutionary text, had only just appeared. Jeffreys lay unread and de Finetti was still only available in Italian. The statistical world, at least in the United States, looked to be soundly governed by the Wald-Neyman-Pearson school. Few had doubts that power, confidence intervals and unbiased estimates were not completely sound. Basu had not produced his counter-examples.

Today, the travellers would surely look at the philosophy of statistics and its implication for practice. They would not be quite so confident that their methods were sound. Fisher still flourishes like an awkward government that is slightly suspect. The upstart Bayesian movement is being contained, largely by being ignored, but represents a continual threat to the establishment. Even the arithmetic mean has fallen from its pedestal and we argue about whether or not to shrink our census returns.

To leave the travel-guide analogy, there are three features that would be present in a contemporary survey yet are omitted by Cox. First, there would be a discussion about computers; about their ability to handle large data sets, to perform more complicated and larger analyses than hitherto, to simulate in procedures like the bootstrap and Gibbs sampling. Second, the topic of probability would loom larger. The ideas of influence diagrams, expert systems and artificial intelligence have led to an appreciation of probability manipulations, and especially of independence, that are important. Third, there would be some consideration of decision-making. Cox's view of a statistician's role was passive; we observe and report. There is increasing awareness today, for example in Ron Howard's recent address, to the more active statistician who contemplates risk and utility, and is prepared to advise not just about beliefs but about the actions that might spring from those beliefs.

F. Mosteller

Gertrude Cox loved travel so much that we are not surprised that she chose this analogy for her paper. Although statisticians have made progress on many of the issues that she mentions in 1956, her list leaves plenty of room for a decade more of thoughtful doctoral dissertations in the last decade of her century.

One omission I note is that in dealing with descriptive statistics, both graphical and tabular, she does not invoke the need for behavioral science to help us decide what methods of presentation come through to the viewers as especially helpful. We have had little progress in this area, though Cleveland's group has made some contributions. I look forward to big progress as computer technology offers us plenty of options for flexible and attractive presentations and for easy assessment.

An example of the kind of research needed is given in Ibrekk and Morgan (1987) where these authors explore for nontechnical users the communication merits of nine pictorial displays related to the uncertainity of a statistic.

In learning how to use graphics to improve analysis, statistics alone may well be adequate, but in improving presentation, we have to find out what methods are better at communicating, and for this nothing can replace the findings for actual users.

Reference

H. Ibrekk and M. G. Morgan, Graphical communication of uncertain quantities to nontechnical people, *Risk Analysis*, 1987, **7**: 519–529.

P.K. Sen

Looking back at this remarkable article written almost thirty-five years ago, I have nothing but deep appreciation for the utmost care with which (the late) Gertrude M. Cox depicted the *statistical universe* (in 1956) as well as for her enormous foresight. In fact, to appreciate fully this (ASA) presidential address delivered to a very wide audience (from all walks in statistical methodology and applications), it would be very appropriate to bear the remembrance of her prime accomplishments in creating such a universe in the Research Triangle Park in the heart of North Carolina, and even after 35 years, we are proudly following her footsteps.

The three major *fortresses* in her *statistical frontiers* are (i) descriptive methods, (ii) design of experiments and investigations, and (iii) analysis and theory. Collection of data, their summarization and presentation in textual/tabular/graphical forms constitute the first aspect. The advent of modern computer and statistical packages has made this job somewhat easier and mechanical, albeit the abuses of such statistical packages have been increasing at an alarming rate. The main burden lies with a good *planning* (synonymous to design) of experiments/investigations, so that the collected data convey meaningful information, can be put to valid statistical analysis, and suitable statistical packages can be incorporated in that context. In spite of the fact that most of us have our bread and butter from statistical theory and analysis, we often digress from applicable methodology onto the wilderness of abstractions. Gertrude was absolutely right in pointing out that there is a compelling need to ensure that statistical methodology is theoretically sound and at the same time adoptable in diverse practical situations. The scenario has not changed much in the past three decades, although introduction of new disciplines has called for some shifts in emphasis and broadening of the avenues emerging from the Cox fortresses.

The genesis of statistical sciences lies in a variety of disciplines ranging from agricultural science, anthropometry, biometry, genetics, sociology, economics, physical and engineering sciences, and bio-sciences to modern medicine, public health and nascent bio-technology. While Cox's thoughtful observations pertain to a greater part of this broad spectrum, there may be some need to examine minutely some of the frontiers which were mostly annexed to the Cox universe later on. In this respect, I would like to place the utmost emphasis on Energy, Ecology and Environmetrics. Our planet is endangered with the thinning of the ozone layer, extinction of several species, massive atmospheric pollution, nuclear radiation, genotoxicity, ecological imbalance and numerous other threatening factors. The thrust for energy-sufficiency and economic stability has led to global tensions, and the mankind is indeed in a perilous state. Statisticians have a basic role to play in conjunction with the scientists in other disciplines in combating this extinction. The design of such investigations may differ drastically from that of a controlled experiment. The collection of data may need careful scrutiny in order that valid statistical analysis can be done, and more noticably, novel statistical methodology has to be developed to carry out such valid and efficient statistical analysis. Lack of a control, development of proper scientific instruments to improve the measurement system, proper dimension reduction of data for efficient analysis and above all good modelling are essential factors requiring close attention from the statisticians. To a lesser extent, similar problems cropped up in the area of epidemiological investigations including clinical trials and retrospective studies, and the past two decades have witnessed phenomenal growth of the literature of statistical methodology to cope with these problems. Non-stationarity of concomitant variates (over time or space), measurement errors, doubts about the appropriateness of linear, log-linear or logistic models, and above all, the relevance of 'random sampling' schemes (particularly, equal probability sampling with/without replacement) all call for non-standard statistical analysis, for which novel methodology need to be developed. As statisticians, we have the obligation to bridge the gap between the classical theory and applicable methodology, so that valid statistical conclusions can be made in a much broader spectrum of research interest. Last year, at the Indian Science Congress Association Meeting in Madurai, I have tried to summarize this concern, and as such, I would not go into the details. Rather, I would like to conclude this discussion with the remark that most of the problems relating to multivariate analysis, nonparametric methods and sequential analysis referred to in this Cox address has been satisfactorily resolved in the past three decades, and we need to march forward beyond these traditional quarters onto the rough territories which are as yet deprived of the statistical facilities, and towards this venture, we need to accommodate a plausible shift in our statistical attitude too. Nevertheless, the Cox milestone remains a good exploration point.

Reference

Sen, P.K. (1989). Beyond the traditional frontiers of statistical sciences: A challenge for the next decade. Platinum Jubilee Lecture in Statistics, Indian Science Congress Association Meeting, Madurai. *Inst. Statist., Univ. N. Carolina Mimeo. Rep.* 1861.

Introduction to Fisher (1922) On the Mathematical Foundations of Theoretical Statistics

Seymour Geisser
University of Minnesota

1. General Remarks

This rather long and extraordinary paper is the first full account of Fisher's ideas on the foundations of theoretical statistics, with the focus being on estimation. The paper begins with a sideswipe at Karl Pearson for a purported general proof of Bayes' postulate. Fisher then clearly makes a distinction between parameters, the objects of estimation, and the statistics that one arrives at to estimate the parameters. There was much confusion between the two since the same names were given to both parameters and statistics, e.g., mean, standard deviation, correlation coefficient, etc., without an indication of whether it was the population or sample value that was the subject of discussion. This formulation of the parameter value was certainly a critical step for theoretical statistics [see, e.g., Geisser (1975), footnote on p. 320 and Stigler (1976)]. In fact, Fisher attributed the neglect of theoretical statistics not only to this failure in distinguishing between parameter and statistic but also to a philosophical reason, namely, that the study of results subject to greater or lesser error implies that the precision of concepts is either impossible or not a practical necessity. He sets out to remedy the situation, and remedy it he did. Indeed, he did this so convincingly that for the next 50 years or so almost all theoretical statisticians were completely parameter bound, paying little or no heed to inference about observables.

Fisher states that the purpose of statistical methods is to reduce a large quantity of data to a few that are capable of containing as much as possible of the relevant information in the original data. Because the data will generally supply a large number of "facts," many more than are sought, much information in the data is irrelevant. This brings to the fore the Fisherian dictum that statistical analysis via the reduction of data is the process of extracting the relevant information and excluding the irrelevant information. A way of accomplishing this is by modeling a hypothetical population specified by relatively few parameters.

Hence, the critical problems of theoretical statistics in 1920, according to Fisher, were (1) specification, choice of the hypothetical parametric distribution; (2) estimation, choice of the statistics for estimating the unknown parameters of the distribution; (3) sampling distributions, the exact or approximate distributions of the statistics used to estimate the parameters. For a majority of statisticians, these have been and still are the principal areas of statistical endeavor, 70 years later. The two most important additions to this view are that the parametric models were, at best, merely approximations of the underlying process generating the observations, and in view of this, much greater emphasis should be placed on observable inference rather than on parametric inference.

2. Foundational Developments

In this paper, Fisher develops a number of concepts relevant to the estimation of parameters. Some were previously introduced but not generally developed, and others appear for the first time. Here, also, the richness of Fisher's *lingua statistica* emerges, yielding poignant appelatives for his concepts, vague though some of them are. This activity will continue throughout all his future contributions. First he defines consistency: A statistic is consistent if, when calculated from the whole population, it is equal to the parameter describing the probability law. This is in contradistinction to the usual definition which entails a sequence of estimates, one for each sample size, that converges in probability to the appropriate parameter. While Fisher consistency is restricted to repeated samples from the same distribution, it does not suffer from the serious defect of the usual definition. That flaw was formally pointed out later by Fisher (1956): Suppose one uses an arbitrary value A for an estimator for $n < n_1$, where n is as large as one pleases, and for $n > n_1$ uses an asymptotically consistent estimator T_n. The entire sequence, now corrupted by A for $n < n_1$ and then immaculately transformed to T_n thereafter, remains a useless, but perfectly well-defined, consistent estimator for any n. Fisher is not to be trifled with!

