ATLANTIS PRESS
Journal of Statistical Theory and Applications, Vol. 17, No. 2 (June 2018) 193–205

On Evolution of Statistical Inference

D A S Fraser *

June 2, 2018

Received 11 February 2018
Accepted 1 March 2018

Abstract

The foundations of statistics have evolved over many centuries, perhaps millennia, with major paradigm shifts of the form described in Kuhn (1962). We briefly consider these important transitions and how they have led to major shifts in the foundations of statistical inference. Clearly there is no conventional mathematical or axiomatic basis. But there is a progressive clarification in the processes of statistical inference so that current theory can now coherently and definitively handle a wide range of inference problems.

1 Introduction

Statistics theory with many variants has been evolving for many centuries, from primitive counts and tabulations, to averages and scalings, through the use of probability to construct statistical models, to Bayes methodology, to testing and decision theory, to Fisher design and analysis, to exploratory and empirical methodologies, to saddlepoint and likelihood approximations. These many interim steps have sometimes been in agreement, sometimes in conflict, sometimes politicized or promotional, but often with need for more unity and integrity.

*D A S Fraser, Department of Statistical Sciences, University of Torontodfraser@utstat.toronto.edu, This research was partially supported by the Natural Sciences and Engineering Research Council of Canada and Senior Scholars Funding at York University.

We focus on the key evolutionary steps where major transitions of emphasis or substance have occurred in the statistics theory; we do not attempt a detailed history of the developments themselves.

2 Records counts and tabulation

The emergence of the written word must surely mark the beginning of records, with recorded counts and tabulations soon to follow. Counts and tabulations gained importance with the rise of official statistics in turn providing background material for life insurance and other such endeavours. Modern big data is just a reemergence of counts and tabulations on a massive scale. The need for statistical methods for such data is just now gaining the attention it deserves.

3 Statistical models

Probability arose from the assessment of gambling results in the 17th century and earlier, and developed as part of mathematics from the 18th century onward. Its introduction to counts and tabulations served to address possible alternatives to any particular tabulation, thus giving us the statistical model, with probability for the possible alternatives, and parameters for the unknowns in the deterministic and probabilistic components. Such a model can then be presented in a density form as $f(y;\theta)$ often with a Euclidean differential dy. But models can also be presented in many other forms, say using distribution functions, or even structural equations that embed the randomness in specified error distributions. The statistical problem can then be viewed as the analysis of the combination $\{f(y;\theta), y^0\}$ where y^0 designates the observed data on the variable y. The inference problem is then to ascertain what that model data combination says about the value of the unknown parameter θ.

4 Early inference

With tabulations and curiosity concerning possible patterns, the average and root mean square deviation allowed the assessment of patterns such as "central

tendency" to be addressed pragmatically. The Central limit theorem or its early form as DeMoivre's (1738) version for the Binomial distribution led to a more formal assessment of departures from what one might expect under specified conditions.

5 The Bayes dilemma

Bayes (1763), as published in 1763, had been examining the Binomial Model $f(y;p) = (\frac{n}{y})p^y q^{n-y}$ with observed data and, being aware of the conditional probability formula, proffered a random source $\pi(p)$ for the realized p value and even suggested a modified roulette wheel with behaviour $\pi(p)$ as having produced that realized value p; this gave him an alleged joint model $\pi(p)f(y;p)$ for (p,y) and then the conditional distribution $c\pi(p)f(y^0;p)$ in the presence of observed data. Mathematics majors of course know that to use a theorem with a missing ingredient you can't just make it up and then assert that the results of the theorem apply. At that earlier time many prominent mathematicians, Boole, Venn, Chrystal. De Morgan and others, saw the flagrant flaw in this Bayes approach and objected with varying vehemence; for a lively discussion of the dialogue see Fisher (1956). Laplace (1812), however, in 1812 with his powerful involvement in science and perhaps anticipating confidence was mildly supportive but sought outside principles, such as that of non-informative prior information, to support the choice of a prior distribution.

The Bayes approach however is still widely practised despite serious criticisms. It might well have had greater influence had it been promoted as an exploratory methodology.

