Mathematical Modelling and Analysis
Volume 15 Number 1, 2010, pages 137–151
Doi:10.3846/1392-6292.2010.15.137-151
© Vilnius Gediminas Technical University, 2010

ISSN 1392-6292 print
ISSN 1648-3510 online
www.vgtu.lt/mma/

Two-Sample Problems in Statistical Data Modelling

J. Valeinis, E. Cers and J. Cielens

University of Latvia
Zellu 8, LV-1002, Riga, Latvia
E-mail(*corresp.*): `valeinis@lu.lv`

Received August 31, 2009; revised December 21, 2009; published online February 15, 2010

Abstract. A common problem in mathematical statistics is to check whether two samples differ from each other. From modelling point of view it is possible to make a statistical test for the equality of two means or alternatively two distribution functions. The second approach allows to represent the two-sample test graphically. This can be done by adding simultaneous confidence bands to the probability-probability $(P-P)$ or quantile-quantile $(Q-Q)$ plots. In this paper we compare empirically the accuracy of the classical two-sample t-test, empirical likelihood method and several bootstrap methods. For a real data example both $Q-Q$ and $P-P$ plots with simultaneous confidence bands have been plotted using the smoothed empirical likelihood and smoothed bootstrap methods.

Keywords: two-sample problem, t-test, empirical likelihood, quantile-quantile plot, probability-probability plot, simultaneous bands.

AMS Subject Classification: 62G10; 62G15; 62G09.

1 Introduction

Many statistical applications deal with two groups of observations of the same kind that originate from two possibly different model distributions. One of the most common question in statistical applications is whether these two distributions have different expectations. More specifically, let $X_1, X_2, \ldots, X_{n_1}$ and $Y_1, Y_2, \ldots, Y_{n_2}$ be two independent samples with distribution functions F_1 and F_2 and expectations μ_1 and μ_2, respectively. In this case we wish to test the hypothesis

$$H_0 : \mu_1 = \mu_2 \quad \text{against} \quad H_1 : \mu_1 \neq \mu_2. \tag{1.1}$$

The alternative hypothesis H_1 can be set also to be one-sided ($\mu_1 > \mu_2$ or $\mu_1 < \mu_2$) depending on the practical applications. Among the statistical methods dealing with the testing problem (1.1) we have to mention the classical two-sample t-test, empirical likelihood method and different standard bootstrap methods. The question is which method to use for a particular practical problem under consideration. t-test is known to be a robust test although it is based on a very restrictive assumption that the modelling distributions F_1 and F_2 should be normal. Empirical likelihood and bootstrap methods are non-parametric statistical methods, which do not have the restrictive assumption of normality.

A common method to compare different tests dealing with the same problem is to make empirical power comparison by Monte-Carlo simulations with respect to different alternatives. In this paper we will alternatively compare empirical coverage accuracy of pointwise confidence intervals for the parameter $\mu_2 - \mu_1$ for all methods mentioned above. For the chosen significance level α if the confidence interval does not contain 0 then the null hypothesis (1.1) is rejected at this level. Therefore confidence intervals for $\mu_2 - \mu_1$ not only contain the answer to the testing problem, but also give an additional information where the true parameter lies with some prescribed confidence $(1 - \alpha)$.

One may also approach the two–sample problem from an alternative point of view. We can test the hypothesis

$$H_0 : F_1 = F_2 \quad \text{against} \quad H_1 : F_1 \neq F_2, \tag{1.2}$$

which of course is stronger than (1.1). There are many statistical tests dealing with this problem, among them most well-known are Kolmogorov–Smirnov, Cramer–von Mises and Anderson–Darling tests. However, to check (1.2) we propose to construct simultaneous confidence intervals for the probability – probability $(P - P)$ or quantile – quantile $(Q - Q)$ plot of the two distribution functions. $P - P$ plot is defined as the plot of the function $\{F_1(F_2^{-1}(y)) : y \in (0, 1)\}$ and $Q - Q$ plot is defined as $\{F_1^{-1}(F_2(x)) : x \in \mathbb{R}\}$. Obviously when both distributions F_1 and F_2 are equal, the $P - P$ and $Q - Q$ plot should lie on the 45-degree line. Adding simultaneous confidence intervals gives a formal two-sample test. The procedure is:

1. Draw an empirical $P - P$ or $Q - Q$ plot.

2. Add simultaneous bands at some chosen confidence level.

3. If the diagonal $y = x$ fits into the bands at every point do not reject the null hypothesis.

Almost every statistical package provides a possibility to construct the empirical $Q - Q$ and $P - P$ plots with pointwise confidence intervals. One of our goals was to develop and implement a code in statistical package R for constructing the simultaneous bands (the code is available on the author's homepage).

We construct simultaneous confidence bands using the smoothed empirical likelihood and bootstrapped empirical $P - P$ and $Q - Q$ processes. Empirical likelihood, introduced by Owen, [18, 19] has nice properties, especially for confidence intervals (see, for example, [11, 20]). It does not involve any prescribed assumptions about the shape of intervals, which is fully determined by the data. Moreover, it is Bartlett correctable in most cases. Thus a simple correction for scale improves the coverage accuracy from order n^{-1} to n^{-2}, where n denotes the sample size. For the mean difference corresponding to the testing problem (1.1) the Bartlett correction has been investigated in [16] and [17]. For two-sample problems in a general framework the empirical likelihood method has been introduced by [22] and [24].

The order of quantile and distribution function in definition of $P-P$ and $Q-Q$ plots make them quite different (see [9, 13]). If Y is a linear function of X then its $Q-Q$ plot will still be linear but with possible changed location and slope. This property is not shared by $P-P$ plots. On the other hand the range of a $P-P$ plot is always the same, i.e. a rectangle with the diagonal connecting $(0,0)$ and $(1,1)$ which makes them comparable. When there is a positive or negative shift between two samples, the $P-P$ plot is above or below the diagonal. This indicates a positive or negative treatment effect. According to Holmgren [13] $P-P$ plots, among other advantages, are to be preferred when outliers are present.

