Confidence and Likelihood

Tore Schweder
Department of Economics, University of Oslo
Nils Lid Hjort
Department of Mathematics, University of Oslo

January 2001

Abstract

Independent data are efficiently integrated by adding their respective log-likelihoods. Instead of Bayesian updating of information, we propose to use the likelihood directly as a vehicle for coherent learning. Data concerning a one-dimensional interest parameter might be summarised in a likelihood function reduced of nuisance parameters. This reduced likelihood is combined with the likelihood of future data to update information. In the frequentist tradition, statistical reporting is often done in the format of confidence intervals. The confidence distribution, with quantiles specifying all possible confidence intervals provides a more complete report than a 95% interval, say, or the p-value of a test. The concept of confidence distribution is discussed, and a new version of the Neyman–Pearson lemma is provided.

Confidence distributions based on prior data represent frequentist analogues to Bayesian priors. These confidence distributions need to be converted to likelihoods before they can be integrated with the new data likelihood. It is the statistical model, usually through a pivot, that dictates both the confidence distribution and the reduced likelihoods. There is not a one-to-one correspondence between the two. Confidence distributions resulting from the integrated analysis, along with their probability bases, represent the frequentist analogue to the Bayesian posterior distributions. Asymptotics or bootstrapping is used to find pivots and their distributions, and hence reduced likelihoods and confidence distributions. A simple form of inverting bootstrap distributions to approximate pivots of the abc type is proposed. The issue of non-informative Bayesian priors is also visited.

The material is illustrated in a number of examples and in an application to multiple capture data for bowhead whales. Here it is argued that the confidence distribution depends on the study protocol, even for identical data from the same statistical model.

KEY WORDS: *abc correction, bootstrapping likelihoods, capture-recapture data, confidence distributions and densities, frequentist posteriors and priors, integrating information, Neyman–Pearson lemma, pivots, reduced likelihood, study protocols*

1 Introduction

Confidence intervals and p-values are the primary formats of statistical reporting in the frequentist tradition. The close relationship between p-values and confidence intervals allows a unification of these concepts in the *confidence distribution*. Let the one-dimensional parameter of interest be ψ. A confidence distribution for ψ is calculated from the data within the statistical model. The cumulative confidence distribution function, C, provides $C(\psi_0)$ as the p-value when testing the one-sided hypothesis $H_0\colon \psi \leq \psi_0$ whatever value ψ_0 takes. Any pair of confidence quantiles constitutes, on the other hand, a confidence interval $(C^{-1}(\alpha), C^{-1}(\beta))$ with degree of confidence $\beta - \alpha$.

The likelihood function is a minimal sufficient statistic. Since it generally is difficult to interpret, the information contained in the likelihood function concerning a parameter of interest needs

to be extracted in an intelligible format. Distributions are the eminent format of presenting uncertain information. Much of the attraction of the Bayesian approach is due to the use of distributions as the format of presenting information, e.g. prior and posterior distributions. Fisher (1930) introduced fiducial probability distributions as an alternative to the Bayesian posterior distribution as a format of presenting what has been learned from the data in view of the model; see Fisher (1973) for his final understanding of fiducial probability and the fiducial argument. Quantiles of a fiducial distribution are endpoints of fiducial intervals. Following Neyman rather than Fisher in understanding fiducial intervals as confidence intervals, we adopt the term *confidence distribution* from Efron (1998) and others.

The likelihood function is the pre-eminent tool for integrating diverse data. Bayesians and frequentists all agree on this issue. Old and new data are also best integrated via the likelihood function. As an alternative to Bayesian updating of information regarding a parameter of interest, the likelihood of the old and the new data are thus simply multiplied together. In the presence of nuisance parameters, statistical reporting of the information regarding the interest parameter ψ might be done both in the format of a confidence distribution and in the format of a reduced likelihood function. The confidence provides the interpretation and the reduced likelihood allows the essential information in the present data regarding ψ to be integrated with new data at a later stage. Such updating of information might be termed *likelihood updating.*

As distinct from the Bayesian view, we will distinguish between probability as frequency, termed *probability,* and probability as information/uncertainty, termed *confidence.* A prior distribution in our frequentist world is then to be understood as a confidence distribution. To achieve likelihood updating, the likelihood representing the prior confidence distribution needs to be identified. This likelihood is an ordinary likelihood of the past data underlying the prior confidence distribution, but reduced to that statistic. There might have been other parameters involved in the model when analyzing those past data, and the full likelihood was then a function of all the parameters. As Fisher (1922) used a two-stage procedure to obtain the likelihood of σ from $\mathrm{N}(\mu, \sigma^2)$ data by reduction to the empirical standard deviation, we use the term *reduced likelihood* for the likelihood of data suitably reduced to a statistic informative of the interest parameter only. Exact reduced likelihoods are only available in nice models. Our proposal is thus to use an approximate reduced likelihood when updating the information expressed by the (approximate) prior confidence distribution with the likelihood of the new data. As confidence distributions are found from (approximate) *pivots,* so are reduced likelihoods. When the pivot is additive and normally distributed, as often is the case with large data, the reduced log-likelihood is proportional to the squared normal score of the confidence distribution. This likelihood, called the *normal-based reduced likelihood,* agrees with the so-called implied likelihood of Efron (1993).

Only when the pivot is additive in the statistic is the reduced likelihood proportional to the confidence density. In the general case, this is not the case. A given confidence distribution can arise from a multitude of pivots. By an example, we show that a given confidence distribution can be related to different reduced likelihoods, depending on the pivot it arises from. Sections 2–4 are devoted to developing this basic material, of updating information by likelihoods, representing information in the format of confidence distributions, and of using appropriate (approximate) pivots to identify both the reduced likelihood and the confidence distribution.

Efron (1998) expects a revival of the Fisherian paradigm of statistical inference based on the concepts of likelihood, fiducial distribution (confidence distribution) and the many other useful ideas introduced by Fisher. With bootstrapping and techniques for constructing confidence intervals and thus confidence distributions, Efron has added an important tool for statistical inference of the frequentist tradition of Fisher and Neyman. In our context of parametric models and inference,

parametric bootstrapping is often the natural technique to use when the information contained in a likelihood function is to be converted to a confidence distribution for an interest parameter. To allow parametric bootstrapping of the likelihood function obtained by likelihood updating, it is necessary to know how to bootstrap the 'prior' likelihood summarising the old data. Together with the reduced likelihood to be used in later likelihood updating, sufficient information must be given to allow correct bootstrapping. As confidence distributions and reduced likelihoods are found from pivotal constructs and their distributions, the pivot provides the key to parametric bootstrapping of these statistics. This is dealt with in Section 5.

In Section 6 a version of the Neyman–Pearson lemma is provided, explaining the frequentist optimality of the confidence distribution in one-parameter models with monotone likelihood ratio. This also leads to optimal constructions of confidence distributions in higher-dimensional parametric families of the exponential kind, via conditioning on ancillary statistics. These confidence distributions become uniformly most reliable in a sense made precise in Section 6. Other notions of optimality are briefly discussed in Section 7, including the use of equivariance.

It is desirable to develop methods for obtaining approximate confidence distributions in situations where exact constructions either become too intricate or do not exist. In Section 8 we discuss various approximations, the simplest of which being based on the traditional delta method for asymptotic normality. Better versions emerge via corrections of various sorts. In particular we develop an acceleration and bias corrected bootstrap percentile interval method for constructing improved confidence distributions. It has an appealing form and is seen to perform well in terms of accuracy. It also leads to good approximations for reduced likelihoods.

In Section 9 our apparatus is tested on a real data problem involving capture-recapture photo-identification data for bowhead whales. Finally, supplementing remarks and discussion are found in Section 10. Among the points argued there is the suggestion that the confidence density is a very useful summary for any parameter of interest and may serve as the frequentist analogue of the Bayesian's posterior density. We also discuss our work in the context of what Hald (1998) terms the three (so far) revolutions in parametric statistical inference.

2 Confidence distributions

Before relating confidence distributions to likelihoods, it is worthwhile having a closer look at the concept as a format of reporting statistical inference.

2.1 Confidence and statistical inference

Our context is a parametric model with an interest parameter ψ for which inference is sought. The interest parameter is assumed to be scalar, and to belong to a finite or infinite interval on the real line. The space of the parameter is thus linearly ordered. With inference we shall understand statements of the type '$\psi > \psi_0$', '$\psi_1 \leq \psi \leq \psi_2$', etc., where ψ_0, ψ_1 etc. are values usually computed from the data. To each statement, we would like to associate how much confidence the data allow us to have in the statement.

As the name indicates, the confidence distribution is related to confidence intervals, which are interval statements with the confidence fixed *ex ante*, and with endpoints calculated from the data. A one-sided confidence interval with (degree of) confidence $1-\alpha$ has right endpoint the corresponding quantile of the confidence distribution. If C is the cumulative confidence distribution calculated from the data, the left-sided confidence interval is $(-\infty, C^{-1}(1-\alpha))$. A right-sided confidence interval $(C^{-1}(\alpha), \infty)$ has confidence $1-\alpha$, and a two-sided confidence interval $[C^{-1}(\alpha), C^{-1}(\beta)]$ has

confidence $\beta-\alpha$. Two-sided confidence intervals are usually equi-tailed in the sense that $\alpha = 1-\beta$.

Definition 1 *A (one-dimensional) confidence distribution for ψ with cumulative distribution function (cdf) C is a statistic such that $C(\psi)$ has a uniform distribution over $(0,1)$ under the probability distribution $P_{\psi,\chi}$, where χ is the remaining (nuisance) parameter.*

By this definition, the (stochastic) confidence quantiles are endpoints of confidence intervals with degree of confidence given by the stipulated confidence. For one-sided intervals $(-\infty, \psi_\alpha)$, where $\psi_\alpha = C^{-1}(\alpha)$, the coverage probability is, in fact, $P_{\psi,\chi}\{\psi \leq \psi_\alpha\} = P_{\psi,\chi}\{C(\psi) \leq C(\psi_\alpha)\} = P_{\psi,\chi}\{C(\psi) \leq \alpha\} = \alpha$.

Being an invertible function of the interest parameter, and having a uniform distribution independent of the full parameter, $C(\psi)$ is a pivot (Barndorff-Nielsen and Cox 1994). On the other hand, whenever a pivot $\mathrm{piv}(Y,\psi)$ is available, taken to be increasing in ψ, and having cumulative distribution function F independent of the parameter,

$$C(\psi) = F(\mathrm{piv}(Y,\psi)) \tag{1}$$

is uniformly distributed and is thus the cdf of a confidence distribution for ψ. If the natural pivot is decreasing in ψ, then $C(\psi) = 1 - F(\mathrm{piv}(Y,\psi))$.

Exact confidence distributions represents *valid* inference in the sense of statistical conclusion validity (Cook and Campbell, 1979). The essence is that the confidence distribution is free of bias in that any confidence interval $(\psi_\alpha, \psi_\beta)$ has exact coverage probability $\beta - \alpha$. The *reliability* of the inference represented by C is basically a question of the spread of the confidence distribution. We return to the issue of reliability, and optimal reliability, in Section 6.

Hypothesis testing and confidence intervals are closely related. Omitting the instructive proof, this relation is stated in the following lemma.

Lemma 2 *The confidence of the statement '$\psi \leq \psi_0$' is the cumulative confidence distribution function value $C(\psi_0)$, and is equal to the p-value of a test of $H_0\colon \psi \leq \psi_0$ versus the alternative $H_1\colon \psi > \psi_0$.*

The opposite statement '$\psi > \psi_0$' has confidence $1 - C(\psi_0)$. Usually, the confidence distributions are continuous, and '$\psi \geq \psi_0$' has the same confidence as '$\psi > \psi_0$'.

Some care is needed when calculating and interpreting the confidence for statements determined *ex ante*. When ψ_0 is fixed, the statement '$\psi \neq \psi_0$' should, preferably, have confidence given by one minus the p-value when testing $H_0\colon \psi = \psi_0$. This can be calculated from the observed confidence distribution, and is $1 - 2\min\{C(\psi_0), 1 - C(\psi_0)\}$. It is, however, questionable whether $\psi_0 = C^{-1}(\frac{1}{2}(1+c))$ or $\psi_0 = C^{-1}(\frac{1}{2}(1-c))$, where c is chosen *ex ante*, makes the statement '$\psi \neq \psi_0$' have confidence c.

Confidence intervals are invariant w.r.t. monotone transformations. This is also the case for confidence distributions.

Lemma 3 *Confidence distributions based essentially on the same statistic are invariant with respect to monotone continuous transformations of the parameter: If $\rho = r(\psi)$, say, with r increasing, and if C^ψ is based on T while C^ρ is based on $S = s(T)$ where s is monotone, then*

$$C^\rho(\rho) = C^\psi(r^{-1}(\rho)).$$

To a large extent statistical inference is being carried out as follows. From optimality or structural considerations, an estimator of the parameter of interest, and possibly of the remaining

(nuisance) parameters in the model, is determined. Then, the sampling distribution of the estimator is calculated, possibly by bootstrapping. Finally, statements of inference, e.g. confidence intervals, are extracted from the sampling distribution and its dependence on the parameter.