6 Sufficiency and Likelihood

The fundamental concepts of sufficiency and likelihood were introduced by Fisher and formally discussed in Fisher (1922) as part of fundamental new directions for statistics theory. A statistic t is sufficient if the conditional distribution $f(y|t;\theta)$ given the statistic is free of the parameter θ, that is, the conditional density can be written as just $f(y|t)$. The development of sufficiency theory could also have

been expressed as: t is sufficient if the the parameter only affects the statistic t, thus $f(y;\theta) = g(t,\theta)h(y)$. As sufficiency reduces the data size from that of the full data, say n, to that of the statistic t say d, there is serious interest quite naturally in finding the statistic that makes the maximum reduction, thus extracting all the information available in as simple a form as possible; such a statistic is called a minimum sufficient statistic.

The likelihood function from a model data combination records how probability at the observed data depends on the parameter: thus $L(\theta) = cf(y^0;\theta)$ where c is an arbitrary positive constant that forces likelihood to record only relative values of probability at the observed data y^0.

There is also a simple set theory viewpoint that directly gives the likelihood function. From the model and data we have of course the function $f(y;\theta)$; and we know the value $y = y^0$: so put them together and of course the information is just $f(y^0;\theta)$. In addition however there is always implicitly a support metric that is specified but just up to a constant multiple. We thus obtain the likelihood function $L(y^0;\theta)$ by an almost trivial argument.

It was only much later, that Barndorff-Nielsen (Barndorff-Nielsen) in 1976 noted that the mapping from the data space $\{y\}$ to the possible likelihood functions provided the minimal sufficient statistic immediately. Thus the two concepts, sufficiency and likelihood, are intrinsically related, almost in a mild equivalence. And then even further there is multi-faceted irony in finding much later that Fisher (1934) much earlier in 1934 had stated this as obvious; see the last complete paragraph on the page 300.

7 Reject or Accept

As mentioned in Fisher (1956), Chapter 4, tests of significance have been in wide usage in the early years of the 20th century, for example, the Pearson χ^2 test and the Student t-test. In these tests some measure of departure of data from what was expected was compared with the corresponding distribution for such under the hypothesis being examined, to see if the data are in some reasonable range of what might be expected under the hypothesis. Fisher's purpose in discussing the tests of significance was to draw critical attention to a 1933 paper by Neyman

and Pearson (1933) that proposed a mechanization of tests of significance: the hypothesis being tested would be Rejected or Accepted according as whether or not the value of a departure measure was beyond some critical value often $\alpha = 5\%$ under the hypothesis distribution. In addition, the critical value would be determined so as to maximize the probability of rejection under alternatives to the hypothesis. This automatic Reject or Accept was viewed by Fisher as totally at odds with appropriate scientific practise. However, this automatic process was well supported by the mathematics community at that time so negative criticism of the decision theoretic approach was easily ignored. It was not until 1959 that Sterling (1959) gave serious and well documented criticism of the Reject Accept procedure.

However more recently there has been much increased attention to the use and misuse of p-values. Even the Editors of one journal in the social sciences have declared they will not allow p-values in new submissions to the journal; that is, not allow the Reject-Accept version for p-values. And the American Statistical Association has had a committee to address the role of p-values; and others are considering whether the rejection level should be moved to 0.5%. But this seems to overlook the fact that deciding whether hypotheses are true or not based on a break point in some measured variable is a violation of scientific and common sense. Indeed Rozeboom (1960) cites a philosophical epigram that the null hypothesis significance test was the "glory of science and the scandal of philosophy". We return to this with a discussion of the p-value function in §10 and §11 .

8 Subjective priors

In 1954 Savage (1953) discussed the use of subjective priors, priors that represented feelings, judgements, perceptions concerning an unknown parameter of a statistical model in an investigation. The prior would be used with the given model in a conditional probability analysis. This is not what Bayes had done when he proposed what can be viewed as a mathematical prior to initiate a statistical analysis. Rather it accepts a prior with its subjective content as representing some background information concerning the unknowns in the problem,

and then with the observed data does standard probability analysis: this leads to a presentation of a nominal conditional probability model as a description of the unknown parameter value. As such it is not what Bayes had proposed, but is standard model building but with different stages in the construction of a composite model and varying support for the different stages. Nonetheless many with Bayesian sympathies quickly embraced the approach as a serious extension of the standard Bayes analysis of the time.