The paper is organized as follows. In Sections 2 and 3 the empirical likelihood method has been introduced in the one and two-sample cases, respectively. Furthermore its smoothed version is defined in Section 4. Empirical $P-P$ and $Q-Q$ plot processes are discussed in Section 5. Section 6 deals with the construction of simultaneous confidence bands for $P-P$ and $Q-Q$ plot functions. Finally, empirical coverage accuracy using Monte Carlo simulations is analyzed in Section 7, where also a real data example is considered.

2 One–Sample Empirical Likelihood Method

First we will give a brief overview on the empirical likelihood function introduced by Owen [18, 19] in the one-sample case. Let $X_1, X_2, \ldots, X_{n_1}$ be a sample with some unknown distribution function F_1. The empirical (or nonparametric) likelihood function is

$$L(F) = \prod_{i=1}^{n} dF(X_i) = \prod_{i=1}^{n} p_i, \tag{2.1}$$

where $p_i = dF(X_i) = P(X = X_i)$. The difference from the usual (parametric) likelihood function is obvious. We do not assume any parametric form of the density or distribution. Instead we model the data by discrete distributions having positive atom probabilities on the data points. This is in fact also a kind of parametric likelihood method – we model our data with the family of multinomial distributions.

Obviously (2.1) is maximized by the empirical distribution function

$$F_{1n_1}(x) = n_1^{-1} \sum_{i=1}^{n_1} I(X_i \le x).$$

This motivates the usual statistical plug-in estimator technique. Often estimators of parameters can be seen as estimators of some statistical functionals $T(F)$. For example, the mean $\mu := E(X) = \int x dF_1(x)$. The plug-in estimator is $T(F_{1n})$ which in case of the mean leads to the usual sample mean estimator $\bar{X} = n_1^{-1} \sum_{i=1}^{n_1} X_i$. This is also a nonparametric likelihood estimator for the mean. Consider for simplicity the hypothesis testing for the mean in the one-sample case, that is,

$$H_0 : \mu = \mu_0 \quad \text{against} \quad H_1 : \mu \neq \mu_0. \tag{2.2}$$

For the hypothesis testing (2.2) the most powerful is the likelihood ratio test, which is based on the likelihood function (see, for example, [5]). Analogously let us define the empirical likelihood ratio $R(F) = L(F)/L(F_n) = \prod_{i=1}^{n} np_i$ and the profile empirical likelihood ratio function for the mean

$$R_E(\mu) = \sup\Big\{\prod_{i=1}^{n} np_i \Big| \, p_i \geq 0, \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i X_i = \mu\Big\}.$$

For the general case let us define the estimating function $w_1(X, \mathit{\Delta})$ such that $E_{F_1}(w_1(X, \mathit{\Delta})) = 0$, where the parameter of interest $\mathit{\Delta} \in \mathbb{R}^d$. In this case (2.1) is maximized under the constraints

$$p_i \geq 0, \quad \sum_{i=1}^{n_1} p_i = 1, \quad \sum_{i=1}^{n_1} p_i w_1(X_i, \mathit{\Delta}) = 0.$$

For the special case of hypothesis testing (2.2) we have $w_1(X, \mathit{\Delta}) = X_i - \mu$. Therefore the maximization problem is determined by constraints on estimating function $w_1(X, \mathit{\Delta})$. If 0 is inside the convex hull of the points $w_1(X_1, \mathit{\Delta})$, $w_1(X_2, \mathit{\Delta}), \ldots, w_1(X_n, \mathit{\Delta})$ then the unique maximum exists (see, [21]).

Using the standard Lagrange multiplier method one can obtain

$$p_i = \frac{1}{n_1(1 + \lambda_1 w_1(X_i, \mathit{\Delta}))},$$

where $\lambda_1 = \lambda_1(\mu)$ can be found as a solution of

$$\sum_{i=1}^{n} \frac{X_i - \mu}{n_1(1 + \lambda_1 w_1(X_i, \mathit{\Delta}))} = 0.$$

Finally, Qin and Lawless [21] have shown under some smoothness assumptions on the estimating function $w_1(X_i, \mathit{\Delta})$ that a nonparametric analogue of Wilk's theorem holds. Thus minus two multiplied by the logarithm of the likelihood ratio statistic converges in distribution to the chi-squared distributed random variable. More precise,

$$-2 \log R(\mathit{\Delta}) = 2 \sum_{i=1}^{n} \log(1 + \lambda_1(\mathit{\Delta}) w_1(X_i, \mathit{\Delta})) \rightarrow_d \chi_d^2, \tag{2.3}$$

where the degree of freedom d denotes the dimension of the parameter $\mathit{\Delta}$. Similarly as in the case of likelihood ratio statistic in the parametric case we can derive the confidence intervals for the true parameter $\mathit{\Delta}_0$. The confidence interval will contain such $\mathit{\Delta}$ values for which $\{R(\mathit{\Delta}) > c\}$, where c can be determined from (2.3).

To conclude, the empirical likelihood method has no assumptions on underlying distribution of population. The obtained confidence intervals are not symmetric as usually obtained by standard methods but fully determined by the data. For a recent review of the empirical likelihood methods see [20].