A sharp distinction should be drawn between the (estimated) sampling distribution and the confidence distribution. The sampling distribution of the estimator is the *ex ante* probability distribution of the statistic under repeated sampling, while the confidence distribution is calculated *ex post* and distributes the confidence the observed data allow to be associated with different statements concerning the parameter. Consider the estimated sampling distribution of the point estimator $\widehat{\psi}$, say as obtained from the parametric bootstrap. If ψ^* is a random estimate of ψ obtained by the same method, the estimated sampling distribution is the familiar

$$S(\psi) = \Pr\{\psi^* \leq \psi; \widehat{\psi}\} = F_{\widehat{\psi}}(\psi).$$

The confidence distribution is also obtained by (theoretically) drawing repeated samples, but now from different distributions. The interest parameter is, for the confidence distribution, considered a control variable, and it is varied in a systematic way. When $\widehat{\psi}$ is a reasonable statistic and the hypothesis H_0: $\psi \leq \psi_0$ is suspect when $\widehat{\psi}$ is large, the p-value is $\Pr\{\psi^* > \widehat{\psi}; \psi_0\}$. The cumulative confidence distribution is then

$$C(\psi) = \Pr\{\psi^* > \widehat{\psi}; \psi\} = 1 - F_{\psi}(\widehat{\psi}). \tag{2}$$

The sampling distribution and the confidence distribution are fundamentally different entities. The sampling distribution is a probability distribution, while the confidence distribution, *ex post*, is not a distribution of probabilities but of confidence – obtained from the probability transform of the statistic used in the analysis.

The confidence densities we deduce or approximate in the following would presumably be equivalent to the infamous fiducial distributions in the sense of Fisher, at least in cases where Fisher would have considered the mechanism behind the confidence limits to be inferentially correct; see the discussion in Efron (1998, Section 8). In view of old and on-going controversies and confusion surrounding this theme of Fisher, and the fact that such fiducial distributions sometimes have been put forward in ad hoc fashions and with vague interpretation, we emphasise that our distributions of confidence are actually derived from certain principles in a rigorous framework, and with a clear interpretation. Our work can perhaps be seen as being in the spirit of Neyman (1941). We share the view expressed in Lehmann (1993) that the distinction between the Fisherian and the Neyman–Pearson tradition is unfortunate. The unity of the two traditions is illustrated by our version of the Neyman–Pearson lemma as it applies to Fisher's fiducial distribution (confidence distribution). Note also that we in Section 3, in particular, work towards establishing confidence distributions that are inferentially correct.

Example 1. Consider the exponentially distributed variate T with probability density $f(t;\psi) = (1/\psi)\exp(-t/\psi)$. The cumulative confidence distribution function for ψ is $C(\psi; t_{\rm obs}) = \exp(-t_{\rm obs}/\psi)$. The confidence density is thus $c(\psi; t_{\rm obs}) = (\partial/\partial\psi)C(\psi; t_{\rm obs}) = t_{\rm obs}\psi^{-2}\exp(-t_{\rm obs}/\psi)$, which not only has a completely different interpretation from the sampling density of the maximum likelihood estimator, T, but also has a different shape. ■

Example 2. Suppose the ratio $\psi = \sigma_2/\sigma_1$ between standard deviation parameters from two different data sets are of interest, where independent estimates of the familiar form $\widehat{\sigma}_j^2 = \sigma_j^2 W_j/\nu_j$ are available, where W_j is a $\chi^2_{\nu_j}$. The canonical intervals, from inverting the optimal tests for single-point hypotheses $\psi = \psi_0$, take the form

$$[\widehat{\psi}/K^{-1}(1-\alpha)^{1/2},\ \widehat{\psi}/K^{-1}(\alpha)^{1/2}],$$

where $\widehat{\psi} = \widehat{\sigma}_2/\widehat{\sigma}_1$ and $K = K_{\nu_2,\nu_1}$ is the distribution function for the F statistic $(W_2/\nu_2)/(W_1/\nu_1)$. Thus $C^{-1}(\alpha) = \widehat{\psi}/K^{-1}(1-\alpha)^{1/2}$. This corresponds to the confidence distribution function $C(\psi;\text{data}) = 1 - K(\widehat{\psi}^2/\psi^2)$, with confidence density

$$c(\psi;\text{data}) = k(\widehat{\psi}^2/\psi^2)2\widehat{\psi}^2/\psi^3,$$

expressed in terms of the F density $k = k_{\nu_2,\nu_1}$. See also Section 7.1 for an optimality result of the confidence density used here, and Section 8.3 for a very good approximation based on bootstrapping. ■

2.2 Linear regression

In the linear normal model, the n-dimensional data Y of the response is assumed $\mathrm{N}(X\beta, \sigma^2 I)$. With SSR being the residual sum of squares and with $p = \text{rank}(X)$, $S^2 = \text{SSR}/(n-p)$ is the traditional estimate of the residual variance. With S_j^2 being the mean-unbiased estimator of the variance of the regression coefficient estimator $\widehat{\beta}_j$,

$$V_j = (\widehat{\beta}_j - \beta_j)/S_j$$

is a pivot with a t-distribution of $\nu = n - p$ degrees of freedom. Letting $t_\nu(\alpha)$ be the quantiles of this t-distribution, the confidence quantiles for β_j are the familiar $\widehat{\beta}_j + t_\nu(\alpha)S_j$. The cumulative confidence distribution function for β_j is seen from this to become

$$C(\beta_j;\text{data}) = 1 - G_\nu((\widehat{\beta}_j - \beta_j)/S_j) = G_\nu((\beta_j - \widehat{\beta}_j)/S_j),$$

where G_ν is the cumulative t-distribution with ν degrees of freedom. Note also that the confidence density $c(\beta_j;\text{data})$ is the t_ν-density centred at $\widehat{\beta}_j$ and with the appropriate scale.

Now turn attention to the case where σ, the residual standard deviation, is the parameter of interest. Then the pivot $\text{SSR}/\sigma^2 = \nu S^2/\sigma^2$ is a χ^2_ν, and the cumulative confidence distribution is found to be

$$C(\sigma;\text{data}) = \Pr\{\chi^2_\nu > \text{SSR}/\sigma^2\} = 1 - \Gamma_\nu(\nu S^2/\sigma^2),$$

where Γ_ν is the cumulative distribution function of the chi-square with density γ_ν. The confidence density becomes

$$c(\sigma;\text{data}) = \gamma_\nu\Big(\frac{\nu S^2}{\sigma^2}\Big)\frac{2\nu S^2}{\sigma^3} = \frac{S^\nu}{2^{\nu/2}\Gamma(\frac{1}{2}\nu)}\sigma^{-(\nu+1)}\exp(-\tfrac{1}{2}\nu S^2/\sigma^2),$$

which again is different from the likelihood. The likelihood, for the SSR part of the data, is the density of $\text{SSR} = \sigma^2\chi^2_\nu$, which is proportional to

$$L(\sigma) = \sigma^{-\nu}\exp(-\tfrac{1}{2}\nu S^2/\sigma^2).$$

This is the two-stage likelihood for σ, in the spirit of Fisher (1922), and we term it the reduced likelihood for σ. Taking logarithms, the pivot is brought on an additive scale, $\log S - \log\sigma$, and in the parameter $\tau = \log\sigma$ the confidence density is proportional to the likelihood. The log-likelihood also has a nicer shape in τ than in σ, where it is less neatly peaked.

It is of interest to note that the improper prior $\pi(\sigma) = \sigma^{-1}$, regarded as the canonical 'non-informative' prior for scale parameters like the present σ, yields when combined with the likelihood L the confidence distribution as the Bayes posterior distribution. See also the more general comment in Section 11.

2.3 Approximate confidence distributions for discrete data

To achieve exact say 95% coverage for a confidence interval based on discrete data is usually impossible without artificial randomisation. The same difficulty is encountered when constructing tests with exactly achieved significance level. Confidence distributions based on discrete data can never be exact. Since the data are discrete, any statistic based on the data must have a discrete distribution. The confidence distribution is a statistic, and $C(\psi)$ cannot have a continuous uniform distribution. Half-correction is a simple device to achieve a reasonably approximate confidence distribution. When T is the statistic on which p-values and hence the the confidence distribution is based, half-correction typically takes the form

$$C(\psi) = \Pr_\psi\{T > t_{\rm obs}\} + \tfrac{1}{2}\Pr_\psi\{T = t_{\rm obs}\}.$$

For an illustration, let T be Poisson with parameter ψ. Then the density of the half-corrected confidence distribution simplifies to

$$c(\psi) = \tfrac{1}{2}\Big\{\frac{\psi^{t_{\rm obs}-1}}{(t_{\rm obs}-1)!}e^{-\psi} + \frac{\psi^{t_{\rm obs}}}{t_{\rm obs}!}e^{-\psi}\Big\} \qquad \text{provided } t_{\rm obs} \geq 1.$$

Although the confidence distribution has a discrete probability distribution *ex ante*, it is a continuous distribution for ψ *ex post*.

A confidence distribution depends on the probability model, not only on the likelihood. The Bayesian posterior distribution, depends on the other hand only on the observed likelihood. This point is understood by frequentists. It is illustrated by the following.

Example 3. Let T_x be the waiting time until x points is observed in a Poisson process with intensity parameter ψ, and let X_t be the number of points observed in the period $(0, t)$. The two variables are respectively gamma-distributed with shape parameter x and Poisson distributed with mean ψt. In one experiment, T_x is observed to be t. In another, X_t is observed to be x. The observed log-likelihood is then identical in the two experiments, namely $\ell(\psi) = x\log(t\psi) - t\psi$. From the identity $\Pr\{T_x > t\} = \Pr\{X_t < x\}$, and since ψT_x is a pivot, the confidence distribution based on T_x has cdf $C_t(\psi) = 1 - F(x-1; \psi t)$ where F is the cdf of the Poisson distribution with mean ψt. This is not an exact confidence distribution if the experiment was to observe X_t. It is, in fact, stochastically slightly smaller than it should be in that case. In fact, in that experiment $\mathrm{E}C_t(\psi) = \frac{1}{2}\Pr\{X_t \neq Y_t\} < \frac{1}{2}$, where Y_t is an independent copy of the Poisson variate X_t. As noted above, no non-randomised exact confidence distribution exists in the latter experiment. ■

3 Likelihood related to confidence distributions

To combine past reported data with new data, and also for other purposes, it is advantageous to recover a likelihood function or an approximation thereof from the available statistics summarising the past data. The question we ask is whether an acceptable likelihood function can be recovered from a published confidence distribution, and if this is answered in the negative, how much additional information is needed to obtain a usable likelihood. An example will show that a confidence distribution is in itself not sufficient to determine the likelihood of the reduced data, T, summarised by C. A given confidence distribution could, in fact, result from many different probability models, each with a specific likelihood.

Frequentist statisticians have discussed at length how to obtain confidence distributions for one-dimensional interest parameters from the likelihood of the data in view of its probability basis. Barndorff-Nielsen and Cox (1994) discuss adjusted likelihoods and other modified likelihoods based

on saddle-point approximations. Efron and Tibshirani (1993) and Davison and Hinkley (1997) present methods based on bootstrapping and quadratic approximations. The reverse problem, finding an approximate likelihood of the reduced data represented by the confidence distribution, has received less interest. Fisher's two-stage likelihood is an exception, and we follow in his footsteps.

By definition, a likelihood is a probability density regarded as a function of the parameters, keeping the data at the observed value. A confidence distribution can not be interpreted as a probability distribution. It distributes confidence and not probability. The confidence density is therefore not usually a candidate for the likelihood function we seek. It is the probability distribution of the confidence distribution, regarded as the data, which matters. We will now demonstrate by means of a simple example that a given confidence distribution can relate to many different likelihoods, according to the underlying statistical model.

Example 4. Consider a uniform confidence distribution for ψ over $(0,1)$. It is based on the statistic T with observed value $t_{\text{obs}} = \frac{1}{2}$. We shall consider three different models leading to this confidence distribution, and we calculate the likelihood function in each case.

The first model is a shift-uniform model with pivot $\psi - T + \frac{1}{2} = U$ where U has a uniform probability distribution over $(0,1)$. Thus, $C(\psi) = \psi$ for $0 \leq \psi \leq 1$ representing the uniform confidence distribution. Further, T is uniform over $(\psi - \frac{1}{2}, \psi + \frac{1}{2})$ and the likelihood is $L_{\text{shift}}(\psi) = I_{(0,1)}(\psi)$, the indicator function.

Second, consider the scale model with pivot $\frac{1}{2}\psi/T = U$. Again, the confidence distribution is the uniform. The probability density of T is easily found, and the likelihood based on $T = \frac{1}{2}$ comes out as $L_{\text{scale}}(\psi) = 2\psi I_{(0,1)}(\psi)$.

