Principles of Statistical Inference

N. Reid
Department of Statistics, University of Toronto, Canada
and
D. R. Cox
Nuffield College, Oxford, UK

Abstract

Statistical theory aims to provide a foundation for studying the collection and interpretation of data that does not depend on the particular details of the substantive field in which the data is being considered. This gives a systematic way to approach new problems, and a common language for summarizing results; ideally the foundations and common language ensure that statistical aspects of one study, or of several studies on closely related phenomena, can in broad terms be understood by the non-specialist. 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.

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. 3) for remarks on the need for a range of forms of statistical inference.

We shall not in this short paper discuss statistical decision theory, important though that is. Of course many investigations involve a decision-making element, but most commonly the role of statistics is to summarize evidence in a clear and cogent form, not explicitly 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 essentially arbitrarily. Note though that the best action in such cases may be a search for a third decision, so far overlooked; the formulations of general theory are rarely really closed. Further, formal treatments of decision theory are typically based on maximizing 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 exclude also 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.

In what follows we concentrate on formal issues connected with the assessment of uncertainty. There are of course many challenging aspects of statistical work that are not covered by this.

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 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 axiomatization 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. We shall not discuss the main alternative forms of axioms, 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 latter is available in Yager and Liu (2008).

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.

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 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 above, combined with an explicit description in idealized form of the physical, biological, . . . data generating process;

- either or both of the first two approaches may be used solely to describe the randomization used in experimental design or in sampling an existing population, leading to 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.

We shall not discuss the last aspect in this paper, important though it is for the rather special applications in which it is relevant. The common feature of the other approaches is that they represent features of the "real" world, of course in somewhat idealized 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.

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 strength of belief in some uncertain proposition, in a statistical context perhaps 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:

- 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 of course need care, often requiring some notion of conditioning.[1]

- Probability may measure rational, supposedly impersonal, degree of belief given relevant information. This has a long history, the most notable account being that of Jeffreys (1961).

- 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 I.J. Good (1950).

2.3 Brief assessment

We will in this section simply outline comments on the issues implicitly raised by these distinctions.

[1] Note that the formal accept-reject paradigm of the Neyman-Pearson theory, if taken literally as defining a mode of implementation, would be an instance of decision analysis and as such outside the immediate discussion.

Even if an empirical frequency-based view 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 were much of the time to be poorly *calibrated*: that is, if it 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 is it always sound to regard a probability p found from careful investigation of a real-world effect as equivalent to a personal judgment 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.

A great attraction of Bayesian arguments is that all calculations are by the rules of probability theory. Another attractive feature, in principle at least, is the possibility of assimilating external evidence. This is at the heart of personalistic approaches, but a great many applications of Bayesian arguments 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.

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.

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 the word *prior* does not necessarily mean *previous in time*. Thus, particularly in studies that last for a long time period, the prior may change from that used in planning the study and in particular cases 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 which, 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, of course, often 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 in Cox (2006, Ch. 5). A lively discussion of calibration of Bayesian approaches is given from several points of view in Berger (2006), Goldstein (2006), Browne and Draper (2006) and the 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. Xie and Singh (2013) and the accompanying discussion is a good entry point to this literature.

3 Simple test of significance

While discussions of the meaning of probability have proved 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 though 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, which specifies numerically the distribution of either the full data or certain aspect of the data. We wish to examine consistency with that null hypothesis. Further we suppose chosen a 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.[2]

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 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, of course, a substantial literature on the interpretation and misinterpretation of p-values.

4 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 has been or is to be observed that is provisionally assumed to follow some member of the family of probability models. These are, of course, formidable assumptions from an applied viewpoint. McCullagh (2002) emphasizes the importance of careful delineation of

[2] Fisher considered that at least for discrete problems the test statistic could be minus the probability of the data; see, for example, Fisher (1956, pp. 37-). This could, however, lead to some artificiality.

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 factorization 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 emphasized 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 factorization of their joint distribution of the form
$$f(s;\theta) \propto f_1(s_1 \mid s_2;\theta)f_2(s_2), \tag{3}$$
that 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 that only the directly observed likelihood is relevant, thus excluding dependence on the sampling distribution of statistics derived from the likelihood function.

