ELSEVIER

Journal of Statistical Planning and Inference 105 (2002) 35–65

journal of statistical planning and inference

www.elsevier.com/locate/jspi

Reconciling frequentist properties with the likelihood principle

Peter Walley

Received 6 October 1999; received in revised form 6 March 2000; accepted 25 January 2001

Abstract

For the general problem of parametric statistical inference, several frequentist principles are formulated, including principles of hypothesis testing, set estimation, and conditional inference. These principles guarantee that, whatever the true parameter value, statistical procedures have little chance of producing misleading inferences. The frequentist principles are shown to be compatible with the likelihood principle and with principles of coherence. Two general methods are studied which satisfy both the likelihood and frequentist principles in finite samples. One method produces posterior upper and lower probabilities from a very large set of prior probability measures, which can be taken to be an ε-contamination neighborhood with ε slightly larger than $\frac{1}{2}$. The second method derives inferences from a normalized version of the observed likelihood function. Because inferences from the two methods encompass a wide range of frequentist, likelihood and Bayesian inferences, they are conservative and they have relatively low power. More powerful methods can be obtained by weakening the frequentist principles and making weak assumptions about the sampling rule. The results show that there are methods of statistical inference, based on particular types of imprecise probability model, which satisfy the likelihood principle, are coherent, and have good frequentist properties under a range of sampling models.

MSC: primary 62A01; secondary 62F03; 62F25

Keywords: Conditional inference; Consistency function; Contamination neighborhood; Frequentist principle; Foundations of statistics; Imprecise beta model; Imprecise probability; Upper probability

1. Introduction

Theories of statistical inference can be divided into two broad classes: those that satisfy the likelihood principle, and those in which inferences have a frequentist or repeated-sampling interpretation. Theories that satisfy the likelihood principle include likelihood inference, Bayesian inference, robust Bayesian inference, and some theories

E-mail address: pwalley@eudoramail.com (P. Walley).

0378-3758/02/$ - see front matter
PII: S0378-3758(01)00203-8

of imprecise probability. In these theories, measures of uncertainty, such as posterior probabilities or likelihood ratios, do not depend on the likelihoods of possible observations that did not occur. Frequentist theories include the Neyman–Pearson theory of hypothesis testing and set estimation and Fisher's theory of significance testing. In these theories, the uncertainty of inferences is measured by quantities such as p-values or confidence coefficients, which are averages over possible data sets that might have been observed. [1]

In the case of interval estimation, for instance, a Bayesian or robust Bayesian procedure for producing a 95% credible interval is defined so that the posterior probability that the true parameter value belongs to the interval, with respect to a particular posterior distribution or set of distributions, is at least 0.95. Such procedures may have poor frequentist properties, in the sense that, under some possible parameter values, the chance (frequentist probability) of obtaining a credible interval which covers the true parameter value may be much less than 0.95.

On the other hand, a frequentist procedure for producing a 95% confidence interval is defined so that, before the statistical data are obtained, the chance of obtaining an interval which will cover the true parameter value is at least 0.95, no matter what the true parameter value may be. Such procedures may have poor properties from a conditional point of view, e.g., it may be certain, after the data are obtained, that the 95% confidence interval does not contain the true parameter value, or the procedure may have 'relevant subsets' which suggest that a specific numerical interval is less likely to contain the true value (Buehler, 1959; Robinson, 1979; Lehmann, 1986). There is a similar difference in philosophy between the two approaches to hypothesis testing.

There are strong arguments to support each approach. Briefly, the first approach is supported by arguments in favor of the likelihood principle (Birnbaum, 1962; Basu, 1975; Berger and Wolpert, 1984) and principles of coherence (de Finetti, 1974; Walley, 1991). The main arguments in favor of the frequentist approach are that it gives a physical interpretation to measures of uncertainty and it guarantees that, whatever the true parameter value, statistical procedures have little chance of producing misleading conclusions (Fisher, 1956; Birnbaum, 1969; Cox and Hinkley, 1974; Neyman, 1977). For comparisons of the two approaches, see Barnett (1982), Kyburg (1974), and Cox and Hinkley (1974).

It seems desirable to satisfy both the likelihood principle and frequentist principles. However, it is widely believed that, in general, these principles are incompatible. This belief seems to stem, at least in part, from work of Birnbaum. Birnbaum (1969, p. 114) distinguished two kinds of criteria for concepts of statistical evidence, which correspond to the two approaches I have outlined, and summarized his conclusion

[1] The most popular theories of statistical inference fit into one of these two classes, but some methods do not, e.g., Jeffreys' method of choosing a 'noninformative' prior distribution, which violates both likelihood and frequentist principles, and apparently also Fisher's fiducial method.

about their incompatibility:

> It has seemed to some (including this writer) that any adequate concept of statistical evidence must meet at least certain minimum versions of both of the criteria just indicated. But the difficulties of developing such a concept have become increasingly apparent, and it now seems rather clear that no such adequate concept of statistical evidence can exist.

I share Birnbaum's desire for a theory of statistical inference that satisfies 'at least certain minimum versions' of both the likelihood and frequentist principles, but I believe that his negative conclusion was premature and not justified by the examples he discussed in his paper. Here is a simplified version of the crucial example of Birnbaum (1969, pp. 127–128).

Example 1.1 (*Cox and Hinkley*, 1974, *pp.* 51–52). Let the parameter space be $\Theta = \{0, 1, \ldots, 100\}$, let the sample space be $\mathscr{X} = \{1, 2, \ldots, 100\}$, and define the sampling distribution P_θ to be the uniform probability distribution on $\mathscr{X}$ if $\theta = 0$, or the degenerate distribution at $x = \theta$ if $\theta = 1, 2, \ldots, 100$. Then, whatever value of x is observed, the likelihood function L_x satisfies $L_x(x) = 1$, $L_x(0) = 0.01$, and $L_x(\theta) = 0$ for all other values of θ. Let H_0 denote the hypothesis that $\theta = 0$. If we use the generalized likelihood ratio $L_x(0)/\sup\{L_x(\theta)\colon \theta \in \Theta\} = 0.01$ to measure the strength of evidence against H_0 then, even if H_0 is true, we are certain to obtain strong evidence against H_0. These inferences are incompatible with fundamental frequentist principles and with common sense.

Although Birnbaum (1969, p. 126) seems to deny that inferences based on the likelihood function must involve any particular measure of evidence in favor of a composite hypothesis, in his example he implicitly used the generalized likelihood ratio as the measure of evidence. There are other measures of evidence for a composite hypothesis, to be studied in this paper, which give very different answers in this problem and which are consistent with frequentist principles. Under the contamination models defined in Section 3 (Theorem 3.1), based on a uniform distribution Q, the posterior upper probability for H_0 is identical to its prior upper probability and is greater than $\frac{1}{2}$. Under the normalized-likelihood models in Eq. (5.2) and in Section 5.4, the degree of consistency of H_0 with the data x is one, whatever value of x is observed. According to these methods, which satisfy both the likelihood principle and frequentist principles, the data provide no evidence at all against H_0. So the Birnbaum–Cox–Hinkley example does not establish that the likelihood principle is incompatible with frequentist principles.

My aims in this paper are to give a mathematical formulation of the likelihood and frequentist principles and to show that there are very general statistical methods which satisfy both criteria. The obvious difficulty in finding such methods is that frequentist properties depend on the entire sampling model, whereas inferences which satisfy the likelihood principle can depend only on the observed likelihood function.

For a method which satisfies the likelihood principle to also satisfy the frequentist properties, it must do so irrespective of how the sampling probabilities are defined for unobserved data, e.g., irrespective of the stopping rule in the case of sequential observations. I will show that this can be achieved by making sufficiently cautious inferences.

The approach that I adopt in most of this paper, excluding Section 5, is to derive inferences from an *imprecise probability* model (Walley, 1991, 1999), which is based on a set of prior probability distributions for the unknown parameter. Inferences are made by first applying Bayes' rule to each prior distribution in the set, and then computing upper and lower posterior probabilities by maximizing and minimizing over the set. This is also the inference method used in *robust Bayesian* inference (Berger, 1994; Wasserman, 1997). But whereas robust Bayesians aim to model uncertainty about a correct Bayesian prior distribution, and they regard the distributions in the basic set as 'plausible' prior distributions, the aim here is to produce methods which satisfy frequentist principles and there is no assumption that any prior probability distributions are 'plausible'. For discussion of the differences between these two approaches, see Walley (1991, Sections 2.10, 5.9) and Pericchi and Walley (1991).

The imprecise probability models studied here can be regarded as ways of reconciling frequentist, likelihood and Bayesian inferences. Inferences from an imprecise probability model automatically satisfy the likelihood principle and principles of coherence, and I will show that the basic frequentist principles can be satisfied by choosing the set of prior distributions to be sufficiently large. For example, 95% credible sets are defined to have posterior probability at least 0.95 with respect to every prior distribution in the set, and therefore they are valid (but conservative) Bayesian 95% credible sets under a range of different prior distributions. If the set of prior distributions is sufficiently large then the 95% credible sets can also be shown to be valid (but conservative) frequentist 95% confidence regions under a range of different sampling models. They can also be chosen to be likelihood sets (a set of parameter values whose likelihood exceeds a threshold). Similarly, in testing any null hypothesis (simple or composite), the consistency values produced by these methods are upper bounds for certain frequentist p-values, for the generalized likelihood ratio, and for a range of Bayesian posterior probabilities.

Of course, there have been many previous attempts to reconcile different statistical theories. As far as I am aware, none of these previous studies have considered the possibility that the frequentist and likelihood principles could be satisfied simultaneously in very general settings. There is a large literature on two particular kinds of reconciliation. The first body of work concerns *exact* agreement between frequentist and Bayesian (or likelihood or fiducial) inferences for particular types of parametric sampling model; see Lindley (1958), Welch and Peers (1963), Thatcher (1964), Bartholomew (1971), Edwards (1972, Chapter 9), Box and Tiao (1973), and Chang and Villegas (1986). For example, frequentist inferences about a real-valued location parameter, conditional on a maximal ancillary statistic, formally agree with Bayesian inferences based on an improper uniform prior density and with fiducial inferences (Fisher,

1934; Cox and Hinkley, 1974, p. 221). But such results are very limited. Consider, for example, a random sample from a normal distribution with known variance. Provided the sample size is fixed, standard frequentist inferences about the mean are formally the same as Bayesian inferences based on a uniform prior density. But if the sample size may depend on the observations then, in general, frequentist inferences will vary with the stopping rule, to satisfy frequentist principles, whereas Bayesian inferences, in order to satisfy the likelihood principle, cannot depend on the stopping rule.

The second large body of work concerns *asymptotic* agreement between frequentist, Bayesian and likelihood inferences as the sample size tends to infinity, and the use of one inference method to *approximate* another. See, for example, Welch and Peers (1963), Pratt (1965), Dawid (1991), Fraser (1991), Severini (1991), Efron (1993), Nicolaou (1993), and Datta and Ghosh (1995). This work shows that, under some restrictive assumptions, there are methods that satisfy both the frequentist and likelihood principles asymptotically (Dawid, 1991). However, these asymptotic results are of doubtful relevance to finite samples. The methods studied in this paper satisfy both the frequentist and likelihood principles in very general settings and for all finite sample sizes, not just asymptotically or approximately.

The paper is organized as follows. Section 2 presents a mathematical formulation of the likelihood principle and five frequentist principles, and characterizes the close relationships between the frequentist principles. Two general methods which satisfy both the likelihood and frequentist principles are described. The first method, defined in Section 3, produces posterior upper and lower probabilities from a very large set of prior probability measures, which is taken to be an ε-contamination neighborhood with ε slightly bigger than $\frac{1}{2}$. Some examples of these inferences are given in Section 4. A second method, in which inferences are based on a normalized version of the observed likelihood function, is described in Section 5. Because inferences from the two methods encompass a wide range of frequentist and Bayesian inferences, they have relatively low power. It is possible to obtain more powerful inferences in particular problems, one of which is discussed in Section 6, by weakening the frequentist principles and making weak assumptions about the sampling model. One way of generalizing that approach is outlined in the concluding Section 7.

2. Some principles of statistical inference

We are concerned with the standard problem of parametric statistical inference: it is known or assumed that statistical data x are generated by a sampling model which belongs to a family $\{P_\theta\colon \theta \in \Theta\}$, where θ is a parameter which indexes the possible sampling models P_θ and Θ is the parameter space, and we wish to draw some conclusions from x concerning the true sampling model or (equivalently) the true parameter value. Throughout the paper I use the term *chance* to refer to the probability of an event under the true sampling model P_θ. Such a probability can be interpreted as a physical propensity or as a long-run relative frequency.