The third model is based on a normally distributed pivot, $\Phi^{-1}(\psi) - \Phi^{-1}(T) = Z$, where Z has a standard normal distribution with cdf Φ. For the observed data, the confidence distribution is the same uniform distribution. Calculating the probability density of T, we find the likelihood of the observed data $L_{\text{norm}}(\psi) = \exp[-\frac{1}{2}(\Phi^{-1}(\psi))^2]$.

These three possible log-likelihoods consistent with the uniform confidence distribution are shown in Figure 1. Other log-likelihoods are also possible. ■

In the Poisson/gamma example we saw that different models for the same data lead to different confidence distributions, despite the fact that the resulting likelihood functions were identical. This is the reverse of the situation in Example 4, where different likelihoods were associated with the same confidence distribution. More dramatic examples of this phenomenon are possible.

Example 5. The data point is again $t_{\text{obs}} = \frac{1}{2}$, but now the likelihood function is the flat one over $(-2,2)$ obtained by a uniform shift model leading to the uniform confidence distribution over the same interval. In the alternative model, T has probability density $f(t) = 1 + \psi(t - \frac{1}{2})$ for $0 \leq t \leq 1$, and the cdf is $F(t) = t - \frac{1}{2}\psi t(1-t)$. The observed datum yields a flat likelihood over the parameter set $(-2,2)$. The cdf of the confidence distribution is found as the upper tail-probability, $C(\psi) = 1 - t + \frac{1}{2}\psi t(1-t)$, cf. (2). The observed data thus yield a confidence distribution with point mass $1/4$ at $\psi = -2$ and $\psi = 2$, and the remaining confidence uniformly distributed over the interior of the parameter set. This example does also illustrate the fact that proper confidence distributions are not always available, i.e. when the parameter set is the open interval, the confidence distribution has only total mass $\frac{1}{2}$. ■

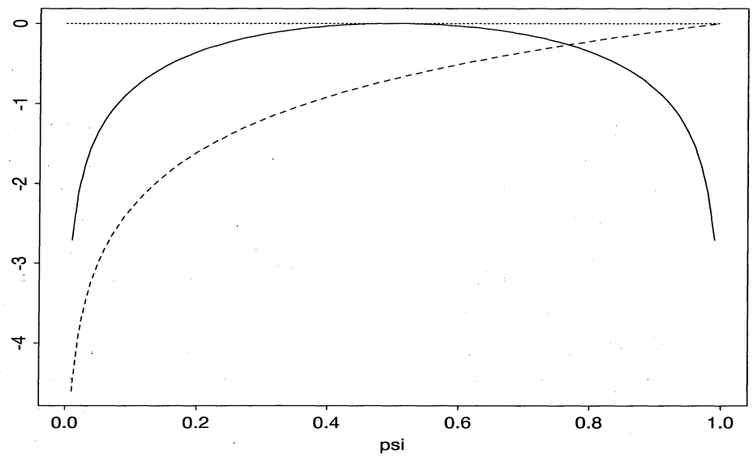

FIGURE 1: Three log-likelihoods consistent with a uniform confidence distribution over $[0, 1]$. *'Many likelihoods informed me of this before, which hung so tottering in the balance that I could neither believe nor misdoubt.'* – SHAKESPEARE.

4 Confidence and likelihoods based on pivots

Assume that the confidence distribution $C(\psi)$ is based on a pivot piv with cumulative distribution function F and density f. Since ψ is one-dimensional, the pivot is typically a function of a one-dimensional statistic T in the data X. The probability density of T is then

$$f^T(t;\psi) = f(\mathrm{piv}(t;\psi)) \Big| \frac{\partial \mathrm{piv}(t;\psi)}{\partial t} \Big|.$$

Since $\mathrm{piv}(T;\psi) = F^{-1}(C(\psi))$ we have the following.

Proposition 4 *When the statistical model leads to a pivot* $\mathrm{piv}(T;\psi)$ *in a one-dimensional statistic* T*, increasing in* ψ*, the likelihood is*

$$L(\psi;T) = f\big(F^{-1}(C(\psi))\big) \Big| \frac{\mathrm{dpiv}(T;\psi)}{\mathrm{d}T} \Big|.$$

The confidence density is also related to the distribution of the pivot. Since one has $C(\psi) = F(\mathrm{piv}(T;\psi))$,

$$c(\psi) = f(\mathrm{piv}(T;\psi)) \Big| \frac{\mathrm{dpiv}(T;\psi)}{\mathrm{d}\psi} \Big|.$$

Thus, the likelihood is in this simple case related to the confidence density by

$$L(\psi;T) = c(\psi) \Big| \frac{\partial \mathrm{piv}(T;\psi)}{\partial T} \Big| \Big/ \Big| \frac{\partial \mathrm{piv}(T;\psi)}{\partial \psi} \Big|. \tag{3}$$

There are important special cases. If the pivot is additive in T (at some measurement scale), say $\mathrm{piv}(T;\psi) = \mu(\psi) - T$ for a smooth increasing function μ, the likelihood is $L(\psi;T) = f\big(F^{-1}(C(\psi))\big)$. When furthermore the pivot distribution is normal, we will say that the confidence distribution has a *normal probability basis.*

Proposition 5 (Normal-based likelihood) *When the pivot is additive and normally distributed, the reduced log-likelihood related to the confidence distribution is* $\ell(\psi) = -\frac{1}{2}\left\{\Phi^{-1}(C(\psi))\right\}^2$.

The normal-based likelihood might often provide a good approximate likelihood. Note that classical first order asymptotics leads to normal-based likelihoods. The conventional method of constructing confidence intervals with confidence $1-\alpha$,

$$\{\psi\colon 2(\ell(\widehat{\psi}) - \ell(\psi)) < \Phi^{-1}(1-\tfrac{1}{2}\alpha)\}$$

where $\widehat{\psi}$ is the maximum likelihood estimate, is equivalent to assuming the likelihood to be normal-based. The so-called ABC confidence distributions of Efron (1993), concerned partly with exponential families, have asymptotic normal probability basis, as have confidence distributions obtained from Barndorff-Nielsen's r^* (Barndorff-Nielsen and Wood, 1998). Efron (1993) used a Bayesian argument to derive the normal-based likelihood in exponential models. He called it the implied likelihood.

In many applications, the confidence distribution is found by simulation. One might start with a statistic T which, together with an (approximate) ancillary statistic A, is simulated for a number of values of the interest parameter ψ and the nuisance parameter χ. The hope is that the conditional distribution of T given A is independent of the nuisance parameter. This question can be addressed by applying regression methods to the simulated data. The regression might have the format

$$\mu(\psi) - T = \tau(\psi)V \tag{4}$$

where V is a scaled residual. Then $\mathrm{piv}(T;\psi) = (T-\mu(\psi))/\tau(\psi)$, and the likelihood is

$$L(\psi) = f\big(F^{-1}(C(\psi))\big)\big/\tau(\psi).$$

The scaling function τ and the regression function μ might depend on the ancillary statistic.

Example 6. Let T be Poisson with mean ψ. The half-corrected cumulative confidence distribution function is

$$C(\psi) = 1 - \sum_{j=0}^{t_{\mathrm{obs}}} \exp(-\psi)\psi^j/j! + \tfrac{1}{2}\exp(-\psi)\psi^{t_{\mathrm{obs}}}/t_{\mathrm{obs}}!.$$

Here $Y = 2(\sqrt{\psi} - \sqrt{T})$ is approximately $\mathrm{N}(0,1)$ and is accordingly approximately a pivot for moderate to large ψ. From a simulation experiment, one finds that the distribution of Y is slightly skewed, and has a bit longer tails than the normal. By a little trial and error, one finds that $\exp(Y/1000)$ is closely Student distributed with $\mathrm{df} = 30$. With Q_{30} being the upper quantile function of this distribution and t_{30} the density, the log-likelihood is approximately $\ell_s(\psi) = \log t_{30}(Q_{30}(C(\psi)))$. Examples are easily made to illustrate that the $\ell_s(\psi)$ log-likelihood quite closely approximates the real Poisson log-likelihood $\ell(\psi) = t_{\mathrm{obs}} - \psi + t_{\mathrm{obs}}\log(\psi/t_{\mathrm{obs}})$. Our point here is to illustrate the approximation technique; when the exact likelihood is available we will of course that one. ■

Usually, the likelihood associated with a confidence distribution is different from the confidence density. The confidence density depends on the parametrisation. By reparametrisation, the likelihood can be brought to be proportional to the confidence density. This parametrisation might have additional advantages.

Let $L(\psi)$ be the likelihood and $c(\psi)$ the confidence density for the chosen parametrisation, both assumed positive over the support of the confidence distribution. The quotient $J(\psi) = L(\psi)/c(\psi)$

has an increasing integral $\mu(\psi)$, with $(\partial/\partial\psi)\mu = J$, and the confidence density of $\mu = \mu(\psi)$ is $L(\psi(\mu))$. There is thus always a parametrisation that makes the likelihood proportional to the confidence density. When the likelihood is based upon a pivot of the form $\mu(\psi) - T$, the likelihood in $\mu = \mu(\psi)$ is proportional to the confidence density of μ.

Example 7. Let $\widehat{\psi}/\psi$ be standard exponentially distributed. Taking the logarithm, the pivot is brought on translation form, and $\mu(\psi) = \log\psi$. The likelihood and the confidence density is thus $c(\mu) \propto L(\mu) = \exp(\widehat{\mu} - \mu - \exp(\widehat{\mu} - \mu))$. Bootstrapping this confidence distribution and likelihood is achieved by adding the bootstrap residuals $\log V^*$ to $\widehat{\mu}$ above, where V^* is standard exponentially distributed. The log-likelihood has a more normal-like shape in the μ parametrisation than in the canonical parameter ψ. Also, being a translation family in μ, the likelihood and the confidence density are easily interpreted. ■

When the likelihood equals the confidence density, the pivot is in broad generality of the translation type. The cumulative confidence distribution function is then of translation type, with $C = F(\mu - \widehat{\mu})$, and so is the likelihood, $L = c = f(\mu - \widehat{\mu})$. In this case, bootstrapping amounts to drawing bootstrap values from the confidence distribution, and substituting these for the point estimate $\widehat{\mu}$. Normal-based confidence distributions that are Gaußian are of the translation type, and are thus particularly easy to bootstrap, as are their likelihoods.

5 Bootstrapping confidence distributions and reduced likelihoods

Bootstrapping has emerged as an indispensable tool in statistical inference. When working with reduced likelihoods it is often desirable to mimic the result of bootstrapping the original data underlying the reduced likelihood and the prior confidence distribution. A bootstrap replicate would then result in a perturbed confidence distribution, and a perturbed reduced likelihood.

Assume the pivot to be invertible in the statistic T, allowing the reduced likelihood to exist. The obvious parametric bootstrap of this statistic at the parameter $\widetilde{\psi}$ solves $\mathrm{piv}(T^*, \widetilde{\psi}) = V^*$, where V^* is a draw from the pivotal distribution F. Then, the parametric bootstrap of the confidence distribution at $\widetilde{\psi}$ is $C^*(\psi; \widetilde{\psi}) = F(\mathrm{piv}(T^*, \psi))$, and the corresponding parametric bootstrap of the reduced likelihood function is

$$L^*(\psi; \widetilde{\psi}) = f(\mathrm{piv}(T^*, \psi))\Big|\frac{\partial \mathrm{piv}(T^*, \psi)}{\partial T^*}\Big|.$$

In the location and scale model (4), $T^* = \mu(\widetilde{\psi}) + \tau(\widetilde{\psi})V^*$, and

$$L^*(\psi; \widetilde{\psi}) = f\Big(\frac{\mu(\widetilde{\psi}) - \mu(\psi) + \tau(\widetilde{\psi})V^*}{\tau(\psi)}\Big)\frac{1}{\tau(\psi)}.$$

When the reduced likelihood is normal-based, the parametric bootstrap of the log-likelihood is

$$\ell^*(\psi; \widetilde{\psi}) = -\tfrac{1}{2}\big\{\mu(\widetilde{\psi}) - \mu(\psi) + Z^*\big\}^2 = -\tfrac{1}{2}\Big\{\Phi^{-1}\big(C^*(\psi; \widetilde{\psi})\big)\Big\}^2,$$

where $Z^* \sim N(0, 1)$. This leads to the bootstrap cumulative confidence distribution function

$$C^*(\psi) = F\Big(\frac{T^* - T_{\mathrm{obs}}}{\tau(\psi)} + F^{-1}(C(\psi))\Big).$$

