Anthony W. F. Edwards, Anders Hald,
George A. Barnard

THREE CONTRIBUTIONS TO THE HISTORY OF STATISTICS

3 Preprint
March
1994

Institute of Mathematical Statistics
University of Copenhagen

Symposium on The History of Statistics

On the occasion of the 80th birthday of Professor Anders Hald a Symposium on the History of Statistics was held at the University of Copenhagen on June 3-4 1993.

The symposium was organised by The Institute of Mathematical Statistics, University of Copenhagen, where Anders Hald held a chair in statistics in the years 1960-1982.

Speakers at the conference were Anthony W. F. Edwards, University of Cambridge, Anders Hald, University of Copenhagen, Stephen Stigler, University of Chicago, and George A. Barnard, University of Essex.

Niels Keiding, University of Copenhagen and Anthony W. F. Edwards, University of Cambridge, served on the programme committee. Hans Brøns, Inge Henningsen and Karl Vind, all from University of Copenhagen, comprised the organising committee, and Søren Johansen, University of Copenhagen, has been the editor of this collection of preprints.

The Symposium received support from the following institutes at the University of Copenhagen: The Institute of Economics, The Laboratory of Actuarial Mathematics, The Institute of Statistics and The Statistical Research Unit as well as The Danish Society for Theoretical Statistics. Private funds covered travel expenses for the speakers.

Institute of Mathematical Statistics
University of Copenhagen

Symposium
on
The History of Statistics
3 - 4 June 1993

Programme for a Symposium on the History of Statistics to be held in auditorium 4 at the H. C. Ørsted Institute Universitetetsparken 5, 2100 Copenhagen Ø.

Thursday 3rd June

14.00-15.15 Anthony W. F. Edwards, University of Cambridge: What did Fisher mean by 'Inverse' Probability in 1912-1922?

15.15-15.45 Coffee.

15.45-17.00 Anders Hald, University of Copenhagen: The three Revolutions in Statistical Inference and their Importance for my Life as a Statistician.

19.00- Dinner.

Friday 4th June

10.00-11.15 Stephen M. Stigler, University of Chicago: Bernoulli and Likelihood.

11.15-11.45 Coffee.

11.45-13.00 George A. Barnard, University of Essex: Pivotal Models and the Fiducial Argument.

All interested are welcome to participate.

What did Fisher mean by 'inverse probability' in 1912-1922?

A.W.F.Edwards*

1 Introduction

In his 1992 paper *On the Mathematical Foundations of Theoretical Statistics* Fisher made a rather puzzling remark:

> I must indeed plead guilty in my original statement of the Method of Maximum Likelihood (1912) 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.

The remark is puzzling because in the 1912 paper Fisher is entirely clear that the entity he is maximising (not yet called the likelihood) "is a relative probability only, suitable to compare point with point, but incapable of being interpreted as a probability distribution over a region, or of giving any estimate of absolute probability". Moreover, contrary to his assertion in 1922, the 1912 paper does not contain any argument as such, but merely the magisterial statement (after dismissing least squares and the method of moments) "But we may solve the real problem directly", followed a few lines later by the assertion that "The most probable set of values for the [parameters] will make [the likelihood] a maximum". (In the original the corresponding mathematical symbols were employed rather than the words in square brackets, but it will sometimes be convenient in this account to use Fisher's later terminology anachronisticly.)

G.A. Barnard has suggested to me that Fisher's comments about relative probability in his 1912 paper might have been something of an afterthought; they are indeed confined to the sixth and last section of the paper. By contrast, the phrase "inverse probability system" is used in section 5 to describe the graph of the likelihood function for the mean m and dispersion parameter h of a Normal distribution (in modern notation, $h = \frac{1}{2}\sigma^2$). Fisher says that a Mr. T.L. Bennett, in a printed technical lecture, has integrated out m in order to derive a function

*Presented to Professor Anders Hald in honour of his 80th birthday

of h to maximise for variation in h alone. But "We shall see (in §6) that the integration with respect to m is illegitimate and has no definite meaning with respect to inverse probability", a comment which might have been added at the time he drafted the additional section 6. The interpretation is an attractive one, the more so because it explains the wording of the last sentence of all: "In conclusion I should like to acknowledge the great kindness of Mr. J.F.M. Stratton [*sic*; F.J.M. Stratton], to whose criticism and encouragement *the present form* of this note is due" [my italics].

The literal conclusion from Fisher's two uses of the phrase "inverse probability" in 1912 is that he meant by it what he later called the likelihood, because (1) it was analytically equal to the likelihood, an (2) it could not be integrated. But "the principle of inverse probability" is not mentioned.

I only know of two published comments on the 1922 remark. Nearly twenty years ago (Edwards, 1974a) I interpreted it as an admission by Fisher that he "was using the phrase "inverse probability" incorrectly", whilst in his recent introduction to the 1922 paper Geisser (1992) says that in 1912 "[Fisher] had taken a Bayesian approach because the maximising procedure resembled the calculation of the mode of a posterior probability". I believe both of these comments are wide off the mark, mine because it does not reflect what Fisher actually wrote, and Geisser's because it does not mention Fisher's clear understanding in 1912 that the probabilistic entity he was maximising was not an ordinary probability but a kind of "relative" one which did not obey the traditional law. The problem evidently needs looking at afresh, and in the present paper I shall try to examine exactly what Fisher meant by "inverse probability" in his youth. I should mention in passing that Zabell (1989), in his very informative paper "R.A. Fisher on the history of inverse probability", notes that 1922 remark, but simply takes it at its face value (R.L. Plackett, in his comment appended to Zabell's paper, says "Zabell passes quickly over Fisher's statement in 1912 that his absolute criterion, known later as the method of maximum likelihood, is derived from the principle of inverse probability").

There is a famous remark of Fisher's from 1936 about "the theory of inverse probability", that "I may myself say that I learned it at school as an integral part of the subject, and for some years saw no reason to question its validity". Alas, we do not know how long "some years" were, though another historical remark two years later (published as a note to Jeffreys, 1938) gives a clue:

> From a purely historical standpoint it is worth noting that the ideas and nomenclature for which I am responsible were developed only after I had innured myself to the absolute rejection of the postulate of Inverse Probability,..

2 Inverse Probability

Where does the phrase "inverse probability" come from, and has it always meant the same thing? It does not seem to have been used by Hume (1739), but Hume's contemporary, David Hartley, in his *Observations on Man* published in 1749, wrote

> An ingenious Friend has communicated to me a Solution of the Inverse Problem, in which he has shewn what the Expectation is, when an Event has happened p times, and failed q times, that the original Ratio of the Causes for the Happening or Failing of an Event should deviate in any given Degree from that of p to q.

This, of course, is earlier than Bayes (1764), a fact which prompted Stigler (1983) to suggest that someone other than Bayes had discovered his theorem. The need for such an explanation only arises if the "Solution of the inverse Problem" is the Bayesian solution, and in response to Stigler I argued (Edwards, 1986) that it was not, writing "I myself doubt that the passage refers to the Bayesian solution at all, believing it more likely to refer to one of the non-Bayesian attempts at a solution discussed first by James Bernoulli (1713) and then de Moivre (1738)". Dale (1988) agrees. When Thodhunter came to write about *Ars Conjectandi* in 1865 he too wrote of "the *inverse* use of James Bernoulli's theorem" [his italics], whereas in his chapter on Bayes he did not use the word inverse at all (contrary to a statement of mine in 1974a).

In 1809 Charles Hutton, George Shaw, and Richard Pearson published an abridgement of the Philosophical Transactions spanning the period which included Bayes's paper, about which he wrote:

> The problem is to this effect: "Having given the number of times an unknown event has happened and failed; to find the chance that the probability of its happening should lie somewhere between any two named degrees of probability". In its full extent and perfect mathematical solution, this problem is much too long and intricate, to be at all materially and practically useful, and such as to authorize the reprinting it here; especially as the solution of a kindred problem in de Moivre's Doctrine of Chances, *p.* 243, and the rules there given, may furnish a shorter way of solving the problem. See also the demonstration of these rules at the end of Mr. Simpson's treatise on "The Nature and Laws of Chance".

Although we do not here find the word "inverse", it is interesting that the authors regarded Bayes's solution and de Moivre's as alternatives.

It is important to note that in these early quotations the word "inverse" refers to the problem itself, and not necessarily to a particular solution of it, and that

when it does refer to a particular solution it might not be to Bayes's. Richard Price, we may here remark, called the problem "the converse problem" in his introduction to Bayes's *Essay.*

For a continuation of this enquiry into the origin of the phrase "inverse probability" Dale's *A History of Inverse Probability* (1991) is invaluable. Dale reports that it was the Danish statistician Arne Fisher who traced the first "modern" use of "*inverse*" to Augustus de Morgan's *Essay on Probabilities* of 1838. De Morgan employs the phrase "the inverse method" in his preface to describe what is required if one is to reason "from the happening of an event to the probability of one or other cause". A little later on he writes:

> De Moivre, nevertheless, did not discover the inverse method. This was first used by the Rev. T. Bayes, in *Phil. Trans.*liii.370.; and the author, though now almost forgotten, deserves the most honourable remembrance from all who treat the history of science.

Then, in Chapter I, de Morgan goes on to say:

> ...causes are likely or unlikely, just in the same proportion that it is likely or unlikely that observed events should follow from them. The most probable cause is that from which the observed event could most easily have arisen.

Now this statement is remarkable for its absence of any mention of equal prior probabilities for the causes. De Morgan similarly makes no mention of prior probabilities when, later on, he gives the rule for applying this method. Dale cites two examples from de Morgan's subsequent writing which also omit any appeal to equal prior probabilities:

> When an event has happened, & may have happened in 2 or 3 different ways, that way which is most likely to bring about the event is most likely to have been the cause, (1843)

and

> This inversion of circumstances, this conclusion that the circumstances under which the event did happen are most probably those which would have been most likely to bring about the event, is of the utmost evidence to our minds. (1847)

These last three quotations from de Morgan seem to suggest a novel principle, namely that when considering causes one might simply choose the one with the highest likelihood without computing posterior probabilities explicitly, and hence without having to consider prior probabilities. (The first of then, it is true, goes

further by suggesting a measure of "likely or unlikely" and not just a maximising principle.)

Such a principle was not entirely new, for de Moivre himself in the preface to the first edition of *The Doctrine of Chances* published in 1718 observed that in order to compare the hypotheses of "Chance" and "Design" "we may calculate what Probability there is, that ... Events should be rather owing to one than to the other", whilst Daniel Bernoulli (1778) had written: "Of all the innumerable ways of dealing with errors of observation one should choose the one which has the highest degree of probability for the complex of observations as a whole".

De Morgan was, of course, writing after the publication of Laplace's influential 1774 *Memoire sur la probabilite des causes par les evenements* in which he had introduced the following fundamental principle (independently of both Bayes and D. Bernoulli):

> If an event can be produced by a number n of different causes, the probabilities of these causes given the event are to each other as the probabilities of the event given the causes, and the probability of the existence of each of these is equal to the probability of the event given that cause, divided by the sum of all the probabilities of the event given each of these causes. (*Translation of Stigler, 1986.*)

Laplace did not mention the implicit equal prior probabilities any more than de Morgan did, but by the time of his *Essai philosophique sur les probabilites* (1814-1825; see Edwards, 1974a, and Dale, 1991) he was quite explicit about the need to take unequal prior probabilities into account. He seems, however, never to have written of *inverse probability* (though Stigler, 1986, entitles his translation "Laplace's 1774 memoir on inverse probability"). Dale notes that the heading *Methode inverse des probabilites* does occur in a Paris lecture summary from the turn of the century.

Jumping many years ahead it is interesting to compare Jeffrey's statement "if we must choose between two definitely stated alternatives we should naturally take the one that gives the larger likelihood" (1948), and that in turn brings to mind Ramsey's 1928 notes (Ramsey, 1931).

3 A digression on F.P. Ramsey

Ramsey had clearly been much impressed by Fisher's 1925 paper *Theory of statistical estimation*, read by title to the Cambridge Philosophical Society on 4 May. Although critical of the "stupid fiction" of an infinite population, he was enthusiastic about the Method of Maximum Likelihood, whose "significance is in suggesting a theory or set of chances. Proportion of infinite population should be replaced by chance". "Chances", wrote Ramsey, "are degrees of belief within a certain system of beliefs and degrees of belief". Then comes a set of notes of which two deserve quoting in full:

(13) in choosing a system [of chances] we have to compromise between two principles: subject always to the proviso that the system must not contradict any facts we know, we choose (other things being equal) the simplest system, and (other things being equal) we choose the system which gives the highest chance to the facts we have observed. This last is Fisher's "Principle of Maximum Likelihood", and gives the only method of verifying a system of chances.

(15) Statistical science must be briefly dealt with from our point of view; it has three parts:-

(a) Collection and arrangement of selections from the multitudinous data.

(b) Induction = forming a system of chances from the data by means of the Principle of Maximum Likelihood.

(c) Causal analysis; e.g. this die falls so often this way up, therefore its centre of gravity must be displaced towards the opposite face.

Then, in enlarging on (c) Causal analysis, Ramsey makes an odd use of the phrase "Principle of Indifference":

e.g. I conclude from the fact that more boys are born than girls to some superiority in the number, mobility or capacity for fertilization of male-bearing spermatoza or one of a thousand other possible causes, because of the Principle of Indifference, which is part of my fundamental system, the observed inequality would be so unlikely if there were no such difference.

He seems to be arguing that it is the Principle of Indifference which justifies the rejection of a hypothesis on the test-of-significance grounds that its acceptance would render the data exceedingly improbable. However, G. Shafer has suggested to me that perhaps Ramsey only meant that the Principle justified the hypothesis of "no such difference" and that his syntax is misleading.

