International Statistical Review

International Statistical Review (2015), 83, 2, 293–308 doi:10.1111/insr.12067

On Some Principles of Statistical Inference

Nancy Reid[1] and David R. Cox[2]

[1]*Department of Statistics, University of Toronto, Toronto, Canada*
E-mail: reid@utstat.utoronto.ca
[2]*Nuffield College, Oxford, UK*
E-mail: david.cox@nuffield.ox.ac.uk

Summary

Statistical theory aims to provide a foundation for studying the collection and interpretation of data, a foundation that does not depend on the particular details of the substantive field in which the data are being considered. This gives a systematic way to approach new problems, and a common language for summarising results; ideally, the foundations and common language ensure that statistical aspects of one study, or of several studies on closely related phenomena, can be broadly accessible. We discuss some principles of statistical inference, to outline how these are, or could be, used to inform the interpretation of results, and to provide a greater degree of coherence for the foundations of statistics.

Key words: Ancillary; Bayesian; conditional; likelihood; models; *p*-values; sufficient.

1 Introduction

A healthy interplay between theory and application is crucial for statistics, as no doubt for other fields. This is particularly the case when by theory we mean foundations of statistical analysis, rather than the theoretical analysis of specific statistical methods. The very word *foundations* may, however, be a little misleading in that it suggests a solid base on which a large structure rests for its entire security. But foundations in the present context equally depend on and must be tested and revised in the light of experience and assessed by relevance to the very wide variety of contexts in which statistical considerations arise. It would be misleading to draw too close a parallel with the notion of a structure that would collapse if its foundations were destroyed.

The idea that the essence of all such general considerations can be captured within a simple framework, let alone a simple set of mathematical axioms, seems dangerously naive. See, for example, Fisher (1956, Ch. 5) for remarks on the need for a range of forms of statistical inference.

In what follows, we concentrate on formal issues connected with the assessment of uncertainty. There are many challenging aspects of statistical work that are not covered by this. We shall not in this essay discuss statistical decision theory, important though that is. We exclude also comments on prediction of future observations as contrasted with estimation of unknown parameters. These aspects are touched on in the concluding section.

We discuss first the role of probability, which is central to most but not all formulations of statistical issues; see for example Breiman (2001) for a more algorithmic emphasis. We then

discuss some of the classical concepts of statistical theory, some insights on principles that can be gained from asymptotic analysis and some thoughts on the relevance of these concepts for current developments in statistics and the analysis of data.

2 Role of Probability

Kolmogorov's axiomatisation of probability theory liberated the theory of probability from discussions of the meaning of probability, enabling it in particular to become a vibrant part of modern pure mathematics. Statisticians do not have the luxury of escaping such concerns with meaning: indeed, in a sense, most discussions of the last 200 years and more of the basis of statistical inference have centred around the relation between contrasting views of the meaning of probability.

Very particularly, statistical theory continues to focus on the interplay between the roles of probability as representing physical haphazard variability, what Jeffreys (1961) called *chances*, and as encapsulating in some way, directly or indirectly, aspects of the uncertainty of knowledge, often referred to as epistemic, or epistemological, probability.

2.1 Probability as Representing Empirical Variability

There are at least four related but different approaches to the connections between data and a target underlying the object of study:

- The data are regarded as a random sample from a hypothetical infinite population, frequencies within which are probabilities, some aspects of which encapsulate the target of inference.
- The data form part of a long real or more commonly somewhat hypothetical process of repetition under constant conditions, limiting frequencies in which are probabilities, again some aspects of which represent the target of inference.
- Either or both of the preceding items, combined with an explicit description in idealised form of the physical, biological, . . ., data generating process.
- Either or both of the first two approaches may be used solely to describe the randomisation used in experimental design or in sampling an existing population, leading to the so-called design-based analysis.

Fisher (1956, pp. 31–36) was emphatic that he intended the first of these, not the second. For discussions of, for example, climate change, the second or the third would be appropriate; the stochastic process of interest need not be required to be stationary.

It is important that the hypothetical population of the first approach, and the hypothetical repetition in the second approach, be recognised as idealisations, but this does not limit their usefulness. For the second approach, quantities may be defined by the procedures to be followed in measuring them, even if this measurement may be impracticable. For example, a geophysicist may contemplate the value of the acceleration due to gravity at sea level under Mount Everest: this has an operational definition, although the possibility of carrying out the operation is remote. In other applications, the direct notion of repetition may be so hypothetical as to be meaningless, for example, an analysis of literary style based on the complete works of Plato. The most satisfactory approach in such cases may be to hypothesise a data generating mechanism that produces observations *as if* from some physical probabilistic mechanisms. One may then hope that the underlying parameters of this notional data generating mechanism provide a summary of underlying properties of the real system free of certain accidental features.

The common feature of the first three approaches is that they represent features of the 'real' world, in a somewhat idealised form, and, given suitable data, are subject to empirical test and improvement. Conclusions of statistical analysis are in the first place expressed in terms of interpretable parameters describing such a probabilistic representation of the system under study.

We touch briefly in Section 3 on the very important principle of randomisation: one aspect of which is to provide an approach to inference useful for the somewhat specialised applications in which it is relevant.

2.2 Probability as Uncertain Knowledge

The form of probability outlined in the previous section is related to, but sharply different from, the consideration of probability as measuring uncertainty of knowledge about a specified proposition given incomplete information about it. In a statistical context, this might be expressed by the statement that an unknown parameter of interest lies in a specified range. There are at least three broad ways in which this issue can be addressed.

First, we may avoid the need for a different version of probability by appeal to a notion of calibration, as measured by the behaviour of a procedure under hypothetical repetition. That is, we study assessing uncertainty, as with other measuring devices, by assessing the performance of proposed methods under hypothetical repetition. Within this scheme of repetition, probability is defined as a hypothetical frequency. The precise specification of the assessment process does need care, often requiring some notion of conditioning. Secondly, probability may measure a rational, supposedly impersonal, degree of belief, given relevant information. This has a long history, the most notable account being that of Jeffreys (1961). Finally, probability may measure a particular person's degree of belief, subject typically to some constraints of self-consistency, an idea going back to F.P. Ramsey (1926) and developed to a refined level by de Finetti (1937) and Savage (1954). This approach seems intimately linked with personal decision-making. A broad-ranging view embracing all these perspectives was given by Good (1950).