Let X denote the random variable (possibly multidimensional) whose observed value is x. Assume that each possible sampling model P_θ is a probability measure defined on $\mathscr{A}(\mathscr{X})$, a σ-field of subsets of the sample space $\mathscr{X}$ which includes all the singleton sets $\{x\}$. I also assume that a σ-finite measure ν is defined on $\mathscr{A}(\mathscr{X})$, each probability measure P_θ has a probability density function f_θ with respect to ν, and $f_\theta(x)$ is measurable as a function of (θ,x) with respect to the product σ-field $\mathscr{A}(\Theta)\times\mathscr{A}(\mathscr{X})$. Here $\mathscr{A}(\Theta)$ is a σ-field of subsets of Θ which contains all the subsets of interest, including all the singleton sets $\{\theta\}$. Apart from these regularity assumptions, there are no restrictions on the sample space, parameter space or sampling models.

2.1. Consistency functions

The first task is to formalize the two criteria for evaluating statistical methods that were outlined in the introduction. To do so I shall assume that, after observing data x, the conclusions of statistical inference are summarized in the form of a *consistency function* $\rho(\cdot|x)$, where the nonnegative real number $\rho(A|x)$ is defined for all subsets $A\in\mathscr{A}(\Theta)$. To simplify the notation I write $\rho(\theta|x)$ instead of $\rho(\{\theta\}|x)$.

For $A\in\mathscr{A}(\Theta)$, let H_A denote the hypothesis that the true parameter value θ belongs to the set A. I shall call $\rho(A|x)$ the *degree of consistency* between the hypothesis H_A and the data x. It could also be interpreted as the (inverse) *strength of evidence* against H_A provided by the data x, or as the *plausibility* of H_A given data x. These three interpretations are mutually compatible: lower consistency of H_A with x corresponds to stronger evidence against H_A and to lower plausibility of H_A given x. In particular, if the data x provide no real evidence for discriminating between H_A and its negation then both $\rho(A|x)$ and $\rho(A^{\mathrm{c}}|x)$ may be high, meaning that both hypotheses are consistent with the data and both hypotheses are plausible given the data.

For the inference methods studied later in this paper which are based on imprecise probability models, $\rho(A|x)$ is taken to be the *posterior upper probability* of H_A given x, which has another, more practical, interpretation as a posterior betting rate for betting *against* H_A: lower consistency of H_A with x corresponds to offering longer odds against H_A after observing x.[2] When both $\rho(A|x)$ and $\rho(A^{\mathrm{c}}|x)$ are high, we are unwilling to bet against either hypothesis. For the inference method studied in Section 5, $\rho(A|x)$ is the maximum value of a *normalized likelihood function* that is attained on A.

Other types of consistency function are possible. For example, $\rho(A|x)$ could be taken to be a frequentist *p-value* for testing H_A: $\theta\in A$ (which is often regarded as

[2] The term 'plausibility' is sometimes used as a synonym for 'upper probability' (Shafer, 1976), but it suggests, to some people, a subjective or psychological concept, whereas $\rho(A|x)$ is intended to be objective. The term 'consistency' is widely used in the frequentist interpretation of p-values, whereas 'plausibility' and 'upper probability' are used in approaches based on the likelihood principle (imprecise probability and robust Bayesian theory). Each of these terms is appropriate here, since the functions $\rho(\cdot|X)$ are required to satisfy both the likelihood principle and the frequentist properties of p-values, but it may be preferable to use a term that is neutral between the frequentist and likelihood approaches. The term 'strength of evidence' does seem to be neutral, but it could be misleading because it is inversely related to $\rho(A|x)$, i.e. higher values of $\rho(A|x)$ represent weaker evidence. On balance, I have chosen to use the term 'consistency'.

a measure of the consistency of the null hypothesis with the data), a Bayesian *posterior probability* $P(A|x)$, or the *generalized likelihood ratio* for testing H_A. However, these three alternative measures do not satisfy, in general, all the frequentist and likelihood properties that I shall require of $\rho(\cdot|X)$. The two types of consistency function to be studied in Sections 3–5 are essentially upper bounds for the three alternative measures, i.e., they produce higher degrees of consistency and more cautious inferences.

In order to define frequentist properties of the consistency function, I assume that $[\rho(A|X) \leqslant \alpha] = \{x \in \mathscr{X}: \rho(A|x) \leqslant \alpha\}$ is in $\mathscr{A}(\mathscr{X})$ whenever $A \in \mathscr{A}(\Theta)$ and $0 \leqslant \alpha \leqslant 1$. I also assume the *monotonicity* condition

$$\rho(\theta|x) \leqslant \rho(A|x) \quad \text{whenever } \theta \in A,\ A \in \mathscr{A}(\Theta) \text{ and } x \in \mathscr{X}. \tag{2.1}$$

2.2. Likelihood principle

Given data x, define the *observed likelihood function* L_x on Θ by $L_x(\theta) = f_\theta(x)$. According to the likelihood principle, L_x can be replaced by any positive multiple cL_x without changing inferences.[3]

The first approach to statistical inference that was discussed in the introduction is based on the *likelihood principle*, which can be formalized as follows.

LP: Suppose that either of two experiments can be performed and the sampling model for each experiment is completely determined by the same unknown parameter θ. If x and y are possible outcomes of the experiments whose corresponding likelihood functions are proportional, then the consistency functions $\rho(\cdot|x)$ and $\rho(\cdot|y)$ that would result from each outcome should be identical. That is, if $L_x \propto L_y$ then $\rho(\cdot|x) = \rho(\cdot|y)$.

Essentially, LP tells us that the consistency function $\rho(\cdot|x)$, which summarizes the inferences from data x, should depend on the sampling model and data only through the observed likelihood function L_x. There are strong arguments in favor of LP. For example, LP is implied by sufficiency and conditionality principles (Birnbaum, 1962), and by general principles of coherence (Walley, 1991, Section 8.6). Other arguments in favor of LP are discussed in Basu (1975) and Berger and Wolpert (1984). All the methods studied later in the paper satisfy LP.

[3] Strictly, 'the likelihood function' should be regarded as an equivalence class of mutually proportional functions. Unless the sample space $\mathscr{X}$ is discrete, difficulties can arise from the nonuniqueness of L_x: a probability density function f_θ can be changed on a zero-probability subset of $\mathscr{X}$ without changing the probability distribution P_θ. To reduce these difficulties, in problems where a continuous sample space is an idealization of a discrete measurement process, $f_\theta(x)$ should be defined not as a Radon–Nikodym derivative, but as a limit of normalized probabilities of decreasingly small neighbourhoods of x which correspond to the discrete measurement process (Walley, 1991, Section 8.6). In most practical applications where x is a vector of real-valued observations x_i, each $f_\theta(x_i)$ is taken to be the derivative at x_i of the cumulative distribution function of X_i under P_θ, which usually exists at all except finitely many points x_i. To avoid difficulties arising from any remaining nonuniqueness of L_x, $f_\theta(x)$ should be chosen in a consistent way across comparable sampling models, to allow the likelihood principle to be applied (e.g. see Section 2.7).

2.3. Fundamental frequentist principle

The most fundamental frequentist principle appears to be the following: whatever the true value of the parameter θ, there should be little chance of obtaining data x that have low degree of consistency $\rho(\theta|x)$ with the true θ. Many authors have suggested similar principles, including the 'confidence concept' of Birnbaum (1969, 1977), the 'repeated sampling principles' of Cox and Hinkley (1974), the motivation for significance testing in Fisher (1956), and the motivation for the Neyman–Pearson approach to hypothesis testing and interval estimation in Neyman (1977). See also the 'production principle' of Dawid (1991). I do not know whether there have been previous attempts to formulate these principles mathematically. One way to do so, and to give a frequentist interpretation to low degrees of consistency, is to require that, for any sufficiently small value of α and whatever the true parameter value θ, the chance of obtaining data x under which $\rho(\theta|x)$ is no greater than α should be no greater than α. We therefore require the following *fundamental frequentist principle*:

FFP: $P_\theta[\rho(\theta|X) \leqslant \alpha] \leqslant \alpha$ whenever $\theta \in \Theta$ and $0 \leqslant \alpha \leqslant \alpha_0$.

The main argument for FFP is that it guarantees that there is little chance of making a misleading inference by declaring that a parameter value is 'inconsistent with the data' when it happens to be the true value. The reason for including the restriction $\alpha \leqslant \alpha_0$ is that the argument applies only to small consistency values, which (unlike probabilities) are not symmetric with large consistency values. The following argument shows that consistency values behave like p-values, which are interpreted as significant evidence against a null hypothesis only when they are sufficiently small, which usually means smaller than 0.1. Similarly, in set estimation, confidence coefficients are usually chosen to be at least $0.9 = 1 - 0.1$. To be consistent with the usual types of frequentist inference, it therefore suffices to take $\alpha_0 = 0.1$ in FFP. Of course, larger values of α_0, producing stronger versions of FFP, can also be used.

A second argument for FFP is that it gives an operational meaning and calibration to degrees of consistency, in terms of long-run frequencies. By substituting $\alpha = \rho(\theta|x)$, FFP implies that, when $\rho(\theta|x) \leqslant \alpha_0$, the observed degree of consistency $\rho(\theta|x)$ has the well known property of p-values: whatever the true parameter value, the chance of obtaining a degree of consistency no greater than the observed value $\rho(\theta|x)$ is no greater than $\rho(\theta|x)$. If this kind of calibration is required to hold across the whole consistency scale, we need $\alpha_0 = 1$ in FFP. Such a requirement is not compelling, however, because of the asymmetry between low and high consistency values. Of the two methods studied later in the paper, the first requires $\alpha_0 < 1$ but the second has $\alpha_0 = 1$.

Frequentist properties such as FFP can be interpreted in terms of repeated sampling from the same sampling model, or more generally in terms of repeated use of statistical methods. Suppose that a method for calculating a consistency function $\rho(\cdot|x)$ is used in a long run of applications, where the data are independent across different applications. Provided the method satisfies FFP, it follows from the weak law of large numbers that, with chance arbitrarily close to one when the number of applications is

sufficiently large, the proportion of applications in which the true parameter value has low degree of consistency will be small, irrespective of the true parameter value in each application.

2.4. Hypothesis testing

FFP is essentially concerned with testing the simple null hypothesis that θ is the true parameter value: it tells us that, when this hypothesis is true, there is no more than a small chance that the data will be inconsistent with it. This can be extended from simple to composite hypotheses. Consider the completely general (simple or composite) hypothesis H_A: $\theta \in A$, where A is a nontrivial subset of Θ and $A \in \mathscr{A}(\Theta)$. To test H_A we calculate $\rho(A|x)$, which measures the consistency of H_A with the data x. (Usually, we would also want to measure the consistency of the complementary hypothesis by calculating $\rho(A^c|x)$.) The next principle is called the *hypothesis testing principle*:

HTP: $P_\theta[\rho(A|X) \leqslant \alpha] \leqslant \alpha$ whenever $\theta \in A$, $A \in \mathscr{A}(\Theta)$ and $0 \leqslant \alpha \leqslant \alpha_0$.

Since smaller values of $\rho(A|x)$ represent lower consistency and therefore stronger evidence against H_A, HTP tells us that, if H_A is true, the chance of obtaining such strong evidence against H_A as we actually observed was no greater than $\rho(A|x)$. Thus $\rho(A|x)$ has the standard property of p-values. Indeed we can regard $\rho(A|x)$ as a *test statistic* for testing H_A and define the p-value of the test to be $\sup_{\theta \in A} P_\theta[\rho(A|X) \leqslant \rho(A|x)]$. Then HTP requires that, at least when $\rho(A|x) \leqslant \alpha_0$, $\rho(A|x)$ is an upper bound for this p-value.[4] This ensures that we will declare H_A to be 'inconsistent with the data' only when there is strong evidence against it, in the frequentist sense that the test has a small p-value. Alternatively, if we adopt the Neyman–Pearson formulation and we reject H_A at level α if and only if $\rho(A|x) \leqslant \alpha$, where α is a suitably low threshold, HTP guarantees that if H_A is true then the chance of rejecting it at level α is at most α.

It will be shown in Sections 2.7 and 4.3 that, assuming LP and HTP, $\rho(A|x)$ is an upper bound for the p-values from certain conditional tests of H_A and for the generalized likelihood ratio statistic, whenever $\rho(A|x) \leqslant \alpha_0$.

