Statistical Methods & Applications
https://doi.org/10.1007/s10260-020-00520-y

REVIEW PAPER

Can Bayesian, confidence distribution and frequentist inference agree?

Erlis Ruli[1] · Laura Ventura[1]

Accepted: 1 March 2020

Abstract
We discuss and characterise connections between frequentist, confidence distribution and objective Bayesian inference, when considering higher-order asymptotics, matching priors, and confidence distributions based on pivotal quantities. The focus is on testing precise or sharp null hypotheses on a scalar parameter of interest. Moreover, we illustrate that the application of these procedures requires little additional effort compared to the application of standard first-order theory. In this respect, using the R software, we indicate how to perform in practice the computation with three examples in the context of data from inter-laboratory studies, of the stress–strength reliability, and of a growth curve from dose–response data.

Keywords Credible interval · First-order theory · Full Bayesian significance test · Higher-order asymptotics · Likelihood inference · Marginal posterior distribution · Matching prior · Pivotal quantity · Precise null hypothesis · *p* value · Tail area probability

1 Introduction

In recent years, the interplay between Bayesian and frequentist inference has lead to several connections. Some instances are, among others, the use of pseudo-likelihoods for Bayesian inference; the derivation of default priors, such as matching priors; the development of higher-order asymptotics for likelihood

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10260-020-00520-y) contains supplementary material, which is available to authorized users.

✉ Erlis Ruli
ruli@stat.unipd.it

Laura Ventura
ventura@stat.unipd.it

[1] Department of Statistical Sciences, University of Padova, Padua, Italy

Published online: 07 April 2020

methods, posterior distributions and related quantities; the definition of fiducial inference and confidence distributions, without the need for prior information; see, among others, Fraser and Reid (2002), Reid (2003), Xie and Singh (2013), Ventura and Reid (2014), Nadarajaha et al. (2015), Ventura and Racugno (2016), Hjort and Schweder (2018) and references therein.

To first-order, Bayesian, frequentist and confidence distribution (CD) inference may be based on familiar large sample theory for the maximum likelihood estimator and the Wald statistic. This theory involves simple approximations, that may be justified when the sample size is large, and in this case the three modes of inference agree. However, it is well-known that first-order approximations can be inaccurate in many applications, in particular when the sample is small, and that standard first-order theory can be readily improved. The purpose of this review article is to investigate and characterise higher-order relationships between Bayesian, frequentist and CD inference based on: a posterior distribution, when a suitable objective prior is used; modern likelihood methods; a CD based on a pivotal quantity. The focus is on testing precise or sharp null hypotheses on a scalar parameter of interest and it is of interest to relate frequentist and Bayesian significance indices. In particular, the procedures involved in this paper are:

1. higher-order likelihood inference based on the profile modified likelihood root (see, e.g., Brazzale et al. 2007; Pierce and Bellio 2017);
2. higher-order approximations of the measure of evidence for the full Bayesian significance test (see, e.g., Madruga et al. 2003), when using matching priors (Cabras et al. 2015):
3. CD inference based on higher-order pivots (see, e.g, Xie and Singh 2013; Nadarajaha et al. 2015; Hjort and Schweder 2018; Fraser et al. 2018).

From a practical point of view, it is shown how these procedures can be easily applied in practical problems using the `likelihoodAsy` package (Bellio and Pierce 2018) of the statistical software `R`.

The rest of paper is organised as follows. Section 2 illustrates first-order agreement between frequentist, Bayesian and CD significance indices, while higher-order connections are discussed in Sect. 3. Section 4 illustrates how to perform in practice the computations with three examples in the context of: data from inter-laboratory studies, the stress–strength reliability, and a growth curve from dose–response data. Concluding remarks are given in Sect. 5.

2 First-order agreement between frequentist, Bayesian and CD inference

Consider a random sample $y = (y_1, \ldots, y_n)$ of size n from a parametric model with probability density function $f(y; \theta)$, indexed by a d-dimensional parameter θ. Write $\theta = (\psi, \lambda)$, where ψ is a scalar parameter for which inference is required and λ represents the remaining $(d-1)$ nuisance parameters. We wish to test the precise (or sharp) null hypothesis

$$H_0 : \psi = \psi_0 \quad \text{against} \quad H_1 : \psi \neq \psi_0. \tag{1}$$

Possible examples occur, for instance, when the parameter of interest is the stress–strength reliability and the null hypothesis is $H_0 : \psi = 0.5$, or in regression problems, when ψ is a regression coefficient and the null hypothesis is $H_0 : \psi = 0$.

Likelihood inference Let $\ell(\theta) = \log L(\theta) = \log f(y;\theta)$ be the log-likelihood function, maximized by the maximum likelihood estimator (MLE) $\hat{\theta} = (\hat{\psi}, \hat{\lambda})$, assumed to be finite. Moreover, let $\hat{\theta}_\psi = (\psi, \hat{\lambda}_\psi)$ be the MLE when ψ is held fixed. Well-known bases to test (1) are the profile Wald pivot

$$w_p(\psi) = \frac{\hat{\psi} - \psi}{\sqrt{j_p(\hat{\psi})^{-1}}}, \tag{2}$$

where $j_p(\psi) = -\partial^2 \ell(\hat{\theta}_\psi)/\partial\psi^2$ is the profile observed information, and the profile likelihood root

$$r_p(\psi) = \text{sign}(\hat{\psi} - \psi)\sqrt{2(\ell(\hat{\theta}) - \ell(\hat{\theta}_\psi))}. \tag{3}$$

Both have a first-order standard normal null distribution. When testing (1), the p values based on (2) and (3) are, respectively,

$$p_w = 2(1 - \Phi(|w_p(\psi_0)|)) \quad \text{and} \quad p_r = 2(1 - \Phi(|r_p(\psi_0)|)),$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution.