The role of calibration seems essential: even if an empirical frequency-based view of probability is not used directly as a basis for inference; it is unacceptable if a procedure yielding regions of high probability in the sense of representing uncertain knowledge would, if used repeatedly, give systematically misleading conclusions.

The standard accounts of probability assume total ordering of probabilities. For some purposes, this may be reasonable, but for interpretation it seems unsound to regard a probability p found from careful investigation of a real-world effect as equivalent to a personal judgement based on scant or no direct evidence. That is, are the standard axioms of probability theory applicable when totally different types of evidence are mixed?

Personalistic approaches merge seamlessly what may be highly personal assessments with evidence from data possibly collected with great care. This may well be essential for personal decision-making but is surely unacceptable for the careful discussion of the data and the presentation of conclusions in the scientific literature. This is in no way to deny the role of personal judgement and experience in interpreting data; it is the merging that may be unacceptable.

Finally, a view that does not accommodate some form of model checking, even if very informally, is inadequate. Note very particularly that this includes mutual consistency of data and prior where a Bayesian formulation is used. Clear discrepancy may indicate a systematic flaw in the data, a mis-formulation of the statistical model or a misconception in formulating the prior. Priors that are consistent with all possible data configurations presumably play a merely formal role in the analysis.

A great attraction of Bayesian arguments, based as they are on the notion of probability as uncertain knowledge, is that all calculations are governed by the rules of probability theory. Another attractive feature, in principle at least, is the possibility of assimilating external evidence. While this is at the heart of personalistic approaches, many and perhaps most current applications of Bayesian methods rely explicitly or implicitly on some form of reference prior representing vague knowledge; these are also called objective, or non-informative priors. This is increasingly questionable as the dimension of the parameter space increases, as it leads to well-established difficulties with marginalisation and with calibration (Dawid *et al.*, 1973; Fraser, 2011). A very simple and striking example of this was put forward by Stein (1959): if $X \sim N(\mu, 1)$ is a d-dimensional random vector and we are interested in the parameter $||\mu||^2$, Bayesian marginal inference with a flat prior for μ is completely mis-calibrated, the error increasing with the dimension d (Stein, 1959; Cox & Hinkley, 1974, Ex. 2.39).

In principle, in most Bayesian arguments, the prior distribution aims to encapsulate all relevant information apart from that in the data under analysis; as such *prior* does not necessarily mean previous in time. Thus, particularly in studies that last for a long period, the prior may change from that used in planning the study and may be influenced either by the experience of collecting the data or even by the data themselves. As an extreme example, suppose the prior depends in part on a theoretical calculation of likely outcomes and a clear clash with that theory leads to the discovery of a mathematical mistake in the theory that, when corrected, resolves the discrepancy. The prior then depends on the data in a totally rational way. The assumption that the prior remains constant in time, which is typically not part of formal Bayesian theory, is called temporal coherency and has strong consequences; it will often, but not always, be reasonable. A more general comment about external or prior information is that the choice is not between Bayesian arguments that include it and non-Bayesian arguments that ignore it. Rather it is between including such information quantitatively by a probability distribution and merging it seamlessly with the data versus using it largely or entirely qualitatively.

An expansion of these comments is given by Cox (2006, Ch. 5). A lively discussion of calibration of Bayesian approaches is given from several points of view by Berger (2006), Goldstein (2006), Browne & Draper (2006) and extensive discussions. Wasserman (2008) considers this further, in the context of models and methods relevant for machine learning.

A non-Bayesian approach to interval estimation was set out by Fisher (1930) and, subject to some monotonicity conditions, leads for continuous distributions to a formal distribution for the unknown parameter, termed by Fisher a fiducial distribution. Indeed, a single such statement about a parameter and a single probability statement about an event seem evidentially essentially equivalent. The idea became controversial only later when such distributions were manipulated as probability distributions: Lindley (1958) showed this to be inappropriate in general. There is recently a renewal of interest in such approaches, variously described as generalised fiducial inference (Weerahandi, 1993; Hannig, 2009) and confidence distributions (Cox, 1958; Efron, 1993; Xie & Singh, 2013).

In summary, it seems necessary to recognise explicitly the dual role of probability in statistical inference, as representing variability and as a measurement of uncertainty. The conventional approach involving confidence intervals aims to use a hypothetical frequency-based probability in an epistemological sense; while this causes some tension, it seems on the whole to be broadly successful. At the other extreme, the approach that attempts to incorporate all forms of probability in a personalistic viewpoint does not seem to succeed in general scientific applications, even if it may sometimes be appropriate as a basis for personal decisions. A hybrid method of inference that uses Bayesian reasoning with impersonal priors, if the results are well calibrated in a frequency sense, may be the ideal, but to date, the construction of these priors is elusive.

We have not discussed the main alternative forms of axioms for probability, which concern, for instance, possible modifications needed in quantum mechanics, the development of upper and lower probabilities (Walley, 1990) and the development of belief functions, often called Dempster–Shafer theory; an overview of the last is available in Yager & Liu (2008); see also Hannig & Xie (2012).

3 Randomisation Inference

We now discuss briefly a somewhat different approach based not on a probabilistic model describing natural variability but rather on randomisation used in study design. We concentrate here on randomisation in experimental design; broadly parallel considerations arise in sampling a known population. By randomisation, we mean the use of an impersonal selection or allocation device with simple agreed probabilistic properties.

There are four broad reasons for randomising; their relative importance depends on the context. First, randomisation with appropriate concealment may be needed to avoid personal selection biases or measurement errors and to provide public assurance of such avoidance. Separate randomisations may be needed at different stages of an investigation.

A second and somewhat different issue of avoidance of bias aims to eliminate the systematic influence of confounders, that is, explanatory variables that are prior to the treatments under comparison and may affect the outcome.

Third, randomisation combined with a non-stochastic assumption of unit-treatment additivity may be used to justify the estimation of standard errors in many experimental designs. That is, it indicates a default analysis of variance for standard experimental designs. For example, it shows that the analyses of data from, for example, Latin squares, randomised blocks and balanced incomplete block designs, do not require distinct assumptions about the physical structure of error for each design.