When the probability basis is normal and the scale τ is constant (and then chosen as unity), the bootstrapped confidence distribution is

$$C^*(\psi) = \Phi\big(\Phi^{-1}(C(\psi)) + T^* - T_{\mathrm{obs}}\big),$$

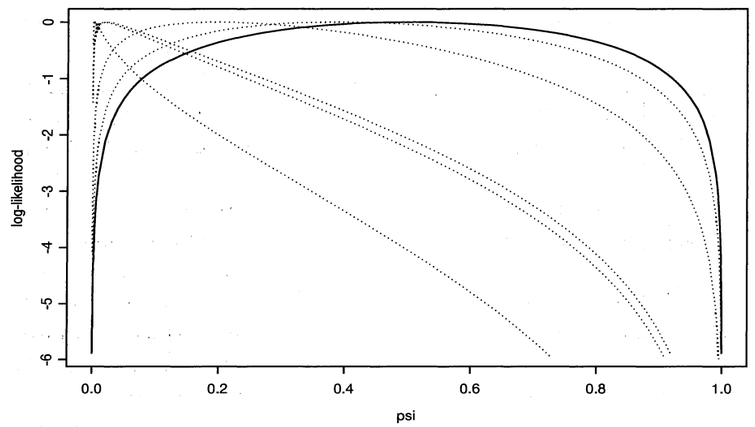

FIGURE 2: Normal-based log-likelihood related to a uniform confidence distribution over $(0, 1)$ (solid line), along with five replicated bootstrap log-likelihoods (dotted).

where T^* is a bootstrap replicate of the normal score of the original statistic, T. On the normal score scale, $T^* - T_{\text{obs}}$ is then normally distributed, and since bias has been removed through the confidence estimation, we may take $T^* - T_{\text{obs}} = Z^* \sim \text{N}(0, 1)$. In this case, the bootstrapped log-likelihood is

$$\ell^*(\psi) = -\tfrac{1}{2}\{\Phi^{-1}(C(\psi)) + Z^*\}^2.$$

Example 4 (cont.). Figure 2 shows the normal-based log-likelihood related to the uniform confidence distribution described in Example 4, together with five bootstrapped log-likelihoods at $\widetilde{\psi} = 0.1$. ■

6 Confidence level and confidence reliability

Let $C(\psi)$ be the cumulative confidence distribution. The intended interpretation of C is that its quantiles are endpoints of confidence intervals. For these intervals to have correct coverage probabilities, the cumulative confidence at the true value of the parameter must have a uniform probability distribution. This is an *ex ante* statement. Before the data have been gathered, the confidence distribution is a statistic with a probability distribution, often based on another statistic through a pivot.

The choice of statistic on which to base the confidence distribution is unambiguous only in simple cases. Barndorff-Nielsen and Cox (1994) are in agreement with Fisher when emphasising the structure of the model and the data as a basis for choosing the statistic. They are primarily interested in the logic of statistical inference. In the tradition of Neyman and Wald, emphasis has been on inductive behaviour, and the goal has been to find methods with optimal frequentist properties. In nice models like exponential families it turns out that methods favoured on structural and logical grounds usually also are favoured on grounds of optimality. This agreement between the Fisherian and Neyman–Wald schools is encouraging and helps to reduce the distinction between

the two schools. See, however, Section 9.2 where we argue that the study protocol might dictate another inference than that based on Neyman–Pearson optimality or Fisherian logic. Nevertheless, the Neyman–Pearson theory is core material in theoretical statistics. This core is in the following reformulated in terms of confidence distributions.

6.1 Reliability and power

A method is *reliable* when it leads to similar conclusions for repeated samples. The more reliable, the less variability in results. A method that is both exact and reliable gives results that vary little, and which are centred at the truth. A cumulative confidence distribution is monotone: at $\psi > \psi_{\text{true}}$, one should have $C(\psi) \geq C(\psi_{\text{true}})$, etc. When C is exact, $C(\psi_{\text{true}}) \sim U.$ (uniform on the unit interval), and above the true value, $C(\psi)$ must be stochastically larger than U (have cumulative distribution function less than that of U). Since $1 \geq C(\psi)$, the more the *ex ante* probability distribution of $C(\psi)$ is shifted towards its upper limit, the less variability it has in repeated samples. For $\psi < \psi_{\text{true}}$, it is desirable to have the probability distribution of $C(\psi)$ concentrated as much as possible towards low values.

The tighter the confidence intervals are, the better, provided they have the claimed confidence. *Ex post*, it is thus desirable to have as little spread in the confidence distribution as possible. Standard deviation, inter-quantile difference or other measures of spread could be used to rank methods with respect to their discriminatory power. The properties of a method must be assessed *ex ante*, and it is thus the probability distribution of a chosen measure of spread that would be relevant. The assessment of the information content in a given body of data is, however, another matter, and must clearly be discussed *ex post*.

In the standard Neyman–Pearson theory, the focus is on spread-measures of the indicator type, $\Gamma(t) = I(t > \psi_1)$ etc. When testing $H_0 : \psi = \psi_0$ versus $H_1 : \psi > \psi_0$, one rejects at level α if $C(\psi_0) < \alpha$. The power of the test is $\Pr\{C(\psi_0) < \alpha\}$ evaluated at a point $\psi_1 > \psi_0$. Cast in terms of p-values, the power distribution is the distribution at ψ_1 of the p-value $C(\psi_0)$. The basis for test-optimality is monotonicity in the likelihood ratio based on a sufficient statistic, S,

$$\text{LR}(\psi_1, \psi_2; S) = L(\psi_2; S)/L(\psi_1; S) \text{ is increasing in } S \text{ for } \psi_2 > \psi_1. \tag{5}$$

From Schweder (1988) we have the following.

Lemma 6 (Neyman–Pearson for p-values) *Let S be a one-dimensional sufficient statistic with increasing likelihood ratio whenever $\psi_1 < \psi_2$. Let the cumulative confidence distribution based on S be C^S and that based on another statistic T be C^T. In this situation, the cumulative confidence distributions are stochastically ordered:*

$$C^S(\psi_0) \overset{ST(\psi)}{\geq} C^T(\psi_0) \text{ at } \psi > \psi_0 \quad \text{and} \quad C^S(\psi_0) \overset{ST(\psi)}{\leq} C^T(\psi_0) \text{ at } \psi < \psi_0.$$

Now, every natural measure of spread in C around the true value of the parameter, ψ_0, can be expressed as a functional $\gamma(C) = \int_{-\infty}^{\infty} \Gamma(\psi - \psi_0)\, C(\mathrm{d}\psi)$, where $\Gamma(0) = 0$, Γ is non-increasing to the left of zero, and non-decreasing to the right. Here $\Gamma(t) = \int_0^t \gamma(\mathrm{d}u)$ is the integral of a signed measure γ.

Agree to say that a confidence distribution C^S is uniformly more reliable in expectation than C^T if

$$\mathrm{E}_{\psi_0}\gamma(C^S) \leq \mathrm{E}_{\psi_0}\gamma(C^T)$$

holds for all spread-functionals γ and at all parameter values ψ_0. With this definition, the Neyman–Pearson lemma yields the following.

Proposition 7 (Neyman–Pearson for power in the mean) *If S is a sufficient one-dimensional statistic and the likelihood ratio* (5) *is increasing in S whenever $\psi_1 < \psi_2$, then the confidence distribution based on S is uniformly most reliable in the mean.*

Proof. By partial integration,

$$\gamma(C) = \int_{-\infty}^{0} C(\psi+\psi_0)\,(-\gamma)(\mathrm{d}\psi) + \int_{0}^{\infty} (1-C(\psi+\psi_0))\,\gamma(\mathrm{d}\psi). \tag{6}$$

By the Neyman–Pearson lemma, $\mathrm{E}C^S(\psi+\psi_0) \leq \mathrm{E}C^T(\psi+\psi_0)$ for $\psi < 0$ while $\mathrm{E}(1-C^S(\psi+\psi_0)) \leq \mathrm{E}(1-C^T(\psi+\psi_0))$ for $\psi > 0$. Consequently, since both $(-\gamma)(\mathrm{d}\psi)$ and $\gamma(\mathrm{d}\psi) \geq 0$,

$$\mathrm{E}_{\psi_0}\gamma(C^S) \leq \mathrm{E}_{\psi_0}\gamma(C^T).$$

This relation holds for all such spread measures that have finite integral, and for all reference values ψ_0. Hence C^S is uniformly more reliable in the mean than any other confidence distribution. ■

The Neyman–Pearson argument for confidence distributions can be strengthened. Say that a confidence distribution C^S is uniformly most reliable if, *ex ante*, $\gamma(C^S)$ is stochastically less than or equal to $\gamma(C^T)$ for all other statistics, T, for all spread-functionals γ, and with respect to the probability distribution at all values of the true parameter ψ_0.

Proposition 8 (Neyman–Pearson for confidence distributions) *If S is a sufficient one-dimensional statistic and the likelihood ratio* (5) *is increasing in S whenever $\psi_1 < \psi_2$, then the confidence distribution based on S is uniformly most reliable.*

Proof. Let S be probability transformed to be uniformly distributed at the true value of the parameter, set at $\psi_0 = 0$ for simplicity. Write $\mathrm{LR}(\psi_0, \psi; S) = \mathrm{LR}(\psi; S)$. By conditioning, and using the sufficiency of S, $C^T(\psi) = 1 - \mathrm{E}_\psi F_0(T \mid S) = 1 - \mathrm{E}_0\left[F_0(T \mid S)\mathrm{LR}(\psi; S)\right]$. Thus, from (6),

$$\gamma(C^T) = \mathrm{E}_0\left[(1-F_0(T\mid S)\int_{-\infty}^{0} \mathrm{LR}(\psi;S)\,(-\gamma)(\mathrm{d}\psi)\right] + \mathrm{E}_0\left[F_0(T\mid S)\int_{0}^{\infty} \mathrm{LR}(\psi;S)\,\gamma(\mathrm{d}\psi)\right]$$

provided these integrals exist. Now, from the sign of γ and from the monotonicity of the likelihood ratio, $h_-(S) = \int_{-\infty}^{0} \mathrm{LR}(\psi;S)(-\gamma)(\mathrm{d}\psi)$ is decreasing in S while $h_+(S) = \int_0^\infty \mathrm{LR}(\psi;S)\gamma(\mathrm{d}\psi)$ is increasing in S. The functions φ_- and φ_+ of S that stochastically minimise

$$\mathrm{E}_0\{\varphi_-(S)h_-(S) + \varphi_+(S)h_+(S)\}$$

under the constraint that both $\varphi_-(S)$ and $\varphi_+(S)$ are uniformly distributed at $\psi_0 = 0$, are $\varphi_-(S) = 1-S$ and $\varphi_+(S) = S$. This choice corresponds to the confidence distribution based on S, and we conclude that $\gamma(C^S)$ is stochastically no greater than $\gamma(C^T)$. ■

6.2 Uniformly most powerful confidence for exponential families

Conditional tests often have good power properties in situations with nuisance parameters. In the exponential class of models it turns out that valid confidence distributions must be based on the conditional distribution of the statistic which is sufficient for the interest parameter, given the remaining statistics informative for the nuisance parameters. That conditional tests are most powerful among power-unbiased tests is well known, see e.g. Lehmann (1959). There are also other broad lines of arguments leading to constructions of conditional tests, see e.g. Barndorff-Nielsen and Cox (1994). Presently we indicate how and why also the most reliable confidence distributions are of such conditional nature.

Proposition 9 *Let ψ be the scalar parameter and χ the nuisance parameter vector in an exponential model, with a density w.r.t. Lebesgue measure of the form*

$$p(y) = \exp\{\psi S(y) + \chi_1 A_1(y) + \cdots + \chi_p A_p(y) - k(\psi, \chi_1, \ldots, \chi_p)\},$$

for data vector y in a sample space region not dependent upon the parameters. Assume (ψ, χ) is contained in an open $(p+1)$-dimensional parameter set. Then, for ψ and hence for all monotone transforms of ψ, there exist exactly valid confidence distributions, and the uniformly most reliable of these takes the conditional form

$$C_{S\,|\,A}(\psi) = \Pr_{\psi,\chi}\{S > S_{\rm obs} \,|\, A = A_{\rm obs}\}.$$

Here $S_{\rm obs}$ and $A_{\rm obs}$ denote the observed values of S and A.

A minor discontinuity correction amendment is called for in case of a discrete distribution, as discussed in Section 2.3.