The approximate normal distribution of $w_p(\psi)$ provides the most common basis for inference about ψ. It is well-known that true significance levels may differ substantially from their nominal values, because the shape of the log-likelihood about its maximum is not accommodated. Moreover, inference based on $w_p(\psi)$ is not invariant to transformations of the parameter. On the contrary, $r_p(\psi)$ takes potential asymmetry of the log-likelihood into account and is invariant to reparametrizations. However, when the sample size is relatively small, in general first-order approximations are often inaccurate, especially if the dimension of the nuisance parameter λ is high with respect to n. In these situations, further accuracy can be achieved by resorting to modern likelihood theory based on higher-order asymptotics (see, e.g., Brazzale et al. 2007; Lozada-Can and Davison 2010; Pierce and Bellio 2017).

Bayesian inference Given a prior density $\pi(\theta) = \pi(\psi, \lambda)$ for θ, Bayesian inference for ψ is based on the marginal posterior density

$$\pi_m(\psi|y) = \int \pi(\psi, \lambda|y)\, d\lambda \propto \int \pi(\psi, \lambda) L(\psi, \lambda)\, d\lambda, \tag{4}$$

provided the integral on the right hand side of (4) is finite. The usual Bayesian testing procedure is based on the well-known Bayes factor (BF), defined as the ratio of the posterior to the prior odds in favour of H_0. We decide in favour of H_0 whenever the BF, or the corresponding weight of evidence log(BF), assumes high value. However, it is well known that, when improper priors are used, the BF can be undetermined and, when the null hypothesis is precise, the BF can lead to the so-called Jeffreys–Lindley's paradox; see, e.g. Kass and Raftery (1995). Moreover, the BF is not calibrated, i.e. its finite sampling distribution is unknown and it may depend on the nuisance parameter.

Alternative to the BF, Pereira and Stern (1999, 2001) provide an intuitive measure of evidence for the full Bayesian significance test (FBST) in favour of H_0. This measure is the posterior probability related to the less probable points of the parametric space, and it favours the null hypothesis whenever it is large; see, e.g. Madruga et al. (2001, 2003) and references therein. Moreover, the FBST is based on a specific loss function, and thus the decision made under this procedure is the action that minimizes the corresponding posterior risk. Specifically, consider the marginal posterior distribution $\pi_m(\psi|y)$ for the parameter of interest ψ and consider the set $T_y(\psi) = \{\psi : \pi_m(\psi|y) \geq \pi_m(\psi_0|y)\}$. Starting from $\pi_m(\psi|y)$, the Pereira–Stern measure of evidence in favour of H_0 can be computed as (see Cabras et al. 2015, and Fig. 1)

$$p_\pi = 1 - P_\pi(\psi \in T_y(\psi)),$$

where $P_\pi(\cdot)$ denotes posterior probability. The null hypothesis H_0 is accepted whenever p_π is large enough.

A first-order approximation for p_π, but without any notion of a prior distribution involved, is simply given by (Pereira et al. 2008; Diniz et al. 2012)

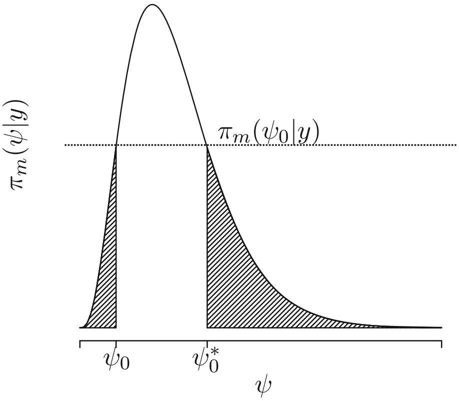


Fig. 1 The measure of evidence p_π for the precise hypothesis $H_0 : \psi = \psi_0$ (shaded area)

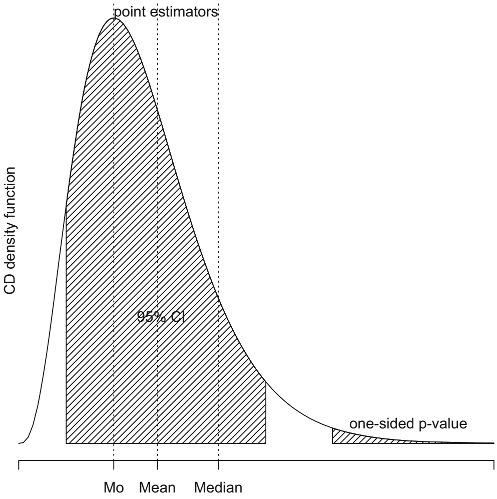
$$p_\pi \doteq 2\Phi\left(\left|\frac{\psi_0 - \hat{\psi}}{\sqrt{j_p(\hat{\psi})^{-1}}}\right|\right), \qquad (5)$$

where the symbol '$\doteq$' indicates that the approximation is accurate to $O(n^{-1/2})$. Thus, to first-order, p_π agrees with p_w, i.e. the p value based on the Wald statistic. In practice, this approximation of p_π is often inaccurate, in particular when the dimension of λ is large with respect to the sample size, because it forces the marginal posterior distribution to be symmetric. Thus, in order to have a more accurate evaluation of p_π, it may be useful to resort to higher-order approximations based on tail area approximations (see, e.g., Reid 2003; Ventura and Reid 2014; Ruli et al. 2014).

