Statistical Papers (2024) 65:1233–1257
https://doi.org/10.1007/s00362-023-01445-w

REGULAR ARTICLE

A unifying framework for rank and pseudo-rank based inference using nonparametric confidence distributions

Jonas Beck[1] · Arne C. Bathke[1]

Received: 25 May 2022 / Revised: 10 February 2023 / Published online: 6 April 2023

Abstract

Nonparametric confidence distributions estimate statistical functionals by a distribution function on the parameter space, instead of the classical point or interval estimators. The concept bears analogy to the Bayesian posterior, but is nevertheless a completely frequentist concept. In order to ensure the desired statistical properties, we require that the cumulative distribution function on the parameter space is, evaluated at the true parameter, uniformly distributed over the unit interval. Our main focus lies on developing confidence distributions for the nonparametric relative effect and some natural extensions thereof. We develop asymptotic, range preserving and—especially important in the case of small sample sizes—approximate confidence distributions based on rank and pseudo-rank procedures. Due to the close relationship between point estimators, confidence intervals and p-values, these can all be approached in a unified manner within the framework of confidence distributions. The main goal of our contribution is to make the powerful theory of confidence distributions available in a nonparametric context, that is, for situations where methods relying on parametric assumptions are not justifiable. Application of the proposed methods and interpretation of the results is demonstrated using real data sets, including ordinal, non-metric data.

Keywords Confidence distribution · Non-parametric statistical inference · Rank statistic · Unweighted relative effect · Pseudo-rank statistic

✉ Jonas Beck
jonas.beck@plus.ac.at

Arne C. Bathke
arne.bathke@plus.ac.at

[1] Department of Artificial Intelligence and Human Interfaces, Paris-Lodron-University of Salzburg, Hellbrunner Straße, 5020 Salzburg, Austria

1 Introduction

The key idea underlying confidence distributions (CD) is to use a distribution function on the parameter space in order to estimate a certain parameter—in a way analogue to using a Bayesian posterior distribution, but without employing a prior distribution. Thus, confidence distributions are a completely frequentist concept, but they do carry the possibility to be a unifier for concepts from Bayesian, Frequentist, and Fiducial statistics (Thornton and Xie 2020). The major advantage of a confidence distribution is that it contains a wealth of information (like a posterior) for a parameter of interest compared to the commonly used point estimates, confidence intervals, or p-values.

Due to the close relationship between the two latter concepts, they can also be unified under the framework of confidence distributions or even be derived using CD.

As stated in the ICH E9 Guideline "Estimates of treatment effects should be accompanied by confidence intervals" (The International Council for Harmonisation 1998), decisions based on confidence intervals and the corresponding hypothesis tests should coincide.

In order to ensure that this is the case, confidence distributions provide a natural and simple solution. In addition to computing the confidence distribution (or the p-value function), it is technically not even necessary to specify a concrete null hypothesis or a confidence level in advance, which could allow for large flexibility in practice, but requires much care in interpreting the results.

1.1 Motivating example

Let us give an example demonstrating why the unifying framework provided by confidence distributions is useful in the context of rank and pseudo-rank procedures. In this example, undesired toxic effects on male Wistar rats were examined for a drug that was administered in four increasing dose levels. In the trial, $n_0 = 8$ rats received the placebo and $n_1 = 7$, $n_2 = 8$, $n_3 = 7$, and $n_4 = 8$ animals received the respective increasing dose levels of the drug. The relative liver weight (as percentage of body weight) is the primary endpoint.

For this ratio variable, the normality assumption or a classical shift effect model do not seem justifiable. Thus, the data are analyzed based on a nonparametric model framework, with methods employing (pseudo-)ranks. Here, the following research question should be answered.

1. What is an estimate of the relative effect $\theta := P(X_1 < X_2) + \frac{1}{2}P(X_1 = X_2)$ for each drug level against the placebo?
2. How variable are the estimators, and what are the confidence intervals for the relative effect θ at difference confidence levels?
3. Is there a significant effect on the relative liver weights at each dose level of the drug, and what are the associated p-values?

The major advantage of using a nonparametric confidence distribution approach as presented in this paper is that we can answer these questions basically within just one graph (and the associated quantitative information).

1.2 Review on confidence distributions and rank- and pseudo-rank-based procedures

The concept captured by CD has quite a long history, dating back to Fisher‘s fiducial distribution (Fisher 1935), and the first usage of the term “confidence distribution” by Cox (1958). However, CD were still rarely used till their rediscovery by Singh et al. (2005) and Schweder and Hjort (2002). Many of the recent applications of confidence distribution theory are in the field of fusion learning (e.g. Liu et al. 2014, 2015; Shen et al. 2020; Cai et al. 2021; Liu et al. 2021), prediction (e.g. Shen et al. 2018; Xie and Zheng 2021; Tian et al. 2021), and fields such as Fisher randomization tests (Luo et al. 2021) and survival studies (Tian et al. 2011).

Despite their colorful history and the well demonstrated usefulness of confidence distributions along with their many applications, it appears that up to now most published research on confidence distributions is still within the area of parametric statistics. The goal of the present paper is to change this and to develop results for confidence distributions in a nonparametric context, so that we can use the well developed theory for confidence distributions without having to make many assumptions on the data.

To this end, we consider the statistical functional $\theta := P(X_1 < X_2) + \frac{1}{2}P(X_1 = X_2)$, the *nonparametric relative effect* or *probabilistic index*, which goes back to the classical Mann–Whitney–Wilcoxon Test (Wilcoxon 1945; Mann and Whitney 1947), and we will determine the corresponding (asymptotic) confidence distribution. In order to do this, we make use of the theory developed by Brunner et al. (2018) and the references therein.

One of the major advantages of this functional is its flexible use. The nonparametric relative effect θ requires very few assumptions on the data, and accordingly it is possible to use the resulting methods for metric, as well as ordinal data, as we will see in the following. Consequently, we obtain a much more robust confidence distribution than the one based on parametric statistical models. Indeed, due to their generality and wide applicability, related rank-based nonparametric methods have been receiving much popularity in practice, in particular also in the life sciences (Zimmermann et al. 2019).

In the following, we will develop confidence distributions for this nonparametric two sample relative effect and some natural extensions thereof. They meet the desired conditions asymptotically, and in one case also approximately, which is especially important in the case of small sample sizes. As it is quite clear that the relative effect can only take values in [0, 1], we would also want our confidence distribution to place all its probability mass within this theoretically possible range. In order to guarantee this, we are proposing to use range preserving confidence distributions in situations where the estimated effect is close to the interval boundaries. This follows a similar idea as Efron‘s range preserving confidence interval (Efron and Tibshirani 1994).

Because of the generality of our confidence distribution approach, we can expand it to the case of more than two samples. However, in the case of three or more groups, classical pairwise ranking methods can produce quite paradoxical results. This is due to the fact that they lack an important transitivity property as is shown by Thangavelu and Brunner (2007) and Brunner et al. (2021). A solution to this problem is using the unweighted mean of the distribution functions $F_1, \ldots, F_d$ $G := \frac{1}{d}\sum_{i=1}^{d} F_i$ and comparing each of the d distributions to this reference distribution (Brunner et al. 2017). This strategy also solves issues arising for weighted mean distributions in unbalanced designs that have recently resulted in the development of procedures that are based on pseudo-ranks instead of ranks (Brunner et al. 2021).

The remainder of this paper is organized as follows. Section 2 gives a short review of the theory of confidence distributions and their usefulness in many fields of statistical inference. Additionally, we give an example of a confidence distribution for the classical parametric Behrens–Fisher problem. Based on this consideration, we continue in Sect. 3 with a confidence distribution for the nonparametric relative effect. Specifically, we develop asymptotic, approximate, and range preserving versions of the CD. In the fourth section, we extend this to the case of three or more samples by using procedures based on pseudo-ranks. In Sect. 5 we evaluate the performance of the derived confidence distributions using an extensive simulation study. In the sixth section we apply our approach to the toxic effect study on Wistar rats. We conclude with a short summary and ideas for future research.

2 Confidence distributions

In this section, we provide a short summary regarding the concept of confidence distributions. For further reading, see for example Xie and Singh (2013) and Schweder and Hjort (2016).

After the definition and an explanation of the main statistical properties, we illustrate the concept by giving a confidence distribution for the parametric Behrens–Fisher Problem.

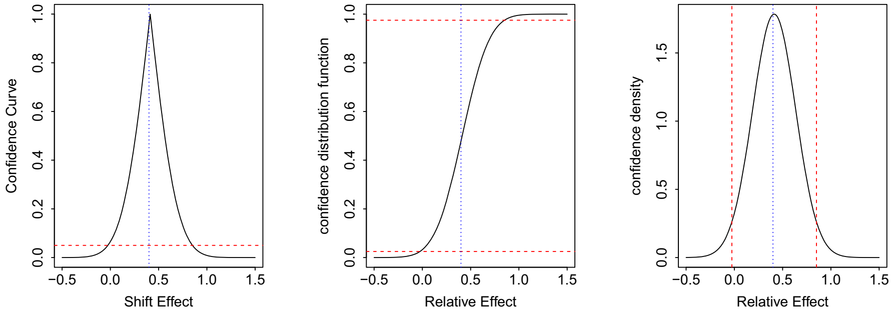
Let us start by defining confidence distributions, using the following definition which was first proposed by Singh et al. (2005).