4.1 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, for example Cox &

Hinkley (1974, Ch. 2) or Lehmann & Romano (2005, Ch 1). Sufficiency is closely tied to the theory of exponential families, as in general these are the models which permit substantial dimension reduction via sufficiency. A mathematical discussion of the sufficiency of the *likelihood map*, i.e. the equivalence class of functions $L(\theta; \cdot)$ is given in Barndorff-Nielsen et al. (1976), and extended in Fraser et al. (1997) and Fraser and Naderi (2007).

4.2 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 in Fraser (1961, 1968). In a few cases non uniqueness in the choice of ancillary statistics has to be resolved by somewhat *ad hoc* criteria.

4.3 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 which are consistent with that likelihood principle are a non-probabilistic use of the likelihood function (see, e.g. Royall, 1997 and Edwards, 1960) as defining regions of the parameter space that are relatively more or less likely, or Bayesian inference, which derives its probabilities via the prior distribution.

If there are no nuisance parameters the non-probabilistic approach indicates, for example, graphical summarization 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, in Berger and Wolpert (1984). One point of interest noted there 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 y, and sufficiency is automatically incorporated, since the likelihood function depends the data only through the sufficient statistic.

4.4 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 nonlinear 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 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 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 factorizations 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 in 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 §2, will in any case rarely be available for complex models with large numbers of parameters. 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 and Bernardo, 1992; Berger et al., 2009; Fraser et al., 2010).

The confidence distributions briefly mentioned at the end of §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. While important, the lack of a general recipe for constructing pivotal quantities, has meant that they receive somewhat less attention in studies of theoretical statistics. There has been a recent revival of interest in confidence distribution functions; see Xie and Singh (2013), and Schweder and 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.

5 Asymptotic theory

Consideration of distributions of inferential quantities as a notional sample size or amount of information increases, and the approximations for use in inference suggested by these, both simplifies and complicates the discussion. For example, notions of approximate sufficiency and approximate ancillarity have been developed; see for example Cox (1980) and McCullagh (1984), as well as Barndorff-Nielsen & Cox (1994, Ch. 7). 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.

In contrast to approximate sufficiency and ancillarity, the details of which are complex, approximate pivotal quantities are used nearly routinely in applied work, thanks in part to the development of robust software for optimization 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 $\ell_{\mathrm{p}}(\psi) = \log L(\psi, \hat{\lambda}_\psi)$, the standardized maximum likelihood estimator

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

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

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

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 more detailed study of the asymptotic expansions involved, and when the parameter of interest is a scalar an improved version of (5) is

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

where $r(\psi)$ is the square root of (5), with the appropriate sign attached, and $Q(\psi)$ is a related pivotal quantity with the property that it has the same limiting distribution as $r(\psi)$, i.e. 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.

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

$$r^*_B(\psi) = r(\psi) + \frac{1}{r(\psi)} \log\{\frac{Q^\pi_B(\psi)}{r(\psi)}\}, \tag{7}$$

where $Q^\pi_B(\psi)$ depends on the prior π, as well as the first and second derivatives of the profile 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 $Q^\pi_B(\psi)$ is

$$Q^\pi_B(\psi) = -\ell'_{\mathrm{p}}(\psi) j_{\mathrm{p}}^{-1/2}(\hat{\psi}) \left\{ \frac{|j_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)|}{j_{\lambda\lambda}(\hat{\psi}, \hat{\lambda})|} \right\}^{1/2} \frac{\pi(\hat{\psi}, \hat{\lambda})}{\pi(\psi, \hat{\lambda}_\psi)}, \tag{8}$$

where $j_{\mathrm{p}}(\psi) = -\partial^2 \ell_{\mathrm{p}}(\psi)/\partial\psi^2$ is the analogue of Fisher information based on the profile log-likelihood, $j_{\lambda\lambda}(\theta) = -\partial^2 \log L(\psi, \lambda)/\partial\lambda\partial\lambda^{\mathcal{T}}$ is the sub-matrix of the full Fisher information matrix corresponding to the nuisance parameter λ, and, as above, $\hat{\lambda}_\psi$ 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 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 in Brazzale et al. (2008, Chs. 3–7), where asymptotically equivalent versions due to Barndorff-Nielsen (1991) and Fraser (1991) 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 standardized maximum likelihood estimator of ψ, with a nuisance parameter adjustment:

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

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

Detailed study of these approximations leads to the following 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 in Welch & Peers (1963), and articulated in this context in Pierce & Peters (1994).

