Bootstrap tests for robust means of asymmetric distributions with unequal shapes

A. Marazzi

Institut universitaire de médecine sociale et préventive,

University of Lausanne, Lausanne, Switzerland

Abstract

Bootstrap procedures for testing equality of robust means in the one-, two- , and multi-sample problems for asymmetrically distributed data with unequal shapes are described. The emphasis is on parametric procedures, but some results are provided for nonparametric procedures as well. In the parametric framework, it is assumed that a model with two parameters, shape and scale, can be used to approximatively describe the populations. Examples are contaminated Lognormal, Weibull, Gamma, and Pareto distributions. Robust estimators of the parameters are supposed to be available; the robust mean is then defined as the asymptotic value of the robust estimate of the model mean. In the nonparametric framework, the robust mean is the asymptotic value of some estimate that does not depend on a parametric model, e.g., a trimmed mean. A studentized test statistic is explored with the help of examples on simulated and real data. In order to estimate the null model, criteria for robust constrained model fitting, the constraint being the null hypothesis, are introduced and discussed. In the nonparametric case, a robust version of exponential tilting is provided. Parametric, semiparametric, and nonparametric bootstrap schemes for the computation of finite sample distributions are are considered. The examples illustrate procedures that can be useful in practice.

Keywords. Asymmetric distributions, constrained estimation, exponential tilting, robust inference, bootstrap inference, equality of means, tests for means.

Running title. Bootstrap tests for robust means.

Address of the corresponding author. Alfio Marazzi, Institut universitaire de médecine sociale et préventive, Bugnon 17, CH 1005 Lausanne, Switzerland; E-mail: alfio.marazzi@unil.ch

1 Introduction

Positive random variables with asymmetric distributions arise in many statistical applications and often the population mean (i.e., the expected value of the random variable) is the parameter of interest. We consider the problem of testing the hypothesis that the mean of a single population equals a given value (one-sample problem) or that the means of several populations are identical (multi-sample problem) and, in both cases, we use the notation $\mathcal{H}_0$ to denote this hypothesis. Unfortunately, testing $\mathcal{H}_0$ is a difficult problem when the data contain outliers.

For example, the length of stay (LOS) is often used as an indicator of hospital resource consumption and LOS means of medically homogeneous groups of patients are used for budgeting, the cost of such a group of patients being proportional to its mean LOS. It is common to compare LOS means among different hospitals or over different periods of time and it is easy to give examples where common tests of mean (e.g. the t-test on transformed or untransformed LOS) lead to the opposite decision when a single atypical stay is removed from the data set. Procedures for automatic outlier detection, robust mean estimation, and robust mean comparison are therefore of great interest in this framework.

Many robust nonparametric techniques for comparing asymmetric populations are available (Hettmansperger and McKean, 1998). The usual assumption underlying these methods is that the populations differ by at most shifts in location. In the one-sample case, $\mathcal{H}_0$ is then equivalent to the hypothesis that the location measure (e.g., the median) equals a given value; in the two-sample case, the hypothesis of a null shift is equivalent to $\mathcal{H}_0$. However, in many applications the distributions differ in shape and the mean is not a location parameter. The problem of testing $\mathcal{H}_0$ in this case has been considered by some authors: Cressie and Whitford (1986) propose corrections of the two sample t-test to account for nonzero skewness and kurtosis; Shiue et al.(1988) describe tests of equal Gamma distribution means with unequal shapes. Unfortunately, these tests are based on classical measures of moments and are not outlier resistant. There appears to be few results on tests based on robust estimates of means of asymmetric distributions; see Tiku (1982), Wilcox (1990; with a brief overview) and Wilcox et al. (1998).

On some occasions one may find a symmetry-creating transformation (e.g., the log) and consequently use robust tools, such as trimmed means and M-estimates, that are then available (Hampel et al, 1986; Staudte an Sheather, 1990). This approach – recently followed by Guo and Luh (2000) – also has a few drawbacks: for example, unless two transformed samples are homoscedastic, the hypothesis of a null shift in the log scale is not equivalent to the target hypothesis of equality of means and it may be difficult to find a simple transformation for both symmetry and homoscedasticity. Moreover, in the asymmetric case, dispersion is usually a main component of mean, not a nuisance parameter as in the symmetric case; therefore, the usual robust estimates of location and dispersion lead to inefficient estimates (and low power tests) for means when transformed back to the original scale.