2.5. Set estimation

Say that the random set $C(X)$ is a *set estimator* for θ when $C(x) \in \mathscr{A}(\Theta)$ for all $x \in \mathscr{X}$, and $[\theta \in C(X)] = \{x \in \mathscr{X}: \theta \in C(x)\} \in \mathscr{A}(\mathscr{X})$ for all $\theta \in \Theta$. There are several properties that a set estimator may be expected to satisfy. The first property is that, after observing x, there should be strong evidence that $C(x)$ contains the true parameter value; or, in terms of consistency, the hypothesis that $C(x)$ does not contain the true

[4] Note that $\rho(A|x)$ is not necessarily an upper bound for the p-values obtained from other frequentist tests of H_A, because the frequentist test statistic may produce a very different ordering of $\mathscr{X}$ from $\rho(A|x)$, and different orderings generally produce different p-values. But the values $\rho(A|x)$ produced by the two methods in Sections 3 and 5 do seem to be upper bounds for many standard frequentist p-values.

value should be inconsistent with the data x. A second property is that $C(x)$ should contain all the parameter values that are reasonably consistent with x, i.e., whose degree of consistency exceeds a suitable threshold. These two properties correspond to two different ways of defining a credible set and are formalized below. They are based on the likelihood principle, since they refer only to the observed data x. A third property, from a frequentist point of view, is that the set estimator $C(X)$ should have a high chance of containing the true parameter value, whatever the true value may be. The following principles WSEP and SSEP require that set estimators which are constructed to satisfy one of the first two properties must also satisfy the frequentist property.

When $A \in \mathscr{A}(\Theta)$, the quantity $1-\rho(A^c|x)$ is called the *credibility* of the set A (given x). A set estimator for θ is called a *credible set estimator* at *credibility level* $1-\alpha$ when the credibility of $C(x)$ is at least $1-\alpha$, i.e., $\rho(C(x)^c|x) \leqslant \alpha$, for all $x \in \mathscr{X}$. Thus $C(x)$ is said to be a credible set estimate for θ when the hypothesis that $C(x)$ does not contain the true parameter value is sufficiently inconsistent with the data x.

A set estimator for θ is called a *confidence set estimator* at *confidence level* $1-\alpha$ when $P_\theta[\theta \in C(X)] \geqslant 1-\alpha$ for all $\theta \in \Theta$. A natural frequentist requirement is that a credible set estimator should have a high chance of including the true value of θ, whatever it may be. That suggests the following *weak set estimation principle*.

WSEP: If $C(X)$ is a credible set estimator for θ at credibility level $1-\alpha$, where $0 \leqslant \alpha \leqslant \alpha_0$, then $C(X)$ should be a confidence set estimator for θ at confidence level $1-\alpha$.

If $\alpha_0 \geqslant 0.05$, for instance, WSEP ensures that when we use a 95% credible set estimator, the chance that it will cover the true parameter value is at least 0.95, whatever the true value may be. If we apply such estimators in a long run of independent problems, they will be successful in at least 95% of cases. Again this gives a frequentist interpretation to degrees of consistency.

A second way to construct a credible set for θ, using the well known connection between set estimation and hypothesis testing, is to include θ in a credible set (at level $1-\alpha$) if and only if we would not reject (at level α) the null hypothesis that θ is the true parameter value, which is equivalent to the condition $\rho(\theta|x) > \alpha$. This produces the credible set $C^*(x) = \{\theta \in \Theta: \rho(\theta|x) > \alpha\}$, the set of all parameter values whose consistency with the data exceeds the threshold α. The proof of Theorem 2.1 shows that, assuming (2.1), every credible set estimate for θ at credibility level $1-\alpha$ must contain $C^*(x)$, but $C^*(x)$ itself is not necessarily a credible set estimate for θ at level $1-\alpha$. However, it is still possible to require that $C^*(X)$ be a confidence set estimator for θ at confidence level $1-\alpha$, as in the following *strong set estimation principle*.

SSEP: If $C^*(x) = \{\theta \in \Theta: \rho(\theta|x) > \alpha\}$ for all $x \in \mathscr{X}$, where $0 \leqslant \alpha \leqslant \alpha_0$, then $C^*(X)$ should be a confidence set estimator for θ at confidence level $1-\alpha$.

2.6. Relationships between the frequentist principles

The four frequentist principles in the preceding subsections are closely related.

Theorem 2.1. *Given the sampling models* $\{P_\theta\colon \theta\in\Theta\}$, *suppose that a consistency function* $\rho(\cdot|x)$ *is defined for each* $x\in\mathscr{X}$. *Then FFP is equivalent to SSEP. Assuming that the monotonicity condition* (2.1) *holds, FFP, HTP and SSEP are equivalent and each implies WSEP.*

Proof. Using the equation $P_\theta[\theta\in C^*(X)]=P_\theta[\rho(\theta|X)>\alpha]=1-P_\theta[\rho(\theta|X)\leqslant\alpha]$, it follows that SSEP is equivalent to FFP.

By taking $A=\{\theta\}$ in HTP, it is clear that HTP implies FFP. Assuming the monotonicity condition (2.1), if $\theta\in A$ then $P_\theta[\rho(A|X)\leqslant\alpha]\leqslant P_\theta[\rho(\theta|X)\leqslant\alpha]$, and therefore FFP implies HTP.

To show that, given (2.1), SSEP implies WSEP, suppose that $C(X)$ is a credible set estimator for θ at credibility level $1-\alpha$, where $0\leqslant\alpha\leqslant\alpha_0$. Using (2.1), it follows that, whenever $x\in\mathscr{X}$ and $\theta\in C(x)^c$, $\rho(\theta|x)\leqslant\rho(C(x)^c|x)\leqslant\alpha$. Hence $C(x)^c\subseteq\{\theta\in\Theta\colon \rho(\theta|x)\leqslant\alpha\}$, i.e., $C(x)\supseteq C^*(x)=\{\theta\in\Theta\colon \rho(\theta|x)>\alpha\}$, for all $x\in\mathscr{X}$. Thus $C(X)$ contains $C^*(X)$. By SSEP, $C^*(X)$ is a confidence set estimator for θ at confidence level $1-\alpha$, hence so is $C(X)$. This establishes WSEP. □

The inference methods studied in this paper all satisfy the monotonicity condition (2.1). Theorem 2.1 shows that it is enough to verify that these methods satisfy FFP.

2.7. Conditional inference

The sampling probabilities involved in the preceding frequentist principles are unconditional probabilities. It is widely accepted amongst frequentist statisticians that, in many statistical problems, sampling probabilities should be calculated conditionally on the value of an appropriate ancillary statistic; see Fisher (1956), Cox (1958, 1988), Cox and Hinkley (1974), Seidenfeld (1979) and Lehmann (1986). A statistic $T:\mathscr{X}\to\mathscr{T}$ is called an *ancillary statistic* if its sampling distribution under P_θ does not depend on θ. The value of an appropriate ancillary statistic determines a subsequence of repetitions of the statistical experiment in which "the long run of trials considered is like the data" (Cox and Hinkley, 1974, p. 49).

This suggests that we should modify FFP, by replacing the unconditional sampling probabilities by probabilities conditional on an ancillary statistic, to give the *conditional frequentist principle*:

CFP: If $T:\mathscr{X}\to\mathscr{T}$ is any ancillary statistic, then $P_\theta[\rho(\theta|X)\leqslant\alpha|T(X)=t]\leqslant\alpha$ whenever $\theta\in\Theta$, $t\in\mathscr{T}$ and $0\leqslant\alpha\leqslant\alpha_0$.

The following theorem shows that, if an inference method satisfies LP, then FFP and CFP are equivalent. To simplify the statement and proof of the theorem, I assume that we use a particular version of the likelihood function,[5] which is not uniquely defined when $P_\theta(\{x\})=0$. Given a sampling model P_θ and any $t\in\mathscr{T}$, define a new

[5] This assumption can be dropped if we allow the constraint on error probabilities in CFP to be violated on a subset of $\mathscr{T}$ that has probability zero.

sampling model P_θ^t on the same sample space $\mathscr{X}$ by conditioning P_θ on the event that $T(X)=t$. Thus $P_\theta^t(B)=P_\theta[B|T(X)=t]$ for all $B\in\mathscr{A}(\mathscr{X})$. Because T is ancillary, if $T(x)=t$ then the likelihood function generated by P_θ^t can be taken to be proportional to that generated by P_θ. If P_θ and T have probability density functions f_θ and g, respectively, where g is independent of θ since T is ancillary, then P_θ^t has probability density function $f_\theta^t(x)=f_\theta(x)/g(t)$ whenever $T(x)=t$ and $g(t)>0$. Hence $f_\theta^t(x)\propto f_\theta(x)$, so that P_θ^t and P_θ generate proportional likelihood functions (given x), whenever $T(x)=t$ and $g(t)>0$. When $g(t)=0$, which has probability zero under each P_θ, the likelihood function for P_θ^t is taken to be proportional to the likelihood function for P_θ (given x). Because of this proportionality, LP implies that the two sampling models P_θ^t and P_θ must produce the same inferences.

Theorem 2.2. *Consider an inference method that produces a consistency function $\rho(\cdot|X)$ for any given sampling model with parameter θ. If the method satisfies the constraint in CFP for some statistic T (which need not be ancillary) then it also satisfies FFP. If the method satisfies LP and FFP then it satisfies CFP. Thus, assuming that the method satisfies LP, CFP is equivalent to FFP.*

Proof. Suppose that CFP holds for some T, so $P_\theta[\rho(\theta|X)\leqslant\alpha|T(X)=t]\leqslant\alpha$ for all $t\in\mathscr{T}$. The unconditional sampling probability can be expressed as the expected value of the conditional probability, averaged over all possible values of t, and therefore $P_\theta[\rho(\theta|X)\leqslant\alpha]\leqslant\alpha$. Thus FFP holds.

To prove the second statement, suppose that P_θ is a sampling model on $\mathscr{X}$ and $T:\mathscr{X}\to\mathscr{T}$ is an ancillary statistic. Define a new sampling model P_θ^t by $P_\theta^t(B)=P_\theta[B|T(X)=t]$ for all $B\in\mathscr{A}(\mathscr{X})$. Let $\rho(\theta|x)$ and $\rho^t(\theta|x)$ denote the consistency functions generated by data x when the inference method is applied to the respective sampling models P_θ and P_θ^t. When $x\in\mathscr{X}$ and $T(x)=t$, the two sampling models generate proportional likelihood functions, so that LP implies $\rho^t(\theta|x)=\rho(\theta|x)$ for all $\theta\in\Theta$. Hence, whenever $\theta\in\Theta$, $t\in\mathscr{T}$ and $0\leqslant\alpha\leqslant\alpha_0$,

$$P_\theta[\rho(\theta|X)\leqslant\alpha|T(X)=t]=P_\theta[\rho^t(\theta|X)\leqslant\alpha|T(X)=t]=P_\theta^t[\rho^t(\theta|X)\leqslant\alpha]\leqslant\alpha,$$

by applying FFP to the sampling model P_θ^t. Thus $\rho(\cdot|X)$ satisfies CFP. □

Conditional versions of the other frequentist principles (HTP, WSEP, SSEP) can be formulated in a similar way to FFP, by conditioning the sampling probabilities on $T(X)=t$. For inference methods that satisfy LP, each conditional frequentist principle is equivalent to its unconditional version. (That can be proved by modifying the proof of Theorem 2.2.) For the inference methods considered in this paper, which satisfy LP, it therefore does not matter whether we evaluate the frequentist sampling probabilities unconditionally, or conditionally on some ancillary statistic. In studying these methods, there is no need to select a particular ancillary statistic T or to explicitly consider conditional frequentist properties.

As in Section 2.4, HTP and LP together imply (through the conditional version of HTP) that, when $A \in \mathscr{A}(\Theta)$ and $\rho(A|x) \leqslant \alpha_0$, $\rho(A|x)$ is an upper bound for the p-values of certain conditional tests of H_A: $\theta \in A$, in which the test statistic is $\rho(A|x)$ and the p-values are conditional on the observed value of any ancillary statistic.