Indicating that many statistics for the same parameter can be Fisher-consistent, in particular, the sample standard deviation and sample mean deviation for the standard deviation of a normal population, he goes on to suggest a criterion for efficiency. It is a large sample definition. Among all estimators for a parameter that are Fisher-consistent and whose distributions are asymptotically normal, the one with the smallest variance is efficient. Later, he shows that when the asymptotic distribution of the method of moments estimator is normal for the location of a uniform distribution while that of the "optimum" estimator is double exponential, he realizes that the variance does not necessarily provide a satisfactory basis for comparison, especially for small samples. Thus, he also recognizes that his large sample definition of intrinsic accuracy (a measure of relative efficiency) should not be based on variances and a definition appropriate for small samples is required. In later papers, e.g., Fisher (1925), vague concepts of intrinsic accuracy will be replaced by the more precise amount of information per observation. At any rate, the large sample criterion is incomplete and needs to be supplemented by a sufficiency criterion. The "remarkable" property of this concept was previously pointed out when introduced for a special case without giving it a name [Fisher (1920)]. A statistic, then, is sufficient if it contains all the information in the sample regarding the parameter to be estimated; that is, given a sufficient statistic, the distribution of any other statistic does not involve the parameter. This compelling concept of his, including the factorization result, is still in vogue. Assuming a sufficient statistic and any other statistic whose joint distribution is asymptotically bivariate normal with both means being the parameter estimated, he then "demonstrates" that the sufficient statistic has an asymptotic variance smaller than that of the other statistic by a clever conditioning argument that exploits the correlation between the statistics. Hence, he claims that a sufficient* statistic satisfies the criterion of (large sample) efficiency. This "proof" of course could only apply to those statistics whose asymptotic bivariate distribution with the sufficient statistic was normal.

He comments further on the method of moments estimation procedure. While ascribing great practical utility to it, he also exposes some of its shortcomings. In particular, in estimating the center of a one-parameter Cauchy distribution, he points out that the first sample moment, the sample mean, which is the method of moments estimator is not consistent but the median is. He also cautions against the statistical rejection of outliers unless there are other substantive reasons. Rather than outright rejection, he proposes that it seriously be considered that the error distribution is not normal. Fisher effectively argues that the specification of the underlying probability law will generally require the full set of observations. A sufficient reduction is only meaningful once the probability law has been adequately established.

3. Maximum Likelihood

Fisher begins this part of his discourse acknowledging, first, that properties such as sufficiency, efficiency, and consistency per se were inadequate in directly obtaining an estimator. In solving any particular problem, we would

* In the author's note, Fisher (1950), there is a handwritten correction to the definition of intrinsic accuracy replacing sufficiency by efficiency, possibly based on his later recognition that maximum likelihood estimators were not always sufficient.

require a method that would lead automatically to the statistic which satisfied these criteria. He proposes such a method to be that of maximum likelihood, while admitting dissatisfaction with regard to the mathematical rigor of any proof that he can devise toward that result. Publication would have been withheld until a rigorous proof was found, but the number and variety of new results emanating from this method pressed him to publish. With some uncharacteristic humility, he says, "I am not insensible of the advantage which accrues to Applied Mathematics from the cooperation of the Pure Mathematician and this cooperation is not infrequently called forth by the very imperfections of writers on Applied Mathematics." This totally disarming statement would preclude any harsh commentary on the evident lack of rigor in many of his "proofs" here. Such evident modesty and good feelings toward mathematicians would never again flow from his pen.

Fisher (1912) had earlier argued for a form of maximum likelihood estimation. He had taken what superficially appeared to be a Bayesian approach because the maximizing procedure resembled the calculation of the mode of a posterior probability. In the present paper, he is very concerned to differentiate it from the Bayesian approach. He also argues against the "customary" Bayesian use of flat priors on the grounds that different results are obtained when different scales for the parameters are considered.

To illustrate Fisher's argument, suppose x denotes the number of successes out of n independent trials with probability of success; then the likelihood function is

$$L(p) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \qquad (0 < p < 1),$$

which is maximized when p is chosen to be x/n. Now, if a uniform distribution on (0, 1) is taken to be the prior distribution of p, then Bayesian analysis would yield

$$\pi(p) \propto p^x (1-p)^{n-x}$$

as the posterior density of p. But if we parameterize this Bernoulli process in a different way, say, in terms of θ with $\sin\theta = 2p - 1$, then the likelihood function of θ is

$$L(\theta) = \frac{n!}{x!(n-x)!} \frac{(1+\sin\theta)^x}{2^x} \frac{(1-\sin\theta)^{n-x}}{2^{n-x}} \qquad \left(-\frac{\pi}{2} < \theta < \frac{\pi}{2}\right),$$

which, when maximized with respect to θ, gives $\sin\hat{\theta} = (2x - n)/n = 2\hat{p} - 1$. Thus, the maximum likelihood estimate is invariant under a 1-1 transformation. For the Bayes approach, he questions the assignment of a prior assigned to θ. The uniformity of θ on $(-\pi/2, \pi/2)$ leads to the posterior density of p as

$$\pi(p) \propto p^{x-1/2}(1-p)^{n-x-1/2},$$

which is different from the previous result above. Due to this inconsistency and other reasons, Fisher derides the arbitrariness of the Bayes prior and mation. But $\hat{\theta}$ has a smaller mean squared error for a sufficiently large sample size.

The method of maximum likelihood appears to have been anticipated by Edgeworth (1908–9) according to Pratt (1976). Although there is less than universal consensus for this view, there is ample evidence that Edgeworth derived the method in the translation case directly and also using inverse probability. It appears he also conjectured the asymptotic efficiency of the method without giving it a name.

4. Other Topics

The remainder of the paper contains mainly applications of maximum likelihood techniques and various relative efficiency calculations. There is a long discussion of the Pearson system of frequency curves. This section serves mainly to display Fisher's analytic virtuosity in handling the Pearson system, also displaying graphs that serve to characterize the system in a more useful form than previously. This enables him to calculate for the various Pearson frequency curves,‡ regions for varying percent efficiencies of the method of moments estimators of location and scale. He also determines the conditions that make them fully efficient. In the latter case, he shows that if the log of a density is quartic, under certain conditions it will be approximately normal and fit the Pearson system. In dealing with the Pearson-type III curve, he now demonstrates that the asymptotic variance of the maximum likelihood estimators of scale a and shape p is smaller than that of their method of moments counterparts. However, he fails to remark or perhaps notice the anomaly of the nonregular case. Here the asymptotic variance of the maximum likelihood estimator of a is larger when m and p are given than when only p is given. Similarly, the maximum likelihood estimator of p has smaller asymptotic variance when a and m are unknown than when a and m are known.

Interest in the Pearsonian system has declined considerably over the years, being supplanted by so-called nonparametric and robust procedures, and revival appears unlikely unless Bayesians find use for them. The final part of the paper looks at discrete distributions, where the method of minimum chi-square is related to maximum likelihood, and the problem of Sheppard's correction for grouped normal data is addressed in detail. This and the material on the Pearson system actually make up the bulk of the paper. No doubt of considerable interest 70 years ago, it is of far less interest than the preceding work on the foundations. Fisher implies as much in his author's note. However, there is a final example that deals with the distribution of observations in a dilution series that is worthy of careful examination.

‡ The density on the bottom of page 342 as well as the one on page 343 of the original paper contain misprints. The section involving this material has been omitted in the abridged version of Fisher's paper which follows.

After earlier displaying the potential lack of efficiency inherent in an uncritical application of the method of moments, Fisher in an ingenious *volte-face* produces an estimation procedure for a dilution series example, which, though inefficient, is preferable to a fully efficient one essentially for economic and practical reasons. To be sure, in later years Fisher fulminated against the wholesale introduction of utility or decision theory into scientific work, but rarely again were such principles so elegantly and unobtrusively applied to such a significant practical problem. The analysis here represents a peerless blend of theory and application.

An important monitoring procedure, of ongoing interest and wide applicability, used in this instance for estimating the density of protozoa in soils, was brought to Fisher's attention. A series of dilutions of a soil sample solution were made such that each is reduced by a factor a. At each dilution, a uniform amount of the solution is deposited on s different plates containing a nutrient. After a proper incubation period, the number of protozoa on each plate is to be counted. A reasonable model for such a situation is that the chance of z protozoa on a plate is Poisson-distributed with expected value θ/a^x, where θ is the density or number of protozoa per unit volume of the original solution, and x the dilution level. A large number of such series were made daily for a variety of organisms. It proved either physically impossible or economically prohibitive to count the number of such organisms on every plate for many such series in order to estimate θ. First, Fisher suggests that only those plates containing no organisms be counted; the chance of such an occurrence at level x is $p_x = \exp(-\theta/a^x)$. By this device, an experimentally feasible situation is attained that produces a joint likelihood for Y_x, the number of sterile plates at level x, as

$$L = \prod_{x=0}^{k} \binom{s}{y_x} p_x^{y_x}(1 - p_x)^{s-y_x}$$

for dilution levels $x = 0, 1, \ldots, k$. He then calculates the contribution of a plate at level x to the information about $\log \theta$ to be

$$w_x = p_x(1 - p_x)^{-1}(\log p_x)^2.$$

This is informative as to the number of dilution levels necessary in such experiments. Further, the total expected information is approximately given as

$$s \sum_x w_x \approx \frac{s\pi^2}{(6 \log a)}.$$

The maximum likelihood solution to the problem, however, required a heavy investment of time and effort given the computational facilities of 1922. (Of course, it can easily be done today.)

At this point, Fisher employs a second wrinkle that makes the problem tractable. He suggests that the expected total number of sterile plates be equated to the observed total number in order to obtain an estimate of θ. This "rough" procedure has expected information with respect to $\log \theta$ of approximate value

$$\frac{s}{\log 2 \log a}.$$

This results in a very quick and easy procedure possessing an efficiency, independent of the dilution factor, of about 88%.