The trimmed mean of the untransformed data is an appealing nonparametric robust estimate of the population mean after removal of a fixed proportion α of extremely large values. Its breakdown point is α, which can be as high as 50% (but the estimate cannot be interpreted as a mean in this case). Usual trimming proportions range from 5% to 10% and can be satisfactory for mildly to moderately contaminated data. In case of moderate to severe contaminations, and when a parametric model is a reasonable description of the data majority, it is be preferable to use a parametric robust estimate. For example, Victoria-Feser and Ronchetti (1994, 1997) use M-estimators of the Gamma distribution to describe income data; Marazzi et al. (1998) use the same kind of estimators to model hospital length of stay with Lognormal, Weibull, and Gamma distributions. Highly robust and efficient parametric estimates of the population mean have been recently developed. One of them is the "truncated mean" proposed by Marazzi and Ruffieux (1999). This estimate is similar to a trimmed mean but the trimming points are provided by a parametric model. Unlike the common trimmed mean, the truncated mean does not remove a fixed proportion of extreme observations; it is asymptotically unbiased under the model, has an efficiency of about 80%, and can attain the maximum 50% breakdown point.

When a certain parametric or nonparametric estimator has been chosen, it is natural to test hypotheses concerning "robust means" defined through this estimator. It is the purpose of this paper to study this kind of test. Formally, let Y_j, $j = 1, \ldots, k$, be random variables with unknown cdf-s F_j, k being the number of samples. In the parametric case, we assume that the F_j-s can be approximately described with the help of some, possibly contaminated, parametric models with shapes α_j and scales σ_j. Examples are Lognormal, Weibull, Gamma, and Pareto distributions. Let $\theta_j = (\alpha_j, \sigma_j)$ and denote by F_{θ_j} the corresponding cdf with density f_{θ_j}. We consider the functional $\mu(F) = \int y dF(y)$, use the abbreviation $\mu(\theta_j)$ in place of $\mu(F_{\theta_j})$, and assume that $\mu(\theta_j) = \sigma_j \xi(\alpha_j)$ for some positive function $\xi(\cdot)$. The examples just mentioned satisfy this condition. Let $y_j = (y_{j1}, \ldots, y_{jn_j})$,

$j = 1, \ldots, k$, be samples of iid observations from F_j and let F_{n_j,y_j} be the empirical cdf of y_j. We assume that some procedure for the computation of a robust Fisher consistent and asymptotically normal estimator $\hat{\theta}(F_{n_j,y_j}) = (\hat{\alpha}(F_{n_j,y_j}), \hat{\sigma}(F_{n_j,y_j}))$ of θ_j with asymptotic value $\hat{\theta}(F_j)$ and asymptotic covariance matrix $V(\hat{\theta}, F_j)$ is available (see references above), that $\hat{\sigma}$ is scale equivariant, and define $\hat{\mu}(F_j) = \mu(F_{\hat{\theta}(F_j)})$. In the nonparametric case, $\hat{\mu}(F_{n_j,y_j})$ is an estimator that does not depend on a parametric model, e.g., a sample trimmed mean; we assume that $\hat{\mu}(F_{n_j,y_j})$ is scale equivariant and asymptotically normal with mean $\hat{\mu}(F_j)$ and variance $V(\hat{\mu}, F_j)$.

We define the parameter of interest in terms of the population value of its estimate (Huber 1981, pp. 6-7). In other words, we make inferences for $\hat{\mu}(F_j)$ (in place of $\mu(F_j)$) and call $\hat{\mu}(F_j)$ a *robust mean* of F_j. In this context, the robust means represent approximations of the population means after removal of the extreme values (see also Wilcox et al., 1998, p. 133). We are interested in testing the hypothesis that the robust mean of F_1 equals a given value or that the F_j-s have a common robust mean, without assuming identical shapes or scales. We use simple studentized test statistics that can be based on parametric or nonparametric estimators $\hat{\mu}(F_{n_j,y_j})$ and the bootstrap as a general tool to compute the finite sample null distribution of the test statistic.