The conditional version of WSEP says that, if $C(X)$ is a credible set estimator for θ at credibility level $1-\alpha$, $0 \leqslant \alpha \leqslant \alpha_0$, and T is any ancillary statistic, then $P_\theta[\theta \in C(X)|T(X)=t] \geqslant 1-\alpha$ whenever $\theta \in \Theta$ and $t \in \mathscr{T}$. The well-known concept of a *relevant subset* (Buehler, 1959; Robinson, 1979; Lehmann, 1986; Walley, 1991, Section 7.5) involves a similar type of conditioning, but on a subset of $\mathscr{X}$ rather than an ancillary statistic. Let $C(X)$ be a credible or confidence set estimator for θ at level $1-\alpha$. A subset $B \in \mathscr{A}(\mathscr{X})$ is called a *negatively biased relevant subset* for $C(X)$ when $P_\theta(B)>0$ for all $\theta \in \Theta$ and $\sup\{P_\theta[\theta \in C(X)|X \in B]: \theta \in \Theta\} < 1-\alpha$. The frequentist interpretation (Fisher, 1956) is that B determines a recognizable subsequence of repetitions of the experiment (those in which B occurs) in which the limiting relative frequency of coverage of $C(X)$ is uniformly smaller than $1-\alpha$. Typically, the indicator function of B is not an ancillary statistic because $P_\theta(B)$ depends on θ, so the conditional version of WSEP does not rule out the possibility of negatively biased relevant subsets. But if degrees of consistency $\rho(\cdot|x)$ are identified with the posterior upper probabilities $\overline{P}(\cdot|x)$ from some coherent model, as in Section 3, and $C(X)$ is a credible set estimator for θ at credibility level $1-\alpha$, then there cannot be any negatively biased relevant subset for $C(X)$. That can be proved by modifying the proof of Lemma 7.5.6 of Walley (1991), to show that if there was a negatively biased relevant subset then the assessments $\underline{P}(C(x)|x) \geqslant 1-\alpha$ (for all $x \in \mathscr{X}$) would be incoherent.

3. Contamination neighborhoods

Let Q be any probability measure on $\mathscr{A}(\Theta)$ such that $0 < \int_\Theta L_x(\theta)Q(\mathrm{d}\theta) < \infty$ for all $x \in \mathscr{X}$. We could regard Q as a prior probability measure and use it to calculate Bayesian inferences; these inferences would satisfy LP, but they would not satisfy the frequentist principles in general. However, we can satisfy the frequentist principles by replacing the single probability measure Q by a suitably large neighborhood of it.

In this section I study the inferences produced by an ε-*contamination neighborhood* of Q (Huber, 1973; Berger and Berliner, 1986). This is defined to be the set of all probability measures of the form $(1-\varepsilon)Q + \varepsilon P$ where ε is fixed $(0<\varepsilon<1)$ and P can be any probability measure on $\mathscr{A}(\Theta)$. The ε-contamination neighborhood assigns weight $1-\varepsilon$ to the probability measure Q and allows the remaining probability ε to be distributed anywhere on Θ. The ε-contamination neighborhoods are also called *gross error models* (Huber, 1973) and *linear-vacuous mixtures* (Walley, 1991). These models are used in both frequentist and Bayesian studies of robustness (Huber, 1981; Berger and Berliner, 1986; Pericchi and Walley, 1991), and they are natural candidates to reconcile frequentist properties with coherence and the likelihood principle.

3.1. Posterior upper and lower probabilities

The posterior probability of any set A in $\mathscr{A}(\Theta)$ is maximized by assigning the free probability ε to a point θ in A which maximizes the observed likelihood $L_x(\theta)$ over A. The maximized posterior probability, which is called the *posterior upper probability* of A, is therefore given by the formula

$$\overline{P}(A|x) = \frac{\int_A L_x(\theta)Q(\mathrm{d}\theta) + \tau \sup\{L_x(\theta)\colon \theta \in A\}}{\int_\Theta L_x(\theta)Q(\mathrm{d}\theta) + \tau \sup\{L_x(\theta)\colon \theta \in A\}} \tag{3.1}$$

for all $A \in \mathscr{A}(\Theta)$, where $\tau = \varepsilon/(1-\varepsilon)$. An equivalent formula was given by Huber (1973). The conjugate lower probabilities are defined by

$$\underline{P}(A|x) = 1 - \overline{P}(A^c|x) = \frac{\int_A L_x(\theta)Q(\mathrm{d}\theta)}{\int_\Theta L_x(\theta)Q(\mathrm{d}\theta) + \tau \sup\{L_x(\theta)\colon \theta \in A^c\}}. \tag{3.2}$$

It is possible that $\sup\{L_x(\theta)\colon \theta \in A\} = \infty$ in these formulae, since L_x may be unbounded. In that case, use the conventions $\infty/\infty = 1$, and $k/\infty = 0$ if k is finite.

An upper probability $\overline{P}(A|x)$ can be interpreted as a marginally acceptable betting rate for betting *against* A, after observing the data x. This behavioral interpretation is compatible with the more abstract interpretations of $\rho(A|x)$ that were suggested in Section 2.1: $\overline{P}(A|x)$ measures the *degree of consistency* between x and $\mathrm{H}_A\colon \theta \in A$, or the *plausibility* of H_A after observing x, and it is inversely related to the *strength of evidence* against H_A. I therefore identify $\overline{P}(\cdot|x)$ with the consistency function $\rho(\cdot|x)$ considered in Section 2.

An important property of the ε-contamination model is that, under the previous regularity assumptions, the inferences produced by formulae (3.1) and (3.2) satisfy strong properties of coherence; see Walley (1991, Theorem 7.8.1).

3.2. Frequentist properties

The following theorem shows that the likelihood and frequentist principles of Section 2 are satisfied when ε is sufficiently large.

Theorem 3.1. *For each $A \in \mathscr{A}(\Theta)$, define the degree of consistency $\rho(A|x) = \overline{P}(A|x)$ to be the posterior upper probability* (3.1) *produced by an ε-contamination neighborhood of Q, where $\varepsilon \geqslant (2-\alpha_0)^{-1}$. Then inferences satisfy LP and the five frequentist principles in Section* 2: *FFP, HTP, WSEP, SSEP and CFP.*

Proof. Inferences (3.1) satisfy LP, since $\overline{P}(A|x)$ depends on $\{P_\theta\colon \theta \in \Theta\}$ only through L_x and is unchanged when L_x is multiplied by a positive constant. To verify FFP, use (3.1) to give

$$\overline{P}(\{\theta\}|x) = \frac{[Q(\{\theta\}) + \tau]L_x(\theta)}{\int_\Theta L_x(\psi)Q(\mathrm{d}\psi) + \tau L_x(\theta)} \geqslant \frac{\tau L_x(\theta)}{\int_\Theta L_x(\psi)Q(\mathrm{d}\psi) + \tau L_x(\theta)}.$$

Hence $\overline{P}(\{\theta\}|x) \leqslant \alpha$ implies that $L_x(\theta) \leqslant \tau^{-1}(\alpha/(1-\alpha)) \int_\Theta L_x(\psi) Q(\mathrm{d}\psi)$. Let $B(\theta) = \{x \in \mathscr{X}: \overline{P}(\{\theta\}|x) \leqslant \alpha\}$. Using $L_x(\theta) = f_\theta(x)$ and $\int_\mathscr{X} f_\psi(x) \nu(\mathrm{d}x) = 1$ (since f_ψ is a probability density function with respect to ν), and using Fubini's Theorem to change the order of integration,

$$
\begin{aligned}
P_\theta[\overline{P}(\{\theta\}|X) \leqslant \alpha] &= \int_{B(\theta)} f_\theta(x) \nu(\mathrm{d}x) \\
&\leqslant \int_{B(\theta)} \tau^{-1}\left(\frac{\alpha}{1-\alpha}\right) \int_\Theta f_\psi(x) Q(\mathrm{d}\psi) \nu(\mathrm{d}x) \\
&\leqslant \tau^{-1}\left(\frac{\alpha}{1-\alpha}\right) \int_\mathscr{X} \int_\Theta f_\psi(x) Q(\mathrm{d}\psi) \nu(\mathrm{d}x) \\
&= \tau^{-1}\left(\frac{\alpha}{1-\alpha}\right) \int_\Theta \int_\mathscr{X} f_\psi(x) \nu(\mathrm{d}x) Q(\mathrm{d}\psi) \\
&= \tau^{-1}\left(\frac{\alpha}{1-\alpha}\right) = \left(\frac{1-\varepsilon}{\varepsilon}\right)\left(\frac{\alpha}{1-\alpha}\right) \\
&\leqslant (1-\alpha_0)\left(\frac{\alpha}{1-\alpha}\right) \leqslant \alpha,
\end{aligned}
$$

using $\varepsilon \geqslant (2-\alpha_0)^{-1}$ and $0 \leqslant \alpha \leqslant \alpha_0$ for the last two inequalities. Thus FFP holds. Also upper probabilities satisfy the monotonicity condition (2.1). Using Theorems 2.1 and 2.2, inferences satisfy the other frequentist principles HTP, WSEP, SSEP and CFP. □

It is remarkable that the sufficient condition for the frequentist principles, $\varepsilon \geqslant (2-\alpha_0)^{-1}$, is independent of both the sampling models and the probability measure Q. For example, Q can be taken to be a degenerate distribution which assigns probability one to a point θ_0, provided $L_x(\theta_0) > 0$ for all $x \in \mathscr{X}$.

The lower bound $\varepsilon_0 = (2-\alpha_0)^{-1}$ in Theorem 3.1 is the sharpest possible bound that is independent of the size of Θ. If Θ is any continuous parameter space, L is any bounded likelihood function on Θ, Q is any continuous probability distribution on Θ, and inferences are calculated from an ε-contamination neighborhood of Q with $\varepsilon < \varepsilon_0$, then there are 'unfavorable' sampling models which could have generated the given likelihood function L but for which FFP is not satisfied (Walley, 1998, Example 4.3). For continuous Θ, it is therefore not possible to improve the lower bound in Theorem 3.1 by exploiting the observed likelihood function or by choosing a particular distribution Q.

The lower bound ε_0 can be improved when Θ has finite cardinality r. Let Q be the uniform probability distribution on Θ, which is optimal in the sense that it minimizes the threshold value ε_0' below. Then the proof of Theorem 3.1 can be modified to show that inferences from an ε-contamination neighborhood of Q satisfy LP, FFP and the other frequentist principles, provided that $\varepsilon \geqslant \varepsilon_0' = (r-2+\alpha_0)/(r-1)(2-\alpha_0)$. Here ε_0' is an increasing function of α_0 and of r. It is smaller than the lower bound ε_0 in

Theorem 3.1 and approaches ε_0 as $r \to \infty$, but ε_0' can be much smaller than ε_0 when r is small, e.g., when $r=2$ and $\alpha_0=0.1$, $\varepsilon_0=10/19$ but $\varepsilon_0'=1/19$. See Walley and Moral (1999) for other properties of these models in the case of finite Θ.

3.3. Choice of α_0 and Q

Inferences are quite insensitive to the choice of α_0 and Q. The value $\alpha_0=0.1$, which reflects the usual practice in frequentist inference of regarding p-values as significant evidence only when they are smaller than 0.1, gives $\varepsilon_0=(2-\alpha_0)^{-1}=\frac{10}{19}$, slightly bigger than $\frac{1}{2}$. The corresponding value of τ is $\tau_0=(1-\alpha_0)^{-1}=\frac{10}{9}$. It makes little difference to inferences if α_0 is changed from 0.1 to 0.2, 0.05 or 0: the effect is to change τ in (3.1) from $\frac{10}{9}$ to $\frac{5}{4}$, $\frac{20}{19}$ or 1. It is necessary that $\alpha_0<1$, since $\alpha_0=1$ gives $\varepsilon_0=1$ and vacuous inferences.

Whatever positive value of α_0 is used, ε must be greater than $\frac{1}{2}$ and τ greater than 1. Under any such ε-contamination model, every simple hypothesis $\{\theta\}$ has prior upper probability greater than $\frac{1}{2}$, since $\overline{P}(\{\theta\})=(1-\varepsilon)Q(\{\theta\})+\varepsilon \geqslant \varepsilon > \frac{1}{2}$ for all $\theta \in \Theta$. In behavioral terms, this means that we are initially unwilling to bet against any point hypothesis unless the odds are better than even money. This is a kind of prior ignorance property. It indicates that the set of prior probability measures is very large and that prior beliefs about the parameter are very cautious.

Inferences are also insensitive to the choice of Q. Because the ε-contamination neighborhoods are so large when $\varepsilon > \frac{1}{2}$, any two measures Q_1 and Q_2 produce overlapping neighborhoods and compatible inferences. It may seem that the choice of Q introduces a subjective element into the model, contrary to frequentist aims. But Q could be determined objectively, e.g., by taking it to be a *uniform* probability distribution when the parameter space is finite or bounded. It is unclear how Q should be chosen for unbounded parameter spaces. In any case, the frequentist properties hold irrespective of Q.

3.4. Hypothesis testing and set estimation

When Θ is a continuous space and Q has no point masses, the consistency function is

$$\overline{P}(\{\theta\}|x)=\frac{L_x(\theta)}{\tau^{-1}\int_\Theta L_x(\psi)Q(\mathrm{d}\psi)+L_x(\theta)}. \tag{3.3}$$

This is an order-preserving transformation of the likelihood function L_x. The parameter value with maximum degree of consistency is the maximum likelihood estimate. When θ is real valued, (3.3) can be graphed to show the posterior plausibilities of simple hypotheses.