Confidence distribution. Recently, the concept of CD has received a great attention (see, e.g., Xie and Singh 2013; Nadarajaha et al. 2015; Veronese and Melilli 2015, and references therein, and the special issue by Hjort and Schweder 2018).

A CD is a distribution estimator and, conceptually, is not different from a point estimator or a confidence interval. As in Fisher's fiducial development, pivotal functions play a crucial role to the derivation of a CD. The inversion of a pivot gives probabilities on the parameter space. More precisely, let $q(y, \psi)$ be a pivot function, monotone in ψ. Then, the sample-dependent distribution function on the parameter space $H(\psi) = F(q(y, \psi))$ is a CD for ψ, where $F(\cdot)$ is the cumulative distribution function of the pivot quantity $q(y, \psi)$. The confidence density for ψ is thus

Fig. 2 Illustration of making inference using a confidence density



$$h(\psi; y) = \frac{dH(\psi)}{d\psi}.$$

The distribution $H(\psi) = F(q(y, \psi))$ is called an asymptotic CD if $F(\cdot)$ is the asymptotic cumulative distribution function of the pivot $q(y, \psi)$.

The plot in Fig. 2 gives an illustration on making inference using a CD: point estimators (mode, median and mean), 95% confidence interval and one-sided p value. Since the CD by design is unbiased, then the confidence median is a median unbiased point estimator. To test (1), the p value is given by (see, e.g., Xie and Singh 2013)

$$p_{cd} = 2\min\{H(\psi_0), 1 - H(\psi_0)\}.$$

Under H_0 it is immediate that $p_{cd} \sim U(0,1)$, since $H(\psi_0) \sim U(0,1)$ by the definition of a CD.

First-order pivot functions are $w_p(\psi)$ and $r_p(\psi)$, which can be used to derive an asymptotic CD (see, e.g., Schweder and Hjort 2016). For instance, using the Wald statistic we have $H(\psi) \doteq \Phi(w_p(\psi))$ and a first-order CD p value is (5), while using $r_p(\psi)$ we have $H(\psi) \doteq \Phi(r_p(\psi))$ and a first-order CD p value is p_r.

In summary ... to first-order the p value p_w agrees with both p_π and p_{cd}. That is, when the sample size is sufficiently high, Bayesian, confidence distribution and frequentist significance indices agree. But what happens when using a higher-order pivot function and objective Bayes?

3 Higher-order agreement between frequentist, Bayesian and CD inference

Modern likelihood inference Improved likelihood inference may be obtained through higher-order asymptotics, on which there is a large literature (see, among others, Severini 2000; Reid 2003; Brazzale et al. 2007, and references therein). One key formula is the modified profile likelihood root

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

which has a third-order standard normal null distribution. In (6), the quantity $q_p(\psi)$ is a suitably defined correction term (see, e.g., Severini 2000, Chapter 9).

The quantity $r_p^*(\psi)$ is a higher-order pivot obtained as a refinement of the likelihood root $r_p(\psi)$, which allow us to obtain accurate p values and confidence limits for ψ. The p value based on (6) is

$$p_r^* = 2(1 - \Phi(|r_p^*(\psi_0)|)). \tag{7}$$

The computation of $q_p(\psi)$ is straightforward in simple models, such as exponential families, but in general it is awkward and approximations must be derived; see Reid and Fraser (2010) for a discussion on different versions of $q_p(\psi)$. However, inference based on $r_p^*(\psi)$ can be implemented in practice for many commonly used parametric models by using the package `likelyhoodAsy` of the `R` software (Bellio and Pierce 2018), which adopts the version of $q_p(\psi)$ developed by Skovgaard (1996, 2001). In practice, the advantage of using this of the package is that it does not require the function $q_p(\psi)$ explicitly but it only requires the code for computing the log-likelihood function and for generating data from the assumed model.