Finally, randomisation may be used to justify an exact test of a null hypothesis that the outcomes are unaffected by treatment allocation. By extension, non-parametric confidence intervals can be obtained for parameters representing simple effects, such as changes in location or in scale.

In contexts where conscious or unconscious selection biases are possible in treatment allocation or implementation or in the measurement of outcome, the first of the aforementioned roles of randomisation may be of crucial importance. The second is one of the distinguishing features of the distinction between experiments and broadly analogous observational studies aimed at establishing some form of causality.

The third role provides an elegant and uniform approach to the analysis of many simple and not-so-simple experimental designs without *ad hoc* assumptions specific to each design.

The fourth role was emphasised by Fisher at an early stage of his discussions of experimental design, in particular to answer criticisms that his parametric procedures might be too sensitive to a normality assumption. For a period around the late 1930s, there was some discussion on whether the randomisation distribution, although at the time computationally not feasible, was the 'ideal solution', or whether it merely provided some security to, and indicated the form of, the related normal-theory-based analysis. This debate is in some sense unresolvable, as the two approaches are unlikely to give very different results except in quite small samples, where either approach is likely to be problematic. There are echoes in recent work on bootstrap-based inference, and in particular in a body of work showing that bootstrap inference is in moderate-sized samples typically very close to model-based inference (e.g. DiCiccio & Efron, 1996; Horowitz, 2001; DiCiccio & Young, 2008).

Randomisation tests should be distinguished from the numerically identical permutation tests. The latter are based on an assumed probabilistically formulated structure for the data generating procedure involving independent and identically distributed random variables, whereas randomisation tests are based on the design procedure.

While in some situations, failure to randomise may be a serious defect of design, there are others where it is unwise. Randomisation may entail complex administrative steps in allocating material according to a detailed and ever-varying set of instructions, which will be destructive to good control. For the probability distribution over the randomisation to be a reasonable basis for interpretation, the design used has to be not special in some relevant sense. This restricts the usefulness of randomisation in small experiments in which each arrangement is likely to have distinctive features.

The immediate formal inference from randomisation theory is restricted to the particular units in the experiment, except in the unlikely possibility that these were selected in a suitably specified way from a larger population of units. That is, there might be clear evidence that in aggregate, the particular group of patients would have had better prognosis if all had received treatment A than they would have had all received B, or a whole particular area treated with fertiliser C would have had a higher yield than it would have had the whole received fertiliser D. In Fisher's view, the conclusions would apply also to the hypothetical infinite population of individuals from which the patients or plots had been drawn at random, but the specific implications of this are unclear. This is connected to the subtly interconnected issues of generalisability and specificity. That is, in the first setting, do the conclusions apply to a different population of patients, and secondly, why should they be relevant to a specific patient being advised by a clinician? Answers to both include demonstration of the absence of interaction of the treatment effect with key features and understanding of the fundamental underlying pharmacology or soil science, respectively. A discussion of some of these issues, with emphasis on clinical trials, is given by Zheng & Zelen (2008).

4 Simple Test of Significance

While discussions of the meaning of probability have proven difficult to resolve, there is more widespread agreement on the importance of some statistical concepts that serve as a basis for development of statistical theory, even if there is some disagreement about how the principles should be implemented.

The great majority of the formal discussion is based on the specification that there is a family of probability models, one of which has, to an adequate approximation, generated the data under analysis. We start, however, from a more primitive viewpoint, namely that we have a null hypothesis, H_0, that specifies numerically the distribution of either the full data or certain aspects of the data. We wish to examine consistency with that null hypothesis. Further, we suppose a chosen test statistic, $t(y)$, such that the larger its value, the stronger the discrepancy of concern and such that the distribution of the random variable T under H_0 is known.

In other words, we specify, largely qualitatively, the type of departure from H_0 of potential interest; any monotonic function of T would be equivalent. To assess consistency with H_0, we have an observed value of t, a probability distribution for T were H_0 to be true and the specification that the larger the t, the poorer the consistency. There seems little choice in this formulation but to use the p-value, that is,

$$p(t) = P(T \geq t; H_0). \tag{1}$$

If this is a modest number, the data are as consistent with H_0 as could reasonably be expected. If p is small, it is suggestive of inconsistency with H_0 in the direction indicated by large values of T. The observed value $p(t)$ can be given the hypothetical interpretation that if the observations were regarded as just decisive against H_0, then $p(t)$ would be the long-run proportion of times in which H_0 would be falsely rejected when true.

There are two broad situations in which this formulation may be relevant. In one, H_0 is a subject-matter hypothesis, suggested perhaps by theory, that may to a reasonable approximation be true. The other is where adequacy of a formal model, itself forming H_0, is to be assessed.

This is conceptually quite different from, although formally related to, other formulations, such as that of Neyman–Pearson theory, Bayesian testing theory and formal two-decision problems. A parallel can be established by extending the null hypothesis distribution into an exponential family form with a factor $\exp(t\lambda)$ (Cox, 2006, §3.5), but this may seem very contrived, particularly in testing model adequacy.

There is a substantial literature on the interpretation and misinterpretation of p-values. One criticism, also applied to confidence intervals, is that these are widely misinterpreted to represent epistemic probabilities, although the increase in general statistical literacy seems to have assuaged this concern somewhat. Another is that p-values may become meaningless in the context of testing very large numbers of similar hypotheses, common now in many fields of application. This problem is being addressed through a number of methods generally related to the notion of false discovery rates (Storey, 2002; Cox & Wong, 2004; Efron, 2010).

5 Classical Principles for Inference

We from now on assume that a probability model, in the form of a distribution function $F(y;\theta)$ or a density function $f(y;\theta)$, has been formulated; that θ ranges over a space Θ, leading to a family of such models; and that data that are provisionally assumed to follow some member of the family of probability models have been or are to be observed. These are formidable assumptions from an applied viewpoint. McCullagh (2002) emphasises the importance of careful delineation of design, covariate and treatment variables as an essential part of the correct specification of a statistical model. We do not consider here models in which the parameter is an unspecified function, and hence infinite dimensional.