The quality of prior information can be a major factor as to whether it can be used reliably in such an analysis. In addition in many contexts, such prior information should explicitly not be used. For example in Experimentation and Sampling such information would not be used as it is in conflict with physical randomization, the standard procedure for validation in experimental design and analysis. Accordingly those involved with Bayes type analysis fall generally into two groupings: those with strong views on the merits of the subjective approach, and those supportive of the mathematical priors of Bayes and Laplace recently called objective priors, in contrast to the subjective. The term objective is a substantial misnomer as the corresponding priors are clearly not objective, although they may have objective intentions. More recently the term calibrate has been invoked, where the priors lead to results that have objective validity, in the sense of reproducibility or confidence. As such one might say that Laplace's strong support for certain mathematical priors was in substance pragmatic and thus an anticipation of confidence, Fisher (1935).

9 Confidence

Fisher (1935) in an article entitled "Inverse probability" wrote destructively of the Bayes approach (here §5), and introduced fiducial probability as an alternative. In its simplest form the model for a variable y can be given in distribution function form as say $u = F(y;\theta)$. The variable u under the model is known to have a uniform distribution on the interval $(0,1)$. Fisher then in effect substituted the observed y^0 obtaining $u = F(y^0;\theta)$ and allowed the Uniform distribution for u to backwardly induce a distribution for θ, to be called the fiducial distribution based on that model with data. The right tailed distribution function, the sur-

vivor function, for this fiducial distribution is then given as $s(\theta; y^0) = F(y^0; \theta)$. Neyman (1937) scolded Fisher for not following "the classical theory of probability", and then restricted the fiducial inversion to sets on the pivot space $\{u\}$, and renamed the resulting fiducial intervals as confidence intervals. The credit for confidence should truly go to Fisher but the renaming of the fiducial intervals somehow gave Neyman the credit for the fiducial/confidence methodology; nevertheless Neyman's point was well made. One can note that the probabilities called confidence are just those of a sub-algebra in the model data structure.

More generally the fiducial/confidence procedure begins with a pivotal quantity say $z = z(y, \theta)$ that has a fixed distribution under the model and then uses that distribution for z to backwardly induce a distribution for θ using $z = z(y^0, \theta)$. However appealing or unappealing the argument may be, the resulting confidence argument is now well established in the profession, provided one uses the Neyman pivotal sets. Recently the BFF conferences organized by Xiao-Li Meng at Harvard University are seeking some proper recognition for parameter distributions including fiducial distributions; the abbreviation BBF stands for Bayesian, Fiducial, and Frequentist but also for Best Friends Forever. Thus small social events deprived Fisher of the major credits for confidence.

10 Approximate statistical models

Likelihood and sufficiency (§6) when available can often provide a large reduction in the size of the data variable, and pivotal quantities (§9) can similarly provide a reduction as part of the formation of confidence intervals and regions. But the availability of these for particular models is unfortunately too rare.

In 1954 Daniels (1954) introduced the saddlepoint method to statistics from applied mathematics. The saddlepoint method can produce highly accurate approximations for many types of statistical model, in particular exponential models. An exponential model has the general form

$$f(y; \theta)dy = \exp\{\varphi^T(\theta)s(y) - k[\varphi(\theta)]\}H(y)dy. \tag{1}$$

It can be seen that only the parameter φ influences the distribution and then

only through the variable $s(y)$. Accordinly the effective model js

$$g(s;\varphi)ds = \exp\{\varphi^T s - k(\varphi)\}h(s)ds. \tag{2}$$

The importance of the saddlepoint approach was not recognized for some 25 years until Barndorff-Nielsen and Cox, (Barndorff-Nielsen and Cox, 1979) discussed a broad range of examples; in particular the exponential model (1) or (2) can be rewritten as

$$h(s;\varphi)ds = \frac{e^{k/n}}{(2\pi)^{p/2}} \exp\{-r^2/2\}|\jmath_{\varphi\varphi}(\hat{\varphi})|^{-1/2}ds \tag{3}$$

using the familiar statistical quantities: $r^2/2 = \ell(\hat{\varphi};s) - \ell(\varphi : s) = \hat{\ell} - \ell$ is the log-likelihood ratio quantity; $\hat{\varphi} = \hat{\varphi}(s)$ is the maximum likelihood estimate that maximizes the log-likelihood $\ell(\varphi;s)$; $\jmath_{\varphi\varphi}(\hat{\varphi})$ is the second derivative $-(\partial/\partial\varphi)^T(\partial/\partial\varphi)\ell(\varphi;s)|_{\hat{\varphi}}$ of the negative likelihood; and p is the dimension of the parameter. The use of the familiar statistical quantities gives the formula incredible power and applicability, essentially replacing sufficiency and pivotal quantities for statistical analysis. But even more importantly the corresponding distribution function is available Barndorff-Nielsen (1991) in the scalar parameter case with full third order accuracy,