Objective Bayesian inference Let us consider a so-called *strong* matching prior (Fraser and Reid 2002; Ventura et al. 2013), i.e. a prior such that a frequentist p value coincides with a Bayesian posterior survivor probability to a high degree of approximation, in the marginal posterior density (4). In this case, the tail area of the marginal posterior for ψ can be approximated to third-order as

$$\int_{\psi_0}^{\infty} \pi_m(\psi|y)\, d\psi \mathrel{\dot{\ddot{=}}} \Phi(r_p^*(\psi_0)), \tag{8}$$

where the symbol '$\dot{\ddot{=}}$' indicates that the approximation is accurate to $O(n^{-3/2})$. Following Ventura et al. (2013), the marginal posterior density can be written, to second-order, as

$$\pi_m(\psi|y) \mathrel{\ddot{\propto}} \exp\left(-\frac{1}{2} r_p^*(\psi)^2\right) \left|\frac{s_p(\psi)}{r_p(\psi)}\right|, \tag{9}$$

where $s_p(\psi) = \ell_p'(\psi)/j_p(\hat{\psi})^{1/2}$ is the profile score statistic. Using (9), an asymptotic equi-tailed credible interval for ψ can be computed as $\{\psi : |r_p^*(\psi)| \geq z_{1-\alpha/2}\}$, i.e., as a confidence interval for ψ based on (6) with approximate level $(1-\alpha)$, where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$-quantile of the standard normal distribution. Note from (8) that, the posterior median of $\pi_m(\psi|y)$ can be computed as the solution in ψ of the estimating equation $r_p^*(\psi) = 0$, and thus it coincides with the frequentist estimator defined as the zero-level confidence interval based on $r_p^*(\psi)$. Such an estimator has been shown to be a refinement of the MLE $\hat{\psi}$ (Giummolé and Ventura 2002).

Using the tail area approximation (8), a third-order approximation of the measure of evidence p_π is (Cabras et al. 2015)

$$p_\pi^* = 1 - \Phi(r_p^*(\psi_0)) + \Phi(r_p^*(\psi_0^*)), \tag{10}$$

with ψ_0^* the value of the parameter such that $\pi_m(\psi_0^*|y) = \pi_m(\psi_0|y)$ (see Fig. 1). The measure p_π^* is calibrated to second order with respect to the U(0,1) distribution. Note that $\Phi(r_p^*(\psi_0)) - \Phi(r_p^*(\psi_0^*)) \mathrel{\dot{\ddot{=}}} \int_{\psi_0^*}^{\psi_0} \pi_m(\psi|y)\, d\psi$ in (10) gives the posterior probability of the HPD (High Posterior Density) credible interval (ψ_0, ψ_0^*). Therefore the higher-order measure of evidence (10) differs from (7), since the former is a density-based measure while the latter is a quantile-based quantity. However, if

$\pi_m(\psi|y)$ is symmetric, (10) reduces to $2(1-\Phi(|r_p^*(\psi_0)|))$, and thus it coincides with p_r^*.

Asymptotically third-order accurate CD Starting from the profile modified likelihood root (6), it is easy to derive an asymptotically third-order accurate CD, i.e. with error of order $O(n^{-3/2})$. Indeed, using $r_p^*(\psi)$, the CD can be expressed as

$$H(\psi) \,\dddot{=}\, \Phi(r_p^*(\psi)). \tag{11}$$

Thus, the corresponding $h(\psi;y)$ coincides with (9), that is (9) is both a confidence density and a posterior density for ψ. In view of this, both the posterior and the confidence medians are the solution of $r_p^*(\psi)=0$ and coincide with the frequentist estimator defined as the zero-level confidence interval based on $r_p^*(\psi)$. Moreover, the $(1-\alpha)$ equi-tailed credible interval $\{\psi : |r_p^*(\psi)| \leq z_{1-\alpha/2}\}$ coincides with the $(1-\alpha)$ confidence interval for ψ, which also coincides with the higher-order likelihood-based confidence interval for ψ. Finally, to test (1), the CD p value based on (11) is given by

$$p_{cd}^* = 2(1-\Phi(|r_p^*(\psi_0)|)),$$

which coincides with p_r^*. If (9) is asymmetric, p_{cd}^* is not equal to p_π^*.

In summary ... this shows that when using strong matching priors and higher-order asymptotics, there is an agreement between Bayesian, CD and frequentist point and interval estimation. However, when focus is on significance indices, even though the CD density $h(\psi;y)$ coincides with the marginal posterior (9), the inferential meaning of the two distributions remains different. Indeed, a p value is a tail evaluation of the sampling distribution under H_0, while the measure of evidence for the FBST is a tail evaluation of the posterior distribution conditional on the observed sample. Furthermore, while the tail for the p value evaluation starts at the observed value of the test statistic, the tail for p_π starts at the sharp null hypothesis.

4 Three examples

We discuss three examples in the context of data from inter-laboratory studies, stress–strength reliability, and growth curves from dose–response data. From a practical point of view, it is shown how all these procedures can be easily applied in practical problems using the `likelihoodAsy` package of the statistical software `R`, which does not require to derive explicitly the quantity $q_p(\psi)$ involved in the modified profile likelihood root (Bellio and Pierce 2018).

The focus is on highlighting inferential agreement and disagreement reached when considering higher-order asymptotics, matching priors, and confidence distributions based on pivotal quantities. The `R` code and the data used for the examples can be found in the supplementary material.