The *sufficiency principle* supposes that there is a factorisation of the model of the form

$$f(y;\theta) \propto f_1(s;\theta) f_2(y \mid s), \tag{2}$$

with minimal s. The first and most commonly emphasised part of the principle is that inference about θ should be based on the statistic $s = s(y)$, which is sufficient for θ in this model. The second part is that the conditional distribution of the data, given s, being a fixed and known distribution, is available for assessing model adequacy, for example, in the way outlined in the previous section.

The *conditionality principle* states that if the minimal sufficient statistic can be split into components (s_1, s_2) such that there is a factorisation of their joint distribution of the form

$$f(s;\theta) \propto f_1(s_1 \mid s_2;\theta) f_2(s_2), \tag{3}$$

then inference about θ should be based on the conditional distribution of s_1, given the *ancillary statistic* s_2.

The *likelihood principle* states that inference should be based on the likelihood function, more precisely, the equivalence class of functions of θ determined by the model, in which the observed data are fixed:

$$L(\theta) \propto f(y;\theta). \tag{4}$$

We take this in the strong form that only the directly observed likelihood is relevant, thus excluding dependence on the sampling distribution of statistics derived from the likelihood function.

5.1 Sufficiency

The primary role of sufficiency is essentially that of simplification by dimension reduction; it enables inference to proceed based on a reduction of the set of observed or observable values to a potentially much smaller number of quantities, without loss of information. The interpretation of a sufficient reduction as giving a direct partitioning of the sample space is outlined in many textbooks (e.g. Cox & Hinkley, 1974, Ch. 2; Lehmann & Romano, 2005, Ch. 1). Sufficiency is closely tied to the theory of exponential families, as in general, these are the models that permit substantial dimension reduction via sufficiency. A mathematical discussion of the sufficiency of the *likelihood map*, that is, the equivalence class of functions $L(\theta;\cdot)$, is given by Barndorff-Nielsen *et al.* (1976) and extended by Fraser *et al.* (1997) and Fraser & Naderi (2007).

5.2 Conditionality

Conditionality is in a sense the least technical of the principles and at the same time the most elusive to formulate. The motivation is that if we want to calibrate methods of statistical analysis by their performance in hypothetical repetition, it is important that the repetitions match in some sense the very particular set of data under analysis. This demands conditioning on features that might distinguish in some important respect the ensemble of repetitions from the data; these are sometimes called 'relevant subsets'. The most important example of this idea is to normal-theory linear models in which also the explanatory variables have a probability distribution. Ancillarity shows that under rather general circumstances, inference about the regression parameters should be conditional on the observed values of the explanatory variables. Another important application is to the class of transformation models, in which a unique ancillary statistic can be obtained by considerations of invariance, and the conditional inference in such models was called structural inference by Fraser (1961, 1968). In a few models, non-uniqueness in the choice of ancillary statistics has to be resolved by somewhat *ad hoc* criteria.

5.3 Likelihood Principle

This is formulated in (4) in its strongest form: the data should be used only in terms of the observed likelihood function. The only inferences that are consistent with that likelihood principle are a non-probabilistic use of the likelihood function as defining regions of the parameter space that are more or less likely, or Bayesian inference, which derives its probabilities via the prior distribution. Accounts of the former are given by Edwards (1960) and especially Royall (1997) and several subsequent publications.

If there are no nuisance parameters, the non-probabilistic approach indicates, for example, graphical summarisation of the data by plotting the likelihood function by a curve or contour

plot. This is ineffectual, however, if there are nuisance parameters, particularly if there are many such.

The use of the strong likelihood principle in the development of Bayesian inference is discussed, with many examples, by Berger & Wolpert (1984). One point of interest they noted is that Bayesian approaches with priors based on model characteristics, that is, most 'non-informative' priors, are not consistent with the strong likelihood principle.

Conditionality does not arise as a specific issue, because inference is conditioned on the full data, and sufficiency is automatically incorporated, as the likelihood function depends on the data only through the sufficient statistic.

5.4 General Comments

In nearly all applied work, the parameter θ will be comprised of parameters of direct interest to the problem at hand, and additional parameters typically representing aspects of secondary interest; for example, the parameters of interest may govern the mean response, possibly as a linear or non-linear function of some auxiliary variables, while secondary parameters might be related to the variability, and/or other aspects of the distribution such as the shape, or tail weight, or other features relevant to the problem. Such parameters may be essential to complete the specification, but not be themselves the focus of subject-matter concern. Different phases of the analysis of a single set of data may well involve different choices of the parameter of interest.

In its simplest form, we may write $\theta = (\psi, \lambda)$, with ψ as the parameters of interest and λ usually referred to as nuisance parameters. Unfortunately, the definitions of sufficiency and ancillarity for ψ immediately become more difficult, because it is rarely the case that factorisations analogous to (2) and (3) can be obtained. In the ideal case, where

$$f(y; \psi, \lambda) \propto f_1(s; \psi) f_2(t \mid s; \lambda),$$

or possibly $f(y; \psi, \lambda) \propto f_1(s; \psi) f_2(t \mid s; \eta)$, where $\eta = \eta(\psi, \lambda)$ and the parameter spaces for (ψ, η) and (ψ, λ) are the same, inference for ψ can be cleanly based on the model f_1; s is sufficient for ψ and ancillary for λ. This ideal case rarely obtains, more usually, either

$$f(y; \psi, \lambda) \propto f_1(s; \psi) f_2(t \mid s; \psi, \lambda) \quad \text{or} \quad f(y; \psi, \lambda) \propto f_1(s; \psi, \lambda) f_2(t \mid s; \psi),$$

and the situation is much less clear. Some aspects of this are discussed in more detail by Reid (1995).

Bayesian methods, being based on the observed data, avoid this consideration, but at the expense of specification of prior probabilities for a possibly large number of parameters, which entails another set of difficulties. Subjective information, the relevance of which we have argued against in Section 2, will in any case rarely be available for complex models with large numbers of parameters, at least not in the scientific context. The extensive development of reference priors and other forms of priors meant to be uninformative with respect to the parameters clearly indicates that such priors must be targeted to the parameters of interest (Berger & Bernardo, 1992; Berger *et al.*, 2009; Fraser *et al.*, 2010).