To test a hypothesis H_A: $\theta \in A$ using the contamination model, we simply calculate $\overline{P}(A|x)$, using (3.1) with $\tau=(1-\alpha_0)^{-1}$, and interpret it as a marginally acceptable betting rate for betting against H_A after observing x. If $\overline{P}(A|x)$ is small then there is

strong evidence against H_A, both in the frequentist sense that a particular test of H_A has p-value less than or equal to $\overline{P}(A|x)$ (Section 2.4), and in the robust Bayesian sense that the posterior probability of H_A, under any prior distribution in the contamination class, is less than or equal to $\overline{P}(A|x)$.

In this approach, simple (point) hypotheses do not require any different treatment from composite hypotheses. For every null hypothesis H_A, $\overline{P}(A|x)$ is an upper bound for a frequentist p-value for testing H_A. That can be achieved even for a simple hypothesis H_0: $\theta = \theta_0$ because, by (3.1), the contamination model assigns a positive value to $\overline{P}(\{\theta_0\}|x)$ whenever $L_x(\theta_0) > 0$.

There are two ways to construct level $1-\alpha$ confidence sets from the contamination model. The simpler method, and the one which produces smaller sets, is to define

$$C^*(x) = \{\theta \in \Theta\colon \overline{P}(\{\theta\}|x) > \alpha\}, \tag{3.4}$$

where $\overline{P}(\{\theta\}|x)$ is computed from (3.1) using $\tau = (1-\alpha)^{-1}$. When Q has no point masses, $\overline{P}(\{\theta\}|x)$ is given by (3.3) and we obtain

$$C^*(x) = \left\{\theta \in \Theta\colon L_x(\theta) > \alpha \int_\Theta L_x(\psi) Q(\mathrm{d}\psi)\right\}. \tag{3.5}$$

Thus $C^*(x)$ consists of those parameter values whose likelihood exceeds a threshold. Such sets are called *likelihood sets*, so I call $C^*(x)$ a *likelihood confidence set*. Because the contamination model satisfies SSEP, the set estimator $C^*(X)$ defined by (3.4) or (3.5) is a confidence set estimator at confidence level $1-\alpha$.

The second method, using WSEP, is to define $C(X)$ to be a credible set estimator for θ at credibility level $1-\alpha$. By (3.2), this requires that, for all $x \in \mathscr{X}$,

$$\underline{P}(C(x)|x) = \frac{\int_{C(x)} L_x(\theta) Q(\mathrm{d}\theta)}{\int_\Theta L_x(\theta) Q(\mathrm{d}\theta) + \tau \sup\{L_x(\theta)\colon \theta \in C(x)^c\}} \geqslant 1-\alpha. \tag{3.6}$$

If Q is a continuous distribution and $C(x)$ is required to satisfy (3.6), then the 'size' of $C(x)$, measured by $Q(C(x))$, is minimized when $C(x)$ is a likelihood set,[6] i.e., $C(x) = \{\theta \in \Theta\colon L_x(\theta) > \kappa\}$, where the threshold κ can be determined by solving $\underline{P}(C(x)|x) = 1-\alpha$ numerically. It was shown in the proof of Theorem 2.1 that every credible set at level $1-\alpha$ must contain $C^*(x)$, and usually strict containment is needed.

4. Examples

4.1. Bernoulli trials (Walley, 1996, Section 4)

Suppose that a sequence of marbles is drawn from a bag with replacement, and we observe five non-red marbles followed by one red. Let θ be the unknown chance

[6] To prove that, let q be the minimized value of $Q(C(x))$. Subject to $Q(C(x)) = q$, the numerator of (3.6) is maximized and the denominator is minimized when $C(x)$ is a likelihood set.

of drawing a red marble. The observed likelihood function is $L_x(\theta) = \theta(1-\theta)^5$. Let Q be the uniform probability distribution on $\Theta = (0,1)$, so $\int_\Theta L_x(\theta)Q(\mathrm{d}\theta) = \int_\Theta \theta(1-\theta)^5\,\mathrm{d}\theta = \frac{1}{42}$.

First, consider testing the hypothesis H_0: $\theta \geqslant \frac{1}{2}$ versus H_1: $\theta < \frac{1}{2}$. Let $A = \{\theta \in \Theta: \theta \geqslant \frac{1}{2}\}$. Using $\varepsilon = \frac{10}{19}$, which corresponds to $\alpha_0 = 0.1$, in (3.1), the posterior upper probability is $\overline{P}(A|x) = 0.458$. This value is small enough to lead us to accept an even-money bet against H_0, but it indicates only very weak evidence against H_0. The posterior lower probability, given by (3.2), is $\underline{P}(A|x) = 0.015$, which indicates that there is no evidence in favor of H_0.

These inferences are more cautious than standard frequentist or Bayesian inferences. Frequentist inferences depend on the stopping rule for the experiment, which has not been specified. Assuming that it was decided in advance to draw six marbles (binomial sampling), the standard frequentist p-value for testing H_0 is $\frac{7}{64} = 0.109$, the chance (when $\theta = \frac{1}{2}$) of obtaining no more than one red marble in six drawings. A p-value greater than 0.1 is interpreted as no more than weak evidence against H_0. But if we assume that the experiment was designed to continue until the first red marble was drawn (negative binomial sampling) then the p-value is $\frac{1}{32} = 0.031$, the chance (when $\theta = \frac{1}{2}$) that it will take at least six drawings to obtain the first red marble. This would be interpreted as moderately strong evidence against H_0. Bayesian methods using the standard 'noninformative' priors would conclude that there is strong evidence against H_0: the posterior probability of A is $\frac{1}{16}$ using the uniform prior for θ or $\frac{1}{32}$ using Haldane's improper prior.

According to the contamination model, H_0 is quite consistent with observing one red marble in six drawings. Suppose that we make n drawings from the bag and obtain only one red marble. How large must n be for the data to be inconsistent with H_0? Using the ε-contamination model with $\varepsilon = \frac{10}{19}$, we find that $\overline{P}(A|x)$ is 0.112 when $n = 10$, and 0.042 when $n = 12$. So there is quite strong evidence against H_0 when $n = 10$, leading us to bet at odds of about 8 to 1 against H_0, and very strong evidence against H_0 when $n = 12$. Nevertheless, these inferences still appear to be very cautious.

Now consider set estimation of θ. By (3.5), the likelihood interval derived from the ε-contamination model is $C^*(x) = \{\theta \in \Theta: L_x(\theta) > \alpha/42\}$, which is a confidence interval for θ at confidence level $1-\alpha$. This gives $C^*(x) = (0.0012, 0.722)$ as a 95% confidence interval and $C^*(x) = (0.0024, 0.677)$ as a 90% confidence interval for θ.

These can be compared with standard frequentist and Bayesian intervals. The method of Clopper and Pearson (1934), based on the assumption of binomial sampling, gives $(0.0042, 0.641)$ as a conservative 95% confidence interval for θ. Bayesian 95% highest posterior density credible intervals are considerably shorter. For example, the uniform prior distribution for θ gives $(0.013, 0.527)$ and Haldane's improper prior gives $(0, 0.451)$ as 95% credible intervals for θ.

As the sample size n becomes large, the width of the likelihood confidence intervals $C^*(x)$ tends to zero, but at a slightly slower rate than the standard binomial confidence intervals. Writing r_n for the observed relative frequency of successes in a sample of n, a large-sample approximation to the standard binomial confidence interval is

$(r_n - z_{\alpha/2}\delta_n, r_n + z_{\alpha/2}\delta_n)$, where $\delta_n = n^{-1/2}[r_n(1-r_n)]^{1/2}$ and $z_{\alpha/2}$ is the upper quantile of a standard normal distribution. A comparable large-sample approximation to $C^*(x)$ is obtained by replacing $z_{\alpha/2}$ by $\{-\log(2\pi\alpha^2 r_n[1-r_n]n^{-1})\}^{1/2}$. The asymptotic behavior of the intervals $C^*(x)$ is therefore slightly different from the behavior of the standard frequentist, Bayesian and likelihood intervals, which agree asymptotically. The ratio of the length of $C^*(x)$ to the length of the standard interval increases very slowly with n. Taking $\alpha = 0.05$ and $r_n = 0.3$, for example, this ratio is 1.8 when $n = 10^3$, 2.2 when $n = 10^6$, and 2.7 when $n = 10^{10}$. Since both intervals are short when n is very large, the difference in asymptotic behavior appears to have no practical importance.

4.2. Clinical data (Begg, 1990)

In a clinical trial discussed in Begg (1990), an adaptive randomized design (a type of 'play the winner' rule) was used to allocate patients to treatments. The design produced a highly unbalanced allocation: 11 patients were given the experimental treatment, all outcomes being successful, and only one patient received the control treatment, which was a failure. Let θ_c and θ_e denote the chances of success under the control and experimental treatments, respectively, with $\Theta = (0,1)^2$. The likelihood function is proportional to $L_x(\theta_c, \theta_e) = (1-\theta_c)\theta_e^{11}$. Here the proportionality constant depends on the particular design that was used to allocate patients to treatments but, according to LP, the design can be ignored in making inferences about the parameters.

The aim of the trial was to test whether the experimental treatment had a higher chance of success than the control. We therefore wish to test H_0: $\theta_c \geqslant \theta_e$ versus H_1: $\theta_c < \theta_e$. Let $A = \{(\theta_c, \theta_e) \in \Theta: \theta_c \geqslant \theta_e\}$, and let Q be the joint uniform probability distribution for (θ_c, θ_e). By double integration of the likelihood function, $\int\int_\Theta L_x(\theta_c, \theta_e)\, d\theta_c\, d\theta_e = \frac{1}{24}$ and $\int\int_A L_x(\theta_c, \theta_e)\, d\theta_c\, d\theta_e = \frac{1}{2184}$. Also $\sup\{L_x(\theta_c, \theta_e): \theta_c \geqslant \theta_e\} = \max\{(1-\theta)\theta^{11}: 0 \leqslant \theta \leqslant 1\} = 11^{11}/12^{12} = 0.032$, achieved by $\theta_c = \theta_e = 11/12$. Using the ε-contamination model with $\varepsilon = 10/19$, (3.1) gives the posterior upper probability $\overline{P}(A|x) = 0.466$. The conclusion is that, although there is sufficient evidence to support a bet against H_0 at even money, there is not strong evidence against H_0. The posterior lower probability of H_0, found from (3.2), is $\underline{P}(A|x) = 0.00040$, an extremely small value which shows that (not surprisingly) the data provide no evidence in favor of H_0.

In Begg (1990) and the ensuing discussion, a variety of p-values, from 0.00049 to 0.62, were suggested, based on different test statistics and conditioning variables. Most of these p-values were computed by conditioning on the observed sequence of responses. For example, a standard randomization test defines the p-value to be the probability of the observed sequence of treatment allocations conditional on the observed sequence of responses, which is found to be $1/22 = 0.045$, apparently indicating moderately strong evidence against H_0. [See the comments of Cox and Royall in the discussion of Begg (1990).] However, Begg shows that if we also condition on the total numbers of patients that were allocated to each treatment then the p-value rises to 0.62, which indicates no evidence at all against H_0. It is not clear in this

example which of the frequentist tests is most appropriate, and since they lead to different conclusions it is unclear how a frequentist should assess the strength of evidence against H_0.

A Bayesian analysis based on a uniform prior distribution for (θ_c, θ_e) gives posterior probability $P(A|x) = 24/2184 = 1/91 = 0.011$, which indicates extremely strong evidence against H_0. This conclusion seems unreasonable. Although the data provide some evidence against H_0, the evidence is not conclusive because only one patient received the control treatment.

4.3. Generalized likelihood ratio

To get some further insight into the properties of consistency functions, consider a very simple example of a sampling model, P_θ^*, which can generate a likelihood function proportional to L_x, where L_x is any bounded likelihood function on a parameter space Θ. Define P_θ^* on $\mathscr{X} = \{x, x'\}$ by

$$P_\theta^*(x) = L_x(\theta)/\sup\{L_x(\psi)\colon \psi \in \Theta\}, \qquad P_\theta^*(x') = 1 - P_\theta^*(x), \quad \text{for all } \theta \in \Theta. \tag{4.1}$$

Since $\sup\{P_\theta^*(x)\colon \theta \in \Theta\} = 1$, P_θ^* maximizes $P_\theta(x)$ amongst all sampling models that generate a likelihood function (given x) proportional to L_x.

