Econometrics and Statistics xxx (xxxx) xxx

Contents lists available at ScienceDirect

Econometrics and Statistics

journal homepage: www.elsevier.com/locate/ecosta

On approximate robust confidence distributions

Elena Bortolato, Laura Ventura*

Department of Statistical Sciences, University of Padova, Italy

ARTICLE INFO

Article history:
Received 15 September 2022
Revised 23 April 2023
Accepted 24 April 2023
Available online xxx

Keywords:
Bayesian simulation
Confidence density
Discrepancies
M-estimators
Pivotal quantity
Robustness
Non-inferiority and superiority testing

ABSTRACT

A confidence distribution is a complete tool for making frequentist inference for a parameter of interest based on an assumed parametric model. Indeed, it provides point estimates, along with confidence intervals, allows to define rejection regions for testing unilateral and bilateral hypotheses, to assign measures of evidence or levels of confidence to prespecified regions of the parameter space, and to compare the parameter of interest with other parameters from other studies. The aim is to discuss robust confidence distributions derived from unbiased *M*-estimating functions, which provide bounded-influence inference for a parameter of interest, when the assumed central model is just an approximate parametric model or in the presence of deviant values in the observed data. Paralleling likelihood-based results and extending results available for robust scoring rules, two methods are proposed for deriving robust confidence distributions: the first one uses the asymptotic theory of robust pivotal quantities and the second one is based on simulation methods. An application and a simulation study are illustrated in the context of non-inferiority and superiority testing.

1. Introduction

In recent years there has been considerable interest in frequentist inference based on confidence distributions (CDs) and confidence curves (CCs); see, among others, Xie and Singh (2013); Schweder and Hjort (2016); Hjort and Schweder (2018), and references therein. In practice, a confidence distribution analysis is much more informative than providing a $(1-\alpha)\%$ interval or a p-value for an associated hypothesis test.[1]

Let the scalar parameter of interest be ψ. With inference on ψ we shall understand statements of the type "$\psi > \psi_0$" or "$\psi_1 \leq \psi \leq \psi_2$", where ψ_0, ψ_1 and ψ_2 are given fixed values. To each statement, a CD allows us to associate how much confidence the data have in the statement. The plot in Fig. 1 gives an illustration on making inference using a confidence density for a scalar parameter of interest ψ: point estimators, $(1-\alpha)\%$ confidence interval, one-sided p-value and measure of evidence for "$\psi_1 \leq \psi \leq \psi_2$".

The standard theory for CDs evolves around the use of likelihood methods for a scalar parameter of interest ψ of a parametric model. Typically, to first-order, CD inference may be based on familiar large-sample theory for the maximum likelihood estimator (MLE), the Wald statistic and the likelihood-ratio test. However, it is well-known that likelihood-based methods are not robust when the assumed distribution is just an approximate parametric model or in the presence of deviant values in the observed data. In this case, it may be preferable to base inference on procedures that are more resistant,

* Corresponding author.
E-mail address: ventura@stat.unipd.it (L. Ventura).

[1] R code used for simulation studies is available online on https://github.com/elenabortolato/RobustConfidenceDistributions

https://doi.org/10.1016/j.ecosta.2023.04.006
2452-3062/

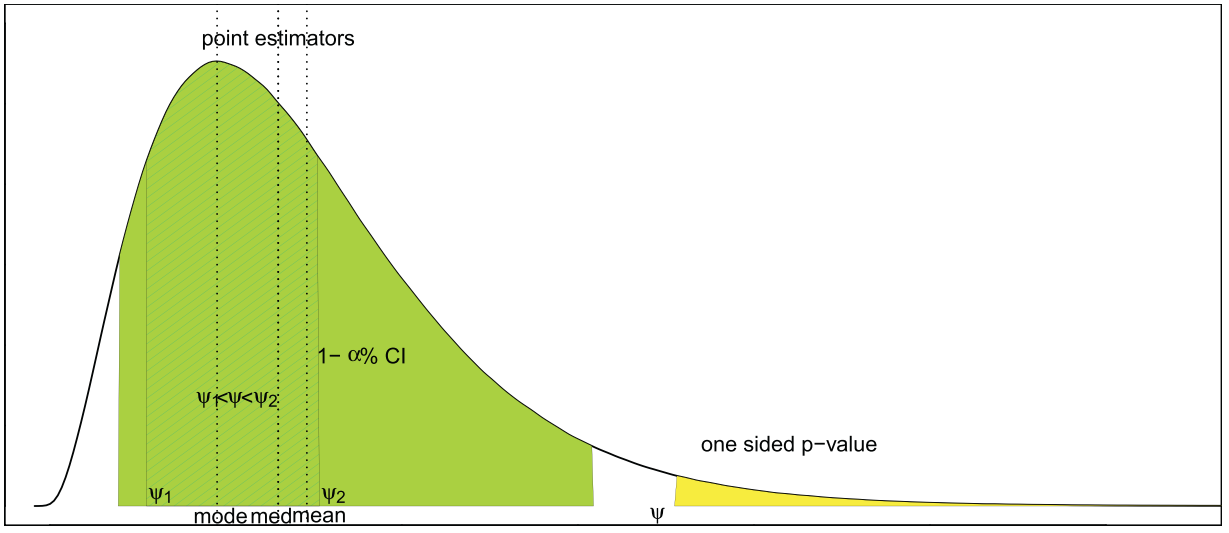

Fig. 1. Illustration of making inference on a scalar parameter of interest ψ using a confidence density: point estimators (mode, median, mean), $(1-\alpha)\%$ quantile-type confidence intervals, one-sided p-value and measure of evidence for $\psi_1 < \psi < \psi_2$.

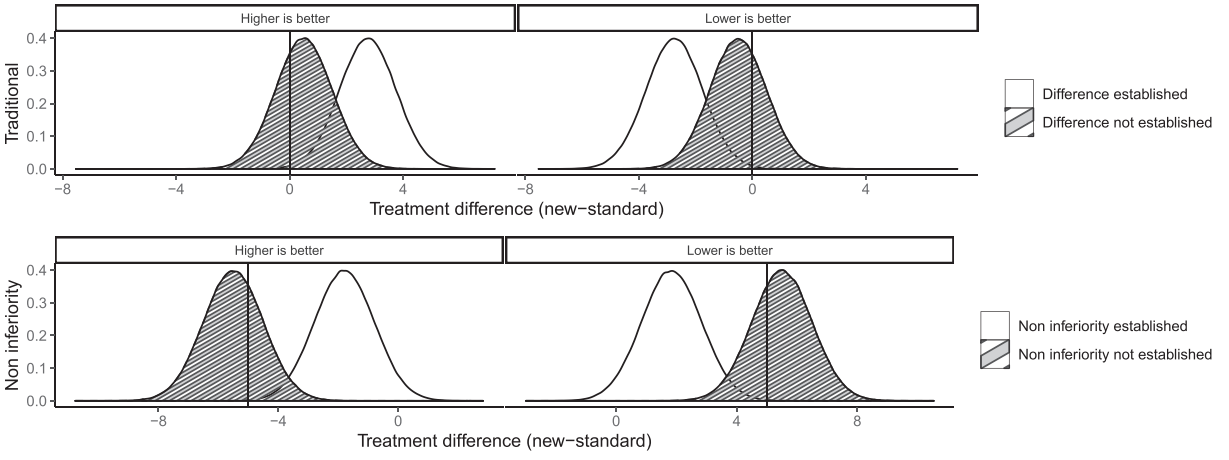

Fig. 2. Testing procedure with traditional comparative studies and non-inferiority studies using confidence densities: vertical lines represent the equivalence/non-inferiority margin ($\delta = -5$).

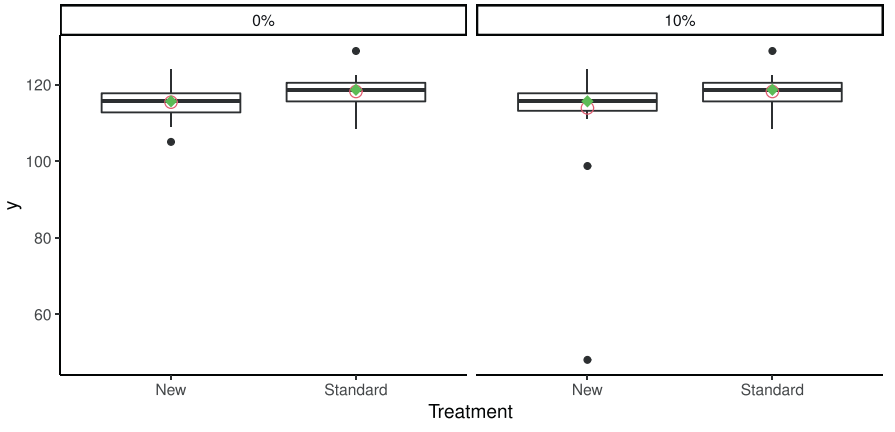

Fig. 3. Boxplots of the new treatment group and of the standard group under the two scenarios. Red dots indicate group means, green dots group medians.

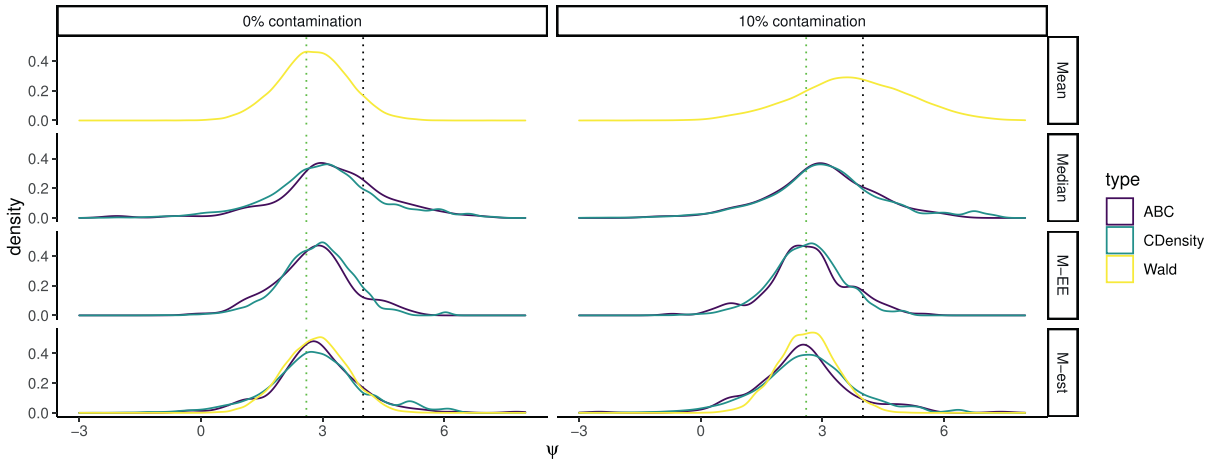

Fig. 4. Confidence densities for ψ based on 10^5 proposals, without (left column) and with contamination (right column). Results according to a different types of statistic are reported for each row, inferential techniques are represented with different colors. Black vertical dotted line is δ, green vertical dotted line is ψ_0.

that is which specifically take into account the fact that the assumed models used by the analysts are only approximate. In order to produce statistical procedures that are stable with respect to small changes in the data or to small model departures, robust statistical methods can be considered (see, e.g., Hampel et al., 1986; Huber and Ronchetti, 2009; Heritier et al., 2009; Farcomeni and Ventura, 2012; Ronchetti, 2020; Ronchetti, 2021, and references therein).