The confidence distributions briefly mentioned in Section 2.2 are typically obtained by inversion of a pivotal quantity, which is a function of the data y and parameter of interest ψ, with a known distribution. Using this known distribution enables us to obtain a set of p-values for different values of ψ, variously called a significance function or p-value function. Slightly more generally, a set of confidence regions at various confidence levels can be used to define a

confidence distribution for ψ (Cox, 1958); in nearly all treatments, these regions are assumed to be nested. The usual t-statistic of normal theory is a simple example of a pivot leading to a p-value function or providing a set of nested confidence intervals for the unknown mean of a normal distribution. The Fieller pivotal quantity for inference about the ratio of the means of two independent normal distributions, $(\bar{y}_1 - \psi\bar{y}_2)/\{\sigma_0\sqrt{(1/n_1 + \psi^2/n_2)}\}$, where σ_0^2 is the common known variance, is another example, which is however non-monotone in ψ so does not lead directly to a confidence distribution. Anomalies like this, and the lack of a general recipe for constructing pivotal quantities, have meant that they have received somewhat less attention in studies of theoretical statistics. There is a recent revival of interest in confidence distribution functions; see Xie & Singh (2013), Hannig (2009) and Schweder & Hjort (2002) for overviews and further references. For most problems, the notion of an approximate pivotal quantity is needed, and these can be obtained from asymptotic theory, to which we turn next.

6 Principles and Asymptotic Theory

Consideration of distributions of inferential quantities as a notional sample size or amount of information increases both simplifies and complicates the discussion. While asymptotic theory is often viewed as a means of generating approximate inference, for a general theoretical discussion, it is perhaps more important for the insight it gives into some foundational aspects.

Notions of approximate sufficiency and approximate ancillarity have been developed (e.g. Cox, 1980; McCullagh, 1984; as well as Barndorff-Nielsen & Cox, 1994, Ch. 7), but in general, the details are relatively complex.

In contrast, approximate pivotal quantities are used nearly routinely in applied work, thanks in part to the development of robust software for optimisation and root finding. So, for example, letting $\hat{\lambda}_\psi$ denote the maximum likelihood estimator of the nuisance parameter λ when the parameter of interest ψ is fixed and defining the profile log-likelihood function by $\ell_{\mathrm{p}}(\psi) = \log L\left(\psi, \hat{\lambda}_\psi\right)$, the standardised maximum likelihood estimator

$$\left(\hat{\psi} - \psi\right) j_{\mathrm{p}}^{1/2}\left(\hat{\psi}\right),$$

where $j_{\mathrm{p}}(\psi) = -\partial^2\ell_{\mathrm{p}}(\psi)/\partial\psi\partial\psi'$, is an approximately pivotal quantity, as its asymptotic distribution is known to be, under suitable regularity conditions, normal with mean 0 and covariance matrix the identity. Similarly,

$$r^2(\psi) = 2\left\{\ell_{\mathrm{p}}\left(\hat{\psi}\right) - \ell_{\mathrm{p}}(\psi)\right\} \tag{5}$$

is an approximate pivotal quantity following a χ^2_d distribution, where d is the dimension of ψ. Either or both of these can be inverted to give confidence regions for ψ at any desired level of confidence.

Improved approximations can be developed from a more detailed study of the asymptotic expansions involved, and when the parameter of interest is a scalar, an improved version of (5) is

$$r^*(\psi) = r(\psi) + \frac{1}{r(\psi)}\log\left\{\frac{Q(\psi)}{r(\psi)}\right\}, \tag{6}$$

where $r(\psi)$ is the square root of (5), with the appropriate sign attached, and $Q(\psi)$, discussed briefly later, is a related pivotal quantity with the property that it has the same limiting distribution as $r(\psi)$, that is, standard normal. In continuous models, the distribution of $r^*(\psi)$, under the model $f(y;\theta)$ is also standard normal, but with a relative error of $O(n^{-3/2})$ in terms of the sample size for independent observations from the model, whereas the relative error in (5) is $O(n^{-1/2})$. In other words, (6) is a large deviation result: the practical implication of this is that the approximation often works very well in the tails of the distribution, where small p-values are of interest. The inferential basis for using the pivotal quantity $r^*(\psi)$ is that it approximates the signed square root of the log-likelihood ratio statistic for an approximate conditional or marginal log-likelihood function; these latter are obtained by implementing the ideas of approximate sufficiency and ancillarity mentioned earlier.

A similar asymptotic analysis of the marginal posterior distribution in a Bayesian analysis leads to

$$r_B^*(\psi) = r(\psi) + \frac{1}{r(\psi)} \log\left\{\frac{Q_B^\pi(\psi)}{r(\psi)}\right\}, \tag{7}$$

where $Q_B^\pi(\psi)$ depends on the prior π, as well as the first and second derivatives of the log-likelihood function. The distribution of $r_B^*(\psi)$, in the posterior distribution $\pi(\theta \mid y) \propto f(y;\theta)\pi(\theta)$, is standard normal with a relative error of $O(n^{-3/2})$ (DiCiccio *et al.*, 1990).

The approximately pivotal quantity $Q_B^\pi(\psi)$ is

$$Q_B^\pi(\psi) = \ell_{\mathrm{p}}'(\psi) j_{\mathrm{p}}^{-1/2}(\hat\psi) \left\{\frac{\left|j_{\lambda\lambda}\left(\psi, \hat\lambda_\psi\right)\right|}{\left|j_{\lambda\lambda}\left(\hat\psi, \hat\lambda\right)\right|}\right\}^{1/2} \frac{\pi\left(\hat\psi, \hat\lambda\right)}{\pi\left(\psi, \hat\lambda_\psi\right)}, \tag{8}$$

where $j_{\lambda\lambda}(\theta) = -\partial^2 \log L(\psi,\lambda)/\partial\lambda\partial\lambda^{\mathcal{T}}$ is the sub-matrix of the full Fisher information matrix corresponding to the nuisance parameter λ, and, as in previous discussion, $\hat\lambda_\psi$ is the constrained maximum likelihood estimator of λ when ψ is fixed. The factor in braces in (8) comes from integrating out the nuisance parameters by Laplace approximation.