Definition 1 A function $H_n(\cdot) = H_n(X, \cdot)$ on $\mathcal{X} \times \Theta \to [0, 1]$ is called a *confidence distribution* (CD) for a parameter θ, if the following two conditions hold.

(i) For each given $X \in \mathcal{X}$, $H_n(\cdot)$ is a cumulative distribution function on Θ.
(ii) At the true parameter value $\theta = \theta_0$, $H_n(\theta_0) = H_n(x, \theta_0)$, as a function of the sample value x, follows the uniform distribution $U[0, 1]$.

As convenient variations to this definition, the function $H_n(\cdot)$ is called an *asymptotic confidence distribution* or an *approximate confidence distribution* if the second requirement only holds asymptotically or approximately, respectively. When it exists, we call $h_n(\theta) = H_n'(\theta)$ a *confidence density*. A CD function can be expressed alternatively as a *confidence curve* or *p-value function*: $CV_n(\theta) = 1 - 2|H_n(\theta) - 0.5| = 2\min\{H_n(\theta), 1 - H_n(\theta)\}$

In the definition of a confidence distribution, the first assumption ensures that H_n can be a data-dependent “distribution estimator" of the parameter, and the second one guarantees its desired statistical properties. Based on H_n, one can easily obtain one- and two-sided confidence intervals at each desired level. Indeed, they are simply given by $(-\infty, H_n^{-1}(1-\alpha)],\ [H_n^{-1}(\alpha), \infty)$, and $[H_n^{-1}(\alpha/2), H_n^{-1}(1-\alpha/2)]$.



(a) Confidence Curve

(b) Confidence Distribution

(c) Confidence Density

Fig. 1 Confidence curve, confidence distribution function, and confidence density function for the shift effect in a two-sample problem

The confidence curve is a very useful graphical tool to visualize the main inferential evidence regarding a parameter. Every value on the y-axis, which indicates the respective α-level, corresponds to two points on the curve representing the equal-tailed $1-\alpha$ confidence interval for the parameter under consideration.

From the confidence distribution H_n, one can also derive point estimators in a straightforward way. Similar to the way that estimators are constructed based on (Bayesian) posterior distributions, sensible point estimators based on a (frequentist) confidence distribution are, for example, the median $M_n = H_n^{-1}(1/2)$, the mean $\bar{\theta}_n = \int_{-\infty}^{\infty} t dH_n(t)$, and the mode $\hat{\theta}_n = \arg\max h_n(\theta)$ of the CD. It can be proven that these point estimators are consistent under mild conditions (see Theorem 3.1−3.3 in Singh et al. (2007)).

For the usage of confidence distributions in hypothesis testing Singh et al. (2007) introduced two measures of support for a null hypothesis $K_0 : \theta \in C$ versus the alternative $K_1 : \theta \in C^c$: *Strong support* is defined as $p_s(C) := \int_C dH_n$, while *weak support* is given by $p_w(C) := \sup_{\theta \in C} 2\min\{H_n(\theta), 1 - H_n(\theta)\}$. Strong support lends itself for use in case of interval type null hypotheses, while weak support may be used in simple (singleton) null hypotheses. If the null hypothesis is of the type $(-\infty, H_n^{-1}(1-\alpha)]$ or $[H_n^{-1}(\alpha), \infty)$ it can be shown (Theorem 3.5 in Singh et al. (2007)) that strong support agrees with classical p-values. A similar result can be obtained for the weak support and the p-value in the singleton case. Thus, there is a close connection to the concept of p-value functions (Fraser 1991). Let us now consider an illustrative example.

Example 1 We have two independent normally distributed samples $X_{1,k}$ and $X_{2,k}$, $k = 1, \ldots, n$ with mean μ_1 and μ_2. For now, assume that they both have variance σ^2, and we define $\delta = \mu_1 - \mu_2$ and $\hat{\delta} = \bar{X}_{1,n} - \bar{X}_{2,n}$. In the case of a known σ, a CD for δ is given by $H_n(\delta) = \Phi(\frac{\delta - \hat{\delta}}{\sqrt{2}\sigma/\sqrt{n}})$, a confidence curve $CV_n(\delta) = 2\min(\Phi(\frac{\delta - \hat{\delta}}{\sqrt{2}\sigma/\sqrt{n}}), 1 - \Phi(\frac{\delta - \hat{\delta}}{\sqrt{2}\sigma/\sqrt{n}}))$, and a confidence density $h_n(\delta) = \frac{1}{2\sqrt{\pi\sigma^2/n}} \exp(-\frac{(\delta - \hat{\delta})^2}{4\sigma^2/n})$. In Fig. 1a, b, simulated versions of these functions are shown for $\mu_1 = 0.4$, $\mu_2 = 0$, and $\sigma = 1/\sqrt{2}$.

In all three figures, the red line indicates an equal-tailed 95% confidence interval for the shift effect δ, and the vertical blue line gives the true value of δ.

For unknown σ, we have $H_n(\delta) = F_{t_{2n-2}}(\frac{\delta-\hat{\delta}}{\sqrt{2}s/\sqrt{n}})$ as CD for δ, where $F_{t_{2n-2}}$ is the cdf of a t-distribution with $2n-2$ degrees of freedom and s the sample standard deviation.

For different sample sizes n_1 and n_2 and different variances σ_1 and σ_2 (the classical parametric Behrens–Fisher Problem), we do have an approximate CD. Here, the denominator of H_n changes to $\sqrt{s_1^2/n_1 + s_2^2/n_2}$ and the degrees of freedom estimator has the well-known form

$$\hat{f} = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1-1) + (s_2^2/n_2)^2/(n_2-1)},$$

where s_1^2 and s_2^2 are the unbiased estimators of the respective population variance.

3 Confidence distribution for the nonparametric relative effect in two samples

Most literature on confidence distributions assumes parametric models, and there are only very few works considering the concept of confidence distributions outside a parametric context (Chapter 11 in Schweder and Hjort (2016), Thornton and Xie (2020), Liu et al. (2021)). Among these is a generalization of Example 1 that consists of dropping the normality assumption and thus considering the location shift effect in a more flexible, semiparametric way. For the resulting confidence distribution, see Schweder and Hjort (2016) and Thornton and Xie (2020).

In the present section, we are extending the scope of confidence distributions towards an even more general framework for the comparison of two samples, and with as few assumptions on the respective underlying distributions F_1 and F_2 as possible. Inference will be based on the nonparametric relative effect which is defined as follows. For two independent random variables $X_1 \sim F_1$ and $X_2 \sim F_2$, the probability

$$\theta := P(X_1 < X_2) + \frac{1}{2}P(X_1 = X_2) = \int F_1 dF_2$$

is called the *non-parametric relative effect* of X_2 with respect to X_1. It can be calculated without assuming a specific parametric model or a location shift effect. Indeed, it may even be calculated for ordinal data, thus providing a substantial generalization compared to the normal distribution model considered in Example 1. However, in order to develop an intuition for this functional, it is instructive to calculate the nonparametric relative effect within the framework of such a parametric example. In the concrete situation of Example 1, the relative effect can be written as

$$\theta = \Phi\left(\frac{\mu_1 - \mu_2}{\sqrt{\sigma_1^2 + \sigma_2^2}}\right). \tag{1}$$

From the general definition of θ it follows that if $\theta < 1/2$ the observations from F_1 tend to have greater values than those from F_2, and vice versa for $\theta > 1/2$. As one can clearly see in the previous parametric example, a value of $\theta = 1/2$ does not imply that X_1 and X_2 are identically distributed. Indeed, in the Behrens–Fisher problem with normal distributions, only the means must be identical for $\theta = 1/2$ to hold, but not the variances. A value of $\theta = 1/2$ may be interpreted as no tendency to larger or smaller values when comparing one distribution against the other.