The aim of this paper is to discuss the derivation of robust confidence distributions, together with their associated densities, focusing on the robustness approach based on the influence function (see, e.g., Hampel et al., 1986; Huber and Ronchetti, 2009). Paralleling likelihood-based results, we first illustrate that asymptotic robust CDs can be obtained by using pivotal quantities derived from unbiased M-estimating functions, extending the theory of robust scoring rules discussed in Hjort and Schweder (2018); Ruli et al. (2022) to the more general setting of M-estimating functions. Secondly, we explore two simulation-based approaches to derive robust CDs. The first one relies on a frequentist reinterpretation of Approximate Bayesian Computation (ABC) techniques (see, e.g., Rubio and Johansen, 2013; Bee et al., 2017; Ruli et al., 2020; Thornton et al., 2022). The second approach leverages a Monte Carlo rejection algorithm for obtaining a one-sided p-value function. An application and a simulation study are illustrated in the context of non-inferiority (and superiority) testing (see, e.g., Rothman *et al.*, 2012), in which null hypotheses of the form $H_0 : \psi \leq \psi_0$ are of interest in order to establish if a new product is not unacceptably worse than a product already in use. Finally, in conclusion, we mention the possibility to resort to a non-parametric derivation of CDs based on integral probability semimetrics (Muller, 1997) or pseudo-metrics (Huber and Ronchetti, 2009, Chap. 2). This approach makes use of discrepancies among cumulative density functions as nonparametric pivots for performing inference on ψ, and a preliminary example and simulation study are illustrated in a misspecification context, as in Legramanti et al. (2022).

The paper is organized as follows. Section 2 reviews some background on CDs. Section 3 discusses the derivation of first-order robust CDs from unbiased M-estimating functions and of simulation-based CDs. An application and a simulation study are presented in Section 4 in the context of non-inferiority (and superiority) testing. Finally, Section 5 mentions the non-parametric derivation of CDs based on integral probability semimetrics, and Section 6 discusses some concluding remarks.

2. Background on confidence distributions

2.1. Approximate likelihood-based confidence distributions

Consider a sample $y = (y_1, \ldots, y_n)$ of size n from a random variable Y with assumed parametric model $f(y; \theta)$, indexed by a $d-$dimensional parameter θ. Let $\theta = (\psi, \lambda)$, where ψ is a scalar parameter of primary interest and λ represents the remaining $(d-1)$ nuisance parameters.

A recent definition of a confidence curve $cc(\psi) = cc(\psi, y)$ for ψ can be found, among others, in Xie and Singh (2013) and Schweder and Hjort (2016); see also references therein. Let $\theta_0 = (\psi_0, \lambda_0)$ denote the true parameter point. Then, the random variable $cc(\psi_0) = cc(\psi_0, Y)$ should have a uniform distribution on the unit interval and

$$P_{\theta_0}(cc(\psi_0, Y) \leq \alpha) = \alpha, \quad \text{for all } \alpha.$$

Thus confidence intervals can be read off, at each desired level. When α tends to zero the confidence interval tends to a single point, say ψ^*, the zero-confidence level estimator of ψ. In regular cases, $cc(\psi)$ is decreasing to the left of ψ^* and increasing to the right, in which case the confidence curve $cc(\psi)$ can be uniquely linked to a full confidence distribution

$C(\psi) = C(\psi, y)$, via

$$cc(\psi) = |1 - 2C(\psi, y)| = \begin{cases} 1 - 2C(\psi, y), & \text{if } \psi \le \psi^* \\ 2C(\psi, y) - 1, & \text{if } \psi \ge \psi^*. \end{cases}$$

With $C(\psi)$ a CD, $[C^{-1}(0.05), C^{-1}(0.95)]$ becomes an equi-tailed 90% confidence interval. Also, solving $cc(\psi) = 0.90$ yields two cut-off points for ψ, precisely those of a 90% confidence interval.

A general recipe to derive a CD is based on pivotal quantities. Suppose $q(\psi; y)$ is a function monotone increasing in ψ, with a distribution not depending on the underlying parameter, i.e. $q(\psi; y)$ is a pivot (Barndorff-Nielsen and Cox, 1994). Thus $Q(x) = P_\theta(q(\psi; Y) \le x)$ does not depend on ψ, which implies that

$$C(\psi) = Q(q(\psi; y)) \tag{1}$$

is a CD. The corresponding confidence density for ψ is

$$cd(\psi) = \frac{\partial Q(q(\psi; y))}{\partial q(\psi; y)} \frac{\partial q(\psi; y)}{\partial \psi}.$$

If the natural pivot is decreasing in ψ, then $C(\psi) = 1 - Q(q(\psi; y))$.

In the likelihood framework, there are well-working large-sample approximations for the behavior of pivotal quantities and these lead to constructions of CDs. For instance, if $\hat{\psi}$ is the MLE of ψ, then the CD is derived from the profile Wald statistic

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

with $j_p(\psi)$ profile observed information, and it coincides with the asymptotic first-order Bayesian posterior distribution for ψ (see, for instance, Ruli and Ventura, 2021).

A pivotal quantity that typically works better than (2) is the following. Let $\ell(\theta)$ be the log-likelihood function for θ, and let $\ell_p(\psi) = \ell(\psi, \hat{\lambda}_\psi)$ be the profile log-likelihood for ψ, where $\hat{\lambda}_\psi$ is the MLE for λ given ψ. The profile log-likelihood ratio test $W_p(\psi) = 2(\ell_p(\hat{\psi}) - \ell_p(\psi))$, under mild regularity conditions, has an asymptotic null χ^2_1 distribution. Hence $\Gamma_1(W_p(\psi)) \dot{\sim} U(0, 1)$, with $\Gamma_1(\cdot)$ denoting the χ^2_1 distribution function, and

$$C(\psi) \doteq \Gamma_1(W_p(\psi))$$

is a first-order asymptotic CD, which can reflect asymmetry and also likelihood multimodality in the underlying distributions, unlike the simpler Wald-type confidence distribution. Similarly, the profile likelihood root

$$r_p(\psi) = \text{sign}(\hat{\psi} - \psi)\sqrt{2(\ell_p(\hat{\psi}) - \ell_p(\psi))}$$

can be used to derive a first-order CD, since it has a first-order standard normal null distribution. Improved CD inference based on higher-order asymptotics (see, among others, Severini, 2000; Reid, 2003; Brazzale et al., 2007, and references therein) is discussed, for instance, in Schweder and Hjort (2016, Chap. 7); see also Ruli and Ventura (2021). One key formula is the modified profile likelihood root

$$r^*_p(\psi) = r_p(\psi) + \frac{1}{r_p(\psi)} \log \frac{q_p(\psi)}{r_p(\psi)}, \tag{3}$$

which has a third-order standard normal null distribution. In (3), the quantity $q_p(\psi)$ is a suitably defined correction term (see, e.g., Severini, 2000, Chapter 9). In practice, $r^*_p(\psi)$ is a higher-order pivotal quantity obtained as a refinement of the likelihood root $r_p(\psi)$, which allow us to obtain an asymptotically third-order accurate CD, i.e. with error of order $O(n^{-3/2})$.

2.2. Simulation-based confidence distributions

In the framework of CDs obtained from pivotal quantities, when the distribution of the pivot (1) is not known, approximate confidence distributions can be obtained by bootstrapping (Efron, 1979). Indeed, this method provides an estimate of the sampling distribution of a statistic, and this empirical sampling distribution can be turned into an approximate confidence distribution in several ways (see, e.g., Schweder and Hjort, 2016, Chap. 7).

Bootstrap procedures can be distinguished into two major categories: parametric and non parametric. With parametric bootstrap, the sought distribution is taken to be that of the statistics given pseudo-data, generated by a consistent estimate of the parametric model. With the second, instead, the pseudo-data generating scheme is a multinomial resampling of data points. The error in bootstrap confidence intervals is in most of the cases of order $O_p(n^{-1})$, becoming of order $O_p(n^{-3/2})$ when correction procedures can be implemented, and for $t-$bootstrap (see DiCiccio and Romano, 1988; DiCiccio and Efron, 1996; Schweder and Hjort, 2016, Chap. 7).

For instance, consider a monotone transformation $h(\psi)$ of a scalar parameter of interest, and let $q(\psi, y) = (h(\psi) - h(\hat{\psi}))/\hat{\tau}$ an approximate studentized pivot, where $\hat{\psi}$ is the MLE of ψ and $\hat{\tau}$ is a suitable estimate of the pivot standard

deviation. Let $Q(\cdot)$ be the distribution function of $q(\psi, y)$. Then, a confidence distribution for the parameter of interest is $C(h(\psi)) = Q\Big((h(\psi) - h(\hat{\psi}))/\hat{\tau}\Big)$, with appropriate confidence density $cd(h(\psi)) = \partial C(h(\psi))/\partial\psi$. When $Q(\cdot)$ is unknown, it can be estimated via bootstrapping. Let $h(\psi^*)$ and $\hat{\tau}^*$ be the result of bootstrapping, then the $Q(\cdot)$ distribution can be estimated as $\hat{Q}$, via bootstrapped values of $q^* = q(\psi^*, y^*) = (h(\psi) - h(\hat{\psi}^*))/\hat{\tau}^*$. The approximate CD is then

$$C_{t-boot}(\psi) = \hat{Q}\left(\frac{h(\psi) - h(\hat{\psi})}{\hat{\tau}}\right).$$

This bootstrap method applies even when $q(\psi, y)$ is not a perfect pivot, but is especially successful when it is, because q^* then has exactly the same distribution $Q(\cdot)$ as $q(\psi, y)$. Note that the method automatically takes care of bias and asymmetry in $Q(\cdot)$.