The approximately pivotal quantity $Q(\psi)$ in (6) is more difficult to describe, as it depends in general on the construction of an approximately ancillary statistic. A number of examples are given by Brazzale *et al.* (2008, Chs. 3–7), where asymptotically equivalent versions due to Barndorff-Nielsen (1990) and Fraser (1990) are presented and discussed. In exponential family models, with densities of the form

$$f(y;\psi,\lambda) = \exp\{s_1(y)\psi + s_2^{\mathcal{T}}(y)\lambda - c(\psi,\lambda)\}h(y), \tag{9}$$

the expression for $Q(\psi)$ is the standardised maximum likelihood estimator of ψ, with a nuisance parameter adjustment:

$$Q(\psi) = \left(\psi - \hat\psi\right) j_{\mathrm{p}}^{1/2}\left(\hat\psi\right) \left\{\frac{\left|j_{\lambda\lambda}\left(\hat\psi, \hat\lambda\right)\right|}{\left|j_{\lambda\lambda}\left(\psi, \hat\lambda_\psi\right)\right|}\right\}^{1/2}. \tag{10}$$

In regression-scale models, $y_i = x_i^{\mathcal{T}}\beta + \sigma e_i$, with ψ a component of β, $Q(\psi)$ is the standardised score statistic for ψ, modified by a similar adjustment for nuisance parameters. Explicit formulae for Q in a number of regression settings are given by Brazzale *et al.* (2008, Ch. 8).

A detailed study of these approximations leads to a number of insights into foundational aspects.

As $n \to \infty$, the Bayesian and frequentist inferences for ψ are the same, assuming the prior is fixed. This has long been known, sometimes described as the prior being 'washed out' by the data. The point of departure between Bayesian and frequentist inference appears at the next order of approximation. This was discussed from a slightly different point of view by Welch & Peers (1963) and articulated in the context here by Pierce & Peters (1994).

A prior that leads to inferences equivalent to frequentist inferences at this higher order of approximation must satisfy $Q_B^{\pi}(\psi) = Q(\psi)$. These priors are called strong matching priors by Fraser & Reid (2002). These strong matching priors are specific to the parameter of interest, suggesting that any prior that is calibrated in this sense for ψ is unlikely to be calibrated for other components of θ. This also follows from Peers (1965), who considered the extension to nuisance parameter of the results of Welch & Peers (1963). The need to target the prior on the parameter of interest is emphasised in the literature on reference priors (Berger & Bernardo, 1992).

The addition of the approximately pivotal quantity $Q(\psi)$ via (6) means inference is based on more than the profile log-likelihood function. In particular, this quantity includes an adjustment for the estimation of the nuisance parameters, and this is, in practical problems, much more important for the accuracy of the inference than the distributional improvement (Pierce & Peters, 1992). Another key step in the construction of the approximate pivotal $Q(\psi)$ in (6) is measuring the change of the log-likelihood function $\log L(\theta; y)$ with small changes in the data, keeping relevant ancillary or approximately ancillary statistics fixed. As might be expected, this requires that the parameter space and the sample space both be continuous; a slightly different argument is needed for discrete data.

There is no need to work with sufficient statistics in deriving formulae like (6): because it is based on functions of the log-likelihood function, it is automatically a function of the sufficient statistic, although sometimes the calculations are easier after a preliminary reduction by sufficiency. It is however imperative to condition on an exactly or approximately ancillary statistic, except in the case of linear exponential family models (9).

It is shown by DiCiccio & Young (2008), building on work by DiCiccio *et al.* (2001), that to $O(n^{-3/2})$, parametric bootstrap sampling of $r(\psi)$ under the model $f(y; \psi, \hat{\lambda}_{\psi})$ is equivalent to that based on (6) (see also Fraser & Rousseau, 2008). However, the number of replications required to estimate small tail probabilities may be prohibitive.

A development of model checking, using for example $f(t \mid s)$ when factorisation (2) holds, based on these notions of higher-order approximation is to our knowledge not yet available.

7 Discussion

There are several important aspects of statistical science that we have not emphasised here. While many investigations involve a decision-making element, most commonly, the role of statistics is to summarise evidence in a clear and cogent form, rather than to make irrevocable decisions. For example, data may be consistent with two quite different interpretations, indicating the appropriateness of two different decisions. Statistical analysis may end with an indication of the possibilities and associated uncertainties: decision analysis ends with the choice of a single decision, even if this is essentially arbitrary. The best action in such cases may be a search for a third decision, so far overlooked, as the formulations of general theory are rarely really closed. Further, formal treatments of decision theory are typically based on maximising expected utility, whereas Simon's (1956) notion of *satisficing* may be more appropriate,

especially when more than one individual is involved. This is because of the general fragility of formulations that are intrinsically and strongly personalistically based.

We have excluded comments on prediction of future observations as contrasted with estimation of unknown parameters. In Bayesian discussions, there is no formal distinction in that the objective is to find the conditional distribution of the feature of interest given the data and the prior. In frequentist theory too, a formal parallel can be established with testing the consistency of potential future data with the current data. In most formulations, it is assumed that the values to be predicted are generated from the same random system as the data, often a formidable assumption. Stability under perturbation of the generating process may be more important than formal optimality.

We have also not discussed non-parametric methods. If an issue cannot be addressed non-parametrically, a parametric analysis is likely to be hazardous. An illustration is the study of competing risks in the analysis of survival data. The assumption of independent competing risks cannot be tested by non-parametric methods, and any analysis of this by parametric models is likely to be heavily influenced by the particular model family chosen. On the other hand, if a non-parametric analysis is in principle available, a parametric formulation is likely to give a more focused interpretation, and that is our reason for concentrating on parametric issues.

We emphasised in Section 1 that foundations must be continually tested against applications. From this perspective, the strong likelihood principle is found wanting: a great deal of applied work relies on the distribution of quantities based on the likelihood function, such as the maximum likelihood estimator or the likelihood ratio statistic. Similarly, a great deal of applied work with Bayesian methods uses what are hoped to be 'non-influential' priors; the question is whether these really are non-influential, particularly when high-dimensional parameters are involved.