$$H(s;\varphi) = \Phi[r - r^{-1}\log(\frac{1}{r} - \frac{1}{q})] = p(\varphi), \tag{4}$$

where $r = \text{sign}(\hat{\varphi}-\varphi)\{\hat{\ell}-\ell\}^{1/2}$ is the signed likelihood root and $q = (\hat{\varphi}-\varphi)\jmath_{\varphi\varphi}^{1/2}(\hat{\varphi})$ is a standardized Wald statistic. And more generally is available for a scalar parameter in the vector full parameter case with minor modifications; see §11. The distribution function says where the data s is with respect to the targetted parameter and does this in statistical % units, and is called the p-value function.

11 Null distribution for an interest parameter

A familiar problem in statistics is to find a statistic and its distribution in order to assess some hypothesis of interest. There is of course the initial problem of

finding in the applied context an appropriate variable that is sensitive to assessing the validity of the hypothesis. And then quite generally with an appropriate model the familiar sufficiency will be unavailable. As a widely familiar example consider the possible bending of light passing close to the sun; this has been of major scientific interest. In 1919 an opportunity arose with a near perfect eclipse of the sun. Measurements were made at two locations and there was a level of confirmation of Einstein's predictions.

For our considerations here we suppose that the relevant variable say y is available and a satisfactory model is also available say $f(y;\theta)$, and that the model supports a saddlepoint approximation as in (3). Then from likelihood theory it is known that for a fixed value of an interest parameter say $\psi(\theta)$ with dimension say d there is a conditional distribution that assesses the nuisance parameter say λ of dimension d, Also from likelihood theory it is known that Laplace integration can with high accuracy integrate out that condition distribution giving the following null distribution

$$h(s;\psi)ds = \frac{e^{k/n}}{(2\pi)^{p/2}} \exp\{-r^2/2\}|\jmath_{\varphi\varphi}(\hat{\varphi})|^{-1/2}|j^{1/2}_{\lambda\lambda}(\hat{\varphi}_\psi)|^{-1/2}ds, \tag{5}$$

where $\hat{\varphi}_\psi$ is the constrained maximum likelihood estimator given ψ; the variable s is now restricted to a $p-d$ dimensional plane through the observed data and perpendicular to ψ at the constrained maximum likelihood value $\hat{\varphi}_\psi$. And if ψ is scalar the distribution is available immediately on a line. This gives directly the variable and its distribution for assessing a value for ψ. A corresponding distribution function $H(s;\psi) = p(\psi)$ similar to (4) is available and with observed data inserted gives an observed p-value function for that parameter ψ; see for example Fraser (2014). This can be viewed as a resolution of the general problem of inference; see also Fraser (2017).

12 The Bayes direction

As mentioned in §5, many before and Fisher himself Fisher (1956) emphasized that making up a mathematical prior and using it in the conditional probability formula, in itself, would not produce probabilities. At most it could produce a statement that if θ-values had been sourced from the particular prior distribution

then there would be descriptive value to the posterior; it would also imply that with any other prior the conditional probabilities would not in general be valid. We have of course noted that Laplace was a supporter and that he sought other principles that could support posteriors in particular frameworks and that he perhaps was sensing confidence.

There have been many calls that any proposed Bayes posterior should be calibrated, or be reproducible, or to have repeated sample properties. One implicit supporter of this view was Jeffreys (Jeffreys, 1946) who proposed the prior $\pi(\varphi)d\varphi = |j_{\varphi\varphi}(\varphi)|^{1/2}d\varphi$ for an exponential model as at (2) and (3). This prior gives a probability measure that has parameterization invariance, a very attractive property. However later, Jeffreys (1961), (page 182) he noticed that in multiparameter problems the prior had unattractive properties and for location-scale and regression models he proposed an alternative that would supply appropriate degrees of freedom under marginalization. Even this prior proved unreliable; see Fraser (2011).

