Nonparametric Bootstrap Tests for the Generalized Behrens-Fisher Problem

Paul Cotofrei
University of Neuchatel

1 Introduction

Like other concepts and approaches that have dominated the statistical literature of the last two decades, the idea behind the bootstrap did not appear overnight, with the publishing of the article of Efron(1979). But Efron's pioneering paper has structured subsequent discussion and shown the enormous potential of the methodology. The key idea behind the bootstrap is to resample from the original data — either directly or via a fitted model — in order to create replicate datasets, from which the variability of the quantities of interest can be assessed without long-winded and error-prone analytical calculation. Because this approach involves repeating the data analysis procedure many times, bootstrap methods are sometimes called computer-intensive methods. The original name, bootstrap methods, — a very good example of "statistical marketing" — was not chosen randomly, because to use the data to generate more data seems analogous to a trick used by the fictional Baron Münchausen, who, when he found himself at the bottom of a lake, got out by pulling himself up by his bootstraps.

Starting from 1979, the list of research papers, technical reports, textbooks and monographs concerning bootstrap methods and their applications increased strongly year after year. The basic bibliography on this subject must absolutely contain the monographs Efron (1982), Efron and Tibshirani (1993) and Davison and Hinkley (1997) . These books concentrated on ideas rather than their mathematical justification and represent a very good introduction for readers without a strong mathematical background. The monographs by Hall (1992), Shao and Tu (1995) and Politis, Romano and Wolf (1999) may be recommended to mathematical statisticians. Finally, for the practitioners, an excellent guide is represented by Chernick (1999) . This is the first monograph on the bootstrap to provide extensive coverage of real world applications for practitioners in many diverse fields.

Many statistical applications involve significance tests to assess the plausability of scientific hypotheses. Resampling methods are not new to significance testing, since randomisation tests and permutation tests have long been used to provide nonparametric tests, and Monte Carlo tests, which use simulated datasets, are quite common in certain areas of application. Testing is related to constructing confidence sets, but although an hypothesis test can be obtained by constructing an appropriate confidence set, bootstrap hypothesis testing remains an important topic for the following reason.

- First, sometimes finding a test directly is much easier than constructing a confidence set.

- Secondly, tests obtained directly may be better since they usually take account of the special nature of the hypothesis.

- Thirdly, for bootstrap confidence sets, we always generate bootstrap data from an estimated distribution $\hat{F}$ without any restriction. For hypothesis testing, we must generate bootstrap data from either $\hat{F}$ or from an estimated distribution under the restrictions specified by the hypothesis.

- Finally, hypothesis testing requires the calculation of the P-value, and in some cases an estimate of the power of the test.

The proper selection of test statistics before bootstrapping is a key factor in improving the order of correctness of the bootstrap tests. Although a such test may be better than the test based on normal approximation in terms of mean squared error (Liu and Singh, 1987), Bunke et Riemer (1983) demonstrated that a test based on a less pivotal test statistic may result in a large difference between the actual level of the test and the nominal level. Ducharme and Jhun (1986) showed that the performance of a bootstrap test can be improved by using a studentized test statistic. Sutton (1993) suggested using Johnson's modified-t statistics to achieve accurate bootstrap tests about the mean of an asymmetric distribution. More advantages of using a pivot as a test statistic are given in Hall and Wilson (1991). Beran (1988), Hinkley and Shi (1989) , and Hall and Martin (1988) suggested a bootstrap prepivoting method to increase the order of correctness of the bootstrap tests. A general theoretical treatment of power estimation is given by Beran (1986) and a variety of methods for resampling in multiple testing are discussed by Noreen (1989) and Westfall and Young (1993).

This report concentrates on a particular type of bootstrap test, the nonparametric bootstrap test, for a particular type of hypothesis, the generalized Behrens–Fisher problem (or the equality of means of populations when variances are unknown). Here, the main problem is the construction of the null model from which the replicates will be generated. Two situations are possible:

- The null model is based on some initial assumptions, like *equality in distributions* or *symmetry of distributions* — we denote it a semiparametric null model.

- The null model has no such initial assumption — we denote it a fully nonparametric model.

We propose to find answers for two main questions:

1. How accurate are the bootstrap p-values of the different test statistics proposed for the Behrens–Fisher problem? If these test statistics come from a parametrized family, what criteria may be used to determine the parameter value which maximizes this accuracy?

2. To what extent the fact that some of the initial assumptions used in the construction of the null model are false influences the properties of the bootstrap test (such as the difference between its true level and nominal levels or its power against a given alternative hypothesis).

Concerning the structure of this report, the next section contains a brief description of the standard theory of significance tests, of the notion of (non-parametric) bootstrap test, and a short presentation of the Behrens–Fisher problem. In section 3 we analyze a first semiparametric model, based on the assumption that all residuals come from the same distribution, and we consider two test statistics: the first, a test statistic proposed by Davison and Hinkley (1997) and the second, a robust version of the first. Besides experiments conducted to find responses to our two main questions, this section contains also a simulation study designed to verify if the distributions of the two bootstrap tests are indeed independent of the model. In section 4 we propose and study a theoretical framework for the problem of non-parametric combination of bootstrap tests and analyse an application of this procedure to another semiparametric model, based on the assumption of symmetry. The proposed bootstrap tests use different nonparametric combination functions, which do not imply the same performance, as the simulation experiments will show. Section 5 contains the analysis of a fully non-parametric model, to which our contribution is the proposition to use the Cressie–Read power-divergence measure as the divergence measure applied during the minimization procedure giving the null model. Because the Cressie–Read power-divergence measure is a parametrized measure, the simulation studies will try to establish the value of the parameter maximizing the properties of the corresponding bootstrap tests. Finally, a generalization of the notion of nonparametric likelihood function, based on the Cressie–Read power-divergence measure, is proposed in Section 6 and the main conclusions of this report are given in the last section.

2 A General Description

Statistical hypothesis testing can be generally described as follows. Let $X_1, \ldots, X_n$ be random p–vectors (not necessarily independent and identically distributed) having joint distribution $F^{(n)}$, and let $\mathcal{F}^{(n)}$ be the collection of all possible distributions $F^{(n)}$. Let $\mathcal{F}_0^{(n)}$ and $\mathcal{F}_1^{(n)}$ be two disjoint subsets of $\mathcal{F}^{(n)}$. We would like to use the data $X_1, \ldots, X_n$ to determine whether the hypothesis that $F^{(n)} \in \mathcal{F}_0^{(n)}$ is true, i.e., to test

$$H_0 : F^{(n)} \in \mathcal{F}_0^{(n)} \text{ versus } H_1 : F^{(n)} \in \mathcal{F}_1^{(n)}. \tag{2.1}$$

The notation H_0 denotes the null hypothesis and H_1 denotes the alternative hypothesis. In the special but important case where $X_1, \ldots, X_n$ are independent and identical distributed from F, $\mathcal{F}^{(n)}$ is determined by F and (2.1) reduces to

$$H_0 : F \in \mathcal{F}_0 \text{ versus } H_1 : F \in \mathcal{F}_1,$$

where $\mathcal{F}$ is the collection of all possible distributions F, and $\mathcal{F}_0$ and $\mathcal{F}_1$ are disjoint subsets of $\mathcal{F}$. Constructing a test for (2.1) is equivalent to finding a rejection region $\mathcal{R}_n$ such that we reject the null hypothesis H_0 if and only if $(X_1, \ldots, X_n) \in \mathcal{R}_n$. A simple and effective method, called the test statistic approach, is to use a test statistic $T_n = T_n(X_1, \ldots, X_n)$ and to define $\mathcal{R}_n = \{x : T_n(x) \geq c_n\}$, where c_n is called the *critical value*. The *rejection region* $\mathcal{R}_n$ (or the critical value c_n) is determined by controlling the probability of rejecting H_0 when H_0 is in fact true (Type I error),

$$\sup_{F^{(n)} \in \mathcal{F}_0^{(n)}} P\left\{(X_1, \ldots, X_n) \in \mathcal{R}_n \mid F^{(n)}\right\} = \alpha, \tag{2.2}$$

where α is given and $P\left\{ \cdot \mid F^{(n)}\right\}$ is the probability distribution corresponding to $F^{(n)}$. For a fixed n, if (2.2) holds, then the test with rejection region $\mathcal{R}_n$ is an exact level α test. Unless the problem under consideration is simple, an exact level α test is difficult or impossible to obtain. We then consider large n approximation, i.e., replace (2.2) by

$$\lim_{n \to \infty} \sup_{F^{(n)} \in \mathcal{F}_0^{(n)}} P\left\{(X_1, \ldots, X_n) \in \mathcal{R}_n \mid F^{(n)}\right\} = \alpha. \tag{2.3}$$

However, when $\mathcal{F}_0^{(n)}$ is complex, there may exist no tests satisfying (2.3) (Bahadur and Savage, 1956). This leads to the following definition (Shao and Tu, 1995, page 178).

DEFINITION 1 *Let α be a given nominal level. A test with rejection region $\mathcal{R}_n$ is asymptotically correct if* $\lim_{n\to\infty} P\{(X_1,\ldots,X_n)\in\mathcal{R}_n \mid H_0\} = \alpha$. *The test is consistent if* $\lim_{n\to\infty} P\{(X_1,\ldots,X_n)\in\mathcal{R}_n \mid H_1\} = 1$.

DEFINITION 2 *Let α be a given nominal level and T a test statistic. The power of the test when the alternative H_1 to the null hypothesis H_0 is of primary interest is*

$$\pi(\alpha, H_1) = P(T \geq t_\alpha \mid H_1),$$

where t_α is defined by $P(T \geq t_\alpha \mid H_0) = \alpha$.

The notation $P\{\cdot \mid H_0\}$ is equivalent to $P\{\cdot \mid F^{(n)}, \forall F^{(n)} \in \mathcal{F}_0^{(n)}\}$, whereas $P\{\cdot \mid H_1\}$ is equivalent to $P\{\cdot \mid F^{(n)}, \forall F^{(n)} \in \mathcal{F}_1^{(n)}\}$. A notion corresponding to that of a critical value is that of *significance probability*, (or the *error rate* P-value),

$$p = P\{T_n \geq T_n(x_1,\ldots,x_n) \mid H_0\},$$

which expresses the level of evidence against H_0 when H_0 is in fact true.

The distribution of T_n under H_0 is called the null distribution of T_n. Knowing this exactly or approximately means that we can calculate the P-value at least approximately and therefore we can take a decision about acceptance or rejection of H_0. In most parametric problems and all nonparametric problems, the null hypothesis H_0 is composite, that is, it leaves some parameters unknown and therefore does not completely specify $F^{(n)}$. Therefore the P-value is not generally well-defined, because it may depend upon which $F^{(n)}$ satisfy H_0. There are two standard solutions to this difficulty. One is to choose T_n so that its distribution is the same for all $F^{(n)}$ satisfying H_0: examples include the Student-t test for a normal mean with unknown variance and rank tests for nonparametric problems. The second and more widely applicable solution is to eliminate the parameters that remain unknown when H_0 is true by conditioning on the sufficient statistic under H_0. If S denotes this sufficient statistic, then we define the conditional P-value by $p = P\{T_n \geq T_n(x_1,\ldots,x_n) \mid S = s, H_0\}$.

If the distribution of T_n is well defined under H_0, resampling procedures may be used to approximate the exact P-value. The basic Monte Carlo test compares the observed statistics $t = T_n(x_1,\ldots,x_n)$ to R independent values of T_n which are obtained from samples independently simulated under a null hypothesis model. If these simulated values are denoted by $t_1^*,\ldots,t_R^*$, then under H_0 all $R+1$ values, $t, t_1^*,\ldots,t_R^*$ are equally likely values of T_n. That

is, assuming T_n continuous,

$$P\{T_n < T^*_{n(r)} \mid H_0\} = \frac{r}{R+1},$$

where $T^*_{n(r)}$ denotes the r^{th} ordered value. So, the estimated Monte Carlo P-value is defined as

$$p_{mc} = P\{T_n \geq t \mid H_0\} = \frac{1 + \#\{t^*_r \geq t\}}{R+1},$$

where $\#A$ means the number of times the event A occurs. An analysis of the loss of power of the test, expressed as a function of R, suggests that the number of sample replicates must be at least 99 and that $R = 999$ should generally be safe (Davison and Hinkley, 1997, p. 155).

2.1 Parametric Bootstrap Tests

In many problems the distribution of T_n under H_0 will depend upon nuisance parameters, which cannot be conditioned away, so that the Monte Carlo test method does not apply exactly. Then the natural approach is to fit the null model $\hat{F}_0$ and compute the P-values as $p = P\{T_n \geq t \mid \hat{F}_0\}$. For example, for the parametric model where we are testing $H_0 : \psi = \psi_0$ with λ a nuisance parameter, $\hat{F}_0$ would be the CDF of $f(y \mid \psi_0, \hat{\lambda}_0)$ with $\hat{\lambda}_0$ the maximum likelihood estimator of the nuisance parameter when ψ is fixed equal to ψ_0. Calculation of the P-value under $\hat{F}_0$ is referred to as a bootstrap test.

If the P-value cannot be computed exactly, or if there is no satisfactory approximation (normal or otherwise), then we proceed by simulation. That is, R independent samples $y^*_1, \ldots, y^*_n$ are drawn from $\hat{F}_0$, and for the r^{th} such sample the test statistic value t^*_r is calculated. Then the P-value will be approximated by

$$p_{boot} = \frac{1 + \#\{t^*_r \geq t\}}{R+1}.$$

2.2 Nonparametric Bootstrap Tests

The main difficulty in bootstrap hypothesis testing is the choice of the distribution $\hat{F}_0$, true under H_0, such that the distribution of the statistic T_n does not depend on this particular choice. When the model is parametric we saw in the preceding section that there is a natural choice of $\hat{F}_0$. But even if the model is nonparametric, there is a possible choice, using nonparametric maximum likelihood. The core of our work will be the study of how to obtain a nonparametric null model $\hat{F}_0$.

The problem we selected to illustrate some of the proposed procedures is the generalised Behrens–Fisher problem. The standard Behrens–Fisher problem (Behrens, 1929; Fisher, 1935) concerns the equality of the mean values of two normal distributions when their variances are unknown. Formally, we have $X_i \approx N(\mu_i, \sigma_i^2)$, $i = 1, 2$, the hypotheses being $H_0 : \{\mu_1 = \mu_2\}$ against $H_1 : \{\mu_1 \neq \mu_2\}$. We use the following notation: " $\approx$ " means "equal in distribution"; μ_i and $\bar{X}_i$ are the location parameter for the distribution of X_i and its estimate; σ_i and S_i denote the scale parameter, respectively its estimate. The same symbols, but without an index, refer to the pooled data; k is the number of samples, if there are more than two; $n_i \geq 2$ and $n = n_1 + n_2$ are the sample sizes and the pooled sample size.

The generalised Behrens–Fisher problem is obtained when the assumption of normality for X_i is relaxed and the hypothesis H_0 implies the equality of location parameters. The univariate Behrens–Fisher problem admits an exact non-randomised parametric solution for each $m = n_1/n_2$ only when $\rho = \sigma_1/\sigma_2$ is known (Welch, 1938; Nell et al., 1990). If ρ is unknown, an exact solution (at least in a parametric setting) cannot be obtained (Pfanzagl, 1974; Linnik, 1975; Lehmann, 1986). There are, however, simple and practical approximate solutions when the two ratios, m and ρ, are not far from one, like the Aspin–Welch procedure (Welch, 1938; Aspin, 1948; Wang, 1971; Fenstad, 1983; Nell et al., 1990). There is a rich literature concerning approximate solutions. Yuen (1974) proposed a solution based on trimmed estimates of parameters; Tiku and Sing (1981) used robust estimates of means; Jensen (1992) suggested saddlepoint approximations; Berger and Boos (1994) used P-value maximisation over a confidence interval for ρ; Ballin and Pesarin (1990) examined resampling techniques by using Aspin–Welch's types statistics; Pesarin (1995) used permutation tests under the assumption of symmetry. For the generalized Behrens–Fisher problem, a non-parametric approximate solution is given by the median test (Hettmansperger and Malin, 1975; Schlittgen, 1979). Other non-parametric conservative solutions based on ranks have been studied by Potthoff (1963), Flinger and Policello (1981) and Flinger and Rust (1982).