Consider the problem of testing a general hypothesis $H_A\colon \theta \in A$, where $A \in \mathscr{A}(\Theta)$. Given any consistency function $\rho(\cdot|x)$ that satisfies HTP, we can use $\rho(A|x)$ as a test statistic for testing H_A. Assuming that x is observed, the p-value of the test is $\sup\{P_\theta^*[\rho(A|X) \leqslant \rho(A|x)]\colon \theta \in A\}$. Using the trivial inequality $P_\theta^*[\rho(A|X) \leqslant \rho(A|x)] \geqslant P_\theta^*(x)$, the p-value of the test is at least

$$\sup\{P_\theta^*(x)\colon \theta \in A\} = \frac{\sup\{L_x(\theta)\colon \theta \in A\}}{\sup\{L_x(\theta)\colon \theta \in \Theta\}}. \tag{4.2}$$

The right-hand side of (4.2) defines the *generalized likelihood ratio statistic* for testing H_A, which will be denoted by $\lambda(A|x)$.

Now consider a general sampling model. If a consistency function $\rho(\cdot|x)$ satisfies both LP and HTP, and $\rho(A|x) \leqslant \alpha_0$, then $\rho(A|x) \geqslant \lambda(A|x)$. In this case the degree of consistency $\rho(A|x)$ is an upper bound for the generalized likelihood ratio $\lambda(A|x)$. This follows from the facts that (a) using LP, $\rho(A|x)$ must agree with the value it assumes under the sampling model P_θ^* defined in (4.1); (b) using HTP, $\rho(A|x)$ is at least as large as the p-value of the above test of H_A (see Section 2.4); and (c) by the above argument, the p-value under P_θ^* is at least $\lambda(A|x)$.

In Example 4.1, after observing one success in six trials we find that $\lambda(A|x) = 0.233$. This would be interpreted as only weak evidence against H_A, which agrees with the interpretation of $\overline{P}(A|x) = 0.458$ in Example 4.1. In Example 4.2 there is a bigger discrepancy between $\lambda(A|x)$ and $\overline{P}(A|x)$. There $\lambda(A|x) = 0.032$, which is much smaller

than the upper probability $\overline{P}(A|x) = 0.466$ and would be interpreted as strong evidence against H_A. The Birnbaum–Cox–Hinkley example (Example 1.1) shows that using the generalized likelihood ratio statistic as a measure of strength of evidence can violate the frequentist principles and lead to unreasonable conclusions.

4.4. Normal location

Suppose that n independent observations $x_1, x_2, \ldots, x_n$ are obtained from a normal distribution with known variance σ^2 and unknown mean μ. Here $x = (x_1, x_2, \ldots, x_n)$, $\theta = \mu$, and $\Theta = \mathbb{R}$. The likelihood function is proportional to $L_x(\mu) = \exp[-\frac{1}{2}n\sigma^{-2}(\mu - \bar{x})^2]$, where $\bar{x}$ is the sample mean.

Because Θ is unbounded, there is no uniform or 'noninformative' probability distribution Q. To simplify the analysis, I take Q to be a normal distribution with mean ω and variance υ^2. Writing $\xi = (\bar{x} - \omega)/\upsilon$ and $\kappa_n^2 = \sigma^2/n\upsilon^2$ gives $\int_{-\infty}^{\infty} L_x(\mu)Q(\mathrm{d}\mu) = (1 + \kappa_n^{-2})^{-1/2} \exp[-\frac{1}{2}\xi^2/(1 + \kappa_n^2)]$.

The likelihood confidence interval for μ at level $1 - \alpha$ is

$$C^*(x) = \left\{ \mu \in \mathbb{R}\colon L_x(\mu) > \alpha \int_{-\infty}^{\infty} L_x(\mu)Q(\mathrm{d}\mu) \right\}$$
$$= (\bar{x} - n^{-1/2}\sigma\zeta, \bar{x} + n^{-1/2}\sigma\zeta), \tag{4.3}$$

where

$$\zeta^2 = \frac{\xi^2}{1 + \kappa_n^2} + \log(1 + \kappa_n^{-2}) - 2\log\alpha. \tag{4.4}$$

We can see from (4.4) that $C^*(x)$ may be considerably wider than the standard confidence intervals. Since the first two terms in (4.4) are nonnegative, we always have $\zeta^2 \geqslant -2\log\alpha$. (This lower bound for ζ^2 is the limit of (4.4) as $\upsilon \to 0$ with $\omega = \bar{x}$.) When $\alpha = 0.05$, for example, the lower bound for ζ is $(-2\log\alpha)^{1/2} = 2.45$, whereas the corresponding value for the standard 95% confidence interval is 1.96. This means that the 95% likelihood confidence intervals are always at least 1.25 times as wide as the standard 95% confidence intervals, and they can be much wider when $|\xi|$ is large or κ_n is small. For example, the moderate values $\alpha = 0.05$, $\upsilon^2/\sigma^2 = 2$, $n = 10$ and $\xi = 2$ give $\zeta = 3.58$. One reason that the intervals $C^*(x)$ are so wide is that they are valid confidence intervals under all stopping rules that could have generated L_x, whereas the standard intervals are based on the assumption that the sample size n is fixed.

As the sample size $n \to \infty$ with the other parameters $(\sigma^2, \upsilon^2, \xi, \alpha)$ fixed, the width of $C^*(x)$ tends to zero at a slightly slower rate than the standard intervals, at rate $(\log n)^{1/2}n^{-1/2}$ rather than $n^{-1/2}$, as in the Bernoulli Example 4.1. In fact (4.4) implies that, as $n \to \infty$,

$$\zeta^2 = \log n + \xi^2 + \log(\upsilon^2/\sigma^2) - 2\log\alpha + \mathrm{O}(n^{-1}). \tag{4.5}$$

5. Normalized-likelihood inferences

In this section I outline another method of inference that satisfies both the likelihood and frequentist principles. It produces similar inferences to the contamination model (Section 3), but it is more in the spirit of likelihood inference (Barnard, 1967; Edwards, 1972).

Again let Q be any probability measure on $\mathscr{A}(\Theta)$. Define the consistency function $\rho(\cdot|x)$ on Θ to be a normalized version of the likelihood function L_x,

$$\rho(\theta|x) = \frac{L_x(\theta)}{\int_\Theta L_x(\psi)Q(\mathrm{d}\psi)}. \tag{5.1}$$

The normalizing constant $\int_\Theta L_x(\psi)Q(\mathrm{d}\psi) = \int_\Theta f_\psi(x)Q(\mathrm{d}\psi)$ is the marginal probability density (with respect to ν) at x under probability measure Q.

It is proved below (in Theorem 5.1) that function (5.1) satisfies FFP with $\alpha_0 = 1$. By Theorems 2.1 and 2.2, to satisfy the other frequentist principles of Section 2, we merely need to define $\rho(A|x)$ in such a way as to satisfy the monotonicity condition (2.1). To obtain the strongest possible inferences which satisfy the frequentist principles, we need to choose the smallest consistency function $\rho(\cdot|x)$ that agrees with (5.1) on the singleton sets and satisfies (2.1). This function is defined, for all $A \in \mathscr{A}(\Theta)$ and $x \in \mathscr{X}$, by

$$\rho(A|x) = \sup\{\rho(\theta|x)\colon \theta \in A\} = \frac{\sup\{L_x(\theta)\colon \theta \in A\}}{\int_\Theta L_x(\psi)Q(\mathrm{d}\psi)}. \tag{5.2}$$

5.1. Frequentist properties

Theorem 5.1. *The consistency function $\rho(\cdot|x)$, defined in* (5.2), *satisfies LP and the five frequentist principles of Section* 2 (*FFP*, *HTP*, *WSEP*, *SSEP and CFP*) *with* $\alpha_0 = 1$.

Proof. LP holds because (5.2) depends on $\{P_\theta\colon \theta \in \Theta\}$ only through L_x and is unchanged when L_x is multiplied by a positive constant. To verify FFP, let $\theta \in \Theta$, $0 \leqslant \alpha \leqslant 1$, and $B(\theta) = \{x \in \mathscr{X}\colon \rho(\theta|x) \leqslant \alpha\}$. Then, using (5.1) and using Fubini's theorem to change the order of integration,

$$P_\theta[\rho(\theta|X) \leqslant \alpha] = \int_{B(\theta)} f_\theta(x)\nu(\mathrm{d}x) \leqslant \int_{B(\theta)} \alpha \int_\Theta f_\psi(x)Q(\mathrm{d}\psi)\nu(\mathrm{d}x)$$

$$\leqslant \alpha \int_\mathscr{X} \int_\Theta f_\psi(x)Q(\mathrm{d}\psi)\nu(\mathrm{d}x) = \alpha \int_\Theta \int_\mathscr{X} f_\psi(x)\nu(\mathrm{d}x)Q(\mathrm{d}\psi) = \alpha.$$

Thus FFP holds with $\alpha_0 = 1$. Since (5.2) satisfies the monotonicity condition (2.1), it follows from Theorems 2.1 and 2.2 that HTP, WSEP, SSEP and CFP hold with $\alpha_0 = 1$. □

Instead of using $\rho(A|x)$ to measure the degree of consistency, we could truncate the values at 1 and use the upper probability measure $\overline{P}^*(A|x) = \min\{\rho(A|x), 1\}$. Since the

frequentist principles refer only to degrees of consistency smaller than one, $\overline{P}^*(\cdot|x)$ satisfies the frequentist principles in the same way as $\rho(\cdot|x)$. The function $\overline{P}^*(\cdot|x)$ is *coherent*, in the sense of Walley (1991), when regarded as an upper probability function for a fixed x. This property is much weaker than the coherence property of the contamination model, that the whole collection of upper probabilities $\{\overline{P}(\cdot|x)\colon x \in \mathscr{X}\}$ and $\{P_\theta\colon \theta \in \Theta\}$ is coherent.

5.2. Hypothesis testing and set estimation

A general hypothesis $\mathrm{H}_A\colon \theta \in A$ can be tested by calculating $\rho(A|x)$ from (5.2) and interpreting it as a measure of the consistency of H_A with x. The data are inconsistent with H_A when $\rho(A|x)$ is substantially smaller than 1. Since HTP is satisfied, $\rho(A|x)$ is an upper bound for the p-value of a particular test of H_A, and an upper bound for the generalized likelihood ratio (Sections 2.4 and 4.3). From (5.2), $\rho(A|x)$ is also an upper bound for the Bayesian posterior probability $Q(A|x)$ that results from the prior distribution Q.

Also $\rho(A|x)$ is closely related to the posterior upper probability produced by the ε-contamination model, which by (3.1) can be written as

$$\overline{P}(A|x) = \frac{Q(A|x) + \tau\rho(A|x)}{1 + \tau\rho(A|x)}. \tag{5.3}$$

It follows from (5.3) that $\rho(A|x) < \overline{P}(A|x)$ whenever $0 < \overline{P}(A|x) < \alpha_0$ and $\varepsilon \geqslant (2-\alpha_0)^{-1}$. In that case $\rho(A|x)$ produces slightly more informative inferences than $\overline{P}(A|x)$.

Two methods of defining a set estimator were discussed in Section 2.5. These methods are different in general, but for the particular consistency function (5.2) they are essentially the same. A set estimator $C(X)$ for θ is a credible set estimator at credibility level $1 - \alpha$ when $\rho(C(x)^c|x) \leqslant \alpha$ for all $x \in \mathscr{X}$. Using (5.2), this condition is equivalent to: $\rho(\theta|x) \leqslant \alpha$ for all $\theta \in C(x)^c$ and $x \in \mathscr{X}$. Hence the smallest credible set for θ at credibility level $1 - \alpha$ is the likelihood confidence set $C^*(x) = \{\theta \in \Theta\colon \rho(\theta|x) > \alpha\} = \{\theta \in \Theta\colon L_x(\theta) > \alpha \int_\Theta L_x(\psi) Q(\mathrm{d}\psi)\}$, which agrees with the confidence set (3.5) produced by the contamination model.

5.3. Examples

Inferences in the examples of Section 4 are essentially the same for the normalized-likelihood model as for the contamination model. Consider Example 4.1, where one success is observed in six Bernoulli trials. Again let Q be the uniform distribution on $\Theta = (0, 1)$, with $\int_\Theta L_x(\theta) Q(\mathrm{d}\theta) = 1/42$. By (5.1), the consistency function is $\rho(\theta|x) = 42\theta(1 - \theta)^5$. To test $\mathrm{H}_A\colon \theta \in A$, where $A = [\frac{1}{2}, 1)$, we calculate $\rho(A|x) = \sup\{\rho(\theta|x)\colon \theta \geqslant \frac{1}{2}\} = 42(\frac{1}{2})^6 = 21/32 = 0.656$, which indicates little evidence against H_A. If instead we observed one success in 10 Bernoulli trials, we would obtain $\rho(A|x) = 0.107$. The shortest credible interval with credibility $1-\alpha$ is $C^*(x)$, the interval given in Example 4.1.