According to the classical approach to bootstrap tests, simulated samples are generated from estimates $\hat{F}_j$ of the F_j-s under the null hypothesis, i.e., from the null model. Often, this can be avoided by using a pivotal test statistic (see, e.g., Davison and Hinkley, 1997, p. 171) or a depth based test statistics (Liu and Singh, 1997). However, a null model may be useful to compute confidence intervals when the hypothesis is accepted. To compute a nonparametric null model, one often assumes that the $\hat{F}_j$-s are supported on the corresponding sample values and minimizes the aggregated Kullback-Leibler disparity betwen $\hat{F}_j$ and F_{n_j,y_j} subject to the constraint of equal means (Efron and Tibshirani, 1993; Davison and Hinkley, 1997). The constrained nonparametric maximum likelihood estimates $\hat{F}_j$ obtained in this way are then called "exponentially tilted versions" (briefly, "exponential tilts") of F_{n_j,y_j}.

Within the nonparametric setting, we will show how to compute robust exponential tilts of F_{n_j,y_j} under the constraint of identical robust means. The procedure is however rather complex. In the parametric setting we will proceed in an analogous way and consider estimates F_{θ_j} that minimize the aggregated Kullback-Leibler disparity between F_{θ_j} and $F_{\hat{\theta}_j}$ under the same constraint. We use the available full model estimates $F_{\hat{\theta}_j}$ in place of F_{n_j,y_j} avoiding the computation of complex constrained robust estimators. We show however that, if the $\hat{\theta}_j$ are unconstrained maximum likelihood estimates, our criterion provides the exact constrained maximum likelihood estimate in the case of the exponential family and good approximations in other cases. It is therefore a reasonable criterion to derive a robust constrained estimator from unconstrained ones.

The paper is organized as follows. Section 2 describes a simple studentized test statistic that can be based both on parametric and nonparametric estimates. Section 3 considers the computation of the exponential tilts under the null hypotheses for robust means. Section 4 introduces the parametric constrained robust estimation criteria. Section 5 indicates in detail how the constrained estimates are used to set up the null models used in the bootstrap simulations. In Section 6, the procedures are evaluated and compared to their classical counterparts with the help of a real set of hospital length of stays and using simulated data. Since the truncated mean does not strongly depend on the model, it is used to illustrate both the parametric and the nonparametric approaches. The results are summarized in Section 7. Technical details and proofs are given in the appendix.

2 The test statistics

We use the abbreviations $y = (y_1, y_2, \ldots, y_k)$, $F = F_1 \cdot F_2 \cdot \ldots \cdot F_k$, $F_{n,y} = F_{n_1,y_1} \cdot F_{n_2,y_2}$, and $\hat{\mu}_j = \hat{\mu}(F_{n_j,y_j})$, $j = 1, \ldots, k$. We consider simple studentized test statistics $t(F_{n,y})$, that can be based on parametric or nonparametric estimators $\hat{\mu}(F_{n_j,y_j})$, assuming that procedures for the computation of the estimators and their asymptotic variances $V(\hat{\mu}, F)$ for $F = F_{n,y}$ are available. The bootstrap null distribution of t is then obtained using simulated values $t(F_{n,y^*})$, where y^* is drawn from estimates $\hat{F}_0$ of F that satisfy the null hypothesis. Methods to determine $\hat{F}_0$ will be described in the next three sections.

2.1 One-sample problem. For a moment, we drop the suffix "1" from F_1, Y_1, y_{1i}, n_1, θ_1 etc., and consider the hypothesis

$$\mathcal{H}_0: \ \hat{\mu}(F) = \mu_0.$$

Various test statistics could be considered for this hypothesis but, for the purpose of illustration, we restrict our attention to

$$t(F_{n,y}) = d(F_{n,y})/v(d, F_{n,y})^{1/2}, \tag{2.1}$$

where $d(F) = h(\hat{\mu}(F)) - h(\mu_0)$, $v(d, F_{n,y}) = n^{-1} h'(\hat{\mu}(F_{n,y}))^2 V(\hat{\mu}, F_{n,y})$, and h is a variance stabilizing transformation. This is a Wald type test statistic for $h(\hat{\mu}(F))$. In general, h is chosen so that $h'(\hat{\mu})^2 V(\hat{\mu}, F)$ is approximately a constant function of $\hat{\mu}$; an important special case is $h(\cdot) = \ln(\cdot)$. Under $\mathcal{H}_0$, the statistic t is asymptotically normal; moreover, t is approximately a pivot (meaning that its distribution is approximately independent of unknown parameters), and its power function does non depend on h (Appendix 1).