3. Robust confidence distributions

3.1. Asymptotic derivation from M-estimating functions

The class of M-estimators (see, e.g. Huber, 1981; Hampel et al., 1986; Huber and Ronchetti, 2009) is broad and includes a variety of well-known estimators. For example it includes the MLE, the maximum composite likelihood estimator (see e.g. Varin et al., 2011), estimators based on proper scoring rules (see, e.g., Dawid et al., 2016, and references therein), and classical robust estimators (see e.g. Huber and Ronchetti, 2009, and references therein).

Under broad regularity conditions, an M-estimator $\tilde{\theta}$ is the solution of the unbiased estimating equation

$$g(\theta) = \sum_{i=1}^{n} g(y_i; \theta) = 0$$

and it is asymptotically normal, with mean θ and covariance matrix $V(\theta) = K(\theta)^{-1}J(\theta)(K(\theta)^{-1})^{\top}$, where $K(\theta) = E_\theta(\partial g(\theta)/\partial\theta^{\top})$ and $J(\theta) = E_\theta(g(\theta)g(\theta)^{\top})$ are the sensitivity and the variability matrices, respectively. The matrix $V_g(\theta) = V(\theta)^{-1}$ is known as the Godambe information and its form is due to the failure of the information identity since, in general, $K(\theta) \neq J(\theta)$. Let us denote with $G(\theta) = \sum_{i=1}^{n} G(y_i; \theta)$ the function such that $g(\theta)$ is the gradient vector, i.e. $g(y; \theta) = \partial G(y; \theta)/\partial\theta$.

From the general theory of M-estimators, the influence function (IF) of the estimator $\tilde{\theta}$ is given by

$$IF(y; \tilde{\theta}) = K(\theta)^{-1}g(y; \theta), \tag{4}$$

and it measures the effect on the estimator $\tilde{\theta}$ of an infinitesimal contamination at the point y, standardized by the mass of the contamination. The estimator $\tilde{\theta}$ is B-robust if and only if $g(y; \theta)$ is bounded in y. Note that the IF of the MLE is proportional to the score function; therefore, in general, MLE has unbounded IF, i.e. it is not B-robust.

Paralleling likelihood-based results, asymptotic robust inference on the scalar parameter of interest ψ can be based on first-order pivots. With the partition $\theta = (\psi, \lambda)$, the M-estimating function is similarly partitioned as $g(y; \theta) = (g_\psi(y; \theta), g_\lambda(y; \theta))$. Moreover, consider the further partitions

$$K = \begin{bmatrix} K_{\psi\psi} & K_{\psi\lambda} \\ K_{\lambda\psi} & K_{\lambda\lambda} \end{bmatrix}, \quad K^{-1} = \begin{bmatrix} K^{\psi\psi} & K^{\psi\lambda} \\ K^{\lambda\psi} & K^{\lambda\lambda} \end{bmatrix},$$

and similarly for V_g and V_g^{-1}. Finally, let $\tilde{\lambda}_\psi$ be the constrained M-estimate of λ, let $\tilde{\theta}_\psi = (\psi, \tilde{\lambda}_\psi)$, and let $\tilde{\psi}$ be the ψ component of $\tilde{\theta}$. Then, a profile Wald-type statistic for the ψ may be defined as

$$w_R(\psi) = (\tilde{\psi} - \psi)(\tilde{V}_g^{\psi\psi})^{-1/2},$$

and it has an asymptotic $N(0, 1)$ null distribution. Similarly, the profile score-type statistic

$$w_{sR}(\psi) = g_\psi(\tilde{\theta}_\psi)^T K^{\psi\psi}(V_g^{\psi\psi})^{-1}K^{\psi\psi}g_\psi(\tilde{\theta}_\psi)$$

has an asymptotic χ^2_1 null distribution, while the asymptotic distribution of the profile ratio-type statistic for ψ, given by $W_R(\psi) = 2\big(G(\tilde{\theta}_\psi) - G(\tilde{\theta})\big)$, is $\nu\chi^2_1$, where $\nu = (\tilde{K}^{\psi\psi})^{-1}\tilde{V}_g^{\psi\psi}$. In view of this, for the adjusted profile ratio-type statistic to first-order it holds

$$W_R^{adj}(\psi) = \frac{W_R(\psi)}{\nu} \doteq \chi^2_1.$$

Finally, the adjusted profile root, analogous to (3), can be defined as

$$r_R(\psi) = \text{sign}(\tilde{\psi} - \psi)\sqrt{W_R^{adj}(\psi)},$$

which has an asymptotic standard normal distribution. For the general theory of robust tests see Heritier and Ronchetti (1994).

Paralleling results in Section 2.1 for likelihood based CDs, a recipe to derive an asymptotic CD from robust M-estimating functions is based on pivotal quantities, extending the theory illustrated for robust scoring rules in Hjort and Schweder (2018) and Ruli et al. (2022). To this end, let us denote with $q_R(\psi; y)$ a robust pivotal quantity, such as the profile Wald-type statistic $w_R(\psi)$ or the adjusted profile scoring rule root $r_R(\psi)$. Then,

$$C_R^w(\psi) \doteq \Phi\left((\psi - \tilde{\psi})(\tilde{V}_g^{\psi\psi})^{-1/2}\right) \tag{5}$$

and

$$C_R^r(\psi) \doteq \Phi\left(\text{sign}(\psi - \tilde{\psi})\sqrt{W_R^{adj}(\psi)}\right) \tag{6}$$

are first-order asymptotic CDs, and the corresponding confidence densities are, respectively,

$$cd_R^w(\psi) \doteq \frac{\phi\left((\psi - \tilde{\psi})(\tilde{V}_g^{\psi\psi})^{-1/2}\right)}{\sqrt{\tilde{V}_g^{\psi\psi}}}$$

and

$$cd_R^r(\psi) \doteq \phi\left(\text{sign}(\psi - \tilde{\psi})\sqrt{W_R^{adj}(\psi)}\right)\left|\frac{\partial W_R^{adj}(\psi)^{1/2}}{\partial \psi}\right|,$$

where $\phi(\cdot)$ is the density function of the standard normal distribution. Note that the Wald-type based confidence density $cd_R^w(\psi)$ coincides with the asymptotic first-order robust Bayesian posterior distribution for ψ (see, e.g., Greco et al., 2008, and Ventura and Racugno, 2016).

In practice, using for instance (6), the confidence median is $\tilde{\psi}$ and an $(1-\alpha)$ equi-tailed confidence interval can be obtained as $\{\psi : |r_R(\psi)| \le z_{1-\alpha/2}\}$, where $z_{1-\alpha/2}\}$ is the $(1-\alpha/2)$-quantile of the standard normal density. When testing, for instance, $H_0 : \psi = \psi_0$ against $H_1 : \psi < \psi_0$, the p-value is $p = C_R^r(\psi_0)$, while when testing $H_0 : \psi = \psi_0$ against $H_1 : \psi \neq \psi_0$ the p-value is $p = 2(1 - \Phi(|r_R(\psi_0)|))$. Furthermore, a measure of evidence for a statement of the form "$\psi_1 < \psi < \psi_2$" can be computed as $C_R^r(\psi_2) - C_R^r(\psi_1)$.

To study the stability of robust CDs, let us write the robust pivotal quantity more generally as $q_R(\psi; T(\hat{F}_n))$, where $\hat{F}_n$ is the empirical distribution function and $T(F)$ is the functional defined by the unbiased M-estimating equation $\int g(y; T(F))\, dF(y) = 0$, where $F = F(y; \theta)$ is the assumed parametric model. In CD inference the tail area, given by $C_R(\psi) = \Phi(q_R(\psi; T(\hat{F}_n)))$, plays a central role and thus we can consider the tail area influence function (see, e.g., Field and Ronchetti, 1990, and Ronchetti and Ventura, 2001), given by

$$TAIF(y; T) = \frac{\partial}{\partial \varepsilon}\Phi(q_R(\psi; T(F_\varepsilon)))\Big|_{\varepsilon=0},$$

where $F_\varepsilon = (1-\varepsilon)F + \epsilon\Delta_y$ and Δ_y is the probability measure which puts mass 1 at the point y. The $TAIF(y; T)$ thus describes the normalized influence on the CD tail area of an infinitesimal observation at y and, by considering its supremum, it can be used to evaluate the maximum bias of the tail area on the ε-neighborhood of F. It can be shown that

$$TAIF(y; T) = \phi(q_R(\psi; T(F)))\ \frac{\partial q_R(\psi; T(F))}{\partial T(F)}\ \frac{\partial T(F_\varepsilon)}{\partial \varepsilon}\Big|_{\varepsilon=0}, \tag{7}$$

where the last term in (7) is the IF (4) of the M-estimator. Thus, the tail area influence function for the CD tail area at the statistical model F is proportional to the M-estimating function and this gives an immediate handle on robustness. Furthermore, it is bounded with respect to y when the M-estimating function is bounded.

The application of (5) and (6) in the particular context of a robust scoring rule has been discussed in Ruli et al. (2022). In particular, the Tsallis score (Tsallis, 1988) is considered, which is given by

$$G(y; \theta) = (\gamma - 1)\int f(y; \theta)^\gamma\, dy - \gamma f(y; \theta)^{\gamma-1}, \quad \gamma > 1,$$

with corresponding unbiased M-estimating function $g(\theta) = \partial G(y; \theta)/\partial\theta$ (Ghosh and Basu, 2013; Dawid et al., 2016), and with the parameter γ which gives a trade-off between efficiency and robustness.