3 Two–Sample Empirical Likelihood Method

Assume we wish to make statistical inference about the function $\varDelta := \varDelta(t)$ on some interval T. Let θ_0 be some univariate parameter associated with one of the distributions F_1 or F_2 regarded as a nuisance parameter. When dealing with the empirical likelihood method, it is common to assume that information about true parameters $\theta_0, \varDelta_0$ is available in the form

$$E_{F_1} w_1(X, \theta_0, \varDelta_0, t) = 0, \tag{3.1}$$

$$E_{F_2} w_2(Y, \theta_0, \varDelta_0, t) = 0. \tag{3.2}$$

This setup was introduced in [24] allowing to deal also with functions such as $P - P$ and $Q - Q$ plots. If $\varDelta_0 = \theta_1 - \theta_0$, where θ_0 and θ_1 are univariate parameters associated with F_1 and F_2 respectively, we have exactly the setup of [22] allowing to make inference for two-sample parameter differences. For example, for the testing problem of two expectation equality (1.1) choose

$$\theta_0 = \int x\, dF_1(x), \quad \varDelta_0 = \int y\, dF_2(y) - \int x\, dF_1(x).$$

We obtain (3.1) and (3.2) by taking

$$w_1(X, \theta_0, \varDelta_0, t) = X - \theta_0, \quad w_2(Y, \theta_0, \varDelta_0, t) = Y - \theta_0 - \varDelta_0.$$

In the following we define the profile empirical likelihood ratio function

$$R(\varDelta, \theta) = \sup_{p,q} \prod_{i=1}^{n_1} (n_1 p_i) \prod_{j=1}^{n_2} (n_2 q_j), \tag{3.3}$$

where $p = (p_1, \ldots, p_{n_1})$ and $q = (q_1, \ldots, q_{n_2})$ are subject to restrictions

$$p_i \geq 0, \quad \sum_{i=1}^{n_1} p_i = 1, \quad \sum_{i=1}^{n_1} p_i w_1(X_i, \theta, \varDelta, t) = 0,$$
$$q_j \geq 0, \quad \sum_{j=1}^{n_2} q_j = 1, \quad \sum_{j=1}^{n_2} q_j w_2(Y_j, \theta, \varDelta, t) = 0.$$

A unique solution of (3.3) exists, provided that 0 is inside the convex hull of the points $w_1(X_i, \theta, \varDelta, t)$'s and the convex hull of the $w_2(Y_j, \theta, \varDelta, t)$'s. The maximum may be found using the standard Lagrange multipliers method (see, e.g. [20, 24]). Finally, define an estimator $\hat{\theta} = \hat{\theta}(\varDelta)$ of the nuisance parameter θ by minimizing (3.3) over θ for a fixed value of $\varDelta$:

$$\hat{\theta}(\varDelta) = \arg\min_{\theta} \{-2 \log R(\varDelta, \theta)\}. \tag{3.4}$$

Assumptions (Qin and Lawless [22]). Assume that the partial derivatives $\alpha_1(X_i, \theta, \varDelta, t) = \dfrac{\partial w_1(X_i, \theta, \varDelta, t)}{\partial \theta}$ and $\alpha_2(Y_i, \theta, \varDelta, t) = \dfrac{\partial w_2(Y_i, \theta, \varDelta, t)}{\partial \theta}$ exist. Furthermore assume that:

(i) $\theta_0 \in \Theta$ and Θ is an open interval;

(ii) $E_{F_1} w_1^2(X,\theta,\Delta,t) > 0$, $E_{F_2} w_2^2(Y,\theta,\Delta,t) > 0$, $\alpha_1(X,\theta,\Delta,t)$, $\alpha_2(Y,\theta,\Delta,t)$ are continuous in a neighborhood of θ_0, $\alpha_1(X,\theta,\Delta,t)$ and $w_1^3(X,\theta,\Delta,t)$ are bounded by some integrable function $G_1(X)$ in this neighborhood, $\alpha_2(Y,\theta,\Delta,t)$ and $w_2^3(Y,\theta,\Delta,t)$ are bounded by some integrable function $G_2(Y)$ in this neighborhood, and $E_{F_1}\alpha_1(X,\theta,\Delta,t)$ and $E_{F_2}\alpha_2(Y,\theta,\Delta,t)$ are nonzero;

(iii) $n_2/n_1 \to k$ (as $n_2, n_1 \to \infty$) and $0 < k < \infty$.

Theorem 1. *If assumptions (i)-(iii) are satisfied, then*

$$-2\log R(\Delta_0,\hat{\theta}) \to_d \chi_1^2, \tag{3.5}$$

as $n_1, n_2 \to \infty$, where $\to_d$ denotes the convergence in distribution.

Proof. The same as the proof of Theorem 1 in [22]. □

The pointwise empirical likelihood confidence interval for fixed $t \in T$ for the true parameter Δ_0 has the following form $\{\Delta : R(\Delta,\hat{\theta}) > c\}$. The constant c can be calibrated using Theorem 1.

4 Smoothed Empirical Likelihood for $P - P$ and $Q - Q$ Plots

It has been shown in Valeinis [24] (see Section 5.3) that the empirical likelihood method introduced in Section 3 is suitable also for $P - P$ and $Q - Q$ plots. Let $\theta_0 = F_2^{-1}(t)$ and $\Delta_0 = F_1(F_2^{-1}(t))$. In this case forms (3.1) and (3.2) are satisfied by choosing

$$w_1(X,\theta_0,\Delta_0,t) = I_{\{X \le \theta_0\}} - \Delta_0, \quad w_2(Y,\theta_0,\Delta_0,t) = I_{\{Y \le \theta_0\}} - t.$$

Furthermore, for $Q - Q$ plots choose $\theta_0 = F_2(t)$, $\Delta_0 = F_1^{-1}(F_2(t))$ and

$$w_1(X,\theta_0,\Delta_0,t) = I_{\{X \le \Delta_0\}} - \theta_0, \quad w_2(Y,\theta_0,\Delta_0,t) = I_{\{Y \le t\}} - \theta_0.$$

Note, that indicator functions w_1 and w_2 are non-smooth. In order to apply Theorem 1 we propose to use the smoothed empirical likelihood method.