For the estimation of the relative effect and an evaluation of the (asymptotic) properties of the estimator, we take advantage of the theory developed by Brunner and Munzel (2000) and described, for example, by Brunner et al. (2018). Let us assume a statistical model with two independent samples, $X_{ik} \sim F_i$ $i = 1, 2, k = 1, \ldots, n_i$ which contain n_i independent and identically distributed random variables. Further, assume that $N/n_i \leq N_0 < \infty$ for large N, and the variances $\sigma_i^2 = Var(F_j(X_{i1})), i, j = 1, 2, i \neq j$ satisfy $\sigma_i^2 > 0, i = 1, 2$. In other words, for technical reasons, we exclude here the situation that the distributions are fully separated or that one of them is concentrated in one point only. In order to state the main results, we also need the following notation. We call

$$R_{ik} = \frac{1}{2} + \sum_{j=1}^{2} \sum_{l=1}^{n_j} c(X_{ik} - X_{jl})$$

the *rank* (or *overall rank*) of X_{ik} among all $N = n_1 + n_2$ observations, where $c(u) = 0, 1/2, 1$ for $u <, =, > 0$,

$$R_{ik}^{(i)} = \frac{1}{2} + \sum_{l=1}^{n_j} c(X_{ik} - X_{jl})$$

the *internal rank* of X_{ik} among the n_i observations $X_{i1}, \ldots, X_{in_i}$ and $\bar{R}_{i\cdot} = \frac{1}{n_i} \sum_{i=1}^{ni} R_{ik}$ for $i = 1, 2$ the *rank means*. Then, an unbiased and consistent estimator (Proposition 3.1 in Brunner and Puri (2002)) for the relative effect $\theta = P(X_1 < X_2) + \frac{1}{2} P(X_1 = X_2) = \int F_1 dF_2$ is

$$\begin{aligned}\hat{\theta} &= \frac{1}{n_1}\Big(\bar{R}_{2\cdot} - \frac{n_2+1}{2}\Big) = 1 - \frac{1}{n_2}\Big(\bar{R}_{1\cdot} - \frac{n_1+1}{2}\Big) \\ &= \frac{1}{N}(\bar{R}_{2\cdot} - \bar{R}_{1\cdot}) + \frac{1}{2}.\end{aligned}$$

3.1 Asymptotic confidence distribution for the relative effect

In order to derive an asymptotic confidence distribution for the nonparametric relative effect, we first decompose

$$\sqrt{N}(\hat{\theta} - \theta) = \sqrt{N}\Big(\int \hat{F}_1 d\hat{F}_2 - \int F_1 dF_2\Big) = U_N + C_N, \tag{2}$$

where

$$U_N = \sqrt{N}\left(\frac{1}{n_2}\sum_{k=1}^{n_2} F_1(X_{2k}) - \frac{1}{n_1}\sum_{k=1}^{n_1} F_2(X_{1k}) + 1 - 2\theta\right)$$

and

$$C_N = \sqrt{N}\int (\hat{F}_1 - F_1)d(\hat{F}_2 - F_2),$$

which can easily be verified. Brunner and Munzel (2000) has proved the following theorem by invoking the *Asymptotic Equivalence Theorem* (Theorem 2.2 in Akritas and Brunner (1997)):

Theorem 1 *Let $X_{ik} \sim F_i, i = 1, 2, k = 1, \ldots, n_i$ be independent random variables. Additionally $N = n_1 + n_2 \to \infty$ and $N/n_i \leq N_0 < \infty$, $i = 1, 2$. Then $\sqrt{N}(\hat{\theta} - \theta)$ and U_N are asymptotically equivalent.*

Since X_{1k} and X_{2k} are independent, the asymptotic normality of U_N was shown by Brunner and Munzel (2000) using an appropriate central limit theorem. Therefore we have $U_N/\sigma_N \xrightarrow{P} N(0, 1)$, where

$$\sigma_N^2 = \mathrm{Var}(U_N) = \frac{N}{n_1 n_2}\left(n_1\sigma_2^2 + n_2\sigma_1^2\right).$$

The variances σ_1^2 and σ_2^2 are unknown and must be estimated. As X_{1k} and X_{2k} are independent, also $F_2(X_{1k})$ and $F_1(X_{2k})$ are independent. If they were observable, an unbiased and consistent estimator for σ_i would be

$$\tilde{\sigma}_i^2 = \frac{1}{n_i - 1}\sum_{k=1}^{n_i}(F_i(X_{ik}) - \frac{1}{n_i}\sum_{k=1}^{n_i} F_i(X_{ik}))^2. \tag{3}$$

However, as these are not observable, the so-called normed placements are used instead:

$$n_2\hat{F}_2(X_{1k}) = R_{1k} - R_{1k}^{(1)} \text{ and } n_1\hat{F}_1(X_{2k}) = R_{2k} - R_{2k}^{(2)}. \tag{4}$$

Using them, the variance can be estimated by $\hat{\sigma}_N^2 = \frac{N}{n_1 n_2}\left(\frac{S_1^2}{n_2} + \frac{S_2^2}{n_1}\right)$ with

$$S_i^2 = \frac{1}{n_i - 1}\sum_{k=1}^{n_i}\left(R_{ik} - R_{ik}^{(i)} - \bar{R}_{i\cdot} + \frac{n_i + 1}{2}\right)^2 \tag{5}$$

for $i = 1, 2$. Theorem 7.24 in Brunner et al. (2018) proves that this estimator is L_2-consistent. To simplify our notation, we define

$$\hat{\tau}_\theta = \frac{1}{\sqrt{N}}\hat{\sigma}_N = \frac{1}{n_1 n_2}\sqrt{\sum_{j=1}^{2} n_j S_j^2}.$$

Then, based on the above considerations, an asymptotic confidence distribution for the nonparametric relative effect θ is given by

$$H_n^{\Phi}(\theta) = \Phi\Big(\frac{\theta - \hat{\theta}}{\hat{\tau}_\theta}\Big),$$

where Φ is the cumulative distribution function of the standard normal distribution. The corresponding confidence curve (as defined in the previous chapter) is consequently given by $CV_n^{\Phi}(\theta) := 2\min\{H_n^{\Phi}(\theta), 1 - H_n^{\Phi}(\theta)\}$.

3.2 Approximate confidence distribution for the relative effect

For small sample sizes Brunner and Munzel (2000) suggested using quantiles of the t-distribution instead of the normal distribution, in order to improve the test's performance. This idea can be carried over to the concept of confidence distributions, in a similar spirit as in Example 1. However, the result will only be an approximative CD: Even if the original data are normally distributed, the corresponding confidence distribution does not fulfill the uniformity condition exactly. For the derivation, consider again the unobservable variance "estimator"

$$\tilde{\sigma}_N^2 = \frac{N}{n_1 n_2}\sum_{i=1}^{2}(N - n_i)\tilde{\sigma}_i^2,$$

with $\tilde{\sigma}_i^2$ defined as in (3). The distribution of $\tilde{\sigma}_N^2$ can be approximated by a (scaled) χ_f^2/f distribution. Next, substitute again the unobservable terms by the observable normed placements (4).

The idea here is based on the Satterthwaite–Smith–Welch approximation in the classical parametric Behrens–Fisher Problem (see, e.g., Moser and Stevens (1992) for a detailed discussion on this topic). In nonparametric statistic such an approximation was already used for testing by Brunner and Munzel (2000).

In case of normal distributed data (and unknown variance) it is possible to get an exact confidence distribution using a t-distribution with $n - 1$ degrees of freedom (see for example example 1 in Xie and Singh (2013)). This leads to the following considerations: As the absolute quantiles of a t-distribution are always larger than the corresponding absolute quantiles of standard normal distribution, the approximated confidence distribution, based on the t-distribution, is always more conservative than the asymptotic one.

[We have shown in the previous section that under the assumptions of Theorem 1, and if the variances $\sigma_i^2 = Var(F_j(X_{i1})), i, j = 1, 2, i \neq j$ satisfy $\sigma_i^2 > 0, i = 1, 2,$ then the distribution of

$$\sqrt{N}\frac{\theta - \hat{\theta}}{\sigma_N},$$

with σ_N defined as in (2), has asymptotically a standard normal distribution. Due to the consistency of $\hat{\sigma}_N$ for σ_N, the sampling distribution of

$$T_N = \sqrt{N}\frac{\theta - \hat{\theta}}{\hat{\sigma}_N},$$

can be approximated by a standard normal distribution.]

As this one is more or less liberal for small sample sizes, we use the more conservative t-approximation, which converges with increasing sample size to the asymptotic distribution. The degrees of freedom were taken from the parametric Satterthwaite-Smith-Welch approximation:

$$\hat{f} = \frac{\left(\sum_{i=1}^{2} S_i^2/(N - n_i)\right)^2}{\sum_{i=1}^{2}(S_i^2/(N - n_i))^2/(n_i - 1)}. \tag{6}$$

This justification is quite similar to the one of Result 3.22 in Brunner et al. (2018).

Using this we obtain as an approximate confidence distribution

$$H_n^t(\theta) = F_{t_{\hat{f}}}\Big(\frac{\theta - \hat{\theta}}{\hat{\tau}_\theta}\Big),$$

with the degrees of freedom estimator defined in (6). Obviously $\hat{f} \to \infty$, so the approximate confidence distribution converges to the asymptotic confidence distribution. As in the previous section, the corresponding confidence curve is defined by $CV_n^t(\theta) := 2\min\{H_n^t(\theta), 1 - H_n^t(\theta)\}$.

Remark 1 Brunner and Munzel (2000) suggest that a sample size of 10 in each group is needed to get an accurate approximation. The asymptotic results are quite accurate for sample sizes larger than 50. For smaller sample sizes the test is in general too liberal.