It is interesting to note that the published version of Fisher's 1925 paper *Theory of statistical estimation* starts with a *Prefatory Note*: "It has been pointed out to me that some of the statistical ideas employed in the following investigation have never received a strictly logical definition and analysis", and then he attempts to justify the notion of an infinite hypothetical population. It seems altogether probable that Ramsey had commented adversely on this concept when Fisher's paper was read, and this was his reply. The minutes of the Cambridge Philosophical Society show that Fisher's was the fourth paper at the meeting on 4 May, and the first to be read by title. But presumably a copy was available at the meeting for comment then or, privately, later. There is no record of who was present.

Finally, Ramsey's phrase in note (15) "Statistical science" should not be allowed to pass unremarked. It may have been common in Cambridge discussions

at the time, for Fisher used it both in his 1922 paper and in his 1923 review of Keynes's *A Treatise on Probability* (Edwards, 1993).

4 Fisher

The above account suggests caution in interpreting what Fisher might have meant by "inverse probability" in 1922, and it is now time to return to his own writings. The story of his controversy with Karl Pearson over the confusion between maximum likelihood and maximum posterior probability has been told too recently to need repeating in detail (E.S. Pearson, 1968; Edwards 1974a), but we may scan it again for clues.

In his 1915 paper deriving the sampling distribution of the correlation coefficient Fisher also derived the maximum-likelihood estimate of the parameter. "I have given elsewhere (1912) a criterion, independent of scaling, suitable for obtaining the relation between an observed correlation of a sample and the most probable value of the correlation in the whole sample." After the derivation he added "It is now apparent that the most likely value of the correlation will in general be less than that observed, ..." Here "most probable" and "most likely" are clearly synonyms, but the word "inverse" is not used, and the criterion is "independent of scaling". Pearson and his collaborators were not clear about the distinction between maximum probability and Fisher's criterion, and their work (Soper *et. al.*, 1917) prompted Fisher to clarify his criterion by giving the word "likelihood" its technical meaning in 1921. Before then, however, there was another important interchange with Pearson.

The two letters, from the summer of 1916, are preserved in the Fisher archive in the Barr-Smith Library of the University of Adelaide, and have been published by E.S. Pearson (1968). In the first, Fisher offers Karl Pearson the draft of a note for *Biometrika* commenting unfavourably on the method of minimum chi-squared which had been advocated by Kirstine Smith, a Danish pupil of Thiele's then studying under Karl Pearson (E.S. Pearson, 1990). Fisher's note ends:

> There is nothing at all "arbitrary" in the use of the method of moments for the Normal curve; as I have shown elsewhere it flows directly from the absolute criterion ($\Sigma log f$ a maximum) derived from the Principle of Inverse Probability. There is, on the other hand, something exceedingly arbitrary in a criterion which depends entirely upon the manner in which the data happens to be grouped.

Thus in mid-1916 Fisher has already formulated essentially the same perplexing statement as in 1922 about his criterion having been derived from the Principle of Inverse Probability.

The second letter is Pearson's reply, in which not surprisingly in view of Fisher's statement he does not differentiate between Fisher's criterion and maximum probability, which he now refers to as "the Gaussian method". "If you

will write me a defence of the Gaussian method, I will certainly consider it for publication, but if I were to publish your note, it would have to be followed by another note saying that it missed the point,..."

In 1918 Fisher submitted what was presumably his "defence", but after an interval Pearson rejected it in a letter dated 21 Octover 1918 (E.S. Pearson, 1968). Unfortunately no copy of Fisher's paper seems to exist, but probably it had something in common with the last section of Fisher (1921) and section 12 of Fisher (1922).

It is to these two famous papers we must turn in order to find Fisher's clear detachment of the method of maximum likelihood from maximum posterior probability. The first of them had been rejected by Karl Pearson (his letter of 21 August 1920 is in E.S. Pearson, 1968) because he felt that "Under present printing and financial conditions, I am regretfully compelled to exclude all that I think erroneous on my own judgement, because I cannot afford controversy". Fisher was never to submit a paper to *Biometrika* again. Major Leonard Darwin, Fisher's mentor at this time, approached the Royal Statistical Society on his behalf to see if their *Journal* might be interested, but they could not help "because", as Dr. M. Greenwood informed him, "they have to cater for an audience many of whom could not understand it and they therefore have to limit the number of highly technical articles" (Box, 1978). But in Italy Corrado Gini was looking for material for his new journal *Metron*, and it was there that the paper finally appeared. Thus the original definition of *likelihood* is in *Metron* and not *Biometrika* or the *Journal of the Royal Statistical Society*, the two leading British journals of the day.

In the *Introduction* Fisher explains how Soper *et.al.* had incorrectly assumed that his criterion for estimation had been deduced from Bayes's theorem, and that in his opinion "two radically distinct concepts have been confused under the name of "probability" and only by sharply distinguishing these can we state accurately what information a sample does give us respecting the population from which it is drawn." In section 3 Fisher critizises Soper *et.al.* for having assumed that he had appealed to Bayes's theorem in 1915, and adds:

> As a matter of fact, as I pointed out in 1912 (Fisher, 1912) the optimum is obtained by a criterion which is absolutely independent of any assumption respecting the *a priori* probability of any particular value. It is therefore the correct value to use when we wish for the best value *for the given data*, unbiassed by any *a priori* presuppositions.

The paper is dated October 1920. Within nine months, on 25 June 1921, the Royal Society received the manuscript of the 1922 paper with its statement "I must indeed plead guilty in my original statement of the Method of Maximum Likelihood (1912) to having based my argument upon the principle of inverse probability". The only way to reconcile these contemporaneous views of Fisher

about his own undergraduate paper nine years earlier is to suppose that he saw some distinction between the assumption of a uniform prior distribution and the principle of inverse probability, which is just the distinction I detected in the writing of de Morgan.

The 1921 paper ends with the famous *Note on the confusion between Bayes ' Rule and my method of the evaluation of the optimum*, in which *likelihood* is formally defined and differentiated from probability. (It was this heading which emboldened me to introduce the words *evaluate* and *evaluation* in 1972; I still think they ought to be adopted.)

Again in the 1922 paper Fisher repeatedly stresses the difference between likelihood, as newly defined, and probability. The phrase "Method of Maximum Likelihood" occurs for the first time. He shows that he has read Bayes's paper with care, though he did not notice that Bayes had a cunning argument for adopting a uniform distribution for the binominal parameter (Molina, 1931; Edwards, 1974b, 1978). But he does not refer to Bayes's procedure as an example of the application of "inverse probability"; on the contrary, he says "In a less obtrusive form the same species of arbitrary assumption underlies the method known as that of inverse probability", which he expounds for the case of two hypotheses. He states that the method assumes that their post-data probabilities are in the same ratio as the ratio of the probabilities of the data on the two hypotheses, and he notes that this amounts to assuming that the two hypothese have been drawn at random from an infinite population in which each was true half the time. Then comes his admission that in 1912 he had based his Method of Maximum Likelihood on the principle of inverse probability. (The referees of the 1922 paper were A.S. Eddington, Plumian Professor of Astronomy, and G. Udny Yule, University Lecturer in Statistics, both in the University of Cambridge. Their reports are reproduced in Appendix 1 by courtesy of the Royal Society of London.)

There is not much further evidence to be gleaned from Fisher's subsequent writings, but in the Adelaide archives there is a manuscript precis and discussion of Karl Pearson's paper *On the systematic fitting of curves to observations and measurements* (Pearson, 1902). The handwriting is that of Mrs. Fisher, so presumably she was taking dictation; there are a couple of corrections in Fisher's hand. I reproduce the complete note in Appendix 2, with the permission of the University of Adelaide[1]. The second paragraph reads:

> It is noteworthy here, too, that throughout the paper no distinction is drawn between the fitting of frequency curves and that of regression lines. Only the latter had been traditionally treated by least squares. For the former the student might find in Gauss a discussion justifying what is known as the method of maximum likelihood as a general principle, and showing that in fitting a normal frequency

[1]In this version of the paper the two Appendices have been omitted. It is intended to publish them in due course. In the meantime, the author would be happy to supply copies on request.

> curve, this took the form of the method of moments, while in fitting regression lines in the important case of normal and equal variability in the arrays (i.e. of the observations) it took the form of the method of least squares. Gauss's views, though very influential, had not, however, in this matter gained general assent for he derived the method of maximum likelihood, erroneously, from the principle of inverse probability, confidence in which among mathematicians had been dwindling throughout the nineteenth century.

The word "erroneously" is perhaps not so placed as to convey quite the meaning Fisher intended, but the extract serves to confirm the knowledge of the work of Gauss and Pearson. Unfortunately it is undated, but it seems to record a study made for the 1922 paper, but after the naming of the method of maximum likelihood, which would put it into the first half of 1921. This note does not differentiate between the principle of inverse probability and the adoption of a uniform prior distribution, which is how Gauss actually argued, yet Fisher made the distinction in the 1922 paper as I have indicated.

Finally, when Fisher introduced his notion of fiducial probability in 1930 he called the paper "Inverse probability", and it is indeed mostly a criticism of the Bayesian position along lines by now quite familiar, though he does add the new point that if we assume a uniform prior for a parameter, to choose its value by maximising the posterior probability (as he again notes Gauss did) is very odd, for "had the inverse probability distribution any objective reality at all we should certainly, at least for a single parameter, have preferred to take the mean or the median value".

5 Conclusion

We should never look for complete consistency in the writings of any author, especially a young one advancing the frontiers of a subject at a furious rate in the face of uncomprehending elders, but my impression now is in fact that in the decade 1912-1922 Fisher did indeed draw a distinction between inverse probability and fully-blown Bayesian inference which, though "of the same species", starts from a slightly different viewpoint. Bayesian inference delivers a probability distribution for an unknown parameter, which Fisher explicitly and forcefully rejected from the start, whilst the Principle of Inverse Probability only allows (on the present interpretation of his view) the comparison of parameter values "point with point" (1912). After all, if we cut away the historical use of the phrase, "inverse probability" is rather a good term for "likelihood", so long as we understand that it is not an ordinary probability. C.A.B. Smith once observed that "Fisher's choice of the word "likelihood" might have been a little unfortunate, in that in ordinary language the words "likelihood" and "probability" are virtually synonymous" (Smith, 1986). I am inclined to agree with him; some kind of connotation of "support" would have been better (see Edwards, 1972).

To sum up, I believe that in 1912 by *inverse probability* Fisher meant *likelihood*; that in 1916 by the *principle of inverse probability* he meant the Laplace — de Morgan principle which he thought conferred legitimacy on the method of maximum likelihood; and that only as late as 1921-1922 did he fully appreciate that this principle was inescapably Bayesian and had to be rejected.

6 Acknowledgements

This investigation was stimulated by a visit to the University of Adelaide to study the papers of R.A. Fisher. I am grateful to the Royal Society of London for a study grant for the visit, and to the University of Copenhagen for the opportunity to present this paper at the symposium in honour of Professor Anders Hald.

7 References

Bayes, T. (1764): "An Essay Towards Solving a Problem in the Doctrine of Chances." *Phil. Trans. Roy. Society* 53, 370-418. Reprinted in: Studies in the History of Probability and Statistics IX, Thomas Bayes' essay towards solving a problem in the doctrine of chances (with a bibliographical note by G.A. Barnard), *Biometrika* 45, 293-315 (1958), and in Pearson and Kendall (1970).

Bernoulli, D. (1778): *Acta Acad. Petrop.* for 1977, 3-33. Translated as: The most probable choice between several discrepant observations and the formation therefrom of the most likely induction. *Biometrika* 48, 3-13 (1961); reprinted in Pearson and Kendall (1970).

Bernoulli, J. (1713): *Ars Conjectandi*, Basilea: Thurnisius. Fascimile reprint: Bruxelles: *Culture et Civilisation* (1968). Reprinted in *Die Werke von Jakob Bernoulli*, III, 107-286, Birkhäuser, Basel (1975).

Box, J.F. (1978): *R.A. Fisher: The Life of a Scientist* , Wiley, New York.

Dale, A.I. (1988): "On Bayes' Theorem and the Inverse Bernoulli Theorem". *Historica Mathematica* 15, 348-360.

Dale, A.I. (1991): *A History of Inverse Probability*, Springer, New York.

de Moivre, A. (1718): *The Doctrine of Chances*, Pearson, London.

de Moivre, A. (1738): *The Doctrine of Chances* (2nd edition), Woodfall, London. Reprinted (1967) by Cass, London.

de Morgan, A. (1738): *An Essay on Probabilities and their Application to Life Contingencies and Insurance Offices*, Longman, Orme, Brown, Green & Longmans, London.

Edwards, A.W.F. (1972): *Likelihood*, Cambridge University Press. Reprinted in Edwards (1992).

Edwards, A.W.F. (1974a): "The History of Likelihood". *Internat. Statist. Rev.* 42, 9-15. Reprinted in Edwards (1992).

Edwards, A.W.F. (1974b): "A Problem in the Doctrine of Chances". in Barndorff-Nielsen, O. and G. Schou, University of Aarhus (eds.): *Proceedings of the Conference on Foundational Questions in Statistical Inference*, 43-60. Reprinted

in Edwards (1992).

Edwards, A.W.F. (1978): "Commentary on the Arguments of Thomas Bayes". *Scand. J. Statist.* 5, 116-118.

Edwards, A.W.F. (1986): "Is the Reference in Hartley (1749) to Bayesian Inference?" *The American Statistician* 40, 109-110.

Edwards, A.W.F. (1992): *Likelihood* (expanded edition), Johns Hopkins University Press, Baltimore.

Edwards, A.W.F. (1993): *Cambridge University Reporter* 123, 697.

Fisher, R.A. (1912): "On an Absolute Criterion for Fitting Frequency Curves". *Messenger Math.* 41, 155-160.

Fisher, R.A. (1915): "Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Indefinitely Large Population". *Biometrika* 9, 507-521.

Fisher, R.A. (1921): "On the "Probable Error" of a Coefficient of Correlation Deduced from a Small Sample". *Metron* 1, 3-32.

Fisher, R.A. (1922): "On the Mathematical Foundations of Theoretical Statistics". *Phil. Trans. Roy. Soc.* A 222, 309-368.

Fisher, R.A. (1925): "Theory of Statistical Estimation". *Proc. Camb. Phil. Soc.* 22, 700-725.