Many applications of now current statistical ideas involve vast amounts of data, or highly complex models, or both, and the question of whether the principles touched on here continue to be relevant to these settings arises. A principled approach is surely necessary to avoid apparent discoveries based on spurious patterns or correlations. While there are a number of applied contexts, many involving machine learning, where prediction and classification using possibly complex black-box approaches are adequate, for any analysis that hopes to shed light on the structure of the problem, modelling and calibrated inferences about interpretable parameters seem essential.

A recent report (National Research Council, 2013) highlighted the following 'inferential giants' for the study of massive data: assessment of sampling bias, inference about tails, resampling inference, change-point detection, reproducibility of analyses, causal inference for observational data and efficient inference for temporal streams. Sampling bias is an essential aspect of design and analysis of surveys and experiments, topics that we have touched on only briefly, and efficient inference for temporal streams is perhaps mainly an issue of computation. However, theoretical statistics, and the classical concepts discussed earlier, would seem to be important for the remainder. For example, the ideas behind significance testing underlie the development of false discovery rates, and other methods for judging the importance of seemingly large effects when a great many comparisons have been carried out. Sufficiency, or something much like it, is needed for successful implementation of approximate Bayesian computation, which uses simulation to construct the likelihood function in very complex models.

One issue arising when assessment of precision is required from large-data analysis concerns internal correlations and undetected sources of variability, leading to serious underestimation of potential errors if relatively standard methods are used with their attendant strong independence assumptions. Another issue is the information available in large observational data sets, often

collected in such a way that assessment of selection bias is difficult or impossible. Lazer *et al.* (2014) describe a setting where seemingly big data relied for its analysis on what was effectively a small data set, which led to overfitting.

There are also broader strategical issues. How best should a wholly new large set of data be approached? A summary analysis of the whole may be combined with a very detailed analysis of suitably sampled fragments. There are in a real sense theoretical issues involved, although ones possibly not easily captured within a mathematical formalism. The role of principles for inference is to enable abstraction from a range of particular problems to a common underlying theme, and hence to aid the approach to new challenges.

Acknowledgements

This paper is based on a talk given at the World Statistics Congress of the International Statistical Institute in Hong Kong, August 2013. N. R.'s research is partially supported by the Natural Sciences and Engineering Research Council of Canada. We are grateful to reviewers of an earlier version for helpful comments.

References

Barndorff-Nielsen, O.E. (1990). Approximate interval probabilities. *J. Roy. Stat. Soc. B*, **52**, 485–496.
Barndorff-Nielsen, O.E. & Cox, D.R. (1994). *Inference and Asymptotics.* London: Chapman & Hall.
Barndorff-Nielsen, O., Hoffmann-Jorgensen, J. & Pedersen, L. (1976). On the minimal sufficiency of the likelihood function. *Scand. J. Statist.*, **3**, 37–38.
Berger, J.O. (2006). The case for objective Bayesian analysis. *Bayesian Stat.*, **3**, 385–402.
Berger, J.O. & Bernardo, J.M. (1992). On the development of reference priors (with discussion). In *Bayesian Statistics 4*, Eds. J.M. Bernardo, J.O. Berger, A.P. Dawid & A.F.M. Smith. pp. 35–60. Oxford: Oxford University Press.
Berger, J.O., Bernardo, J.M. & Sun, D. (2009). The formal definition of reference priors. *Ann. Stat.*, **37**, 905–938.
Berger, J.O. & Wolpert, R.L. (1984). *The Likelihood Principle.* Hayward: Institute of Mathematical Statistics.
Brazzale, A.R., Davison, A.C. & Reid, N. (2008). *Applied Asymptotics.* Cambridge: Cambridge University Press.
Breiman, L. (2001). Statistical modelling: The two cultures. *Stati. Sci.*, **16**, 199–231.
Browne, W.J. & Draper, D. (2006). A comparison of Bayesian and likelihood-based methods for multi-level models. *Bayesian Anal.*, **3**, 473–514.
Cox, D.R. (1958). Some problems with statistical inference. *Ann. Math. Stat.*, **29**, 357–372.
Cox, D.R. (1980). Local ancillarity. *Biometrika*, **67**, 269–276.
Cox, D.R. (2006). *Principles of Statistical Inference.* Cambridge: Cambridge University Press.
Cox, D.R. & Hinkley, D.V. (1974). *Theoretical Statistics.* London: Chapman & Hall.
Cox, D.R. & Wong, M.Y. (2004). A simple procedure for the selection of significant effects. *J. Roy. Stat. Soc. B*, **66**, 395–400.
Dawid, A.P., Stone, M. & Zidek, J.V. (1973). Marginalization paradoxes in Bayesian and structural inference. *J. Roy. Stat. Soc. B*, **35**, 189–233.
de Finetti, B. (1937). La prévision: ses lois logiques, ses sources subjectives. *Ann. de l'Inst. Henri Poincar*, **7**, 1–68. (Trad.) Foresight, its logical laws, its subjective sources. In: Studies in Subjective Probability (1964, 97–156), ed. H. E. Kyburg & H. E. Smokler. Wiley, New York.
DiCiccio, T.J. & Efron, B. (1996). Bootstrap confidence intervals. *Stat. Sci.*, **11**, 189–212.
DiCiccio, T.J., Field, C.A. & Fraser, D.A.S. (1990). Approximations of marginal tail probabilities and inference for scalar parameters. *Biometrika*, **77**, 77–95.
DiCiccio, T.J., Martin, M.A. & Stern, S.E. (2001). Simple and accurate one-sided inference from signed roots of likelihood ratios. *Canad. J. Stat.*, **29**, 67–76.
DiCiccio, T.J. & Young, G.A. (2008). Conditional properties of unconditional parametric bootstrap procedures for inference in exponential families. *Biometrika*, **95**, 747–758.
Edwards, A.W.F. (1960). *Likelihood.* Oxford: Oxford University Press.
Efron, B. (1993). Bayes and likelihood calculations from confidence intervals. *Biometrika*, **80**, 3–26.