3 Semiparametric null model

A semiparametric model is a model with some but not all features of the underlying distribution described by parameters. Suppose initially that a possible model for data is

$$X_{ij} = \mu_i + \sigma_i \varepsilon_{ij},\ i = 1, \ldots, k,\ j = 1, \ldots, n_i,$$

where ε_{ij} all come from an unknown distribution G with mean zero and variance one. So, the hypothesis can be formalised as

$$H_0: \{\mu_1 = \cdots = \mu_k\} \text{ against } H_1: \{\exists\ i,j\ \ \mu_i \neq \mu_j\},$$

under the condition

$$\left\{ \frac{(X_1 - \mu_1)}{\sigma_1} \approx \frac{(X_2 - \mu_2)}{\sigma_2} \approx \cdots \approx \frac{(X_k - \mu_k)}{\sigma_k};\ \sigma_i \text{ unknown},\ i = 1 \ldots k \right\}.$$

One possible test statistic is

$$T = \sum\nolimits_i w_i (\bar{X}_i - \bar{X})^2, \tag{3.4}$$

where $\bar{X} = \sum_i w_i \bar{X}_i / \sum_i w_i$ and $w_i = n_i / s_i^2$ (Davison and Hinkley 1997, p. 163). But if the proposed model is not true, we may choose to use robust estimators for the mean and variance. Choosing as robust location estimator the median and as robust scale estimator the median deviation ($\tilde{s}_i = \text{median}\{\mid X_{ij} - \tilde{X}_i \mid\}$, where $\tilde{X}_i$ is the median of $\{X_{i\cdot}\}$), we obtain a test statistic of form

$$\tilde{T} = \max_i \tilde{w}_i (\tilde{X}_i - \tilde{X})^2, \tag{3.5}$$

where $\tilde{X} = \text{median}\{\tilde{X}_{ij}\}$ and $\tilde{w}_i = \sqrt{n_i / \tilde{s}_i}$. Two questions arise:

1. What is the fitted null model from which the bootstrap replicates will be simulated?

2. When different statistics are available, how can the bootstrap be used to choose the most efficient statistic for the observed data?

The fitted null model when the statistic (3.4) is used is

$$X_{ij} = \bar{x} + \hat{\sigma}_i \varepsilon_{ij} \tag{3.6}$$

where

$$\hat{\sigma}_i^2 = (n_i - 1) s_i^2 / n_i + (\bar{x}_i - \bar{x})^2$$

and studentized residuals

$$e_{ij} = (x_{ij} - \bar{x}) \left\{ \hat{\sigma}_i^2 - (\sum_i w_i)^{-1} \right\}^{-1/2}.$$

In the second case (statistic (3.5)), the null model became $X_{ij} = \tilde{x} + \tilde{\sigma}_i \varepsilon_{ij}$ with $\tilde{\sigma}_i = \text{median}\{\mid x_{ij} - \tilde{x} \mid\}$ and studentized residuals $e_{ij} = (x_{ij} - \tilde{x}) / \tilde{\sigma}_i$.

Table 1: Eight series of measurements of the acceleration due to gravity g, given as deviations from $980000 \times 10^{-3}\text{cm}s^{-2}$, in units of cm $s^{-2} \times 10^{-3}$ (Cressie, 1982).

Series							
1	2	3	4	5	6	7	8
76	87	105	95	76	78	82	84
82	95	83	90	76	78	79	86
83	98	76	76	78	78	81	85
54	100	75	76	79	86	79	82
35	109	51	87	72	87	77	77
46	109	76	79	68	81	79	76
87	100	93	77	75	73	79	77
68	81	75	71	78	67	78	80
	75	62			75	79	83
	68				82	82	81
	67				83	76	78
						73	78
						64	78

Because we supposed that all residuals come from the same distribution, for a bootstrap test we simulate from the null model $x^*_{ij} = \bar{x} + \hat{\sigma}_i e^*_{ij}$ (respectively $x^*_{ij} = \tilde{x} + \tilde{\sigma}_i e^*_{ij}$) where e^*_{ij} are randomly sampled from the pooled residuals $\{e_{ij}\}$, $i = 1, 2;\ j = 1, \ldots, n_i$. The EDF of the pooled residuals puts probability mass $(\sum n_i)^{-1}$ on each residual.

A first set of data (Table 1) contains eight series of measurements related to the acceleration gravity g, due to Cressie (1982). Table 2 contains a summary of first null model fit, with $\bar{x} = 78.6$ and $t = 21.275$. Table 3 contains the same information, but for the second null model fit, with $\tilde{x} = 78$ and $\tilde{t} = 256.171$. The bootstrap P-value for the statistic T is 0.030 (using 49, 000 replicates) and 0.098 for the statistic $\tilde{T}$. For an error rate $\alpha = 0.05$, we reject the hypothesis based on T and we accept it based on $\tilde{T}$. A normal-error parametric bootstrap gives a P-value close to those of T.

Having different test statistics for the same set of data, a natural choice is to consider the test with the greatest power under the alternative hypothesis H_1. From the bootstrap viewpoint, the power can be estimated by sampling from a distribution $\hat{F}_1$ true under H_1. Bootstrap estimation of power is even more complicated than bootstrap estimation of a P-value: we must know the critical value t_α, which is the $(1 - \alpha)$ quantile of the distribution of T_n under H_0 and we must find a distribution $\hat{F}_1$ such that the distribution of T_n under H_1 depends little on this particular choice.

In our case, the alternative hypothesis is "there are at least two different means". Because we want to study the power of the test against the closer

Table 2. Summary statistics for the eight samples in gravity data, using the first model.

i	$\bar{x}_i$	s_i^2	σ_{i0}^2	w_i
1	66.4	370.6	474.4	0.022
2	89.9	233.9	339.9	0.047
3	77.3	248.3	222.3	0.036
4	81.4	68.8	67.8	0.116
5	75.3	13.4	23.1	0.599
6	78.9	34.1	31.1	0.323
7	77.5	22.4	21.9	0.579
8	80.4	11.3	13.5	1.155

Table 3. Summary statistics for the eight samples in gravity data, using the second model.

i	$\widetilde{x}_i$	$\widetilde{s}_i^2$	$\widetilde{\sigma}_{i0}^2$	$\widetilde{w}_i$
1	72.0	169	90.25	0.784
2	95.0	196	289	0.886
3	76.0	49	25	1.133
4	78.0	20.25	20.25	1.133
5	76.0	4	4	2.000
6	78.0	16	16	1.658
7	79.0	4	1	2.549
8	80.0	9	4	2.081

alternative hypothesis (H_A ="there are exactly two different means"), we choose as alternative model the model (here, for the test statistic (3.4))

$$\begin{cases} X_{ij} &= \bar{x}_{(i_0)} + \hat{\sigma}_i \varepsilon_{ij},\ i = 1, \ldots, k,\ i \neq i_0, \\ X_{i_0 j} &= \bar{x}_{i_0} + \hat{\sigma}_{i_0} \varepsilon_{i_0 j}, \end{cases} \tag{3.7}$$

where

$$\begin{aligned} \bar{x}_{(i_0)} &= \frac{\sum_{i \neq i_0} w_i \bar{x}_i}{\sum_{i \neq i_0} w_i} \\ \hat{\sigma}_{i \neq i_0} &= (n_i - 1)\frac{s_i^2}{n_i} + (\bar{x}_i - \bar{x}_{(i_0)})^2 \\ \hat{\sigma}_{i_0} &= s_{i_0}^2 \end{aligned}$$

and i_0 is the index maximizing

$$\left| \bar{x}_i - \frac{\sum_{j \neq i} w_j \bar{x}_j}{\sum_{j \neq i} w_i} \right|. \tag{3.8}$$

So, the sample with the biggest distance between its average and the common average of all others is considered as coming from a distribution with a different mean. According to this model, we define the distance between H_0 and H_1 as

$$\delta(H_0, H_1) = \left| \frac{\bar{x}_{(i_0)} - \bar{x}_{i_0}}{se(\bar{x})} \right| = \left| (\bar{x}_{(i_0)} - \bar{x}_{i_0}) \sum w_i \right| \tag{3.9}$$

To study the dependence of the power of the two tests on the distance $\delta(H_0, H_1)$ we simulated from the following model, where the sample selected to have a mean different from the common mean of all other samples is one

of the two samples having the largest variance and the largest size:

$$x_{ij} = 5 + \delta_i + \sigma_i \varepsilon_{ij}, \quad i = 1, \ldots, 6,$$
$$\text{where} \quad \delta_{\text{i}} = 0 \quad \text{if} \quad \text{i} \neq 3 \quad \text{and} \quad \delta_3 = \Delta$$
$$\sigma_1 = \sigma_4 = 9, \quad \sigma_2 = \sigma_5 = 16, \quad \sigma_3 = \sigma_6 = 36,$$
$$n_1 = n_4 = 10, \quad n_2 = n_5 = 20, \quad n_3 = n_6 = 30,$$
$$G_i = N(0, 1), \quad i = 1, \ldots, 6.$$

During the first phase of the experiment we estimate the critical values t_α for the statistics T and $\tilde{T}$, under the null hypothesis. To do this we simulate $R = 50000$ data replicates from the previous model having $\Delta = 0$ and we take the $[R(1-\alpha)]^{th}$ ordered value in the sequence $t_1^*, \ldots, t_R^*$. During the second phase, we simulate, for each $\delta(H_0, H_1)$ fixed, one thousand data replicates. For each dataset d, we fit the alternative null model (3.7) for statistic T and the corresponding alternative null model for statistic $\tilde{T}$. Using $R = 10000$ bootstrap replicates, we estimate the power of the two tests for each $\alpha \in \{0.01, 0.05, 0.1\}$. For each $\delta(H_0, H_1)$ and each α, the value of the power will be calculated as the average of 1000 individual values. These values also allow us to estimate the standard error of bootstrap power estimator.

We must remark that the alternative null model does not consider automatically that the third sample of the simulated data d is coming from a distribution with a different mean (as the simulation model is defined). In fact, the alternative null model is data-driven, which implies the fact that a general conclusion of type "the statistic T_1 has a greater power, for a given size, than the statistic T_2" is not true for each dataset.

As we can see in Figure 1, the test T has a greater power than the test $\tilde{T}$ for all values of $\delta(H_0, H_1)$, if the size is 0.1. For smaller sizes and for $\delta(H_0, H_1) \leq 10$, the test $\tilde{T}$ has a greater power than the test T (the standard error of the bootstrap estimation of power, for sizes 0.05 and 0.01, are less than 0.001, so the affirmation "greater" is statistically correct). Hence, we suggest to use the test $\tilde{T}$ if the null hypothesis is rejected for $\alpha = 0.05$ or $\alpha = 0.01$ and if the distance $\delta(H_0, H_1)$ is small enough. But we also must emphasize that this recommendation is based only a single model simulation and therefore a generalisation of our conclusion is hard to defend. Indeed, if we change the distribution of residuals, from normal distribution to Cauchy distribution, and repeat the simulation, we obtain an interesting result, presented in Figure 2. As we can see, the power of the test T remains almost constant for each size, while the power of the test $\tilde{T}$ grows with $\delta(H_0, H_1)$. The reason is due to the fact that Cauchy distribution

has not a finite mean and so the distance between the hypothesis H_0 and H_1 hasn't any influence on the distribution of the statistic T. But as the same Cauchy distribution is symmetric, the median is a sensible estimator for location parameter, and so the distribution of statistic $\tilde{T}$ is influenced by $\delta(H_0, H_1)$. Consequently, for small values of $\delta(H_0, H_1)$, the test T must be preferred, whereas for large values, the test $\tilde{T}$ is more appropriate.

Figure 1: Bootstrap power estimation for tests T and $\tilde{T}$ versus different distances $\delta(H_0, H_1)$, when residuals where simulated from a normal distribution. (LEGEND: left – size = 0.01, center – size = 0.05, right – size = 0.1)

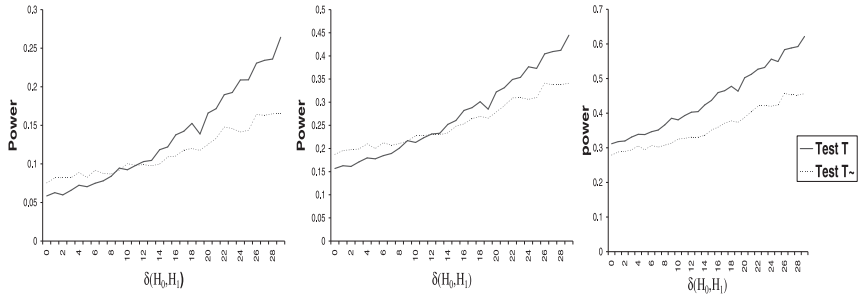
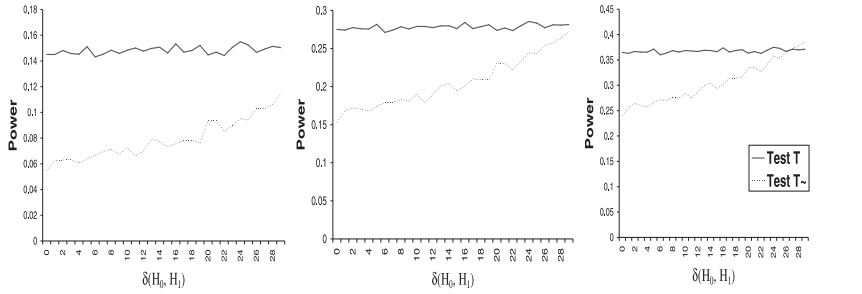


Figure 2: Bootstrap power estimation for tests T and $\tilde{T}$ versus different distances $\delta(H_0, H_1)$, when residuals where simulated from a Cauchy distribution. (LEGEND: left – size = 0.01, center – size = 0.05, right – size = 0.1)



To construct the null model from which the replicates are simulated (see 3.6 as example), two kind of assumption are considered. The first is derived from the null hypothesis H_0 (in our case, the null model considers the same

mean $\bar{x}$ for all samples X_i). The second is derived from initial supposition about the true model (in our case, the fact that all studentized residuals ε_{ij} come from a single distribution). By consequence, the efficiency of the bootstrap test may be estimated regarding the two types of assumptions. The preceding experience, concerning the bootstrap power estimation, was conducted by taking the first assumption false (which implies that the alternative hypothesis H_1 is true) but keeping the second assumption true. Now we will consider the first assumption true and the second false (residuals generated by more than one distribution) and we will examine the influence on the size of two bootstrap tests. For this we simulate 10 000 data sets from a null model with two types of residual generators, having the mean zero and the variance one:

$$\begin{gathered} x_{ij} = 5 + \sigma_i \varepsilon_{ij}, \quad i = 1, \ldots, 6, \\ \sigma_1 = \sigma_4 = 9, \quad \sigma_2 = \sigma_5 = 16, \quad \sigma_3 = \sigma_6 = 36, \\ n_1 = n_4 = 10, \quad n_2 = n_5 = 20, \quad n_3 = n_6 = 30, \\ G_1 = G_2 = G_3 = N(0,1) \text{ and } G_4 = G_5 = G_6 = \sqrt{3}\, U(-1,1). \end{gathered}$$

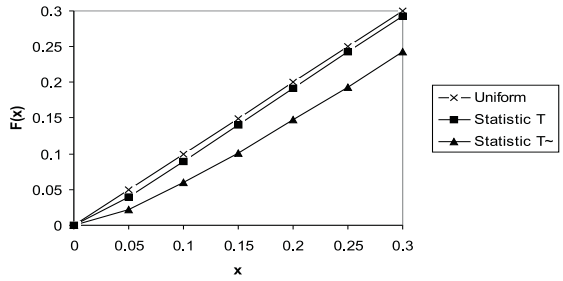
This type of model will be often used within the framework of this report (of course, with some modifications fitted to the concrete situation). We chose a model with more than two samples (here, six) to grow the complexity of the generalized Behrens–Fisher problem. The samples have different sizes (ten, twenty and thirty), varying from approximately small to approximately large and have different variances, but greater for large samples (to introduce more variability in the data). If two distributions are used to generate data, the samples generated by each of them are identical in number, size and variance. If a single distribution is used, we may give up the symmetry condition and vary the variance or the size.

Theoretically, the significance probability has a uniform distribution on (0,1) when the hypothesis H_0 is true. Owing to the fact that the null model used to obtain replicates of data is not exactly correct (the assumption of identical distributions being false), the distribution of the bootstrap P-value (calculated using 19 999 replications) is not uniform for either of the statistics. A Kolmogorov–Smirnoff test for uniformity gives a P-value of 0.13 for the bootstrap P-values of statistic T and a P-value less than 10^{-6} for the bootstrap P-values of statistic $\tilde{T}$.

Figure 3 shows the cumulative empirical distribution for the two samples of bootstrap P-values, restricted to the interval $[0,\ 0.3]$, and, for comparison, the uniform distribution on the same interval. The most important region on x-axis is $(0,\ 0.1)$, because if the P-value falls here, the hypothesis is usually

rejected. As we see, the probability of rejection for the statistic T is greater than for the statistic $\tilde{T}$, so it is preferable to use the robust statistic when we have doubts above the assumption of equal distributions for residuals.

Figure 3: Empirical cumulative distributions for two samples of bootstrap P-values (10000 data), corresponding of the two statistics, T and $\tilde{T}$, restricted to the interval [0, 0.3].