Proof. The claim essentially follows from previous efforts by a reduction to the one-dimensional parameter case, and we omit the details. A key ingredient is that A is a sufficient and complete statistic for χ when $\psi = \psi_0$ is fixed; this parallels the treatment of Neyman–Pearson optimality of conditional tests for the exponential family, as laid out e.g. in Lehmann (1959). Note that the distribution of S given $A = A_{\rm obs}$ depends on ψ but not on $\chi_1, \ldots, \chi_p$. ■

Example 8. Consider pairs (X_j, Y_j) of independent Poisson variables, where X_j and Y_j have parameters λ_j and $\lambda_j\psi$, for $j = 1, \ldots, m$. The likelihood is proportional to

$$\exp\Big\{\sum_{j=1}^m y_j \log\psi + \sum_{j=1}^m (x_j + y_j)\log\lambda_j\Big\}.$$

Write $S = \sum_{j=1}^m Y_j$ and $A_j = X_j + Y_j$. Then $A_1, \ldots, A_m$ become sufficient and complete for the nuisance parameters when ψ is fixed. Also, $Y_j \,|\, A_j$ is a binomial $(A_j, \psi/(1+\psi))$. It follows from the proposition above that the (nearly) uniformly most reliable confidence distribution, used here with a half-correction for discreteness, takes the simple form

$$\begin{aligned} C_{S\,|\,A}(\psi) &= \Pr_\psi\{S > S_{\rm obs} \,|\, A_{1,\rm obs}, \ldots, A_{m,\rm obs}\} + \tfrac12 \Pr_\psi\{S = S_{\rm obs} \,|\, A_{1,\rm obs}, \ldots, A_{m,\rm obs}\} \\ &= 1 - \mathrm{Bin}\Big(S_{\rm obs} \,\Big|\, \sum_{j=1}^m A_{j,\rm obs}, \frac{\psi}{1+\psi}\Big) + \tfrac12 \mathrm{bin}\Big(S_{\rm obs} \,\Big|\, \sum_{j=1}^m A_{j,\rm obs}, \frac{\psi}{1+\psi}\Big), \end{aligned}$$

where $\mathrm{Bin}(\cdot \,|\, n, p)$ and $\mathrm{bin}(\cdot \,|\, n, p)$ are the cumulative and pointwise distribution functions for the binomial. ■

The optimality of the conditional confidence distribution, and thus of conditional tests and confidence intervals, hinges on the completeness of the ancillary statistic A. By completeness, there cannot be more than one exact confidence distribution based on the sufficient statistic. The conditional confidence distribution is exact, and is thus optimal since it is the only exact one. The question is then whether some approximate confidence distributions dominate the conditional one in overall performance in some specified sense. This might be the case in some situations; see Section 9.2.

6.3 Large-sample optimality

Consider any regular parametric family, with a suitable density $f(x, \theta)$ involving a p-dimensional parameter θ. Assume data $X_1, \ldots, X_n$ are observed, with consequent maximum likelihood estimator $\widehat\theta_n$. Let furthermore θ_0 denote the true value of the parameter. It is well known that $\widehat\theta_n$ is

approximately distributed as a normal, centred at θ_0, for large n. The following statement is loosely formulated, but may be made precise in various ways. The above situation, for large n, is approximately the same as that of observing $\widehat{\theta}_n$ from the model with density $\exp\{\sum_{j=1}^p \theta_j u_j(x) - nB(\theta)\}$, where $u_j(x) = \partial \log f(x,\theta_0)/\partial\theta_j$, and $nB(\theta)$ the appropriate normalisation constant. This goes to show that the inference situation is approximated with the form described in Proposition 9. Thus, broadly speaking, the ordinary confidence interval constructions based on maximum likelihood machinery become asymptotically optimal. Section 8.1 offers some insight in first order asymptotics, while Sections 8.2 and 8.3 discuss asymptotic methods that aim at being second order correct.

7 Equivariant and minimax confidence distributions

There are complementary approaches towards constructions of and comparisons between confidence distributions. This section briefly sets down some theory for equivariant confidence distributions and discusses minimax strategies under a natural loss function.

7.1 Equivariance

Suppose data X in sample space $\mathcal{X}$ follow a distribution modelled as P_θ, where $\theta \in \Omega$ is the unknown parameter, and let $\psi = h(\theta)$ be the interest parameter for which a confidence distribution is sought. Assume that transformations $g \in \mathcal{G}$ are such that the problem is left equivariant; when $X \sim P_\theta$, $g(X)$ follow distribution $P_{\bar{g}(\theta)}$, where $g\colon \mathcal{X} \to \mathcal{X}$ and $\bar{g}\colon \Omega \to \Omega$ are 1–1 and surjective. See Lehmann (1983, Ch. 3) for such a framework (for different purposes). In such a situation, it makes sense to restrict attention to confidence distributions $C(\psi) = C(\psi; X)$ that are equivariant, in the sense that

$$C(\psi; X) = C(\bar{\psi}; g(X)), \quad \text{where } \bar{\psi} = h(\bar{g}(\theta)), \quad \text{for all } g \in \mathcal{G}. \tag{7}$$

Constructions obeying (7) have the property that they give the same result each time the statistician is faced with the model and type of data in question.

Equivariance helps to reduce data down to a one-dimensional statistic in fortunate situations. When a pivot exists in this statistic, it determines the confidence distribution. In this connection see also Fraser (1968, 1996). The pivot also determines the reduced likelihood and dictates how to bootstrap these statistics.

Example 9. Assume there are two independent normal samples of sizes n_1 and n_2, with respectively $X_i \sim \mathrm{N}(\mu_1, \sigma_1^2)$ and $Y_j \sim \mathrm{N}(\mu_2, \sigma_2^2)$, and assume that interest focusses on $\psi = \sigma_2/\sigma_1$. Transforming data to $X_i' = aX_i + b$ and $Y_j' = cY_j + d$, where a and c are positive, keeps the model as such intact, with transformed parameters $(a\mu_1 + b, a\sigma_1, c\mu_2 + d, c\sigma_2)$. Write $\bar{X}, \bar{Y}, S_x, S_y$ for the sample means and standard deviations. An equivariant confidence distribution based on this set of sufficient statistics must then obey

$$C(\psi; \bar{X}, \bar{Y}, S_x, S_y) = C((c/a)\psi; a\bar{X} + b, c\bar{Y} + d, aS_x, cS_y) \quad \text{for all } a, b, c, d.$$

Setting $b = -a\bar{X}$ and $d = -c\bar{Y}$, and then for example $a = 1/S_x = c$, leads to $C(\psi)$ being a function of $\widehat{\psi} = S_y/S_x$ alone. Proposition 8 then implies that $C(\psi) = 1 - K(\widehat{\psi}^2/\psi^2)$ is uniformly most reliable among all equivariant confidence distributions, with K the cdf of the F distribution with $n_2 - 1$ and $n_1 - 1$ degrees of freedom, as in Example 2. The reduced equivariant likelihood becomes $L(\psi) = k(\widehat{\psi}^2/\psi^2)/\psi^2$, where $k = K'$ is the density of the F distribution. ■

7.2 Admissible and minimax methods

When data X give rise to a confidence set $A = A(X)$ for the parameter $\psi = h(\theta)$, consider the loss function $L(\theta, A) = k\, m(A) + I\{\psi \notin A\}$. Here m is Lebesgue measure (typically measuring the length of the interval A) while k is a fixed positive constant, possibly modified by a further scale parameter, balancing the two desiderata of good confidence intervals. Using such a loss function amounts to assessing the quality of confidence procedures via their risk functions $R(A, \theta) = k\, \mathrm{E}_\theta m(A) + 1 - \Pr_\theta\{\psi \in A\}$. This is accordingly within the usual decision-theoretic setup, where one may find Bayes solutions, minimax and admissible confidence interval methods, the best invariant procedures, and so on.

As a simple example, consider a sample $X_1, \ldots, X_n$ from $\mathrm{N}(\mu, \sigma^2)$, and let $(k/\sigma)\, m(A) + I\{\mu \notin A\}$ be the loss function for confidence intervals for μ. The risk function for the particular method $A = \bar{X} \pm bS$, where $\bar{X}$ and S are mean and standard deviation, becomes $R(A, \mu, \sigma) = 2kbe_{n-1} + 1 - 2\int_0^{n^{1/2}b} f_{n-1}(u)\, \mathrm{d}u$, where f_{n-1} is the t density with $n-1$ degrees of freedom and $e_{n-1} = \mathrm{E}\{\chi^2_{n-1}/(n-1)\}^{1/2}$. This expression can easily be minimised over b, giving the best interval of this type, say $A_0 = \bar{X} \pm b_0(k)S$. One may show that this interval is minimax and admissible, under the given loss function. It is also of interest to work out Bayes solutions under relevant priors for the parameters.

To connect such an approach to the present development of confidence distributions, one needs to work with a class of loss functions of the above type, where the k in question becomes a function of confidence level α.

8 Approximate confidence distributions and reduced likelihoods

Uniformly most reliable exact inference is only possible in nice models. In a wider class of models, exact confidence distributions are available. The estimate of location based on the Wilcoxon statistic has for example an exact known distribution in the location model where only symmetry is assumed. In more complex models, the statistic upon which to base the confidence distribution might be chosen on various grounds: the structure of the likelihood function, perceived robustness, asymptotic properties, computational feasibility, perspective and tradition of the study. In the given model, with finite data, it might be difficult to obtain an exact confidence distribution based on the chosen statistic. There are, however, various techniques available to obtain approximate confidence distributions and reduced likelihoods.

Bootstrapping, simulation and asymptotics are useful tools in calculating approximate confidence distributions and in characterising their power properties. When an estimator, often the maximum likelihood estimator of the interest parameter, is used as the statistic on which the confidence distribution is based, bootstrapping provides an estimate of the sampling distribution of the statistic. This empirical sampling distribution can be turned into an approximate confidence distribution in several ways. The simplest and most widely used method of obtaining approximate confidence intervals is the delta method. This will lead to first order accuracy properties in smooth models. A more refined method to obtain confidence distributions is via acceleration and bias corrections on bootstrap distributions, as developed below. This method, along with several other venues for refinement, will usually provide second order accuracy properties.

8.1 The delta method

In a sample of size n, let the estimator $\widehat{\theta}_n$ have an approximate multinormal distribution centred at θ and with covariance matrix of the form S_n/n, so that $\sqrt{n}S_n^{-1/2}(\widehat{\theta}_n - \theta) \to_d \mathrm{N}(0, I)$. By the delta method, the confidence distribution for a parameter $\psi = h(\theta)$ is based on linearising h at $\widehat{\theta}$, and yields

$$C_{\mathrm{delta}}(\psi) = \Phi((\psi - \widehat{\psi})/\widehat{\sigma}_n) \tag{8}$$

in terms of the cumulative standard normal. The variance estimate is $\widehat{\sigma}_n^2 = \widehat{g}^{\mathrm{tr}} S_n \widehat{g}/n$ where $\widehat{g}$ is the gradient of h evaluated at $\widehat{\theta}$. Again, this estimate of the confidence distribution is to be displayed post data with $\widehat{\psi}$ equal to its observed value $\widehat{\psi}_{\mathrm{obs}}$.

This confidence distribution is known to be first order unbiased under weak conditions. That $C_{\mathrm{delta}}(\psi)$ is first order unbiased means that the coverage probabilities converge at the rate $n^{-1/2}$, or that $C_{\mathrm{delta}}(\psi_{\mathrm{true}})$ converges in distribution to the uniform distribution at the $n^{1/2}$ rate. Note also that the confidence density as estimated via the delta method, say $c_{\mathrm{delta}}(\psi)$, is simply the normal density $\mathrm{N}(\widehat{\psi}, \widehat{\sigma}_n^2)$.

The additivity of the asymptotically normal pivot implies that the reduced likelihood is Gaußian and actually identical to the confidence density $c_{\mathrm{delta}}(\psi)$. That the reduced likelihood of a linear parameter in a multivariate normal location model is obtained from the marginal normal distribution of its maximum likelihood estimator also makes good sense in view of the factorisation of the joint likelihood.

8.2 The t-bootstrap method

For a suitable monotone transformation of ψ and $\widehat{\psi}$ to $\gamma = h(\psi)$ and $\widehat{\gamma} = h(\widehat{\psi})$, suppose

$$t = (\widehat{\gamma} - \gamma)/\widehat{\tau} \quad \text{is an approximate pivot,} \tag{9}$$

where $\widehat{\tau}$ is proportional to an estimate of the standard deviation of $\widehat{\gamma}$. Let R be the distribution function of t, by assumption approximately independent of underlying parameters (ψ, χ). The approximate confidence distribution for γ is thus $C(\gamma) = 1 - R((\widehat{\gamma} - \gamma)/\widehat{\tau})$, yielding in its turn $C(\psi) = 1 - R((h(\widehat{\psi}) - h(\psi))/\widehat{\tau})$ for ψ, with appropriate confidence density $c(\psi) = C'(\psi)$. Now R would often be unknown, but the situation is saved via bootstrapping. Let $\widehat{\gamma}^* = h(\widehat{\theta}^*)$ and $\widehat{\tau}^*$ be the result of parametric bootstrapping from the estimated model. Then the R distribution can be estimated arbitrarily well as $\widehat{R}$, say, obtained via bootstrapped values of $t^* = (\widehat{\gamma}^* - \widehat{\gamma})/\widehat{\tau}^*$. The confidence distribution reported is then as above but with $\widehat{R}$ replacing R:

$$C_{\mathrm{tboot}}(\psi) = 1 - \widehat{R}\big((h(\widehat{\psi}) - h(\psi))/\widehat{\tau}\big).$$

Example 10. Figure 3 illustrates the t-bootstrap method for the case of the correlation coefficient in the binormal family, using Fisher's zeta transformation $h(\rho) = \frac{1}{2}\log\{(1+\rho)/(1-\rho)\}$ and a constant for $\widehat{\tau}$. The density $c_{\mathrm{tboot}}(\rho)$ is shown rather than its cumulative, and has been computed via numerical derivation. We note that the exact confidence distribution for ρ involves the distribution of the empirical correlation coefficient $\widehat{\rho}$, which however is quite complicated and is available only as an infinite sum. ■