4.1 Heteroscedastic one-way random effects model

The analysis of data from inter-laboratory studies has received a great deal of attention over the past years, and it deals with the one-way random effects model with heteroscedastic error variance (see, e.g., Sharma and Mathew 2011, and references therein). The basic setting is as follow. There are m laboratories, with n_j observations at the jth laboratory, for $j = 1, \ldots, m$. The model is

$$y_{ij} = \mu + b_j + \varepsilon_{ij}, \tag{12}$$

where y_{ij} denotes the ith observation at the jth laboratory, and b_j and ε_{ij} are independent random variables with distribution $b_j \sim N(0, \sigma^2)$ and $\varepsilon_{ij} \sim N(0, \sigma_j^2)$, respectively. Typically, the parameter of interest is the consensus mean μ, which, in case of (12), is also the mean of y_{ij}, $i = 1, \ldots, n_j$ and $j = 1, \ldots, m$. The remaining $(m+1)$ parameters of the model, i.e., within-laboratory variances $(\sigma_1^2, \ldots, \sigma_m^2)$ and the between laboratory variability σ^2, are nuisance parameters.

The log-likelihood function for μ and $\lambda = (\sigma^2, \sigma_1^2, \ldots, \sigma_m^2)$ from model (12) is

$$\ell(\mu, \lambda) = -\frac{1}{2}\sum_{j=1}^{m}\left((n_j - 1)\log\sigma_j^2 - \log\rho_j + \rho_j(\bar{y}_j - \mu)^2 + \frac{(n_j - 1)s_j^2}{\sigma_j^2}\right), \tag{13}$$

with $\rho_j = 1/(\sigma^2 + \sigma_j^2/n_j)$, $\bar{y}_j = \sum_{i=1}^{n_j} y_{ij}/n_j$ and $s_j^2 = \sum_{i=1}^{n_j}(y_{ij} - \bar{y}_j)^2/(n_j - 1)$ for $j = 1, \ldots, m$.

As a numerical example, consider data on the Ki-67 protein on 245 adrenocortical tumors, coming from an inter-laboratory study (Duregon et al. 2013). It is of interest to carry out inference on the mean of the Ki-67 level (on logarithmic scale), i.e. the parameter of interest is the consensus mean μ of the Ki-67 protein (on logarithmic scale). The higher-order approximation (9) to the marginal posterior

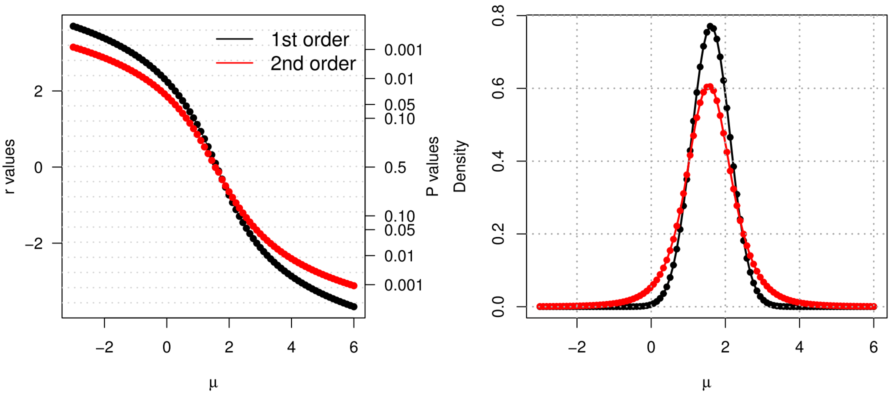


Fig. 3 Inference for the consensus mean μ from the Ki-67 protein data. Left: profile (black) and modified profile (red) likelihood roots as functions of μ; right: higher-order (red) and normal (black) approximations of the marginal posterior distribution (9) of μ. The latters are also confidence distributions for μ (color figure online)

density of μ, which is also the confidence distribution for μ, along with the first-order normal approximation are illustrated in Fig. 3 (right plot).

The higher-order interval $(-0.125, 3.26)$ is both an 0.95 equi-tailed credible interval derived from (9) and a frequentist interval based on $r_p^*(\mu)$; the first-order 0.95 confidence intervals based on the $r_p(\psi)$ and $w_p(\psi)$ pivots are (0.31, 2.85) and (0.60, 2.63), respectively.

It is of interest to test the hypotheses $H_0 : \mu = 0$ versus $H_1 : \mu \neq 0$. The p value p_r^* based on (6) and the p value p_{cd}^* based on the CD (11) are equal to 0.062. In this case, the measure of evidence based on FBST also coincides, up to four decimal places, to $p_r^* = 0.062$, in view of the symmetry of the marginal posterior of μ. The first-order p values p_r and p_w are 0.024 and 0.002, respectively. Hence, first-order results suggest that H_0 must be rejected, whereas third-order accurate significance indices do not.

4.2 Inference for the stress–strength reliability

Assume that X and Y are independent random variables with distributions $F_X(x; \theta_X)$ and $F_Y(y; \theta_Y)$, respectively. A stress–strength model is concerned with the statistical problem of evaluating the reliability $P(X < Y)$ of a component - or a material or a system - subject to a certain environmental stress. Inference about $P(X < Y)$ has revealed an attractive problem in statistical quality control, engineering statistics, medical statistics and biostatistics, among others. For instance, in a reliability study, X is the stress applied to the system, Y is the strength of a system and $P(X < Y)$ measures the chance that the system does not fail. Moreover, in a clinical study, X may be the response of a control group, Y the response of a treatment group and the reliability parameter $P(X < Y)$ measures the effectiveness of the treatment (see Kotz et al. 2003).