Fisher, R.A. (1930): "Inverse Probability". *Proc. Camb. Phil. Soc.* 26, 528-535.

Fisher, R.A. (1936): "Uncertain Inference". *Proc. Amer. Acad. of Arts Sci.* 71, 245-258.

Fisher, R.A. (1938) in Jeffreys (1938).

Geisser, S. (1991): "Introduction to Fisher (1922) On the mathematical foundation of theoretical statistics". In Kotz, S. and N.L. Johnson (eds.): *Breakthroughs in Statistics, Volume 1. Foundations and Basic Theory*, 1-10. Springer, New York.

Hartley, D. (1749): *Observations of Man, His Frame, His Duty , And His Expectations.* Richardson, London. Reprinted (1966) by Scholar's Fascimiles and Reprints, Gainesville, Florida.

Hume, D. (1739): *A Treatise of Human Nature.* Selby-Bigge, L.A. (ed.) (1888 and later printings), Clarendon Press, Oxford.

Hutton, C., Shaw, G. and R. Pearson (1809): *The Philosophical Transactions of the Royal Society of London, from their Commencement, in 1665, to the year 1800, Abridged, with notes and Biographic Illustrations.* Vol. XII from 1763 to 1769. Baldwin, London.

Jeffreys, H. (1938): "Maximum Likelihood, Inverse Probability, and the Method of Moments". *Ann. Eugen.* 8, 146-151.

Jeffreys, H. (1948): *Theory of Probability*, (2nd. edition), Clarendon Press, Oxford.

Keynes, J.M. (1921): *A Treatise on Probability.* Macmillan, London.

Laplace, P.S. de (1774): see Stigler (1986).

Molina, E.C. (1931): "Bayes' Theorem: An Expository Presentation". *Ann . Math. Statist.* 2, 23-37.

Pearson, E.S. (1968): "Some Early Correspondence between W.S. Gosset, R.A. Fisher and Karl Pearson, with notes and comments." *Biometrika* 55, 445-457. Reprinted in Pearson and Kendall (1970).

Pearson, E.S. (1990): *"Student": A Statistical Biography of William Sealy Gosset.* Plackett, R.L. (ed.) with the assistance of G.A. Barnard. Clarendon Press, Oxford.

Pearson, E.S. and M.G. Kendall (1970): *Studies in the History of Statistics and Probability.* Griffin, London.

Pearson, K. (1902): "On the Systematic Fitting of Curves to Observations and Measurements". *Biometrika* 1, 265-303.

Ramsey, F.P. (1931): *The Foundations of Mathematics and other Logical Essays.* Braithwaite (ed.). Routledge and Kegan Paul, London.

Smith, C.A.B. (1986): "The Development of Human Linkage Analysis". *Ann . Hum. Genet.* 50, 293-311.

Soper, H.E., Young, A.W., Cave, B.M., Lee, A. and K. Pearson (1917): "On the Distribution of the Correlation Coefficient in Small Samples. Appendix 2 to the papers of Student and R.A. Fisher. A Comparative Study". *Biometrika* 11, 328-413.

Stigler, S.M. (1983): "Who Discovered Bayes's Theorem?" *American Statistician* 37, 290-296.

Stigler, S.M. (1986): "Laplace's 1774 Memoir on Inverse Probability. *Statistical Science* 1, 359-378.

Todhunter, I. (1865): *A History of the Mathematical Theory of Probability.* Macmillan, Cambridge and London. Reprinted (1965) by Chelsea Publishing, New York.

Zabell, S. (1989): "R.A. Fisher on the History of Inverse Probability". *Statistical Science* 4, 247-263.

The Three Revolutions in Statistical Inference and Their Importance for my Life as a Statistician

Anders Hald

Abstract

A sketch is given of important features of the three revolutions in parametric statistical inference due to Laplace, Gauss, and Fisher. Their importance for my life as a statistician, particularly for my research and teaching, is indicated.

1 The Three Revolutions in Statistical Inference

1.1 Introduction

The three revolutions in parametric statistical inference are due to Laplace (1774), Gauss and Laplace (1809-1811) and Fisher (1922). It took about twenty years and several papers for each of these authors to work out the basic ideas in detail, and it took about a hundred years for the rest of the statistical community to understand and develop the new methods and their applications.

Let $p(x|\theta)$ be a statistical model with a given sample space and parameter space, and let $\underline{x}$ denote a sample of n independent observations. From the model we can find the sampling distribution of the statistic $t = t(\underline{x})$, and from $p(t|\theta)$ we can find probability limits for t for any given value of θ. This is a problem in direct probability, as it was called in the previous century.

In inverse probability the problem is to find probability limits for θ for a given value of $\underline{x}$. Bayes (1764) was the first to realize that a solution is possible only if θ itself is a random variable with a probability density $p(\theta)$. We can then find the conditional distributions $p(\theta|\underline{x})$ and $p(\theta|t)$, which can be used to find probability limits for θ for any given value of $\underline{x}$. Independently of Bayes, Laplace (1774) gave the first general theory of statistical inference based on inverse probability.

1.2 Laplace on Direct Probability 1776-1799

However, at the same time Laplace also developed methods of statistical inference based on direct probability. At the time the problems in applied statistics

were mainly from demography, about rates of mortality and the frequency of male births, and from the natural sciences, where the distribution of errors and relations between variables were studied. It was therefore natural for Laplace to create a theory of testing and estimation comprising relative frequencies, the arithmetic mean, and the linear model.

The error distributions discussed at the time were symmetric with known scale parameter, the most important being the rectangular, triangular, quadratic, cosine, semi-circular, and the double exponential. The normal distribution was not yet invented. These densities were chosen for their mathematical simplicity, nobody compared them with data.

The arithmetic mean was ordinarily used as estimate of the location parameter in the error distribution. Laplace solved the problem of finding the distribution of the mean by means of the convolution formula. However, this was only a solution in principle because all the known error distributions, apart from the rectangular, led to unmanageable distributions of the mean. He also gave the first test of significance of a mean based on the probability of a deviation from the expected value as large or larger than the observed, assuming that the observation are rectangularly distributed.

Let

$$y = X\beta + \epsilon = Xb + e$$

denote the linear model. Three methods of fitting this equation to data without specification of the error distribution apart from symmetry were developed. The method of averages by Mayer and Laplace requiring that $\Sigma w_{ik} e_i = 0$, where as many equations are used as the number of parameters. The method of least absolute deviations by Boscovich and Laplace, where $\Sigma w_i e_i = 0$ and $\Sigma w_i \left| e_i \right|$ is minimized for the two-parameter model. The method of minimizing the largest absolute deviation by Laplace, that is, minimax $\Sigma \left| e_i \right|$. He evaluated the results of such analyses by studying the distribution of the residuals.

1.3 The first Revolution: Laplace 1774-1786

Turning to inverse probability let us first consider two values of the parameter and the corresponding direct probabilities. Laplace's principle says, that if $\underline{x}$ is more probable under θ_2 than under θ_1 and $\underline{x}$ has been observed, then the probability of θ_2 being the underlying value of θ (the cause of $\underline{x}$) is larger than the probability of θ_1. Specifically, Laplace's principle of inverse probability says that

$$\frac{p(\theta_2|\underline{x})}{p(\theta_1|\underline{x})} = \frac{p(\underline{x}|\theta_2)}{p(\underline{x}|\theta_1)} \text{ for all } (\theta_1, \theta_2),$$

or

$$p(\theta|\underline{x}) \propto p(\underline{x}|\theta),$$

that is, inverse probability is proportional to direct probability. In the first instance Laplace formulated the principle intuitively, later he proved it under the supposition that the prior density is uniform on the parameter space.

Introducing the likelihood function Laplace's principle may be related to Fisher's through the diagram

$$\underset{\text{Inverse Probabiliy}}{p(\theta|\underline{x})} \quad \propto \quad \underset{\text{Direct Probability}}{p(\underline{x}|\theta)} \quad \propto \quad \underset{\text{Likelihood}}{L_{\underline{x}}(\theta)}$$

$$\text{Laplace} \qquad | \qquad \text{Fisher}$$

The history of statistical inference is about $p(\underline{x}|\theta)$ and its two interpretations, or in modern terminology about sampling distributions, posterior distributions, and the likelihood function. It is obvious that the mathematical parts of the three topics are closely related, and that a new result in any of the three fields has repercussions in the other two. There has been, and still is, a fruitful interaction between the three fields, despite the harsh language often used.

Based on Laplace's principle it is a matter of mathematical technique to develop a theory of testing, estimation, and prediction, given the model and the observations. Laplace did so between 1774 and 1786.

To implement the theory for large samples Laplace developed approximations by means of asymptotic expansions of integrals, both for tail probabilities and for probability integrals over an interval containing the mode. Using the Taylor expansion about the mode $\widehat{\theta}$ he found

$$\begin{aligned} \ln p(\theta|\underline{x}) &= const. + \ell(\theta) \quad \left[\ell(\theta) = \ln L_{\underline{x}}(\theta)\right] \\ &= const. + \ell\left(\widehat{\theta}\right) + \tfrac{1}{2}(\theta - \widehat{\theta})^2 \ell''(\widehat{\theta}) + ..., \end{aligned}$$

which shows that θ is asymptotically normal with mean $\widehat{\theta}$ and variance $\left[-\ell''(\widehat{\theta})\right]^{-1}$.

In this way Laplace proved for the binomial distribution that the most probable value of θ equals the observed relative frequency h and that θ is asymptotically normal with mean h and variance $h(1-h)/n$. Moreover, to test the significance of the difference $h_1 - h_2$ between two relative frequencies, he showed that $\theta_1 - \theta_2$ is asymptotically normal with mean $h_1 - h_2$ and variance given by $h_1(1-h_1)/n_1 + h_2(1-h_2)/n_2$, which led him to the large-sample test of significance used today.

There is, however, an inconsistency in Laplace's theory of estimation. For the binomial and the multinomial distributions he uses the most probable value as estimate, but in the measurement error model he introduces a new criterion to estimate the location parameter, namely to minimize the posterior expected loss, using the absolute deviation as loss function. He proves that this leads to the median in the posterior distribution as estimator. His justification for this procedure is that the absolute deviation is the natural measure of the goodness

of the estimate and that it corresponds to the gambler's expected loss in a game of chance.

The introduction of a loss function proved to be a serious mistake, which came to hamper the development of an objective theory of statistical inference to the present day. It is of course the beginning of the split between inference and decision theory.

To try out the new method Laplace chose the simplest possible error distribution with infinite support, the double exponential distribution. For three observations he found that the estimate is a root of a polynomial equation of the 15th degree. It must have been a great disappointment for him that the combination of the simplest possible error distribution and the simplest possible loss function led to an unmanageable solution, even for three observations.

In 1799, at the end of the first revolution, one important problem was still unsolved: the problem of the arithmetic mean. Applying all the known methods of estimation to all the known error distributions led to estimates of the location parameter different from the mean. Nevertheless, in practice everybody used the mean.

1.4 The second Revolution: Gauss and Laplace 1809-1828

The second revolution began in 1809-1810 with the solution of this problem, which gave us two of the most important tools in statistics, the normal distribution as a distribution of observations, and the normal distribution as an approximation to the distribution of the mean in large samples.

In 1809 Gauss asked the question: Does there exist an error distribution leading to the mean as estimate of the location parameter according to the principle of inverse probability? Gauss did not make the mistake of Laplace of introducing a loss function, instead he used the most probable value of the parameter as estimate. Setting the posterior mode equal to the arithmetic mean of the observations he got a functional equation with the normal distribution as solution. The normal distribution thus emerged as a mathematical construct, and Gauss did not compare the new error distribution with observations.

Assuming that the observations are normally distributed he found that the most probable value of the location parameter is obtained by minimizing the exponent $\sum(y_i - \theta)^2$, which naturally leads to the mean. If θ is a linear function of m parameters, $\theta = X\beta$, the estimates are found by minimizing the sum of the squared errors $(Y - X\beta)'(Y - X\beta)$. Assuming the variance of the y's to be known, Gauss solved the estimation problems for the linear-normal model and derived the multivariate normal distribution of the parameters.

Before having seen Gauss's book, Laplace in 1810 published a paper in which he derived the first version of the central limit theorem, which says that regardless of the shape of the error distribution, if only the variance is finite, the mean will

be approximately normally distributed in large samples.

As his immediate reaction to Gauss's results Laplace made two remarks in 1810:

(1) If the error distribution is normal, then the posterior distribution is normal and the posterior mean and median are equal. Hence, the method of least squares follows from my method of estimation as a special case.

(2) If the error distribution has finite variance, but is otherwise unknown, then the central limit theorem gives a large-sample justification for the method.

Hence, in the first instance, both Gauss and Laplace used inverse probability in their derivations of the method of least squares.

But already in 1811 Laplace gave an alternative derivation based on direct probability using the asymptotic normality of a linear combination of observations and minimizing the expected absolute error, which for the normal distribution is proportional to the expected squared error.

Between 1823 and 1828 Gauss supplemented Laplace's large-sample frequentist theory by a small-sample theory. Like Laplace he replaced the assumption of normality with the weaker assumption of finite variance, but in contradistinction to Laplace he used squared error as loss function because of its greater mathematical simplicity. He then developed the theory of linear, unbiased, minimum variance estimation for the linear model in the form known today.

Hence, they both gave up the normality assumption as too restrictive.

Gauss's two proofs both became popular and existed beside each other in spite of their contradictory assumptions. One reason for this may be the following argument due to Laplace.

In 1812 Laplace made an important observation on the equivalence of direct and inverse probability for finding large-sample limits for the binomial parameter. Direct probability leads to the limits for the relative frequency h of the form

$$h \sim \theta \pm \sqrt{\theta(1-\theta)/n},$$

disregarding terms of the order of $1/n$. But for this order of approximation the limits may also be written as

$$h \sim \theta \pm \sqrt{h(1-h)/n},$$

which solved for θ gives

$$\theta \sim h \pm \sqrt{h(1-h)/n}.$$

However, these limits are the same as those following from inverse probability.