The smoothed empirical likelihood method for quantile function in one sample case has been first introduced in [3]. It appears that by appropriate smoothing of estimating functions w_1, w_2 the coverage accuracy may be improved from order $n^{-1/2}$ to n^{-1}. Some recent papers have dealt with the smoothed method in the two-sample case. For example, [4] deals with receiver operating characteristic (ROC) curves, in [1] two-sample goodness of fit testing problem is considered and finally [2] is devoted to copulas using the smoothed empirical likelihood method.

For $j = 1, 2$ define $H_j(t) = \int_{u \le t} K_j(u)\,du$, where K_j is a kernel function (typically a density function). Further let $H_{b_j}(t) = H_j(t/b_j)$, where $b_1 = b_1(n_1)$

and $b_2 = b_2(n_2)$ are bandwidth sequences, converging to zero as sample sizes n_1, n_2 grow to infinity. Let $p = (p_1, \dots, p_{n_1})$ and $q = (q_1, \dots, q_{n_2})$ be two vectors consisting of nonnegative numbers adding to one. Define further the estimators

$$\hat{F}_{b_1,p}(x) = \sum_{i=1}^{n_1} p_i H_{b_1}(x - X_i) \quad \text{and} \quad \hat{F}_{b_2,q}(y) = \sum_{j=1}^{n_2} q_j H_{b_2}(y - Y_j). \tag{4.1}$$

For this setting we define the profile two-sample smoothed empirical likelihood ratio function for $\mathit{\Delta}$ as

$$R^{(sm)}(\mathit{\Delta}, \theta) = \sup_{p,q} \prod_{i=1}^{n_1}(n_1 p_i) \prod_{j=1}^{n_2}(n_2 q_j), \tag{4.2}$$

with the smoothed estimating equation for $P - P$ plots

$$w_1(X_i, \theta_0, \mathit{\Delta}_0, t) = H_{b_1}(\theta_0 - X_i) - \mathit{\Delta}_0, \quad w_2(Y_j, \theta_0, \mathit{\Delta}_0, t) = H_{b_2}(\theta_0 - Y_j) - t$$

and for $Q - Q$ plots

$$w_1(X_i, \theta_0, \mathit{\Delta}_0, t) = H_{b_1}(\mathit{\Delta}_0 - X_i) - \theta_0, \quad w_2(Y_j, \theta_0, \mathit{\Delta}_0, t) = H_{b_2}(t - Y_j) - \theta_0.$$

Proposition 1. *Under suitable conditions on bandwidth sequences b_1 and b_2 Theorem 1 holds for the function $R^{(sm)}(\mathit{\Delta}_0, \hat{\theta})$.*

Remark 1. Conditions on bandwidth rates in Proposition 1 will be published elsewhere. For some special cases the rates have been derived in several papers. For example, for ROC curves and $P - P$ plots see [4], for general structural relationship models see [24].

5 Empirical $P - P$ and $Q - Q$ Processes

Let us denote the empirical distribution functions of the X and Y samples by

$$F_{1n_1}(x) = n_1^{-1} \sum_{i=1}^{n_1} I(X_i \le x), \quad F_{2n_2}(x) = n_2^{-1} \sum_{j=1}^{n_2} I(Y_j \le y),$$

respectively. The empirical quantile function is defined as $F_{1n_1}^{-1}(t) = \inf\{x : F_{n_1}(x) \ge t\}$. The classical Kolmogorov-Smirnov two-sample statistic for the hypothesis (1.2) is defined as follows

$$KS = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \sup_{-\infty < x < +\infty} |F_{1n_1}(x) - F_{2n_2}(x)|. \tag{5.1}$$

Under the null hypothesis

$$KS \to_d \sup_{0<t<1} |B(t)|,$$

where $B(t)$ is a Brownian bridge process. Hsieh and Turnbull [15] (see Theorem 2.2) showed that there exists a probability space on which one can define two independent sequences of Brownian bridges $B_1^{(n_1)}$ and $B_2^{(n_1)}$ such that

$$\sup_{0<t<1} P_{n_1,n_2}(t) := \sup_{0<t<1} \sqrt{n_1}|F_{1n_1}(F_{2n_2}^{-1}(t)) - F_1(F_2^{-1}(t))| \to_d$$
$$\sup_{0<t<1} |B_1^{(n_1)}(F_1(F_2^{-1}(t))) + \sqrt{\lambda}\frac{f_1(F_2^{-1}(t))}{f_2(F_2^{-1}(t))}B_2^{(n_2)}(t)|, \quad (5.2)$$

where $n_1/n_2 \to \lambda$ as $n_1, n_2 \to \infty$ and P_{n_1,n_2} denotes the empirical $P-P$ plot process. For the empirical $Q-Q$ plot process similar result holds,

$$\sup_{-\infty<x<\infty} Q_{n_1,n_2}(x) := \sup_{-\infty<x<\infty} \sqrt{n_1}|f_1(F_1^{-1}(F_2(x)))(F_{1n_1}^{-1}(F_{2n_2}(x))$$
$$- F_1^{-1}(F_2(x)))| \to_d \sup_{0<t<1} |B_1^{(n_1)}(F_2(x)) + \sqrt{\lambda}B_2^{(n_2)}(F_2(x))|. \quad (5.3)$$

It is impossible to construct simultaneous bands from (5.2) and (5.3) because the limiting distribution contains the unknown functions f_1, f_2 and F_1, F_2 which have to be estimated. Thus the asymptotic behavior will heavily depend on those estimators. As usual in such situations resampling methods have to be used.

Bootstrap resampling method. Consider, for example, the supremum statistic of $P-P$ empirical plot process defined in (5.2) which we simply denote by T. In order to make statistical inference (hypothesis testing or confidence bands) we have to know the sampling distribution of the statistic. In particular our interest is to find an appropriate $1-\alpha$ quantile (for fixed α, typically equal to 0.05 or 0.01) from the distribution which is a function of the sum of two Brownian bridges and unknown quantities f_1, f_2, F_1, F_2. If we knew the true underlying distributions F_1 and F_2 it would be sufficient to do Monte-Carlo simulations. In this case we would have to replicate say N samples from the population. Then we could form estimates for $P(T \leq p)$ by counting how many of the T_i's are $\leq p$ and dividing by N (as we know that the relative frequency approximates the probability).