3.3 Range preserving confidence distribution

By definition, the relative effect θ can only take values in the interval [0, 1]. However, when we have an estimate $\hat{\theta}$ that is very close to 0 or 1, or when the sample size is rather small, it is possible that the resulting asymptotic or approximate confidence distribution for θ assigns positive probability mass to an area outside the interval [0, 1]. Actually, in our case study example in Sect. 6.1, this occurs for dose levels 3 and 4.

Certainly, it is hard to convey the interpretation of such an inferential result to statistics practitioners, namely how positive confidence can be assigned to sets of impossible parameter values. These concerns will be addressed in this subsection.

Let us call a confidence distribution *range preserving* when the distribution function assigns positive probability mass only to areas that fall within the allowed range—analogously to the concept of range preserving confidence intervals introduced by Efron and Tibshirani (1994).

In order to obtain range preserving confidence distributions, we use the delta-method. Hereby, we transform the open unit-interval (0, 1) which contains all relevant values for the relative effect (ignoring the theoretical possibility of perfectly separated distributions) to $(-\infty, \infty)$ by a transformation-function $g : (0, 1) \to \mathbb{R}$. If the function g has continuous first derivative $g'(\cdot)$ and g' does not vanish at θ, and if additionally the assumptions of Theorem 1 hold, in particular that $N/n_i \leq N_0 < \infty$, $i = 1, 2$, then

$$\frac{\sqrt{N}(g(\hat{\theta}) - g(\theta))}{|g'(\hat{\theta})|\hat{\sigma}_N}$$

is asymptotically $N(0, 1)$ distributed (see, e.g., Chapter 3 in van der Vaart (1998)). We therefore obtain an asymptotic confidence distribution that is range preserving, that is, it assigns all its positive probability mass on the unit interval by

$$H_{n,g}(p) = \Phi\left(\frac{g(\theta) - g(\hat{\theta})}{\hat{\tau}_\theta |g'(\hat{\theta})|}\right).$$

In our example (see Sect. 6.1), we use the logit-function defined by $\text{logit}(x) = \log(x/(1 - x))$ and fulfilling all necessary conditions. Thus, we get an asymptotic confidence distribution for the relative effect by

$$H_{n,\text{logit}}(\theta) = \Phi\left(\frac{\text{logit}(\theta) - \text{logit}(\hat{\theta})}{\hat{\tau}_\theta/\hat{\theta}(1 - \hat{\theta})}\right).$$

There are certainly (infinitely) many possible transformation functions with good analytical properties, so it does make sense to use ones that are already popular in related statistical contexts such as the logit or probit functions. In terms of performance, the particular choice among these is secondary.

4 Confidence distributions for nonparametric effects in several samples

In practice, it is often not sufficient to compare a certain dose level of a verum to a placebo. For example, a new treatment may be compared to a placebo and to an active control. Following a nonparametric paradigm, using only pairwise relative effects in a situation with several samples can yield paradox results, due to the non-transitivity of these pairwise relative effects (see for example Thangavelu and Brunner (2007)). We address this problem by comparing each distribution to the same reference distribution.

In this section, we consider $X_{i1}, \ldots, X_{in_i} \sim F_i(x)$, $i = 1, \ldots, d$ independent and within each group i identically distributed observations, and we denote by $G := \frac{1}{d} \sum_{i=1}^{d} F_i$ the unweighted mean of the distribution functions $F_1, \ldots, F_d$, and define the unweighted relative effects

$$\theta_i := \int G dF_i = P(Z < X_{i1}) + P(Z = X_{i1}),$$

where $Z \sim G$ is a random variable independent of X_{i1}. We use the unweighted effect measure instead of its weighted counterpart because the latter usually requires almost balanced designs in order to provide useful and interpretable results (Brunner et al. 2017).

Also, we assume analogously to the previous part that $N/n_i \leq N_0 < \infty$, where $N = \sum_{i=1}^{d} n_i$ is the total number of observations, and we require that F_i is not a one-point distribution.

To estimate the effect size, we denote by $\hat{F}_i(x) = \frac{1}{n_i} \sum_{k=1}^{n_i} c(x - X_{ik})$ the *empirical distribution function* where c is the *count function* $c(u) = 0, 1/2, 1$, when $u < 0, u = 0, u > 0$ and by

$$R_{ik}^{\Psi} = \frac{1}{2} + \frac{N}{d} \sum_{j=l}^{d} \frac{1}{n_j} \sum_{l=1}^{n_j} c(X_{ik} - X_{jl})$$

the *pseudo-rank* of X_{ik} among all N observations. We refer to the recent literature for more details on the usefulness (Zimmermann et al. 2021) and the efficient computation (Happ et al. 2020) of pseudo-ranks.

Additionally, we define by

$$R_{ik}^{(ir)} = \frac{1}{2} + \sum_{s=i,r} \sum_{l=1}^{n_j} c(X_{ik} - X_{sl})$$

the *paired rank* of X_{ik} among all $n_i + n_r$ observations, by

$$R_{ik}^{(i)} = \frac{1}{2} + \sum_{l=1}^{n_j} c(X_{ik} - X_{il})$$

the *internal rank* of X_{ik} among all n_i observations, and by $\bar{R}_{i\cdot}^{\Psi} = \frac{1}{n_i} \sum_{k=1}^{n_i} R_{ik}^{\Psi}$, $i = 1, \ldots, d$ the *pseudo-rank mean*. Under the stated assumptions, a consistent and unbiased estimator of the unweighted relative effect is given by

$$\hat{\theta}_i = \int \hat{G} d\hat{F}_i = \frac{1}{N}\left(\bar{R}_{i\cdot}^{\Psi} - \frac{1}{2}\right), \ i = 1, \ldots, d.$$

Additionally if we denote by s_i^2 the variance of our test statistic $T_N = \sqrt{N}(\hat{\theta}_i - \theta_i)$, then T_N/s_i converges to the standard normal distribution (Brunner et al. 2017). Consistent estimation of the variance is possible using the previously defined different types of ranks, as stated in the following theorem.

Theorem 2 *If $N/n_i \leq N_0 < \infty$, then*

$$\hat{s}_i^2 = \frac{N}{n_i}\hat{\upsilon}_i^2 + \frac{N}{d^2}\sum_{r \neq i}^{d} \frac{1}{n_r}\hat{\tau}_{r:i}^2, \quad , r \neq i = 1, \ldots, d,$$

with

$$\hat{\upsilon}_i^2 = \frac{1}{N^2(n_i - 1)} \sum_{k=1}^{n_i} \left[R_{ik}^{\Psi} - \bar{R}_{i\cdot}^{\Psi} - \frac{N}{dn_i}\left(R_{ik}^{(i)} - \frac{n_i - 1}{2}\right) \right]^2,$$

$$\hat{\tau}_{r:i}^2 = \frac{1}{n_i^2(n_r - 1)} \sum_{s=1}^{n_r} \left(R_{rs}^{(ir)} - R_{rs}^{(r)} - \bar{R}_{r\cdot}^{(ir)} + \frac{n_r + 1}{2} \right)^2, r \neq i.$$

is a consistent estimator of the asymptotic variance s_i^2 of T_N.

Proof See Theorem 7.40 in Brunner et al. (2018). □

With this result, we obtain an asymptotic confidence distribution for the parameter θ_i as

$$H_n^{\Phi}(\theta_i) = \Phi\left(\frac{\theta_i - \hat{\theta}_i}{\hat{s}_i/\sqrt{N}}\right),$$

where Φ is the cumulative distribution function of the standard normal distribution. As usual, the corresponding asymptotic confidence curve is $CV(\theta_i) := 2\min\{H_n^{\Phi}(\theta_i), 1 - H_n^{\Phi}(\theta_i)\}$.

By definition, the parameter θ_i can only take values in the interval $[\frac{1}{2d}, 1 - \frac{1}{2d}]$. Similar to the previous section we can use the δ-method to construct range preserving confidence distributions, utilizing a transformation function g which fulfills the same conditions as in Sect. 3.3. For example, using the logit transformation, we obtain the range preserving asymptotic confidence distribution for the relative effect as

$$H_{n,\text{logit}}(\theta_i) = \Phi\left(\frac{\text{logit}(\theta_i) - \text{logit}(\hat{\theta}_i)}{\hat{s}_i/\{\hat{\theta}_i(1 - \hat{\theta}_i)\sqrt{N}\}}\right)\Bigg).$$

5 Simulation study

In this part we evaluate the performance of our proposed methods, in particular their accuracy. In the first subsection, we compare the estimated t-approximation and its variability against the empirical distribution function, and in the second subsection, we investigate how the p-values derived by confidence distributions fulfill the desired uniformity property, and examine the usefulness of range preserving transformations for large estimated effects.

5.1 Simulation of the approximated confidence distribution

In this subsection, we analyze the accuracy of the approximative confidence distributions using the t-distribution and the variability of the resulting estimates. For this purpose, we have computed the empirical cumulative distribution function (ecdf) for different simulated data sets and compared it to the t-approximation. We also simulated this using the asymptotic normal approximation, but only show the comparison to the t-approximation, due to negligible differences between both approaches.