Generalizing this argument, the probability limits for the estimate t becomes

$$t \sim \theta \pm \sigma/\sqrt{n},$$

and for the estimate s

$$s \sim \sigma \pm \kappa/\sqrt{n}.$$

Combining these relations we get

$$t \sim \theta \pm s/\sqrt{n}$$

which leads to the limits for θ,

$$\theta \sim t \pm s/\sqrt{n}.$$

This kind of reasoning explains why the methods of direct and inverse probability could coexist in statistical practice without serious conflict for about a hundred years.

For large samples the normal distribution could be used to find probability or confidence limits. For moderately large samples the so-called 3σ-limits became popular, possibly with an appeal to a result by Gauss, which states that for a unimodal symmetric error distribution with finite variance at least 95% of the probability lies within the 3σ-limits. Use of the 3σ-limits became a standard procedure in estimation and testing as a safeguard against deviations from normality.

During the following period the application of statistical methods was extended to the social and biological sciences in which variation among individuals, instead of errors, was studied by means of skew frequency curves, and the measurement error model was replaced by linear regression and correlation.

Two systems of frequency curves were developed: Pearson's system of skew frequency curves, and Kapteyn's system of transformations to obtain normality.

Correspondingly, a new method of estimation was developed which may be called the analogy-method. Pearson equated the empirical moments to the theoretical moments and thus got as many non-linear equations as parameters to be estimated. Kapteyn equated the empirical and theoretical fractiles.

1.5 The Third Revolution: R.A.Fisher 1912-1962

At the beginning of the present century the theory of statistical inference thus consisted of a large number of ad hoc methods, some of them contradictory, and the small-sample theory was only in a rudimentary state. Some important questions were as follows:

How to choose between direct and inverse probability methods?

How to choose between various loss functions?

How to choose between various statistics for use in the analogy-method?

How to find probability limits for the parameters from direct probability methods?

These problems were attacked and most of them solved by Fisher between 1922 and 1936.

He turned the estimation problem upside down by beginning with requirements to estimators. He formulated the criteria of consistency, efficiency, and sufficiency, the last concept being new.

Having thus defined the properties of good estimators he turned to a criticism of the existing methods of estimation.

He showed that the inverse probability estimate depends on the parameterization of the model, which means that the resulting estimate is arbitrary. For a time this argument led to less interest in inverse probability methods.

He rejected the use of loss functions as extraneous to statistical inference.

Turning to analogy-methods he showed that the method of moments in general is inefficient.

Given the model and the observations, he noted that all information on the parameters is contained in the likelihood function, and he proved the asymptotic optimality of the estimates derived from this function, the maximum likelihood estimates. Basing his inference theory on the likelihood function he avoided the arbitrariness introduced by Laplace and Gauss due to loss functions and the assumption of finite variance.

Assuming normality, he derived the t, χ^2, and F distributions, and showed how to use them in testing and interval estimation, thus solving the small-sample problems for the linear-normal model.

He also derived the distribution of the correlation coefficient and the partial correlation coefficients in normal samples.

He initiated the theory of ancillary statistics and conditional inference.

Large-sample probability limits for a parameter were found by what today is called a pivotal statistic. By an ingenious use of the pivotal argument, Fisher derived what he called fiducial limits for a parameter, for example by means of the t distribution.

Blinded by his many successes, he even tried the impossible, namely to find a general theory for probability statements about parameters without considering the parameter as random. To this end he defined the fiducial probability distribution of the parameters.

Fisher explained the new statistical ideas and techniques in an aggressive and persuasive language, which lead to acceptance of his theories within a rather short period of time, not alone among mathematical statisticians, but also among research workers in general.

A large part of mathematical statistics since 1922 has consisted in an elaboration of Fisher's ideas, both in theory and practice.

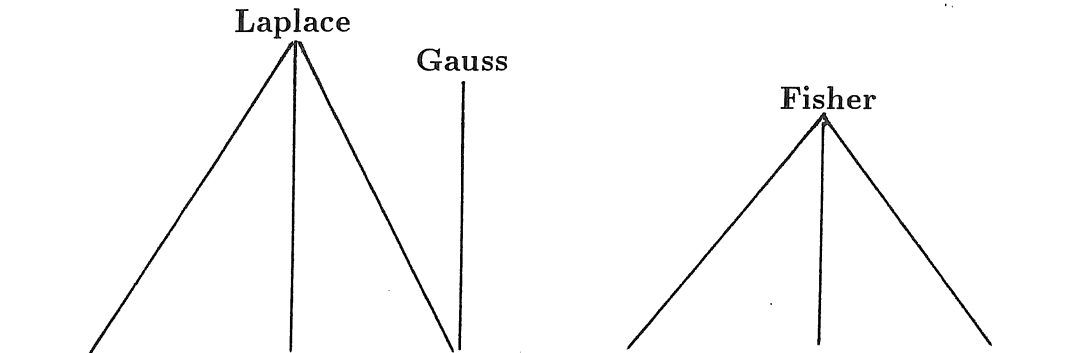
Because of the fundamental relation between the posterior density and the likelihood function many of Fisher's asymptotic results are identical to those of Laplace from a mathematical point of view, only a new interpretation is required. Fisher never acknowledged his debt to Laplace.

The following panorama (handout) indicates how the ideas of Laplace, Gauss, and Fisher, have influenced statistical theory today. As you will see, I have added the names of some prominent statisticians and left some space in which you may add the names of your favourites. The partitions are not watertight, many persons have worked in several fields.

Finally you may add your own name in the appropriate places and contemplate how much you owe to the three giants who founded statistical science.

The three revolutions in statistical inference

Inverse Probability		**Direct Probability**		**Likelihood**
$p(\theta\|\underline{x})$	$\propto$	$p(\underline{x}\|\theta)$	$\propto$	$L_{\underline{x}}(\theta)$



DeFinetti	Jeffreys	Neyman	Bartlett	Barnard	Edwards
Savage	Perks	Wald	Rao		
Lindley		Lehman	Cox		
Box			Barndorff		

2 My Life as a Statistician

2.1 Studying Statistics 1935-1939

Turning to the second part of my lecture I shall begin by mentioning the persons who founded teaching and research of statistics in the University of Copenhagen, see Table 1.

Table 1
Some Teachers of Statistics at the University of Copenhagen

Faculty of Natural Sciences	Faculty of Economics
T.N. Thiele, 1875-1907	H. Westergaard, 1883-1924
J.F. Steffensen, 1919-1943	H. Cl. Nybølle, 1926-1947
W. Simonsen, 1943-1974	A. Hald, 1948-1960
A. Hald, 1960-1982	G. Rasch, 1962-1972

The two columns list the persons working with theoretical statistics in the Faculties of Natural Sciences and Economics, respectively.

Besides the persons listed in Table 1 several others taught statistics, the most important being G. Rasch, who worked as external lecturer in biological and mathematical statistics 1937-1962.

In 1935 only Nybølle was full-time occupied with statistics. He had no assistant. Steffensen had his time divided between actuarial mathematics, numerical analysis, and mathematical statistics. Hence, there was only 1 and 1/3 person engaged in teaching and research in statistics at the higher institutions of learning in Denmark. I guess that today there must be about a hundred.

After having studied mathematics, numerical analysis, economics, and law for $2\frac{1}{2}$ years I began to study actuarial mathematics and statistics in 1935. My teachers were Steffensen and Nybølle. The curriculum consisted of the three books shown in Table 2.

Table 2
Curriculum in Statistics, 1935

H. Westergaard og H.Cl. Nybølle: *Statistikkens Teori i Grundrids.*
G.E.C. Gads Forlag. København 1927
Steffensen J.F.: *Matematisk Iagttagelseslære.*
G.E.C. Gads Forlag. København 1923
Bowley, A.L.: *Elements of Statistics.*
P.S. King and Sons, Ltd. London 1926

From the date of publication of these books it follows that they have been written in the last part of the second revolution. Westergaard and Nybølle's book is an elementary introduction written for students of economics and demography. It was supplemented by Bowley's book, which contains some sections of a more mathematical character.

Steffensen's book is a highly concentrated, mathematical exposition of Pearson's system of frequency curves, the method of moments, and the method of least squares according to Gauss's second proof.

I took my first course in statistics in 1935 by Nybølle, my second in 1936 by Steffensen, and my third in the method of least squares for geodesists by N. E. Nørlund in 1936.

The mathematics was dull, estimation meant routine calculation of point estimates by the method of moments or the method of least squares, there was no unifying idea, there was no enthusiasm, it was at the end of the second revolution, and the revolutionary spirit had disappeared.

On this background you may well ask what induced me to continue studying statistics. The answer is simple: Fisher has the full responsibility. His theory came as a revelation to me, transmitted through Steffensen and Rasch.

In 1936 Steffensen published an important paper on "Free functions and the Student-Fisher theorem" in which he used the method of transformation by means of Jacobians instead of Fisher's geometrical method to derive Fisher's results. In the autumn of 1936 he organized a seminar in which he asked me to discuss his paper. To do so I naturally had to read Fisher.

In the spring of 1937 Nørlund asked Rasch, who had spent the year 1935-36 in Fisher's department in London, to give a short course of lectures on Fisher's results for the linear-normal model as a supplement to his own previous lectures on the method of least squares. Rasch used Steffensen's method to prove Fisher's results.

In the autumn of 1937 Steffensen formulated the problem for the prize essay in mathematics for the year 1938 as "An exposition and critical assessment of the most important of R.A. Fisher's theories." I used the next 14 months to study Fisher and write an essay of 130 pages.

2.2 Consulting, Teaching, and Research 1939-1948

I had of course hoped to go to London and study under Fisher but the war stopped this plan. Instead I got the second best solution, namely to study and work with Rasch at the State Serum Institute from 1939 to 1941.

Between 1941 and 1948 I worked as a consulting actuary and as a statistical consultant, mainly on economic and industrial problems, and also as a private tutor. In 1943 I got a part-time research grant from the University and during the following years I gave some courses on Fisher's theories to actuarial students based on my Essay. And most important for my further work, by request of

the Danish Engineering Association I gave an extensive course of lectures on the application of statistics in industry in 1943-44.

Some of the publications resulting from these activities are listed in Table 3.

Table 3.
Some of my contributions to mathematical statistics
1939 - 1948

En matematisk fremstilling af R. A. Fisher's Teorier. (A mathematical exposition of R.A. Fisher's theories), 131 pp. Prize Essay, 1939. Duplicated version 1945.

On the determination of the phagocytic power of leucocytes (Maximum likelihood estimation in a grouped distribution). *Acta Path.* 1942, 20, 64-81. Together with M. Jersild and G. Rasch.

Nogle Anvendelser af Transformationsmetoden i den normale Fordelings teori (Some applications of the method of transformation in the theory of the normal distribution). *Festskrift til Professor dr. phil. J.F. Steffensen*, 1943, 52-65. Together with G. Rasch.

Den afstumpede normale Fordeling (The truncated normal distribution). *Matematisk Tidsskrift* B, 1946, 83-91.

The Decomposition of a Series of Observations Composed of a Trend, a Periodic Movement and a Stochastic Variable. 134 pp. G.E.C. Gads Forlag, 1948.

Statistiske Metoder med Eksempler på Anvendelser indenfor Teknikken (Statistical Methods with Examples of Engineering Applications). 654 pp. Det private Ingeniørfond, 1948.

Tabel - og Formelsamling til Statistiske Metoder (Tables and Formulas to Statistical Methods). 77 pp. Det private Ingeniørfond, 1948.

As you will see all the publications are strongly influenced by Fisher's ideas, and they thus represent my main written contributions to the dissemination of these ideas in Denmark in the period 1939-1948.

2.3 Professor of Statistics at the Faculty of Economics 1948-1960

In 1948 I succeeded Nybølle as professor of statistics in the Department of Economics. In my naivete I thought that I could now replace my work as a consultant by research, but that turned out to be a great mistake. For about five years I was overburdened with administration and teaching, and with conflicts with the professors of economics about the extent of statistics in the curriculum.

With the support of the Rector of the University I started an Institute of Theoretical Statistics in 1953 with E. Lykke Jensen as assistant. Realizing that I

could not get more support from the Economics Department I made a deal with my old friends, the engineers.

I promised to take up teaching of industrial statistics for members of the Engineering Association, and the Rector of the Technical University then financed two positions, in which I employed Hans Brøns and Erik Harsaae. We then had a nucleus of four, who tried to educate each other in probability and statistics, and to spread the gospel to the outside world.

These activities resulted in the publication of a number of textbooks for engineers and economists as shown in Table 4.

Table 4.
Textbooks on Statistics published 1952-1960

Hald, A.: *Statistical Theory with Engineering Applications.* 783 pp. Wiley, 1952.

Hald, A.: *Statistical Tables and Formulas.* 97 pp. Wiley, 1952.

Hald, A.: *Statistisk Kvalitetskontrol* (Statistical Quality Control). Together with E. Lykke Jensen. 399 pp. Teknisk Forlag, 1954.

Hald, A.: *Statistiske Metoder i Arbejdstudieteknikken* (Statistical Methods in Time Studies). 74 pp. Dansk Rationaliserings Forening, 1955.

Hald, A.: *Elementær Lærebog i Statistisk Kvalitetskontrol* (Elementary Text in Statistical Quality Control). 119 pp. Universitetets Statistiske Institut, 1956.

Hald, A.: *Forelæsninger over forsøgsplanlægning* (Lecture Notes on the Design of Experiments). Together with E. Harsaae. 154 pp. Dansk Ingeniørforenings Industrisektion, 1956.

Hald, A., E. Harsaae and E. Lykke Jensen: *Tilføjelser, økonomiske eksempler og øvelser til Statistical Theory* (Supplements, Economic Examples, and Exercises to Statistical Theory). 318 pp. Universitetets Statistiske Institut, 1959.

Lykke Jensen, E.: *Repræsentative undersøgelsers teori og metode* (Theory and Methods of Sample Surveys). 486 pp. Universitetets Statistiske Institut, 1957 and 1960.