More recently for vector interest parameters with regular models, Fraser et al. (2014) find that only first order inference is generally available by Bayes, but for a scalar interest parameter say ψ full second order inference is available and is found fully on the profile contour for that parameter; in addition full second order information for that scalar parameter is obtained by using the full space Jeffreys but applying it just on the one dimensional profile contour for that parameter, a rather surprising but very useful result.

13 Discussion

Some recent conferences entitled BFF and organized by Xiao-Li Meng of Harvard University are directly concerned with current evolution in statistical inference. The BFF stands for Bayesian, Fiducial, and Frequentist and the conferences have a focus on unity in the theory; as a result they directly address current evolution in theory. In addition, they mention Best Friends Forever, thus emphasizing the concern for unity in a discipline that has a rather fractious history.

Certainly the Bayes approach provides an early and strong emphasis on theory and with having a broader base including mathematical and subjective priors in

addition to true or genuine priors. As such it has grounds for its frequent claim that it is in some sense a universal approach. The fiducial approach also leads to a distribution for the parameter based directly on the model and data. And the frequentist approach too can lead to a distribution for the parameter. Indeed the fiducial and frequentist approaches are almost equivalent, thus differing just in their early claims and reactions to those claims, with emphasis on the use of sets on the pivot space; and of course with who gets credit for confidence. The recent wide acknowledgement that the Bayes approach needs calibration indicates some level of convergence, perhaps to the point where Bayes can be viewed as offering just an exploratory approach to finding confidence, intervals and regions (Fraser et al., 2014).

14 Acknowledgements

This research has been supported by the Natural Sciences and Engineering Research Counsel of Canada and the Senior Scholars Funding of York University.

15 References

References

Barndorff-Nielsen. On the minimal sufficiency of the likelihood function. *Scand. J. Statistics 3*, 37–39.

Barndorff-Nielsen, O. E. (1991). Modified signed log likelihood ratio. *Biometrika 78*, 557–563.

Barndorff-Nielsen, O. E. and D. R. Cox (1979). Edgeworth and saddlepoint approximations with statistical applications. *J. R. Statist. Soc. B 41*, 187–220.

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Phil. Trans. Roy. Soc., London 53*, 370–418.

Daniels, H. E. (1954). Saddlepoint approximations in statistics. *Annals Math. Statist. 46*, 21–31.

Fisher, R. (1922). On the mathematical foundations of theoretical statistics. *Phil. Trans. Roy. Soc. London Series A 222*, 309–368.

Fisher, R. (1934). Two new properties of mathematical likelihood. *Proceedings of the Royal society A 144*, 285–307.

Fisher, R. (1935). The fiducial argument in statistical inference. *Annals of Eugenics 6*, 391–398.

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

Fraser, D. A. S. (2011). Is Bayes posterior just quick and dirty confidence? (with discussion). *Statistical Science 26*, 299–316.

Fraser, D. A. S. (2014). Why does statistics have two theories? In X. Lin, C. Genest, D. L. Banks, G. Molenberghs, D. W. Scott, and J.-L. Wang (Eds.), *Past, Present and Future of Statistical Science*, pp. 237–252. Florida: CRC Press.

Fraser, D. A. S. (2017). The p-value function: The core concept of modern statistical inference. *Annual Reviews of Statistics and its application 4*, 1–14.

Fraser, D. A. S., M. Bedard, W. Lin, A. Wong, and A. M. Fraser (2014). Can Bayes give second order reproducibility? *Statistical Science*.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proc. Roy. Soc. A 186*, 453–461.

Jeffreys, H. (1961). *Theory of Probability 3rd edn.* Oxford: Oxford University Press.

Kuhn, T. (1962). *The Structure of Scientific Revolutions*. The University of Chicago Press.

Laplace, P. S. d. (1812). *Théorie Analytique des Probabilités*. Paris: Courcier.

Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. *Philos. Trans. Roy. Soc. London. Ser. A 236*, 333 – 380.

Neyman, J. and E. S. Pearson (1933). On the problem of the most efficient tests of statistical hypotheses. *Phil. Trans. Roy. Soc. A 231*, 694–706.

Rozeboom, W. (1960). The fallacy of the null-hypothesis significance test. *Psychological Bulletin 57*, 416–428.

Savage, L. J. (1953). *The Foundations of Statistics*. New York: Dover.

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance, or vice versa. *J. Amer. Statist Assoc. 54*, 30–34.