In the last experiment we propose to study the dependence of the distribution of the two statistics on the choice of G distribution. For each distribution G having location parameter equal zero and scale parameter equal one (see Table 4) we generate 10 000 sets of data using the null model $x_{ij} = 5 + i^2\varepsilon_{ij}$. Each set contains ten series, with different lengths, $n_i \in \{2, 2, 3, 3, 5, 5, 7, 7, 9, 9\} \times c$. For $c = 3$, we denote a "small experiment", for $c = 5$ a "medium experiment" and for $c = 9$ a "big experiment". For a more convenient notation, we will denote by (F)distribution the distribution of the test statistics when the residuals follow the F distribution.

Table 4: Different data generator distributions, with mean zero and variance one

Distribution	Type	Mean zero, variance one
Normal(μ, σ)	continuous	N$(0, 1)$
Uniform(a, b)	continuous	$\sqrt{3}\cdot$Uniform(-1, 1)
Beta(r, s)	continuous	$3\sqrt{2}\cdot$(Beta(1,2)-1/3)
Chi-square(n)	continuous	$(2\sqrt{5})^{-1} \cdot (\chi^2(10) - 10)$
Exponential(λ)	continuous	Exp(1)-1
Student(t)	continuous	$\frac{2}{\sqrt{5}}\cdot$ Student(10)
Cauchy(n)	continuous	Cauchy(1)
Laplace(α, β)	continuous	$\sqrt{2}^{-1}\cdot$ Laplace(0,1)

A first conclusion (see Table 5 and Table 6) is that both statistics express a dependence between their distribution and the distribution of the residual

(almost all Kolmogorov–Smirnoff tests of equality of distribution are rejected). There are still some "exceptions", like the pairs (Normal, Student) or (Laplace, Student), but only for the T statistic. We may also remark that if the sets of data contain more and more observations, the dependence between the statistic distribution and the residual distribution diminishes (see Table 5, for the pairs (Normal, Uniform) or (Beta, Chi-square)).

Table 5: The P-values for Kolmogorov–Smirnoff tests of equality in distribution for T statistic, for each pair of data generator distributions (Legend: normal font – small experiment, *italic* – medium experiment, **bold** – big experiment)

P-value	Normal	Beta	Chi-square	Cauchy	Exp.	Laplace	Student
Beta	0						
	0	—					
	0						
Chi-square	0	0.090					
	0	*0.120*	—				
	0	**0.230**					
Cauchy	0	0	0				
	0	*0*	*0*	—			
	0	**0**	**0**				
Exp.	0	0	0	0			
	0	*0*	*0*	*0*	—		
	0	**0**	**0**	**0**			
Laplace	0	0	0	0	0		
	0.127	*0*	*0*	*0*	*0*	—	
	0.348	**0**	**0**	**0**	**0**		
Student	0.890	0	0	0	0	0.001	
	0.600	*0*	*0*	*0*	*0*	*0.154*	—
	0.872	**0.078**	**0.002**	**0**	**0**	**0.139**	
Uniform	0.008	0	0	0	0	0	0.002
	0.175	*0.001*	*0*	*0*	*0*	*0.007*	*0.001*
	0.405	**0.005**	**0**	**0**	**0**	**0.140**	**0.250**

For each generated set of data we calculated the corresponding bootstrap P-value using $R = 9999$ replicates. The plot of "theoretical" P-value (calculated using all 10 000 sets of data) versus the bootstrap P-value gives us some important information. The most "visible" is that, for the statistic T, the bootstrap P-value is almost always greater than the theoretical P-value, whereas for the statistic $\tilde{T}$, the points $x_i =$(theoretical P-value, bootstrap P-value) have a pronounced bi-normal distribution (see Figure 4 and Figure 5).

Table 6: The P-values for Kolmogorov–Smirnoff tests of equality in distribution for $\tilde{T}$ statistic, for each pair of data generator distributions (Legend: normal font – small experiment, *italic* – medium experiment, **bold** – big experiment)

P-value	Normal	Beta	Chi-square	Cauchy	Exp.	Laplace	Student
Beta	0	—					
	0						
	0						
Chi-square	0	0	—				
	0	*0*					
	0	**0**					
Cauchy	0	0	0	—			
	0	*0*	*0.017*				
	0	**0**	**0**				
Exp.	0	0	0	0	—		
	0	*0*	*0*	*0*			
	0	**0**	**0**	**0**			
Laplace	0	0	0	0	0	—	
	0	*0*	*0*	*0*	*0*		
	0	**0**	**0**	**0**	**0**		
Student	0	0	0	0	0	0	—
	0	*0*	*0*	*0*	*0*	*0*	
	0	**0**	**0**	**0**	**0**	**0**	
Uniform	0	0.067	0	0	0	0	0
	0	*0*	*0*	*0*	*0*	*0*	*0*
	0	**0**	**0.002**	**0**	**0**	**0**	**0**

This pattern is especially true for the class of symmetric distributions having finite moments (as in our study, normal, Student, Laplace and uniform distribution). For a second class, of non-symmetric distributions (beta, chi-square and exponential), the pattern of points converges to those of the first class, if the statistic T is used, but converges to an opposite of this pattern, if the statistic $\tilde{T}$ is used (see Figure 6 and Figure 7). This is due to the fact that, for these distributions, the mean is different from the median, and so the model from which the replicates are generated if the statistic $\tilde{T}$ is used (implying the equality of medians), is false. Finally, a last class contains a single distribution, non-symmetric and without finite moments, the Cauchy distribution. In this case, the patterns for both statistics are similar of those of the first class (see Figure 8 and Figure 9).

Figure 4: Theoretical P-value vs. bootstrap P-value for the test using T statistic, when data were generated using a normal distribution. The theoretical P-values were calculated using 10000 data and the corresponding bootstrap P-values using 9999 replicates. (LEGEND: left – small, center – medium, right – large experiment)

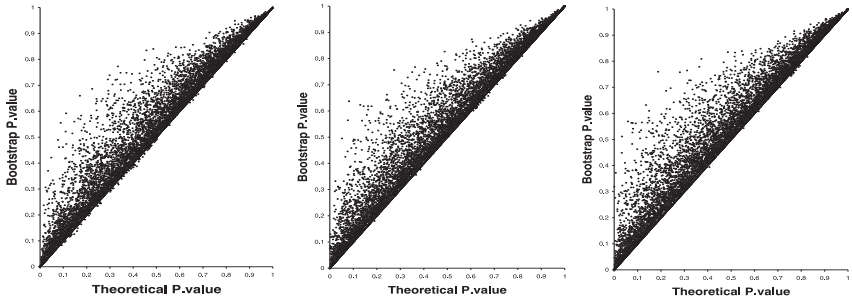


Figure 5: Theoretical P-value vs. bootstrap P-value for the test using $\tilde{T}$ statistic, when data were generated using a normal distribution. The theoretical P-values were calculated using 10000 data and the corresponding bootstrap P-values using 9999 replicates. (LEGEND: left – small, center – medium, right – large experiment)

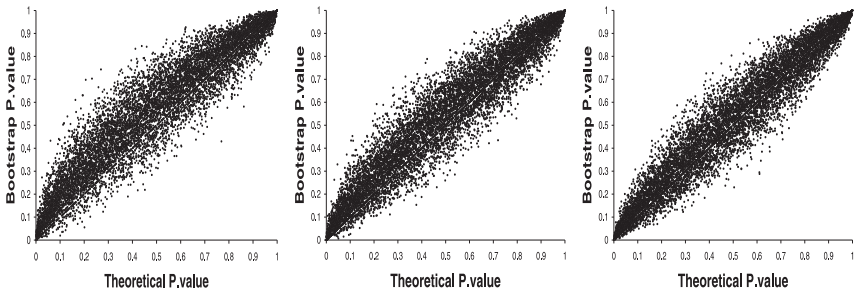
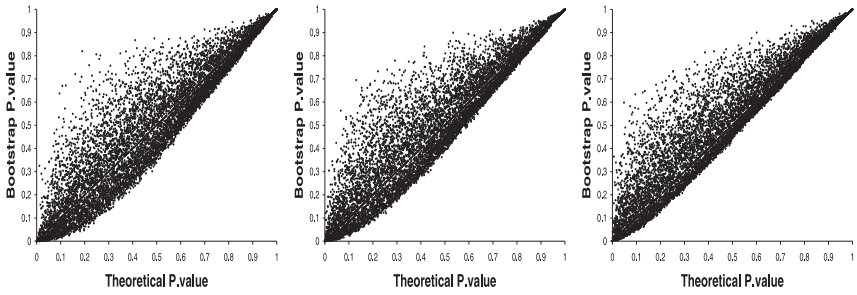


Figure 6: Theoretical P-value vs. bootstrap P-value for the test using T statistic, when data were generated using an exponential distribution. The theoretical P-values were calculated using 10000 data and the corresponding bootstrap P-values using 9999 replicates. (LEGEND: left – small, center – medium, right – large experiment)



To summarize the distribution of the points x_i and to capture the performances of bootstrap tests, we proposed three measures. The first is represented by the true level of the bootstrap tests for given nominal level α ($\alpha \in \{0.01, 0.05, 0.1\}$), i.e Prob(bootstrap P-value $< \alpha$) (see Table 7). In the case of statistic T, the true level is always smaller than the nominal level. More than, as long as the dimension of data grows, the true level goes down to the same values regardless of the data generator distribution. If the statistic $\tilde{T}$ is considered, the true level goes up as the dimension of data grows, for all data generator distributions. But the values towards the true level converges depend on the shape of distributions: if the distribution is symmetric, these values are very close of those obtained when statistic T is considered. If the distribution is non-symmetric, these values are different from those corresponding to statistic T and are larger than the nominal level (up to 0.6 for $\alpha = 0.1$. Therefore, we may conclude that, if data are generated using symmetric distributions, the bootstrap tests T and $\tilde{T}$ will reject the true hypothesis with a probability smaller than a given nominal level α (in other words, the bootstrap tests are more "optimistic" than the theoretical tests). But if data are generated using non-symmetric distributions and the statistic $\tilde{T}$ is considered, the bootstrap test will massively reject the hypothesis.

Because we dispose, for each set of data, of the theoretical P-value and the corresponding bootstrap P-value, we may consider two measures similar with the types of errors defined for a test statistic (type I, i.e. the probability to reject H_0 when the null hypothesis is true, and type II, the probability to accept H_0 when the alternative hypothesis is true). So we define error-like type I measure (denoted EI) as the probability that the bootstrap test rejects the hypothesis when the theoretical test accepts it (or Prob(bootstrap P-value $< \alpha$ and theoretical P-value $> \alpha$)). Because the acceptance/rejection is done in correlation with a given level, we calculate this measure for $\alpha \in \{0.01, 0.05, 0.1\}$. Similarly, we define error-like type II measure (or EII) the probability that the bootstrap test accept the hypothesis whereas the theoretical test reject it, for a given level α. These two measures are calculated in Table 8, for each statistic and each data generator distribution. A first remark is that the behavior of the measures are opposite

Figure 7: Theoretical P-value vs. bootstrap P-value for the test using $\tilde{T}$ statistic, when data were generated using an exponential distribution. The theoretical P-values were calculated using 10000 data and the corresponding bootstrap P-values using 9999 replicates. (LEGEND: left – small, center – medium, right – large experiment)

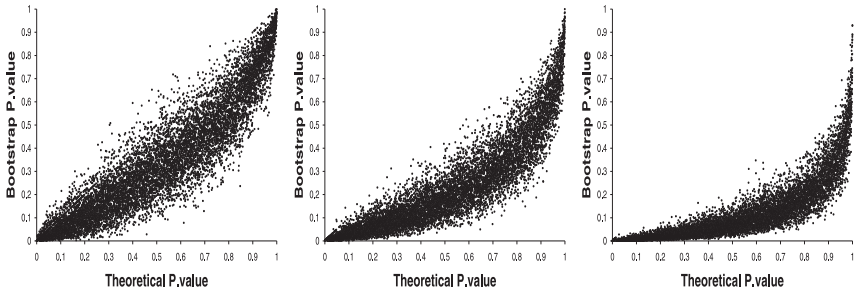


Figure 8: Theoretical P-value vs. bootstrap P-value for the test using T statistic, when data were generated using a Cauchy distribution. The theoretical P-values were calculated using 10000 data and the corresponding bootstrap P-values using 9999 replicates. (LEGEND: left – small, center – medium, right – large experiment)

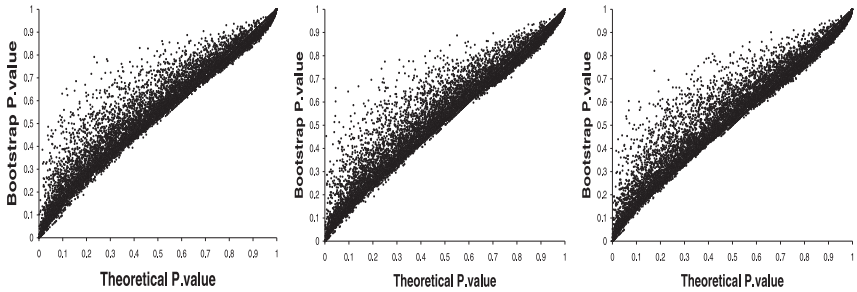


Figure 9: Theoretical P-value vs. bootstrap P-value for the test using $\tilde{T}$ statistic, when data were generated using a Cauchy distribution. The theoretical P-values were calculated using 10000 data and the corresponding bootstrap P-values using 9999 replicates. (LEGEND: left – small, center – medium, right – large experiment)

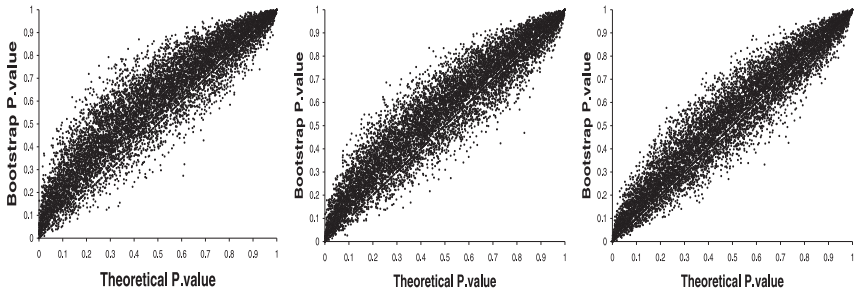


Table 7: The true levels of the bootstrap tests for different nominal levels α and different data generator distributions (LEGEND: normal font — small experiment, *italic* — medium experiment, **bold** — big experiment)

Distribution	Statistic T			Statistic $\tilde{T}$		
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
Beta	0.0811	0.0419	0.0081	0.0746	0.0346	0.0057
	0.0723	*0.0327*	*0.0071*	*0.106*	*0.0498*	*0.0078*
	0.0641	**0.0286**	**0.0059**	**0.1731**	**0.0924**	**0.0156**
Chi-square	0.073	0.0362	0.0081	0.0663	0.0274	0.0042
	0.0677	*0.0327*	*0.0065*	*0.1044*	*0.0482*	*0.0083*
	0.0642	**0.0302**	**0.0067**	**0.192**	**0.0976**	**0.0171**
Cauchy	0.0429	0.0206	0.0063	0.0309	0.0109	0.0012
	0.0447	*0.0190*	*0.0052*	*0.0473*	*0.0180*	*0.0022*
	0.0420	**0.0177**	**0.0038**	**0.0640**	**0.0256**	**0.0051**
Exponential	0.1096	0.0655	0.0218	0.1984	0.1113	0.0285
	0.0914	*0.0508*	*0.0136*	*0.3309*	*0.1858*	*0.0433*
	0.0712	**0.0358**	**0.0086**	**0.6032**	**0.4127**	**0.138**
Laplace	0.0638	0.0291	0.0080	0.0372	0.0124	0.0014
	0.0660	*0.0306*	*0.0065*	*0.0526*	*0.0211*	*0.0032*
	0.0612	**0.0302**	**0.0063**	**0.0649**	**0.0281**	**0.0046**
Normal	0.0678	0.0316	0.0085	0.0444	0.0178	0.0022
	0.0648	*0.0296*	*0.0071*	*0.0614*	*0.0264*	*0.0030*
	0.0640	**0.0297**	**0.0069**	**0.0730**	**0.0309**	**0.0057**
Student	0.0733	0.0347	0.0074	0.0449	0.0167	0.0020
	0.0669	*0.0318*	*0.0082*	*0.0587*	*0.0247*	*0.0043*
	0.0649	**0.0303**	**0.0057**	**0.0686**	**0.0301**	**0.0059**
Uniform	0.0759	0.0385	0.0092	0.0476	0.0193	0.0029
	0.0679	*0.0339*	*0.0079*	*0.0551*	*0.0222*	*0.0042*
	0.0654	**0.0313**	**0.0074**	**0.0729**	**0.0345**	**0.0056**

(if EI goes up, then EII goes down), but are different for each statistic (if the statistic T is considered, then EI goes down if the dimension of data goes up; for statistic $\tilde{T}$, we have an inverse process). As we expected, looking to the pattern of the points x_i for statistic T and symmetric distributions, the EI values are zero and the EII values are greater than zero (always the bootstrap P-value is greater than the theoretical P-value). As an observation, the maximum difference between bootstrap P-value and theoretical P-value is around 0.6, for almost all distributions. On the contrary, for statistic $\tilde{T}$ and non-symmetric distributions, the EII values are close to zero and EI values greater than zero (as we mentioned, because we get replicates from a wrong model, the bootstrap P-values are always smaller than the theoretical P-values).