Efron, B. (2010). *Large-scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction.* Cambridge: Cambridge University Press.
Fisher, R.A. (1930). Inverse probability. *Proc. Cam. Phil. Soc.*, **26**, 528–535.
Fisher, R.A. (1956). *Statistical Methods and Scientific Inference.* Edinburgh: Oliver & Boyd.
Fraser, D.A.S. (1961). The fiducial method and invariance. *Biometrika*, **48**, 261–280.
Fraser, D.A.S. (1968). *The Structure of Inference.* New York: Wiley.
Fraser, D.A.S. (1990). Tail probabilities from observed likelihoods. *Biometrika*, **77**, 65–76.
Fraser, D.A.S. (2011). Is Bayes just quick and dirty confidence? *Stat. Sci.*, **26**, 299–316.
Fraser, D.A.S., McDunnough, P., Naderi, A. & Plante, A. (1997). From the likelihood map to Euclidean minimal sufficiency. *Probability and Mathematical Statistics*, **17**, 223–230. Available at: http://www.utstat.toronto.edu/dfraser/documents/195.pdf, accessed on September 18, 2013.
Fraser, D.A.S. & Naderi, A. (2007). Minimal sufficient statistics emerge from the observed likelihood function. *Int. J. Stat. Sci.*, **6**, 55–61. Available at http://www.utstat.toronto.edu/dfraser/documents/238.pdf, accessed on September 18, 2013.
Fraser, D.A.S. & Reid, N. (2002). Strong matching of frequentist and Bayesian parametric inference. *J. Stat. Plann. Infer.*, **103**, 263–285.
Fraser, D.A.S. & Rousseau, J. (2008). Studentization and deriving accurate p-values. *Biometrika*, **95**, 1–16.
Fraser, D.A.S., Reid, N., Marras, E. & Yi, G.Y. (2010). Default priors for Bayesian and frequentist inference. *J. Roy. Stat. Soc. B*, **72**, 631–654.
Goldstein, M. (2006). Subjective Bayesian analysis: Principles and practice. *Bayesian Anal.*, **3**, 403–420.
Good, I.J. (1950). *Probability and the Weighing of Evidence.* Cambridge: MIT Press.
Hannig, J. (2009). On generalized fiducial inference. *Stat. Sinica*, **19**, 491–544.
Hannig, J. & Xie, M. (2012). A note on Dempster–Shafer recombination of confidence distributions. *Elec. J. Stat.*, **6**, 1943–1966.
Horowitz, J. (2001). The bootstrap. In *Handbook of Econometrics*, Vol. 5, Eds. Heckman, J.J. & Leamer, E.; 3159–3228.
Jeffreys, H. (1961). *Theory of Probability,* 3rd ed. Oxford: Oxford University Press.
Lazer, D., Kennedy, R., King, G. & Vespignani, A. (2014). The parable of Google flu: Traps in big data analysis. *Science*, **343**, 1203–1205.
Lehmann, E.L. & Romano, J.P. (2005). *Testing Statistical Hypotheses,* 3rd ed. New York: Springer.
Lindley, D.V. (1958). Fiducial distributions and Bayes' theorem. *J. Roy. Stat. Soc. B*, **20**, 102–107.
McCullagh, P. (1984). Local sufficiency. *Biometrika*, **71**, 233–244.
McCullagh, P. (2002). What is a statistical model? (with discussion). *Ann. Stat.*, **30**, 1225–1310.
National Research Council. (2013). *Frontiers in Massive Data Analysis.* Washington: National Academies Press. Available at; http://www.nap.edu/catalog.php?record_id=18374, accessed on September 24, 2013.
Peers, H. (1965). On confidence points and Bayesian probability points in the case of several parameters. *J. Roy. Stat. Soc. B*, **27**, 9–16.
Pierce, D.A. & Peters, D. (1992). Practical use of higher-order asymptotics for multiparameter exponential families (with discussion). *J. Roy. Stat. Soc. B*, **54**, 701–737.
Pierce, D.A. & Peters, D. (1994). Higher-order asymptotics and the likelihood principle: One parameter models. *Biometrika*, **81**, 1–10.
Ramsey, F.P. (1926). Truth and probability, Ch. VII. In *Ramsey, F.P. (1931) The Foundations of Mathematics and Other Logical Essays*, Ed. R.B. Braithwaite, pp. 156–198. New York: Harcourt, Brace Company. Electronic version available at: fitelson.org/probability/ramsey.pdf, accessed September 20, 2013.
Reid, N. (1995). The roles of conditioning in inference (with discussion). *Stat. Sci.*, **10**, 138–157.
Royall, R.M. (1997). *Statistical Evidence: A Likelihood Paradigm.* London: Chapman & Hall.
Savage, L.J. (1954). *The Foundations of Statistics.* New York: Wiley.
Schweder, T. & Hjort, N.L. (2002). Confidence and likelihood. *Scand. J. Stat.*, **29**, 309–332.
Simon, H.A. (1956). Rational choice and the structure of the environment. *Psychological Rev.*, **63**, 129–138.
Stein, C. (1959). An example of wide discrepancy between fiducial and confidence intervals. *Ann. Math. Stat.*, **30**, 877–880.
Storey, J.D. (2002). A direct approach to false discovery rates. *J. Roy. Stat. Soc. B*, **64**, 479–498.
Walley, P. (1990). *Statistical Reasoning with Imprecise Probabilities.* London: Chapman & Hall.
Wasserman, L. (2008). Comment on an article by Gelman. *Bayesian Anal.*, **3**, 463–466.
Weerahandi, S. (1993). Generalized confidence intervals. *J. Amer. Stat. Assoc.*, **88**, 899–905.
Welch, B.L. & Peers, H.W. (1963). On formulae for confidence points based in intervals of weighted likelihoods. *J. Roy. Stat. Soc. B*, **25**, 318–329.

Xie, M.-G. & Singh, K. (2013). Confidence distribution, the frequentist distribution estimator of a parameter: a review. *Inter. Stat. Rev.*, **81**, 3–39.
Yager, R.R. & Liu, L. (eds). (2008). *Classic Works of the Dempster–Shafer Theory of Belief Functions.* New York: Springer.
Zheng, L. & Zelen, M. (2008). Multi-center clinical trials: Randomization and ancillary statistics. *Ann. App. Stat.*, **2**, 582–600.

[*Received October 2013, accepted May 2014*]