This t-bootstrap method applies even when t is not a perfect pivot, but is especially successful when it is, since t^* then has exactly the same distribution R as t. Note that the method automatically takes care of bias and asymmetry in R, and that it therefore aims at being more precise

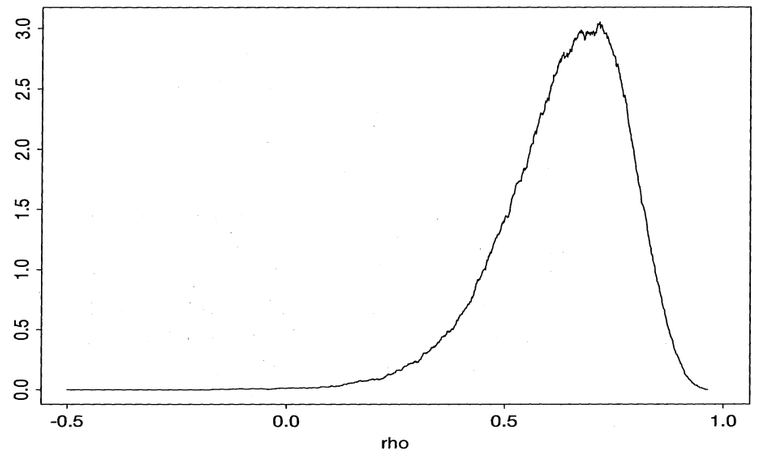

FIGURE 3: Approximate confidence density for a binormal correlation coefficient, having observed $\widehat{\rho} = 0.667$ from $n = 20$ data pairs, via the t-bootstrap method. The confidence density curve is computed via numerical derivation of the $C_{\text{tboot}}(\rho)$ curve, using 5000 bootstrap samples.

than the delta method above, which corresponds to zero bias and a normal R. The problem is that an educated guess is required for a successful pivotal transformation h, and that the interval is not invariant under monotone transformations. The following method is not hampered by these shortcomings.

8.3 The acceleration and bias corrected bootstrap method

Efron (1987) introduced acceleration and bias corrected bootstrap percentile intervals, and showed that these have several desirable aspects regarding accuracy and parameter invariance. Here we will exploit some of these ideas, but 'turn them around' to construct accurate bootstrap-based approximations to confidence distributions.

Suppose that on some transformed scale, from ψ and $\widehat{\psi}$ to $\gamma = h(\psi)$ and $\widehat{\gamma} = h(\widehat{\psi})$, one has

$$(\gamma - \widehat{\gamma})/(1 + a\gamma) - b \sim \mathrm{N}(0, 1) \tag{10}$$

to a very good approximation, for suitable constants a (for acceleration) and b (for bias). Both population parameters a and b tend to be small; in typical setups with n observations, their sizes will be $O(n^{-1/2})$. Assuming $a\widehat{\gamma} > -1$, the pivot in (10) is increasing in γ and $C(\gamma) = \Phi\big((\gamma - \widehat{\gamma})/(1 + a\gamma) - b\big)$ is the confidence distribution for γ. Thus

$$C(\psi) = \Phi\Big(\frac{h(\psi) - h(\widehat{\psi})}{1 + ah(\psi)} - b\Big) \tag{11}$$

is the resulting confidence distribution for ψ. This constitutes a good approximation to the real confidence distribution, say $C_{\text{exact}}(\psi)$, under assumption (10). It requires h to be known, however, as well as values of a and b.

To come around this, look at bootstrapped versions $\widehat{\gamma}^* = h(\widehat{\psi}^*)$ from the estimated parametric model. If assumption (10) holds uniformly in a neighbourhood of the true parameters, then also

$$(\widehat{\gamma}^* - \widehat{\gamma})/(1 + a\widehat{\gamma}) \sim \mathrm{N}(-b, 1)$$

with good precision. Hence the bootstrap distribution may be expressed as

$$\widehat{G}(t) = \Pr_*\{\widehat{\psi}^* \le t\} = \Pr_*\{\widehat{\gamma}^* \le h(t)\} = \Phi\Big(\frac{h(t) - \widehat{\gamma}}{1 + a\widehat{\gamma}} + b\Big),$$

which yields $h(t) = (1 + a\widehat{\gamma})\{\Phi^{-1}(\widehat{G}(t)) - b\} + \widehat{\gamma}$. Substitution in (11) is seen to give the abc formula

$$\widehat{C}_{\mathrm{abc}}(\psi) = \Phi\Big(\frac{\Phi^{-1}(\widehat{G}(\psi)) - b}{1 + a(\Phi^{-1}(\widehat{G}(\psi)) - b)} - b\Big), \tag{12}$$

since $\Phi^{-1}(\widehat{G}(\widehat{\psi})) = b$. Note that an approximation $c_{\mathrm{abc}}(\psi)$ to the confidence density emerges too, by evaluating the derivative of $\widehat{C}_{\mathrm{abc}}$. This may sometimes be done analytically, in cases where $\widehat{G}(\psi)$ can be found in a closed form, or may be carried out numerically.

The reduced abc likelihood is from (10) equal to $L(\gamma) = \phi((\gamma - \widehat{\gamma})/(1 + a\gamma))/(1 + a\gamma)$, which yields the log-likelihood

$$\ell_{\mathrm{abc}}(\psi) = -\tfrac{1}{2}\{\Phi^{-1}(\widehat{C}_{\mathrm{abc}}(\psi))\}^2 - \log[1 + a\{\Phi^{-1}(\widehat{G}(\psi)) - b\}],$$

since the unknown proportionality factor $1 + a\widehat{\gamma}$ appearing in $h(t)$ is a constant proportionality factor in $L_{\mathrm{abc}}(h(\psi))$.

It remains to specify a and b. The bias parameter b is found from $\widehat{G}(\widehat{\psi}) = \Phi(b)$, as noted above. The acceleration parameter a is found as $a = \frac{1}{6}\mathrm{skew}$, where there are several ways in which to calculate or approximate the skewness parameter in question. Extensive discussions may be found in Efron (1987), Efron and Tibshirani (1993, Chs. 14 and 22) and in Davison and Hinkley (1997, Ch. 5). One option is via the jackknife method, which gives parameter estimates $\widehat{\psi}_{(i)}$ computed by leaving out data point i, and use

$$a = (6n^{1/2})^{-1}\mathrm{skew}\{\widehat{\psi}_{(\cdot)} - \widehat{\psi}_{(1)}, \ldots, \widehat{\psi}_{(\cdot)} - \widehat{\psi}_{(n)}\}.$$

Here $\widehat{\psi}_{(\cdot)}$ is the mean of the n jackknife estimates. Another option for parametric families is to compute the skewness of the logarithmic derivative of the likelihood, at the parameter point estimate, inside the least favourable parametric subfamily; see again Efron (1987) for more details.

Note that when a and b are close to zero, the abc confidence distribution becomes identical to the bootstrap distribution itself. In typical setups, both a and b will in fact go to zero with speed of order $1/n^{1/2}$ in terms of sample size n. Thus (12) provides a second order non-linear correction of shift and scale to the immediate bootstrap distribution.

Example 11. Consider again the parameter $\psi = \sigma_2/\sigma_1$ of Example 2. The exact confidence distribution was derived there and is equal to $C(\psi) = 1 - K(\widehat{\psi}^2/\psi^2)$, with $K = K_{\nu_2,\nu_1}$. We shall see how successful the abc apparatus is for approximating the $C(\psi)$ and its confidence density $c(\psi)$.

In this situation, bootstrapping from the estimated parametric model leads to $\widehat{\psi}^* = \widehat{\sigma}_2^*/\widehat{\sigma}_1^*$ of the form $\widehat{\psi}F^{1/2}$, where F has degrees of freedom ν_2 and ν_1. Hence the bootstrap distribution is $\widehat{G}(t) = K(t^2/\widehat{\psi}^2)$, and $\widehat{G}(\widehat{\psi}) = K(1) = \Phi(b)$ determines b. The acceleration constant can be

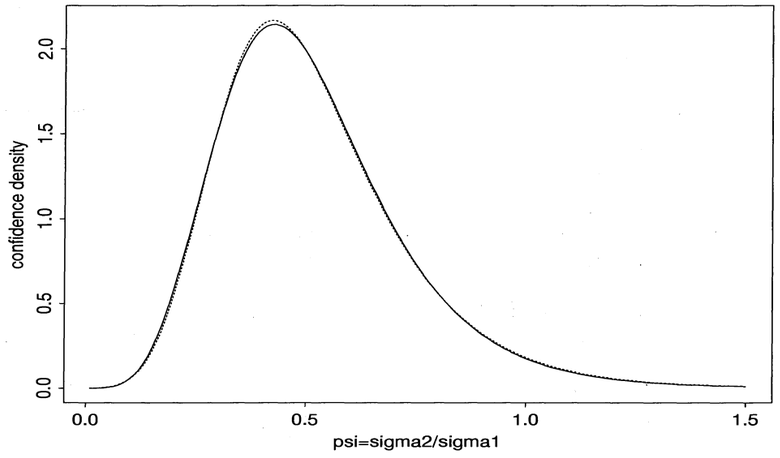

FIGURE 4: True confidence density along with abc-estimated version of it (dotted line), for parameter $\psi = \sigma_2/\sigma_1$ with 5 and 10 degrees of freedom. The parameter estimate in this illustration is $\widehat{\psi} = 0.50$. The agreement is even better when ν_1 and ν_2 are closer or when they are larger.

computed exactly by looking at the log-derivative of the density $\widehat{\psi}$, which from $\widehat{\psi} = \psi F^{1/2}$ is equal to $p(r, \psi) = k(r^2/\psi^2)2r/\psi^3$. With a little work the log-derivative can be expressed as

$$\frac{1}{\psi}\Big\{-\nu_2 + (\nu_1+\nu_2)\frac{(\nu_2/\nu_1)\widehat{\psi}^2/\psi^2}{1+(\nu_2/\nu_1)\widehat{\psi}^2/\psi^2}\Big\} =_d \frac{\nu_1+\nu_2}{\psi}\Big\{\text{Beta}(\tfrac{1}{2}\nu_2, \tfrac{1}{2}\nu_1) - \frac{\nu_2}{\nu_1+\nu_2}\Big\}.$$

Calculating the three first moments of the Beta gives a formula for its skewness and hence for a. (Using the jackknife formula above, or relatives directly based on simulated bootstrap estimates, obviates the need for algebraic derivations, but would give a good approximation only to the a parameter for which we here found the exact value.)

Trying out the abc machinery shows that $\widehat{C}_{\text{abc}}(\psi)$ is amazingly close to $C(\psi)$, even when the degrees of freedom numbers are low and imbalanced; the agreement is even more perfect when ν_1 and ν_2 are more balanced or when they become larger. The same holds for the densities $\widehat{c}_{\text{abc}}(\psi)$ and $c(\psi)$; see Figure 4. ■

8.4 Comparisons

The delta method and the abc method remove bias by transforming the quantile function of the otherwise biased normal confidence distribution, $\Phi(\psi - \widehat{\psi})$. The delta method simply corrects the scale of the quantile function, while the abc method applies a shift and a non-linear scale change to remove bias both due to the non-linearity in ψ as a function of the basic parameter θ as well as the effect on the asymptotic variance when the basic parameter is changed. The t-bootstrap method would have good theoretical properties in cases where the $\widehat{\psi}$ estimator is a smooth function of sample averages, but has a couple of drawbacks compared to the abc method. It is for example not invariant under monotone transformations. Theorems delineating suitable second-order correctness aspects of both the abc and the t-bootstrap methods above can be formulated and proved, with

necessary assumptions having to do with the quality of approximations involved in (9) and (10). Methods of proof would for example involve Edgeworth or Cornish–Fisher expansion arguments. Such could also be used to add corrections to the delta method (8).

Some asymptotic methods of debiasing an approximate confidence distribution involves a transformation of the confidence itself and not its quantile function. From a strict mathematical point of view there is of course no difference between acting on the quantiles or the confidence. But methods like the abc method above are most naturally viewed as a transformation of the confidence for each given value of the parameter.

There are still other methods of theoretical and practical interest for computing approximate confidence distributions, cf. the broad literature on constructing accurate confidence intervals. One approach would be via analytic approximations to the endpoints of the abc interval, under suitable assumptions; the arguments would be akin to those found in DiCiccio and Efron (1996) and Davison and Hinkley (1997, Ch. 5) regarding 'approximate bootstrap confidence intervals'. Another approach would be via modified profile likelihoods, following work by Barndorff-Nielsen and others; see Barndorff-Nielsen and Cox (1994, Chs. 6 and 7) and Barndorff-Nielsen and Wood (1998). Clearly more work and further illustrations are needed to better sort out which methods have the best potential for accuracy and transparency in different situations. At any rate the abc method (12) appears quite generally useful and precise.