As a general conclusion, if we apply bootstrap test T for a particular set of data and the decision is against the null hypothesis, we may be sure that the theoretical test would gave the same result. The same conclusion is valid for bootstrap test $\tilde{T}$, but only if the distribution of residuals is symmetric. On the other hand, if the distribution of residuals is non-symmetric and the bootstrap test $\tilde{T}$ implies an acceptance of the null hypothesis, we may be sure that it is a correct decision.

Table 8: The probability that, for a given set of data, the bootstrap P-value test is less than a nominal level α and the corresponding theoretical P-value is greater than the same level (Error-like type I) and the probability that, for a given set of data, the bootstrap P-value test is greater than a nominal level α and the corresponding theoretical P-value is less than the same level (Error-like type II)(LEGEND: normal font — small experiment, *italic* — medium experiment, **bold** — big experiment)

Distribution	Statistic T						Statistic $\tilde{T}$					
	$\alpha = 0.1$		$\alpha = 0.05$		$\alpha = 0.01$		$\alpha = 0.1$		$\alpha = 0.05$		$\alpha = 0.01$	
	E I	E II	E I	E II	E I	E II	E I	E II	E I	E II	E I	E II
Beta	0.005	0.024	0.048	0.012	0.000	0.002	0.009	0.035	0.007	0.022	0.001	0.005
	0.002	*0.030*	*0.000*	*0.017*	*0.000*	*0.002*	*0.023*	*0.017*	*0.010*	*0.010*	*0.000*	*0.003*
	0.000	**0.035**	**0.000**	**0.021**	**0.000**	**0.004**	**0.075**	**0.003**	**0.044**	**0.002**	**0.005**	**0.000**
Chi-square	0.004	0.031	0.003	0.016	0.000	0.001	0.008	0.041	0.004	0.026	0.000	0.006
	0.002	*0.034*	*0.001*	*0.019*	*0.000*	*0.003*	*0.022*	*0.018*	*0.011*	*0.013*	*0.001*	*0.003*
	0.002	**0.037**	**0.001**	**0.020**	**0.000**	**0.003**	**0.093**	**0.001**	**0.049**	**0.001**	**0.007**	**0.000**
Cauchy	0.000	0.057	0.000	0.029	0.000	0.004	0.001	0.070	0.001	0.040	0.000	0.009
	0.000	*0.055*	*0.000*	*0.031*	*0.000*	*0.005*	*0.002*	*0.054*	*0.001*	*0.033*	*0.000*	*0.008*
	0.000	**0.058**	**0.000**	**0.032**	**0.000**	**0.006**	**0.004**	**0.040**	**0.002**	**0.026**	**0.000**	**0.005**
Exponential	0.032	0.022	0.024	0.008	0.012	0.000	0.102	0.003	0.064	0.003	0.016	0.001
	0.022	*0.031*	*0.016*	*0.016*	*0.006*	*0.002*	*0.231*	*0.000*	*0.136*	*0.000*	*0.033*	*0.000*
	0.012	**0.040**	**0.006**	**0.021**	**0.001**	**0.003**	**0.503**	**0.000**	**0.363**	**0.000**	**0.128**	**0.000**
Laplace	0.000	0.036	0.000	0.021	0.000	0.002	0.001	0.064	0.000	0.038	0.000	0.009
	0.000	*0.034*	*0.000*	*0.019*	*0.000*	*0.003*	*0.003*	*0.050*	*0.002*	*0.030*	*0.000*	*0.007*
	0.000	**0.039**	**0.000**	**0.020**	**0.000**	**0.004**	**0.004**	**0.039**	**0.002**	**0.024**	**0.000**	**0.005**
Normal	0.000	0.032	0.000	0.018	0.000	0.001	0.003	0.058	0.001	0.033	0.000	0.008
	0.000	*0.035*	*0.000*	*0.020*	*0.000*	*0.003*	*0.003*	*0.042*	*0.002*	*0.026*	*0.000*	*0.007*
	0.000	**0.036**	**0.000**	**0.020**	**0.000**	**0.003**	**0.006**	**0.033**	**0.002**	**0.021**	**0.000**	**0.004**
Student	0.000	0.027	0.000	0.015	0.000	0.003	0.003	0.058	0.001	0.034	0.000	0.008
	0.000	*0.033*	*0.000*	*0.018*	*0.000*	*0.002*	*0.004*	*0.045*	*0.001*	*0.026*	*0.000*	*0.006*
	0.000	**0.035**	**0.000**	**0.020**	**0.000**	**0.004**	**0.004**	**0.035**	**0.002**	**0.022**	**0.000**	**0.004**
Uniform	0.000	0.024	0.000	0.011	0.000	0.001	0.003	0.055	0.001	0.032	0.000	0.007
	0.000	*0.032*	*0.000*	*0.016*	*0.000*	*0.002*	*0.002*	*0.047*	*0.001*	*0.029*	*0.000*	*0.006*
	0.000	**0.035**	**0.000**	**0.019**	**0.000**	**0.003**	**0.005**	**0.032**	**0.002**	**0.017**	**0.000**	**0.004**

4 Non-parametric combination of bootstrap tests for multidimensional hypotheses

It is possible to extend the meaning of word *semiparametric* to include some models apparently without parameters. For example, the condition that the residuals ε_{ij} come from an unknown distribution G, for all i, j, can be relaxed by considering different G_i, but all having in common a symmetric distribution. So, if H_0 is true, the variables $Y_i = X_i - \bar{X}$ are symmetrically distributed around zero. The null hypothesis is then equivalent to

$$H_0 : \{P(Y_i < -z) = P(Y_i > z),\ i = 1, \ldots, k,\ \text{for all}\, z \in R\}$$

or

$$H_0 = \{\cap_i H_{0i}\} \text{ where } H_{0i} : \{P(Y_i < -z) = P(Y_i > z),\ \text{for all } z \in R\}.$$

In many statistical analyses of complex hypothesis testing (as in our case), when many parameters are involved or many different aspects are of interest, it is sometimes convenient to first process data by a finite set of $k > 1$ different first-order one-dimensional tests, each being chosen to deal with a particular aspect of interest for the analysis. In a second phase, a second-order one-dimensional test will combine all these. If first-order test statistics were stochastically independent, this combination would be easily solved (Folks, 1984). But, in most situations it is impossible to invoke such a complete independence of first-order test statistics, both because they are functions of the same data X and because component random variables in X are generally assumed to be dependent. Moreover, the underlying dependence structure among first-order tests is known very rarely, except in case of multinormality. And when it is known often it is very difficult to handle. Therefore, this combination must be done nonparametrically, especially with respect to the underlying dependence structure.

The literature does contain some few references on the combination of dependent test statistics. Westberg (1986) considers some cases of robustness with respect to the dependence of an adaptive Tippett combination procedure. Berk and Jones (1978) discuss how Bahadur's asymptotic relative efficiency property relates to a Tippett combination procedure in case of dependence, but they do not provide any practical solution. Conservative solutions could be found via Bonferroni's inequality (Westfall and Young, 1993), in analogy with multiple comparison methods (Zanella, 1973). The major results on the non-parametric combination of several dependent permutation tests were obtained by Pallini and Pesarin (Pallini and Pesarin,

1990; Pallini and Pesarin, 1992) and Pesarin (Pesarin, 1988; Pesarin, 1989; Pesarin, 1990; Pesarin, 1991; Pesarin, 1992).

The main assumptions on the structure of data, on the set of first-order tests and on the set of hypotheses used in the non-parametric combination context are defined as follow:

i. $\mathbf{X} = (X_1, \ldots, X_k)$ represents the data set coming from a sample space $\aleph$, for which a σ-algebra and a family $\mathcal{F}$ of non-degenerate distributions are assumed to exist. Data consist of $k \geq 2$ samples of size $n_i \geq 2$; they are supposed independent, with distributions $F_i \in \mathcal{F}$. In place of independence, sometimes exchangeability may suffice.

ii. H_0 is supposed to be suitably decomposed into m sub-hypotheses $H_{0i}, i = 1, \ldots, m$, each appropriate for partial aspects. Thus H_0 is true if all H_{0i} are jointly true; i.e. $H_0 = \{\cap_{1 \leq i \leq m} H_{0i}\}$.

iii. H_1 states that at least one of the null sub-hypotheses is false; hence it is represented by a union of k sub-alternatives, i.e. $H_1 = \{\cup_{1 \leq i \leq m} H_{1i}\}$. Note that all known or unknown parameters not involved in the statements implied by the system of hypotheses are considered to be nuisance entities and are assumed to maintain the same value under both H_0 and H_1.

iv. $\mathbf{T} = \mathbf{T}(\mathbf{X})$ represents a m-dimensional vector of test statistics whose components $T_i = T_i(\mathbf{X})$, $i = 1, \ldots, m$ represent the first-order univariate and non-degenerate tests appropriate for the component sub-hypothesis H_{0i} against H_{1i}. Without loss of generality, the corresponding first-order bootstrap tests are assumed to be PUOD (Positively Upper Orthant Dependent), (Dharmadhikari and Joag-Dev, 1988), i.e.

   (a) statistically significant for large values, i.e., under H_1 their distributions are stochastically larger than under H_0, thus

   $$P(T_i \leq t \mid H_{0i}) = P(T_i \leq t \mid H_{0i} \cap H_i^+) \geq P(T_i \leq t \mid H_{1i}) = P(T_i \leq t \mid H_{1i} \cap H_i^+),$$

   $i = 1, \ldots, m$, for all $t \in R$, where the irrelevance with respect to the complementary set of hypotheses $H_i^+ = \cup_{i \neq j}(H_{0j} \cup H_{1j})$ means that it does not matter which among H_{0j} and H_{1j}, $j \neq i$ is true when testing the i^{th} sub-hypothesis; and

   (b) marginally unbiased and consistent, i.e. $P(T_i > t_{i\alpha} \mid H_{1i}) \geq \alpha$, for all $\alpha > 0$ and $\lim_{n \to \infty} P(T_i > t_{i\alpha} \mid H_{1i}) = 1$, $i = 1, \ldots, m$.

The non-parametric combination in a unique second-order test $T^c = \psi(p_1, \ldots, p_m)$, where p_i is the P-value of test T_i, is achieved by a suitable non-increasing univariate and non-degenerate real function ψ defined from $(0,1)^m$ into R. In order to be suitable for combination of tests, the combining function ψ must satisfy the following minimal and reasonable properties:

I. It is non-increasing in each argument, i.e. $\psi(\ldots, p_i, \ldots) \geq \psi(\ldots, p_i^{'} \ldots)$ if $p_i < p_i^{'}$, for all $i = 1, \ldots, m$.

II. It attains its supreme value $\bar{\psi}$, which could be infinite, when at least one argument attains zero, i.e. $\psi(\ldots, p_i, \ldots) \to \bar{\psi}$ if $p_i \to 0$, for all $i = 1, \ldots, m$, and

III. Its critical value is finite and satisfies $T_\alpha^c < \bar{\psi}$, for all $\alpha > 0$.

We denote by $\wp$ the class of combining functions having these properties. From the most popular examples of combining functions we may enumerate:

A. Fisher's omnibus algorithm uses the statistic $T_F = -2\sum_i \log p_i$. When the m first-order statistics are independent and continuous, T_F is distributed according to a χ^2_{2m} distribution.

B. A Liptak type algorithm is based on the statistic $T_L = \sum_i \Phi^{-1}(1 - p_i)$, Φ being the standard normal cumulative distribution function. When the m first-order statistics are independent and continuous, T_L is normally distributed with mean zero and variance m.

C. Tippett's type uses the statistic $T_T = \max_i(1 - p_i)$.

D. The Lancaster solution is based on the statistic $T_C = \sum_i \xi_{r,a}^{-1}(1 - p_i)$, where $\xi_{r,a}$ is the cumulative distribution function of a gamma variable with r degrees of freedom and a a given scale parameter. Of course, instead of the cumulative distribution function of a gamma random variable it is possible to use any inverse cumulative distribution function transformation.

E. If all sub-alternatives H_{1i} are bilateral, so that all first-order tests are significant either for large or for small values, a natural combining function is the following quadratic form: $T_Q = \mathbf{U}'(\mathbf{R_U})^{-\mathbf{1}}\mathbf{U}$ where $\mathbf{U}' = [\ldots, \Phi^{-1}(1 - p_i), \ldots]$ and $\mathbf{R_U}$ is the correlation matrix of the U transformations of P-values.

The class $\wp$ also contains Birnbaum's complete class of admissible combining functions of independent tests, i.e. all combining functions whose critical region is convex, all combining functions which non-parametrically take care of the underlying structure among $p_i,\ i = 1, \ldots, m$ and also, all monotone increasing measurable transformations of its members.

4.1 Bootstrap algorithms for nonparametric combination tests

The following algorithm will estimate the bootstrap P-value of the test T^c.

ALGORITHM 3.1 (NONPARAMETRIC COMBINATION TEST)

Step 0. Calculate $T = (T_1(X), \ldots, T_m(X)) = (t_1, \ldots, t_m)$.

For $r = 1, \ldots, R$ do:

Step 1. Generate $X_r^* = (x_{11}^*, \ldots, x_{1n_1}^*, \ldots, x_{k1}^*, \ldots, x_{kn_k}^*)$ from the null model $\hat{F}_0$.

Step 2. Calculate $T_r^* = (T_1(X_r^*), \ldots, T_m(X_r^*)) = (t_{1r}^*, \ldots, t_{mr}^*)$.

Step 3. For $s = 1, \ldots, M$ do:

Step 4. Fit the null distribution $\hat{F}_{0r}^*$ to X_r^*.

Step 5. Generate $X_s^{**} = (x_{11}^{**}, \ldots, x_{1n_1}^{**}, \ldots, x_{k1}^{**}, \ldots, x_{kn_k}^{**})$ from the null model $\hat{F}_{0r}^*$.

Step 6. Calculate $T_s^{**} = (T_1(X_s^{**}), \ldots, T_m(X_s^{**})) = (t_{1s}^{**}, \ldots, t_{ms}^{**})$.

Step 7. Calculate $p_{ir}^* = \dfrac{1 + \#\{t_{is}^{**} \geq t_{ir}^*\}}{B+1},\ i = 1, \ldots, M$.

Step 8. Calculate $T_r^{c*} = \psi(p_{1r}^*, \ldots, p_{mr}^*)$.

Step 9. Calculate $p_i^* = \dfrac{1 + \#\{t_{ir}^* \geq t_i\}}{R+1},\ i = 1, \ldots, m$;

Step 10. Calculate $T^c = \psi(p_1^*, \ldots, p_m^*)$;

Step 11. Calculate $p_{boot} = \dfrac{1 + \#\{T_r^{c*} \geq T^c\}}{R+1}$.

In the standard algorithm for estimating bootstrap P-values, the bootstrap replicates of the statistic t_r^* are compared with t, the value of the statistic on the initial data. In our proposed algorithm, the value of the statistic T^c cannot be calculated exactly on initial data, because the theoretical P-values p_i cannot be obtained, but only estimated (step 9). For

this reason we had to use a second bootstrap procedure (step 3 – step 6) to obtain the bootstrap replicates of the statistic T^c.

Algorithm 3.1 is a generalization of Algorithm 4.2 of Davison and Hinkley (1997), which uses a particular combination function ψ, i.e. the Tippett statistic.

It is also important to emphasise the fact that the null model is fitted on the entire set of data, even if a test T_i concerns only a portion of data. To illustrate this idea more precisely, let's look again at the hypothesis at the beginning of this section, $H_0 = \{\cap_i H_{0i}\}$, where

$$H_{0i} : \{P(Y_i < -z) = P(Y_i > z),\ \forall z \in R\}.$$

For each sub-hypothesis of H_0 we can apply a separate test of symmetry. Such a test statistic (for the i^{th} distribution) is $T_i = \sum_j Y_{ij}$. But to ensure that only large values are significant for the alternative hypothesis (when the variables Y_i are also symmetric, but around positive or negative points), a better choice is $T_i = \left|\sum_j Y_{ij}\right|$ or $T_i = (\sum_j Y_{ij})^2$. Under the null hypothesis, the distributions of X_i are symmetric around μ, so to ensure the same property for the null model $\hat{F}_0 = (\hat{F}_{01}, \ldots, \hat{F}_{0k})$ we will symmetrise the EDF $\hat{F}_i$ about $\bar{X}$ (if x_{ij} is present in the sample, $\bar{x} - (x_{ij} - \bar{x}) = 2\bar{x} - x_{ij}$ also must be present). So, even if the statistic T_i concerns only the i^{th} sample, the null distribution from which the bootstrap replicates of this sample are generated depends on the pooled mean of all samples.