Example 2 In this example we have drawn 10,000 times samples of size 20 for each of the two random variables X_1 and X_2, where $X_1 \sim N(1, 2)$ and $X_2 \sim N(0, 2)$ and computed the estimated relative effect $\theta := P(X_1 < X_2) + \frac{1}{2}P(X_1 = X_2)$ for all 10,000 draws. The resulting empirical cumulative distribution function is shown by the dotted line in Fig. 2a. Also, for each draw, we computed the t-approximation analogously to Chapter 3. In the plot, the dashed line shows the mean of these 10,000 distribution functions. The grey area gives an equal-tailed 95% confidence interval of these t-distributions. The vertical line gives the exact theoretical value computed by $\theta = \Phi(\frac{\mu_1 - \mu_2}{\sqrt{\sigma_1^2 + \sigma_2^2}})$.

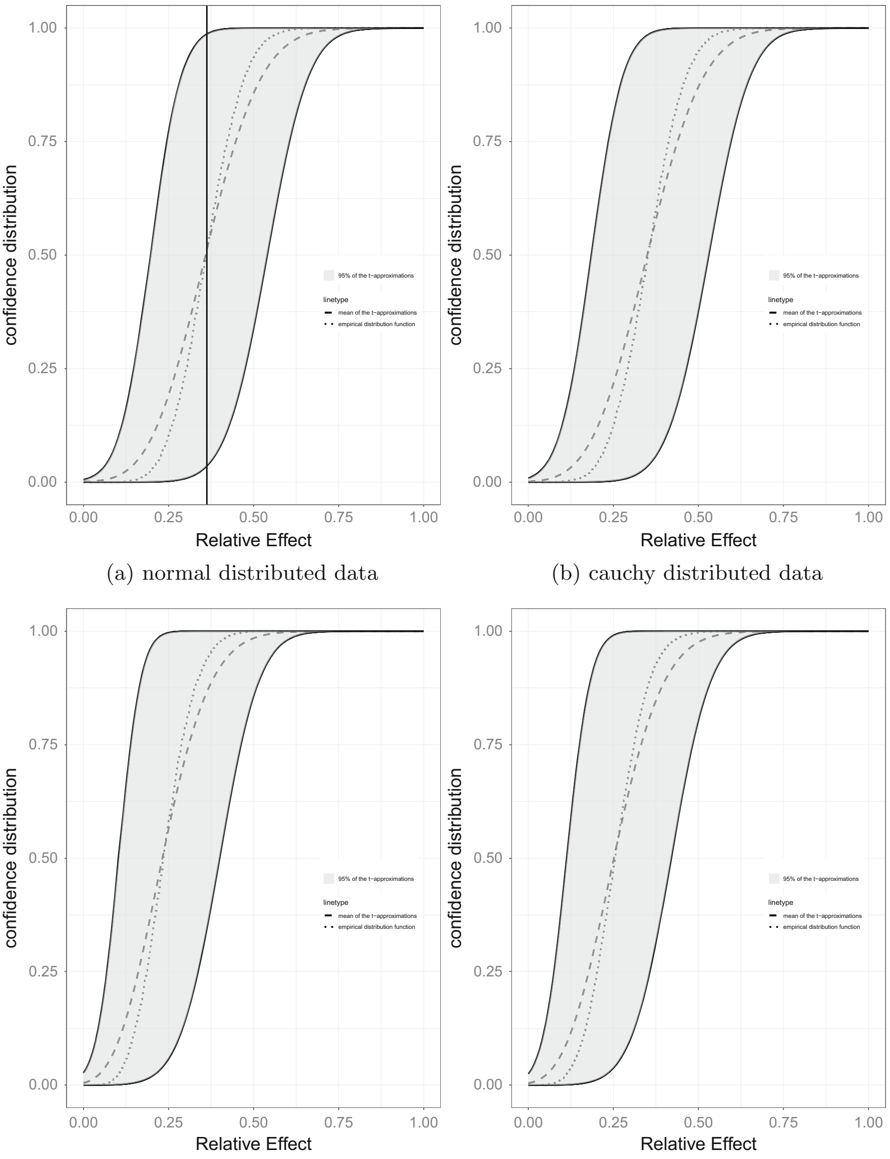
Similar as before, we have drawn 10,000 times samples with size 20 for each of the two random variables X_1 and X_2, but this time X_1 is Cauchy distributed with location 1 and scale 1 and X_2 is also Cauchy distributed with location 2 and scale 1. Again the ecdf for the estimated relative effect is shown in Fig. 2b by the dotted line and the mean of the t-distributions by the dashed one and the grey area the 95% confidence interval of the t-approximations.

In Fig. 2c we have again drawn 10,000 times samples with size 20 for each of the two random variables X_1 and X_2, but here $X_1 \sim LN(1, 1)$ (lognormal distribution) and $X_2 \sim LN(0, 1)$. As before the ecdf of the relative effect and the t-approximations are shown.

In the last Fig. 2d samples with size 20 for each for each group were drawn 10,000 times, but this time from a real data set, the number of implantations after dissection for female Wistar rats with 12 animals in the placebo group and 17 who received the drug. The plot shows the ecdf of the estimated relative effect and the t-approximations as in the previous ones.

The main idea of our simulation is based on a Monte Carlo simulation of confidence intervals. We simulated the range which is inside of 95 % of the confidence intervals based on our t-approximation, but here for all possible confidence levels.

The simulation has been carried out for symmetric (a and b) as well as non-symmetric (c and d) distributions. Typically, heavy tailed distributions such as the Cauchy distribution and the Log-normal distribution yield problems for parametric statistical inference procedures. In our simulations, we see almost no differences between



(a) normal distributed data

(b) cauchy distributed data

(c) log-normal distributed data

(d) resampling from real data set

Fig. 2 t-Approximation vs. ecdf

the simulations, especially when comparing to the empirical drawn samples from a real data set (Fig. 2d), confirming the robustness of the nonparametric approach pursued in this paper.

5.2 Simulation of range preserving confidence distribution

In this subsection, we examine the Range Preserving Confidence Distribution (RP-CD) defined in Sect. 3.3. By definition the confidence distribution should follow the uniform distribution on [0, 1]. Therefore, we evaluate the confidence curves at the true value of the relative effect and compare them to the uniform distribution on [0, 1] by means of a QQ-plot.

Example 3 Random samples of size 20 were simulated from a normal distribution with variance 2 and means 1, 3, and 5, respectively. We then compute the p-value function of the two sided test for the relative effect regarding data from a normal distribution with mean 0 and variance 2. The QQ-plot indicates how the p-value function evaluated at the true value of the relative effect (computed by formula (1)) coincides with the uniform distribution on [0, 1]. On the left side of Fig. 3, we show the plots derived by the asymptotic confidence distribution (aCD), and on the right side the ones derived by the RP-CD.

In the first row, the relative effect of 0.64 is close enough to the center of the (0, 1)-interval, and thus the transformation results in almost no difference. Both approaches meet the uniformity condition quite well. In the second row, the theoretical relative effect is at 0.86, and we can already see a small difference between the two procedures. The asymptotic confidence distribution without range preserving transformation tends to be a bit too liberal. In the third row, we have a relative effect of 0.96, and we can now see a major difference in the performance. The aCD is much too liberal, and the RP-CD performs still quite well (although the performance is not as good as in the case of relative effects closer to 0.5). Our simulation shows that the range preserving confidence distribution (RP-CD) is only needed in cases where the relative effect takes values very close to 0 or 1. In all other cases such an adjustment is not necessary.