This may very well be one of the earliest statistical applications of a decision like approach to the analysis of data.

5. Summary

Clearly Fisher's paper was a landmark event in theoretical statistics. While it suffered from a lack of mathematical rigor, long analytic excursions into areas of lesser interest, and some confusion in parts, the novelty and number of ideas expressed here, both those developed from previous work and newly introduced, are still compelling for most statisticians. Although this paper is infrequently cited, its influence completely pervades the subsequent paper [Fisher (1925)],§ which presents a clearer exposition of his views. However, he poured into the 1922 paper, pell-mell, all his creative thinking and work on the foundations of statistics, the major exception being the fiducial argument. This work, filtered through the 1925 paper, has had a profound impact on statistical thinking unto this day. One has only to scan any serious work on the foundations to see that these ideas still have relevance in statistical theory, although citation is almost always to the 1925 paper.

References

Barnard, G., Jenkins, G.M. and Winsten, C.B. (1963). Likelihood inference and time series, *Jo. Roy. Statist. Soc., Ser. A*, **125**, 321–372.

Boole, G. (1854). *The Laws of Thought*. Dover, New York.

Chrystal, G. (1886). *Algebra*. Adam and Charles Black, London.

Edgeworth, F.Y. (1908–9). On the probable errors of frequency-constants, *J. Roy. Statist. Soc.*, **71** 381–397, 499–512, 651–678. Addendum ibid. **72**, 81–90.

Edwards, A.W.F. (1972). *Likelihood*. Cambridge University Press, New York.

Fisher, R.A. (1912). On an absolute criterion for fitting frequency curves, *Messenger of Math.*, **41**, 155–160.

Fisher, R.A. (1920). A mathematical examination of the methods of determining the accuracy of an observation by the mean error, and by the mean square error, *Monthly Notices Roy. Astro. Soc.*, **80**, 758–770.

§ This paper is cited a number of times in Fisher (1956), his final work on the foundations, while the seminal 1922 paper is not mentioned. In fact, the dozen or so times that Fisher subsequently cites the 1922 paper, he misdates it about half the time as 1921. This Fisherian slip, making him a year younger at its publication, accords with the author's note attributing certain deficiencies in the paper to youth.

Fisher, R.A. (1925). Theory of statistical estimation, *Proc. Cambridge Philos. Soc.*, **22**, 700–725.

Fisher, R. A. (1950). *Contributions to Mathematical Statistics.* Wiley, New York.

Fisher, R.A. (1956). *Statistical Methods and Scientific Inference.* Oliver and Boyd, Edinburgh.

Geisser, S. (1975). The predictive sample reuse method with applications, *Jo. Amer. Statist. Assoc.*, **70**, 320–328.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems, *Proc. Roy. Soc. London, Ser. A*, **186**, 453–454.

Pratt, J.W. (1976). F.V. Edgeworth and R.A. Fisher on the efficiency of maximum likelihood estimation, *Ann. Statist.*, 501–514.

Savage, L.J. (1976). On Rereading R.A. Fisher (with discussion), *Ann. Statist.*, **4**, 441–500.

Stigler, S.M. (1976). Discussion of Savage (1976), *Ann. Statist.*, **4**, 498–500.

Venn, J. (1866). *The Logic of Chance.* Macmillan, London.

On the Mathematical Foundations of Theoretical Statistics

R.A. Fisher
Fellow of Gonville and Caius College,
Chief Statistician, Rothamsted Experimental Station

Definitions

Centre of Location. That abscissa of a frequency curve for which the sampling errors of optimum location are uncorrelated with those of optimum scaling. (9.)

Consistency. A statistic satisfies the criterion of consistency, if, when it is calculated from the whole population, it is equal to the required parameter. (4.)

Distribution. Problems of distribution are those in which it is required to calculate the distribution of one, or-the simultaneous distribution of a number, of functions of quantities distributed in a known manner. (3.)

Efficiency. The efficiency of a statistic is the ratio (usually expressed as a percentage) which its intrinsic accuracy bears to that of the most efficient statistic possible. It expresses the proportion of the total available relevant information of which that statistic makes use. (4 and 10.)

Efficiency (Criterion). The criterion of efficiency is satisfied by those statistics which, when derived from large samples, tend to a normal distribution with the least possible standard deviation. (4.)

Estimation. Problems of estimation are those in which it is required to estimate the value of one or more of the population parameters from a random sample of the population. (3.)

Intrinsic Accuracy. The intrinsic accuracy of an error curve is the weight in large samples, divided by the number in the sample, of that statistic of location which satisfies the criterion of efficiency. (9.)

Isostatistical Regions. If each sample be represented in a generalized space of which the observations are the co-ordinates, then any region throughout elementary of statistical concepts. It is a parameter which specifies a simple dichotomy in an infinite hypothetical population, and it represents neither more nor less than the frequency ratio which we imagine such a population to exhibit. For example, when we say that the probability of throwing a five with a die is one-sixth, we must not be taken to mean that of any six throws with that die one and one only will necessarily be a five; ar that of any six million throws, exactly one million will be fives; but that of a hypothetical population of an infinite number of throws, with the die in its original condition, exactly one-sixth will be fives. Our statement will not then contain any false assumption about the actual die, as that it will not wear out with continued use, or any notion of approximation, as in estimating the probability from a finite sample, although this notion may be logically developed once the meaning of probability is apprehended.

The concept of a *discontinuous frequency distribution* is merely an extension of that of a simple dichotomy, for though the number of classes into which the population is divided may be infinite, yet the frequency in each class bears a finite ratio to that of the whole population. In *frequency curves*, however, a second infinity is introduced. No finite sample has a frequency curve: a finite sample may be represented by a histogram, or by a frequency polygon, which to the eye more and more resembles a curve, as the size of the sample is increased. To reach a true curve, not only would an infinite number of individuals have to be placed in each class, but the number of classes (arrays) into which the population is divided must be made infinite. Consequently, it should be clear that the concept of a frequency curve includes that of a hypothetical infinite population, distributed according to a mathematical law, represented by the curve. This law is specified by assigning to each element of the abscissa the corresponding element of probability. Thus, in the case of the normal distribution, the probability of an observation falling in the range dx, is

$$\frac{1}{\sigma\sqrt{2\pi}}e^{-(x-m)^2/2\sigma^2}\,dx,$$

in which expression x is the value of the variate, while m, the mean, and σ, the standard deviation, are the two parameters by which the hypothetical population is specified. If a sample of n be taken from such a population, the data comprise n independent facts. The statistical process of the reduction of these data is designed to extract from them all relevant information respecting the values of m and σ, and to reject all other information as irrelevant.

It should be noted that there is no falsehood in interpreting any set of independent measurements as a random sample from an infinite population; for any such set of numbers are a random sample from the totality of numbers produced by the same matrix of causal conditions: the hypothetical population which we are studying is an aspect of the totality of the effects of these conditions, of whatever nature they may be. The postulate of randomness thus resolves itself into the question, "Of what population is this a random sample?" which must frequently be asked by every practical statistician.

It will be seen from the above examples that the process of the reduction of data is, even in the simplest cases, performed by interpreting the available observations as a sample from a hypothetical infinite population; this is *a fortiori* the case when we have more than one variate, as when we are seeking the values of coefficients of correlation. There is one point, however, which may be briefly mentioned here in advance, as it has been the cause of some confusion. In the example of the frequency curve mentioned above, we took it for granted that the values of both the mean and the standard deviation of the population were relevant to the inquiry. This is often the case, but it sometimes happens that only one of these quantities, for example the standard deviation, is required for discussion. In the same way an infinite normal population of two correlated variates will usually require five parameters for its specification, the two means, the two standard deviations, and the correlation; of these often only the correlation is required, or if not alone of interest, it is discussed without reference to the other four quantities. In such cases an alteration has been made in what is, and what is not, relevant, and it is not surprising that certain small corrections should appear, or not, according as the other parameters of the hypothetical surface are or are not deemed relevant. Even more clearly is this discrepancy shown when, as in the treatment of such fourfold tables as exhibit the recovery from smallpox of vaccinated and unvaccinated patients, the method of one school of statisticians treats the proportion of vaccinated as relevant, while others dismiss it as irrelevant to the inquiry. (3.)

3. The Problems of Statistics

The problems which arise in reduction of data may be conveniently divided into three types:

(1) Problems of Specification. These arise in the choice of the mathematical form of the population.

(2) Problems of Estimation. These involve the choice of methods of calculating from a sample statistical derivates, or as we shall call them statistics, which are designed to estimate the values of the parameters of the hypothetical population.

(3) Problems of Distribution. These include discussions of the distribution of statistics derived from samples, or in general any functions of quantities whose distribution is known.

It will be clear that when we know (1) what parameters are required to specify the population from which the sample is drawn, (2) how best to calculate from the sample estimates of these parameters, and (3) the exact form of the distribution, in different samples, of our derived statistics, then the theoretical aspect of the treatment of any particular body of data has been completely elucidated.

As regards problems of specification, these are entirely a matter for the practical statistician, for those cases where the qualitative nature of the hypothetical population is known do not involve any problems of this type. In other cases we may know by experience what forms are likely to be suitable, and the adequacy of our choice may be tested *a posteriori.* We must confine ourselves to those forms which we know how to handle, or for which any tables which may be necessary have been constructed. More or less elaborate forms will be suitable according to the volume of the data. Evidently these are considerations the nature of which may change greatly during the work of a single generation. We may instance the development by Pearson of a very extensive system of skew curves, the elaboration of a method of calculating their parameters, and the preparation of the necessary tables, a body of work which has enormously extended the power of modern statistical practice, and which has been, by pertinacity and inspiration alike, practically the work of a single man. Nor is the introduction of the Pearsonian system of frequency curves the only contribution which their author has made to the solution of problems of specification: of even greater importance is the introduction of an objective criterion of goodness of fit. For empirical as the specification of the hypothetical population may be, this empiricism is cleared of its dangers if we can apply a rigorous and objective test of the adequacy with which the proposed population represents the whole of the available facts. Once a statistic, suitable for applying such a test, has been chosen, the exact form of its distribution in random samples must be investigated, in order that we may evaluate the probability that a worse fit should be obtained from a random sample of a population of the type considered. The possibility of developing complete and self-contained tests of goodness of fit deserves very careful consideration, since therein lies our justification for the free use which is made of empirical frequency formulae. Problems of distribution of great mathematical difficulty have to be faced in this direction.