A natural question arises: what combination function is "best"? Usually, having available different tests, we must choose those with the greatest power under an alternative hypothesis. We will consider two alternative hypothesis. The first, H_{A1}, is that all the variables X_i are symmetric around μ_i, but there is one index j such that $\mu_j \neq \mu$ and $\mu_i = \mu$ for all $i \neq j$. This index minimizes the expression (3.8). The second alternative hypothesis, H_{A2}, is that all location parameters μ_i are equal, but there is a variable X_i having a asymmetric distribution.

ALGORITHM 3.2 (BOOTSTRAP ESTIMATION OF POWER)

Step 0. Using a slight modified algorithm 3.1, estimate, for a given α, the $(1-\alpha)$ quantile of the distribution of T^c under H_0, denoted t_α.

For $r = 1, \ldots, R$ do:

Step 1. Generate $X_r^* = (x_{11}^*, \ldots, x_{1n_1}^*, \ldots, x_{k1}^*, \ldots, x_{kn_k}^*)$ from the alternative null model $\hat{F}_1$.

Step 2. Calculate $T_r^* = (T_1(X_r^*), \ldots, T_m(X_r^*)) = (t_{1r}^*, \ldots, t_{mr}^*)$.

Step 3. For $s = 1, \ldots, M$ do:

Step 4. Fit the null distribution $\hat{F}_{0r}^*$ to X_r^*.

Step 5. Generate $X_s^{**} = (x_{11}^{**}, \ldots, x_{1n_1}^{**}, \ldots, x_{k1}^{**}, \ldots, x_{kn_k}^{**})$ from the null model $\hat{F}_{0r}^*$.

Step 6. Calculate $T_s^{**} = (T_1(X_s^{**}), \ldots, T_m(X_s^{**})) = (t_{1s}^{**}, \ldots, t_{ms}^{**})$.

Step 7. Calculate $p_{ir}^* = \dfrac{1 + \#\{t_{is}^{**} \geq t_{ir}^*\}}{B+1}$, $i = 1, \ldots, M$.

Step 8. Calculate $T_r^{c*} = \psi(p_{1r}^*, \ldots, p_{mr}^*)$.

Step 9. Calculate $\pi(\alpha, H_A) = \dfrac{\#\{T_r^{c*} \geq t_\alpha\}}{R}$

The bootstrap algorithm for power estimation in the context of nonparametric combination of bootstrap tests requires a fine analysis. The statistic T^c must take greater values for data coming from an alternative model, which implies, by the definition of combination function, that the P-values of the tests T_i are smaller for these data. On the other hand, for the same data, the values taken by the statistics T_i are greater compared with *those obtained for data coming from the null model* and consequently, the corresponding P-values will be smaller. So, to estimate correctly the bootstrap P-values of the tests T_i, in Step 1 of the algorithm 3.2 we will generate data from the alternative model, but in Step 4 of the same algorithm, for the second bootstrap procedure, we will keep the null model. There is a strong analogy with the power formulae, where the statistic calculated on data coming from an alternative model is compared with quantiles of the same statistic, but calculated on null model data.

4.2 Simulation Studies

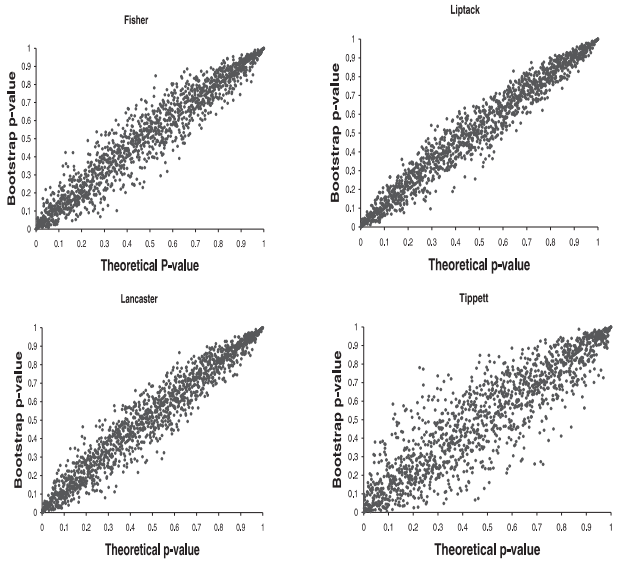
The first experiment is designed to analyse the performances of bootstrap tests for the generalized Behrens–Fisher problem when different nonparametric combining functions are considered. For this purpose we generated 1500 data sets from the following null model, which uses two different, symmetric distributions for data generation (see also the remarks for a such

model in the previous section):

$$x_{ij} = 5 + \sigma_i \varepsilon_{ij}, \quad i = 1, \ldots, 6,$$
$$\sigma_1 = \sigma_4 = 9, \quad \sigma_2 = \sigma_5 = 16, \quad \sigma_3 = \sigma_6 = 25,$$
$$n_1 = n_4 = 10, \quad n_2 = n_5 = 20, \quad n_3 = n_6 = 30,$$
$$G_i = N(0,1), \quad i = 1, \ldots, 3 \qquad G_i = U(-1,1), \quad i = 4, \ldots, 6.$$

For each data set we calculate four bootstrap P-values corresponding to four statistics: Fisher – denoted T_F, Liptack – denoted T_L, Tippett – denoted T_T and Lancaster, (using chi-square distribution with six degrees of freedom) – denoted T_A. Algorithm 3.1 is applied with the parameters $R = 999$ and $M = 999$. The empirical distributions of the same statistics are calculated using all 1500 corresponding values, which allows us to calculate a "theoretical" P-value for each data set and for each statistic. To avoid simulation errors due to different runs of bootstrap algorithm during the comparison process, we used the same set $p^*_{ir},\ i = 1, \ldots, M,\ r = 1, \ldots, R$ (see Step 7 and Step 9 of Algorithm 3.1) for all combination functions. The plots of theoretical P-values versus bootstrap P-values, for each statistic, are presented in Figure 10.

Figure 10: Theoretical P-value vs. bootstrap P-value for the tests using different non-parametric combination functions , for data generated from a null model. The theoretical P-values were calculated using 1500 data and the corresponding bootstrap P-values using a double bootstrap algorithm having $R = 999$ and $M = 999$.



As a first observation, the distribution of points (x_i, y_i) seems to be symmetric with respect to the lines $x0y$ and $x + y = 1$, but the standard deviation along the second axe is not the same in all cases: in the case of T_T statistics, it is about 0.31 (the greatest value), whereas for T_L statistics, it is about 0.14 (the smaller value). Table 9 contains the summary of the three measures considered also in Section 2, i.e. the true level of bootstrap test, the probability that the bootstrap test to accept the null hypothesis whereas the theoretical test to reject it (EI) and the probability that the bootstrap test to reject the null hypothesis whereas the theoretical test to accept it (EII). As we can see, the true level of bootstrap test, for all combination functions, is less than the given nominal level. On the other hand, the bootstrap test using Liptak function presents the smallest values for EI and EII measures, followed close for Lancaster function. The greatest values, especially for $\alpha = 0.1$, are found for the Tippett function, conclusions which confirm what one see in the Figure 10.

Table 9: The true level of bootstrap test for a given nominal level α (Size), the probability that the bootstrap P-value is less than a nominal level α and the corresponding theoretical P-value is greater than the same level (EI) and the probability that the bootstrap P-value test is greater than a nominal level α and the corresponding theoretical P-value is less than the same level (EII)

Distribution	$\alpha = 0.1$			$\alpha = 0.05$			$\alpha = 0.01$		
	Size	Error I	Error II	Size	Error I	Error II	Size	Error I	Error II
Fisher	0.083	0.012	0.027	0.044	0.013	0.018	0.006	0.001	0.004
Liptak	0.09	0.008	0.015	0.041	0.007	0.014	0.006	0	0.003
Lancaster	0.088	0.012	0.022	0.043	0.009	0.012	0.005	0	0.004
Tippett	0.084	0.026	0.041	0.045	0.02	0.023	0.004	0.001	0.006

The next two experiments will analyze the performance of the same bootstrap tests when one of the assumptions used in null model construction is false. Firstly, we keep the assumption of symmetry true and relax the assumption of mean equality. In this case, the measure used to estimate the performance of bootstrap tests is the power against the alternative hypothesis H_{A1}, calculated for size 0.01, 0.05 and 0.1, and for different distance between the null and the alternative hypothesis. The model from which data are generated is very similar with those used to analyse the same problem in Section 3, but the variance of the sample selected to have a different

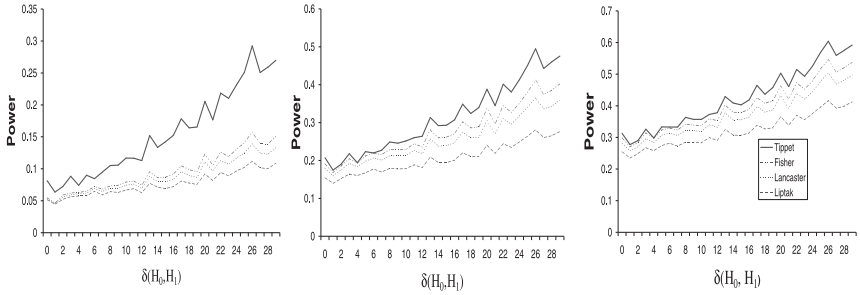
mean (sample three) is set close to the average of the other variances.

$$x_{ij} = 5 + \delta_i + \sigma_i \varepsilon_{ij}, \quad i = 1, \ldots, 6,$$
$$\text{where} \quad \delta_{\text{i}} = 0 \quad \text{if} \quad \text{i} \neq 3 \quad \text{and} \quad \delta_3 = \Delta$$
$$\sigma_1 = 9, \quad \sigma_2 = 16, \quad \sigma_3 = 25, \quad \sigma_4 = 36, \quad \sigma_5 = 49, \quad \sigma_6 = 64$$
$$n_1 = n_4 = 10, \quad n_2 = n_5 = 20, \quad n_3 = n_6 = 30,$$
$$G_i = N(0, 1), \quad i = 1, \ldots, 6.$$

During the first phase of the experiment we estimate the critical values t_α for the four statistics T_F, T_T, T_L and T_A, under the null hypothesis, by using 10000 data replicates from the previous model with $\Delta = 0$ and by taking the $[R(1-\alpha)]^{th}$ ordered value in the sequence $t_1^*, \ldots, t_R^*$. During the second phase, we simulate, for each $\delta(H_0, H_1)$ fixed, one hundred data replicates. For each dataset d, we apply the algorithm 3.2 for $\alpha \in \{0.01, 0.05, 0.1\}$. The value of the power will be calculated, for each $\delta(H_0, H_1)$ and each α, as the average of 100 individual values. As in the previous experiment, to avoid simulation errors, we used the same set $p_{ir}^*,\ i = 1, \ldots, M,\ r = 1, \ldots, R$ (see Step 8 of Algorithm 3.2) for all combination functions.

As we can see in Figure 11, the order of the bootstrap tests according to the power criterion is totaly inversed from the previous experiment: the Tippett function gives the best results, whereas on the last place we find the Liptak function. We repeated the same experience using the Cauchy distribution instead of the normal distribution, in the null model, and we obtained the same results.

Figure 11: Bootstrap power estimation for tests using Tippett, Fisher, Lancaster and Liptak function. (LEGEND: left – size = 0.01, center – size = 0.05, right – size = 0.1)



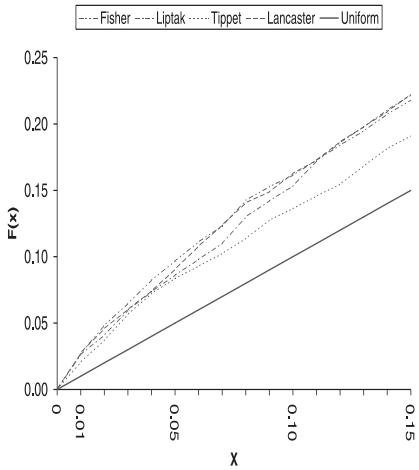
In the last experiment we generate data from a null model which does not respect the assumption of symmetry, but only the assumption of mean equality. The distributions considered are the normal distribution and Chi-square

distribution with one degree of freedom (a non-symmetric distribution).

$$x_{ij} = 5 + \sigma_i \varepsilon_{ij}, \quad i = 1, \ldots, 6,$$
$$\sigma_1 = \sigma_4 = 9, \quad \sigma_2 = \sigma_5 = 16, \quad \sigma_3 = \sigma_6 = 25,$$
$$n_1 = n_4 = 10, \quad n_2 = n_5 = 20, \quad n_3 = n_6 = 30,$$
$$G_i = N(0,1), \quad i = 1, \ldots, 3 \qquad G_i = \chi^2(1), \quad i = 4, \ldots, 6.$$

Using 2000 data sets we calculate the corresponding bootstrap P-values for each of the four statistics. Looking at the corresponding distributions of these P-values (which should be uniform on $(0, 1)$) we observe (see Figure 12) that the probability of rejecting the null hypothesis is always greater than a given nominal level $\alpha \in (0, 0.1)$, for each test. But between these tests, the smallest effective size is obtained by the test using Tippett function.

Figure 12: Empirical cumulative distributions for the samples of bootstrap P-values corresponding to tests using Tippet, Fisher, Lancaster and Liptak function, restricted to the interval [0,0.15].



As a general conclusion, the Liptak combination function gives the best results concerning the accuracy of the bootstrap P-value (for data generated from a normal distribution), whereas the Tippett combination function gives the worst results. But in the same time the Tippett combination functions seems to be the most robust against the departure from the assumptions of symmetry or mean equality. So, we recommend the use of this combination function, even if we must again emphasize that this conclusion is not based on a large set of simulation studies.

The very good accuracy of the bootstrap test using Liptak combination function for data generated from a normal distribution seems to be connected to the fact that the Liptak function uses the inverse normal distribution (the same distribution as data). Consequently, we propose the next procedure:

A. Firstly, verify the assumptions of symmetry (using, as example, a consistent bootstrap test, proposed by Schuster and Barker (1987)).

B. If this assumption is rejected, use the Tippett combination function

C. If is not rejected, try to determine if data were generated from a known distribution(s) (uniform, chi-squared, student, etc.)

D. If a such distribution(s) is found, use a combination function implying the inverse cumulative function of this(these) distribution(s).

E. If a such distribution is not found, use a combination function implying the inverse cumulative function of the empirical sample distributions.

4.3 Unbiasedness and consistency of T^c

In order to examine some asymptotic aspects of T^c, we make the following assumptions:

a) When n diverges, all the $n_j,\ j = 1, \ldots, k$ also diverge;

b) the number of iterations, R and M, diverge; and

c) k and α are fixed.

THEOREM 1 *If $T = \{T_i,\ i = 1, \ldots, m\}$ are consistent first-order bootstrap tests for H_{0i} against H_{1i}, then for all $\psi \in \wp$, $T^c = \psi(p_1^*, \ldots, p_m^*)$ is a consistent combined test for H_0 against H_1.*

Proof. To be consistent, the combined test must reach its critical region with probability one if H_1 is true, or equivalent, if at least one of the hypotheses $H_{1i},\ i = 1, \ldots, m$ is true. Supposing that H_{1j} is true, the consistency of T_j implies that $p_j \to 0$ in probability. Conditions a), b) and c) imply that $p_j^* \to p_j$ *a.s.*, so $p_j^* \to 0$ also in probability. From the properties II and III of the combining function ψ , we have $T^c \to \bar{\psi} > T_\alpha^C$ with probability one, for all $\alpha > 0$. ■

LEMMA 1 *If the variable Y is stochastically larger than the variable Z and φ is a non-decreasing real function, then $\varphi(Y)$ is stochastically larger than $\varphi(Z)$.*

Proof. Y stochastically larger than Z is equivalent to $F_Y(t) \leq F_Z(t)$, for all $t \in R$. Then $P(\varphi(Y) \leq t) = F_{\varphi(Y)}(t) = F_Y(\varphi^{-1}(t)) \leq F_Z(\varphi^{-1}(t)) = F_{\varphi(Z)}(t) = P(\varphi(Z) \leq t)$, for all $t \in R$.