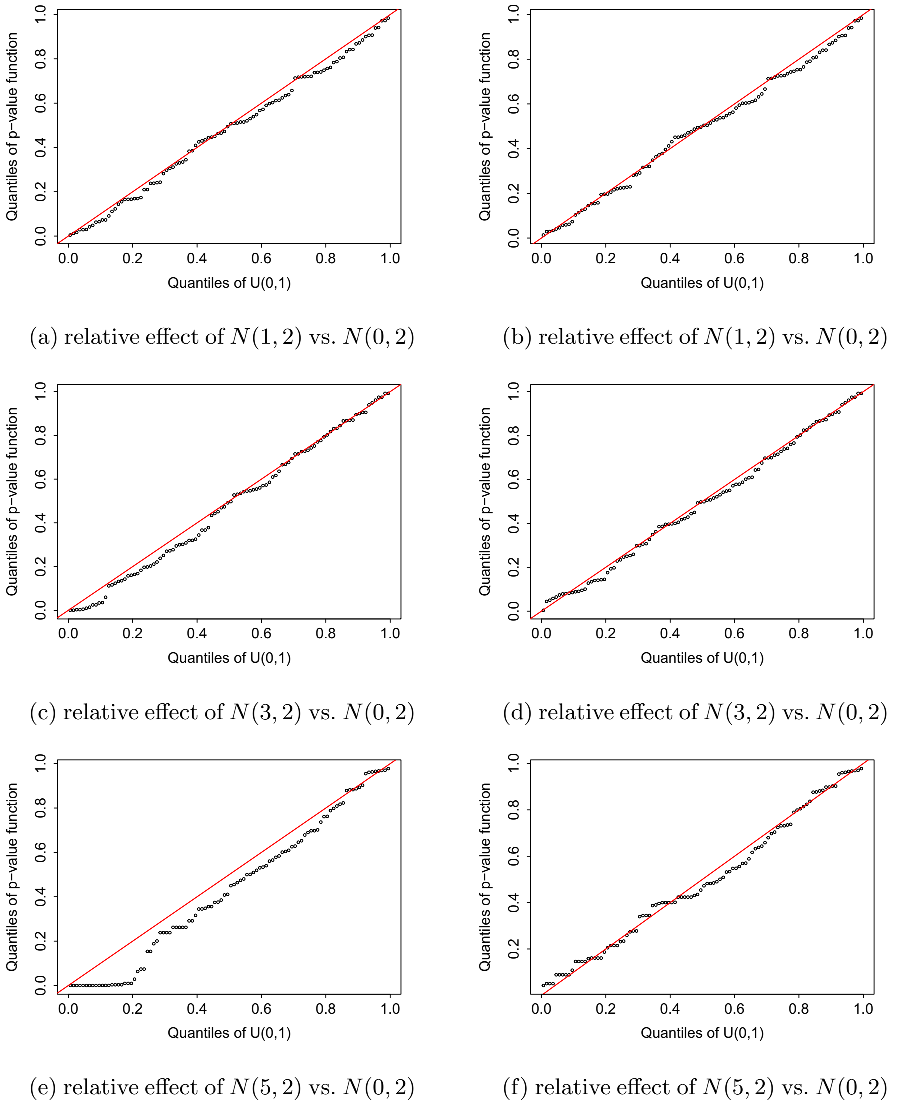
6 Case study

We now come back to the relative liver weights example presented in the introduction, and use this example to illustrate the confidence curve methodology that was introduced in Sects. 3 and 4. Also, we will show that the proposed method can be applied in situations where the main endpoint is an ordinal score,

6.1 Two-sample case (relative liver weights)

Undesired toxic effects of a drug administered in four increasing dose levels to male Wistar rats were to be examined in a pairwise comparison study.

Figure 4 helps to explain how confidence curves can answer our three questions from the introduction. Here we focus on dose level 2 vs. placebo. The confidence curve $2\min\{H_n, 1 - H_n\}$ is shown in Fig. 4a, while Fig. 4b shows the confidence distribution H_n, as well as $1 - H_n$. Note that these figures don't show the range preserving CD (see Fig. 5c). As a consequence, positive probability is assigned outside the interval [0, 1] in this example.



(a) relative effect of $N(1,2)$ vs. $N(0,2)$

(b) relative effect of $N(1,2)$ vs. $N(0,2)$

(c) relative effect of $N(3,2)$ vs. $N(0,2)$

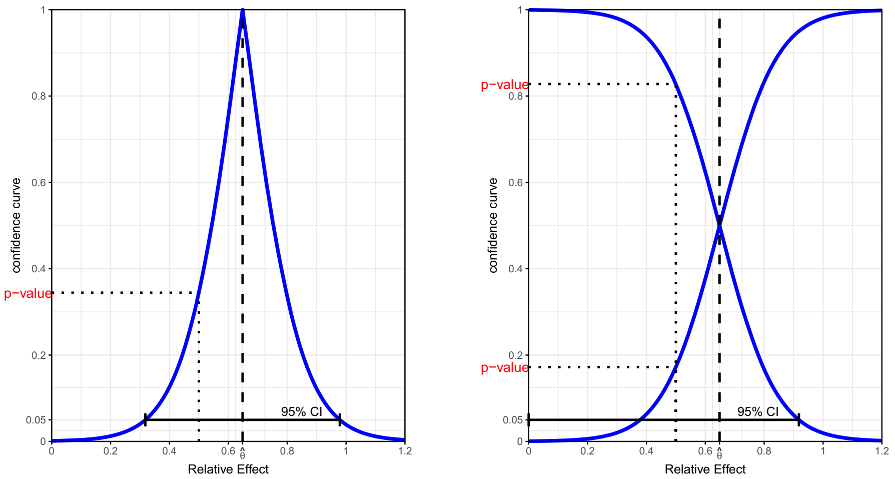
(d) relative effect of $N(3,2)$ vs. $N(0,2)$

(e) relative effect of $N(5,2)$ vs. $N(0,2)$

(f) relative effect of $N(5,2)$ vs. $N(0,2)$

Fig. 3 QQ-plots: p-values derived by the confidence distribution against the theoretical quantiles of an uniform distribution: aCD on the left and RP-CD on the right

The top of the graph in Fig. 4a (intersection point in Fig. 4b) marks the point estimator $\hat{\theta}$ of the relative effect θ on the x-axis (dashed line, $\hat{\theta} = 0.648$). The variability of this estimator is shown by confidence intervals for all possible confidence levels. The y-axis in Fig. 4a gives the α-level for $1 - \alpha$ equal-tailed two-sided confidence intervals. Here the 95% confidence interval $[0.318, 0.979]$ is shown explicitly (solid line). The solid line in Fig. 4b shows a one-sided upper bound 95% confidence interval $[0, 0.918]$. Of course a lower bound confidence interval can derived analogously.



(a) Confidence Curve (b) Confidence Distribution

Fig. 4 Confidence curve and confidence distribution for the relative effect of drug vs. placebo regarding relative liver weights

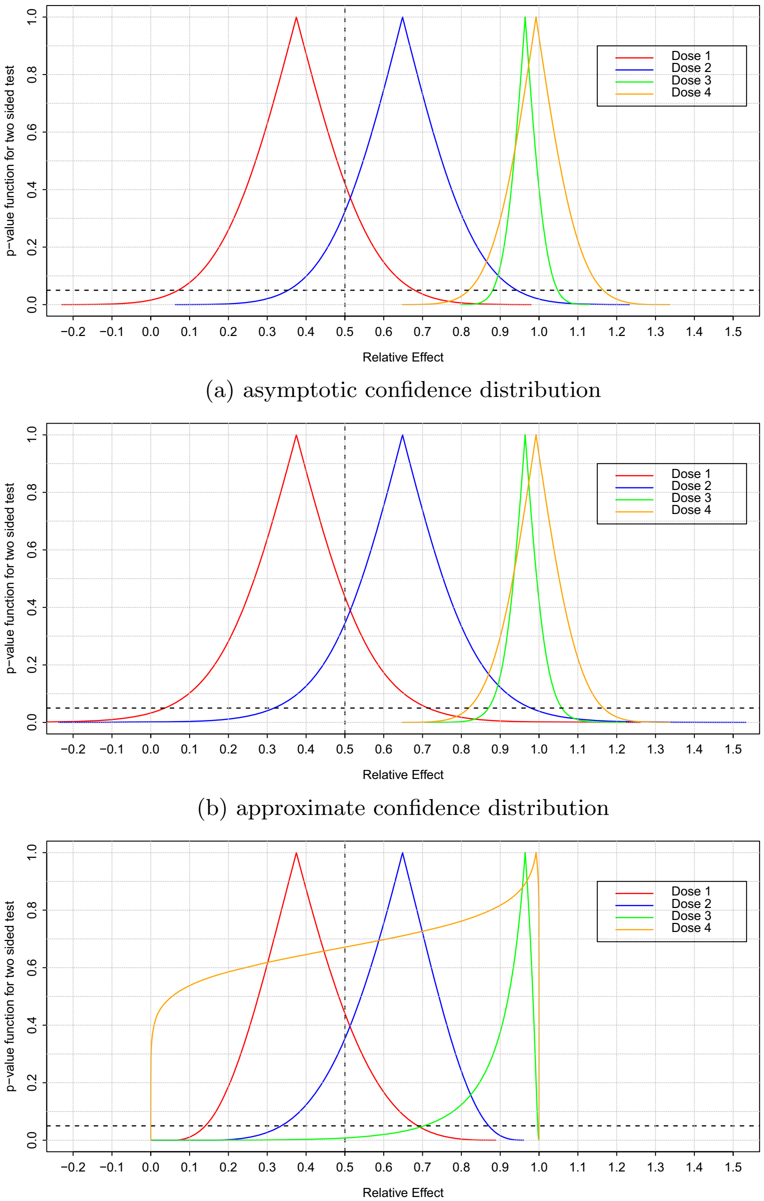
To evaluate the significance of an effect, the confidence curve provides us with the following information for two-sided null hypotheses: for every null hypothesis value θ on the x-axis, the corresponding p-value can be read on the y-axis. This justifies the alternative term *p-value function* for the confidence curve. In Fig. 4a, the dotted line shows the p-value for the null hypothesis $K_0 : \theta = 1/2$ (here: 0.344). For the one-sided null hypothesis $K_0 : \theta \leq \theta_0$ the confidence distribution (the increasing function in Fig. 4b) gives the p-value, and similarly the decreasing one for $K_0 : \theta \geq \theta_0$. For example, the p-value for $K_0 : \theta \leq 1/2$ is 0.172, and for $K_0 : \theta \geq 1/2$ the p-value is 0.828 (dotted lines).