3.2. Derivation of robust confidence distributions via simulation

In the context of robust procedures, suitable modifications of the bootstrap have been explored for improving the stability of inference. In fact, both families of parametric and non parametric bootstrap presents some drawbacks. For instance, with the non parametric bootstrap, despite the direct specification of the data generating process is avoided and this robustifies the analysis, the distribution of estimators might be unstable since the amount of outliers of pseudo-samples can be higher than that in the original dataset. For remedying this drawback, some authors proposed the use of weights associated

to observations before performing repeatedly likelihood maximization under the assumed model, as in weighted likelihood bootstrap (Newton and Raftery, 1994). Lyddon et al. (2019) showed that the corresponding estimator is asymptotically normal and with covariance matrix with the classical structure of sandwich form. Also, Chen and Zhou (2020) introduced a similar approach based on estimating equations and provided non-asymptotic guarantees for the resulting errors. Moreover, a further problem when using a bootstrap approach together with robust procedures as M-estimating equations is the need of repeatedly solving numerically non-convex or complex optimization problems, which may be computationally expensive.

In this section we inspect two alternative methods for computing CDs based on robust M-estimating functions, that go beyond some limitations of the bootstrap. Broadly speaking, the first method is based on a frequentist reinterpretation of the ABC machinery (see, e.g., Bee et al., 2017; Ruli et al., 2020; Thornton et al., 2022), whose properties have been derived by Rubio and Johansen (2013) in a general setup. The idea consists in generating candidate parameter values from an uniform distribution, computing a robust suitable summary statistic using the simulated data and then accepting only the parameter values such that the corresponding summary statistic is "close" to its observed counterpart (see Algorithm 1).

Algorithm 1 Accept-reject robust ABC.

Input: proposal $p(\psi, \lambda)$, number of iterations R, robust summary statistic $t(\cdot)$, $t^{\text{obs}} = t(y^{\text{obs}})$, where y^{obs} is the observed sample, tolerance ε, distance $\rho(\cdot;\cdot)$

for $j \in 1, \dots, R$ **do**
  Sample $(\psi_j^*, \lambda_j^*) \sim p(\psi, \lambda)$and $y_j^* \sim f(y; \psi_j^*, \lambda_j^*)$
  Compute $t_j^* = t(y_j^*)$
  Accept ψ_j^* if $\rho(t_j^*; t^{\text{obs}}) \leq \varepsilon$ else reject
end for
resample the accepted (ψ^*, λ^*) with probability $\propto 1/p(\psi^*, \lambda^*)$
return robust approximate normalized pseudo-likelihood $\propto$ confidence density $\hat{cd}_R^{abc}(\psi)$

In Algorithm 1, the summary statistics of Soubeyrand and Haon-Lasportes (2015) or of Ruli et al. (2016,2020) can be used. In particular, the first one is based directly on the M-estimator $\tilde{\psi}$ as the summary statistic $t(y)$ and a, possibly rescaled, distance among the observed and the simulated value of the statistic. In the second one, a rescaled version of the M-estimating function $g(\theta)$, evaluated at a fixed value of the parameter, is used as a summary statistic $t(y)$; this avoids repeated evaluations of the consistency correction involved in the M-estimating function. Note also that when using the M-estimator as a summary statistic, the algorithm for solving the estimating equation might not converge after a prefixed number of iterations, thus causing additional noise in the results. For a single parameter of interest, with the partition $\theta = (\psi, \lambda)$, we propose to modify the algorithm of Ruli et al. (2020) by using the profile estimating equation $g_\psi(y; \psi, \lambda)$ and plugging in the value of proposals λ^* for nuisance parameters used to generate pseudo-data. The treatment of the nuisance parameters resembles as a generalized profile likelihood computation.

Note that, assuming the regularity assumptions of Soubeyrand and Haon-Lasportes (2015) and the usual regularity conditions on M-estimators (Huber and Ronchetti, 2009, Chap. 4), then for $n \to \infty$ the robust confidence densities derived via simulation are asymptotically equivalent to the Wald-type confidence density $cd_R^w(\psi)$. Moreover, following Ruli et al. (2020), if $g(y;\theta)$ is bounded in y, i.e. if the M-estimator is B-robust, then asymptotically the posterior mode, as well as other posterior summaries of the robust confidence density $\hat{cd}_R^{abc}(\psi)$ have bounded IF.

The second method for computing CDs based on robust M-estimating functions is similar but aims at computing directly a Monte Carlo p-value or a significance function using a different acceptance rule (see Bortolato and Ventura, 2022): again the parameter values are sampled from a uniform distribution, a robust summary statistic $t(y^*)$ using the simulated data is obtained, and then the proposed parameter is accepted if the robust summary statistic t^* is greater of its observed counterpart t^{obs} (see Algorithm 2). For obtaining the confidence density, if the obtained CD is indeed monotone increasing, hence it is a proper CD, Algorithm 3 can be used.

Algorithm 2 Accept-reject confidence curve and confidence distribution computing.

Input: proposal $p(\psi, \lambda)$, number of iterations R, robust summary statistic $t(\cdot)$, $t^{\text{obs}} = t(y^{\text{obs}})$, where y^{obs} is the observed sample.

for $j \in 1, \dots, R$ **do**
  Sample $(\psi_j^*, \lambda_j^*) \sim p(\psi, \lambda)$and $y_j^* \sim f(y; \psi_j^*, \lambda_j^*)$
  Compute $t_j^* = t(y_j^*)$
  Accept ψ_j^* if $t_j^* \geq t^{\text{obs}}$ else reject
end forreturn ψ^* with density $\propto cc_R(\psi)$ a confidence curve/distribution

Algorithm 3 Confidence density.

Input: robust confidence distribution $C_R(\psi)$, desired size R,grid of values $G = \{\frac{1}{R}, \frac{2}{R}, \dots, \frac{R-1}{R}, 1\}$
for $j \in 1, \dots, R$ **do**
Compute the empirical quantiles of the CD:$\psi_j^* = C_R^{-1}(G_j)$
end for
return $\psi^* \sim \hat{cd}_R(\psi)$.

As a final remark, we observe that, for obtaining stable results with the accept-reject schemes, it is suggestable to increase the number of proposals as the dimension of the parameter space increases, in order to have reasonable acceptance rate with ABC-type algorithm, and in general for obtaining more precise estimation of the confidence distributions.

4. Applications to non-inferiority and superiority testing

The aim of this section is to introduce and apply CDs inference in the context of non-inferiority (and superiority) testing, in which interest is in establishing if a new product is not unacceptably worse than a product already in use. For instance, applications of non-inferiority testing has revealed an attractive problem in medical statistics, biostatistics, statistical quality control and engineering statistics, among others. Here we focus in non-inferiority clinical trials where the aim is to show that an experimental treatment is not (much) worse than a standard treatment. Clinical practice, however, is not the only field of application of these tests: in comparing the performance of sensors in industrial environment, for instance, the margin may be linked to some difference in costs due to sensor functioning. Other applications can be found in machine learning literature, where instead the meaningful margin is related to the accuracy or to the speed in classification tasks.

In the process of evaluating the efficacy of an experimental treatment, it is common to develop studies in which the two arms are the new and the standard therapy, respectively, rather than the new and the placebo. This is because it is considered unethical to deprive patients from a therapy that has already been proven to be beneficial. The underlying research hypothesis to be verified is that new therapies have equivalent or non-inferior efficacy to the ones currently in use. Both non-inferiority and superiority tests are examples of directional (one-sided) tests (see, e.g., D'Agostino et al., 2003; Rothmann et al., 2012, and references therein). In particular, the *non-inferiority test* wants to test that the treatment mean μ_N is not worse than the reference mean μ_S by more than a given equivalence margin δ. The actual direction of the hypothesis depends on the response variable being studied. This question can be formulated into a test procedure for which the null hypothesis is

$$H_0 : \mu_S - \mu_N \geq \delta,$$

where $\delta > 0$ is the equivalence margin, when higher values of the response variable mean better results, versus

$$H_1 : \mu_S - \mu_N < \delta.$$

The scalar parameter of interest in this context is thus $\psi = \mu_S - \mu_N$, and non-inferiority is claimed when the null hypothesis is rejected.

The equivalence margin δ corresponds to the practical acceptable difference and should be pre-specified before the data is recorded (see e.g. Garret, 2003). An overly conservative margin might result in a high risk of not being able to claim non-inferiority when it actually is non-inferior. Conversely, overly liberal margins could result in a high risk of claiming non-inferiority when it actually is not non-inferior. A reasonable margin would be best derived from a combination of factors: the expected event rate, the duration of follow-up, and the number and nature of the events. However, arbitrary clinical judgment and the sponsor budget are of a great influence, resulting in a somewhat subjective non-inferiority margin. It is not clear in some situations how to perform the choice, and multiple thresholds could be plausible; in this respect, CDs are particularly useful to perform sensitivity analyses. Indeed, in this situation a confidence distribution on the difference $\psi = \mu_S - \mu_N$ will simultaneously show the evidence of the p-value against the null for a series of values δ, and decide for a reasonable δ with the nominal control of the rejection level and possible alternatives.

Here we consider an example of trial where higher levels of the response variable mean that the new treatment is effective. The aim is verifying that the new treatment (N) is not unacceptably worse to the standard (S). Let us assume that $n = 80$ patients are randomized into two groups, and the model for the data is assumed to be

$$Y_S = \mu_N + \psi + u, \quad Y_N = \mu_N + u, \quad u \sim N(0, \sigma^2). \tag{8}$$

The normal distribution on the error term is often the basis of statistical analyses in medicine, genetics and in related sciences. Under this assumption, parametric inferential procedures based on the sample means, standard deviations, two-samples t-test, and so on, are the most efficient. However, it is well known that they are not robust when the normal distribution is just an approximate parametric model or in the presence of deviant values in the observed data (see, e.g., Farcomeni and Ventura, 2012). In the framework described by (8), we inspect the effect of model misspecification. In particular, misspecification considers a mixture model, where one mixes in a small amount (say 10%) of contamination in the data of the new treatment group. In particular, in the contaminated scenario, 10% of the error terms in the new treatment group are half-Cauchy distributed (see Fig. 3).

Table 1
Confidence measures of evidence for the null hypothesis $H_0 : \psi > \delta$, with $\delta = 4$ associated to Fig. 4, without and with contamination.