2.2 Two-sample problem. We consider the hypothesis

$$\mathcal{H}_0 : \ \hat{\mu}(F_1) = \hat{\mu}(F_2),$$

define $d(F) = h(\hat{\mu}_2) - h(\hat{\mu}_1)$, $v(d, F_{n,y}) = \sum n_j^{-1} h'(\hat{\mu}_j)^2 V(\hat{\mu}_j, F_{n_j, y_j})$, and use (2.1) to define the test statistic. The comments following (2.1) are still appropriate.

2.3 Multi-sample problem. To test

$$\mathcal{H}_0 : \ \hat{\mu}(F_1) = \ldots = \hat{\mu}(F_k)$$

the statistics t can be extended in various ways, e.g,

$$t(F_{n,y}) = \sum w_j (h(\hat{\mu}_j) - h(\hat{\mu}_0))^2,$$

where $\hat{\mu}_0 = \sum w_j h(\hat{\mu}_j) / \sum w_j$ and $w_j = n_j / [h'(\hat{\mu}_j)^2 V(\hat{\mu}_j, F_{n_j, y_j})]$ (Welch, 1951). Asymptotically, t has a χ^2 distribution with $k-1$ degrees of freedom when $\mathcal{H}_0$ is true.

3 Exponential tilting for a robust constraint

In this section, we modify the methods described in Efron and Tibshirani (1993, p. 235) and Davison and Hinkley (1997, p. 166) to compute constrained nonparametric maximum likelihood estimators of the F_j-s. In the classical setup the constraint is $\mu(F_1) = \mu_0$ (one-sample problem) or $\mu(F_1) = \mu(F_2)$ (two-sample problem); here, we replace μ with a robust estimator $\hat{\mu}$. The generalization from the two-sample to the multi-sample problem is straightforward and is left to the reader.

3.1 The one-sample case. As in Section 2.1, we drop the suffix "1" from F_1, Y_1, y_{1i}, n_1, etc. Let $\mathcal{P}$ be the set of discrete distributions $F_p = \sum p_i \Delta_{y_i}$, where Δ_{y_i} denote the cdf of a point mass at y_i, $p_i \geq 0$ for all i, $p = (p_1, \ldots, p_n)$, and let $\mathcal{P}_0 = \{F_p \in \mathcal{P} \mid \hat{\mu}(F_p) = \mu_0\}$. We look for a distribution $F_{\tilde{p}} \in \mathcal{P}_0$, such that

$$d_{KL}(F_{n,y}, F_{\tilde{p}}) = \min_{\mathcal{P}_0} d_{KL}(F_{n,y}, F_p),$$

where $d_{KL}(F_{n,y}, F_p) = \sum p_i \ln(np_i)$ denotes the Kullback-Leibler disparity between $F_{n,y}$ and F_p. Unlike the classical problem, the constraint $\hat{\mu}(F_p) = \mu_0$ is nonlinear in p.

In order to solve this optimization problem, the influence function (IF, Hampel et al., 1986) of $\hat{\mu}$ must be computed. By definition,

$$IF(y; \hat{\mu}, F) = \lim_{t \downarrow 0} [\hat{\mu}((1-t)F + t\Delta_y) - \hat{\mu}(F)]/t$$

where y denotes a value of Y and F the cdf of Y. We use the abbreviation $I_i(F) = IF(y_i; \hat{\mu}, F)$, the infinitesimal influence of y_i on $\hat{\mu}(F)$. The following proposition (proved in Appendix 2) gives a necessary condition for optimality.

PROPOSITION 1. If $F_{\tilde{p}}$ minimizes $d_{KL}(F_{n,y}, F_p)$ under the constraint $\hat{\mu}(F_p) = \mu_0$ and $0 < \tilde{p}_i < 1$ for all i, then $F_{\tilde{p}}$ is an *exponential tilt* of $F_{n,y}$, i.e.,

$$\tilde{p}_i = \exp(\lambda I_i(F_{\tilde{p}})) / \sum_i \exp(\lambda I_i(F_{\tilde{p}})). \tag{3.1}$$

The probabilities $\tilde{p}_i$ depend on the scalar λ that must be determined so that $\hat{\mu}(F_{\tilde{p}}) = \mu_0$.