By the definition of reliability, $P(X < Y)$ can be evaluated as a function of the parameter $\theta = (\theta_X, \theta_Y)$, through the relation

$$\psi = \psi(\theta) = P(X < Y) = \int F_X(t; \theta_X)\, dF_Y(t; \theta_Y).$$

Theoretical expressions for ψ are available under several distributional assumptions both for the stress and the strength (see Kotz et al. 2003). For instance, let us assume that X and Y are independent normal random variables; that is $X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$. In this setting, $\theta = (\mu_X, \sigma_X^2, \mu_Y, \sigma_Y^2)$ and the reliability parameter is

$$\psi = \psi(\theta) = \Phi\left(\frac{\mu_Y - \mu_x}{\sqrt{\sigma_X^2 + \sigma_y^2}}\right).$$

This expression can be extended to include linear model formulations by assuming that μ_X and μ_Y depend on some covariates (see for instance Guttman et al. 1988). In this case, the log-likelihood function for θ can be written as

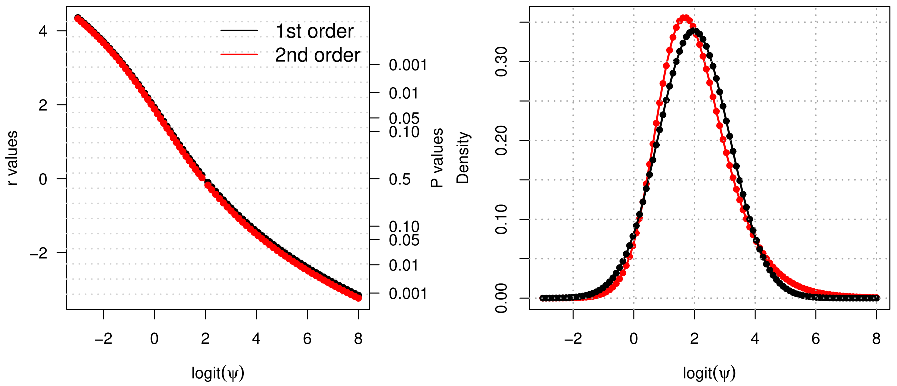


Fig. 4 Inference for the reliability parameter ψ (in logit scale) from the rocket motor data (Guttman et al. 1988) with $z = 500$. Left: profile (black) and modified profile (red) likelihood roots as functions of ψ; right: higher-order (red) and normal (black) approximations of the marginal posterior distribution (9) of logit(ψ). The latters are also confidence distributions for logit(ψ) (color figure online)

$$\ell(\theta) = \frac{1}{2}\left\{ -n_X \log \sigma_X^2 - n_Y \log \sigma_Y^2 - \frac{1}{\sigma_X^2}\sum_{i=1}^{n_X}(x_i - \mu_{Xi})^2 - \frac{1}{\sigma_Y^2}\sum_{i=1}^{n_Y}(y_i - \mu_{Yi})^2 \right\}. \tag{14}$$

As a numerical example consider the rocket motor dataset (Guttman et al. 1988; Kotz et al. 2003, pp. 205–207), which provides $n_X = 51$ stress measures of operating pressure (x) and ambient temperature (z), and $n_Y = 17$ strength data which are measures of chamber burst (y). We make the reasonable assumption (Guttman et al. 1988) that the stress depends on the ambient temperature according to a linear model with two parameters. The aim is to make inference on the reliability of a rocket motor case, at a given ambient temperature value z. Following Guttman et al. (1988), we consider $z = 500$. The first-order and the higher-order approximations to the marginal posterior density of logit(ψ), which are also the first- and higher-order CD for ψ, are illustrated in Fig. 4 (right plot).

The MLE of logit(ψ) is logit($\hat{\psi}$) = 1.982 and the 0.95 higher-order confidence interval, the 0.95 credible interval and the 0.95 CD interval for logit(ψ) coincide and are all equal to $(-0.076, 4.86)$. The first-order 0.95 confidence intervals based on the $r_p(\psi)$ and $w_p(\psi)$ pivots are $(-0.031, 5.03)$ and $(-0.324, 4.287)$, respectively.

For the stress–strength reliability it is often of interest to test the null hypothesis H_0 : logit(ψ) = 0 against H_1 : logit(ψ) $\neq$ 0. In this case, the accurate likelihood p value p_r^* based on (6) coincides with the CD p value p_{cd}^* based on (11) and is equal to 0.06. The measure of evidence p_π^* for H_0 is 0.093. The first-order p values p_r and p_w are 0.054 and 0.092 respectively. Hence, from a practical perspective, the conclusion about H_0 is roughly the same, that is H_0 cannot be rejected.

4.3 Dose–response regression model

A dose–response model is a nonlinear regression model in which the response y_i is related to an explanatory variable x_i as $y_i = \mu(x_i;\beta) + \sigma_i \varepsilon_i$, $i = 1, \ldots, n$, where $\mu(x_i;\beta)$ is the mean function, x_i is fixed, the regression parameters are the $p \times 1$ vector β, and the ε_i are independent and generated from a known continuous density function. The standard normal density is widely used, especially for dose–response curves in bioassays. The variance can be modelled as $\sigma_i^2 = \sigma^2 V(x_i;\beta,g)$, where σ^2 and the $q \times 1$ vector g are variance parameters and $V(\cdot)$ is a given positive function.