Method	0% cont.	10% cont.
Wald/Mean	0.08	0.42
Wald/M-test	0.07	0.03
ABC/Median	0.24	0.20
ABC/M-EE	0.11	0.09
ABC/M-est	0.12	0.09
CDensity/Median	0.20	0.21
CDensity/M-EE	0.08	0.05
CDensity/M-est	0.14	0.11

It is of interest to compare CDs inference for ψ based on the following approaches (abbreviations are also used in Fig. 4 and in the following) used to derive confidence densities:

1. exact classical Wald-type confidence density based on $w_p(\psi)$, which is related to the classical two sample t-test (Wald/Mean)
2. robust asymptotic Wald-type confidence density $cd_R^w(\psi)$ based on the Huber's estimator (Wald/M-test)
3. approximate confidence density based on ABC (Algorithm 1) with robust Huber's estimator as summary statistics (ABC/M-est)
4. approximate confidence density based on ABC (Algorithm 1) with the robust Huber's estimating equation as summary statistic (ABC/M-EE)
5. simulated confidence density (Algorithm 2) based on the robust Huber's estimator (CDensity/M-est)
6. simulated confidence density (Algorithm 2) based on the robust Huber's estimating equation (CDensity/M-EE)
7. approximate confidence density based on ABC (Algorithm 1) with the difference of medians as summary statistics (ABC/Median)
8. simulated confidence density (Algorithm 2) based on the difference of medians (CDensity/Median).

The nominal value of the mean difference between the treatment effects is ψ_0 is fixed to 2.6, and for simulation-based confidence distribution as well as for those obtained by the ABC-type algorithm we used 10^5 proposals and a tolerance level of 0.1. In the Huber's estimator we fix the tuning constant which controls the desired degree of robustness to 1.345, which imply that the estimator is 5% less efficient than the corresponding MLE under the assumed model.

From the resulting confidence densities illustrated in Fig. 4 we note that, when the data come from the central model (left column) all the confidence densities are in reasonable agreement, even if the confidence densities based on the medians behave slightly worse, with a greater variability. Under model misspecification (right column), the non robust confidence density Wald/Mean is less trustworthy as it drifts away from the true parameter value (green dotted line). This is not the case however for the robust confidence densities which remains centered around the true parameter value. We further note that in the contaminated case, the robust confidence densities based on the M-estimating equation (ABC/M-EE and CDensity/M-EE) are the ones with the smallest variability. For all these confidence densities, Table 1 gives the measures of evidence for the equivalence margin δ taken equal to 4 (black dotted line in Fig. 4), that is for the statement "$\psi > \delta$". As a reference, consider the result derived by the exact $t-$distribution of the exact classical Wald-type confidence density in the non contaminated case, which is 0.08. The results, without and with the contamination, confirm the behavior of the confidence densities in Fig. 4, in particular the non robustness of the likelihood-based confidence density (Wald/Mean). The most stable values under contamination seem to be those obtained with M-EE approaches (0.09 with ABC/M-EE and 0.05 with CD/M-EE). The same analysis could be done in principle, given the CD, for any margin δ.

4.1. Simulation study

For investigating the behavior of the several confidence densities, we perform a simulation study under two sample sizes $n = 40, 80$ (20, 40 per group) and for each of them we investigate two main scenarios: one in which the assumptions of the model in (8) are met by the true data generating model, and the second one where 10% of the error terms in the new treatment group are half-Cauchy distributed (as in Fig. 3). The proposals for ψ were taken Uniform in $[-3, 9]$, for the parameter μ_N we sample from a Uniform [110,130], while for σ we generated values from a Uniform [1,8]. In the simulation study, we also let the nominal values of the parameter of interest vary ($\psi_0 = 2.6$ and $\psi_0 = 1$), to test the stability of the proposals distributions for simulation-based CDs.

The families of methods to derive the confidence densities considered are the same as in the example above, and confidence distributions construction are again based on exact and asymptotic pivotal quantities or simulation-based. We also compare the results with those obtained with Bootstrap methods: in particular basic parametric bootstrap (Boot/Basic), parametric bootstrap with normal intervals (Boot/Norm) and parametric bootstrap with percentiles intervals (Boot/Perc). For Rejection-ABC-type confidence distributions, the tolerance for the discrepancy was set to 0.1. For all the techniques using $M-$estimators or $M-$estimating equation, the Huber's tuning constant is fixed to 1.345. For computing a single confidence

Table 2
Empirical coverages in a simulation study without and with 10% of contamination and $n = 40$.

Contamination		0%		10%	
$n = 40$		95% CI	90% CI	95% CI	90% CI
ψ_0=2.6	Wald/Mean	93.9	89.1	97.1	94.0
	Wald/M-test	93.7	88.4	94.1	88.3
	ABC/Median	97.1	93.4	97.7	93.7
	ABC/M-EE	92.7	87.2	93.7	88.9
	ABC/M-est	97.0	93.1	97.6	93.9
	CDensity/Median	99.5	97.6	99.2	98.0
	CDensity/M-EE	95.8	90.5	96.7	92.1
	CDensity/M-est	99.4	97.3	99.2	98.0
	Boot/basic	93.4	88.0	92.3	86.1
	Boot/Norm	93.5	88.2	92.4	86.0
	Boot/Perc	93.4	87.9	92.3	86.1
ψ_0=1	Wald/Mean	95.4	89.9	97.0	93.0
	Wald/M-test	94.3	88.4	94.1	88.9
	ABC/Median	95.8	91.7	95.5	91.6
	ABC/M-EE	95.4	89.9	97.0	93.0
	ABC/M-est	95.5	90.5	95.7	91.8
	CDensity/Median	98.9	97.6	99.4	97.9
	CDensity/M-EE	95.1	91.0	96.8	92.9
	CDensity/M-est	99.3	97.0	99.4	98.2
	Boot/basic	92.7	87.7	92.3	86.0
	Boot/Norm	92.6	87.7	92.3	86.0
	Boot/Perc	92.8	87.7	92.3	86.2

Table 3
Empirical coverages in a simulation study without and with 10% of contamination and $n = 80$.

Contamination		0%		10%	
$n = 80$		95% CI	90% CI	95% CI	90% CI
ψ_0=2.6	Wald/Mean	95.5	90.0	95.9	92.2
	Wald/M-test	95.1	89.6	93.9	87.9
	ABC/Median	97.2	93.5	97.3	93.5
	ABC/M-EE	93.3	86.9	92.7	87.3
	ABC/M-est	97.5	93.6	97.2	93.9
	CDensity/Median	99.1	97.6	99.2	97.5
	CDensity/M-EE	95.9	89.4	96.4	92.5
	CDensity/M-est	99.1	97.3	99.2	97.7
	Boot/Basic	94.1	89.5	92.3	87.5
	Boot/Norm	94.2	89.5	92.4	87.4
	Boot/Perc	94.3	89.6	92.3	87.5
$\psi_0 = 1$	Wald/Mean	94.9	89.7	96.8	92.8
	Wald/M-test	94.5	89.4	94.4	89.3
	ABC/Median	95.3	90.4	95.5	92.0
	ABC/M-EE	94.9	89.7	96.8	92.8
	ABC/M-est	95.4	90.2	95.7	91.4
	CDensity/Median	99.4	97.7	99.0	97.3
	CDensity/M-EE	96.2	92.0	96.0	92.2
	CDensity/M-est	99.1	97.3	99.1	97.0
	Boot/Basic	94.1	89.5	92.3	87.5
	Boot/Norm	94.2	89.5	92.4	87.4
	Boot/Perc	94.3	89.6	92.3	87.6

distribution, R =4000 values were generated from the proposals, while for each scenario simulation experiments were replicated 2000 times. Results obtained for the empirical coverages of 90% and 95% equi-tailed confidence intervals are synthesized in Tables 2 and 3. We also report in Tables 4 and 5 the error associated to confidence median point estimators, in terms of bias ($b = \sum_{r=1}^{R} \tilde{\theta}_r - \theta_0$), probability of underestimation ($PU = \sum_{r=1}^{R} 1_{\{\tilde{\theta}_r < \theta_0\}}$) and type I error with $\alpha = 0.05$. We note that, under the central model, the Wald/Mean CD shows a good performance, as well as some robust CDs (Wald/M-test, CDensity/M-EE and ABC/M-EE) and bootstrap-based CDs. With contaminated data, the Wald/Mean CD and bootstrap-based CDs tend to be affected by contamination, whereas the robust CDs perform substantially better, with the CDs based on M-estimating equations being preferred over those based on M-estimators. Asymptotic symmetric confidence densities based on Wald-type robust CDs and ABC-type confidence densities seem to be affected more by bias than the simulated CDs (see Tables 4 and 5). Note finally that ABC-type results, even if behaving well, depend on a tolerance choice, hence the results may degradate when the latter is not well calibrated.

Table 4
Measures of stability of CDs: absolute bias ($|b|$), probability of underestimation (PU) and type I error ($\alpha = 0.05$) of confidence estimators (medians) in the simulation study with $n = 40$.

Contamination		0%			10%		
$n = 40$		$\lvert b\rvert$	PU	I type err.	$\lvert b\rvert$	PU	I type err.
$\psi_0 = 2.6$	Wald/Mean	0.01	0.51	0.06	5.57	0.65	0.03
	Wald/M-test	0.00	0.51	0.06	0.23	0.42	0.08
	ABC/Median	0.03	0.52	0.01	0.09	0.46	0.01
	ABC /M-EE	0.00	0.51	0.03	0.23	0.42	0.03
	ABC/M-est	0.01	0.51	0.01	0.23	0.42	0.01
	CDensity/Median	0.15	0.55	0.01	0.03	0.51	0.01
	CDensity/M-EE	0.11	0.55	0.03	0.11	0.46	0.03
	CDensity/M-est	0.13	0.56	0.01	0.09	0.46	0.01
	Boot/Basic	0.01	0.50	0.06	0.28	0.58	0.10
	Boot/Norm	0.01	0.50	0.06	0.28	0.58	0.10
	Boot/Perc	0.01	0.51	0.06	0.27	0.58	0.09
$\psi_0 = 1$	Wald/Mean	0.01	0.49	0.05	2.90	0.66	0.04
	Wald/M-test	0.01	0.49	0.06	0.20	0.43	0.07
	ABC/Median	0.10	0.53	0.00	0.00	0.50	0.01
	ABC/M-EE	0.00	0.50	0.01	0.18	0.44	0.01
	ABC/M-est	0.00	0.57	0.01	0.16	0.44	0.02
	CDensity/Median	0.30	0.56	0.01	0.20	0.57	0.01
	CDensity/M-EE	0.20	0.55	0.01	0.00	0.49	0.01
	CDensity/M-est	0.20	0.56	0.01	0.10	0.52	0.02
	Boot/Basic	0.01	0.50	0.06	0.28	0.58	0.10
	Boot/Norm	0.01	0.50	0.06	0.28	0.58	0.10
	Boot/Perc	0.01	0.51	0.06	0.27	0.58	0.09

Table 5
Measures of stability of CDs: absolute bias ($|b|$), probability of underestimation (PU) and type I error ($\alpha = 0.05$) of confidence estimators (medians) in the simulation study with $n = 80$.