2.4 Professor of Statistics at the Faculty of Natural Sciences 1960-1982

By 1958 the importance of statistics had been firmly established among economists, engineers, and mathematicians. Personally I had paid the price for this by being out of research for about ten years. I realized that if I ever should get back I had to make a new start.

Instead of returning to statistical inference I turned to decision theory, partly influenced by Wald's book and partly by my work in industrial statistics. Since the mostly used and simplest rules in statistical decision theory are those of sampling inspection by attributes I began investigating the theoretical foundation for the existing rules and to study the properties of a new model based on the hypergeometric distribution combined with a prior distribution and a linear loss function. I presented the first results at a discussion meeting at Imperial College in London in 1960.

For some time my activity on this project was slowed down because in 1958 it was decided to reorganize the teaching of statistics at the University. A new position as professor of mathematical statistics and a corresponding Institute was created under the Faculty of Natural Sciences, and together with Brøns I left the old institute to take over this new job.

During the following 15 years I published many papers on the properties of attribute sampling plans in cooperation with my colleagues at the Institute. I carried the project to its bitter end by publishing a book in two instalments in 1976 and 1978, later published by Academic Press in 1981 as *Statistical Theory of Sampling Inspection by Attributes* , 515 pp. (Tables together with U. Møller). I think it contains all that is worth knowing on the theory of sampling inspection by attributes - and a little more.

This work may be considered as an application of the principles laid down by Laplace, which I did not know in detail at the time. The asymptotic properties of the sampling plans follow from the Laplace expansion of the loss integrals. The analysis of this model has proved to be of great importance for my understanding of the first revolution.

Just after having completed the book in 1978 I was asked to write a short history of mathematical statistics in Denmark to be published by the University on the occasion of its 500th anniversary in 1979. A serious problem in this project was to understand how Thiele found the canonical form of the linear-normal model. To get the background for Thiele's work I read Gauss and wrote a detailed annotated transcription in modern notation of his papers on the method of least squares, which I hope to include, after suitable revision, in a book on the second revolution. My paper on Thiele's contributions to statistics was published in 1981.

2.5 Retirement and the History of Statistics 1982-1993

When I retired in 1982 I continued my research on the history of statistics, which so far has resulted in several papers and the book *A History of Probability and Statistics and Their Applications before 1750*, 586 pp., Wiley, 1990, giving the background for the first revolution. Since then I have nearly completed a manuscript on the first revolution and done some work on the second. Time will show if it will be completed.

Pivotal Models and the Fiducial Argument

George A. Barnard*

Abstract

A sketch of the history of Fisher's fiducial argument is accompanied by a version, called "pivotal inference" which, it seems, may give a consistent revised version of the ideas involved.

1 Some History

The first public exposition of the fiducial argument was given in 1929 by Dr. J.O. Irwin who, having asked Fisher for help with a talk he was to give to that summer's meeting of the British Association for the Advancement of Science, was told of the idea which had occurred to Fisher after a conversation with E.J. Maskell (1930: Journal of Tropical Agriculture, vii, 101-104, 125-131). Maskell's idea was, that one could invert a test of significance to give an interval within which a parameter could be expected to lie. A similar idea had been put by Egon Pearson to W.S. Gosset in a letter dated 7 November 1927, the year in which E.B. Wilson rediscovered Laplace's method of obtaining confidence limits for a binomial probability p and corrected Laplace by giving a correct account of their coverage property. Indeed a clear account of how limits can be put on an unknown probability p from the observation that, in n independent trials, an event of probability p has occurred r times, is to be found in Cournot (1843). Irwin told me of his British Association talk more than once, but he never referred to the existence of any text so I presume that no record of exactly what he said has been kept. In any case, Fisher's paper sent to the Cambridge Philosophical Society on the 23rd July, 1930, is the first careful general discussion of the "fiducial frequency distribution."

Fisher's introduction of the notion is preceded by a re-emphasis of the point he had made several years before, that his "likelihood" is not a probability — in particular that it is a point function, not an interval function:

*For Anders Hald on his 80th birthday in the hope that some of this may prove useful in his later volumes

> If A and B are mutually exclusive possibilities, the probability of "A or B" is the sum of probabilities of A and of B, but the likelihood of A or B means no more than "the stature of Jackson or Johnson"; you do not know what it is until you know which is meant.... there are, however, certain cases in which statements in *terms* of probability (my stress — G.A.B.) can be made with respect to the parameters of the population.

He goes on to cite the case where

> the random sampling distribution of a statistic T, calculable directly from the observations, is expressible solely in terms of a single parameter, of which T is the estimate found by the method of maximum likelihood [1]. If T is a statistic of continuous variation, and P the probability that T should be less than any specified value, we have then a relation of the form
> $$P = F(T, \theta).$$

After specifying the necessary continuity and monotonicity conditions, Fisher says

> we may express the relationship by saying that the true value of θ will be less than the fiducial 5 per cent value corresponding to the observed value of T in exactly 5 trials in 100....This then is a definite probability statement about the unknown parameter θ, which is true irrespective of any assumption as to its *a priori* distribution.

Later in this paper he writes "The fiducial frequency distribution will in general be different numerically from the inverse probability distribution obtained from any particular hypothesis as to *a priori* probability. Since such a hypothesis may be true [2], it is obvious that the two distributions must differ not only numerically, but in their logical meaning." It will be suggested below that the

[1] The fact that Fisher here introduced a requirement of "efficiency" for the estimate T has not always been noted.

[2] In the Collected Papers, revised by Fisher shortly before his death, there is a footnote to the phrase "Since such an hypothesis may be true":

It must not be known to be true - R.A.F.

The footnote does not appear in the reprint of the paper in the 1950 "Contributions to Mathematical Statistics" and I think the later addition reflects Fisher's stronger appreciation of the fact that in statistical inference it is as important to specify what is assumed to be unknown as to specify what is assumed to be known. Many of the conversations I had with him centered on the fact that purely mathematical theories of statisical inference must suffer the logical defect that in the logical structure of pure mathematics there is no distinction to be drawn between "true" and "known to be true". In statistical inference the distinction is of course vital.

"difference in logical meaning" to which Fisher refers is illustrated by the fact that the fiducial distribution assigned to a parameter does not make the parameter into a random variable in the sense defined by Kolmogoroff in his 1933 classic "Grundlagen der Wahrscheinlichkeitsrechnung." Specifically, it is an immediate consequence of Kolmogoroff's definition that any measurable function of a Kolmogoroff random variable (KRV) is itself a KRV; but it is only under special conditions that we can derive a fiducial distribution for a measurable function $f(\theta)$ from that of θ.

The next major development in Fisher's thought on these matters is given in his 1934 "Two New Properties of Mathematical Likelihood". This contains the beginning of "Conditional Inference", with its demonstration of the fact that problems of location λ and scale σ can be solved for continuous distributions of arbitrary form by taking any location and scale sample functions — such as the first observation x_1 and the difference $d = x_2 - x_1$ between the first observation and the second — forming what we would now call the "pivotals" $b = (x_1 - \lambda)/\sigma$ and $c = d/\sigma$ and considering the joint distribution of b and c conditioned on the sample configuration a whose i^{th} component is $(x_i - x_1)/d$.

In retrospect this paper seems unduly preoccupied with further promoting the likelihood function as carrying all the sample information about the parameters. Considerable space is taken up, for example, in proving that the average Fisher information in the observed x_{10} and d_0 referred to their conditional distribution is equal to the average information in the original sample. In these days of routine telemetering we can easily imagine that instead of learning the values x_{i0} in suffix order, we could first learn the observed value a_0 of the configuration, and afterwards learn the observed x_{10} and the observed d_0. The distribution of a does not involve either of the two parameters, and so can carry no direct information about these. It conditions the distribution of the pair (b, c). It then seems obvious that when we learn the values x_{10} and x_{20} of x_1 and x_2 our information about λ and σ will be fully contained in $b_0 = (x_{10} - \lambda)/\sigma$ and $c_0 = (x_{20} - x_{10})/\sigma$, referred to their joint conditional distribution.

In case anyone should doubt this, one may cite a "Turing type" argument: Suppose we had a person P with "second sight" who knew the values of any parameters like λ and σ which might come under consideration, and who knew the joint density $\varphi(.)$ of $p = (x - \lambda.1)/\sigma$. Such a person could programme his computer to generate a random sample a_0 from the know distribution of a. For each such sample he could programme the conditional density $\xi(b, c|a_0)$ and using a *completely independent* source of random numbers he could generate values of b and c. Knowing λ and σ he could convert these into values of x_{10} and x_{20}. Then using a_0 he could generate an artificial x_0. As a source of information about λ and σ, P would be indistinguishable from the genuine experimental source. But since neither λ nor σ were involved in generating the artificial a_0, its value can, like the experimenter's height, be ignored. The argument used by Thomas Bayes to establish his Propositions 4 and 5 has some connection with these considerations.

2 Pivotal Models

I now break into the historical account with an instance of valid fiducial reasoning as it now appears to me. Consider a clinical trial on the effect of a (change in) treatment intended to change the value of a quantity measurable for each of $2n$ patients who come for treatment in matched pairs. One patient, chosen at random for the pair, is given the changed treatment while the other is given the standard treatment. Distinct pairs are assumed to be independent of each other.

The effect of the change in treatment may be additive or multiplicative. Both cases can be treated along related lines, but for brevity I deal only with the additive case. Let us denote by x_i, $i = 1, 2, ..., n$ the difference between the measures of a matched pair. The effects of the treatment on the population of presenting patients can be described by two "pivotal parameters" μ and σ and a density φ such that in the population of matched pairs each of the functions

$$p_i = (x_i - \mu)/\sigma \tag{1}$$

has, independently, the density φ. Some knowledge of φ must be assumed, and we take φ to be approximable by a member of the Fechner family of unimodal densities

$$\varphi(u) = K \exp\{-\frac{1}{2}M^\alpha(u)\} \tag{2}$$

where the function $M^\alpha(u)$ is defined for $1 \leq \alpha < \infty$ and $M > 0$ by

$$M^\alpha(u) = u^\alpha \text{ when } u \geq 0 \text{ and } = (-Mu)^\alpha \text{ when } u \leq 0 \tag{3}$$

Here, as throughout the paper, K denotes a norming constant determined in each case by the condition that the density integrates to 1. In the present case its values is given by

$$1/K = (1 + 1/M)\, 2^{1/\alpha}\Gamma(1/\alpha)/\alpha \tag{4}$$

The parameter M may be thought of as a skewness parameter, while α may be thought of as a tail or kurtosis parameter since it determines thickness, or otherwise, of the tails. The pair (M, α) are "shape parameters", contrasting with the pivotal parameters (μ, σ). The wide range of shapes obtainable by varying M and α is illustrated at the end of this paper. The density (2) has its mode at 0, so the parameter μ denotes the mode of the distribution of the differences x_i. μ is then the most likely value for the effect of the treatment if it is to be applied to another patient randomly drawn from the same population. In the case of a medical treatment we are concerned with effects on individual patients. The sum of a set of differences has no operational meaning, and the modal effects would seem to be the value of primary interest. But if x_i represented an effect on the yield of a crop it could well be that we would be more interested in assessing

the mean change in total yield. In such a case we might specify the mean of the p_i, $\int p\varphi(p)\,dp$, to be zero, instead of the mode. [3]

We call our model a *pivotal model* because it defines, as *known* functions of observables and parameters, the *basic pivotals* p_i. Since their joint density $\psi(p) = \Pi_i \varphi(p_i)$ is *to some extent* known, knowledge of the values of the parameter would enable us to determine, to a corresponding extent, the distributions of the relevant functions of the observations. In our case the relevant functions of the observations are the differences x_i between treated and untreated pairs. When the observations are made the basic pivotals become determinate, though still unknown because the parameters are unknown. The basic pivotals *are* then partially known quantities whose distributions *were* to some extent known. Pivotal inference relies on the fact that functions of the basic pivotals may exist which do not involve the unknown parameters. When the observations have been made these functions will be known quantities. Their known values can be used to condition the distribution of the remaining unknowns, and the conditional distribution may turn out to be better known than the original distribution of the basic pivotals.

Pivotal inference is rigorously applicable only to data whose distribution is taken to be continuous. It is sometimes asserted that because the result of any measurement can be given to only a finite number of significant figures, any theory of statistical inference in which a fundamental distinction is drawn between continuous distributions and discrete distributions must on that account be flawed. Such an argument fails to take account of the fact that continuous transformations of continuous data constitute fundamental tools used in inference; and as soon as we depart from simple addition and subtraction any discretisation that has been supposed to be applied to a measurement of x will have to be changed if we are to treat *knowing* the value of x_0 as logically equivalent to *knowing* the value of x itself. [4]

The fact that "known" versus "unknown" is a distinction of importance in statistical inference, but not in pure mathematics, means that the Dirichlet-Bourbaki (DB) concept of a function, as any correspondence f between sets A and B such that to each x in A there corresponds just one $f(x)$ in B is not generally appropriate in statistics. The expression (1) above defines p_i *algorithmically* — a means is given whereby, for any triplet (x_i, μ, σ) we can effectively calculate the value of p_i. From it we can deduce that, knowing the *distribution* of p_i, and the

[3] Physicians would ultimately be interested not only in μ but also in the variability of the effect, as measured by σ ; they should also be interested in the shape of the distribution of effects, as indicated by M and α. Thus while μ will be the parameter of *primary* interest, estimates of the remaining three parameters should also be recorded for possible future use.

[4] "I do not wish to deny either that measurements, however accurate, are in the strictly mathematical sense discontinuous, or that counts may be of numbers so large that they could without sensible error be treated as measurements of continuous variables." (Fisher - Statistical Methods and Scientific Inference (SMSI) 3rd. Ed.p.53).

value of θ, we could deduce the *distribution* of x_i. This might not be the case were the "function" given only DB-extensively. We shall see that it is of importance in pivotal inference to know, for example, whether a function of the basic pivotals is, or is not, parameter-free, and whether it does or does not depend solely on a specific parameter. If the function is algorithmically given, such questions can be immediately settled from the form of the algorithm. With functions given in DB-extensive form, such questions may remain open unless the set A is finite. In what follows we understand "function" to mean "algorithmically given function".