REMARK. When $\hat{\mu}(F) = \mu(F)$, $I_i(F) = y_i$ and Proposition 1 characterizes the classical exponential tilt of $F_{n,y}$.

Two difficulties must be faced in order to use this proposition. First, the IF-s of many robust estimators (e.g., the trimmed means, Hampel et al 1986, p. 108-110) do not exist when F is not continuous; second, condition (3.1) is a highly dimensional nonlinear system of equations that must be solved numerically.

In order to compute $I_i(F_p)$ for a discrete F_p, we propose to replace the original estimator $\hat{\mu}$ with a "smoothed version" as follows. Let $\bar{F}_p(y) = \sum p_i G_{y_i}^{\delta_i}(y)$ be a smooth of F_p, where $G_{y_j}^{\delta_i}$ denotes a continuous cdf (for simplicity, we take a normal cdf) with location y_i and scale δ_i; the scales δ_i, $i = 1, \ldots, n$ are fixed by the user. The smooth of $\hat{\mu}$ is then defined as

$$\bar{\hat{\mu}}(F_p) = \hat{\mu}(\bar{F}_p).$$

It then follows that (Huber, 1981, p.37)

$$IF(y_j;\bar{\hat{\mu}},F_p)=\int IF(y;\hat{\mu},\bar{F}_p)dG^{\delta}_{y_j}(y).$$

Clearly, in order to maintain the robustness properties of $\hat{\mu}$, the δ_i-s must be chosen as small as possible, while preserving the stability of computations. As a guideline, the values of $\hat{\mu}(\bar{F}_p)$ and $\bar{\hat{\mu}}(F_p)$ as well as the influences $IF(y_j;\bar{\hat{\mu}},F_p)$ and $IF(y;\hat{\mu},\bar{F}_p)$ should be close. An example is considered in Section 6.

A simple algorithm to solve (3.1) applies the Picard iteration

$$\tilde{p}_i^{(k+1)}=\exp(\lambda I_i(F_{\tilde{p}^{(k)}}))/\sum_i\exp(\lambda I_i(F_{\tilde{p}^{(k)}}))$$

until convergence, for various values of λ. This provides a set of approximate solutions $\tilde{p}_i(\lambda)$, from which the one that solves $\hat{\mu}(F_{\tilde{p}(\lambda)})=\mu_0$ is chosen. Unfortunately, the convergence of this algorithm is impeded by the high dimensionality and because the influences have to be evaluated with numerical integration. However, when $\mathcal{H}_0$ is true, $F_{\tilde{p}}$ is close to $F_{n,y}$ and we may replace the nonlinear function $\hat{\mu}(F_p)$ with the linear approximation $\hat{\mu}(F_{n,y})+\sum p_i\hat{I}_i$, where $\hat{I}_i=I_i(F_{n,y})$ (Hampel et al., 1986, p.85) in the constraint. Note that $\hat{I}_i$ does not depend on $\tilde{p}$ and that the necessary condition then becomes more explicit,

$$\tilde{p}_i=\exp(\lambda\hat{I}_i)/\sum_i\exp(\lambda\hat{I}_i),$$

where λ satisfies $\hat{\mu}(F_{n,y})+\sum\tilde{p}_i\hat{I}_i=\mu_0$. An example is considered in Section 6.

3.2 The two-sample case. In this subsection we consider the problem of minimizing the aggregated Kullback-Leibler disparity $\sum_{j=1}^{2}\sum_{i=1}^{n_j}p_{ji}\ln(p_{ji})$ under the constraint $\hat{\mu}(F_{p_1})=\hat{\mu}(F_{p_2})$ where the F_{p_j}, $j=1,2$, are discrete distributions that put probabilities p_{ji} on y_{ji} for each i. Linearizing the constraint, one obtains the necessary conditions

$$\begin{aligned}\tilde{p}_{1i}&=\exp(\lambda\hat{I}_{1i})/\sum\exp(\lambda\hat{I}_{1i}), && i=1,\dots,n_1,\\ \tilde{p}_{2j}&=\exp(-\lambda\hat{I}_{2j})/\sum\exp(-\lambda\hat{I}_{2i}), && j=1,\dots,n_2,\end{aligned}\tag{3.2}$$

where $\hat{I}_{ji}=IF(y_{ji};\hat{\mu},F_{n_j,y_j})$ and λ satisfies $\hat{\mu}(F_{n_1,y_1})+\sum\tilde{p}_{1i}\hat{I}_{1i}=\hat{\mu}(F_{n_2,y_2})+\sum\tilde{p}_{2i}\hat{I}_{2i}$. Finally, note that the approximation

$$\hat{\mu}(F_{\tilde{p}_1})\approx\hat{\mu}(F_{n_1,y_1})+\sum\tilde{p}_{1i}\hat{I}_{1i}\tag{3.3}$$

provides an estimate to the common robust mean of F_1 and F_2.