For the clinical data in Example 4.2, H_0 has degree of consistency $\rho(A|x) = \sup\{L_x(\theta_c, \theta_e)\colon \theta_c \geqslant \theta_e\} / \int\int_\Theta L_x(\theta_c, \theta_e)\, d\theta_c\, d\theta_e = 0.0320 \times 24 = 0.768$, indicating that H_0 is consistent with the data.

5.4. Finite parameter spaces

Suppose that Θ has finite cardinality r. In that case, as with the contamination model, the consistency function (5.2) can be sharpened. Let Q be the uniform probability distribution on Θ. Then (5.1) gives $\rho(\theta|x) = L_x(\theta)/[r^{-1}\sum_{\psi\in\Theta} L_x(\psi)]$. Define a sharper consistency function by

$$\rho'(\theta|x) = \frac{L_x(\theta)}{(r-1)^{-1}\sum_{\psi\neq\theta} L_x(\psi)} \tag{5.4}$$

and extend it to subsets of Θ by $\rho'(A|x) = \sup\{\rho'(\theta|x)\colon \theta \in A\}$. This consistency function satisfies LP, and the frequentist principles can be verified by modifying the proof of Theorem 5.1. Also $\rho'(\theta|x) < \rho(\theta|x)$ whenever $0 < \rho(\theta|x) < 1$, so the new function does produce sharper inferences in the cases of interest. Because the denominator in (5.4) depends on θ, $\rho'(\cdot|x)$ is not a normalized version of the likelihood function.

In the case $\Theta = \{\theta, \psi\}$, we obtain $\rho'(\theta|x) = L_x(\theta)/L_x(\psi)$, the observed likelihood ratio. It is well known that the likelihood ratio has good frequentist properties in the case $r = 2$ (e.g., see Birnbaum, 1969, p. 129).

6. More powerful inferences from Bernoulli data

The examples in Sections 4 and 5 show that the methods considered there produce cautious inferences: degrees of consistency tend to be large and credible intervals tend to be wide. To satisfy both LP and FFP, inferences need to be sufficiently conservative to allow for all conceivable ways of embedding the observed likelihood function in a complete sampling model. From a frequentist point of view, the methods have relatively low power. Frequentist theories generally include some requirement that not only should statistical procedures satisfy frequentist principles like those in Section 2, but also they should be as powerful as possible. Maximizing power is emphasized especially in the Neyman–Pearson theory.

To obtain more powerful and less conservative inferences, it appears that we must give up strict adherence to either LP or FFP. The approach that I favor is to retain LP and to look for methods with good frequentist properties, albeit weaker than FFP, and reasonably good power. I will give one example of this approach, concerning Bernoulli data.

6.1. Imprecise beta model

Suppose that the statistical data are the outcomes of a sequence of Bernoulli trials, the parameter θ is the chance of success in a single trial, and $\Theta = (0, 1)$. Let X_n

and Y_n, respectively, denote the number of successes and failures in the first n trials, $X_n + Y_n = n$. Suppose that the experiment is stopped as soon as (X_n, Y_n) reaches some set $\mathscr{S}$. Say that the stopping rule for the experiment is *monotone* when $\mathscr{S}$ satisfies the condition: if $(u, v) \in \mathscr{S}$, $u \leqslant x$ and $v \leqslant y$ then $(x, y) \in \mathscr{S}$. The most common stopping rules, including the binomial and negative binomial, are monotone. An example of a stopping rule that is not monotone is 'stop as soon as $|X_n - Y_n| \geqslant k$', where k is a positive integer, since $(k, 0) \in \mathscr{S}$ but $(k, k) \notin \mathscr{S}$.

It was suggested by Bernard (1996) and Walley (1996) that a particular imprecise probability model, the *imprecise beta model* (IBM) with hyperparameter 1, should be used to make inferences from Bernoulli data; see also Walley (1991, Section 5.3). The prior IBM is the following set of beta probability density functions for θ:

$$\{\pi_t\colon 0 < t < 1\}, \quad \text{where } \pi_t \text{ is beta}(t, 1-t),\ \pi_t(\theta) \propto \theta^{t-1}(1-\theta)^{-t}. \tag{6.1}$$

I want to show here that inferences from the IBM have good frequentist properties under all monotone stopping rules. Suppose that the degree of consistency $\rho(A|x)$ is identified with $\overline{P}(A|x)$, the posterior upper probability from the IBM. Since $\overline{P}(\{\theta\}|x) = 0$ (beta distributions are absolutely continuous), inferences from the IBM do not satisfy FFP and SSEP, and nor do they satisfy HTP and WSEP in complete generality. But assuming that the stopping rule is monotone, the following results show that inferences do satisfy HTP for one-sided tests, and WSEP for one-sided and equitailed two-sided credible intervals. These results cover the most common types of inference.

6.2. Hypothesis testing

First, consider testing a one-sided hypothesis H_0: $\theta \leqslant \theta_0$ against H_1: $\theta > \theta_0$. Under the IBM, the posterior upper probability of H_0, after observing x successes and y failures, is given by the formulae

$$\overline{P}(H_0|x, y) = P[\text{beta}(x, y+1) \leqslant \theta_0] = P[\text{binomial}(x+y, \theta_0) \geqslant x], \tag{6.2}$$

using a basic relationship between the binomial and beta distributions. It follows from the last formula that $\overline{P}(H_0|x, y)$ is equal to the p-value for testing H_0, assuming that the stopping rule is binomial ('stop after $x + y$ trials'). The next theorem shows that $\overline{P}(H_0|x, y)$ is an upper bound for the p-value, assuming only that the stopping rule is monotone. Its proof relies on the following property of monotone stopping rules.

Lemma 6.1. *Let $\mathscr{B}$ denote the boundary of $\mathscr{S}$, i.e., the set of all points in $\mathscr{S}$ which can be reached from outside $\mathscr{S}$ by making one extra observation. Suppose that the stopping rule is monotone and $(x, y) \in \mathscr{B}$. Consider a sequence of Bernoulli trials with 'stopping point' $(u, v) \in \mathscr{B}$. Imagine that the sequence is continued until at least $x + y$ observations have been made (continuing past the 'stopping point', if necessary), and suppose that the first $x + y$ observations produce fewer than x successes. Then $u \leqslant x$ and $v \geqslant y$.*

Proof. Suppose that $u > x$. Then the stopping point has not been reached after $x + y$ observations, in which there are more than y failures, hence $v > y$. Since $(x, y) \in \mathscr{S}$, monotonicity implies $(u-1, v) \in \mathscr{S}$ and $(u, v-1) \in \mathscr{S}$. But the sequence (X_n, Y_n) must pass through either $(u-1, v)$ or $(u, v-1)$ to reach (u, v), so the sequence must stop before reaching (u, v), which contradicts the assumption $(u, v) \in \mathscr{B}$. Thus $u \leqslant x$. Similarly, if $v < y$ then we must stop before observation $x+y$, and $u < x$. Hence $(u, v) \in \mathscr{S}$ implies $(x-1, y) \in \mathscr{S}$ and $(x, y-1) \in \mathscr{S}$ by monotonicity, and so $(x, y) \notin \mathscr{B}$, contradicting the assumption. □

Theorem 6.1. *Let T be a test statistic for testing* H_0: $\theta \leqslant \theta_0$ *against* H_1: $\theta > \theta_0$, *where $T(u, v)$ measures the strength of evidence against* H_0 *provided by observations (u, v) and satisfies the weak condition: if $u \leqslant w$, $v \geqslant z$ and $(u, v) \neq (w, z)$ then $T(u, v) < T(w, z)$. If (x, y) are the observed numbers of successes and failures, the p-value for testing* H_0 *is defined to be* $\sup\{P_\theta[T(X, Y) \geqslant T(x, y)]: \theta \leqslant \theta_0\}$. *Then, for every monotone stopping rule, the p-value is no greater than $\overline{P}(\mathrm{H}_0 | x, y)$. Here the bound $\overline{P}(\mathrm{H}_0 | x, y)$ is the sharpest possible, since it is achieved when the stopping rule is binomial.*

Proof. Let $(x, y) \in \mathscr{B}$ denote the data obtained in the experiment. Consider a random realization of the experiment and let $(u, v) \in \mathscr{B}$ be its stopping point. By Lemma 6.1, if the first $x + y$ observations produce fewer than x successes then $u \leqslant x$, $v \geqslant y$ and $(u, v) \neq (x, y)$, and so $T(u, v) < T(x, y)$ by the assumption about T. Thus $T(u, v) \geqslant T(x, y)$ implies that the first $x + y$ observations produce at least x successes. Hence the p-value is

$$
\begin{aligned}
p &= \sup\{P_\theta[T(X, Y) \geqslant T(x, y)]: \theta \leqslant \theta_0\} \\
&\leqslant \sup\{P_\theta[\text{at least } x \text{ successes in the first } x + y \text{ trials}]: \theta \leqslant \theta_0\} \\
&= \sup\{P[\text{binomial}(x + y, \theta) \geqslant x]: \theta \leqslant \theta_0\} \\
&= P[\text{binomial}(x + y, \theta_0) \geqslant x] = \overline{P}(\mathrm{H}_0 | x, y). \qquad \square
\end{aligned}
$$

6.3. Interval estimation

A similar result holds for one-sided credible intervals. At credibility level $1 - \alpha$, the IBM produces the one-sided credible interval for θ: $C_1(x, y) = (0, \theta^*(x, y))$, where $\theta^*(x, y) = F_a^{-1}(1 - \alpha)$, F_a is the beta $(x + 1, y)$ cumulative distribution function, and $\theta^*(x, 0) = 1$. This interval $C_1(x, y)$ agrees with the frequentist one-sided confidence interval at confidence level $1 - \alpha$, based on the assumption of binomial sampling. The following theorem shows that the one-sided credible interval is a valid confidence interval under any monotone stopping rule. The proof is similar to the proof of Theorem 6.1.

Theorem 6.2. *For every monotone stopping rule, the one-sided credible interval estimator $C_1(X,Y)$ is a confidence interval estimator for θ at confidence level $1-\alpha$, i.e., $P_\theta[\theta \in C_1(X,Y)] \geqslant 1-\alpha$ for all $\theta \in \Theta$.*

At level $1-\alpha$, the equitailed two-sided credible interval for θ produced by the IBM is $C_2(x,y) = (\theta_*(x,y), \theta^*(x,y))$ with $\theta^*(x,y) = F_a^{-1}(1-\alpha/2)$ and $\theta_*(x,y) = F_b^{-1}(\alpha/2)$, where F_a and F_b are the beta $(x+1, y)$ and beta $(x, y+1)$ cumulative distribution functions, $\theta^*(x,0) = 1$ and $\theta_*(0,y) = 0$. It follows from Theorem 6.2 that, for any monotone stopping rule, $C_2(X,Y)$ is a confidence interval estimator for θ at confidence level $1-\alpha$. In fact, the credible intervals $C_2(x,y)$ agree with the widely used confidence intervals of Clopper and Pearson (1934) for a binomial parameter. The binomial stopping rule is *least favorable* amongst the monotone stopping rules in the sense that it minimizes $P_\theta[\theta \in C_2(X,Y)]$.

6.4. Examples

To see that the IBM produces more powerful inferences than the contamination and normalized-likelihood models, consider the Bernoulli Example 4.1. The IBM produces an equitailed two-sided 95% credible interval for θ which agrees with the Clopper–Pearson interval $(0.0042, 0.641)$. It is somewhat shorter than the 95% likelihood confidence interval $(0.0012, 0.722)$.

In the problem of testing H_0: $\theta \geqslant \frac{1}{2}$ versus H_1: $\theta < \frac{1}{2}$, the IBM gives $\overline{P}(H_0|x,y) = 7/64 = 0.109$, equal to the p-value assuming binomial sampling, which is much smaller than the values 0.458 and 0.656 from the contamination and normalized-likelihood models. A clearer illustration of the improvement in power is obtained in the case where we observe one success in 10 trials: then the IBM gives $\overline{P}(H_0|x,y) = 0.011$, whereas the values given by the two earlier models are 0.112 and 0.107.

The good frequentist properties of the IBM extend to contingency tables and predictive inferences. Suppose that data are obtained from two populations in the form of a 2×2 contingency table. Let θ_1 and θ_2 be the chances of success in each population, and apply the IBM to θ_1 and θ_2 independently. Then the posterior upper probability of H_0: $\theta_1 \geqslant \theta_2$ agrees with the one-sided p-value from Fisher's exact test (Walley, 1996). In Example 4.2, for instance, $\overline{P}(H_0|x) = 1/12 = 0.083$. This is much smaller than the values 0.466, obtained from the contamination model in Example 4.2, and 0.768, obtained from the normalized-likelihood model in Example 5.3.