In the joint plot of the asymptotic confidence curves for all pairwise comparisons of dose levels vs. placebo, we see in Fig. 5a that for all sensible confidence levels, doses 3 and 4 exhibit a relevant effect, while for the dose levels 1 and 2, such a statement is not possible.

In addition to asymptotic confidence curves, we also calculate approximate confidence curves for the relative liver weights. As can be seen in Fig. 5b, there are only small differences between both approaches.

Finally, we show the use of asymptotic range preserving confidence curves for the nonparametric pairwise relative effect. Here, the confidence distribution is by construction restricted to the unit interval. We can see in Fig. 5c that the confidence curve is now asymmetric, and especially for the fourth dose level, we obtain for almost all points in [0, 1] large p-values and quite large confidence intervals. However, for the other dose levels, the previously discussed statements are still valid.

Additionally in Table 1 we compared the different confidence curves evaluated at 1/2 (p-value for testing H_0: $\theta = 1/2$) with the studentized permutation test (SPT) (Pauly et al. 2016) and the Wilcoxon-Mann–Whitney (WMW) test equality of two



(a) asymptotic confidence distribution

(b) approximate confidence distribution

(c) range preserving confidence distribution

Fig. 5 Confidence curve for the relative effect different drug doses against the placebo on relative liver weights

Table 1 p-Values for the relative effect of different drug doses on the relative liver weights against the placebo

Drug	$CV_n^{\Phi}(1/2)$	$CV_n^{t}(1/2)$	$CV_n^{\text{logit}}(1/2)$	P-val SPT	P-val WMW
1	0.421	0.437	0.441	0.418	0.447
2	0.323	0.344	0.353	0.338	0.342
3	0.000	0.000	0.008	0.004	0.001
4	0.000	0.000	0.671	0.006	0.000

Table 2 p-Values for the relative effect of different drug doses on the relative liver weights against the mean distribution

Drug	$CV_n^{\Phi}(1/2)$	$CV_n^{\text{logit}}(1/2)$	P-val KW Test
0	0.000	0.001	0.000
1	0.024	0.040	
2	0.013	0.018	
3	0.000	0.000	
4	0.000	0.000	

distribution functions. Except the range preserving confidence curve for drug (dose level) 4, the p-values are more or less the same.

6.2 Several sample case (relative liver weights)

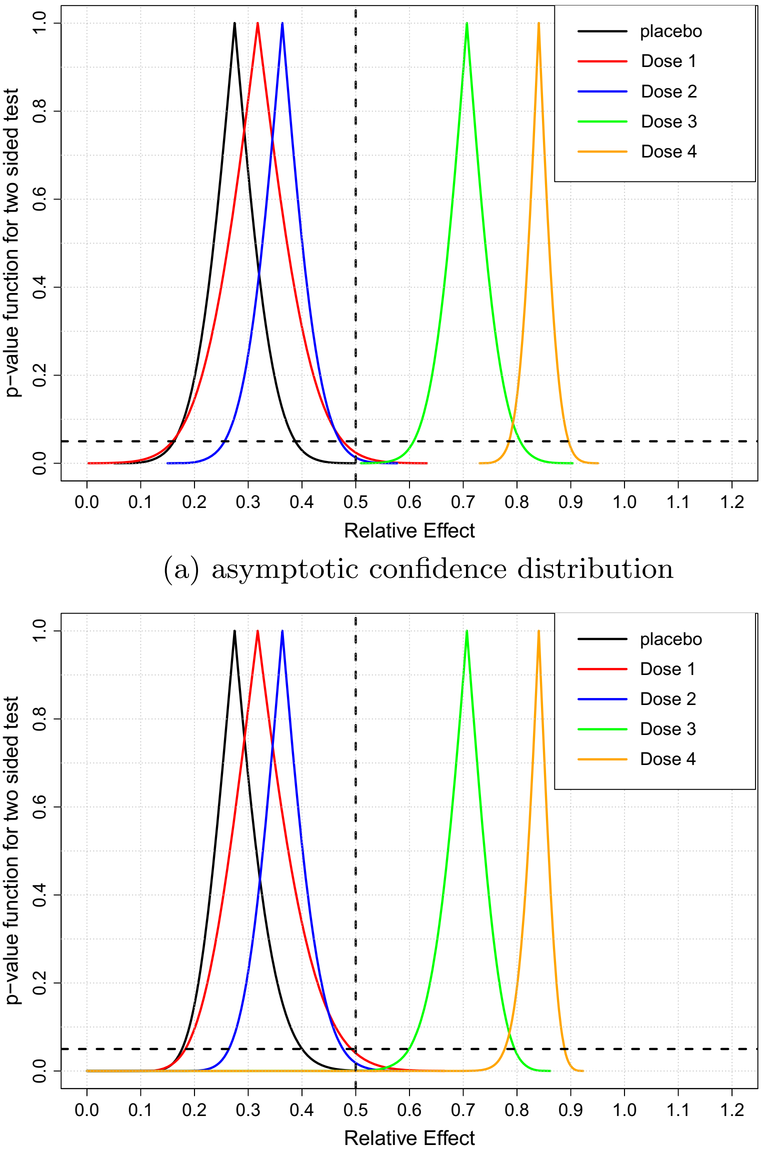
As explained in the fourth section, comparing different dose levels in a pairwise comparison to the placebo does not always suffice. Therefore, we compare every dose level and the placebo to the mean distribution of them. In Fig. 6a, we us the asymptotic confidence curve, and in Fig. 6b the range preserving confidence curve. Here we can see only a small difference between them, especially compared to the two-sample case in the previous section. In both plots, we see a clear effect for the dose levels 3 and 4. For dose levels 1 and 2, we see a tendency to smaller values except for very small confidence levels.

Table 2 compares the confidence curves evaluated at 1/2 (p-value for testing H_0: $\theta = 1/2$) with the approximate Kruskal–Wallis (KW) test for the equality of all distributions (Hollander et al. 2013).

6.3 Several sample case for non-metrical data

As explained in the introduction, one major advantage of the used rank- and pseudo-rank-based methods is the general validity of them. In addtion to metric data, one may also apply these methods to ordinal data, as demonstrated in this example.

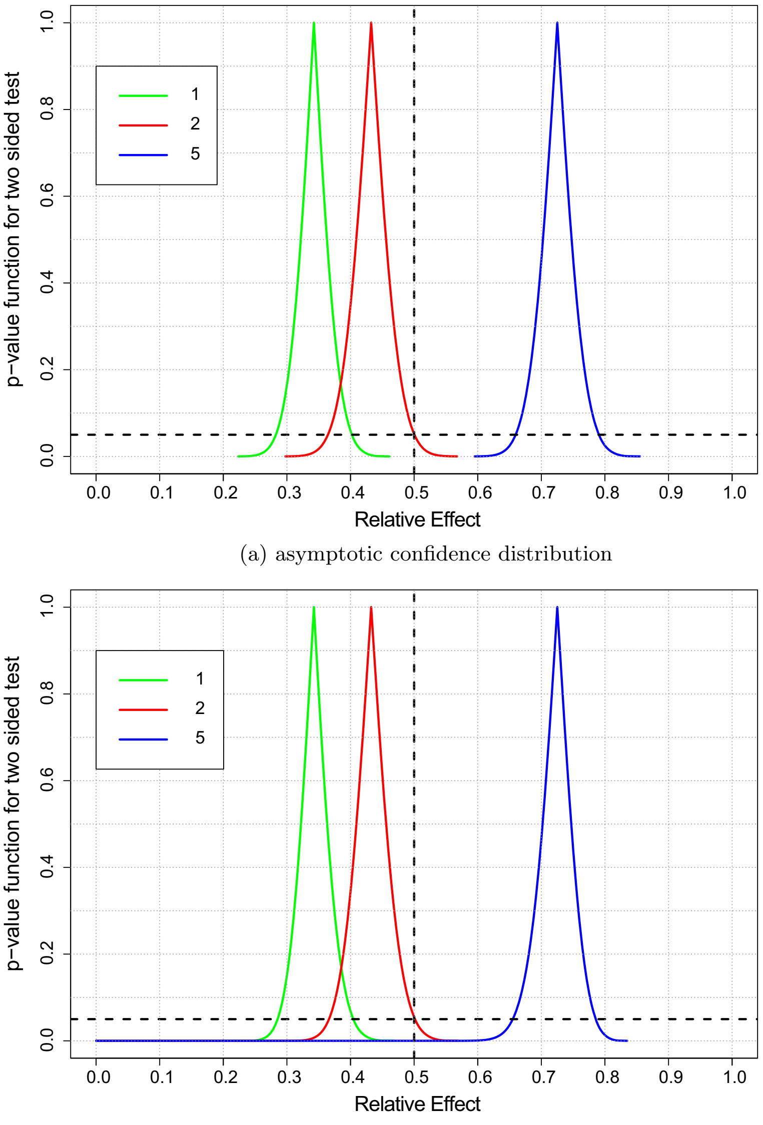
Here, we analyze how one gaseous substance in three different concentrations irritated or damaged the nasal mucous membrane of 25 DBA/1J mice each after sub-chronic inhalation. The damage was measured by a defect score from 0 to 4 (0 = “no



(a) asymptotic confidence distribution

(b) range preserving confidence distribution

Fig. 6 Confidence curve for the relative effect different drug doses compared to the mean distribution on relative liver weights



(a) asymptotic confidence distribution

(b) range preserving confidence distribution

Fig. 7 Confidence curve for the relative effect different drug doses compared to the mean distribution on the Nasal Mucosa

irritation", 1 = "mild irritation", 2 = "strong irritation", 3 = "severe irritation", 4 = "irreversible damage").