4 Parametric constrained robust estimation

We now consider methods to compute parametric robust estimates of the F_j-s, under the constraints given by the one- and the multi-sample null hypotheses. As in the nonparametric case we look for maximum likelihood type estimators.

Let $\ell(G, \Delta_y)$ denote the log-likelihood of a model with cdf G with respect to an observation y and Δ_y the cdf of point mass at y. We define $\ell(G, H) = \int \ell(G, \Delta_y) dH(y)$, where H is an arbitrary cdf. The Kullback-Leibler disparity (Lindsey, 1996, p. 101) between G and H is then

$$d_{KL}(G, H) = \ell(H, H) - \ell(G, H).$$

In the following, we consider the problem of minimizing $d_{KL}(G, H)$ over G for given H, under a certain constraint. This problem is equivalent to the problem of maximizing $\ell(G, H)$ under the same constraint.

4.1 The one-sample problem. As in Section 2.1, we drop the suffix "1" from the notations F_1, Y_1, y_{1i}, n_1, θ_1 etc. We consider the problem of fitting a parametric model F_θ to the empirical cdf $F_{n,y}$ under the constraint

$$\mathcal{K}_0 : \ \mu(\theta) = \mu_0,$$

where μ_0 is a given value. Let $\Theta_0 = \{\theta \mid \mu(\theta) = \mu_0\}$ be the parameter space under $\mathcal{K}_0$.

The *constrained maximum likelihood (CML) estimator* $\bar{\theta}_c = (\bar{\alpha}_c, \bar{\sigma}_c)$ of θ is defined as

$$\bar{\theta}_c = \text{argmin}_{\Theta_0} \ d_{KL}(F_\theta, F_{n,y}). \tag{4.1}$$

Since $\mathcal{K}_0$ can be re-expressed as $\sigma = \mu_0/\xi(\alpha)$, $\bar{\theta}_c$ is best computed by solving

$$\sum_{i=1}^{n} s_0(\alpha; y_i) = 0,$$

where $s_0(\alpha; y)$ is the score function of the reduced model: $s_0(\alpha; y) = \partial/\partial\alpha \ \ell(F_{\alpha,\mu_0/\xi(\alpha)}; y)$. (Obvious changes can be made if it is more convenient to express α as a function of σ.)

A natural robust extension of the CML-estimator is the *constrained M-estimator* $\hat{\theta}_c = (\hat{\alpha}_c, \hat{\sigma}_c)$, where $\hat{\alpha}_c$ is a solution of

$$\sum_{i=1}^{n} \psi[s_0(\alpha; y_i) - B(\alpha)] = 0, \tag{4.2}$$

and $\hat{\sigma}_c = \mu_0/\xi(\hat{\alpha}_c)$. $B(\alpha)$ is an implicitly defined bias correction (Hampel et al., 1986, p. 117) such that

$$\int \psi[s_0(\alpha; y) - B(\alpha)] f_{(\alpha, \mu_0/\xi(\alpha))}(y) dy = 0;$$

$\psi(\cdot)$ is some bounded function, e.g. Huber's function $\max[-b, \min(b, x)]$) with tuning constant b. Unfortunately, this estimator is computationally awkward, especially in the multisample case, and programs are not available. Users of other types of estimators (e.g., the truncated mean) encounter anologous difficulties. We propose therefore a very simple criterion to compute a constrained estimator when an unconstrained one, say $\hat{\theta}$, is available. The idea is to replace $F_{n,y}$ by $F_{\hat{\theta}}$ in (4.1).

Criterion C. Use the constrained estimator $\check{\theta}_c = (\check{\alpha}_c, \check{\sigma}_c)$ defined as the value of θ that satisfies

$$\check{\theta}_c = \text{argmin}_{\Theta_0}\, d_{KL}(F_\theta, F_{\hat{\theta}}), \tag{4.3}$$

where $\hat{\theta}$ is an available unconstrained robust estimator. We call $\check{\theta}_c$ the *constrained optimal disparity estimator based on* $\hat{\theta}$, abbreviated the *C-estimator*.