Although problems of estimation and of distribution may be studied separately, they are intimately related in the development of statistical methods. Logically problems of distribution should have prior consideration, for the study of the random distribution of different suggested statistics, derived from samples of a given size, must guide us in the choice of which statistic it is most profitable to calculate. The fact is, however, that very little progress has been made in the study of the distribution of statistics derived from samples. In 1900 Pearson (15) gave the exact form of the distribution of χ^2, the Pearsonian test of goodness of fit, and in 1915 the same author published (18) a similar result of more general scope, valid when the observations are regarded as subject to linear constraints. By an easy adaptation (17) the tables of probability derived from this formula may be made available for the more numerous cases in which linear constraints are imposed upon the hypothetical population by the means which we employ in its reconstruction. The distribution of the mean of samples of n from a normal population has long been known, but in 1908 "Student" (4) broke new ground by calculating the distribution of the ratio which the deviation of the mean from its population value bears to the standard deviation calculated from the sample. At the same time he gave the exact form of the distribution in samples of the standard deviation. In 1915 Fisher (5) published the curve of distribution of the correlation coefficient for the standard method of calculation, and in 1921 (6) he published the corresponding series of curves for intraclass correlations. The brevity of this list is emphasised by the absence of investigation of other important statistics, such as the regression coefficients, multiple correlations, and the correlation ratio. A formula for the probable error of any statistic is, of course, a practical necessity, if that statistic is to be of service: and in the majority of cases such formulae have been found, chiefly by the labours of Pearson and his school, by a first approximation, which describes the distribution with sufficient accuracy if the sample is sufficiently large. Problems of distribution, other than the distribution of statistics, used to be not uncommon as examination problems in probability, and the physical importance of problems of this type may be exemplified by the chemical laws of mass action, by the statistical mechanics of Gibbs, developed by Jeans in its application to the theory of gases, by the electron theory of Lorentz, and by Planck's development of the theory of quanta, although in all these applications the methods employed have been, from the statistical point of view, relatively simple.

The discussions of theoretical statistics may be regarded as alternating between problems of estimation and problems of distribution. In the first place a method of calculating one of the population parameters is devised from common-sense considerations: we next require to know its probable error, and therefore an approximate solution of the distribution, in samples, of the statistic calculated. It may then become apparent that other statistics may be used as estimates of the same parameter. When the probable errors of these statistics are compared, it is usually found that, in large samples, one particular method of calculation gives a result less subject to random errors than those given by other methods of calculation. Attacking the problem more thoroughly, and calculating the surface of distribution of any two statistics, we may find that the whole of the relevant information contained in one is contained in the other: or, in other words, that when once we know the other, knowledge of the first gives us no further information as to the value of the parameter. Finally it may be possible to prove, as in the case of the Mean Square Error, derived from a sample of normal population (7), that a particular statistic summarises the whole of the information relevant to the corresponding parameter, which the sample contains. In such a case the problem of estimation is completely solved.

4. Criteria of Estimation

The common-sense criterion employed in problems of estimation may be stated thus:—That when applied to the whole population the derived statistic should be equal to the parameter. This may be called the *Criterion of Consistency.* It is often the only test applied: thus, in estimating the standard deviation of a normally distributed population, from an ungrouped sample, either of the two statistics—

$$\sigma_1 = \frac{1}{n}\sqrt{\frac{\pi}{2}}S(|x - \bar{x}|) \qquad \text{(Mean error)}$$

and

$$\sigma_2 = \sqrt{\frac{1}{n}S(x - \bar{x})^2} \qquad \text{(Mean square error)}$$

will lead to the correct value, σ, when calculated from the whole population. They both thus satisfy the criterion of consistency, and this has led many computers to use the first formula, although the result of the second has 14 per cent. greater weight (7), and the labour of increasing the number of observations by 14 per cent. can seldom be less than that of applying the more accurate formula.

Consideration of the above example will suggest a second criterion, namely: —That in large samples, when the distributions of the statistics tend to normality, that statistic is to be chosen which has the least probable error.

This may be called the *Criterion of Efficiency.* It is evident that if for large samples one statistic has a probable error double that of a second, while both are proportional to $n^{-1/2}$, then the first method applied to a sample of $4n$ values will be no more accurate than the second applied to a sample of any n values. If the second method makes use of the whole of the information available, the first makes use of only one-quarter of it, and its efficiency may therefore be said to be 25 per cent. To calculate the efficiency of any given method, we must therefore know the probable error of the statistic calculated by that method, and that of the most efficient statistic which could be used. The square of the ratio of these two quantities then measures the efficiency.

The criterion of efficiency is still to some extent incomplete, for different methods of calculation may tend to agreement for large samples, and yet differ for all finite samples. The complete criterion suggested by our work on the mean square error (7) is:

That the statistic chosen should summarise the whole of the relevant information supplied by the sample.

This may be called the *Criterion of Sufficiency.*

In mathematical language we may interpret this statement by saying that if θ be the parameter to be estimated, θ_1 a statistic which contains the whole of the information as to the value of θ, which the sample supplies, and θ_2 any other statistic, then the surface of distribution of pairs of values of θ_1 and θ_2, the four moments of the grouped population

$$_1\mathrm{A}_0 = \int_{-\infty}^{\infty} xf(x)\,dx,$$

$$_2\mathrm{A}_0 = \int_{-\infty}^{\infty} \left(x^2 + \frac{a^2}{12}\right) f(x)\,dx,$$

$$_3\mathrm{A}_0 = \int_{-\infty}^{\infty} \left(x^3 + \frac{a^2 x}{4}\right) f(x)\,dx,$$

$$_4\mathrm{A}_0 = \int_{-\infty}^{\infty} \left(x^4 + \frac{a^2 x^2}{2} + \frac{a^4}{80}\right) f(x)\,dx.$$

If we ignore the periodic terms, these equations lead to the ordinary Sheppard corrections for the second and fourth moment. The nature of the approximation involved is brought out by the periodic terms. In the absence of high contact at the ends of the curve, the contribution of these will, of course, include the terms given in a recent paper by Pearson (8); but even with high contact it is of interest to see for what degree of coarseness of grouping the periodic terms become sensible.

Now

$$\begin{aligned}
\mathrm{A}_\mathrm{S} &= \frac{1}{\pi} \sum_{p=-\infty}^{p=\infty} \int_0^{2\pi} \sin s\theta\, d\theta \int_{\xi-(1/2)a}^{\xi+(1/2)a} \xi^k f(x)\,dx, \\
&= \frac{2}{a} \int_{-\infty}^{\infty} \sin \frac{2\pi s\xi}{a}\, d\xi \int_{\xi-(1/2)a}^{\xi+(1/2)a} \xi^k f(x)\,dx, \\
&= \frac{2}{a} \int_{-\infty}^{\infty} f(x)\,dx \int_{\xi-(1/2)a}^{\xi+(1/2)a} \xi^k \sin \frac{2\pi s\xi}{a}\, d\xi.
\end{aligned}$$

But

$$\frac{2}{a} \int_{x-(1/2)a}^{x+(1/2)a} \xi \sin \frac{2\pi s\xi}{a}\, d\xi = -\frac{a}{\pi s} \cos \frac{2\pi s x}{a} \cos \pi s,$$

therefore

$$_1\mathrm{A}_\mathrm{S} = (-)^{s+1} \frac{a}{\pi s} \int_{-\infty}^{\infty} \cos \frac{2\pi s x}{a} f(x)\,dx;$$

similarly the other terms of the different moments may be calculated.

For a normal curve referred to the true mean

$$_1\mathrm{A}_\mathrm{S} = (-)^{s+1} \frac{2\varepsilon}{s} e^{-(s^2\sigma^2/2\varepsilon^2)},$$

$$_1\mathrm{B}_\mathrm{S} = 0,$$

in which

$$a = 2\pi\varepsilon.$$

The error of the mean is therefore

$$-2\varepsilon(e^{-(\sigma^2/2\varepsilon^2)}\sin\theta - \tfrac{1}{2}e^{-(4\sigma^2/2\varepsilon^2)}\sin 2\theta + \tfrac{1}{3}e^{-(9\sigma^2/2\varepsilon^2)}\sin 3\theta - \cdots).$$

To illustrate a coarse grouping, take the group interval equal to the standard deviation: then

$$\varepsilon = \frac{\sigma}{2\pi},$$

and the error is

$$-\frac{\sigma}{\pi}e^{-2\pi^2}\sin\theta$$

with sufficient accuracy. The standard error of the mean being $\frac{\sigma}{\sqrt{n}}$, we may calculate the size of the sample for which the error due to the periodic terms becomes equal to one-tenth of the standard error, by putting

$$\frac{\sigma}{10\sqrt{n}} = \frac{\sigma}{\pi}e^{-2\pi^2},$$

whence

$$n = \frac{\pi^2}{100}e^{4\pi^2} = 13{,}790 \times 10^{12}.$$

For the second moment

$$\mathrm{B}_s = (-)^s\, 4\left(\sigma^2 + \frac{\varepsilon^2}{s^2}\right)e^{-(s^2\sigma^2/2\varepsilon^2)},$$

and, if we put

$$\frac{\sqrt{2}\sigma^2}{10\sqrt{n}} = 4\sigma^2 e^{-2\pi^2},$$

there results

$$n = \tfrac{1}{800}e^{4\pi^2} = 175 \times 10^{12}.$$

The error, while still very minute, is thus more important for the second than for the first moment.