- A prior which 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 in Fraser & Reid (2002).

- These strong matching priors are specific to the parameter of interest, suggesting that any prior which is calibrated in this sense for ψ is unlikely to be calibrated for other components of θ. This also follows from Peers (1965), which considered the extension to nuisance parameter of the results in Welch & Peers (1963). The need to target the prior on the parameter of interest is emphasized 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 adjusts for the estimation of the nuisance parameters, and this adjustment is, in practical problems, much more important for the accuracy of the inference than the distributional improvement (Pierce & Peters, 1992).

- A 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): since 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 in 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 factorization (2) holds, based on these notions of higher order approximation is to our knowledge not yet available.

6 Discussion

We emphasized in §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 or not there really are non-influential, particularly when high-dimensional parameters are involved.

Many applications of statistical ideas now current involve vast amounts of data, or highly complex models, or both, and the question arises whether the principles touched on here continue to be relevant to these settings. A principled approach is surely necessary to avoid continued '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 of course an essential aspect of design and analysis of surveys and experiments, topics that we are not addressing here, and efficient inference for temporal streams is perhaps mainly an issue of computation, but theoretical statistics, and the classic concepts discussed above would seem to be important for the remainder. For example, the ideas behind significance testing underly 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.

An issue arising if 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. There are also broader strategical issues. How best should a wholly new large set of data be approached; summary analysis of the whole may be combined with 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.

Acknowledgments

This paper is based on a talk given at the World Statistics Congress of the International Statistical Institute in Hong Kong, August 2013. NR's research is partially supported by the Natural Sciences and Engineering Research Council of Canada.

References

Barndorff-Nielsen, O.E. and Cox, D.R. (1994). *Inference and Asymptotics.* Chapman & Hall, London.

Barndorff-Nielsen, O., Hoffmann-Jorgensen, J., and Pedersen, L. (1976). On the minimal sufficiency of the likelihood function. *Scandinavian Journal of Statistics*, **3**, 37–38.

Berger, J.O. and Bernardo, J.M. (1992). On the development of reference priors (with discussion). In *Bayesian Statistics 4*, J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith, eds., 35–60. Oxford University Press, Oxford.

Berger, J.O. (2006). The case for objective Bayesian analysis. *Bayesian Statistics* **3**, 385–402.

Berger, J.O., Bernardo, J.M. and Sun, D. (2009). The formal definition of reference priors. *Ann. Statist.* **37**, 905–938.

Berger, J.O. and Wolpert, R.L. (1984). *The Likelihood Principle.* Institute of Mathematical Statistics, Hayward.

Breiman, L. (2001). Statistical modelling: the two cultures. *Statistical Science* **16**, 199–231.

Browne, W.J. and Draper, D. (2006). A comparison of Bayesian and likelihood-based methods for multi-level models. *Bayesian Analysis* **3**, 473–514.

Cox, D.R. (1958). Some problems with statistical inference. *Annals of Mathematical Statistics* **29**, 357–372.

Cox, D.R. (1980). Local ancillarity. *Biometrika* **67**, 269–276.

Cox, D.R. (2006). *Principles of Statistical Inference.* Cambridge University Press, Cambridge.

Cox, D.R. and Hinkley, D.V. (1974). *Theoretical Statistics.* Chapman & Hall, London.

de Finetti, B. (1974). *Probability, Induction and Statistics.* Wiley, New York.

DiCiccio, T.J., Field, C.A. and 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. and Stern, S.E. (2001). Simple and accurate one-sided inference from signed roots of likelihood ratios. *Canadian Journal of Statistics* **29**, 67–76.