EXAMPLE. Suppose that Y has a Lognormal distribution $F_{(\lambda,\tau)}$ and let $G_{(\lambda,\tau)}$ denote the normal distribution of $X = \ln(Y)$ with mean λ and scale τ. The scale parameter of $F_{(\lambda,\tau)}$ is $\sigma = \exp(\lambda)$ and the shape parameter is $\alpha = \tau$. Suppose that $(\hat{\lambda}, \hat{\tau})$ is a robust estimate of (λ, τ), e.g., the sample median and median absolute deviation. With $z = (x - \lambda)/\tau$, we obtain

$$\begin{aligned}
\ell(G_{(\lambda,\tau)}, \Delta_x) &= -z^2/2 - \ln(\tau) - \ln(\sqrt{2\pi}), \\
\ell(G_{(\lambda,\tau)}, G_{(\hat{\lambda},\hat{\tau})}) &= -\frac{1}{2\tau^2}[\hat{\tau}^2 + (\lambda - \hat{\lambda})^2] - \ln(\tau) - \ln(\sqrt{2\pi}), \\
\ell(F_{(\lambda,\tau)}, F_{(\hat{\lambda},\hat{\tau})}) &= \ell(G_{(\lambda,\tau)}, G_{(\hat{\lambda},\hat{\tau})}) - \hat{\lambda}.
\end{aligned}$$

Using the constraint $\lambda + \tau^2/2 = \ln(\mu_0)$, we have $\partial/\partial\tau \ell(G_{(\ln(\mu_0)-\tau^2/2,\tau)}, G_{(\hat{\lambda},\hat{\tau})}) = [\hat{\tau}^2 + (\ln(\mu_0) - \hat{\lambda})^2]/\tau^3 - \tau/4 - 1/\tau$ and

$$\begin{aligned}
\check{\lambda}_c &= \ln(\mu_0) - \check{\tau}_c^2/2, \\
\check{\tau}_c^2 &= 2\{[1 + \hat{\tau}^2 + (\hat{\lambda} - \ln(\mu_0))^2]^{1/2} - 1\},
\end{aligned}$$

provides the constrained maximum of $\ell(F_{(\lambda,\tau)}, F_{(\hat{\lambda},\hat{\tau})})$. Further examples can be found in Appendix 4.

Obviously, we may take the maximum likelihood estimator (ML) as a particular non-robust unconstrained estimator $\hat{\theta}$. We can then prove (Appendix 3) the following proposition.

PROPOSITION 2. Let $F_\theta(y)$ denote the cdf of a member of the exponential family of distributions in the natural parametrization and let $\hat{\theta}$ be the maximum-likelihood estimator of θ. Any estimator that maximizes $\ell(F_\theta, F_{n,y})$ under some constraint for θ coincides with the estimator that maximizes $\ell(F_\theta, F_{\hat{\theta}})$ under the same constraint.

It then follows that the C-estimator based on the unconstrained ML-estimator coincides with the CML-estimator, when F_θ belongs to the exponential family of distributions. This result obviously extends to models that can be transformed to members of the exponential family, e.g. the Lognormal distribution. For the Weibull model, numerical comparisons show that the C-estimator is a very good approximation. Since good robust estimators are close to the ML-estimator in the absence of outliers, we conjecture that the criterion (4.3) is a promising simple way to obtain a constrained robust estimator when an unconstrained one is already available. The asymptotic normality of the C-estimator readily follows from the assumed properties of the unconstrained estimates; the asymptotic covariance matrix can be computed with the help of the IF. Appendix 5 shows that the IF of the C-estimator is proportional to the IF of $\hat{\theta}$ and, therefore, bounded. Clearly, the C-estimator inherits the breakdownpoint of $\hat{\theta}$, which may be as high as 50%.

Finally, we consider a crude short cut that requires no extra computational effort when $\hat{\theta}$ is available, but that performs well in many applications (Section 5 and Section 6).