Contamination		0%			10%		
$n = 80$		$\lvert b\rvert$	PU	I type err.	$\lvert b\rvert$	PU	I type err.
$\psi_0 = 2.6$	Wald/Mean	0.02	0.49	0.05	1.76	0.58	0.05
	Wald/M-test	0.01	0.49	0.05	0.19	0.42	0.08
	ABC/Median	0.02	0.50	0.02	0.05	0.49	0.02
	ABC/M-EE	0.02	0.48	0.03	0.19	0.42	0.03
	ABC/M-est	0.01	0.49	0.01	0.19	0.42	0.02
	CDensity/Median	0.07	0.53	0.01	0.02	0.51	0.02
	CDensity/M-EE	0.08	0.53	0.03	0.11	0.45	0.03
	CDensity/M-est	0.08	0.54	0.01	0.11	0.46	0.02
	Boot/Basic	0.03	0.49	0.05	0.39	0.33	0.09
	Boot/Norm	0.03	0.49	0.05	0.39	0.33	0.09
	Boot/Perc	0.03	0.49	0.05	0.39	0.33	0.09
$\psi_0 = 1$	Wald/Mean	0.05	0.47	0.06	1.40	0.59	0.04
	Wald/M-test	0.05	0.48	0.06	0.16	0.41	0.07
	ABC/Median	0.02	0.50	0.01	0.02	0.50	0.01
	ABC/M-EE	0.04	0.48	0.01	0.15	0.42	0.01
	ABC/M-est	0.04	0.48	0.01	0.16	0.41	0.02
	CDensity/Median	0.20	0.54	0.01	0.30	0.58	0.01
	CDensity/M-EE	0.10	0.53	0.01	0.00	0.49	0.01
	CDensity/M-est	0.10	0.54	0.02	0.10	0.51	0.02
	Boot/Basic	0.03	0.49	0.05	0.39	0.67	0.09
	Boot/Norm	0.03	0.49	0.05	0.39	0.67	0.09
	Boot/Perc	0.03	0.49	0.05	0.39	0.67	0.09

As a final remark, note that an interesting aspect of this simulation study was the difference among the approach of using robust M-estimating functions instead of robust M-estimates, especially in the treatment of nuisance parameters. For the M-est CD based on $\tilde{\psi}$, for each fixed values (ψ^*, λ^*), the corresponding acceptance probability is $Pr_{\psi^*,\lambda^*}\left(\tilde{\psi}^* > \tilde{\psi}\right)$. This might be recognized to be similar to a bootstrap p-value, with the exception of what is the model considered for the simulation. In contrast, for the CD with the profile M-estimating function as summary statistic, the acceptance probability for same values (ψ^*, λ^*) is $Pr_{\psi^*,\lambda^*}\left(g_\psi(y; \tilde{\psi}, \lambda^*) \geq 0\right)$, as if λ^* was the oracle estimate. Hence, the CD for each ψ is associated to an average p-value, i.e. $CD(\psi) = \int Pr_{\psi,\lambda}(g_\psi(y; \tilde{\psi}, \lambda) \geq 0) d\lambda$. Note that under the true model, with (ψ_0, λ_0) true parameter value, the corresponding p-value would be equal to the one without the nuisance parameter $Pr_{\psi_0,\lambda_0}(g_\psi(y; \tilde{\psi}, \lambda_0) \geq 0)$. Note that, instead, keeping fixed the nuisance parameters in the simulations would correspond to consider them as known.

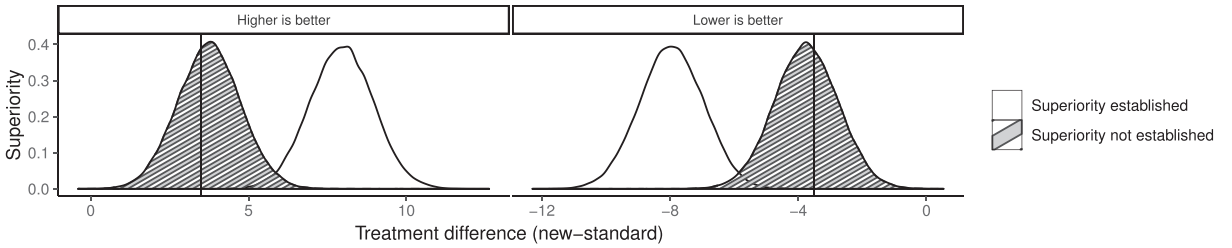

Fig. 5. Example of making inference with confidence densities in superiority tests, with margin $\delta = -3.5$.

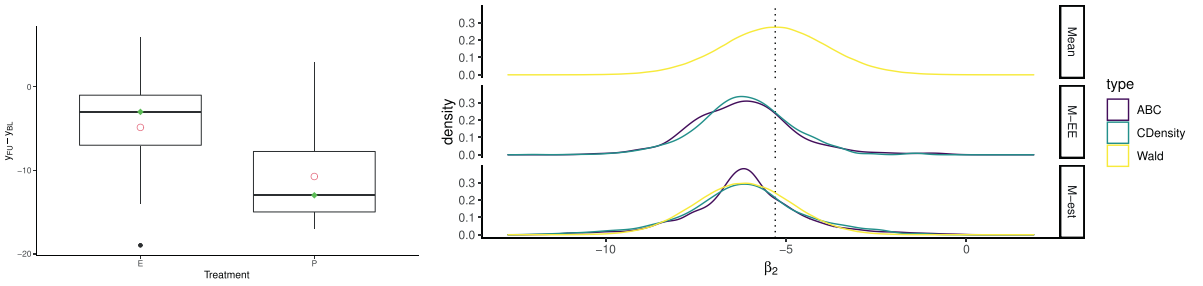

Fig. 6. Real data example: boxplots of pre-post differences of scores in the group using psilocybin (P) versus escitalopram (E) (left), confidence densities for the parameter β_2, measuring the difference in efficacy of the two therapies (right).

Table 6
Measures of evidence for the hypothesis "$\beta_2 > \delta$" for several margins.

$-\delta$	-3.5	-4	-4.5	-5	-5.3
Wald/Mean	0.11	0.18	0.29	0.41	0.49
Wald/M-test	0.02	0.05	0.10	0.19	0.26
ABC/M-est	0.00	0.00	0.14	0.14	0.29
ABC/M-EE	0.08	0.12	0.18	0.24	0.30
CDensity/M-est	0.03	0.09	0.14	0.21	0.28
CDensity/M-EE	0.07	0.14	0.17	0.22	0.28

4.2. Real data application

A class of problems requiring similar considerations to those of non-inferiority tests, i.e. sensitivity analysis with respect to the reference margin δ, is that of superiority studies (see Fig. 5).

Here we analyze the data collected in a randomized controlled trial (see Carhart-Harris et al., 2021; Nayak et al., 2022) with the aim of assessing the superiority of a new therapy with psilocybin (P) versus that with escitalopram (E), in treating major depressive disorder. The dataset contains the scores obtained by $n = 57$ patients on a questionnaire, before and after a 6-week period of therapy.

The model considered for the scores at the time of follow-up (FU) is the following

$$y_{FU} = \beta_0 + \beta_1 y_{BL} + \beta_2 \mathrm{P} + u, \quad u \sim N(0, \sigma^2),$$

where y_{BL} represents the value at the baseline and P is a dummy variable that equals 1 if the subject belongs to the group treated with the new therapy (psilocybin), and thus the coefficient relates to the additional change with respect to the control group (escitalopram) after the therapy. A reduction of the score indicates a clinical improvement; thus superiority is claimed if the estimate of the coefficient β_2 is sufficiently lower than 0. In particular, in order to conclude in favor of meaningful superiority, the clinicians considered as reference a margin $\delta = -5.3$. It is of interest to provide stable measures of evidence for the statement "$\beta_2 > \delta$", with $\delta = -5.3$ (H_0).

The MLE for the parameter β_2 and its standard error are, respectively, -5.32 and 1.44, while the robust counterparts are -6.18 and 1.33. Note that after removing two outliers the MLE become -6.23, with standard error 1.35. We resume the whole confidence densities based on Wald-type methods together with simulated confidence densities based on Huber's estimators and Huber's estimating equations in Fig. 6. As it can be noted the classical confidence density (Wald/Mean) is shifted to the right, because of the presence of outliers. Evidence measures for different margins are reported in Table 6. Using the margin chosen by the clinicians (-5.3) there is no evidence of superiority at level $\alpha = 0.1$; however note that the measure of evidence computed with the Wald-type confidence density (Wald/Mean) is the double of the ones computed with the robust confidence densities. With a margin of $\delta = -3.5$ all the robust procedure would agree in claiming superiority

with $\alpha = 0.1$, while according to classical Wald-type confidence density (Wald/Mean) there would not be enough evidence to conclude superiority.

5. Confidence distributions based on integral probability semimetrics

For mitigating the effects of small departures from model assumptions, and possible dramatic changes of inferential conclusions due to inconvenient choices of pivotal quantities and summaries, the use of non parametric procedures may be an alternative.

Here, we mention the construction of CDs based on discrepancy measures defined on the space of distributions, belonging to the class of integral probability semimetrics (Muller, 1997) or pseudo-metrics (Huber and Ronchetti, 2009, Chapt. 2). These divergences are classically associated to the concept of stability, used as global tests and studied in the context of misspecified models where the meaningfulness of model features is uncertain, whereas directly comparing the distributions happens to be more natural (Bernton et al., 2019; Frazier et al., 2020; Legramanti et al., 2022). In particular, we focus on the Kolmogorov-Smirnov distance (d_{KS}) and the Wasserstein distance (d_W) for one-dimensional distributions, defined respectively as

$$d_{KS}(P, Q) = \sup_y |P(y) - Q(y)|,$$

$$d_W(P, Q) = \int_Y |P(y) - Q(y)| dy.$$

The discrepancy furnishes global indication of potential agreement between the two distributions P and Q, analogously to a likelihood ratio test as in the likelihood-based inference for correctly specified models. For estimating the distances, empirical cumulative distribution functions ($\hat{P}_n(y)$ and $\hat{Q}_n(y)$, respectively) are used.