THEOREM 2 *If $T = \{T_i,\ i = 1, \ldots, m\}$ are unbiased first-order bootstrap tests for H_{0i} against H_{1i}, then for all $\psi \in \wp$, $T^c = \psi(p_1^*, \ldots, p_m^*)$ is an unbiased combined test for H_0 against H_1.*

Proof. Unbiasedness of T_i implies $P(p_i \leq z \mid H_{0i}) \geq P(p_i \leq z \mid H_{1i})$, for all $z \in [0, 1]$. Moreover, the assumed PUOD property of component tests implies that $(p_i \mid H_{1i})$ is stochastically smaller than $(p_i \mid H_{0i})$. By the non-decreasing property of ψ and Lemma 1, $\psi(\ldots, p_i, \ldots \mid H_{1i})$ is stochastically larger than $\psi(\ldots, p_i, \ldots \mid H_{0i})$. Hence, by iterating for $i = 1, \ldots, m$, unbiasedness of T^c is achieved. ■

5 Fully nonparametric null model

Now suppose that plotting residuals we do not obtain similar shapes, not even symmetric shapes, so a semiparametric model is not adequate. Therefore, we must determine a wholly nonparametric null model, $\hat{F}_0$. Let also consider the univariate Behrens–Fisher problem, i.e. $H_0 : \{\mu_1 = \mu_2\}$. But this simple hypothesis can be rewritten, under the assumption of the existence of the mean, as $\{E(X_1) = E(X_2)\}$ or $\{t(F_1) = t(F_2);\ t(F) = \int x dF\}$ or

$$\{T(F_1, F_2) = 0,\ T(F_1, F_2) = t(F_1) - t(F_2),\ t(F) = \int x dF\}.$$

In a general case suppose that H_0 has the form $T(F_1, \ldots, F_k) = 0$. The nonparametric maximum likelihood estimate of F_i is the usual EDF, which puts mass n_i^{-1} on each x_{ij}, j=1,...,n_i. The EDF is a multinomial distribution, $Mult(n_i, \hat{p}_i)$ where $\hat{p}_i = (n_i^{-1}, \ldots, n_i^{-1})$. If we restrict the class of possible estimators for F_i to the class of multinomial distributions, then an estimate for F_i will be $Mult(n_i\ ,\ p_i)$, where $p_i = (p_{i1}, \ldots, p_{in_i})$, $\sum_j p_{ij} = 1$. Suppose now that we have a divergence measure between two vectors of probabilities, $d(p, q)$, such that $d(p_i, \hat{p}_i)$ attains its minimum when $p_i = \hat{p}_i$ with the only constraint $\sum_j p_{ij} = 1$. Because there is a one-to-one relationship between the multinomial estimator of F_i and the vector p_i, the constraint $T(F_1, \ldots, F_k) = 0$ may be rewritten as $T(p_1, \ldots, p_k) = 0$. Consequently the nonparametric null model is given by the vectors of probabilities

$p_i, i = 1, \ldots, k$ which minimise (Davison and Hinkley, 1997, p. 165)

$$\sum_{i=1}^{k} d(p_i, \hat{p}_i) - \omega T(p_1, \ldots, p_k) - \sum_{i=1}^{k} \alpha_i \left(\sum_{j=1}^{n_i} p_{ij} - 1 \right), \tag{5.10}$$

where $\omega,\ \alpha_i$ are Lagrange multipliers. We will see that the condition, which restricts the class of estimators for F to the class of multinomial distributions, is not restrictive, in fact.

Many divergence measures have been proposed. One of the first was Kullback's directed divergence (Kullback, 1959)

$$K(p, q) = \sum\nolimits_i p_i \log_2(p_i/q_i),$$

with the symmetrised form $J(p, q) = K(p, q) + K(q, p)$ (Jeffreys, 1948). Other measures, generalisations of $K(p, q)$, are the additive directed divergence of order α (Rényi, 1961)

$$R^{\alpha}(p, q) = (\alpha - 1)^{-1} \log_2 \left(\sum\nolimits_i p_i^{\alpha} q_i^{1-\alpha} \right), \quad \alpha \neq 1,$$

and the non-additive directed divergence of order-α (Rahtie and Kannappan, 1972)

$$\tilde{I}^{\alpha}(p, q) = (2^{\alpha-1} - 1)^{-1} \left(\sum\nolimits_i p_i^{\alpha} q_i^{1-\alpha} - 1 \right), \ \alpha \neq 1.$$

But the most general divergence measure, restricted to the class of discrete distributions, is the power divergence (Cressie and Read, 1984; Cressie and Read, 1988)

$$I^{\delta}(p, q) = \frac{1}{\delta(\delta + 1)} \sum p_i \left\{ \left(\frac{p_i}{q_i} \right)^{\delta} - 1 \right\}, \quad -\infty < \delta < \infty.$$

The values for $\delta = 0$ and $\delta = -1$ are taken to be the continuous limits for $\delta \to 0$ and $\delta \to -1$ respectively, for which we obtain the reverse information distance

$$I^{0}(p, q) = \sum p_i \log \left(\frac{p_i}{q_i} \right),$$

and the aggregation information distance

$$I^{-1}(p, q) = \sum q_i \log \left(\frac{q_i}{p_i} \right).$$

The power divergence clearly satisfies the condition $I^{\delta}(p, q) \geq 0$ and $I^{\delta}(p, q) = 0$ if and only if $p = q$ (Jensen's inequality). The power divergence is not a

true distance for either value of δ. However, the square root of $I^{-1/2}(p,q)$ satisfies all the conditions for a distance and is known as Matusita distance (Matusita, 1955) $M = \left(\sum_i (\sqrt{p_i} - \sqrt{q_i})\right)^{1/2}$.

The expression (5.10) for $d(p,q) = I^{\delta}(p,q)$ and for the initial H_0 has form:

$$\frac{1}{\delta(\delta+1)} \sum_{i=1}^{2} \sum_{j=1}^{n_i} (p_{ij}^{\delta+1} n_i^{\delta} - p_{ij}) - \omega \left(\sum_{j=1}^{n_1} x_{1j} p_{1j} - \sum_{j=1}^{n_2} x_{2j} p_{2j} \right) - \alpha_1 \left(\sum_{j=1}^{n_1} p_{1j} - 1 \right) - \alpha_2 \left(\sum_{j=1}^{n_2} p_{2j} - 1 \right) \quad (5.11)$$

Calculating the derivatives with respect with p_{ij} and setting them equal with zero, we obtain the expressions for the two vectors of probabilities, which minimise (5.11),

$$\begin{cases} p_{1j,0} = (\delta(\alpha_1 + \omega\, x_{1j}) + (\delta+1)^{-1})^{1/\delta} n_1^{-1} \\ p_{2j,0} = (\delta(\alpha_2 - \omega\, x_{1j}) + (\delta+1)^{-1})^{1/\delta} n_2^{-1} \end{cases}, \quad (5.12)$$

where $\alpha_1,\ \alpha_2$ and ω are determined by the equations

$$\sum_{j=1}^{n_1} p_{1j} = 1,\ \sum_{j=1}^{n_2} p_{2j} = 1,\ \sum_{j=1}^{n_1} x_{1j} p_{1j} = \sum_{j=1}^{n_2} x_{2j} p_{2j}.$$

For $\delta = 0$, the expressions for p_{ij} become

$$p_{1j,0} = \frac{e^{\omega\, x_{1j}}}{\sum e^{\omega\, x_{1j}}}, \quad p_{2j,0} = \frac{e^{-\omega\, x_{2j}}}{\sum e^{-\omega\, x_{2j}}}. \quad (5.13)$$

These solutions are always positive. The value of ω is determined from the third constraint,

$$\frac{\sum x_{1j} e^{\omega\, x_{1j}}}{\sum_j e^{\omega\, x_{ij}}} = \frac{\sum x_{2j} e^{-\omega\, x_{2j}}}{\sum_j e^{-\omega\, x_{2j}}}.$$

The family of distributions having the form (5.13) is called an empirical exponential family.

For $\delta = -1$, the probabilities are calculated by

$$p_{1j,0} = (n_1(\alpha_1 + \omega\, x_{1j}))^{-1},\ p_{2j,0} = (n_2(\alpha_2 - \omega\, x_{2j}))^{-1},$$

but the existence and the positivenes of these solutions is ensured only under some conditions.

After choosing a value for δ and solving (5.12), we obtain the distribution estimators $p_{1,0}$ and $p_{2,0}$ (or equivalently $\hat{F}_{1,0}$ and $\hat{F}_{2,0}$), which satisfy the null nonparametric model. Therefore the general bootstrap algorithm can be applied:

`Algorithm 4.1 Estimation of the bootstrap` P`-value`

`Step 1. Generate` $x_{11}^*, \ldots, x_{1n_1}^*$ `from` $\hat{F}_{1,0}$ `by random sampling.`

`Step 2. Generate` $x_{21}^*, \ldots, x_{2n_2}^*$ `from` $\hat{F}_{2,0}$ `by random sampling.`

`Step 3. Calculate the test statistic` t^*.

`Step 4. Repeat steps 1-3 (R-1) times.`

`Step 5.` $p = \dfrac{1 + \#\{t_r^* \geq t\}}{R+1}$.

For each δ we obtain other estimators $p_{1,0}$ and $p_{2,0}$ and finally another $P(\delta)$-value (making abstract for the simulation errors). The next question arises naturally: is there a connection between the parameter δ and the power of the test under an alternative hypothesis? In our case

$$H_1 = \{\mu_1 \neq \mu_2\} = \{T(F_1, F_2) = \theta,\ \theta \neq 0\}.$$

So, after choosing a value for θ we can apply the same procedure, with the same δ, to determine the estimates $\hat{F}_{1,1}$ and $\hat{F}_{1,2}$, true under H_1. A natural choice for θ is $\bar{X}_1 - \bar{X}_2$ (in this particular case, $\hat{F}_{i,1} = EDF$ of F_i). Expression (5.11) becomes:

$$\sum_{i=1}^{2} \frac{1}{\delta(\delta+1)} \sum_{j=1}^{n_i} (p_i^{\delta+1} n_i^{\delta} - p_{ij}) - \omega \left(\sum_{j=1}^{n_1} x_{1j} p_{1j} - \sum_{j=1}^{n_2} x_{2j} p_{2j} - \theta \right) - \alpha_1 \left(\sum_{j=1}^{n_1} p_{1j} - 1 \right) - \alpha_2 \left(\sum_{j=1}^{n_2} p_{2j} - 1 \right),$$

with the same expression (5.12) for $p_{1j,1}$ and $p_{2j,1}$ but with parameters ω, α_1, α_2 satisfying a different constraint system.

$$\sum_{j=1}^{n_1} p_{1j} = 1,\ \sum_{j=1}^{n_2} p_{2j} = 1,\ \sum_{j=1}^{n_1} x_{1j}p_{1j} - \sum_{j=1}^{n_2} x_{2j}p_{2j} = \theta.$$

The plot of $\pi(\alpha, \theta, \delta \mid H_1)$ versus δ gives us the necessary information about a possible connection between these two measures. Another possible measure of efficiency for the test statistic T_δ is the reciprocal of its estimated variance under the null model, which is $v_0 = \sum_{i=1}^{2} n_i^{-1} \sum_{j=1}^{n_i} (x_{ij} - \hat{\mu}_{i0})^2 p_{ij,0}$, where $\hat{\mu}_{i0} = \sum_j x_{ij}p_{ij,0}$.

If $H_0 = \{\mu_1 = \mu_2 = \cdots = \mu_k\}$ then a possible choice for $T(F_1, \ldots, F_k)$ would be $T = \sum_{i=1}^{k-1} (t(F_{i+1}) - t(F_i))^2$ because $T = 0 \Leftrightarrow t(F_i) = t(F_j)$, for all $i \neq j$. But as k grows, the minimisation problem becomes more and more difficult. A different approach is to use a combination test. Because $H_0 = \bigcap_{i<j}^{c} H_{0(ij)}$ and $H_1 = \bigcup_{i<j}^{c} H_{1(ij)}$ (where $c = \binom{k}{2}$ and $H_{0(ij} = \{\mu_i = \mu_j\}$, $H_{1(ij)} = \{\mu_i <\neq> \mu_j\}$), and because each individual test is unbiased and consistent, we can apply one of the combination functions already mentioned.

5.1 Simulation Studies

Before we describe the goals of the experiments we made, we must make some remarks. Firstly, to obtain the estimators $p_{1j,0}, p_{2j,0}$, for a given δ, we must solve a minimization problem under constraints, by using a numerical algorithm. There is no guarantee that for each dataset the numerical algorithm will properly stop, but we can control the error rate of the numerical solution, if this is obtained. We set this error rate at 10^{-6}. If, during the simulations, the numerical algorithm[1] stops with an warning message, the corresponding dataset is deleted. Secondly, we deliberately restricted the size of the samples because we think that only for small sample size is it worth trying to find the "best" δ (and, of course, it is computationally feasible). Thirdly, the employed statistic is not exactly $T = (\bar{X}_1 - \bar{X}_2)^2$, but the statistic (3.4), which becomes, for $n = 2$, $T_w = w_1w_2(w_1 + w_2)^{-1} \left(\bar{X}_1 - \bar{X}_2\right)^2 = w_1w_2(w_1 + w_2)^{-1} \times T$. Of course, this choice has no effect on the constraints of the minimization problem.

During the first experiment we tried to determine if there is a connection between the value of δ and the accuracy of the bootstrap P-value of the cor-

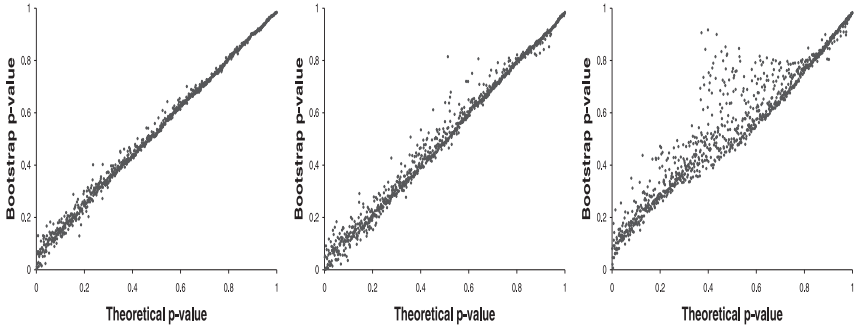
[1] function **imsl_d_min_con_nonlin()** from **IMSL** library, implemented in C

responding test. For this we generated 1000 data sets from the following null model, which uses a symmetric distribution, with mean zero and variance one:

$$\begin{gathered} x_{ij} = 5 + \sigma_i \varepsilon_{ij}, \quad i = 1, \ldots, 2, \quad j = 1, \ldots, n_i, \\ \sigma_1 = 4, \quad \sigma_2 = 9, \\ n_1 = 4, \quad n_2 = 6, \\ G_1 = N(0,1), \quad G_2 = N(0,1). \end{gathered}$$

We denote this model $N1$. Another null model, denoted $N2$, is obtained by taking $G_2 = \text{Exp}(1) - 1$ (a non-symmetric distribution, with mean zero and variance one) and a third null model ($N3$) is obtained for $G_2 = \text{Cauchy}(1)$ (a symmetric distribution, with an infinite first moment). For each dataset and for each δ between -20 and 20 (integer values) we applied Algorithm 4.1, with $R = 100,000$, to obtain the bootstrap p(δ)-value. The theoretical P-value is estimated from the sample of 1000 values of the statistic T_w. The plot of the bootstrap P-value (for $\delta = 0$) vs. the theoretical P-value is presented in Figure 13. We chose $\delta = 0$ because there are not significant differences for other values.

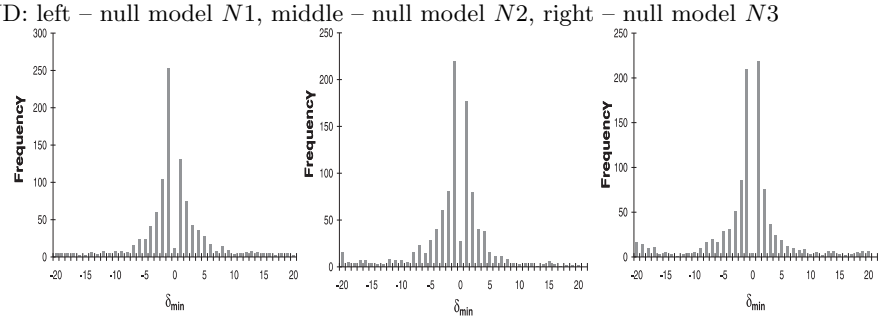
Figure 13: Theoretical P-value vs. bootstrap p(δ)-value for the test T_w and $\delta = 0$. The theoretical P-values were calculated using 1000 data and the corresponding bootstrap P-values using 100000 replicates. LEGEND: left – null model $N1$, middle – null model $N2$, right – null model $N3$



As we may observe, the bootstrap P-value overestimates the theoretical P-value, especially if the last measure is less than 0.2. The plot for the null model $N3$ (on the right) is a special case, because the statistic T_w is not defined for a Cauchy distribution. A more deep analysis of the dependence between δ and the accuracy of the bootstrap $P(\delta)$-value is obtained

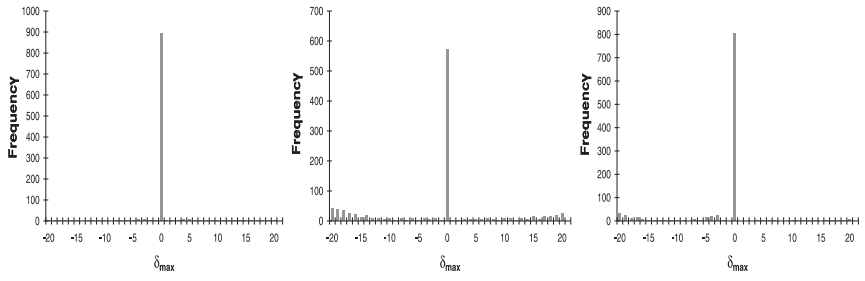
by calculating a particular histogram. More precisely, for each dataset we calculate the minimum of the absolute difference between the theoretical P-value and the bootstrap $P(\delta)$-value, for each $\delta \in [-20, 20]$. We set as δ_{min} the value of δ for which this minimum was obtained and, if there are several such values, we take the value nearest to zero. The histogram of all 1000 δ_{min} values for all null models is represented in Figure 14, whereas the same histogram, but for δ_{max}, is represented in Figure 15.