As a numerical example consider data from a bioassay study taken to estimate the root length of perennial ryegrass (Inderjit Streibig and Olofsdotter 2002). The `ryegrass` dataset reports a covariate x, the dose of ferulic acid (in mM), and the response y, the root length of perennial ryegrass (in cm) for $n = 24$ plants. The concentration–response relationship is modelled by means of the four-parameter logistic function

$$\mu(x;\beta) = \beta_1 + \frac{\beta_2 - \beta_1}{1 + \exp\{\beta_3(\log x - \log \beta_4)\}}, \quad x > 0.$$

The parameter $\beta_4 > 0$ is also denoted ED50 and it is the dose producing a response half-way between the upper limit, β_2, and lower limit, β_1. The parameter β_3 denotes the relative slope around β_4. To allow for more flexibility, we permit the variance of each observation to vary according to the following power function $\sigma_i^2 = \sigma^2 \mu(x_i;\beta)^g$, where g is a scalar real parameter and $\sigma^2 > 0$. In this example we are concerned with inference on g and the aim is to asses the homoscedasticity hypothesis $H_0 : g = 0$ versus $H_1 : g \neq 0$.

The first-order and the higher-order approximations to the marginal posterior density of g, which are also the first- and higher-order CD for g, are illustrated in Fig. 5 (right plot). For this model we found that $\hat{g} = 0.836$ and the interval

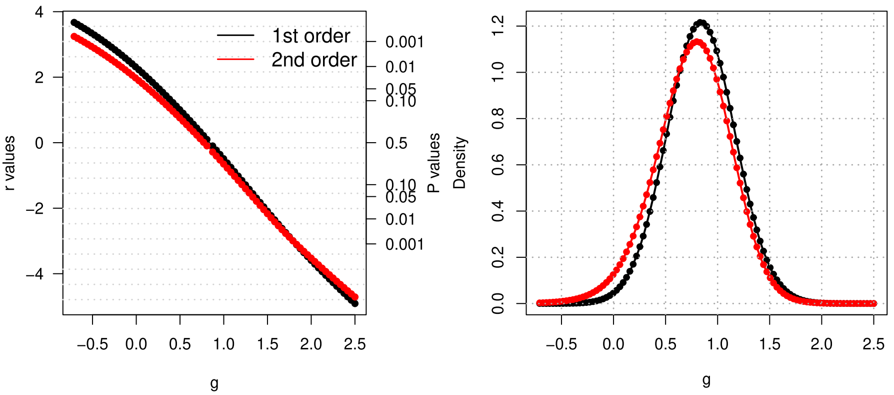


Fig. 5 Inference for the power parameter of the variance function from the `ryegrass` data (Inderjit Streibig and Olofsdotter 2002). Left: profile (black) and modified profile (red) likelihood roots as functions of g; right: higher-order (red) and normal (black) approximation of the marginal posterior distribution (9) of g. The latters are also confidence distributions for g (color figure online)

$(-0.001, 1.43)$ is a higher-order 0.95 equi-tailed credible interval, a 0.95 CD interval and a higher-order 0.95 confidence interval. The first-order 0.95 confidence intervals based on the $r_p(\psi)$ and $w_p(\psi)$ pivots are (0.14, 1.46) and (0.2, 1.48), respectively.

The higher-order frequentist p value p_r^* and the higher-order CD p value p_{cd}^* for the null hypothesis $H_0 : g = 0$ are both equal to 0.0502, whereas the measure of evidence is equal to $p_\pi^* = 0.042$. The first-order p values p_r and p_w are 0.022 and 0.011 respectively. Even though p_r^* and p_π^* do not exactly coincide, they both indicate a similar evidence for H_0, while with the first-order significance indices inferential conclusions would be altered. Thus, also in this example, higher-order approximation can produce appreciably better inferences when sample sizes are small or moderate.

5 Final remarks

To first-order, Bayesian, frequentist and CD inference - point and interval estimation and significance indices - agree. This theory involves approximations that may be justified when the sample size is large and its accuracy is often questionable. Standard first-order theory can be readily improved in Bayesian, frequentist and CD inference, and higher-order approximations can produce appreciably better inferences.

When using objective Bayesian procedures based on strong matching priors and higher-order asymptotics, there is an agreement between Bayesian, CD and frequentist point and interval estimation, but not - in general - in significance measures. This shows that, even when using objective Bayesian procedures based on strong matching priors, again the eternal half-disagreement between Bayesians and frequentists may arise. There is a strong connection between Bayesian and frequentist significance indices, but while a p value is a tail evaluation of the sampling distribution under the null hypothesis, the measure of evidence for the FBST is a tail evaluation of the posterior distribution conditional on the observed sample.

In this paper we also outline how approximate computational tools have a role to play in the modern era of frequentist and Bayesian statistics. From a practical point of view, modern likelihood inference, CD based on higher-order pivots and approximate Bayesian computations are available at little additional computational cost over simple first-order approximations. Indeed, all the computations involved in this paper are performed by using the `likelihoodAsy` package of the statistical software `R`, with only the log-likelihood function (and its derivative) as the user-provided input.