In general models, the non-asymptotic distributions of the statistics $d_{KS}(\cdot, \cdot)$ and $d_W(\cdot, \cdot)$ are complex, and numerical methods are employed to compute exact p-values. Here we suggest to rely on Algorithm 1 to derive CDs, once identified as summary statistics suitable discrepancies with distribution stochastically monotone in a scalar parameter of interest θ. In particular, let us consider the observed sample y^{obs} and a fixed reference sample y^{ref}, drawn from a completely known model $f(y; \theta^{\text{ref}})$. A sequence of unilateral tests can be built by using as observed summary statistic in Algorithm 2 the quantity $d(y^{\text{obs}}, y^{\text{ref}}) = d(\hat{P}(y^{\text{obs}}), \hat{Q}(y^{\text{ref}}))$, where $d(\cdot, \cdot)$ may be the Kolmogorov-Smirnov distance (d_{KS}) or the Wasserstein distance (d_W). Then, the CD is obtained with the Accept-Reject scheme of Algorithm 2, evaluating

$$Pr_{\theta^*}(d(y^*, y^{\text{ref}}) > d(y^{\text{obs}}, y^{\text{ref}})),$$

where y^* is simulated from the central model $y^* \sim f(y; \theta^*)$. Also, by Algorithm 3 a confidence density can be retrieved. To obtain a proper confidence distribution, the distribution of the summary statistic should be stochastically ordered in the parameter of interest. Hence it is convenient to draw y^{ref} from the model $f(y; \theta')$, with θ' being the supremum of the proposal distribution support in Algorithm 2.

Otherwise, a series of bilateral tests, straightly comparing $d(y^*, y^{\text{obs}})$ to zero, without a reference sample, can also be performed, for obtaining a confidence curve instead of a proper confidence distribution (Legramanti et al., 2022).

5.1. Simulation study

As in Legramanti et al. (2022) we consider a contamination study. Compared to the approximate posteriors in Legramanti et al. (2022), the CD based on Wasserstein distance seems even more stable.

Let us denote with $\tilde{\theta}^m$ the confidence median and let us focus on the one-dimensional Wasserstein distance. Under the non contaminated sample (y_{θ_0}) the confidence median satisfies

$$Pr(d_W(y_{\tilde{\theta}^m}, y_{\theta^{\text{ref}}}) > d_W(y_{\theta_0}, y_{\theta^{\text{ref}}})) = 0.5. \tag{9}$$

When considering a ϵ-contaminated sample, assuming that the contamination c has negligible mass under the central model, we look for θ^* that satisfies

$$Pr(d_W(y_{\theta^*}, y_{\theta^{\text{ref}}}) > d_W(y_{\theta_0^{c_\epsilon}}, y_{\theta^{\text{ref}}})) = 0.5. \tag{10}$$

The difference $\theta^* - \tilde{\theta}^m$ is the shift due to the contamination. Writing $d_W(y_{\theta_0^{c_\epsilon}}, y_{\theta^{\text{ref}}})$ as

$$d_W(y_{\theta_0^{c_\epsilon}}, y_{\theta^{\text{ref}}}) = (1 - \epsilon) d_W(y_{\theta_0}, y_{\theta^{\text{ref}}}) + \epsilon \cdot d_W(c, y_{\theta^{\text{ref}}}),$$

we can rewrite (10) as

$$Pr\left(d_W(y_{\theta^*}, y_{\theta^{\text{ref}}}) > d_W(y_{\theta_0}, y_{\theta^{\text{ref}}}) + \underbrace{\epsilon[d_W(c, y_{\theta^{\text{ref}}}) - d_W(y_{\theta_0}, y_{\theta^{\text{ref}}})]}_{\Delta} \right) = 0.5.$$

Table 7
Average confidence medians: true $\theta_0 = 1$..

n	ϵ	$\theta^{\text{ref}} = 3$		$\theta^{\text{ref}} = 4$		$\theta^{\text{ref}} = 6$		$\theta^{\text{ref}} = 8$		$\theta^{\text{ref}} = 10$	
		KS	W	KS	W	KS	W	KS	W	KS	W
20	0.00	1.02	1.03	1.01	1.04	1.41	1.06	3.96	1.08	5.04	1.10
	0.05	1.19	0.99	1.35	1.20	2.37	1.61	4.21	1.86	6.07	2.26
	0.10	1.35	1.01	1.61	1.21	2.75	1.60	4.35	1.86	6.11	2.22
	0.15	1.46	1.02	1.70	1.20	2.77	1.60	4.35	1.86	6.11	2.21
	0.20	1.46	1.03	1.69	1.20	2.76	1.60	4.35	1.86	6.11	2.22
100	0.00	1.06	1.07	1.08	1.09	1.55	1.17	3.97	1.38	5.02	1.76
	0.05	1.23	1.07	1.41	1.31	2.46	1.76	4.35	2.18	6.28	2.82
	0.10	1.38	1.09	1.67	1.29	2.82	1.71	4.50	2.15	6.31	2.80
	0.15	1.48	1.10	1.75	1.30	2.84	1.71	4.50	2.15	6.31	2.79
	0.20	1.49	1.10	1.74	1.30	2.84	1.71	4.50	2.14	6.31	2.78

Table 8
Average ABC-posterior medians: true $\theta_0 = 1$.

n	ϵ	$\theta^{\text{ref}} = 3$		$\theta^{\text{ref}} = 4$		$\theta^{\text{ref}} = 6$		$\theta^{\text{ref}} = 8$		$\theta^{\text{ref}} = 10$	
		KS	W	KS	W	KS	W	KS	W	KS	W
20	0.00	1.38	1.04	1.40	1.11	1.60	1.50	2.02	2.00	2.51	2.51
	0.05	1.05	1.16	1.13	1.21	1.51	1.52	2.00	2.00	2.51	2.51
	0.10	1.12	1.29	1.18	1.32	1.51	1.55	2.00	2.00	2.51	2.51
	0.15	1.19	1.41	1.24	1.46	1.53	1.63	2.00	2.01	2.51	2.51
	0.20	1.28	1.55	1.31	1.61	1.56	1.75	2.00	2.05	2.51	2.51
100	0.00	1.33	1.00	1.34	1.05	1.52	1.50	2.00	2.00	2.51	2.51
	0.05	1.00	1.14	1.05	1.16	1.50	1.50	2.00	2.00	2.51	2.51
	0.10	1.07	1.27	1.10	1.30	1.50	1.51	2.00	2.00	2.51	2.51
	0.15	1.15	1.40	1.17	1.44	1.50	1.57	2.00	2.00	2.51	2.51
	0.20	1.23	1.52	1.25	1.57	1.50	1.68	2.00	2.01	2.51	2.51

As the term $\Delta \to 0$, the confidence median is recovered. In particular this happens in the trivial case, when $\epsilon \to 0$ or if θ^{ref} minimizes $d_W(c, y_{\theta^{\text{ref}}}) - d_W(y_{\theta_0}, y_{\theta^{\text{ref}}})$, that means it parametrizes the model which corresponds to the barycenter between the central one and the model that generates the contamination. The optimal value cannot be known in advance, but as an initial guess a nonrobust estimate could be considered.

To investigate the impact of the reference parameters on the stability of the confidence distribution, we perform a simulation study, aiming at comparing inference with ABC-posterior approximations and confidence distributions. We consider different reference parameters ($\theta^{\text{ref}} = 3$, $\theta^{\text{ref}} = 4$, $\theta^{\text{ref}} = 6$, $\theta^{\text{ref}} = 8$, $\theta^{\text{ref}} = 10$), different percentages of contaminated data ($\epsilon = 0\%$, $\epsilon = 5\%$, $\epsilon = 10\%$, $\epsilon = 15\%$, $\epsilon = 20\%$), and two sample sizes ($n = 20,\ n = 100$). The central model is the Gaussian $N(1, 1)$ and the contamination is drawn from a Cauchy distribution. For each scenario, we consider 1000 simulations, proposing for each $R = 1000$ values from the Uniform $[0, \theta^{\text{ref}}]$. Note that the choice of the reference parameter has an impact in both the ABC and CD-based procedures. In ABC, this is used as the right endpoint of the Uniform proposal distribution, thus, once fixed the proportion of accepted values equal to 10%, the sequence of distributions obtained with higher θ^{ref} becomes more and more diffuse, as the proposal, with a consequent location-shift. In the approach based on CDs, the reference parameter relates to the position of the confidence median. Tables 7 and 8 display the averaged point estimators (confidence medians and approximate-posterior medians). Simulations are highly stable regardless the choice of the sample size. When non contaminated samples are considered, the Wasserstein-confidence medians remain close in average to the nominal value for any choice of the reference parameter (from 1.03 to 1.10 when $n = 20$, from 1.06 to 1.76 when $n = 100$). In contrast, Wasserstein-ABC medians are more sensitive to the proposal distribution (Wasserstein: from 1.38 to 2.51 when $n = 20$, from 1.33 to 2.51 when $n = 100$). When considering misspecification, the reference parameter emerges as the object that controls the breakdown point in CD-based estimators. Confidence medians constructed with $\theta^{\text{ref}} = 3$ are highly stable even with 20% of contaminated data (average medians from 0.99 to 1.03 when $n = 20$ and from 1.07 to 1.10 when $n = 100$). Instead, KS distance is less resistant to the contamination impact. The results are similar if compared to the ABC-approach (Table 8).