Figure 14: The histogram of 1000 δ_{min} (values of δ for which the minimum absolute difference between the bootstrap $P(\delta)$-value and the theoretical P-value is attained). LEGEND: left – null model $N1$, middle – null model $N2$, right – null model $N3$



As we may observe, for all null models the histograms are almost identical. The most accurate bootstrap test is obtain for $\delta = -1$ or $\delta = 1$ (*remark*: for many data sets, the minimum absolute distance was attained for both δ, but because our computational procedure retained the first δ_{min}, i.e. -1, the mode is bigger for this value). Concerning the less accurate bootstrap test, this is obtained for $\delta = 0$.

Figure 15: The histogram of 1000 δ_{max} (values of δ for which the maximum absolute difference between the bootstrap $P(\delta)$-value and the theoretical P-value is attained). LEGEND: left – null model $N1$, middle – null model $N2$, right – null model $N3$



Trying to go deeper into this analysis (for the moment we looked only at integer values for δ), we repeat the simulation from null model $N1$ by tacking δ between -2 and 2, with a step set to 0.1. We construct the same histograms, for δ_{min} and δ_{max}, but this time in two different contexts. In the first case, we consider all 1000 values of δ_{min} (respectively δ_{max}). In the second case, we consider only those values obtained for data sets for which the theoretical P-value is less or equal to 0.1. And as we may see in Figure 16 and 17, the histogram of the subsample is very different from the histogram of the entire sample (*remark*: this situation was not present for δ taking integer values). A first conclusion is, looking to the entire sample of δ_{min}, that the particular values $\delta = -1$ and $\delta = 1$ lost their good properties concerning the accuracy, and are now replaced by $\delta = -2$ or 0 or 2. More than, the probability mass of the histogram is more dispersed than in the precedent case (δ tacking integer values). But the most interesting result is the fact that for $\delta = 0$, the bootstrap $P(\delta)$-value is always far from the theoretical P-value, if the last one is less than 0.1 (see the plot from right of Figure 16). The difference between the two histograms is due to the fact that, for large values of the theoretical P-value, the corresponding bootstrap $P(\delta)$-values are very similar, whatever is the value of δ. Consequently, very often δ_{min} and δ_{max} takes the value zero (remember that we decided to select, from a set of candidates for δ_{min} or for δ_{max}, those candidate closest to zero). But if the theoretical P-value is small, the corresponding bootstrap $P(\delta)$-values are more dispersed and the values of δ_{min} are more concentrated around the points -2 and 2.

Figure 16: The histogram of δ_{min} (values of δ for which the minimum absolute difference between the bootstrap $P(\delta)$-value and the theoretical P-value is attained), for data generated from null model $N1$) . LEGEND: left – all 1000 values of δ_{min}, right – subsample of δ_{min}, for data having the theoretical P-value ≤ 0.1

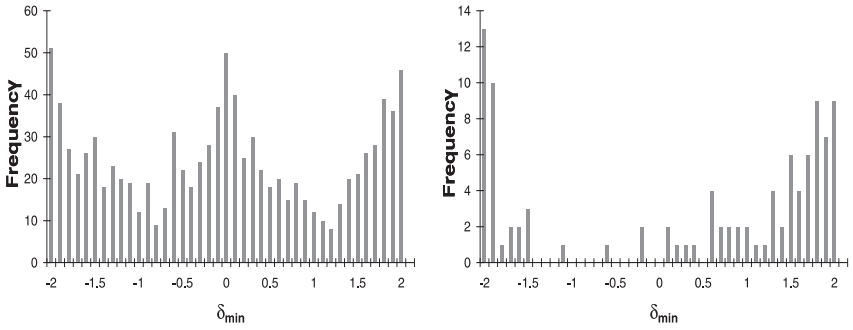
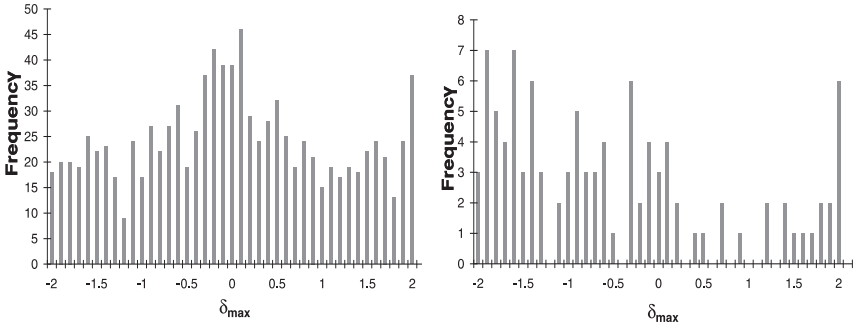


Figure 17: The histogram of δ_{max} (values of δ for which the maximum absolute difference between the bootstrap $P(\delta)$-value and the theoretical P-value is attained), for data generated from null model $N1$. LEGEND: left – all 1000 values of δ_{max}, right – subsample of δ_{max}, for data having the theoretical P-value ≤ 0.1



As a conclusion, we may affirm that in addition to dependence between the value of the parameter δ and the accuracy of the bootstrap $P(\delta)$-value, there is also dependence between the theoretical P-value of a given data set and the accuracy of the corresponding bootstrap p(δ)-value. Using non-integer values for δ does not improve the accuracy of bootstrap p(δ)-value (more than, our calculus showed a certain degradation). The value $\delta = 0$ (implying the use of an empirical exponential distribution by the bootstrap algorithm) seems to be the worst choice, whereas the values -1 or 1 proved very good properties concerning the accuracy (*remark*: from a computational viewpoint, we prefer the value $\delta = 1$).

The second experiment was designed to analyse the (possible) connection between δ and the power of the bootstrap test against the alternative hypothesis H_1. The alternative null model from which data are generated is

$$x_{1j} = 5 + \sigma_1\varepsilon_{1j}, \quad x_{2j} = 5 + \Delta + \sigma_2\varepsilon_{2j}$$
$$\sigma_1 = 4, \quad \sigma_2 = 9, \quad n_1 = 4, \quad n_2 = 6,$$
$$G_1 = N(0,1), \quad G_2 = N(0,1).$$

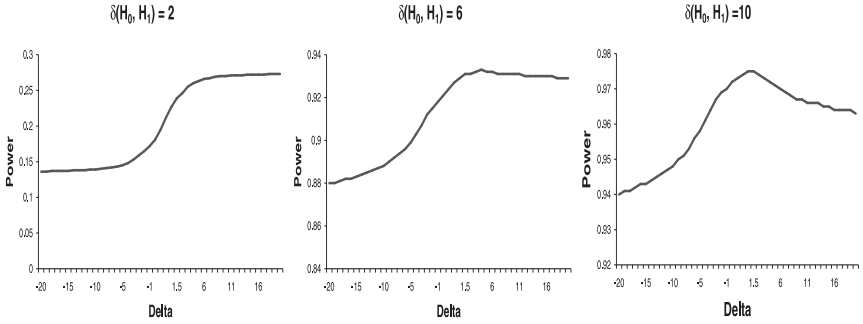
and the distance between the null and the alternative hypothesis is defined as

$$\delta(H_0, H_1) = \left| \frac{\bar{x}_1 - \bar{x}_2}{se(\bar{x})} \right| = |\Delta(w_1 + w_2)|,$$

which in our case become $\delta(H_0, H_1) = 0.32|\Delta|$. The critical values t_α ($\alpha \in \{0.01, 0.05, 0.1\}$) for the statistic T_w were calculated using 100 000

data replicates from the null model $N1$. For a given $\delta(H_0, H_1)$ and a given δ, the power of the bootstrap test is estimated using 100000 replicates generated from the alternative model. The distance $\delta(H_0, H_1)$ takes values in $\{1, 2, \ldots, 10\}$ and the parameter δ takes values in $\{-20, -19, \ldots, 19, 20\}$. The plot of the bootstrap estimate of the power vs. δ, for a fixed $\delta(H_0, H_1)$ and for $\alpha = 0.1$ is presented in Figure 18. It is interesting to see that, for small values of $\delta(H_0, H_1)$, the maximum power is attained for large values of δ, whereas when the distance grows, the maximum power shifts to $\delta = 1$. We repeated the experiment using an alternative null model derived from the null model $N2$ (i.e. $G_2 = Exp(1) - 1$) and the conclusions were similar. But by using an alternative null model implying the Cauchy distribution (see null model $N3$), the dependence of the bootstrap power on δ change for large values of $\delta(H_0, H_1)$ (see Figure 19). As we may observe, the power grows slowly for $\delta \leq -2$, there is a sudden growth for δ between -2 and 2, after that the power continues to grow slowly. Therefore, we may recommend the value $\delta = 1$ as an optimum choice to obtain a maximum power against the alternative hypothesis H_1.

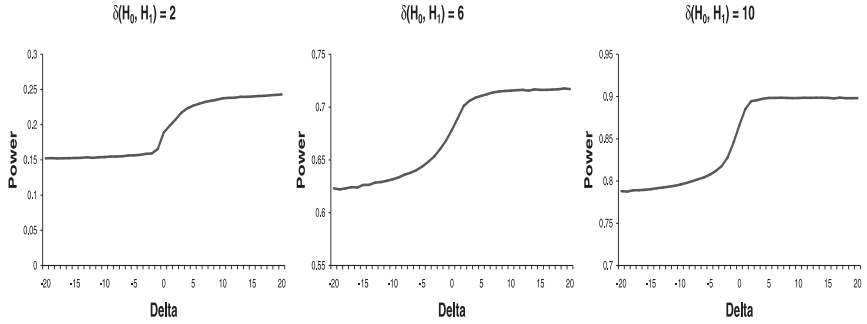
Figure 18: Bootstrap power estimation of the test T_w, for fixed $\delta(H_0, H_1)$ and different δ values, when size= 0.1 and data are generated from a normal distribution. LEGEND: left – $\delta(H_0, H_1) = 2$, middle – $\delta(H_0, H_1) = 6$, right – $\delta(H_0, H_1) = 10$



6 Empirical likelihood

When the data are supposed to come from a distribution F_ψ, ψ being a vector of unknown parameters, the likelihood for ψ evaluated at x is the corresponding density $L(\psi) = f_\psi(x)$ and the relative plausibility of other

Figure 19: Bootstrap power estimation of the test T_w, for fixed $\delta(H_0, H_1)$ and different δ values, when size= 0.1 and data are generated from a Cauchy distribution. LEGEND: left – $\delta(H_0, H_1) = 2$, middle – $\delta(H_0, H_1) = 6$, right – $\delta(H_0, H_1) = 10$



values may be measured by the likelihood ratio statistic

$$T(\psi) = 2(\log L(\hat{\psi}) - \log L(\psi)), \tag{6.14}$$

where $\hat{\psi}$ is the maximum likelihood estimate. Under regularity conditions, if ψ_0 is the true value of ψ, then $T(\psi_0)$ has approximately a chi-square distribution with d degrees of freedom, where d is the length of vector ψ. So the statistic $T(\psi_0)$ is a natural test statistic for $H_0 = \{\psi = \psi_0\}$ versus $H_1 = \{\psi \neq \psi_0\}$, having an approximate P-value $p = P(\chi^2_d \geq t)$. When we have no parametric model, a parameter ψ for F could be any $t(F)$, where t is a statistical functional. If we can construct a function similar to the likelihood from the parametric case, we can obtain a similar test statistic $T(\psi)$. Such a construction has strong similarities to use of a full nonparametric model.

Having the observed data $\{x_1, \ldots, x_n\}$, suppose that a possible distribution F for data is chosen only from the multinomial family. By consequence $F = Mult(n, p)$, $p = (p_1, \ldots, p_n)$ being a vector of probabilities. The likelihood for p is $L(p) = \prod_i p_i$. Because F must be supported by data, we have $\psi = t(F) = t(p)$, the last equality being a consequence of the one-to-one relationship between the multinomial distribution and the vector p, the profile likelihood for ψ will be

$$L_e(\psi) = \max_{p:t(p)=\psi} \prod_{i=1}^{n} p_i,$$

called the empirical likelihood (Owen, 1988; Owen, 2001). The value of ψ which maximises $L_e(\psi)$ is just $\hat{\psi} = t(n^{-1}, \ldots, n^{-1})$, the vector of probabilities corresponding to EDF, and so the corresponding statistic will be

$T(\psi) = 2\{\log L_e(\hat{\psi}) - \log L_e(\psi)\}$. But looking closely at $\log L_e(\psi)$, denoted $l_e(\psi)$, we can write the next equivalence

$$\begin{aligned} l_e(\psi) &= \max_{p:t(p)=\psi} \sum \log(p_i) \\ &= -n \min_{p:t(p)=\psi} \sum\nolimits_i \frac{1}{n} \log(\frac{1}{np_i}) - n\log(n) \\ &= -n \min_{p:t(p)=\psi} I^0(\frac{1}{n}, p) - n\log(n). \end{aligned}$$

In conclusion, to calculate the empirical log likelihood is equivalent to solve the minimisation problem $\min I^0(\frac{1}{n}, p)$ with constraints $t(\psi) = p$ and $\sum_i p_i = 1$. But because $I^0(\frac{1}{n}, p)$ is just a particular case of $I^\delta(\frac{1}{n}, p)$, we can generalise the concept of log likelihood, normalised to have maximum zero, as (Cotofrei, 1998; Cotofrei, 1999):

DEFINITION 3 *The δ-empirical divergence log likelihood for parameter $\psi = t(p)$ is the function*

$$l_{e(\delta)}(\psi) = -n \min_{p;t(p)=\psi} I^\delta(n^{-1}, p).$$

As a consequence, the empirical exponential family loglikelihood (Efron, 1982; DiCiccio et al., 1989) is just $l_{e(-1)}(\psi)$. A similar generalization was proposed by Corcoran (1998), but starting from the expression of statistic (6.14). The corresponding test statistic is

$$T_\delta(\psi_0) = 2\{l_{e(\delta)}(\hat{\psi}) - l_{e(\delta)}(\psi_0)\} = 2n \min_{p;t(p)=\psi_0} I^\delta(\frac{1}{n}, p).$$

Even in this general case it may be proved that the limiting distribution for $T_\delta(\psi_0)$ is still chi-squared with d degrees of freedom (Cressie and Read, 1988; Baggerly, 1998). It is straightforward to show, using the independence of F_i and the fact that the distributions are restricted to the class of multinomial family, that the generalisation of the Definition 3 for multivariate case is:

DEFINITION 4 *The δ-empirical divergence log likelihood for the parameter $\psi = t(\mathbf{p})$ where $\mathbf{p} = (p_1, \ldots, p_k)$, $\sum_{j=1}^{n_i} p_{ij} = 1$ is the function*

$$l_{e(\delta)}(\psi) = - \min_{\mathbf{p}:t(\mathbf{p})=\psi} \sum_{i=1}^{k} n_i I^\delta(\frac{1}{n_i}, p_i).$$

For the generalised Behrens–Fisher problem, the parameter that must be considered is $\psi = t(F_1) - t(F_2)$, $t(F) = \int x dF$. The system of hypotheses is $H_0 = \{\psi = 0\}$ against $H_1 = \{\psi \neq 0\}$. Definition 4 is applied in this case for $\psi = \tilde{t}(\mathbf{p}) = \sum_j x_{1j} p_{1j} - \sum_j x_{2j} p_{2j}$. For δ equal to -1, respectively 0, the test statistic (6.14) has the form $T_{-1}(\psi) = 2 \sum_{i=1}^{2} \sum_{j=1}^{n_i} n_i p_{ij} \log(p_{ij} n_i)$, where

$$p_{1j} = \frac{e^{-(n_1)^{-1}(\alpha_1 + \omega\, x_{1j})}}{\sum_j e^{-(n_1)^{-1}(\alpha_1 + \omega\, x_{1j})}},\ p_{2j} = \frac{e^{-(n_2)^{-1}(\alpha_2 - \omega\, x_{2j})}}{\sum_j e^{-(n_2)^{-1}(\alpha_2 - \omega\, x_{2j})}},$$

respectively $T_0(\psi) = 2 \sum_{i=1}^{2} \sum_{j=1}^{n_i} \log(n_i p_{ij})$, where

$$p_{1j} = \frac{1}{\alpha_1 + \omega\, x_{1j}},\ p_{2j} = \frac{1}{\alpha_2 - \omega\, x_{2j}}.$$

For $\delta \neq -1,\ 0$, the test statistic has the general form

$$T_\delta(\psi) = \frac{2}{\delta(\delta+1)} \sum_{i=1}^{2} \sum_{j=1}^{n_i} \left((n_i p_{ij})^{-\delta} - 1 \right)$$

where

$$\begin{cases} p_{1j} = (n_1^\delta (\delta+1)(\alpha_1 + \omega\, x_{1j}))^{-1/(\delta+1)} \\ p_{2j} = (n_2^\delta (\delta+1)(\alpha_2 - \omega\, x_{2j}))^{-1/(\delta+1)} \end{cases}.$$

In all these formulas, the coefficients $\alpha_1, \alpha_2, \omega$ are determined by the equations $\sum_j p_{ij} = 1$ and $\sum_j x_{1j} p_{1j} = \sum_j x_{2j} p_{2j}$. Although the limiting distribution for $T_\delta(\psi_0)$ is the same as that of $T(\psi_0)$ under a correct parametric model, such asymptotic results are typically less useful in a non-parametric setting. This suggests that the bootstrap be used to calibrate δ-empirical divergence log likelihood, by using quantiles of bootstrap replicates of $T_\delta(\psi_0)$, i.e. quantiles of $T_\delta^*(\hat{\psi}_0)$.