In the 1930's it was noted by Irving Segal (1938) and, independently, by Paul Levy, that any continuous probability model which specifies that the observables x_i, $i = 1, 2, ..., n$ with density $\xi(x_1, x_2, ..., x_n; \theta_1, \theta_2, ..., \theta_s)$ depending on the parameters $\theta_1, \theta_2, ..., \theta_s$, can be expressed in pivotal form. We first find $F_1(x_1; \theta_1, \theta_2, ..., \theta_s)$, the cdf of the marginal distribution of x_1; then we find $F_2(x_2|x_1; \theta_1, \theta_2, ..., \theta_s)$, the cdf of x_2 given x_1; then we calculate the distribution $F_3(x_3|x_1, x_2; \theta_1, \theta_2, ..., \theta_s)$, and so on. Then we obtain a specification logically equivalent to the original ξ by saying that the point $(F_1, F_2, ..., F_n)$ is uniformly distributed over the unit n -cube. Such a transformation to pivotal form is clearly not unique, since we may change the order in which the x_i are taken; but so long as the ξ is taken to be known exactly, all the $n!$ possibly distinct pivotal forms are logically equivalent to each other.

It is otherwise in the more realistic case when, as here, the forms of our densities are not known precisely. Thus the precise forms of the functions F_i in the Levy-Segal transformation cannot be known, and the whole procedure loses precision. We have no guarantee that the transformed specification is logically equivalent to the original one. An advantage of initially specifying the model in pivotal form is that it allows us, as here for example, to specify quite precisely that it is the mode μ which we wish to estimate without having to specify precisely what form the observational distributions take. Measurements are invariably made to finite accuracy, and are finite in number, so that we can hardly ever know our distribution shapes exactly; but in pivotal inference we avoid the pretense of exactness otherwise necessary in our model formulations.

Another logically important consequence of the pivotal mode of specifying a model is, that it defines the class of functions with which we have to deal — the "fundamental probability space" of random variables under consideration. Continuous functions of the basic pivotals will themselves be pivotals with distributions known to a degree of accuracy determinable from the form of the function in question and the degree of accuracy with which the distribution of the basic pivotals is known. This will not be true of functions of the observations *and/or* parameters which are not expressible as continuous functions of the basic pivotals.

In case this should seem a severe limitation on the range of application of the pivotal mode of inference, we may point out that *any* function of observations *and/or* parameters can be added to the basic pivotal if it is judged to have a distribution known a priori to an accuracy comparable with that of the distribu-

tion of the basic pivotals. Extending the basic pivotal in this way will, of course, mean extending the assumptions on which our inference is based. The objective correlative of the basic pivotal is the real random mechanism which generates our observations. Our basic pivotal and what we say about its distribution should express what we claim to know about this mechanism. [5]

3 Pivotal Inference

In pivotal inference we try to transform the basic pivotals p to a new set $T(p) = (a, b, c)$ by a continuous 1-1 transformation T to a form $T(p) = (a, b, c)$ in which a is the maximal function of p which involves only the observations, b is the maximal function of p which involves only that function μ of the pivotal parameters which we wish to estimate, while c is the "complementary" pivotal required to make the transformation T $1-1$. In our medical example we set

$$p \Rightarrow T(p) = c(b\underline{1} + a) \tag{5}$$

subject to the conditions

$$\underline{1}'a = 0 \text{ and } a'a = n(n-1) \tag{6}$$

Here $\underline{1}$ denotes an n-vector of 1's and $'$ denotes transpose. Then a little algebra gives the inverse transformation as

$$a = (x - \bar{x}.1)\sqrt{n}/s_x,\ b = (\bar{x} - \mu)\sqrt{n}/s_x, c = s_x/\sigma\sqrt{n} \tag{7}$$

while the Jacobian is

$$J(a, c) = \Delta(a)\, c^{n-1} \tag{8}$$

where

$$\Delta(a) = n(n-1)/|a_n - a_{n-1}|,$$

(with a_n and a_{n-1} expressed in terms of a_j, $j = 1, 2, ..., n-2$ using (6)) is a determinant involving only n and a.

It is easy to see that in the model we have taken for our example, a is the maximal parameter-free function of the basic pivotals p, while b is the maximal function of p which involves the parameter function of interest, μ, but no other parameter. In this example the separation into the form (a, b, c) is complete. The joint probability density of (a, b, c) is

$$K\Delta(a)\, c^{n-1} \exp\{-\frac{1}{2}\Sigma_i M^{\alpha}(c(b + a_i))\} \tag{9}$$

[5] Those who take the view that all probabilities represent personal betting odds may here replace "claim to know about" by "are agreed about".

Since the function M^α is homogeneous of degree α, we can write this density as

$$K\Delta(a)\, c^{n-1} \exp\{-\frac{1}{2}c^\alpha H(b,a)\} \tag{10}$$

where

$$H(b,a) = \Sigma_i M^\alpha(b + a_i) = \left(\sqrt{n}/s_x\right)^\alpha \Sigma_i M^\alpha(x_i - \mu) \tag{11}$$

We now introduce the suffix $_0$ to denote the operation of substituting, in a pivotal, the observed values of all the observables in it. With this notation, in our example, a becomes a numerically known quantity $a_0 = (x_0 - \bar{x}_0.1)\sqrt{n}/s_{x0}$, similarly b becomes a numerically known function $b_0 = (\bar{x}_0 - \mu)\sqrt{n}/s_{x0}$ of the unknown parameter μ, while c becomes a numerically known function $c_0 = s_{x0}/\sigma\sqrt{n}$ of σ. The values of b_0 and c_0 remain unknown so long as nothing further is known about μ or σ. If M and α were known exactly, then H_0 would become the numerically known function $(\sqrt{n}/s_{x0})^\alpha \Sigma_i M^\alpha(x_{i0} - \mu)$. With M and α remaining unknown, H_0 becomes a family of functions which may or may not change drastically over the plausible range of (M, α).

Now suppose we are told the value of a_0 of a before we are told the values of b or c. Since neither μ nor σ are involved in a, knowledge of a gives us no direct information about either parameter. But the joint density of (b, c) is now the conditional:

$$K\ c^{n-1} \exp\{-\frac{1}{2}c^\alpha H(b,a_0)\} \tag{12}$$

and unless we gain further information about the value of c, for any set B the probability $Pr\{b_0 \in B\}$ is found by integrating over B the marginal density:

$$\zeta(b|a_0) = K\int c^{n-1} \exp\{-\frac{1}{2}c^\alpha H(b,a_0)\}dc = K/\{H(b,a_0)\}^{n/\alpha},$$

absorbing into K the Gamma function into which the integral converts. So long as nothing further is known about σ, c_0 remains unknown and the density $\zeta(b|a_0)$ remains applicable.

By way of comparison with a classical result, suppose we were prepared to assume $M = 1$ and $\alpha = 2$ – *i.e.* that the density of the basic p is spherical standard normal $N(0, I)$. Then

$$\begin{aligned} H(b,a_0) = (n/s_x^2)\,\Sigma_i (x_i - \mu)^2 &= (n/s_x^2)\left[n(\bar{x} - \mu)^2 + (n-1)s_x^2\right] \\ &= (n-1)\left[1 + b^2/(n-1)\right]^{n/2} \end{aligned}$$

from which it can be seen that the conditional density of b is Student's density with $n-1$ degrees of freedom, $K/\left[1 + b^2/(n-1)\right]^{n/2}$. The fact that a_0 does not

appear explicitly here is due to the fact that for the normal distribution, $(\bar{x}, s^2)$ is a sufficient pair.

Whatever the values of M and α $\{H(b, a_0)\}^{n/\alpha}$ can be calculated from (11) as a function of b and of a_0, so that the normalised density of b

$$\zeta(b|a_0) = [H(b, a_0)]^{-n/\alpha} / \int [H(b, a_0)]^{-n/\alpha} \, db \tag{13}$$

can be evaluated by a single quadrature. Given any set B, a further single quadrature (which can be programmed as part of the previous one) will therefore give $Pr\{b \in B|a_0\}$ and $Pr\{^{\sim}b \in B|a_0\}$.

Now when the observations $\bar{x} = \bar{x}_0$, $s_x = s_{x0}$, are known, for any set B the proposition $b_0 \in B$ is logically equivalent to $\mu \in \{\mu : (\bar{x}_0 - \mu)\sqrt{n}/s_{x0} \in B\}$. The 1-1 mapping

$$B \Rightarrow F_0(B) = \{\mu : (\bar{x}_0 - \mu)\sqrt{n}/s_{x0} \in B\}$$

maps the field $\{B\}$ of B sets onto the field $\{F\}$ of μ-sets. Associating with any element F of $\{F\}$ the numerical value of $Pr\{b \in B|a_0\}$, where b is the corresponding element of $\{B\}$ defines a probability measure on $\{F\}$, called the fiducial probability measure for μ.

We might be interested in σ as well as in μ. If so, we would take b to be the vector pivotal $((\bar{x} - \mu)\sqrt{n}/\sigma, s_x/\sigma\sqrt{n})$, and then c would be null. The joint density of the two components of b would be given by (12) and then to any subset B of the half-plane we could attach a probability $Pr\{b \in B\}$. When the observed values $\bar{x}_0$, s_{x0} are known, this probability can attach to the set

$$\{(\mu, \sigma) : \left[b_{10} = (\bar{x}_0 - \mu)\sqrt{n}/\sigma, b_{20} = s_{x0}/\sigma\sqrt{n}, \ (b_{10}, b_{20}) \in B\right]\}.$$

In practice M and α will not be known exactly. The programme for computing $H(b, a_0)$ can, however, be written with M and α as programme parameters. These can then be varied over their plausible range to asses the accuracy which can be attached to the fiducial distribution.

Perhaps it should be added that we have taken a simple concrete example to avoid heavy notation. In fact the approach used to this problem extends to most of what is called "Linear statistical inference".

4 Fiducial Probability and its Frequency Interpretation

The fiducial probabilities thus defined are mathematical probability measures in the sense of Feller (Vol.II,pp.111-2). As indicated in Feller's book, we can define probability densities $f(\theta)$ for the parameters involved with the property that for any parameter set F, the fiducial probability attached to F is $\int_F f(\theta) \, d\theta$. But to give the fiducial probabilities their objective, aleatory, or statistical character in

the sense used by Hald in Volume I (p. 28) of his history [6] , we must imagine the clinical trial under discussion to be one of a series of independent clinical trials, or other experiments with similar logical structure.

There are several ways of giving a frequency interpretation to an event's having probability p. Perhaps the broadest is that which refers to a "long run" of N independent trials with events having probabilities $p, p', p'', ..., p^N$, when the proportion of events which will occur will be close to the average $\bar{p}_N = \left(p + p' + p'' + ... + p^N\right)/N$. A simpler version takes $p = p' = p'' = ... = \bar{p}$, corresponding to repeated independent "throws" of the "same" die give a long run frequency approaching $\bar{p}$. Unnecessary restriction to this simpler version has led to the use of the phrase to which Fisher objected so strongly: "Repeated sampling from the same population" (RSSP). As he used to say, when a scientist seriously "repeats" an experiment, he always has in mind at least the possibility that it may turn out *not* to be the same population from which he is sampling — if he *knew* it was the same he would think of himself either as enlarging his original experiment, or wasting his time. For the frequency interpretation or fiducial probability we would therefore think of a series of experiments involving *various* pivotal parameters θ, θ', θ'',... not known to be connected one with another. We are therefore throwing different dice, and may very well consider different events, with probabilities p, p', p",... Independence will still imply long run convergence of the average frequency of occurrence to the long run average of the $p's$. The frequency interpretation of fiducial probabilities involves repeated sampling from *different* populations — RSDP instead of RSSP. In appealing primarily to RSDP instead of RSSP for its frequency interpretation, fiducial probability differs in its logical nature from Neyman's concept of "confidence". As will be noted below, insistence on taking RSSP as basic can lead to anomalies.

Another important difference between fiducial probability, which involves RSDP, and probabilities involving RSSP, is that any single statement of fiducial probability relates to a proposition $b_0 \in B$ that is already true or false, though we do not know which of these possibilities holds. We shall not know whether it is in fact true or false until we know the value of the parameter involved — and *typically* we never know this exactly.

The fact that the b involved is required to be a function of the basic pivotals does not only exclude some of the difficulties from which the fiducial argument has been judged to suffer; it may put a severe restriction on the parameter functions for which fiducial distributions can exist. The fiducial probability attached

[6] Hald adds "or epistemic" to "subjective or personal". I fear that the additional word is open to misunderstanding. The probability of "1" with an unbiased die is 1/6. If it becomes known that the number showing is odd, this probability becomes 1/3. Because of the intervention of "known", it is arguable that the 1/3 probability is "epistemic". I imagine that Hald would agree with me that it remains aleatory. The Greek philosophers who prided themselves on being $\epsilon\pi\iota\sigma\tau\eta\mu o\nu o\iota$ did not, so far as I know, distinguish between private knowledge obtained by mystical or other such means, and the type of public knowledge characteristic of science.

to a set Θ to which a parameter function θ may belong is equal to the probability, before the observed values entering into the pivotal were known, that the corresponding pivotal $t(\theta)$ belonged to the set T in the space of pivotals. This can be determined only if there is a function $t(b)$ which, given the observations, involves only θ. With the pivotals (b_1, b_2) referred to above, for example, if we take $\theta = \mu + 3\sigma$, we can set $t(\theta) = (b_1 - 3\sqrt{n})/b_2$ and obtain a fiducial distribution for $\mu + 3\sigma$. But for $\theta = \mu^2$, no such t exists and we cannot derive a fiducial distribution for μ^2. *When we know* $\bar{x}_0$, we can attach a fiducial probability to the statement $\mu \in \{(-4 < \mu < -1) \cup (1 < \mu < 4)\}$, and we may abbreviate this to $1 < \mu^2 < 2$; but the equivalence of this abbreviation to a statement about a function μ^2 of the basic pivotals will cease when $\bar{x}$ takes another value.