We can easily see in Fig. 7a and b that the highest concentration yields a significantly larger damage than the "mean" of all treatment levels, and a significantly smaller damage is observed at level 1. Again, we only have small differences between the

Table 3 p-Values for the relative effect of different drug doses on the Nasal Mucosa against the mean distribution

Drug	$CV_n^{\Phi}(1/2)$	$CV_n^{\text{logit}}(1/2)$	P-val KW Test
1	0.000	0.000	0.000
2	0.052	0.055	
5	0.000	0.000	

asymptotic confidence curves in Fig. 7a and the range preserving confidence curve in Fig. 7b.

Quite interesting in this example is the interpretation of the middle dose-level. For a confidence level of 0.05 or smaller we have no tendency to smaller or larger values but for a larger α-level, we have a tendency to smaller values for the middle dose-level, which shows how useful communicating the confidence curve, instead of a confidence interval for a certain confidence level, may prove.

Table 3compares again the confidence curves at $1/2$ with the approximate Kruskal Wallis (KW) test for the equality of all distributions (Hollander et al. 2013).

7 Discussion

We have introduced confidence distributions within a complete nonparametric framework using relative effects and their estimators based on ranks and pseudo-ranks. The nonparametric framework requires very few assumptions regarding the model and the data, thus significantly extending the scope of confidence distributions and their application. The resulting CD are more robust and more widely applicable than the parametric ones. We hope that our contribution will contribute to making the powerful and appealing theory of confidence distributions with its large field of applications, such as fusion learning (e.g. Liu et al. 2021; Shen et al. 2020; Cai et al. 2021) or predictive inference (e.g. Shen et al. 2018; Xie and Zheng 2021; Tian et al. 2021) available for situations in which standard parametric assumptions are not justifiable.

An enormous challenge in applied statistics, especially in biostatistics is the communication to practitioners. Here, confidence distributions could provide a major simplification in communication, as a unifier for the basic and essential inferential methods, such as p-values, point estimators, and confidence intervals.

The approach presented here provides asymptotic confidence distributions for large samples, and also approximative confidence distributions for small samples. For the case of a confidence distribution with positive probability mass outside the possible parameter space, we are proposing range preserving confidence distributions. And for the case of three or more samples, we recommend pseudo-rank procedures in order to alleviate possible interpretative problems due to non-transitivity and other issues that may arise with rank-based tests in unbalanced designs.

Most literature on confidence distribution assumes univariate parameter spaces. The development of unifying multiple testing and simultaneous confidence intervals within the confidence distribution theory may be a interesting extension to be pursued in future.

We have developed a unifying framework for statistical inference with the nonparametric relative effect based on confidence distributions. Since the relative effect is a nonparametric location measure, an extension of our theory to a nonparametric dispersion measure could be an interesting direction for future research.

Funding Open access funding provided by Paris Lodron University of Salzburg.

References

Akritas MG, Brunner E (1997) A unified approach to rank tests for mixed models. J Stat Plan Inference 61:249–277

Brunner E, Munzel U (2000) The nonparametric behrens-fisher problem: asymptotic theory and a small-sample approximation. Biom J 42(1):17–25

Brunner E, Puri M (2002) A class of rank-score tests in factorial designs. J Stat Plan Inference 103:331–360

Brunner E, Konietschke F, Pauly M, Puri ML (2017) Rank-based procedures in factorial designs: hypotheses about non-parametric treatment effects. J R Stat Soc B 79(5):1463–1485

Brunner E, Bathke AC, Konietschke F (2018) Rank and pseudo-rank procedures for independent observations in factorial designs. Springer, Cham

Brunner E, Konietschke F, Bathke AC, Pauly M (2021) Ranks and pseudo-ranks-surprising results of certain rank tests in unbalanced designs. Int Stat Rev 89(2):349–366

Cai C, Chen R, Xie M (2021) Individualized group learning. J Am Stat Assoc

Cox DR (1958) Some Problems Connected with Statistical Inference. Ann Math Stat 29(2):357–372

Efron B, Tibshirani RJ (1994) An introduction to the bootstrap. Chapman & Hall/CRC monographs on statistics & applied probability. Taylor & Francis, New York

Fisher RA (1935) The fiducial argument in statistical inference. Ann Eugen 6(4):391–398

Fraser D (1991) Statistical inference: Likelihood to significance. J Am Stat Assoc 86(414):258–265

Happ M, Zimmermann G, Brunner E, Bathke AC (2020) Pseudo-ranks: how to calculate them efficiently in r. J Stat Softw 95(1):1–22

Hollander M, Wolfe DA, Chicken E (2013) Nonparametric statistical methods Wiley series in probability and statistics. Wiley, New York

Liu D, Liu RY, Xie M (2014) Exact meta-analysis approach for discrete data and its application to 2 x 2 tables with rare events. J Am Stat Assoc 109(508):1450–1465

Liu D, Liu RY, Xie M (2015) Multivariate meta-analysis of heterogeneous studies using only summary statistics: Efficiency and robustness. J Am Stat Assoc 110(509):326–340

Liu D, Liu RY, Xie M (2021) Nonparametric fusion learning for multiparameters: synthesize inferences from diverse sources using data depth and confidence distribution. J Am Stat Assoc 1–19

Luo X, Dasgupta T, Xie M, Liu RY (2021) Leveraging the fisher randomization test using confidence distributions: inference, combination and fusion learning. J R Stat Soc B 83(4):777–797

Mann HB, Whitney DR (1947) On a test of whether one of two random variables is stochastically larger than the other. Ann Math Stat 18(1):50–60

Moser BK, Stevens GR (1992) Homogeneity of variance in the two-sample means test. Am Stat 46(1):19–21

Pauly M, Asendorf T, Konietschke F (2016) Permutation- based inference for the auc: a unified approach for continuous and discontinuous data. Biom J 58:1319

Schweder T, Hjort N (2002) Confidence and likelihood. Scand J Stat 29:309–332

Schweder T, Hjort NL (2016) Confidence, likelihood, probability, vol 41. Cambridge University Press, Cambridge

Shen J, Liu RY, Xie MG (2018) Prediction with confidence-a general framework for predictive inference. J Stat Plan Inference 195:126–140

Shen J, Liu RY, Xie M (2020) ifusion: individualized fusion learning. J Am Stat Assoc 115(531):1251–1267

Singh K, Xie M, Strawderman WE (2005) Combining information from independent sources through confidence distributions. Ann Stat 33(1):159–183

Singh K, Xie M, Strawderman WE (2007) Confidence distribution (cd)–distribution estimator of a parameter. In: Complex datasets and inverse problems, vol. 54, pp. 132–150

Thangavelu K, Brunner E (2007) Wilcoxon-Mann-Whitney test for stratified samples and Efron's paradox dice. J Stat Plan Inference 137(3):720–737 . Special Issue on Nonparametric Statistics and Related Topics: In honor of M.L. Puri

The International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH): ICH Topic E9. Statistical Principles for Clinical Trials (1998)

Thornton S, Xie M (2020) Bridging bayesian, frequentist and fiducial (bff) inferences using confidence distribution. arXiv preprint

Tian L, Wang R, Cai T, Wei L-J (2011) The highest confidence density region and its usage for joint inferences about constrained parameters. Biometrics 67(2):604–610

Tian Q, Meng F, Nordman DJ, Meeker WQ (2021) Predicting the number of future events. J Am Stat Assoc 117:1296

van der Vaart AW (1998) Asymptotic statistics. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge

Wilcoxon F (1945) Individual comparisons by ranking methods. Biom Bull 1(6):80–83

Xie M, Singh K (2013) Confidence distribution, the frequentist distribution estimator of a parameter: a review. Int Stat Rev 81(1):3–39

Xie M, Zheng Z (2021) Homeostasis phenomenon in conformal prediction and predictive distribution functions. Int J Approx Reason 141:131

Zimmermann G, Bolter LM, Sluka R, Höller Y, Bathke A, Thomschewski A, Leis S, Lattanzi S, Brigo F, Trinka E (2019) Sample sizes and statistical methods in interventional studies on individuals with spinal cord injury: a systematic review. J Evid-Based Med 12:200

Zimmermann G, Brunner E, Brannath W, Happ M, Bathke AC (2021) Pseudo-ranks: the better way of ranking? Am Stat 76:124