7 Conclusions

(1) The main problem in bootstrap hypothesis testing is the choice of the distribution $\hat{F}_0$, true under H_0, such that the distribution of the statistic T_n does not depend too strongly on this particular choice. This is difficult, especially when the null model is semiparametric or fully non-parametric. Simulations designed to study the dependence between

the distribution of the test statistic T (see (3.4)) and a particular distribution F included in the null model showed that this independence is not always achieved, especially if the sample size is small (i.e. less than 100). This independence is achieved only inside a family of distributions. This lack of independence explains the answer we found for the first main question, i.e. how accurate are the bootstrap p-values of the different test statistics proposed for Behrens–Fisher problem. We could prove that the distribution (lets call it F_{bt}) of the points (x_i, y_i), where x_i is the bootstrap P-value for a given dataset and y_i is the corresponding theoretical P-value, depends on the data distribution, even if the null model from which the replicates are generated is the same for all bootstrap tests. We expect the distribution F_{bt} to be approximately bi-normal, symmetric around the first bisection line and we found that the test statistic T induces a non-symmetric distribution F_{bt} and that, for this test, always the bootstrap P-value overestimates the theoretical P-value. To estimate the accuracy of the bootstrap p-values, we defined two measures:

(a) Error like type I measure (or EI), i.e. the probability that the bootstrap test reject the hypothesis whereas the theoretical test accept it (or Prob(bootstrap P-value $< \alpha$ and theoretical P-value $> \alpha$)).

(b) Error like type II measure (or EII), i.e. the probability that the bootstrap test accept the hypothesis whereas the theoretical test reject it, for a given level α.

These measures are calculated in correlation with a given level, for $\alpha \in \{0.01, 0.05, 0.1\}$ and present an opposite behaviour, similar to the classical types of errors for a test statistic (i.e., if EI go up, then EII go down).

(2) If a multidimensional hypothesis may be decomposed into a set of one-dimensional hypotheses, $H_0 = \{\cap_{1 \leq i \leq m} H_{0i}\}$, and the alternative hypothesis states that at least one of the one-dimensional hypotheses is not true, a non-parametric combination of bootstrap test is applicable. The bootstrap algorithm developed to estimate the P-value for a combination test is the only possible computational methodology, especially because the value of the combination test depends on the P-values of the one-dimensional tests, which are, usually, unknown. For this reason, a first level bootstrap iteration is used to estimate the P-values of the one-dimensional tests and a second level bootstrap

iteration is used to obtain the necessary data replications. The null model must consider the set of all sub-hypotheses and must reflect the fact that the multidimensional hypothesis is true if and only if all sub-hypotheses are true. From a theoretical viewpoint, we showed that the properties of the one-dimensional tests (unbiasedness and consistency) are kept by the combination test due to the properties of the combination function. From a practical viewpoint, the simulations showed that the Tippett combination function induces a less accurate bootstrap test than the other combination functions, but offers greater power against the alternative hypothesis H_A ("there are at least two samples with different means") or against the non-respect of the assumption of symmetry, whereas the Liptack combination function has the opposite behaviour. Following this observation we proposed a combination function (which we call the "fitted combination") having the form: $T_{FC} = \sum_i f_i^{-1}(1 - p_i)$, f_i being the distribution of the i^{th} one-dimensional test. Of course, this function respects all the necessary conditions imposed on combination functions. Computationally speaking, if the distribution f_i is unknown or no asymptotic form is available, then a third-level bootstrap iteration is needed to estimate it.

(3) A fully nonparametric model does not impose any restriction on the form of the distribution F_0, but practically we must restrict it to a well-defined family of distributions. For bootstrap resampling, the most useful family is the multinomial family, but in the literature other families are mentioned , like the generalised lambda distribution (Karian and Dudewicz, 2000), which is continuous. The problem of fitting the null model with respect to data, hypothesis H_0 and the selected family of distributions is equivalent to a minimization problem with constraints. Our contribution to this problem is the choice of the metric: we proposed the power divergence measure

$$I^{\delta}(p, q) = \frac{1}{\delta(\delta + 1)} \sum p_i \left\{ \left(\frac{p_i}{q_i} \right)^{\delta} - 1 \right\}, \quad -\infty < \delta < \infty,$$

which includes as particular cases the reverse information distance and the aggregation information distance. A real difficulty here is to find the best value for the parameter δ according to some criterion, like the distance between the bootstrap p-value and the theoretical P-value or the power against a given alternative. The computational effort is immense, because it's about to obtain a numerical solution for

an optimisation problem. A simulation study showed that, for small samples, the values $\delta = -1$ or $\delta = 1$ give the most accurate bootstrap tests, whereas $\delta = 0$ conducts to the worst results. Concerning the power, the experiences proved a clear difference between the results obtained for $\delta < 0$ and those for $\delta > 0$, with a "peak" placed around $\delta = 1$.

(4) A similar approach applied to empirical log likelihood showed that this concept might be also considered as a minimization problem with constraints, using the power divergence measure with $\delta = 0$. This observation conducted us to a generalisation of the empirical likelihood, named the $\delta-$empirical divergence likelihood, even if we can no longer talk about a logarithmic function. The concept of likelihood is a basic notion in statistics, with applications in parameter estimation, confidence intervals and hypothesis testing, and the possibility to extend this notion for nonparametric models represents an important gain. Looking closely, the generalised likelihood ratio statistic is equivalent, up for a constant, to the fitting of a fully non-parametric model and therefore all the observations made in the preceding paragraph are also applicable for the $\delta-$empirical divergence likelihood.

(5) All conclusions and observations from this report were obtained using simulation studies. The null models from which data were generated depend practically on many parameters, like the number of samples, the size and the standard error of samples, the distribution generating each sample, etc. To limit the computational effort to a reasonable time complexity, we set many of these parameters to some predefined values and we varied especially the generator distributions. Therefore, these conclusions should be generalised with care.

References

Aspin, A. (1948). An examination and further developments of a formula arising in the problem of comparing two mean values. *Biometrika*, 35:88–96.

Baggerly, K. (1998). Empirical likelihood as a goodness-of-fit measure. *Biometrika*, 85:535–547.

Bahadur, R. R. and Savage, L. J. (1956). The nonexistence of certain statistical procedures in nonparametric problems. *Annals of Mathematical Statistics*, 27:1115–1122.

Ballin, M. and Pesarin, F. (1990). Una procedura di ricampionamento e di combinazione nonparametrica per il problema di Behrens–Fisher multivariato. In *Atti Della Societa Italiana Di Statistica*, volume 2, pages 351–358, CEDAM, Padova.

Behrens, B. V. (1929). Ein beitrag zur fehlerberechnung bei wenige beobachtungen. *Landwirtch. Jb.*, 68:807–837.

Beran, R. (1986). Simulated power functions. *Annals of Statistics*, 14:151–173.

Beran, R. (1988). Prepivoting test statistics: A bootstrap view of asymptotic refinements. *Journal of the American Statistical Association*, 83:687–697.

Berger, R. and Boos, D. (1994). *P*-values maximized over a confidence set for the nuissance parameter. *Journal of the American Statistical Association*, 89:1012–1016.

Berk, R. and Jones, D. (1978). Relatively optimal combination of tests. *Scandinavian Journal of Statistics*, 15:158–162.

Bunke, O. and Riemen, S. (1983). A note on bootstrap and other empirical procedures for testing linear hypotheses without normality. *Statistics*, 14:517–526.

Chernick, M. R. (1999). *Bootstrap Methods*. John Wiley and Sons, New York.

Corcoran, S. A. (1998). Bartlett adjustment of empirical discrepancy statistics. *Biometrika*, 85:967–972.

Cotofrei, P. (1998). A bootstrap test for the generalized Behrens–Fisher problem. Technical report, Università di Padova, Dipartimento di Scienze Statistiche.

Cotofrei, P. (1999). A possible generalisation of the empirical likelihood. In *Bulletin of the ISI*, pages 225–227, Helsinki. 52nd Session, Tome LVIII, Book 1.

Cressie, N. A. (1982). Playing safe with misweighted means. *Journal of the American Statistical Association*, 77:754–759.

Cressie, N. A. and Read, T. R. C. (1984). Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society series B*, 46:440–464.

Cressie, N. A. and Read, T. R. C. (1988). *Goodness-of-fit statistics for discrete multivariate data*. Springer Series in Statistics. Springer-Verlang.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap methods and their application*. Cambridge University Press.

Dharmadhikari, S. and Joag-Dev, K. (1988). *Unimodality, convexity and applications*. Academic Press, Boston.

DiCiccio, T. J., Hall, P., and Romano, J. P. (1989). Comparison of parametric and empirical likelihood functions. *Biometrika*, 76:465–476.

Ducharme, G. and Jhun, M. (1986). A note on the bootstrap procedure in testing linear hypotheses. *Statistics*, 17:527–531.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, 7:1–26.

Efron, B. (1982). The jackknife, the bootstrap and other resampling plans. In *CBMS-NSF Regional Conference Series in Applied Mathematics, No. 38*, Philadelphia. SIAM.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the bootstrap*. Chapmann and Hall, New-York.

Fenstad, G. (1983). Comparison between the U and the V test in the Behrens–Fisher problem. *Biometrika*, 70:300–302.

Fisher, R. (1935). The fiducial argument in statistical inference. *Annals of Eugenics*, 11:141–172.

Flinger, M. and Policello, G. (1981). Robust rank procedures for the Behrens–Fisher problem. *Journal of the American Statistical Association*, 76:162–168.

Flinger, M. and Rust, S. (1982). A modification of Mood's median test for the generalized Behrens–Fisher problem. *Biometrika*, 69:221–226.

Folks, J. (1984). *Handbook of Statistics*, chapter Combination of independent tests, pages 113–121. Elsevier Science Publishers.

Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*. Springer-Verlag, New York.

Hall, P. and Martin, M. (1988). On bootstrap resampling and iteration. *Biometrika*, 75:661–671.

Hall, P. and Wilson, S. (1991). Two guidelines for bootstrap hypothesis testing. *Biometrics*, 47:327–337.

Hettmansperger, T. and Malin, J. (1975). A modified Mood's test for location with no shape assumptions on the underlying distributions. *Biometrika*, 62:527–529.

Hinkley, D. and Shi, S. (1989). Importance sampling and the nested bootstrap. *Biometrika*, 76:435–446.

Jeffreys, H. (1948). *Theory of Probability*. University Press, Oxford, 2nd edition.

Jensen, J. L. (1992). The modified signed likelihood statistics and saddlepoint approximations. *Biometrika*, 79:693–703.

Kullback, S. (1959). *Information Theory and Statistics*. John Wiley, New York.

Lehmann, E. (1986). *Testing Statistical Hypotheses*. John Wiley, New York, 2nd edition.

Linnik, Y. V. (1975). *Problems of Analytical Statistics*. Statistical Publishing Society, Calcutta.

Liu, R. Y. and Singh, K. (1987). On a partial correction by the bootstrap. *Annals of Statistics*, 15:1713–1718.

Matusita, K. (1955). Decision rules based on the distance, for problems of fit, two samples, and estimation. *Annals of Mathematical Statistics*, 26:631–640.

Nell, D. G., Van Der Merwe, G. A., and Moser, B. K. (1990). The exact distribution of the univariate and multivariate Behrens–Fisher statistics with a comparison of several solutions in the univariate case. *Communications in Statistics, Theory and Methods*, 19:279–298.

Noreen, E. W. (1989). *Computer Intensive Methods for Testing Hypotheses: An Introduction*. Wiley, New York.

Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75:237–249.

Owen, A. B. (2001). *Empirical Likelihood*. Monographs on Statistics and Applied Probability. Chapman and Hall/CRC.

Pallini, A. and Pesarin, F. (1990). Combinazione di test dipendenti: una soluzione basata su tecniche di ricampionamento. *Proc. Riun. Scient. SIS*, 2:367–374.

Pallini, A. and Pesarin, F. (1992). A class of combinations of dependent tests by a resampling procedure. *Lecture Notes in Economics and Mathematical Systems*, 376:93–97.

Pesarin, F. (1988). Combinazione non parametrica di testi dipendenti. Technical report, Giornate di Metodologia Statistica, Università di Padova.

Pesarin, F. (1989). Nonparametric combination method for dependent permutation tests. Technical report, Università di Padova.

Pesarin, F. (1990). On a nonparametric combination method for dependent permutation tests with applications. *Psychotherapy and Psychosomatics*, 54:172–179.

Pesarin, F. (1991). Some multidimensional testing problems for missing values via resampling procedure. *Statistica Aplicata*, 3:569–577.

Pesarin, F. (1992). A resampling procedure for nonparametric combination of several dependent permutation tests. *Journal of the Italian Statistical Society*, 1:87–101.

Pesarin, F. (1995). A new solution for the generalized Behrens–Fisher problem. *Statistica*, 2:131–146.

Pfanzagl, J. (1974). On the Behrens–Fisher problem. *Biometrika*, 61:39–47.

Politis, D. N., Romano, J. P., and Wolf, M. (1999). *Subsampling*. Springer, New York.

Potthoff, R. F. (1963). Use of the Wilcoxon statistics for a generalized Behrens–Fisher problem. *Annals of Mathematical Statistics*, 34:1596–1599.

Rahtie, P. N. and Kannappan, P. (1972). A directed-divergence function of type β. *Information and Control*, 20:38–45.

Rényi, A. (1961). On measures of entropy and information. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 1:547–561.

Schlittgen, R. (1979). Use of the median test for the generalized Behrens–Fisher problem. *Metrika*, 26:95–103.

Schuster, E. F. and Barker, R. C. (1987). Using the bootstrap in testing symmetry versus asymmetry. *Communication in Statistics B*, 16:68–84.

Shao, J. and Tu, D. (1995). *The Jackknife and Bootstrap*. Springer-Verlag, New York.

Sutton, C. (1993). Computer-intensive methods for tests about the mean of an asymmetric distribution. *Journal of American Statistical Association*, 88:802–810.

Tiku, L. M. and Sing, M. (1981). Robust test for means when population variances are unequal. *Communications in Statistics, Theory and Methods, A10*, 20:2057–2071.

Wang, Y. Y. (1971). Probabilities of the type I errors of the Welch test for the Behrens–Fisher problem. *Journal of the American Statistical Association*, 66:605–608.

Welch, B. L. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika*, 29:350–362.

Westberg, M. (1986). Tippett–adaptive method of combining independent statistical tests. Technical Report 3-1986, Department of Statistics, Goteborg.

Westfall, P. H. and Young, S. S. (1993). *Resampling-based multiple testing: Examples and methods for p-value adjustment*. John Wiley, New York.

Yuen, K. (1974). An approximate degrees of freedom solution to the multivariate Behrens-Fisher problem. *Biometrika*, 61:165–170.

Zanella, A. (1973). Sulle procedure di classificazione simultanea. *Statistica*, 33:63–120.