For the third moment

$$\mathrm{A}_s = (-)^s\,\frac{6\sigma^4 s}{\varepsilon}\left\{1 + \frac{\varepsilon^2}{s^2\sigma^2} - \frac{\varepsilon^4}{3s^4\sigma^4}(\pi^2 s^2 - 6)\right\}e^{-(s^2\sigma^2/2\varepsilon^2)};$$

putting

$$\frac{\sqrt{15}\sigma^3}{10\sqrt{n}} = 12\pi\sigma^3 e^{-2\pi^2},$$

$$n = \frac{1}{960\pi^2}e^{4\pi^2} = 147 \times 10^{12}.$$

While for the fourth moment

$$\mathrm{B}_s = (-)^{s+1}\frac{8\sigma^6 s^2}{\varepsilon^2}\left\{1 - (\pi^2 s^2 - 3)\frac{\varepsilon^4}{s^4\sigma^4} - (\pi^2 s^2 - 6)\frac{\varepsilon^6}{s^6\sigma^6}\right\}e^{-(s^2\sigma^2/2\varepsilon^2)},$$

so that, if we put,

$$\frac{\sqrt{96}\sigma^4}{10\sqrt{n}} = 32\pi^2\sigma^4 e^{-2\pi^2},$$

$$n = \frac{3}{3200\pi^4}e^{4\pi^2} = 1.34 \times 10^{12}.$$

In a similar manner the exact form of Sheppard's correction may be found for other curves; for the normal curve we may say that the periodic terms are exceedingly minute so long as a is less than σ, though they increase very rapidly if a is increased beyond this point. They are of increasing importance as higher moments are used, not only absolutely, but relatively to the increasing probable errors of the higher moments. The principle upon which the correction is based is merely to find the error when the moments are calculated from an infinite grouped sample; the corrected moment therefore fulfils the criterion of consistency, and so long as the correction is small no greater refinement is required.

Perhaps the most extended use of the criterion of consistency has been developed by Pearson in the "Method of Moments." In this method, which is without question of great practical utility, different forms of frequency curves are fitted by calculating as many moments of the sample as there are parameters to be evaluated. The parameters chosen are those of an infinite population of the specified type having the same moments as those calculated from the sample.

The system of curves developed by Pearson has four variable parameters, and may be fitted by means of the first four moments. For this purpose it is necessary to confine attention to curves of which the first four moments are finite; further, if the accuracy of the fourth moment should increase with the size of the sample, that is, if its probable error should not be infinitely great, the first eight moments must be finite. This restriction requires that the class of distribution in which this condition is not fulfilled should be set aside as "heterotypic," and that the fourth moment should become practically valueless as this class is approached. It should be made clear, however, that there is nothing anomalous about these so-called "heterotypic" distributions except the fact that the method of moments cannot be applied to them. Moreover, for that class of distribution to which the method can be applied, it has not been shown, except in the case of the normal curve, that the best values will be obtained by the method of moments. The method will, in these cases, certainly be serviceable in yielding an approximation, but to discover whether this approximation is a good or a bad one, and to improve it, if necessary, a more adequate criterion is required.

A single example will be sufficient to illustrate the practical difficulty al-

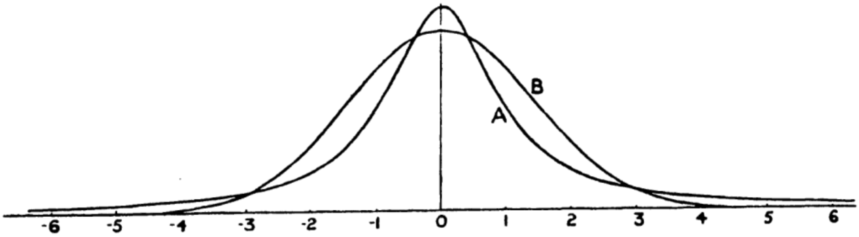

Figure 1. Symmetrical error curves of equal intrinsic accuracy:

$$\text{A}\ldots\ldots df = \frac{1}{\pi}\frac{dx}{1+x^2}.$$

$$\text{B}\ldots\ldots df = \frac{1}{2\sqrt{\pi}}e^{-x^2/4}$$

luded to above. If a point P lie at known (unit) distance from a straight line AB, and lines be drawn at random through P, then the distribution of the points of intersection with AB will be distributed so that the frequency in any range dx is

$$df = \frac{1}{\pi}\cdot\frac{dx}{1+(x-m)^2},$$

in which x is the distance of the infinitesimal range dx from a fixed point 0 on the line, and m is the distance, from this point, of the foot of the perpendicular PM. The distribution will be a symmetrical one (Type VII.) having its centre at $x = m$ (fig. 1). It is therefore a perfectly definite problem to estimate the value of m (to find the best value of m) from a random sample of values of x. We have stated the problem in its simplest possible form: only one parameter is required, the middle point of the distribution. By the method of moments, this should be given by the first moment, that is by the mean of the observations: such would seem to be at least a good estimate. It is, however, entirely valueless. The distribution of the mean of such samples is in fact the same, identically, as that of a single observation. In taking the mean of 100 values of x, we are no nearer obtaining the value of m than if we had chosen any value of x out of the 100. The problem, however, is not in the least an impracticable one: clearly from a large sample we ought to be able to estimate the centre of the distribution with some precision; the mean, however, is an entirely useless statistic for the purpose. By taking the median of a large sample, a fair approximation is obtained, for the standard error of the median of a large sample of n is $\frac{\pi}{2\sqrt{n}}$, which, alone, is enough to show that by adopting adequate statistical methods it must be possible to estimate the value for m, with increasing accuracy, as the size of the sample is increased.

This example serves also to illustrate the practical difficulty which observers often find, that a few extreme observations appear to dominate the value of the mean. In these cases the rejection of extreme values is often advocated, and it may often happen that gross errors are thus rejected. As a statistical measure, however, the rejection of observations is too crude to be defended: and unless there are other reasons for rejection than mere divergence from the majority, it would be more philosophical to accept these extreme values, not as gross errors, but as indications that the distribution of errors is not normal. As we shall show, the only Pearsonian curve for which the mean is the best statistic for locating the curve, is the normal or gaussian curve of errors. If the curve is not of this form the mean is not necessarily, as we have seen, of any value whatever. The determination of the true curves of variation for different types of work is therefore of great practical importance, and this can only be done by different workers recording their data in full without rejections, however they may please to treat the data so recorded. Assuredly an observer need be exposed to no criticism, if after recording data which are not probably normal in distribution, he prefers to adopt some value other than the arithmetic mean.

6. Formal Solution of Problems of Estimation

The form in which the criterion of sufficiency has been presented is not of direct assistance in the solution of problems of estimation. For it is necessary first to know the statistic concerned and its surface of distribution, with an infinite number of other statistics, before its sufficiency can be tested. For the solution of problems of estimation we require a method which for each particular problem will lead us automatically to the statistic by which the criterion of sufficiency is satisfied. Such a method is, I believe, provided by the Method of Maximum Likelihood, although I am not satisfied as to the mathematical rigour of any proof which I can put forward to that effect. Readers of the ensuing pages are invited to form their own opinion as to the possibility of the method of the maximum likelihood leading in any case to an insufficient statistic. For my own part I should gladly have withheld publication until a rigorously complete proof could have been formulated; but the number and variety of the new results which the method discloses press for publication, and at the same time I am not insensible of the advantage which accrues to Applied Mathematics from the co-operation of the Pure Mathematician, and this co-operation is not infrequently called forth by the very imperfections of writers on Applied Mathematics.

If in any distribution involving unknown parameters θ_1, θ_2, θ_3, $\ldots$, the chance of an observation falling in the range dx be represented by

$$f(x, \theta_1, \theta_2, \ldots)\,dx,$$

$$p^x(1-p)^y \frac{dp}{\sqrt{p(1-p)}} = p^{x-1/2}(1-p)^{y-1/2}\,dp,$$

a result inconsistent with that obtained previously. In fact, the distribution previously assumed for p was equivalent to assuming the special distribution for θ,

$$df = \frac{\cos\theta}{2}\,d\theta,$$

the arbitrariness of which is fully apparent when we use any variable other than p.

In a less obtrusive form the same species of arbitrary assumption underlies the method known as that of inverse probability. Thus, if the same observed result A might be the consequence of one or other of two hypothetical conditions X and Y, it is assumed that the probabilities of X and Y are in the same ratio as the probabilities of A occurring on the two assumptions, X is true, Y is true. This amounts to assuming that before A was observed, it was known that our universe had been selected at random from an infinite population in which X was true in one half, and Y true in the other half. Clearly such an assumption is entirely arbitrary, nor has any method been put forward by which such assumptions can be made even with consistent uniqueness. There is nothing to prevent an irrelevant distinction being drawn among the hypothetical conditions represented by X, so that we have to consider two hypothetical possibilities X_1 and X_2, on both of which A will occur with equal frequency. Such a distinction should make no difference whatever to our conclusions; but on the principle of inverse probability it does so, for if previously the relative probabilities were reckoned to be in the ratio x to y, they must now be reckoned $2x$ to y. Nor has any criterion been suggested by which it is possible to separate such irrelevant distinctions from those which are relevant.

There would be no need to emphasise the baseless character of the assumptions made under the titles of inverse probability and Bayes' Theorem in view of the decisive criticism to which they have been exposed at the hands of Boole, Venn, and Chrystal, were it not for the fact that the older writers, such as Laplace and Poisson, who accepted these assumptions, also laid the foundations of the modern theory of statistics, and have introduced into their discussions of this subject ideas of a similar character. I must indeed plead guilty in my original statement of the Method of the Maximum Likelihood (9) to having based my argument upon the principle of inverse probability; in the same paper, it is true, I emphasised the fact that such inverse probabilities were relative only. That is to say, that while we might speak of one value of p as having an inverse probability three times that of another value of p, we might on no account introduce the differential element dp, so as to be able to say that it was three times as probable that p should lie in one rather than the other of two equal elements. Upon consideration, therefore, I perceive that the word probability is wrongly used in such a connection: probability is a ratio of frequencies, and about the frequencies of such values we can know nothing whatever. We must return to the actual fact that one value of p, of the frequency of which we know nothing, would yield the observed result three times as frequently as would another value of p. If we need a word to characterise this relative property of different values of p, I suggest that we may speak without confusion of the *likelihood* of one value of p being thrice the likelihood of another, bearing always in mind that likelihood is not here used loosely as a synonym of probability, but simply to express the relative frequencies with which such values of the hypothetical quantity p would in fact yield the observed sample.