The one-sided and equitailed two-sided prediction sets produced by the IBM agree with the frequentist prediction sets constructed from Fisher's exact test (Thatcher, 1964). The same model applies to sampling without replacement from a finite population, where again the inferences from the IBM have good frequentist properties (Walley and Bernard, 1999, Sections 5.5, 6.3). For other applications of the IBM and its generalization to multinomial data, see Walley et al. (1996), Walley and Bernard (1999), and the paper by Bernard (2002) in this volume.

7. Conclusions

The main point of this paper is that basic frequentist principles, as formulated in Section 2, are compatible with the likelihood principle and principles of coherence. There are general methods of statistical inference which satisfy all these principles. One is the contamination method defined in Section 3. The normalized-likelihood method in Section 5 also satisfies the frequentist and likelihood principles, although it is not fully coherent when degrees of consistency are interpreted as upper probabilities. (For that reason, the contamination method seems preferable.) These two methods can be used to test hypotheses and form set estimates for unknown parameters, and to make predictive inferences about future observations, in virtually any problem of statistical inference where an observed likelihood function can be defined. For extensions of these results to predictive inference, see Walley (1998).

Inferences from the two methods are highly robust. For example, the contamination method produces 95% credible sets which are valid (conservative) frequentist 95% confidence sets under all the sampling models which could have generated the observed likelihood function, and which are also valid (conservative) Bayesian 95% credible sets with respect to all the prior probability distributions in a very large set (an ε-contamination neighborhood, where $\varepsilon > \frac{1}{2}$). That is, the frequentist and Bayesian coverage probabilities are valid across a wide range of stopping rules and prior distributions. The methods do require precise knowledge of the observed likelihood function L_x, but they can be generalized to achieve robustness with respect to the likelihood function by replacing L_x by upper and lower likelihood functions, as in Walley (1991).

It may be objected that, despite these attractive properties, the two methods are too cautious to be useful in most practical problems of statistical inference. But their cautiousness can be beneficial, since it limits inferences to those that are agreed amongst frequentist, likelihood, Bayesian and imprecise probability analyses, and which are therefore highly robust and uncontroversial. The methods may also be useful in some problems where the observed likelihood function is known but the full sampling model is not, because inferences depend only on the likelihood function and satisfy the frequentist properties irrespective of the sampling model. For example, if observations are obtained sequentially but the stopping rule that was used to terminate the experiment is complicated or unknown then it may be impossible to calculate an exact p-value or confidence set. In such cases the methods studied here can be used to make valid frequentist inferences.

The imprecise beta model (IBM) is based on a much smaller set of prior distributions than the contamination model, and it therefore produces more powerful inferences. These inferences do not satisfy the general versions of the frequentist principles given in Section 2, but they do satisfy restricted versions of the hypothesis testing and weak set estimation principles, restricted to one-sided tests and one-sided or equitailed two-sided intervals.

How can the imprecise beta model be extended to other types of likelihood function? To suggest one possible extension, first note that the basic frequentist properties of the IBM are retained if the set of prior distributions (6.1) is replaced by a larger set. To define such a set, first transform the Bernoulli parameter θ to the log odds-ratio $\psi = \log(\theta/[1-\theta])$, and then consider the set of all prior probability density functions (f) for ψ that are everywhere continuous and positive and whose derivative f' satisfies $|f'(\psi)| \leqslant c f(\psi)$ for almost all ψ. Such a set of prior distributions is called a *bounded derivative model* (Walley, 1997). Provided that $c \geqslant 1$, it can be verified that the corresponding set of prior densities for θ contains the IBM set (6.1) and it therefore produces inferences which satisfy the frequentist properties in Theorems 6.1 and 6.2.

The bounded derivative model can be applied in other one-parameter problems by choosing a positive number c and a suitable transformation, ψ, of the parameter. Again degrees of consistency are identified with posterior upper probabilities, and c is chosen to be sufficiently large to guarantee one-sided frequentist properties (i.e., one-sided versions of the hypothesis testing and weak set estimation principles). For example, the bounded derivative model satisfies the one-sided frequentist properties, for the given values of ψ and c, when the data are a random sample of fixed size from the following distributions: Bernoulli, binomial, geometric or negative binomial with success probability θ, where $\psi = \log(\theta/[1-\theta])$, $c \geqslant 1$; Poisson with mean θ, $\psi = \log(\theta)$, $c \geqslant 1$; normal with mean θ and known standard deviation (and other location parameters), $\psi = \theta$, $c > 0$; normal with standard deviation θ and known mean (and other scale parameters), $\psi = \log(\theta)$, $c > 0$.

The contamination, imprecise beta and bounded derivative models illustrate a general approach to reconciling frequentist and likelihood principles that deserves careful investigation. The general approach is to identify degrees of consistency with the posterior upper probabilities that are produced by a set of prior probability distributions, and to find a set of prior distributions that is sufficiently large to satisfy at least one-sided versions of the frequentist principles, yet sufficiently small to have reasonably good power properties. Inferences from such a model are coherent and they satisfy the likelihood principle. By using imprecise probability models in this way, it appears that good frequentist properties can be reconciled with the likelihood principle.

Acknowledgements

I want to thank Departamento de Estadística y Matematica Aplicada, Universidad de Almería, Departamento de Ciencias de la Computación e Inteligencia Artificial, Universidad de Granada, and Dipartimento di Matematica, Università di Perugia, for supporting this research. I am especially grateful to Jean-Marc Bernard for detailed and insightful comments on an earlier version and for sending me relevant papers, to Serafín Moral for stimulating discussions on the topic, and to two anonymous referees for some useful suggestions.

References

Barnard, G.A., 1967. The use of the likelihood function in statistical practice. Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability 1. University of California Press, Berkeley, pp. 27–40.

Barnett, V., 1982. Comparative Statistical Inference, 2nd Edition. Wiley, London.

Bartholomew, D.J., 1971. A comparison of frequentist and Bayesian approaches to inference with prior knowledge (with discussion). In: Godambe, V.P., Sprott, D.A. (Eds.), Foundations of Statistical Inference. Holt, Rinehart and Winston, Toronto, pp. 417–434.

Basu, D., 1975. Statistical information and likelihood (with discussion). Sankhya A 37, 1–71.

Begg, C.B., 1990. On inferences from Wei's biased coin design for clinical trials (with discussion). Biometrika 77, 467–484.

Berger, J.O., 1994. An overview of robust Bayesian analysis (with discussion). Test 3, 5–124.

Berger, J.O., Berliner, L.M., 1986. Robust Bayes and empirical Bayes analysis with ε-contaminated priors. Ann. Statist. 14, 461–486.

Berger, J.O., Wolpert, R.L., 1984. The Likelihood Principle. IMS Lecture Notes, Vol. 6. Institute of Mathematical Statistics, Hayward, CA.

Bernard, J.-M., 1996. Bayesian interpretation of frequentist procedures for a Bernoulli process. Amer. Statist. 50, 7–13.

Bernard, J.-M., 2002. Implicative analysis for multivariate binary data using an imprecise Dirichlet model. J. Statist. Plann. Inference 105, 83–103.

Birnbaum, A., 1962. On the foundations of statistical inference (with discussion). J. Amer. Statist. Assoc. 57, 269–326.

Birnbaum, A., 1969. Concepts of statistical evidence. In: Morgenbesser, S., Suppes, P., White, M. (Eds.), Philosophy, Science, and Method. St. Martin's Press, New York, pp. 112–143.

Birnbaum, A., 1977. The Neyman–Pearson theory as decision theory and as inference theory: with a criticism of the Lindley–Savage argument for Bayesian theory. Synthese 36, 19–49.

Box, G.E.P., Tiao, G.C., 1973. Bayesian Inference in Statistical Analysis. Addison-Wesley, Reading, MA.

Buehler, R.J., 1959. Some validity criteria for statistical inferences. Ann. Math. Statist. 30, 845–863.

Chang, T., Villegas, C., 1986. On a theorem of Stein relating frequentist and classical inferences in group models. Canadian J. Statist. 14, 289–296.

Clopper, C.J., Pearson, E.S., 1934. The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika 26, 404–413.

Cox, D.R., 1958. Some problems connected with statistical inference. Ann. Math. Statist. 29, 357–372.

Cox, D.R., 1988. Some aspects of conditional and asymptotic inference: a review. Sankhya A 50, 314–337.

Cox, D.R., Hinkley, D.V., 1974. Theoretical Statistics. Chapman & Hall, London.

Datta, G.S., Ghosh, J.K., 1995. On priors providing frequentist validity for Bayesian inference. Biometrika 82, 37–45.

Dawid, A.P., 1991. Fisherian inference in likelihood and prequential frames of reference (with discussion). J. Roy. Statist. Soc. B 53, 79–109.

Edwards, A.W.F., 1972. Likelihood. Cambridge University Press, Cambridge.

Efron, B., 1993. Bayes and likelihood calculations from confidence intervals. Biometrika 80, 3–26.

de Finetti, B., 1974. Theory of Probability, Vol. 1. Wiley, London.

Fisher, R.A., 1934. Two new properties of mathematical likelihood. Proc. Roy. Soc. A 144, 285–307.

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

Fraser, D.A.S., 1991. Statistical inference: likelihood to significance. J. Amer. Statist. Assoc. 86, 258–265.

Huber, P.J., 1973. The use of Choquet capacities in statistics. Bull. Internat. Statist. Inst. 45 (Book 4), 181–188.

Huber, P.J., 1981. Robust Statistics. Wiley, New York.

Kyburg, H.E., 1974. The Logical Foundations of Statistical Inference. Reidel, Dordrecht.

Lehmann, E.L., 1986. Testing Statistical Hypotheses, 2nd Edition. Wiley, New York.

Lindley, D.V., 1958. Fiducial distributions and Bayes' theorem. J. Roy. Statist. Soc. B 20, 102–107.

Neyman, J., 1977. Frequentist probability and frequentist statistics. Synthese 36, 97–131.

Nicolaou, A., 1993. Bayesian intervals with good frequentist behaviour in the presence of nuisance parameters. J. Roy. Statist. Soc. B 55, 377–390.

Pericchi, L.R., Walley, P., 1991. Robust Bayesian credible intervals and prior ignorance. Internat. Statist. Rev. 58, 1–23.

Pratt, J.W., 1965. Bayesian interpretation of standard inference statements (with discussion). J. Roy. Statist. Soc. B 27, 169–203.

Robinson, G.K., 1979. Conditional properties of statistical procedures. Ann. Statist. 7, 742–755.

Seidenfeld, T., 1979. Philosophical Problems of Statistical Inference. Reidel, Dordrecht.

Severini, T.A., 1991. On the relationship between Bayesian and non-Bayesian interval estimates. J. Roy. Statist. Soc. B 53, 611–618.

Shafer, G., 1976. A Mathematical Theory of Evidence. Princeton University Press, Princeton, NJ.

Thatcher, A.R., 1964. Relationships between Bayesian and confidence limits for predictions (with discussion). J. Roy. Statist. Soc. B 26, 176–210.

Walley, P., 1991. Statistical Reasoning with Imprecise Probabilities. Chapman & Hall, London.

Walley, P., 1996. Inferences from multinomial data: Learning about a bag of marbles (with discussion). J. Roy. Statist. Soc. B 58, 3–57.

Walley, P., 1997. A bounded derivative model for prior ignorance about a real-valued parameter. Scand. J. Statist. 24, 463–483.

Walley, P., 1998. Reconciling frequentist properties with the likelihood principle. Technical Report DECSAI-98-02-25. Department of Computer Science and Artificial Intelligence, University of Granada, Spain.

Walley, P., 1999. Imprecise probabilities. In: Kotz, S., Read, C.B., Banks, D.L. (Eds.), Encyclopedia of Statistical Sciences, Update Vol. 3. Wiley, New York, pp. 355–359.

Walley, P., Bernard, J.-M., 1999. Imprecise probabilistic prediction for categorical data. Technical Report CAF-9901, Laboratoire Cognition et Activités Finalisées, CNRS ESA 7021, Université de Paris 8, Saint-Denis, France.

Walley, P., Gurrin, L., Burton, P., 1996. Analysis of clinical data using imprecise prior probabilities. Statistician 45, 457–485.

Walley, P., Moral, S., 1999. Upper probabilities based only on the likelihood function. J. Roy. Statist. Soc. B 61, 831–847.

Wasserman, L., 1997. Bayesian robustness. In: Kotz, S., Read, C.B., Banks, D.L. (Eds.), Encyclopedia of Statistical Sciences, Update Vol. 1. Wiley, New York, pp. 45–51.

Welch, B.L., Peers, H.W., 1963. On formulae for confidence points based on integrals of weighted likelihoods. J. Roy. Statist. Soc. B 25, 318–329.