In case of unknown distributions in order to approximate the asymptotic limiting distribution the usual nonparametric or empirical bootstrap method proposes the following procedure. Let us draw B replicated samples with replacement from the initial data set with the probability n^{-1} of drawing each separate observation. More formally this means that the bootstrapped resamples $X_1^*, X_2^*, \ldots, X_{n_1}^*$ and $Y_1^*, Y_2^*, \ldots, Y_{n_2}^*$ have been drawn from the empirical distribution functions F_{1n_1} and F_{2n_2}, respectively. Finally, we obtain the bootstrapped test statistic values $T_1^*, T_2^*, \ldots, T_B^*$. We estimate the probability $P(T \leq p)$ again by simple frequency-based estimate as mentioned before. That is, by counting T_i^*'s, which are $\leq p$ and dividing by B.

The bootstrap resampling method introduced by Efron [8] in 1979 has obtained nowadays a large applicability. According to Shao and Tu [23]: "because of the availability of inexpensive and fast computing such computer-intensive methods have caught on very rapidly in recent years and are particularly appreciated by applied statisticians". Moreover, the bootstrap method gives a better approximation as the central limit theorem, which can be shown theoretically using Edgeworth expansion methods, see [6, 7, 23].

In [14] as an alternative to nonparametric bootstrap method, the smoothed bootstrap method has been used to construct simultaneous bands for ROC curve defined as $1 - F_1(F_2^{-1}(1-t))$ for $0 \le t \le 1$. Their results hold also for $P - P$ plots, which can be seen as a simple transformation from the ROC curve function. Define the standard smoothed nonparametric estimators for unknown distribution functions F_1 and F_2 as follows

$$\hat{F}_{1n_1}(x) = \frac{1}{n_1}\sum_{i=1}^{n_1} H_{b_1}(x - X_i) \quad \text{and} \quad \hat{F}_{2n_2}(y) = \frac{1}{n_2}\sum_{j=1}^{n_2} H_{b_2}(y - Y_j)$$

with $H_{b_j}(t)$ the same as in (4.1).

The idea of the smoothed bootstrap is to draw independent random samples $X_1^*, X_2^*, \ldots, X_{n_1}^*$ and $Y_1^*, Y_2^*, \ldots, Y_{n_2}^*$ from $\hat{F}_{1n_1}$ and $\hat{F}_{2n_2}$, respectively. The empirical distributions of X^* and Y^* are denoted by

$$F_{1n_1}^*(x) = n_1^{-1}\sum_{i=1}^{n_1} I(X_i^* \le x), \quad F_{2n_2}^*(y) = n_2^{-1}\sum_{j=1}^{n_2} I(Y_j^* \le y).$$

From Theorem 2.1 in [14] it follows that for a ROC $P - P$ plot process P_{n_1,n_2} with probability 1 holds

$$\sup_x \Bigg| P\left(\sup_{0<t<1} \sqrt{n_1}|F_{1n_1}(F_{2n_2}^{-1}(t)) - F_1(F_2^{-1}(t))| \le x\right) \\ - P^*\left(\sup_{0<t<1} \sqrt{n_1}|F_{1n_1}^*(F_{2n_2}^{*-1}(t)) - F_{1n_1}(F_{2n_2}^{-1}(t))| \le x\right)\Bigg| \to 0 \quad (5.4)$$

as $n_1, n_2 \to \infty$, where conditional probabilities for given initial data are denoted by P^*. Similar conclusion can be done for $Q - Q$ plot processes.

6 Simultaneous Confidence Bands

Pointwise confidence intervals for $P - P$ or $Q - Q$ plots allow to make inference only for these functions at a fixed point. That is the same as to construct confidence intervals for some single parameter of interest, such as mean of the sample. It is clearly preferable to know with, say, 95% confidence where the whole true curve lies. However, using statistical packages such as R it is possible only to add pointwise intervals for $Q - Q$ plots, for example. We will consider two methods for the construction of simultaneous bands.

To construct a simultaneous confidence region for Δ defined in Section 3 over some interval (a, b), we will use the bootstrap confidence region without losing the advantages of the automated shape-determination by the empirical likelihood method. It means that we use empirical likelihood to set the shape of the confidence bands and use the bootstrap to set the level. This method is introduced in [10] and has been also used in [4] and [24].

First method. Define the maximum smoothed empirical likelihood estimator $\hat{\mathit{\Delta}}$,

$$\hat{\mathit{\Delta}} = \arg\max_{\mathit{\Delta}} R^{(sm)}(\mathit{\Delta}, \hat{\theta}),$$

where $\hat{\theta} = \hat{\theta}(\mathit{\Delta})$ is defined in (3.4). For the construction of simulatenous bands over some interval (a, b), first, choose an appropriate bootstrap critical value c^* such that

$$P(-2\log R^{*(sm)}(\hat{\mathit{\Delta}}, \hat{\theta}) \leq c^* \text{ for } a \leq t \leq b\} = 1 - \alpha,$$

where $R^{*(sm)}$ is the likelihood ratio function defined in (4.2) and calculated for bootstrapped resamples $X_1^*, X_2^*, \ldots, X_{n_1}^*$ and $Y_1^*, Y_2^*, \ldots, Y_{n_2}^*$. Second, use c^* for the pointwise confidence bands from Theorem 1. Hence the bootstrap confidence band $\mathcal{C}$ consists of curves $R^{(sm)}(\cdot, \hat{\theta})$ such that the corresponding log likelihood ratio statistic stays below c^* over the interval (a, b), that is,

$$\mathcal{C} = \{-2\log R^{(sm)}(\cdot, \hat{\theta}) : -2\log R^{(sm)}(\mathit{\Delta}, \hat{\theta}) \leq c^* \text{ for } a \leq t \leq b\}.$$

Note that the estimator $\hat{\mathit{\Delta}}$ is based on the initial samples $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$.