The solution of the problems of calculating from a sample the parameters of the hypothetical population, which we have put forward in the method of maximum likelihood, consists, then, simply of choosing such values of these parameters as have the maximum likelihood. Formally, therefore, it resembles the calculation of the mode of an inverse frequency distribution. This resemblance is quite superficial: if the scale of measurement of the hypothetical quantity be altered, the mode must change its position, and can be brought to have any value, by an appropriate change of scale; but the optimum, as the position of maximum likelihood may be called, is entirely unchanged by any such transformation. Likelihood also differs from probability* in that it is not a differential element, and is incapable of being integrated: it is assigned to a particular point of the range of variation, not to a particular element of it. There is therefore an absolute measure of probability in that the unit is chosen so as to make all the elementary probabilities add up to unity. There is no such absolute measure of likelihood. It may be convenient to assign the value unity to the maximum value, and to measure other likelihoods by comparison, but there will then be an infinite number of values whose likelihood is greater than one-half. The sum of the likelihoods of admissible values will always be infinite.

Our interpretation of Bayes' problem, then, is that the likelihood of any value of p is proportional to

$$p^x(1 - p)^y,$$

and is therefore a maximum when

$$p = \frac{x}{n},$$

* It should be remarked that likelihood, as above defined, is not only fundamentally distinct from mathematical probability, but also from the logical "probability" by which Mr. Keynes (21) has recently attempted to develop a method of treatment of uncertain inference, applicable to those cases where we lack the statistical information necessary for the application of mathematical probability. Although, in an important class of cases, the likelihood may be held to measure the degree of our rational belief in a conclusion, in the same sense as Mr. Keynes' "probability," yet since the latter quantity is constrained, somewhat arbitrarily, to obey the addition theorem of mathematical probability, the likelihood is a quantity which falls definitely outside its scope.

which is the best value obtainable from the sample; we shall term this the *optimum* value of p. Other values of p for which the likelihood is not much less cannot, however, be deemed unlikely values for the true value of p. We do not, and cannot, know, from the information supplied by a sample, anything about the probability that p should lie between any named values.

The reliance to be placed on such a result must depend upon the frequency distribution of x, in different samples from the same population. This is a perfectly objective statistical problem, of the kind we have called problems of distribution; it is, however, capable of an approximate solution, directly from the mathematical form of the likelihood.

When for large samples the distribution of any statistic, θ_1, tends to normality, we may write down the chance for a given value of the parameter θ, that θ_1 should lie in the range $d\theta_1$ in the form

$$\Phi = \frac{1}{\sigma\sqrt{2\pi}} e^{-(\theta_1-\theta)^2/2\sigma^2}\, d\theta_1.$$

The mean value of θ_1 will be the true value θ, and the standard deviation is σ, the sample being assumed sufficiently large for us to disregard the dependence of σ upon θ.

The likelihood of any value, θ, is proportional to

$$e^{-(\theta_1-\theta)^2/2\sigma^2},$$

this quantity having its maximum value, unity, when

$$\theta = \theta_1;$$

for

$$\frac{\partial}{\partial\theta} \log \Phi = \frac{\theta_1 - \theta}{\sigma^2}.$$

Differentiating now a second time

$$\frac{\partial^2}{\partial\theta^2} \log \Phi = -\frac{1}{\sigma^2}.$$

Now Φ stands for the total frequency of all samples for which the chosen statistic has the value θ_1, consequently $\Phi = S'(\phi)$, the summation being taken over all such samples, where ϕ stands for the probability of occurrence of a certain specified sample. For which we know that

$$\log \phi = \mathrm{C} + \mathrm{S}(\log f),$$

the summation being taken over the individual members of the sample.

If now we expand $\log f$ in the form

$$\log f(\theta) = \log f(\theta_1) + \overline{\theta - \theta_1} \frac{\partial}{\partial\theta} \log f(\theta_1) + \frac{\overline{\theta - \theta_1}^2}{\underline{|2}} \frac{\partial^2}{\partial\theta^2} \log f(\theta_1) + \cdots,$$

or

$$\log f = \log f_1 + a\overline{\theta - \theta_1} + \frac{b}{2}\overline{\theta - \theta_1}^2 + \cdots,$$

we have

$$\log \phi = \mathrm{C} + \overline{\theta - \theta_1}\,\mathrm{S}(a) + \tfrac{1}{2}\overline{\theta - \theta_1}^2\mathrm{S}(b) + \cdots;$$

now for optimum statistics

$$\mathrm{S}(a) = 0,$$

and for sufficiently large samples S(b) differs from $n\bar{b}$ only by a quantity of order $\sqrt{n\sigma_b}$; moreover, $\theta - \theta_1$ being of order $n^{-1/2}$, the only terms in $\log \phi$ which are not reduced without limit, as n is increased, are

$$\log \phi = \mathrm{C} + \tfrac{1}{2}n\bar{b}\,\overline{\theta - \theta_1}^2;$$

hence

$$\phi \propto e^{(1/2)n\bar{b}\overline{\theta - \theta_1}^2}.$$

Now this factor is constant for all samples which have the same value of θ_1, hence the variation of Φ with respect to θ is represented by the same factor, and consequently

$$\log \Phi = \mathrm{C}' + \tfrac{1}{2}n\bar{b}\,\overline{\theta - \theta_1}^2;$$

whence

$$-\frac{1}{\sigma_{\theta_1}^2} = \frac{\partial^2}{\partial\theta^2}\log \Phi = n\bar{b},$$

where

$$b = \frac{\partial^2}{\partial\theta^2}\log f(\theta_1),$$

θ_1 being the optimum value of θ.

The formula

$$-\frac{1}{\sigma_{\hat{\theta}}^2} = x\overline{\frac{\partial^2}{\partial\theta^2}\log f}$$

supplies the most direct way known to me of finding the probable errors of statistics It may be seen that the above proof applies only to statistics obtained by the method of maximum likelihood.*

* A similar method of obtaining the standard deviations and correlations of statistics derived from large samples was developed by Pearson and Filon in 1898 (16). It is unfortunate that in this memoir no sufficient distinction is drawn between the *population* and the *sample*, in consequence of which the formulae obtained indicate that the likelihood is always a maximum (for continuous distributions) when the *mean* of each variate in the sample is equated to the corre-

But the factorisation of f into factors involving $(\theta, \hat{\theta})$ and $(\hat{\theta}, \theta_1)$ respectively is merely a mathematical expression of the condition of sufficiency; and it appears that any statistic which fulfils the condition of sufficiency must be a solution obtained by the method of the optimum

It may be expected, therefore, that we shall be led to a sufficient solution of problems of estimation in general by the following procedure. Write down the formula for the probability of an observation falling in the range dx in the form

$$f(\theta, x)\,dx,$$

where θ is an unknown parameter. Then if

$$\mathrm{L} = \mathrm{S}(\log f),$$

the summation being extended over the observed sample, L differs by a constant only from the logarithm of the likelihood of any value of θ. The most likely value, $\hat{\theta}$, is found by the equation

$$\frac{\partial \mathrm{L}}{\partial \theta} = 0,$$

and the standard deviation of $\hat{\theta}$, by a second differentiation, from the formula

$$\frac{\partial^2 \mathrm{L}}{\partial \theta^2} = -\frac{1}{\sigma_{\hat{\theta}}^2};$$

this latter formula being applicable only where $\hat{\theta}$ is normally distributed, as is often the case with considerable accuracy in large samples. The value $\sigma_{\hat{\theta}}$ so found is in these cases the least possible value for the standard deviation of a statistic designed to estimate the same parameter; it may therefore be applied to calculate the efficiency of any other such statistic.

When several parameters are determined simultaneously, we must equate the second differentials of L, with respect to the parameters, to the coefficients of the quadratic terms in the index of the normal expression which represents the distribution of the corresponding statistics. Thus with two parameters,

$$\frac{\partial^2 \mathrm{L}}{\partial \theta_1^2} = -\frac{1}{1 - r_{\hat{\theta}_1\hat{\theta}_2}^2}\cdot\frac{1}{\sigma_{\hat{\theta}_1}^2}, \qquad \frac{\partial^2 \mathrm{L}}{\partial \theta_2^2} = -\frac{1}{1 - r_{\hat{\theta}_1\hat{\theta}_2}^2}\cdot\frac{1}{\sigma_{\hat{\theta}_2}^2},$$

$$\frac{\partial^2 \mathrm{L}}{\partial \theta_1 \partial \theta_2} = +\frac{1}{1 - r_{\hat{\theta}_1\hat{\theta}_2}^2}\cdot\frac{r}{\sigma_{\hat{\theta}_1}\sigma_{\hat{\theta}_2}},$$

or, in effect, $\sigma_{\hat{\theta}}^2$ is found by dividing the Hessian determinant of L, with respect to the parameters, into the corresponding minor.

The application of these methods to such a series of parameters as occur in the specification of frequency curves may best be made clear by an example....

12. Discontinuous Distributions

The applications hitherto made of the optimum statistics have been problems in which the data are ungrouped, or at least in which the grouping intervals are so small as not to disturb the values of the derived statistics. By grouping, these continuous distributions are reduced to discontinuous distributions, and in an exact discussion must be treated as such.