9 Confidence inference for capture-recapture data

Consider a closed population of N individuals. Captured individuals might be marked in the course of the study, or they might have unique natural marks that are observed, say on photographs. Captures are made on four occasions, with X_t different individuals captured on occasion $t = 1, 2, 3, 4$ and with X unique captures made in the combined sample. We seek a confidence distribution and a likelihood for the population size N.

9.1 A multinomial recapture model

Assuming captures to be stochastically independent between occasions and letting all individuals having the same capture probability p_t on occasion t, we have the multinomial multiple-capture model of Darroch (1958). The likelihood is

$$L(N, p_1, \ldots, p_4) \propto \binom{N}{X} \prod_{t=1}^{4} p_t^{X_t} (1 - p_t)^{N - X_t},$$

and $\{X_t\}$ is thus ancillary for N. By Fisher's inductive logic, inference on N should therefore be based on the conditional distribution of X given $\{X_t\}$. For fixed N, $\{X_t\}$ is sufficient and complete for $\{p_t\}$. The conditions of the extended Neyman–Pearson lemma are therefore satisfied, except that the data are discrete and not continuously distributed. The confidence distribution for N based on X in the conditional model given $\{X_t\}$ is therefore also suggested by the extended Neyman–Pearson lemma. With half-correction due to discreteness, the cdf of the confidence distribution is

$$C(N) = \Pr_N\{X > X_{\text{obs}} | \{X_t\}\} + \tfrac{1}{2}\Pr_N\{X = X_{\text{obs}} | \{X_t\}\}.$$

It is nearly optimal in the sense of being uniformly most powerful among exact confidence distributions. Since the data are discrete, the conditional confidence distribution is not exact, and we are precluded from stating exact optimality.

	X_1	X_2	X_3	X_4	X
Immature	15	32	9	11	62
Mature	44	20	49	7	113

Table 1: Observed numbers of individuals.

The conditional distribution is computed via the hypergeometric distribution. Let R_t be the number of recaptures on occasion t relative to previous captures. Set $R_1 = 0$. The total number of recaptures is $R = \sum_{t=1}^{4} R_t = X - \sum_{t=1}^{4} X_t$. Given the number of unique captures previous to t, $\sum_{i=1}^{t-1}(X_i - R_i)$, R_t has a hypergeometric distribution. In obvious notation, the conditional distribution is therefore

$$\Pr\{X = x \mid \{x_t\}\} = \sum_{r_2=0}^{r} \sum_{r_3=0}^{r-r_2} \prod_{t=2}^{4} \binom{N - \sum_1^{t-1}(x_i - r_i)}{x_t - r_t} \binom{\sum_1^{t-1}(x_i - r_i)}{r_t} \Big/ \binom{N}{x_t}. \qquad (13)$$

The present approach generalises to an arbitrary number of recaptures, but it assumes the population to be closed and homogeneous with respect to capturing, which is independent over capturing occasions. We will return to this multinomial multiple-recapture model in Section 9, noting that conditioning on the numbers of captures over occasions is sensible for one type of study protocol. For other protocols for such studies, quite different pivots and resulting confidence distributions and reduced likelihoods are appropriate, despite the near 'optimality'.

Application: Bowhead whales in Alaska. In the summers and autumns of 1985 and 1986, photographs were taken of bowhead whales north of Alaska (see da Silva et al., 2000 and Schweder, 2000). We shall mainly be concerned with the immature component of the population that had natural marks on their bodies. The numbers of identified individuals in photographs taken on each of the four sampling occasions and in the pooled set of photographs are given in Table 1. The table also gives data for the marked mature whales. The confidence distribution for number of immature whales is $C(N) = \Pr_N\{X > 62\} + \frac{1}{2}\Pr_N\{X = 62\}$, calculated in the conditional distribution (13).

The conditional probability provides a reduced likelihood for N, $L(N) = \Pr_N\{X = 62\}$. The likelihood happens to be extremely close to the normal-based likelihood calculated from $C(N)$. See Figure 5. It is also quite close to the profile likelihood. This agreement is due to the underlying conditional pivot being in the conditional maximum likelihood estimate which is approximately normal and additive in a function of N. To an amazing accuracy, we find $C(N) \approx \Phi(5.134 - 87.307N^{-1/2})$. The natural parameter is thus $\mu(N) = 1/N^{1/2}$. Due to the nonlinearity in the natural parameter, the likelihood is different from the confidence density (taking N to be continuous); in this case the difference is actually substantial, see Figure 5.

The same picture emerges for mature whales. Here we find the conditional confidence distribution to be rather accurately given by $N^{-1/2} \sim \mathrm{N}(0.03734, 0.0064^2)$. Again, the conditional likelihood $L(N) = \Pr_N\{X = 113\}$ is well approximated by the normal-based reduced likelihood.

Passing now to the total number of marked whales, $N = N_i + N_m$ where N_i is the number of marked immatures and N_m the number of marked mature whales, the problem is to estimate N. A simple approach is now to bootstrap each of the two normal-based likelihoods and to calculate the maximum likelihood estimate of N for each replicate. Due to the additivity of the pivots in the two natural parameters, this amounts to drawing bootstrap replicates from the joint confidence distribution for (N_i, N_m) and then add.

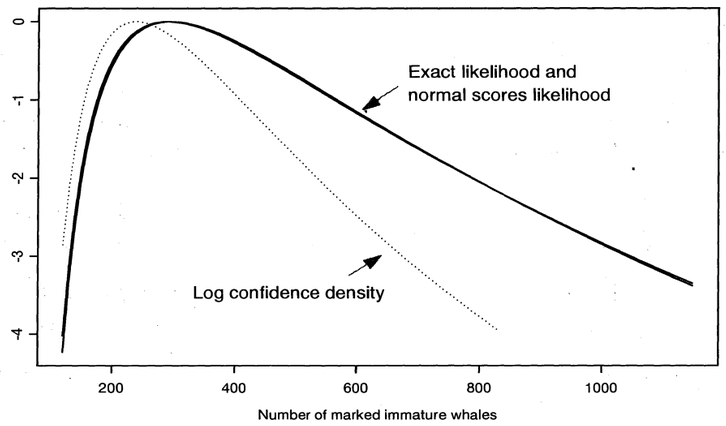

FIGURE 5: The exact likelihood is shown together with the normal scores likelihood, for the number of marked immature whales; these curves are almost identical. Also shown is the log-confidence density (dotted).

9.2 On the importance of the study protocol

The Bayesian approach is to integrate the prior distribution with the likelihood, as if they both were probability distributions over the joint space of the parameter and the data. In practice, the Bayesian posterior distribution is based on the observed likelihood function. What could have been observed is of no consequence. The same applies to the purist likelihoodist. They both agree that the observed likelihood function carries all the information, and contrafactual thoughts of data that could have been realised but were not, is of no concern. The frequentist departs from this by insisting that his 95% confidence interval should cover the true parameter with probability 0.95 in a hypothetical repetition of the experiment, regardless of the true state of nature, i.e. the value of the unknown parameter. The confidence distribution is a truly frequentist concept, and it is not a function of the observed likelihood alone. As seen in Examples 3 and 5 a given observed likelihood can lead to a different confidence distribution when the model is changed. Whether a Poisson process is observed in t units of time until x points have been realised, or whether the number of points x is what is observed over a given time period of length t, should have been decided ahead of the experiment and stated in the study protocol (we prefer 'study protocol' over the synonymous 'experimental design' since many sets of data are generated by an observational process rather than a manipulated experiment). The study protocol is a description of the observer, and a good statistical model reflects the study protocol as well as characteristics of the system under study. The study protocol mattered in the Poisson process situation, but not a great deal. In some situations, the statistical model is formally identical across studies, but the study protocol dictates how to use the model and thus how to obtain a confidence distribution and a reduced likelihood as the case may be. The study protocol might matter a great deal, as seen in the following example.

Application (cont.). Consider four different hypothetical study protocols for the multinomial multiple-recapture process of immature marked whales. The situation is as discussed above.

Protocol	Pivot	C	$c_{.025}$	$c_{.5}$	$c_{.975}$
$X_t = x_t$	$\widehat{N}^{-1/2} - aN^{-1/2} = \sigma Z$	$N^{-\frac{1}{2}} \sim .0588 + .014Z$	152	289	752
$\mathrm{E}X_t = x_t$	$\widehat{N}^{-1/2} - aN^{-1/2} = \sigma Z$	$N^{-\frac{1}{2}} \sim .0596 + .0127Z$	140	282	830
$\mathrm{E}X_t = x_t(N/289)^{1/2}$	$(N/\widehat{N})^{1/2} - a = \sigma Z$	$N^{\frac{1}{2}} \sim 16.8 + 3.67Z$	92	282	575
$\mathrm{E}X_t = x_t N/289$			0	281	495

Table 2: Results for immature marked whales under four hypothetical study protocols. The last columns gives confidence quantiles.

Under the first protocol, sampling continues on occasion t until $X_t = x_t$, where x_t is the observed sample size given in Table 1. In the second case, the expected sample size is given, $\mathrm{E}X_t = Np_t = x_t$. The third protocol aims at a given precision in the resulting abundance estimate. This entails $p_t = c_t/N^{1/2}$ where c_t are fixed constants. For comparison, assume $\mathrm{E}X_t = x_t(N/289)^{1/2}$. The fourth protocol is that of a given sampling effort (perhaps determined by the budget of the study). Now, p_t is independent of N, and we set $\mathrm{E}X_t = x_t N/289$ for easy comparison. The first and last protocols are practical, while the middle two are more difficult to deploy in practice.

The four protocols are given in Table 2, together with approximate pivots, confidence distributions and confidence quantiles. Whatever the protocol, assume the observed data to be the same. The statistical model for the study is formally unaffected, and given by (13). To proceed with conditional inference given $\{X_t\}$ is sensible under the first two protocols. It is, however, less sensible under the two other protocols. Then the expected number of individuals captured on a given occasion tends to increase in N. It is thus not obvious that conditional inference is sensible, even though $\{X_t\}$ is ancillary for N, despite Fisher's inductive logic, as supported by Barndorff-Nielsen and Cox (1994) and others, and despite the extended Neyman–Pearson lemma.

Table 2 is obtained as follows. In the first case, the pivot is found from a simulation study with $\widehat{N}$ as the conditional maximum likelihood estimator, and the confidence distribution is found from the half-corrected tail probability as discussed above. In the remaining cases, $\widehat{N}$ is the maximum likelihood estimator, and simulation is carried out to identify the pivot and the confidence distribution. The search for a pivot was unsuccessful in the constant effort case. The confidence distribution is found to be improper, with a point mass larger than 0.025 at 0, and no closed form was found.

Conditional inference (first protocol) leads essentially to the same results as when expected sample size is fixed (second protocol). The confidence distribution is slightly less dispersed when conditioning, as expected. The confidence distribution is markedly skewed, with a long tail to the right. This makes sense, since if the population is large, the fixed number of captured will lead to very few recaptures, and eventually $\sum X_t = X$ with high probability, with very little information on N. The other two confidence distributions are centred at the same point estimate, but they are differently skewed. Under the last protocol of constant sampling effort, there is hardly any information in the data if the population is small and hence the number of captures is small. It is therefore sensible to be cautious towards small values. On the other hand, many captures will be made if the population is large, with consequent high information on the population size. This explains the short right tail of the confidence distribution in this case.

10 Discussion

The confidence distribution is an attractive format for reporting statistical inference for parameters of primary interest. To allow future good use of the results it is desirable to allow a likelihood to be

constructed from the confidence distribution. An alternative is to make the original data available, or to present the full likelihood. However, the work invested in reducing the original data to a confidence distribution for the parameter of interest would then be lost. To convert the posterior confidence distribution to a likelihood, and to allow future correct bootstrapping, the probability basis for the confidence distribution must be reported. This is often achieved by reporting the underlying pivot and its distribution. Our suggestion is accordingly to extend current frequentist reporting practice from only reporting a point estimate, a standard error and a (95%) confidence interval for the parameters of primary interest. To help future readers, one should report the confidence distribution fully, and supplement it with information on its probability basis.

10.1 Advantages with our approach

The advantages of representing the information contained in a confidence distribution in the format of (an approximate) likelihood function are many and substantial.

By adding the log-likelihoods of independent confidence distributions for the same parameter, a combined confidence distribution is obtained, usually by bootstrapping the integrated likelihood and using the maximum likelihood estimator as the basic statistic. The merging of independent confidence intervals has attracted considerable attention, and the use of reduced likelihoods presents a solution to the problem. One might, for example, wish to merge independent confidence intervals for the same parameter to one interval based on all the data. When the probability basis and the confidence distribution are known for each data set, the related log-likelihoods can be added, and an integrated confidence distribution, accompanied by its pivot and likelihood, is obtained.

A related problem is that of so-called meta-analyses. If independent confidence distributions are obtained for the same parameter, the information is combined by adding the reduced log-likelihoods. A frequent problem in meta-analysis is, however, that the interest parameter might not have exactly the same value across the studies. This calls for a model that reflects this variation, possibly by including a random component. In any event, the availability of reduced likelihood functions from the various studies facilitates the meta-analysis, whether a random component is needed or not.