Second method. For $P - P$ plots equation (5.4) provides the bootstrap approximation of the critical value c_2^* such that

$$\begin{aligned} P(F_{1n_1}(F_{2n_2}^{-1}(t)) - c_2^* n_1^{-1/2} &\leq F_1(F_2^{-1}(t)) \\ &\leq F_{1n_1}(F_{2n_2}^{-1}(t)) + c_2^* n_1^{-1/2}, \quad a \leq t \leq b) \to 1 - \alpha, \end{aligned} \tag{6.1}$$

as $n_1, n_2 \to \infty$ on some interval $0 < a < b < 1$ with $0 < \alpha < 1$. For $Q - Q$ plot processes a similar result is true.

7 Simulation Study and Applications

We have implemented the generalized two-sample empirical likelihood method proposed by Valeinis [24] in statistical package *R*. Program codes will be available on the author's website (*http : //home.lanet.lv/ valeinis/index.html*).

Table 1 shows comparative 95% coverage accuracies using the empirical likelihood method, four standard kinds of bootstrap methods, and the two-sample t-test. More specifically, we will use the percentile bootstrap (B.P.), normal bootstrap (B.N.), basic bootstrap (B.B) and bias-corrected bootstrap methods (see, for example, [6, 23]).

First, we compare two normal distributions $N(0, 1)$ and $N(1, 1)$ for which, as expected, the t-test is better among other methods. Note, that both empirical likelihood and the various bootstrap procedures produce acceptable coverage accuracies, with the empirical likelihood having a slight edge. The coverage accuracies using all methods quickly converges towards 0.95 in this case.

Further we examine a family of log-normal distributions $\log N(\mu, \sigma^2)$ with increasing values of parameter σ^2 and fixed $\mu = 0$. Due to the asymmetry, log-normal distributions are known to be problematic for the t-test. We find, that

Table 1. 95% coverage accuracies for two-sample mean differences $\mu_2 - \mu_1$, comparing. Empirical likelihood (E.L.), two–sample t-test, four bootstrap methods (B.N. – normal bootstrap; B.B. – basic bootstrap; B.P. – percentile bootstrap; B.A. – bias–corrected bootstrap). The coverage accuracies are based on 10,000 pseudorandom samples from F_1 and F_2 for each combination of sample sizes n_1 and n_2.

F_1, F_2	n_1	n_2	E.L.	B.N.	B.B.	B.P.	B.A.	t-test
$F_1 = N(0,1)$ $F_2 = N(1,1)$	15	15	0.933	0.919	0.922	0.920	0.917	0.951
	20	20	0.940	0.930	0.931	0.930	0.928	0.952
	30	30	0.944	0.934	0.936	0.935	0.934	0.950
	50	50	0.947	0.940	0.941	0.941	0.940	0.951
	100	100	0.952	0.949	0.951	0.952	0.949	0.954
	20	30	0.939	0.937	0.937	0.937	0.937	0.948
	30	20	0.940	0.935	0.938	0.934	0.934	0.951
	20	50	0.937	0.936	0.937	0.935	0.934	0.950
	50	20	0.940	0.934	0.937	0.937	0.934	0.951
$F_1 =$ $LogN(0,0.5)$ $F_2 =$ $LogN(0,0.5)$	15	15	0.922	0.935	0.941	0.927	0.911	0.958
	20	20	0.923	0.934	0.941	0.927	0.908	0.954
	30	30	0.937	0.940	0.949	0.940	0.925	0.955
	50	50	0.940	0.941	0.947	0.939	0.928	0.950
	100	100	0.945	0.945	0.950	0.943	0.937	0.949
	20	30	0.935	0.941	0.949	0.938	0.925	0.956
	30	20	0.927	0.935	0.942	0.932	0.915	0.950
	20	50	0.930	0.934	0.936	0.931	0.919	0.945
	50	20	0.932	0.935	0.941	0.934	0.920	0.949
$F_1 = LogN(0,1)$ $F_2 = LogN(0,1)$	15	15	0.878	0.941	0.958	0.908	0.855	0.969
	20	20	0.892	0.947	0.961	0.921	0.868	0.966
	30	30	0.911	0.950	0.962	0.929	0.885	0.963
	50	50	0.919	0.948	0.961	0.933	0.897	0.956
	100	100	0.931	0.950	0.961	0.938	0.910	0.955
	20	30	0.902	0.947	0.963	0.924	0.877	0.963
	30	20	0.903	0.944	0.957	0.924	0.874	0.959
	20	50	0.890	0.923	0.938	0.907	0.864	0.937
	50	20	0.895	0.924	0.938	0.909	0.873	0.939
$F_1 = LogN(0,2)$ $F_2 = LogN(0,2)$	15	15	0.767	0.967	0.983	0.899	0.780	0.986
	20	20	0.784	0.967	0.983	0.902	0.786	0.983
	30	30	0.811	0.969	0.984	0.906	0.796	0.982
	50	50	0.833	0.970	0.984	0.916	0.816	0.977
	100	100	0.858	0.967	0.982	0.921	0.828	0.973
	20	30	0.788	0.963	0.980	0.896	0.780	0.978
	30	20	0.790	0.963	0.980	0.897	0.784	0.976
	20	50	0.796	0.945	0.965	0.877	0.777	0.956
	50	20	0.792	0.943	0.965	0.873	0.770	0.954

the coverage accuracy for the empirical likelihood method converges towards 0.95 from below, quite in line with the performance of the percentile bootstrap and the bias-corrected bootstrap. The t-test, however, performs very conservatively in the two-sample case and seems to converge very slowly. Interestingly, the same is true for the basic and normal bootstrap methods, which, in fact, show no tendency to converge downwards to 0.95. While for practical purposes a conservative test might be more appropriate in some cases, it must be noted, that conservativeness comes at a cost of wider confidence intervals. For all methods when σ^2 increases much more observations are needed to have reasonable results.