If p_s be the probability of an observation falling in the cell (s), p_s being a function of the required parameters $\theta_1, \theta_2 \ldots$; and in a sample of N, if n_s are found to fall into that cell, then

$$\mathrm{S}(\log f) = \mathrm{S}(n_s \log p_s).$$

If now we write $\bar{n}_s = p_s\mathrm{N}$, we may conveniently put

$$\mathrm{L} = \mathrm{S}\left(n_s \log \frac{n_s}{\bar{n}_s}\right),$$

where L differs by a constant only from the logarithm of the likelihood, with sign reversed, and therefore the method of the optimum will consist in finding the *minimum* value of L. The equations so found are of the form

$$\frac{\partial \mathrm{L}}{\partial \theta} = -\mathrm{S}\left(\frac{n_s}{\bar{n}_s}\frac{\partial \bar{n}_s}{\partial \theta}\right) = 0. \tag{6}$$

It is of interest to compare these formulae with those obtained by making the Pearsonian χ^2 a minimum.

For

$$\chi^2 = \mathrm{S}\frac{(n_s - \bar{n}_s)^2}{\bar{n}_s},$$

and therefore

$$1 + \chi^2 = \mathrm{S}\left(\frac{n_s^2}{\bar{n}_s}\right),$$

so that on differentiating by $d\theta$, the condition that χ^2 should be a minimum for variations of θ is

$$-\mathrm{S}\left(\frac{n_s^2}{\bar{n}_s^2}\frac{\partial \bar{n}_s}{\partial \theta}\right) = 0. \tag{7}$$

Equation (7) has actually been used (12) to "improve" the values obtained by the method of moments, even in cases of normal distribution, and the Poisson series, where the method of moments gives a strictly sufficient solution. The discrepancy between these two methods arises from the fact that χ^2 is itself an approximation, applicable only when $\bar{n}_s$ and n_s are large, and the difference between them of a lower order of magnitude. In such cases discontinuous (as is that of χ^2), but it is not impossible that mathematical research will reveal the existence of effective graduations for the most important groups of cases to which χ^2 cannot be applied.

We shall conclude with a few illustrations of important types of discontinuous distribution.

1. The Poisson Series

$$e^{-m}\left(1, m, \frac{m^2}{2!}, \ldots, \frac{m^x}{x!}, \ldots\right)$$

involves only the single parameter, and is of great importance in modern statistics. For the optimum value of m,

$$\mathrm{S}\left\{\frac{\partial}{\partial m}(-m + x\log m)\right\} = 0,$$

whence

$$\mathrm{S}\left(\frac{x}{\hat{m}} - 1\right) = 0,$$

or

$$\hat{m} = \bar{x}.$$

The most likely value of m is therefore found by taking the first moment of the series.

Differentiating a second time,

$$-\frac{1}{\sigma_{\hat{m}}^2} = \mathrm{S}\left(-\frac{x}{m^2}\right) = -\frac{n}{m},$$

so that

$$\sigma_{\hat{m}}^2 = \frac{m}{n},$$

as is well known.

2. Grouped Normal Data

In the case of the normal curve of distribution it is evident that the second moment is a sufficient statistic for estimating the standard deviation; in investigating a sufficient solution for grouped normal data, we are therefore in reality finding the optimum correction for grouping; the Sheppard correction having been proved only to satisfy the criterion of consistency.

gether of the wrong magnitude, and even in the wrong direction In order to obtain the optimum value of σ, we tabulate the values of $\frac{\partial L}{\partial \sigma}$ in the region under consideration; this may be done without great labour if values of σ be chosen suitable for the direct application of the table of the probability integral (13, Table II.). We then have the following values:

$\frac{1}{\sigma}$	0.43	0.44	0.45	0.46
$\frac{\partial L}{\partial \sigma}$	+15.135	+2.149	−11.098	−24.605
$\Delta^2 \frac{\partial L}{\partial \sigma}$		−0.261	−0.260	

By interpolation,

$$\frac{1}{\hat{\sigma}} = 0.441624$$

$$\hat{\sigma} = 2.26437.$$

We may therefore summarise these results as follows:—

Uncorrected estimate of σ	2.28254
Sheppard's correction	−0.01833
Correction for maximum likelihood	−0.01817
"Correction" for minimum χ^2	+0.07332

Far from shaking our faith, therefore, in the adequacy of Sheppard's correction, when small, for normal data, this example provides a striking instance of its effectiveness, while the approximate nature of the χ^2 test renders it unsuitable for improving a method which is already very accurate.

It will be useful before leaving the subject of grouped normal data to calculate the actual loss of efficiency caused by grouping, and the additional loss due to the small discrepancy between moments with Sheppard's correction and the optimum solution.

To calculate the loss of efficiency involved in the process of grouping normal data, let

$$v = \frac{1}{a} \int_{\xi-(1/2)a}^{\xi+(1/2)a} f(\xi)\, d\xi,$$

when a_σ is the group interval, then

$$\begin{aligned}\frac{\partial^2}{\partial\sigma^2}&\log v\\ &= \frac{1}{\sigma^2} - \frac{3\xi^2}{\sigma^2} + \frac{1}{\sigma^2}\left\{\frac{a^2}{12}(10\xi^2 - 3) - \frac{a^4}{360}(9\xi^4 + 21\xi^2 - 5)\right.\\ &\qquad + \frac{a^6}{30{,}240}(26\xi^6 + 110\xi^4 + 36\xi^2 - 7)\\ &\qquad \left. - \frac{a^8}{1{,}814{,}400}(51\xi^8 + 315\xi^6 + 351\xi^4 - 55\xi^2 + 9) + \cdots\right\},\end{aligned}$$

of which the mean value is

$$-\frac{2}{\sigma^2}\left\{1 - \frac{a^2}{6} + \frac{a^4}{40} - \frac{a^6}{270} + \frac{83a^8}{129{,}600}\cdots\right\},$$

neglecting the periodic terms; and consequently

$$\sigma_{\hat{\sigma}}^2 = \frac{\sigma^2}{2n}\left\{1 + \frac{a^2}{6} + \frac{a^4}{360} - \frac{a^8}{10{,}800}\cdots\right\}.$$

For ungrouped data

$$\sigma_{\hat{\sigma}}^2 = \frac{\sigma^2}{2n},$$

so that the loss of efficiency in scaling due to grouping is nearly $\frac{a^2}{6}$. This may be made as low as 1 per cent by keeping a less than $\frac{1}{4}$.

The further loss of efficiency produced by using the grouped second moment with Sheppard's correction is again very small, for

$$\sigma_{v_2}^2 = \frac{v_4 - v_2^2}{n} = \frac{2\sigma^4}{n}\left(1 + \frac{a^2}{6} + \frac{a^4}{360}\right)$$

neglecting the periodic terms.

Whence it appears that the further loss of efficiency is only

$$\frac{a^8}{10{,}800}.$$

We may conclude, therefore, that the high agreement between the optimum value of σ and that obtained by Sheppard's correction in the above example is characteristic of grouped normal data. The method of moments with Sheppard's correction is highly efficient in treating such material, the gain in efficiency obtainable by increasing the likelihood to its maximum value is trifling, and far less than can usually be gained by using finer groups. The loss of efficiency involved in grouping may be kept below 1 per cent. by making the group interval less than one-quarter of the standard deviation.

$$S = \frac{1}{n} \sum_{\alpha=1}^{n} (X_\alpha - \bar{x})(X_\alpha - \bar{x})'.$$

Hotelling's generalized T^2 is given as

$$T^2 = N\bar{x}'S^{-1}\bar{x}.$$

The analogy of T^2 to $t^2 = N\bar{x}^2/s^2$ is obvious in this vector notation. Hotelling actually carried out his exposition entirely in terms of the components of X_α, $\bar{x}$, S, etc., although now that seems very cumbersome.

The t-statistic, $t = \sqrt{N}\bar{x}/s$, is scale-invariant; that is, if each observation is multiplied by a positive constant c, the sample mean $\bar{x}$ and standard deviation s are both multiplied by c, but t is unaffected. For example, if X_α represents the length in feet of the αth object and $c = 12$ in. per foot, X_α^* is the length measured in inches. The t-statistic does not depend on the units of measurement. Hotelling points out that $T^2 = N\bar{x}'S^{-1}\bar{x}$ is invariant under linear transformations $X_\alpha^* = CX_\alpha$, where C is a nonsingular matrix because $\bar{x}$ is replaced by $\bar{x}^* = C\bar{x}$ and S is replaced by $S^* = CSC'$, leaving T^2 unchanged. In particular, T^2 does not depend on the scale for each component of X. In the univariate case, the real line has a positive and a negative direction, but in the multivariate case, no direction has special meaning. Hotelling's T^2 really corresponds to Student's t^2 since t^2 is invariant with respect to multiplication of X_α by any real constant different from 0.

The statistic T^2 is the only invariant function of the sufficient statistics, $\bar{x}$ and S; that is, any function of $\bar{x}$ and S that is invariant is a function of T^2. An important use of the T^2-statistic is to test the null hypothesis that $\mu = 0$ in $N(\mu, \Sigma)$; the hypothesis is rejected if T^2 is greater than a number preassigned to attain a desired significance level. This test is the uniformly most powerful invariant test. [See Sect. 5.3 of Anderson (1984), for example.] The only invariant of the parameters, μ and Σ, is $\mu'\Sigma^{-1}\mu$. The power of the T^2-test is, therefore, a function of this invariant of the parameters.