DiCiccio, T.J. and 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 University Press, Oxford.

Fisher, R.A. (1930). Inverse probability. *Proceedings of the Cambridge Philosophical Society* **26**, 528–535.

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

Fraser, D.A.S. (1961). The fiducial method and invariance. *Biometrika* **48**, 261–280.

Fraser, D.A.S. (1968). *The Structure of Inference.* Wiley, New York.

Fraser, D.A.S., McDunnough, P., Naderi, A., and 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. and Naderi, A. (2007). Minimal sufficient statistics emerge from the observed likelihood function. *International Journal of Statistical Science* **6**, 55–61. available at `http://www.utstat.toronto.edu/dfraser/documents/238.pdf`, accessed on September 18, 2013.

Fraser, D.A.S. and Reid, N. (2002). Strong matching of frequentist and Bayesian parametric inference. *J. Statist. Plann. Infer.* **103**, 263–285.

Fraser, D.A.S. and Rousseau, J. (2008). Studentization and deriving accurate p-values. *Biometrika* **95**, 1–16.

Fraser, D.A.S., Reid, N., Marras, E. and Yi, G.Y. (2010). Default priors for Bayesian and frequentist inference. *Journal of the Royal Statistical Society* B **72**, 631–654.

Goldstein, M. (2006). Subjective Bayesian analysis: principles and practice. *Bayesian Analysis* **3**, 403–420.

Good, I.J. (1950). *Probability and the weighing of evidence.* MIT Press, Cambridge.

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

Lehmann, E.L. and Romano, J.P. (2005). *Testing Statistical Hypotheses.* 3rd edition. Springer, New York.

Lindley, D.V. (1958). Fiducial distributions and Bayes' theorem. *Journal of the Royal Statistical Society* B **20**, 102–107.

McCullagh, P. (1984). Local sufficiency. *Biometrika* **71**, 233–244.

McCullagh, P. (2002). What is a statistical model? (with discussion) *Annals of Statistics***30**, 1225–1310.

National Research Council (2013). *Frontiers in Massive data Analysis.* National Academies Press, Washington.
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. *Journal of the Royal Statistical Society* B **27**, 9–16.

Pierce, D.A. and Peters, D. (1992). Practical use of higher-order asymptotics for multiparameter exponential families (with discussion). *Journal of the Royal Statistical Society* B **54**, 701–737.

Pierce, D.A. and Peters, D. (1994). Higher-order asymptotics and the likelihood principle: one parameter models. *Biometrika* **81**, 1–10.

Ramsey, F.P. (1926) Truth and probability. in Ramsey, F.P. (1931) *The Foundations of Mathematics and other Logical Essays*, Ch. VII, p.156-198, R.B. Braithwaite,ed. Harcourt, Brace and Company, New York. Electronic version available at `fitelson.org/probability/ramsey.pdf`, accessed September 20, 2013.

Reid, N. (1995). The roles of conditioning in inference (with discussion). *Statistical Science* **10**, 138–157.

Royall, R.M. (1997). *Statistical Evidence: a Likelihood Paradigm.* Chapman & Hall, London.

Savage, L.J. (1954). *The Foundations of Statistics.* Wiley, New York.

Schweder, T. and Hjort, N.L. (2002). Confidence and likelihood. *Scandinavian Journal of Statistics* **29**, 309–332.

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review* **63**, 129–138.

Walley, P. (1990). *Statistical Reasoning with Imprecise Probabilities.* Chapman & Hall, London.

Wasserman, L. (2008). Comment on an article by Gelman. *Bayesian Analysis* **3**, 463–466.

Welch, B.L. and Peers H.W. (1963). On formulae for confidence points based in intervals of weighted likelihoods. *Journal of the Royal Statistical Society* B **25**, 318–329.

Xie, M.-G. and Singh, K. (2013). Confidence distribution, the frequentist distribution estimator of a parameter: a review. *International Statistical Review*, **81**, 3–39.

Yager, R.R. and Liu, L. (eds.) (2008). *Classic Works of the Dempster-Shafer Theory of Belief Functions.* Springer, New York.