Acknowledgements This research work was partially supported by University of Padova (`BIRD197903`) and by PRIN 2015 (Grant `2015EASZFS_003`). The authors wish to thank Nancy Reid for constructive comments on a draft version of the present paper. We are also grateful to the Editor and the referees for their valuable suggestions.

References

Bellio R, Pierce D (2018) likelihoodAsy: functions for likelihood asymptotics. R package version 0.50. https://CRAN.R-project.org/package=likelihoodAsy. Accessed 4 Apr 2020

Brazzale AR, Davison AC, Reid N (2007) Applied asymptotics. Case-studies in small sample statistics. Cambridge University Press, Cambridge

Cabras S, Racugno W, Ventura L (2015) Higher-order asymptotic computation of Bayesian significance tests for precise null hypotheses in the presence of nuisance parameters. J Stat Comput Simul 85:2989–3001

Diniz M, Pereira CA, Polpo A, Stern JM, Wechsler S (2012) Relationship between Bayesian and frequentist significance indices. Int J Uncertain Quantif 2:161–172

Duregon E, Fassina A, Volante M, Nesi G, Santi R, Gatti G, Cappellesso R, Dalino P, Ventura L, Gambacorta M, Dei Tos AP, Loli P, Mannelli M, Mantero F, Berruti A, Terzolo M, Papotti M (2013) The reticulin algorithm for adrenocortical tumors diagnosis: a multicentric validation study on 245 unpublished cases. Am J Surg Pathol 37:1433–1440

Fraser DAS, Reid N (2002) Strong matching of frequentist and Bayesian parametric inference. J Stat Plan Inference 103:263–285

Fraser DAS, Reid N, Lin W (2018) When should modes of inference disagree? Some simple but challenging examples. Ann Appl Stat 2:750–770

Giummolé F, Ventura L (2002) Practical point estimation from higher-order pivots. J Stat Comput Simul 72:419–430

Guttman I, Johnson RA, Bhattacharyya GK, Reiser B (1988) Confidence limits for stress-strength models with explanatory variables. Technometrics 30:161–168

Hjort NL, Schweder T (2018) Confidence distributions and related themes. J Stat Plan Inference 195:1–13

Inderjit Streibig JC, Olofsdotter M (2002) Joint action of phenolic acid mixtures and its significance in allelopathy research. Physiol Plant 114:422–428

Kass R, Raftery A (1995) Bayes factors. J Am Stat Assoc 90:773–795

Kotz S, Lumelskii Y, Pensky M (2003) The stress-strength model and its generalizations: theory and applications. World Scientific, Singapore

Lozada-Can C, Davison AC (2010) Three examples of accurate likelihood inference. Am Stat 64:131–139

Madruga M, Esteves L, Wechslerz S (2001) On the Bayesianity of pereira-stern tests. Test 10:291–299

Madruga M, Pereira C, Stern J (2003) Bayesian evidence test for precise hypotheses. J Stat Plan Inference 117:185–198

Nadarajaha S, Bityukov S, Krasnikov N (2015) Confidence distributions: a review. Stat Methodol 22:23–46

Pereira C, Stern J (1999) Evidence and credibility: full Bayesian significance test for precise hypotheses. Entropy 1:99–110

Pereira C, Stern J (2001) Model selection: full Bayesian approach. Environmetrics 12:559–568

Pereira C, Stern J, Wechsler S (2008) Can a significance test be genuinely Bayesian? Bayesian Anal 3:79–100

Pierce DA, Bellio R (2017) Modern likelihood-frequentist inference. Int Stat Rev 85:519–541

Reid N (2003) The 2000 Wald memorial lectures: asymptotics and the theory of inference. Ann Stat 31:1695–1731

Reid N, Fraser DAS (2010) Mean loglikelihood and higher-order approximations. Biometrika 97:159–170

Ruli E, Sartori N, Ventura L (2014) Marginal posterior simulation via higher-order tail area approximations. Bayesian Anal 9:129–146

Schweder T, Hjort NL (2016) Confidence, likelihood, probability: statistical inference with confidence distributions. Cambridge University Press, Cambridge

Severini TA (2000) Likelihood methods in statistics. Oxford University Press, Oxford

Sharma G, Mathew T (2011) Higher-order inference for the consensus mean in inter-laboratory studies. Biom J 53:128–136

Skovgaard IM (1996) An explicit large-deviation approximation to one-parameter tests. Bernoulli 2:145–165

Skovgaard IM (2001) Likelihood asymptotics. Scand J Stat 28:3–32

Ventura L, Reid N (2014) Approximate Bayesian computation with modified loglikelihood ratios. Metron 7:231–245

Ventura L, Racugno W (2016) Pseudo-likelihoods for Bayesian inference. In: di Battista T, Moreno E, Racugno W (eds) Topics on methodological and applied statistical inference. Studies in theoretical and applied statistics. Springer, Cham, pp 205–220. https://doi.org/10.1007/978-3-319-44093-4_19

Ventura L, Sartori N, Racugno W (2013) Objective Bayesian higher-order asymptotics in models with nuisance parameters. Comput Stat Data Anal 60:90–96

Veronese P, Melilli E (2015) Fifucial and confidence distributions for real exponential families. Scand J Stat 42:471–484

Xie M, Singh K (2013) Confidence distribution, the frequentist distribution estimator of a parameter: a review. Int Stat Rev 81:3–39

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.