Finally, we consider the wet drilling versus dry drilling data from [12]. In this data example wet drilling times (where cuttings are flushed with water) are compared to dry drilling times (the cuttings are flushed with compressed air). The times are given in 1/100 of a minute, for a 5 feet segment. Six series of drilling times for 5 feet segments were given for six drilled holes, three of each wet drilled and three dry drilled. For our example we computed the average segment times for each drilling depth with sample sizes $n_1 = n_2 = 80$. The question is whether dry drilling (let μ_d denote the expected dry drilling time) is faster then wet drilling (where μ_w denotes the expected wet drilling time). Simple box plots suggesting slightly faster dry drilling times are shown in Figure 1.

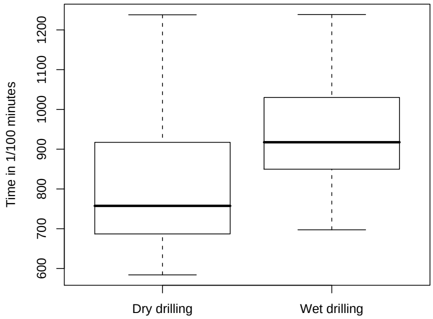

Figure 1. Box-plots of the drilling data samples.

The estimated mean dry and wet drilling times are 805.53 and 943.81, respectively, giving an estimate for the mean difference $\mu_w - \mu_d$ equal to 138.3. This strongly indicates that both samples differ from each other. To find out whether this difference is statistically significant let us further test hypothesis (1.1) for the equality of sample means. This can be done either using the two–sample t-test, or by the empirical log-likelihood ratio statistic (3.5). Both tests reject the null hypothesis with a p-value of < 0.0001.

Next, let us construct the 95% confidence interval for the parameter $\mu_w - \mu_d$ using the t-test, the empirical log-likelihood ratio statistic (3.5) and some standard bootstrap methods. The results are summarized in Table 2. We conclude that all methods produce similar intervals in our case. Confidence intervals not only allow to assess the range of likely values of the difference of means with a given confidence but also provide hypothesis testing for a given significance level.

Next approach is to test hypothesis (1.2) about the equality of the sample distributions F_1 and F_2, which is stronger than hypotheses (1.1) already discussed before. If the two mean values μ_1 and μ_2 differ also the null hypothesis of F_1 and F_2 equality should be rejected, whereas there could be significant differences in the structure of the data, still the mean values being equal. The

Kolmogorov-Smirnov test (5.1) rejects H_0 with a p-value < 0.0001.

Table 2. 95% confidence interval for $\mu_w - \mu_d$ calculated using the empirical likelihood method (E.L.); the two–sample t-test and four kinds of bootstrap (B.N. – normal bootstrap; B.B. – basic bootstrap; B.P. – percentile bootstrap; B.A. – bias–corrected bootstrap).

E.L.	t-test	B.N.	B.B.	B.P.	B.A.
$(84.1, 180.6)$	$(94.6, 181.9)$	$(96.3, 181.5)$	$(95.5, 183.5)$	$(95.1, 181)$	$(93, 180.9)$

Finally, we offer a graphical assessment of the distribution relationships using $P - P$ and $Q - Q$ plots, shown in Figure 2. Coupled with simultaneous confidence bands, theses graphs also allow for a graphical test of hypothesis (1.2). The graphs show both the empirical versions of the plots, and their smoothed counterparts, calculated using the empirical likelihood method. 95% simultaneous confidence bands were constructed using both methods described in Section 6. The diagonal $y = x$ does not fit into the confidence bands. Thus both plots allow to reject H_0.

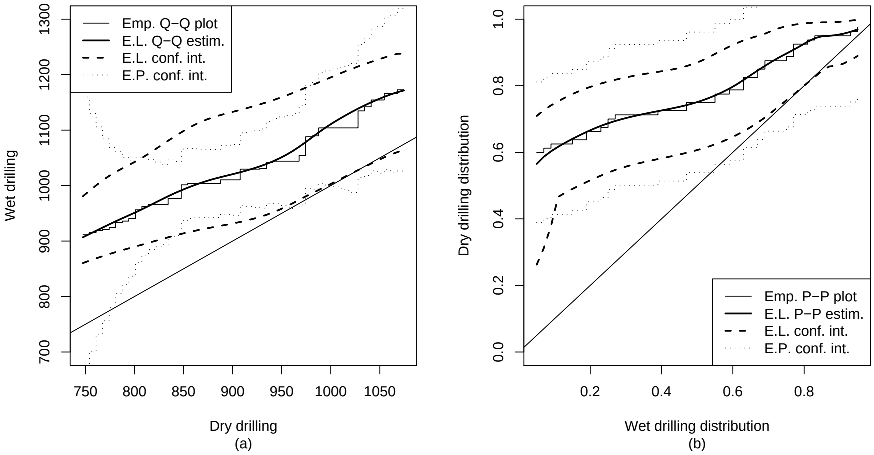

Figure 2. $Q - Q$ (a) and $P - P$ (b) plots of dry drilling versus wet drilling. The plots show the empirical versions together with the smoothed empirical likelihood estimates. Two types of simultaneous 95% confidence bands were constructed using empirical likelihood and empirical processes combined with the bootstrap methods described in Section 6.

A major advantage of the graphical testing method is that the graphs show more details of the distribution differences. For example, from the $P - P$ plot graph we can see that the dry drilling time is stochastically less then that of wet drilling, because the graph lies above the diagonal.