Studies in fields like ecology, economics, geophysics etc. often utilise complex models with many parameters. To the extent results are available for some of these parameters, it might be desirable to include this information in the study. If these previous results appear in the format of confidence distributions accompanied by explicit probability bases, their related likelihoods are perfectly suited to carry this information into the combined likelihood of the new and the previous data. If a confidence distribution is used that is not based on (previous) data, but on subjective judgement, its related likelihood can still be calculated and combined with other likelihood components, provided assumptions regarding its probability basis can be made. This subjective component of the likelihood should then, perhaps, be regarded as a penalising term rather than a likelihood term. Schweder and Ianelli (2000) used this approach to assess the status of the stock of bowhead whales subject to inuit whaling off Alaska.

Finally, being able to obtain the implied likelihood from confidence distributions, and being able to calculate confidence distributions from data summarised by a likelihood within a statistical model, a methodology parallel to and competing with Bayesian methodology emerges. This methodology is frequentist in its foundation. As the Bayesian methodology, it provides a framework for coherent learning and its inferential product is a distribution: a confidence distribution instead of a Bayesian posterior probability distribution.

10.2 Differences from the Bayesian paradigm

It is pertinent to compare our frequentist approach with the Bayesian approach to coherent learning. Most importantly, the two approaches have the same aim: to update distributional knowledge in the view of new data within the frame of a statistical model. The updated distribution could then be subject to further updating at a later stage, etc. In this sense, our approach could be termed 'frequentist Bayesian' (a term both frequentists and Bayesians probably would dislike). There are, however, substantial differences between the two approaches. Compared to the Bayesian approach, we would like to emphasise the following.

Distributions for parameters are understood as confidence distributions and not probability distributions. The concept of probability is reserved for (hypothetically) repeated sampling, and is interpreted frequentistically. To update a confidence distribution it must be related to its probability basis, to obtain the likelihood related to the confidence distribution. To update a distribution for a parameter the frequentist needs more information than the Bayesian, namely its probability basis. Furthermore, the distinction between probability and confidence is basic in the frequentist tradition.

We would like to stress as a general point the usefulness of displaying the confidence density $c(\psi)$, computed from the observed data, for any parameter ψ of interest. This would be the frequentist parallel to the Bayesian's posterior density. We emphasise that the interpretation of $c(\psi)$ should be clear and non-controversial; it is simply an effective way of summarising and communicating all confidence intervals, and does not involve any prior.

One may ask when the $c(\psi)$ curve is identical to a Bayesian's posterior. This is clearly answered by equation (3) in the presence of a pivot; the confidence density agrees exactly with the Bayesian updating when the Bayesian's prior is

$$\pi_0(\psi) = \Big|\frac{\partial \mathrm{piv}(T;\psi)}{\partial \psi}\Big| \Big/ \Big|\frac{\partial \mathrm{piv}(T;\psi)}{\partial T}\Big|. \tag{14}$$

In the pure location case the pivot is $\psi - T$, and π_0 is constant. When ψ is a scale parameter and the pivot is ψ/T, the prior becomes proportional to ψ^{-1}. These priors are precisely those found to be the canonical 'non-informative' ones in Bayesian statistics. In the correlation coefficient example of Section 8.2, the approximate pivot used there leads to $\pi_0(\rho) = 1/(1-\rho^2)$ on $(-1,1)$, agreeing with the non-informative prior found using the so-called Jeffrey's formula. Method (14) may be used also in more complicated situations, for example via abc or t-bootstrap approximations in cases where a pivot is not easily found.

It is possible for the frequentist to start at scratch, without any (unfounded) subjective prior distribution. In complex models, there might be distributional information available for some of the parameters, but not for all. The Bayesian is then stuck, or she has to construct priors. The frequentist will, however, not have principle problems in such situations. The concept of non-informativity is, in fact, simple for likelihoods. The non-informative likelihoods are simply flat. Non-informative Bayesian priors are, on the other hand, a thorny matter. In general, the frequentist approach is less dependent on subjective input to the analysis than the Bayesian approach. But if subjective input is needed, it can readily be incorporated (as a penalising term in the likelihood).

In the bowhead assessment model (Schweder and Ianelli, 2000) there were more prior distributions than there were free parameters. Without modifications of the Bayesian synthesis approach like the melding of Poole and Raftery (1998), the Bayesian gets into trouble. Due to the Borel paradox (Schweder and Hjort, 1996), the Bayesian synthesis will, in fact, be completely determined by the particular parametrisation. With more prior distributions than there are free parameters,

Poole and Raftery (1998) propose to meld the priors to a joint prior distribution of the same dimensionality as the free parameter. This melding is essentially a (geometric) averaging operation. If, however, there are independent prior distributional information on a parameter, it seems wasteful to average the priors. If, say, all the prior distributions happen to be identical, their Bayesian melding will give the same distribution. The Bayesian will thus not gain anything from k independent pieces of information, while the frequentist will end up with a less dispersed distribution; the standard deviation will, in fact, be the familiar $\sigma/k^{1/2}$.

Non-linearity, non-normality and nuisance parameters can produce bias in results, even when the model is correct. This is well known, and has been emphasised repeatedly in the frequentist literature. Such bias should, as far as possible, be corrected in the reported results. The confidence distribution aims at being unbiased: when it is exact, the related confidence intervals have exactly the nominal coverage probabilities. Bias correction has traditionally not been a concern in the Bayesian tradition. There has, however, been some recent interest in the matter. To obtain frequentist unbiasedness, the Bayesian will have to choose her prior with unbiasedness in mind. Is she then a Bayesian? Her prior distribution will then not represent prior knowledge of the parameter in case, but an understanding of the model. Our 'frequentist Bayesianism' solves this problem in principle. It takes as input (unbiased) prior confidence distributions converted to reduced likelihoods and delivers (unbiased) posterior confidence distributions.

Hald (1998) speaks of three revolutions in parametric statistical inference due to Laplace in 1774 (inverse probability, Bayesian methods with flat priors), Gauß and Laplace in 1809-1812 and Fisher in 1922. This is not the place to discuss Fisher's revolution in any detail, other than to note that it partly was a revolt against the Laplacian Bayesianism. When discussing Neyman's 1934 paper on survey sampling, Fisher stated, "All realized that problems of mathematical logic underlay all inference from observational material. They were widely conscious, too, that more than 150 years of dispute between the pros and the cons of inverse probability had left the subject only more befogged by doubt and frustration." To come around the problems associated with prior distributions, Fisher proposed the fiducial distribution as a replacement for the Bayesian posterior. Efron (1998) emphasises the importance of the fiducial distribution, which he prefers reformulated to the confidence distribution discussed in the present paper. The fiducial argument is not without problems (see e.g. Brillinger, 1962, Wilkinson, 1977, Welsh 1996) and has often been regarded as "Fisher's biggest blunder" (see Efron, 1998). By converting to the confidence formulation, as Neyman did in 1941 but which Fisher resisted, Efron holds that the method can be applied to a wider class of problems and that it might hold a key to "our profession's 250-year search for a dependable objective Bayes theory". We agree, and we hope with Efron (1998) and also with Fraser when discussing Efron (1998), that fiducial or confidence distributions will receive renewed interest. By introducing the reduced likelihood associated with a confidence distribution, and by pointing out the importance of the underlying (approximate) pivot for future parametric bootstrapping, a form of objective Bayes methodology has been sketched. Our form of 'frequentist Bayesianism' does not involve Bayes' formula, although we have nothing against using Bayesian techniques to produce confidence distributions with correct frequentist properties. But it seeks to deliver digested statistical information in the format of distributions, and it provides a method for rational updating of such statistical information.

References

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

[2] Barndorff-Nielsen, O.E. and Wood, T.A. (1998). On large deviations and choice of ancillary for p^* and r^*. *Bernoulli* **4**, 35–63.

[3] Berger, J.O., Liseo, B. and Wolpert, R.L. (1999). Integrated likelihood methods for eliminating nuisance parameters. Technical report, Institute of Statistics and Decision Sciences, Duke University.

[4] Brillinger, D.R. (1962). Examples bearing on the definition of fiducial probability with a bibliography. *Annals of Mathematical Statistics* **33**, 1349–1355.

[5] Cook, T.D. and Campbell, D.T. (1979). *Quasi-Experimentation.* Houghton Mifflin Company, Boston.

[6] Darroch, J.N. (1958). The multiple-recapture census. I: Estimation of a closed population. *Biometrika* **45**, 343–359.

[7] DiCiccio, T.J. and Efron, B. (1996). Bootstrap confidence intervals (with discussion). *Statistical Science* **11**, 189–228.

[8] Davison, A.C. and Hinkley, D.V. (1997). *Bootstrap Methods and their Application.* Cambridge University Press, Cambridge.

[9] Edwards, A.W.F. (1992). *Likelihood* (expanded edition). John Hopkins University Press, Baltimore.

[10] Efron, B. (1987). Better bootstrap confidence intervals (with discussion). *Journal of the American Statistical Association* **82**, 171–200.

[11] Efron, B. (1993). Bayes and likelihood calculations from confidence intervals. *Biometrika* **80**, 3–26.

[12] Efron, B. (1998). R.A. Fisher in the 21st century (with discussion). *Statistical Science* **13**, 95–122.

[13] Efron, B. and Tibshirani, R.J. (1993). *An Introduction to the Bootstrap.* Chapman and Hall, London.

[14] Fisher, R.A. (1922). On the mathematical foundation of theoretical statistics. *Philosophical Transactions of the Royal Society of London* **A 222**, 309–368.

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

[16] Fisher, R.A. (1973). *Statistical Methods and Scientific Inference* (3rd ed.). Hafner Press.

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

[18] Fraser, D.A.S. (1996). Some remarks on pivotal models and the fiducial argument in relation to structural models. *International Statistical Review* **64**, 231–236.

[19] Fraser, D.A.S. (1998). Contribution to the discussion of Efron's paper. *Statistical Science* **13**.

[20] Hald, A. (1998). *A History of Mathematical Statistics from 1750 to 1930.* Wiley, New York.

[21] Hald, A. (1999). On the history of maximum likelihood in relation to inverse probability and least squares. *Statistical Science* **14**, 214–222.

[22] Lehmann, E.L. (1959). *Testing Statistical Hypotheses.* Wiley, New York.

[23] Lehmann, E.L. (1983). *Theory of Point Estimation.* Wiley, New York.

[24] Lehmann, E.L. (1993). The Fisher, Neyman–Pearson theories of testing hypotheses: One theory or two? *Journal of the American Statistical Association* **88**, 1242–1249.

[25] Neyman, J. (1941). Fiducial argument and the theory of confidence intervals. *Biometrika* **32**, 128–150.

[26] Poole, D. and Raftery, A.E. (1998). Inference in deterministic simulation models: The Bayesian melding approach. Technical Report no. 346, Department of Statistics, University of Washington.

[27] Raftery, A.E., Givens, G.H. and Zeh, J.E. (1995). Inference from a deterministic population dynamics model for bowhead whales (with discussion). *Journal of the American Statistical Association* **90**, 402–430.

[28] Royall, R.M. (1997). *Statistical Evidence. A Likelihood Paradigm.* Chapman and Hall, London.

[29] Schweder, T. (1988). A significance version of the basic Neyman–Pearson theory for scientific hypothesis testing (with discussion). *Scandinavian Journal of Statistics* **15**, 225–242.

[30] Schweder, T. and Hjort, N.L. (1996). Bayesian synthesis or likelihood synthesis — what does Borel's paradox say? *Reports of the International Whaling Commission* **46**, 475–479.

[31] Schweder, T. and Hjort, N.L. (1999). Frequentist analogues of priors and posteriors. To appear.

[32] Schweder, T. and Ianelli, J.N. (1998). Bowhead assessment by likelihood synthesis: methods and difficulties. Paper IWC/SC/50/AS2.

[33] Schweder, T. (2000). Abundance estimation from photo-identification data: confidence distributions and reduced likelihood for bowhead whales off Alaska. Paper IWC2000/SC/52/AS14 presented to International Whaling Commission (unpublished).

[34] Schweder, T. and Ianelli J.N. (2000). Assessing the Bering-Chukchi-Beaufort Seas stock of bowhead whales from survey data, age-readings and photo-identifications using frequentist methods. Paper IWC2000/SC/52/AS13 presented to International Whaling Commission (unpublished).

[35] da Silva, C.Q., Zeh, J., Madigan, D., Lake, J., Rugh, D., Baraff, L., Koski, W. and Miller, G. (2000). Capture-recapture estimation of bowhead whale population size using photo-identification data. *Journal of Cetacean Reserve Management* **2**, 45–61.

[36] Welsh, A.H. (1996). *Aspects of Statistical Inference.* Wiley, New York.

[37] Wilkinson, G.N. (1977). On resolving the controversy in statistical inference (with discussion). *Journal of the Royal Statistical Society* **B 39**, 119–171.