The theory of pivotal inference thus sketched seems to provide a contradiction free interpretation of Fisher's fiducial argument, which is why I have ventured to use the term "fiducial probability". The key restriction to functions of the basic pivotals secures the uniqueness of the maximal ancillary, since if a_1 and a_2 are both parameter-free functions of the basic pivotals, so is (a_1, a_2). Thus difficulties such as those raised by Mauldon, and Basu, where there appear to be multiplicities of ancillaries, do not arise with the present formulation .[7]

An exactly similar argument shows that the maximal pivotal involving only the parameter function of interest is unique. Thus the difficulties raised by Tukey do not arise either.

Although the unique maximal pivotal involving only the parameter of interest always "exists" it need not give rise to a valid fiducial distribution. In the Behrens-Fisher problem discussed below, the maximal pivotal for the parameter of interest takes only three distinct values, and so cannot give rise to a fiducial distribution. Nor need it be the case that the fiducial complement of a maximal pivotal is free from the parameter function of interest, the maximal pivotal may not contain all the relevant information. Fisher never claimed universal applicability for his fiducial argument; only that the form of statement to which it leads is particularly easy to understand and that its domain of application is sufficiently wide to make it of considerable interest.

5 More History — Bartlett, Fisher, and the Behrens-Fisher Problem

Fisher's next paper on the fiducial argument appeared in 1935 in Annals of Eugenics, the journal he then edited. It contains the passage:

> In general it appears that if statistics $T_1, T_2, T_3...$ contain jointly the whole of the information available respecting parameter $\theta_1, \theta_2, \theta_3...$,

[7]The mere existence of multiplicities of this kind would not be fatal to the theory. Data sets exist from which the probability of a proposition A can validly be assigned two distinct values. se Barnard (1977).

> and if functions $t_1, t_2, t_3, ...$, can be found, the simultaneous distribution of which is independent of the parameters $\theta_1, \theta_2, \theta_3, ...$, then the fiducial distribution of $\theta_1, \theta_2, \theta_3, ...$, simultaneously may be found by substitution.

He went on to propound his solution to the Behrens-Fisher (BF) and the variance ratio (VR) problems in a way which assumes that one can manipulate fiducial distributions as if the parameters involved were random variables having the Kolmogoroff property. That his idea in this area never ceased to evolve is shown by the footnote he added shortly before his death to the reprint in his Collected Papers:

> After "appears", insert "likely" — R.A.F.

But for many years his "solution" to the BF problem was a source of difficulty.

Soon after Fisher's BF solution appeared, Bartlett pointed out, in a paper (1936) published by the Cambridge Philosophical Society (CPS), that with all Fisher's tests up to that time, in a long run of cases in all of which the hypothesis tested was true, the frequency of rejection was equal to the critical P-value α used in making the test; but with Fisher's BF test the hypothesis would be rejected when true with a frequency less than α. Bartlett indicated that it was possible to devise a test which would have the "α-frequency property", though he accepted that his test had unsatisfactory features. Fisher submitted a reply to Bartlett's paper for publication by the CPS, but it was refused publication. Considering himself unfairly treated, Fisher resigned his long-standing membership of the CPS. Bartlett was led to believe that Fisher had resigned, not because they had refused to publish Fisher's paper, but because they had published Bartlett's paper. [Following an appeal by Harold Jeffreys Fisher rejoined].

Another misunderstanding between Bartlett and Fisher was developing on the subject of "sufficiency". Bartlett's definition agreed with that still much in use in requiring that the distribution of the observations, conditional on a statistic sufficient for a parameter θ, should not involve the parameter θ. Thus for normal samples the statistic s_x^2 was not sufficient for σ^2 because the density of $\bar{x}$, $N(\mu, \sigma^2/n)$, itself involves σ^2. For Bartlett it was the "theoretical statistic" $n(\bar{x} - \mu)^2 + (n-1)s_x^2$ that was sufficient for μ. Fisher had not given a formal mathematical definition of sufficiency, though he had referred to s_x^2 as being sufficient for σ^2. His letter to me dated 17 October 1953 suggests that he was still unwilling to be tied down to a precise set of definitions in this regard, as in so many others. The question of the meaning of "sufficiency" resurfaced, with Williams in place of Bartlett, after Fisher was asked to analyse the rock magnetism data that led to the current theory of continental drift cf. Barnard (1963).

It soon became clear that the sufficiency definition used by Bartlett and others could apply only if the densities involved belonged to the exponential family.

The Fisher's insistence on sufficiency as a necessary requirement for the fiducial argument led many to refuse to accept the latter, on the ground that we could never in practice know the observational distribution with the precision required to determine whether or not is was of exponential type. It is an appealing feature of the pivotal mode of inference that it is the precisely known forms of the basic pivotals, — describing the logical structure of the relationship between the pivotal parameters and the observations — not the inexactly known form of the pivotals' joint distribution, which is critical in determining the possibility, or otherwise, of arriving at a fiducial distribution. This conclusion was implicit in Fisher's 1934 paper. And Bartlett certainly saw the implications, at least to some extent — perhaps more clearly than did Fisher himself. He proposed the term "quasi-sufficient" for the sample scale and location parameters as estimates of the population location and scale of a continuous distribution of arbitrary form. And unlike so many of his contemporaries, he clearly rejected Neyman's insistence on prespecification of error rates regardless of the inevitable variations of precision associated with variations in the ancillaries. It is a great pity that this profoundly important paper was published in the Miscellanea section of Biometrika (Vol.31, 1940, pp.391-2) under the less than informative title "A Note on the Interpretation of Quasi-Sufficiency". But any hopes that further discussions such as those organised through the 1930's by the Industrial and Agricultural Research Section of the Royal Statistical Society might eventually result in better mutual understanding between these two great men were dashed by the outbreak of war which scattered the leading personalities far and wide around the country and set them to work on a variety of more immediate tasks.

The second world war saw tremendous expansion both in the applications of statistical methods, and in the use of probability for construction of "operational research" models of situations. The developments in the USA and the UK took place in the absence of an adequate literature in English on mathematical probability. In developing the theory of sequential tests, for example, one has to study papers well over 100 years old concerning the classical "problem of points". The most serious account in English of mathematical probability theory was in Cramer's little book "Random Variables and Probability Distributions" (1937) which does not contain a precise definition of "random variable". Such a definition is still absent from Cramer's 1946 account of mathematical statistics. Kolmogoroff's classic "Grundlagen der Wahrscheinlichkeitsrechnung" was not available in English until 1950. M.G. Kendall had persuaded a number of British mathematical statisticians of the need for a text on the "Advanced Theory of Statistics" in 1942, but he was left on his own to produce the first volume, on Distributions, in 1943. In that year an early version of Wilks's treatise became available in mimeographed form. The full printed version was delayed until 1962.

By the time the war ended interest in the foundations of statistical inference had shifted away from the fiducial argument. Sampling inspection problems had drawn attention once more to Bayes' Theorem; and the likelihood ratio

sequential test led much discussion around the likelihood principle. The use of Haar (invariant) measure to obtain "Frequency justifications" for sequential tests of opposite hypotheses led to suggestions for the 'reconciliation' of the three main theories of statistical inference, due to Fisher, Jeffreys, and Neyman and Pearson. It then seemed to me it might be possible to interpret Fisher's fiducial probability as a likelihood integrated with respect to a relevant Haar measure. A similar idea had occurred to Tukey but when Francis Anscombe in the early 1950's arranged for the two of us to meet, along with one or two others, at Imperial College, we had agreed that this interpretation would not apply to the correlation coefficient, the very first example of fiducial distribution given by Fisher. The group invariance idea was taken up by Fraser, who used the term "structural inference" to describe the central concept.

In 1954 attention again focused on the fiducial argument as a result of work by Monica Creasy on the fiducial distribution of the ratio of two independent normal means μ and $\alpha\mu$. Using arguments similar in form to those used by Fisher to derive his solution to the BF problem, she was lead to fiducial limits for α which sometimes differed quite drastically from those derived in the 1930's by Fieller and expounded in subsequent editions of Fisher's "Statistical Methods for Research Workers" (SMRW). Fieller's limits were also confidence limits, though when interpreted as such they suffered from the anomaly that the "guaranteed coverage frequency" corresponding to confidence γ in RSSP was attained only by reason of the fact that every now and then the limits would be given as $(-\infty, +\infty)$. Since these limits would clearly cover the true value 100% of the time, it immediately follows that the coverage frequency of any *finite* limits must be *less* than $100\gamma\%$. Neither the confidence concept nor the fiducial argument appeared free from difficulties in this problem. [8] Fisher himself did not take part in the discussion of what came to be known as the "Creasy-Fieller" problem. He later pointed out, in effect, that Fieller's treatment, which he had followed, did not require the assumptions made by Fieller and by Creasy. Denoting the sample observables by x and y, it *followed* from their assumptions that the pivotal $y - \alpha x$ was normally distribution about zero with unknown variance. But the converse was not true — their assumptions did not follow from the mere assumption that $y - \alpha x$ was normal with zero mean. Thus Creasy was using further "information", and her inference could well be different. He did not, however, go into details.

Further examples suggesting non-uniqueness of fiducial intervals were given in 1955 by Mauldon and in 1957 by Tukey. The 1956 publication of Fisher's "Statistical Methods and Scientific Inference", with its vigorous restatement and defense of the fiducial argument drew attention to the problems raised. He asserted in a footnote that

> Probability statements derived by arguments of the fiducial type have often been called statements of "fiducial probability". This usage

[8] For a discussion of this problem from the pivotal point of view see Barnard (1994).

> is a convenient one, so long as it is recognised that the concept of probability involved is entirely identical with the classical probability of the early writers, such as Bayes.

But in 1959 Stein pointed to the case where we are given unrelated observations x_i, $i = 1, 2, ..., n$ having means μ_i and unit variance and where the parameter function of interest is $\delta^2 = \Sigma_i \mu_i^2$. The mode of reasoning used by Fisher in the BF problem would lead to the conclusion that the μ_i were spherically normally distributed around the end point of the vector x; but the coverage frequencies estimated from this conclusion could be wildly misleading.

At the Paris session of the ISI in 1961 Fisher's return fare to Adelaide had been paid by the French CNRS; but the Paris to Adelaide ticket had been sent to Adelaide. Daniel Dugué, Fisher, and I, spent the best part of a day touring Paris in a taxi, with Daniel trying to persuade Air France, the CNRS, and other bureaucrats to do something to rescue the situation. We learned then just how near to being gold bullion an airline ticket is. Fisher's wrath was always liable to find target in the nearest person to hand, and I did my best to keep him plied with questions about the precise logical nature of pivotals and the distributions derived from them, partly, of course, to try to clarify my mind, but partly also to prevent him focusing on the air ticket muddle.

That was the last time I saw Fisher. We continued to correspond after his return to Adelaide, and he also corresponded with David Sprott and Donald Fraser. His discussions with Sprott and Fraser centered around the fiducial distribution of the bivariate correlation coefficient, starting from the puzzles arising from the multiplicities to which Mauldon had drawn attention. Details of his correspondence can be studied by reference to the differences between the three editions of SMSI, and to the correspondence between Fisher and Sprott, Fraser and myself, published in Henry Bennet's edition of Fisher's "Correspondence on Statistical Inference and Analysis." Alas, it was only after Fisher's premature death that the idea of "pivotal inference", as sketched above, occurred to me and to David Sprott. What Fisher would have thought of it must therefore remain unknown. The only clue we have is his footnote inserting "likely" after the statement quoted above.

It now appears that Fisher's statement quoted at the beginning of section 5 is true, but only if we note that we cannot assume that the θ_i, $i = 1, 2, 3, ...$ thereby become *Kolmogoroff* random variables, i.e. functions on a probability space, such that functions of them are in turn also random variables. As Bartlett (1937) pointed out, in essence, long ago, it is the *pivotal* which is the Kolmogoroff random variable, not the parameter. Given the fiducial distribution of a (vector) parameter θ based on a (vector) pivotal T it is possible to deduce a fiducial distribution for a function $\eta(\theta)$ only if we can find a function $G(T)$ such that

$$G(T) = H(\underline{x}, \eta(\vartheta))$$

where $\underline{x}$ denotes the observables, and $H(\underline{x}, \eta)$ is of such a form that it can serve as a pivotal for η. In the BF problem the pivotal $\underline{x}$ has four components which can be written:

$$t_1 = \bar{x} - \mu/s,\ t_2 = (\bar{x} + d - \mu - \delta)/rs,\ t_3 = s/\sigma,\ t_4 = rs/\rho\sigma.$$

The parameter of interest is δ. But as proved in Barnard (1982) the only function of T of the form $H(x, \delta)$ takes at most 2 distinct values, and so it cannot serve as a pivotal for δ. The corresponding result for the ratio is proved in Barnard and Sprott (1983).

Had this restriction been recognised earlier, there would have been no need for the considerable controversy that has attached to the BF problem over many years.

6 Fiducial Probability and Confidence

Commenting on a paper submitted to the ISI Review in which I attempted to make some of the points made above, a referee remarked that what I have here ventured to call "fiducial probability" appeared to him to have a strong resemblance to "confidence". I have already mentioned Egon Pearson's 1927 letter to Gosset containing the germ of the confidence interval idea; and perhaps I may say now that in the reminiscent conversations I was able to have with Egon after his retirement to Pendean it became clear to me that in so far as such a gentle character as he was could be said to harbour anything remotely approaching jealousy or resentment, Egon felt that his share in the ideas underlying Neyman's 1937 paper setting out the theory of confidence sets had gone unrecognised. (The perceptive reader of Constance Reid's account of the history of that paper will see that her account by no means conflicts with this view. It should be borne in mind that it was only a few years since Egon had been at pains to persuade Jerzy that he should finally abandon the "classical" Bayesian approach to inference in which he had been trained).

The referee's comment suggests that it will be worth while to make clear some of the ways in which fiducial probability differs from confidence. Principal among these is the idea of RSSP which is not involved in fiducial probability but which is involved in Neyman's confidence concept.