To construct simultaneous confidence bands using both empirical processes and empirical likelihood method one needs to find a suitable bandwidth parameter. Since both distributions of dry and wet drilling were fairly normal, we used the standard rule-of-thumb procedure to select the smoothing parameter. Throughout, 10,000 bootstrap replications were performed. Using the smoothed empirical likelihood method bands the selected bootstrap critical values c^* for the graphs were 7.55 and 7.59 for the $Q-Q$ and the $P-P$ plots respectively. Furthermore using the second method (6.1) the bootstrapped critical values were found to be 2.02 and 1.82.

Note, that while for the $Q-Q$ plot the bands constructed using empirical processes are narrower in the middle, they tend to go to infinity at both ends of the graph. This is a consequence of the density function involved in the definition of the empirical $Q-Q$ process (5.3). This might indicate that the use of empirical likelihood could be preferable here, since in many cases the most 'interesting' data lies near the edges of the graph. For the $P-P$ plot, again empirical likelihood seems preferable, since the bands constructed using it are narrower on the whole graph. A drawback of using empirical likelihood method is the following. In order to calculate bootstrapped critical values rather large samples are needed. Secondly, we have to cut ends of both $P-P$ and $Q-Q$ plots, which has to be considered by each data example separately.

Our findings demonstrate that methods based on empirical likelihood are comparable or in some cases even better than established (bootstrap) methods. A further investigation of both practical and simulated data examples would be of interest here. Moreover, the advantages of $P-P$ plots over $Q-Q$ plots attributed by Holmgren [13] could be analysed further in the context of graphical two-sample tests.

References

[1] R. Cao and I. Van Keilegom. Empirical likelihood tests for two-sample problems via nonparametric density estimation. *The Canadian Journal of Statistics*, **34**(1):61–77, 2006. Doi:10.1002/cjs.5550340106.

[2] J. Chen, L. Peng and Y. Zhao. Empirical likelihood based confidence intervals for copulas. *Journal of Multivariate Analysis*, **100**(1):137–151, 2009. Doi:10.1016/j.jmva.2008.04.005.

[3] S. X. Chen and P. Hall. Smoothed empirical likelihood confidence intervals for quantiles. *The Annals of Statistics*, **21**(3):1166–1181, 1993. Doi:10.1214/aos/1176349256.

[4] G. Claeskens, B-Y. Jing, L. Peng and W.Zhou. Empirical likelihood confidence regions for comparison distributions and ROC curves. *The Canadian Journal of Statistics*, **31**(2):173–190, 2003. Doi:10.2307/3316066.

[5] D.R. Cox and D.V. Hinkley. *Theoretical Statistics.* Chapman & Hall, London, 1974.

[6] A. DasGupta. *Asymptotic Theory of Statistics and Probability.* Springer-Verlag, New York, 2008.

[7] A.C. Davison and D.V. Hinkley. *Bootstrap Methods and Their Application.* Cambridge University Press, 1997.

[8] B. Efron. Bootstrap methods: Another look a the jackknife. *The Annals of Statistics*, **7**(1):1–26, 1979. Doi:10.1214/aos/1176344552.

[9] R. Gnanadesikan and M.B. Wilk. Probability plotting methods for the analysis for the analysis of data. *Biometrika*, **55**(1):1–17, 1968.

[10] P. Hall and A. Owen. Empirical likelihood confidence bands in density estimation. *Journal of Computational and Graphical Statistics*, **2**:273–289, 1993. Doi:10.2307/1390646.

[11] P. Hall and B. La Scala. Methodology and algorithms of empirical likelihood. *International Statistical Review*, **58**(2):109–127, 1990. Doi:10.2307/1403462.

[12] D.J. Hand, F. Daly, A.D. Lunn, K.J. McConway and E. Ostrowski. *A handbook of small data sets*. Chapman & Hall, Boca Raton, FL, 1994.

[13] E. B. Holmgren. The $P-P$ plot as a method for comparing treatment effects. *Journal of the American Statistical Association*, **90**(429):360–365, 1968. Doi:10.2307/2291161.

[14] L. Horvath, Z. Horvath and W. Zhou. Confidence bands for ROC curves. *Journal of Statistical Planning and Inference*, **138**(6):1894–1904, 2008. Doi:10.1016/j.jspi.2007.07.009.

[15] F. Hsieh and B.W. Turnbull. Nonparametric and semiparametric estimation of the receiver operating characteristic curve. *The Annals of Statistics*, **24**(1):25–40, 1996. Doi:10.1214/aos/1033066197.

[16] B-Y. Jing. Two-sample empirical likelihood method. *Statistics & Probability Letters*, **24**(4):315–319, 1995. Doi:10.1016/0167-7152(94)00189-F.

[17] Y. Liu, C. Zou and R. Zhang. Empirical likelihood for the two-sample mean problem. *Statistics & Probability Letters*, **78**(5):548–556, 2008.

[18] A. Owen. Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, **75**(2):237–249, 1988. Doi:10.1093/biomet/75.2.237.

[19] A. Owen. Empirical likelihood ratio confidence regions. *The Annals of Statistics*, **18**(1):90–120, 1990. Doi:10.1214/aos/1176347494.

[20] A. Owen. *Empirical likelihood*. Chapman & Hall, Boca Raton, FL, 2001.

[21] J. Qin and J. Lawless. Empirical likelihood and general estimating equations. *The Annals of Statistics*, **22**(1):300–325, 1994. Doi:10.1214/aos/1176325370.

[22] Y. Qin and L. Zhao. Empirical likelihood ratio confidence intervals for various differences of two populations. *Systems Science and Mathematical Sciences*, **13**(1):23–30, 2000.

[23] J. Shao and D. Tu. *The Jackknife and Bootstrap*. Springer-Verlag, New York, 1995.

[24] J. Valeinis. *Confidence bands for structural relationship models*. Vdm Verlag Dr. Mueller E.K., Germany, 2008.