6. Discussion

Robust statistics is an extension of classical parametric statistics that specifically takes into account the fact that the assumed parametric models used by the researchers are only approximate. The concept of robustness has been widely discussed in the frequentist literature. In this paper, we refer to the robustness approach based on the influence function. The contribution of this paper is to fill the gap between robust inference and confidence distributions analyses. Indeed, in practical applications, CDs are more informative than a simpler confidence interval or a p-value, since they describe the

complete distribution estimator for the parameter of interest, as the posterior distribution for Bayesians. In particular, CDs allow to compute measures of evidence for statements of the type "$\psi > \psi_0$" or "$\psi_1 \leq \psi \leq \psi_2$", which are of particular interest in many real data applications, as for instance the one considered in the paper on non-inferiority testing. Remark that the application to non-inferiority and superiority trials discussed here can be easily extended to the comparison of other parameters than means, such as odds ratios, hazard ratios, etc., where stability of inferential conclusions with respect to model misspecification is essential. Indeed, the proposed procedure can be applied to any parametric statistical model $f(y;\theta)$, even in the presence of censored data (Deléamont and Ronchetti, 2022), for which robust estimation can be performed, for instance, through the Tsallis score (Ruli et al., 2022) or through algorithms for bounded-influence estimation (Bellio, 2007).

The derivation of robust CDs discussed in Section 3.2, based on a frequentist reinterpretation of ABC techniques to obtain a normalized pseudo-likelihood function for the parameter of interest, represents a practical alternative to other robust pseudo-likelihood functions, such as the empirical likelihood or the quasi-likelihood (see e.g. Greco et al., 2008, and references therein). Indeed, the aforementioned functions may present some drawbacks: the empirical likelihood is not computable for small sample sizes and quasi-likelihoods can be easily obtained only for scalar parameters. On the contrary, the approach that directly incorporates robust estimating functions into ABC techniques, with respect to available approaches based on pseudo-likelihoods, can be computationally faster when the evaluation of the estimating function is expensive, can be computed even for small sample sizes and for multidimensional parameters of interest, and the derived normalized pseudo-likelihood has the right curvature (see Ruli et al., 2020). Obviously, the proposed method can be applied to any unbiased robust estimating equations, such as *S*-estimating equations.

Related to the above comment, we highlight that for a scalar parameter of interest in the presence of nuisance parameters we propose to modify the algorithm of Ruli et al. (2020) by using the a profile estimating equation and plugging in the value of proposals for nuisance parameters used to generate pseudo-data. The treatment of nuisance parameters is a theme of ever-renewed attention in frequentist inference. An interesting extension of our proposal, not only limited to robust inference but for general estimating functions, is the further study of ways to handling nuisance parameters with CD-based inference, in particular with the options available with the bootstrap (see DiCiccio and Romano, 1988) and with ABC techniques based on general profile-type estimating functions.

As a final remark, we mention the possibility to adopt non parametric criteria and statistics, other than just centrality measures, for deriving confidence distributions for a scalar parameter of interest in presence of contamination. A central parametric model is assumed but the observed data are evaluated in terms of non parametric pseudo-distances from a reference model, directly based on the empirical cumulative distribution functions. The simulation-based confidence distribution approach presented in this paper can be applied to derive robust estimates based on these pseudo-distances. At present, our preliminary study of this class of methods is limited to models with a scalar parameter of interest, since adapting these kind of test procedures to more complex models to deal with nuisance parameters may require information about the context, situation-dependent considerations and also give rise to confidence curves instead of confidence distributions.

Acknowledgments

The authors acknowledge the Associate Editor and the two anonymous Referees for the useful comments that greatly improved the paper.

References

Barndorff-Nielsen, O.E., Cox, D.R., 1994. Inference and Asymptotics. CRC Press.
Bee, M., Benedetti, R., Espa, G., 2017. Approximate maximum likelihood estimation of the bingham distribution. Computational Statistics & Data Analysis 108, 84–96.
Bellio, R., 2007. Algorithms for bounded-influence estimation. Computational Statistics & Data Analysis 51, 2531–2541.
Bernton, E., Jacob, P.E., Gerber, M., Robert, C.P., 2019. On parameter estimation with the wasserstein distance. Information and Inference: A Journal of the IMA 8, 657–676.
Bortolato, E., Ventura, L., 2022. Confidence distributions and fusion inference for intractable likelihoods. In: Book of short papers SIS 2022. Caserta. 2022, To appear
Brazzale, A.R., Davison, A.C., Reid, N., 2007. Applied asymptotics. Case-studies in small sample statistics. Cambridge University Press.
Carhart-Harris, R., Giribaldi, B., Watts, R., Baker-Jones, M., Murphy-Beiner, A., Murphy, R., Nutt, D.J., 2021. Trial of psilocybin versus escitalopram for depression. New England Journal of Medicine 384, 1402–1411.
Chen, X., Zhou, W., 2020. Robust inference via multiplier bootstrap. The Annals of Statistics 48, 1665–1691.
D'Agostino, R.B., Massaro, J.M., Sullivan, L.M., 2003. Non- inferiority trials: design concepts and issues the encounters of academic consultants in statistics. Statistics in Medicine 22, 169–186.
Dawid, A.P., Musio, M., Ventura, L., 2016. Minimum scoring rule inference. Scand. J. Statist. 43, 123–138.
Deléamont, P.Y., Ronchetti, E., 2022. Robust inference with censored survival data. Scand. J. Statist. doi:10.1111/sjos.12570. In press
DiCiccio, T.J., Efron, B., 1996. Bootstrap confidence intervals. Statistical science 11, 189–228.
DiCiccio, T.J., Romano, J.P., 1988. A review of bootstrap confidence intervals. Journal of the Royal Statistical Society: Series B (Methodological) 50, 338–354.
Efron, B., 1979. Bootstrap methods: another look at the jackknife. The Annals of Statistics 7, 1–26.
Farcomeni, A., Ventura, L., 2012. An overview of robust methods in medical research. Statistical Methods in Medical Research 21, 111–133.
Field, C.A., Ronchetti, E., 1990. Small Sample Asymptotics. IMS Monograph Series, Hayward (CA).
Frazier, D.T., Robert, C.P., Rousseau, J., 2020. Model misspecification in approximate bayesian computation: consequences and diagnostics. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 82, 421–444.
Garret, A.D., 2003. Therapeutic equivalence: fallacies and falsification. Statistics in Medicine 22, 741–762.

Ghosh, M., Basu, A., 2013. Robust estimation for independent non-homogeneous observations using density power divergence with applications to linear regression. Electronic Journal of Statistics 7, 2420–2456.
Greco, L., Racugno, W., Ventura, L., 2008. Robust likelihood functions in bayesian inference. Journal of Statistical Planning and Inference 138, 1258–1270.
Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., Stahel, W.A., 1986. Robust Statistics. The Approach Based on Influence Functions. Wiley.
Heritier, S., Cantoni, E., Copt, S., Victoria-Feser, M.P., 2009. Robust Methods in Biostatistics. Wiley.
Heritier, S., Ronchetti, E.M., 1994. Robust bounded-influence tests in general parametric models. Journal of the American Statististical Association 89, 897–904.
Hjort, N.L., Schweder, T., 2018. Confidence distributions and related themes. Journal of Statististical Planning and Inference 195, 1–13.
Huber, P.J., 1981. Robust Statistics. Wiley, New York.
Huber, P.J., Ronchetti, E.M., 2009. Robust Statistics. Wiley, New York.
Legramanti, S., Durante, D., Alquier, P., 2022. Concentration and robustness of discrepancy-based ABC via rademacher complexity. arXiv:2206.06991.
Lyddon, S.P., Holmes, C.C., Walker, S.G., 2019. General bayesian updating and the loss-likelihood bootstrap. Biometrika 106, 465–478.
Muller, A., 1997. Integral probability metrics and their generating classes of functions. Advances in Applied Probability 29, 429–443.
Nayak, S., Bari, B.A., Yaden, D.B., Spriggs, M.J., Rosas, F., Peill, J.M., Carhart-Harris, R., 2022. A bayesian reanalysis of a trial of psilocybin versus escitalopram for depression. PsyArXiv preprint.
Newton, M.A., Raftery, A.E., 1994. Approximate bayesian inference with the weighted likelihood bootstrap. Journal of the Royal Statistical Society: Series B (Methodological) 56, 3–26.
Reid, N., 2003. The 2000 wald memorial lectures: asymptotics and the theory of inference. Annals of Statistics 31, 1695–1731.
Ronchetti, E., 2020. Accurate and robust inference. Econometrics and Statistics 14, 74–88.
Ronchetti, E., 2021. The main contributions of robust statistics to statistical science and a new challenge. Metron 79, 127–135.
Ronchetti, E., Ventura, L., 2001. Between stability and higher-order asymptotics. Statistics and Computing 11, 67–73.
Rothmann, M.D., Wiens, B.L., Chan, I.S.F., 2012. Design and Analysis of Non-Inferiority Trials. CRC Press.
Rubio, F.J., Johansen, A.M., 2013. A simple approach to maximum intractable likelihood estimation. Electronic Journal of Statistics 7, 1632–1654.
Ruli, E., Sartori, N., Ventura, L., 2016. Approximate bayesian computation with composite score functions. Statistics and Computing 26, 679–692.
Ruli, E., Sartori, N., Ventura, L., 2020. Robust approximate bayesian inference. Journal of Statistical Planning and Inference 205, 10–22.
Ruli, E., Ventura, L., 2021. Can bayesian, confidence distribution and frequentist inference agree? Statistical Methods & Applications 30, 359–373.
Ruli, E., Ventura, L., Musio, M., 2022. Robust confidence distributions from proper scoring rules. Statistics 56, 455–478.
Schweder, T., Hjort, N.L., 2016. Confidence, Likelihood, Probability: Statistical Inference with Confidence Distributions. Cambridge University Press.
Severini, T.A., 2000. Likelihood methods in statistics. Oxford University Press.
Soubeyrand, S., Haon-Lasportes, E., 2015. Weak convergence of posteriors conditional on maximum pseudo-likelihood estimates and implications in ABC. Statistics and Probability Letters 107, 84–92.
Thornton, S., Li, W., Xie, M., 2022. Approximate confidence distribution computing. arXiv:2206.01707.
Tsallis, C., 1988. Possible generalization of boltzmann-gibbs statistics. Journal of Statistical Physics 52, 479–487.
Varin, C., Reid, N., Firth, D., 2011. An overview of composite likelihood methods. Statistica Sinica 21, 5–42.
Ventura, L., Racugno, W., 2016. Pseudo-likelihoods for bayesian inference. In: Topics on methodological and applied statistical inference, series studies in theoretical and applied statistics. Springer-Verlag, pp. 205–220.
Xie, M., Singh, K., 2013. Confidence distribution, the frequentist distribution estimator of a parameter: a review. International Statistical Review 81, 3–39.