Until Bartlett (1935) pointed out that in the Behrens-Fisher problem the probability of rejection of the hypothesis $\delta = 0$ was less than Fisher's P-value, most statisticians (including, for example, Egon Pearson) seem to have thought that the differences of concept between fiducial limits and Neyman's concept of confidence limits were only minor. Fisher noted early on that in allowing the possible use of inefficient statistics Neyman's original proposal could give rise to a multiplicity of "conflicting probability statements". Neyman responded by saying, correctly, that his "confidence" statements were not to be understood as probability statements. In devoting one of his last papers to a method of obtaining

confidence limits for the cross-ratio in a 2x2 table, Fisher acknowledged, if only implicitly, that the confidence concept has its uses. Neyman devoted a paper to demonstrating the difference between his concept and Fisher's, though so far as I know he never asserted that Fisher's "fiducial argument" was fallacious.

In the theory of confidence sets we are to envisage RSSP resulting each time in a set calculated from the new observations on the *same* population. The long run frequency with which the set covers the (fixed) true parameter value θ is to be bounded below by a predetermined positive γ *chosen* by "the experimenter". Subject to this condition the frequency with which any false value is covered is to be minimised, so far as possible. Extremely useful though his idea is, the element of pre-choice involved shows that it constitutes "inductive behaviour", to use Neyman's accurate term, rather than an inferential process such as pivotal inference exemplifies. With fiducial inference anyone is at liberty to nominate the set Θ of possible θ values in which *he/she* is interested, and the fiducial probability associated with Θ will indicate how much confidence (in the non-technical sense) could be associated with a judgment that $\theta \in \Theta$, if we should choose to make it.

That the confidence concept *essentially* involves RSSP has become clearer with work stemming from the late Jack Kiefer's idea of conditional confidence — a brilliant method of avoiding the otherwise disturbing implications sometimes arising from the unconditional approach adopted originally by Neyman. In repeating sampling from a single normal population with unknown mean μ and unknown variance σ^2, the optimum confidence sets are of the form $\{\mu : \bar{x} - us_x < \mu < \bar{x} + us_x\}$ where u is $1/\sqrt{n}$ times the relevant percentage point of Student's distribution. In repeated sampling, the midpoints of these sets vary around μ independently of the varying length $2us_x$ of the intervals. It immediately follows that sets for which s_x is larger than usual must cover the true value more frequently than sets for which s_x is smaller than usual. The increase in coverage probability resulting from conditioning on the event $|\bar{x}|/s_x < K$, for a suitably chosen K, is surprisingly high for sample sizes less than 10. A recent paper by Goutis and Casella (1992) shows that if we make a *guess* g as to the true value of μ, and shift the origin of measurements to g before setting up our confidence procedure (moving the origin back afterwards, if need be), then a good guesser can hope to get a large K and so raise the expected confidence very considerably. There is nothing paradoxical about this — clairvoyance has often been regarded as desirable, if only it were possible. But is does serve to emphasise that the procedure involved cannot be regarded as one of *scientific* inference (at least until repeatable procedures for obtaining clairvoyance have been described and tested.)

In a recent paper (1991) Zabell has suggested that the possibility, not mentioned by Fisher, of conditioning on $|\bar{x}|/s_x < K$ is the Achilles heel of Fisher's argument for the fiducial distribution of a normal mean. It is quite possible that Fisher did simply overlook this possibility; but such an omission more probably

reflects the fact that the repetitions Fisher had in mind would have involved arbitrary variations in the $\mu's$ involved — not RSSP but RSDP. With RSSP we can set a lower bound to the frequency with which $|\bar{x}|/s_x < K$ will occur. But with RSDP we can set no such lower bound, so that if we condition we may be doing so on an event of zero long run frequency. Within the above account of pivotal inference, conditioning on the observed $|\bar{x}|/s_x$ would be disallowed because the function $|\bar{x}|/s_x$ is not expressible as a continuous function of the basic pivotals.

7 Fitting the Fechner

The non-pivotal parameters M and α may be estimated from the ancillary a. We may use the marginal likelihood based on the observed a obtained by integrating out b and c from the expression for the joint density of (a, b, c). If the marginal likelihood functions are routinely plotted for trials of a given type, their combination can provide information helpful in the interpretation of the results and in judgments as to the relative plausibility of various values of M and α.

Alternatively, the moment-ratios g_1 and g_2 of a regarded as a sample can be equated to the expressions in terms of M and α for the corresponding population values γ_1 and γ_2, as suggested in Fisher's "Statistical Methods for Research Workers", Section 14. As another alternative we may take a set of four percentiles of a, $q_1 < q_2 < q_3 < q_4$, and equate the location-scale invariant ratios $(q_4 - q_1) / (q_3 - q_2)$ and $(q_4 - q_3) / (q_2 - q_1)$ to their population values. An advantage of (g_1, g_2) is, that testing these for departure from normality is straightforward, using Fisher's k-statistics. Should no significant departure from normality be indicated, it could be reasonable to assume normality, thus avoiding the necessity for the quadratures indicated around equation (13) above. Before going to the trouble of performing these quadratures in any case a plot of the function $1/H(b_0, a_0)$ for various values of M and α may well indicate that for *the sample to hand* the ancillary is sufficiently near to normality to justify taking $M = 1$ and $\alpha = 2$ even though one may know or suspect that the true density is far from normal. The principle involved here is related to that involved in Efron's "Bootstrap", which suggests it might be called the "Stirrup".

The appended graphs, due to Dr. Bruce Worton and Miss K.J. Thomas, show some of the shapes in the Fechner family. The latter's M.Sc. (Essex) dissertation provides further information. The symmetric densities illustrated in Chapter 3 of Box and Tiao (1973) are also obtainable along with their skew counterparts.

8 References

Barnard, G.A. (1952): "The frequency justification of certain sequential tests", Biometrika 39, 144-150.

—— (1963): "Some Logical Aspects of the Fiducial Argument". Journal of Royal Statistical Society *B* 25, 111-114 (esp. pp. 113-114).

—— (1977): "Pivotal Inference and the Bayesian Controversy". Bulletin of Internat. Statist. Inst. 47 (1), 543-551.

—— (1982): A New Approach to the Behrens-Fisher Problem". Utilitas Mathematica 21 *B*, 261-271.

—— (1985): "Pivotal Inference". Johnson-Kotz Encyclopedia Article.

—— (1977): "Mathematics in Life Sciences", in D.E. Matthews (ed.): *Lecture Notes in Biomathematics*, Springer Verlag, New York, pp. 31-32.

—— (1994): "Pivotal Inference Illustrated on the Darwin Maize Data", in Smith, A.F.M. & P. Freeman (eds.): *Aspects of Uncertainty: Essays in Honour of D.V. Lindley*, John Wiley & Sons Ltd.

Barnard, G.A. & Sprott, D.A. (1983): "The Generalised Problem of the Nile: Robust Confidence Sets for Parametric Functions. Ann. Statist. 11, 104-113. [9]

Bartlett, M.S. (1940): "A Note on the Interpretation of Quasi-Sufficiency". Biometrika 31, 391-392.

Box, G.E.P. & G.C. Tiao (1973): *Bayesian Inference in Statistical Analysis.* Addison Wesley, Reading, Massachusetts.

Cournot, A.A. (1843): *Exposition de la Theorie des Chances et des Probabilites.* Librairie Hachetter, Paris. (See especially Section 96.)

Creasy, M.A. (1954): "Limits for the Ratio of Means". J. Royal Statist. Soc. *B* 16, 186-194 (and discussion 204-222.)

Fieller, E.C. (1954): "Some Problems in Interval Estimation". J. Royal Statist. Soc. *B* 16, 175-185 (and discussion pp. 204-222.)

Goutis, C. & G. Casella (1992): "Increasing the Confidence in Student's t Interval". Ann. Statist. 20, 1501-1513.

Sprott, D.A. & V.T. Farewell (1993): "The Difference Between Two Normal Means". The American Statistican 43, 126-128.

Wilks, S.S. (1940): "On the Problem of Two Samples from Normal Populations with Unequal Variances". Ann. Math. Statist. 11, 475-476. [10]

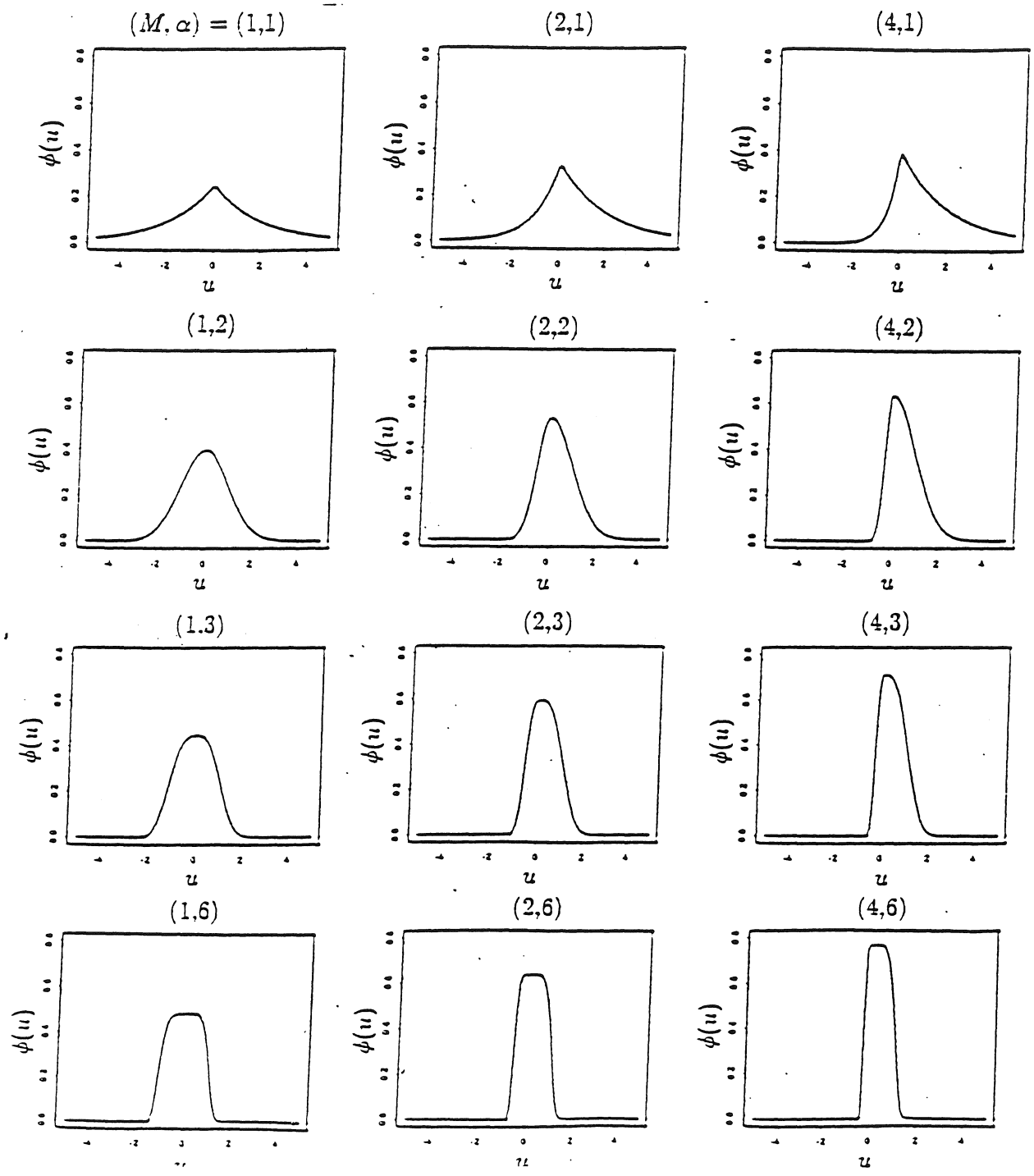
[9] I take this opportunity to acknowledge great assistance in developing the ideas involved in pivotal inference from conversations and correspondance with David Sprott.

[10] This reference is given because it leads to the conjecture that Wilks anticipated the mathematical result given in Barnard 1982. Fisher urged Wilks to publish, but he did not do so.

Standard Fechner densities $\varphi(u; M, \alpha)$ for various values of M and α

$$\varphi(u; M, \alpha) = K \ \exp -M^{\alpha}(u)$$

$$M^{\alpha}(u) = u^{\alpha}, \text{when } u \geq 0, \text{ and } = (-Mu)^{\alpha} \text{ when } u \leq 0$$



Preprints 1993

COPIES OF PREPRINTS ARE OBTAINABLE FROM THE AUTHOR OR FROM THE INSTITUTE OF MATHEMATICAL STATISTICS, UNIVERSITETSPARKEN 5, DK-2100 COPENHAGEN Ø, DENMARK. TELEPHONE +45 35 32 08 99.

No. 1 Hansen, Henrik and Johansen, Søren: Recursive Estimation in Cointegration VAR-Models.

No. 2 Stockmarr, A. and Jacobsen, M.: Gaussian Diffusions and Autoregressive Processes: Weak Convergence and Statistical Inference.

No. 3 Nishio, Atsushi: Testing for a Unit Root against Local Alternatives

No. 4 Tjur, Tue: StatUnit - An Alternative to Statistical Packages?

No. 5 Johansen, Søren: Likelihood Based Inference for Cointegration of Non-Stationary Time Series.

Preprints 1994

COPIES OF PREPRINTS ARE OBTAINABLE FROM THE AUTHOR OR FROM THE INSTITUTE OF MATHEMATICAL STATISTICS, UNIVERSITETSPARKEN 5, DK-2100 COPENHAGEN Ø, DENMARK. TELEPHONE 45 35 32 08 99, FAX 45 35 32 07 72.

No. 1 Jacobsen, Martin: Weak Convergence of Autoregressive Processes.

No. 2 Larsson, Rolf: Bartlett Corrections for Unit Root Test Statistics.

No. 3 Anthony W.F. Edwards, Anders Hald, George A. Barnard: Three Contributions to the History of Statistics.