Hotelling approached the distribution of T^2 when $\mu = 0$ by a geometric method similar to that introduced by Fisher (1915) in finding the distribution of the Pearson correlation coefficient. Since T^2 is invariant with respect to nonsingular transformation $X_\alpha^* = CX_\alpha$, in the case of $\mu = 0$ the distribution of T^2 is the same as the distribution of $T^{*2} = N\bar{x}^{*\prime}(S^*)^{-1}\bar{x}^*$. If X_α has the distribution $N(0, \Sigma)$, then X_α^* has the distribution $N(0, C\Sigma C')$. Inasmuch as C can be chosen so that $C\Sigma C' = I$, Hotelling assumed $\Sigma = I$ when he derived the distribution of T^2 under the null hypothesis $\mu = 0$. In this case, the observed components $x_{i\alpha}$, $i = 1, \ldots, p$, $\alpha = 1, \ldots, N$, are independent and each has the standard normal distribution $N(0, 1)$; the vectors $\mathbf{x}_i = (x_{i1}, \ldots, x_{iN})'$, $i = 1, \ldots, p$, are independently distributed $N(0, I_N)$. The critical feature of $N(0, I_N)$ is that it is a spherical distribution.

Consider $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$ as spanning a p-dimensional linear space V_p and let $\zeta = \mathbf{e} = (1, \ldots, 1)'$ be an N-dimensional vector. Then the point in V_p closest to $\zeta = \mathbf{e}$ is (by usual least squares)

$$\hat{\mathbf{e}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e} = N\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\bar{x}.$$

The length of this vector is

$$\hat{\mathbf{e}}'\hat{\mathbf{e}} = N^2\bar{x}'(\mathbf{X}'\mathbf{X})^{-1}\bar{x}.$$

The squared distance of $\mathbf{e}$ from V_p is

$$(\mathbf{e} - \hat{\mathbf{e}})'(\mathbf{e} - \hat{\mathbf{e}}) = N - N^2\bar{x}'(\mathbf{X}'\mathbf{X})^{-1}\bar{x}.$$

The cotangent of the angle θ between $\mathbf{e}$ and V_p is given by

$$\cot^2\theta = \frac{\|\hat{\mathbf{e}}\|^2}{\|\mathbf{e} - \hat{\mathbf{e}}\|^2} = \frac{N\bar{x}'(\mathbf{X}'\mathbf{X})^{-1}\bar{x}}{1 - N\bar{x}'(\mathbf{X}'\mathbf{X})^{-1}\bar{x}} = \frac{N\bar{x}'(nS + N\bar{x}\bar{x}')^{-1}\bar{x}}{1 - N\bar{x}'(nS_N + \bar{x}\bar{x}')^{-1}\bar{x}}.$$

A little algebra shows that this is $T^2/n = N\bar{x}'(nS)^{-1}\bar{x}$.

The distribution of T^2 when $\mu = 0$ is generated by the vectors $\mathbf{x}_1, \dots, \mathbf{x}_p$; the direction of each of these is random in the sense that its projection on the unit sphere has a uniform distribution. It follows that the distribution of the angle between $\mathbf{e}$ and V_p is the same as the distribution of the angle between an arbitrary vector $\mathbf{y}$ and V_p. In fact, $\mathbf{y}$ can be assigned a distribution $N(0, I_N)$ independent of $\mathbf{x}_1, \dots, \mathbf{x}_p$. The cotangent squared of the angle θ_y between $\mathbf{y}$ and V_p is

$$\cot^2\theta_y = \frac{\mathbf{y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}}{\mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}}.$$

Since $\mathbf{y}$ has a spherical distribution, the distribution of $\cot^2\theta_y$ does not depend on $\mathbf{X}$, that is, V_p. Hotelling then found the distribution of $\cot^2\theta_y$ by representing the projection of $\mathbf{y}$ on the unit sphere in polar coordinates and performing integration. Of course, from normal regression theory, we know that the numerator of $\cot^2\theta_y$ has a χ^2-distribution with p degrees of freedom and the denominator has a χ^2-distribution with $N - p$ degrees of freedom independent of the numerator. Hence, $(N - p)\cot^2\theta_y/p$ has the F-distribution with p and $N - p$ degrees of freedom.

A form of the T^2-statistic can be used to test the hypothesis $\mu = \mu'$ on the basis of observations $X_1, \dots, X_{N_1}$ from $N(\mu, \Sigma)$ and $X'_1, \dots, X'_{N_2}$ from $N(\mu', \Sigma)$. Hotelling defined

$$\xi = \sqrt{\frac{N_1 N_2}{N_1 + N_2}}(\bar{x} - \bar{x}'),$$

which has the distribution $N(0, \Sigma)$ under the null hypothesis, and

$$nS = \sum_{\alpha=1}^{N_1}(X_\alpha - \bar{x})(X_\alpha - \bar{x})' + \sum_{\alpha=1}^{N_2}(X'_\alpha - \bar{x}')(X'_\alpha - \bar{x}')',$$

where $n = N_1 + N_2 - 2$. The matrix nS has the distribution of $\sum_{\alpha=1}^{N_1+N_2-2} Z_\alpha Z'_\alpha$, where the Z_α's are independent and Z_α has the distribution $N(0, \Sigma)$. Then arbitrary p, the author acknowledges the help of Fisher; the method was Fisher's geometric one. Later Wishart and Bartlett (1933) proved the result by characteristic functions.

Following "The generalization of Student's ratio" came many generalizations of univariate statistics. Wilks, who had spent the academic year 1931–32 with Hotelling, published "Certain generalizations in the analysis of variance" (1932). While Hotelling generalized Student's t, Wilks generalized the F-statistic basic to the analysis of variance; one generalization is the likelihood ratio criterion, often called "Wilks' lambda." If $X_\alpha^{(i)}$, $\alpha = 1, \ldots, N_i$, $i = 1, \ldots, q$, is the αth observation from the ith distribution $N(\mu_i, \Sigma)$, we define the "between" covariance matrix as

$$S_1 = \frac{1}{q-1} \sum_{i=1}^{q} [X_\alpha^{(i)} - \bar{X}^{(i)}][X_\alpha^{(i)} - \bar{X}^{(i)}]'$$

and the "within" covariance matrix as

$$S_2 = \frac{1}{\sum_{i=1}^{q} N_i - q} \sum_{i=1}^{q} \sum_{\alpha=1}^{N_i} [X_\alpha^{(i)} - \bar{X}^{(i)}][X_\alpha^{(i)} - \bar{X}^{(i)}]'.$$

In the scalar case ($p = 1$), the F-statistic is $F = S_1/S_2$. In the multivariate case, Wilks' lambda is $N/2$ times

$$\Lambda = \frac{|(\sum_{i=1}^{q} N_i - q)S_2|}{|(q-1)S_1 + (\sum_{i=1}^{q} N_i - q)S_2|},$$

which for $p = 1$ reduces to

$$\frac{1}{(q-1)F/(\sum_{i=1}^{q} N_i - q) + 1}.$$

Wilks found the moments of Λ and the distribution in some special cases. In a paper written the next year (when Wilks was in London and Cambridge), E.S. Pearson and Wilks (1933) treated a more general problem when $p = 2$, testing the homogeneity of covariance matrices as well as of means in the case of $N(\mu_i, \Sigma_i)$, $i = 1, \ldots, q$. Wilks continued his research in multivariate statistics; see S.S. Wilks, *Collected Papers* (1967).

Later Hotelling (1947, 1951) proposed a "generalized T test" for the analysis of variance. In the preceding notation, it was tr $S_1 S_2^{-1}$, known also as T_0^2, to test the hypothesis $\mu_1 = \cdots = u_q$. However, Lawley (1938) had already made this generalization.

Further study was done on the T^2-statistics. Hsu (1938) found the distribution of $T^2 = N\bar{x}'S\bar{x}$ when $\mu \neq 0$. This leads to the power function of a T^2-test. Simaika (1941) proved that of all tests of $H: \mu = 0$ with the power depending only on $N\mu'\Sigma^{-1}\mu$, the T^2-test is uniformily most powerful. Hsu (1945) proved an optimal property of the t^2-test that involves averaging the power over μ and Σ.

Bartlett, who was at Cambridge University with Wishart, developed much of the theory of the multivariate generalization. In his paper (1934), he developed further the geometric approach to multivariate analysis.

The study of properties of the T^2-test, alternatives to the test, and adaptations continue. A key paper was Stein (1956), in which it was shown that the T^2-test is admissible within the class of all tests of $H : \mu = 0$ (not just invariant tests). The proof depended on a more general theorem concerning exponential distributions and closed convex acceptance regions. An alternative proof of the admissibility of the T^2-test is to show that it is a proper Bayes procedure [Kiefer and Schwartz (1965)]. A different kind of test procedure of testing $H : \mu = 0$ is a step-down procedure. Marden and Perlman (1990) showed that a step-down procedure is admissible only if it is trivial, that is, has no step.

For references up to 1966, see Anderson, Das Gupta, and Styan (1972). About 125 papers are listed under the category "Tests of hypotheses about one or two mean vectors of multivariate normal distributions and Hotelling's T^2."

4. Comments

For a modern reader, this paper has the disadvantage of being written explicitly in terms of the components of constituent vectors and matrices. Linear operations and inverses do not seem as natural as in matrix notation. We are accustomed to defining A^{-1} the inverse to a nonsingular matrix A, as the (unique) matrix satisfying $A(A^{-1}) = I$. In keeping with usage at that time, Hotelling defined a as the determinant of (a_{ij}) and an element of the symmetric inverse as

$$A_{ij} = A_{ji} = \frac{\text{cofactor of } a_{ij} \text{ in } a}{a}.$$

The exposition of regression (pages 57–58) is particularly opaque. Let the observation matrix be $X = (x_{i\alpha})$, $i = 1, \ldots, p$, $\alpha = 1, \ldots, n$. Then $X = H + E$, where $\mathscr{E}E = 0$. The model can be written

$$\underset{p \times N}{H} = \underset{p \times q}{Z} \; \underset{q \times N}{G'},$$

where $Z = (\zeta_{is})$ is the matrix of parameters and $G = (g_{\alpha s})$ the matrix of independent variables. The matrix of regression coefficients Z is estimated by minimizing each diagonal element of

$$[X - ZG'][X - ZG']'$$

with respect to the elements of that row of Z.

Hotelling found the distribution of T^2 under the null hypothesis $\mu = 0$, but his approach can be developed to obtain the distribution when $\mu \neq 0$; the