Criterion Q. Use the constrained estimator

$$\acute{\theta}_c = \text{argmin}_{\Theta_0}\ d_\alpha(F_\theta, F_{\hat{\theta}}), \tag{4.4}$$

where $d_\alpha(F_\theta, F_{\hat{\theta}}) = |\alpha - \hat{\alpha}|$ and $\hat{\theta}$ is the available unconstrained robust estimator. In other words, set $\acute{\alpha}_c = \hat{\alpha}$ and $\acute{\sigma}_c = \mu_0/\xi(\hat{\alpha})$. We call $\acute{\theta}_c$ the *quickly constrained estimator based on* $\hat{\theta}$, or simply the *Q-estimator*.

REMARK. If the constrained model is correct and $\hat{\theta}$ is asymptotically unbiased, the estimators provided by criterions C and Q are asymptotically unbiased.

4.3 The multi-sample problem. We now consider the problem of fitting k models F_{θ_j} $(\theta_j = (\alpha_j, \sigma_j),\ j = 1, \dots, k)$ under the constraints

$$\mathcal{K}_0 :\ \mu(\theta_1) = \dots = \mu(\theta_k).$$

We use the abbreviations $\theta = (\theta_1, \dots, \theta_k)$, $\hat{\theta} = (\hat{\theta}_1, \dots, \hat{\theta}_k)$, $F_\theta = F_{\theta_1} \cdot \ldots \cdot F_{\theta_k}$, $F_{n,y} = F_{n_1,y_1} \cdot \ldots \cdot F_{n_k,y_k}$, and define

$$d_{KL}(F_\theta, F_{n,y}) = \sum_{j=1}^{k} n_j d_{KL}(F_{\theta_j}, F_{n_j,y_j}).$$

With this extension of d_{KL}, the *k-sample CML-estimate* $\bar{\theta}_c = ((\bar{\alpha}_{jc}, \bar{\sigma}_{jc}))$ is given again by (4.1), where Θ_0 is the multi-sample parameter space under $\mathcal{K}_0$. Using

$$d_{KL}(F_\theta, F_{\hat{\theta}}) = \sum_{j=1}^{k} n_j d_{KL}(F_{\theta_j}, F_{\hat{\theta}_j})$$

in (4.3) we have the *k-sample constrained C-estimator* $\check{\theta}_c$ based on the available estimates $\hat{\theta}_1, \dots, \hat{\theta}_k$.

We remark that the $2k$-dimensional optimization problem defined by (4.3) is equivalent to the one-dimensional unconstrained problem of computing

$$\min_{\mu_0} \sum_{j=1}^{k} n_j d_{KL}(F_{a_j(\mu_0), z_j(\mu_0)}, F_{\hat{\theta}_j}),$$

where $z_j(\mu_0) = \mu_0/\xi(a_j(\mu_0))$ and $a_j(\mu_0) = \text{argmin}_\alpha d_{KL}(F_{\alpha,\mu_0/\xi(\alpha)}, F_{\hat{\theta}_j}), j = 1, \dots, k$. Moreover, the problem of computing $a_j(\mu_0)$ for a given μ_0 is equivalent to the one-sample version of (4.3). Thus, numerical procedures for computing the one-sample C-estimator, are also useful for the k-sample case. It also follows that the k-sample C-estimator based on the unconstrained ML-estimator coincides with the k-sample CML-estimator (Lognormal model and exponential family) or it is a very good approximation (Weibull model). Moreover, if the constrained model is correct, the estimator provided by criterion C is asymptotically unbiased.

Finally, using

$$d_\alpha(F_\theta, F_{\hat{\theta}}) = \sum n_j d_\alpha(F_{\theta_j}, F_{\hat{\theta}_j})$$

in (4.4), one obtains the *k-sample constrained Q-estimator based on* $\hat{\theta}$, i.e., $\acute{\alpha}_{jc} = \hat{\alpha}_j$, $\acute{\sigma}_{jc} = \mu_0/\xi(\hat{\alpha}_j)$, $j = 1, \ldots, k$, where μ_0 is an arbitrary common mean. An additional criterion should be used to determine μ_0; this is however not necessary for the purpose of testing (Section 5).

5 Setting up the null models

Here, we indicate in detail how the constrained estimates are used to set up the null models $\hat{F}_0$ for the simulation. We follow the general guidelines discussed in Efron and Tibshirani (1993, Chapter 16) and Davison and Hinkley (1997, Chapter 4). In addition, we briefly discuss the computation of confidence intervals. We consider the one- and two-sample cases, the generalisation to several samples being straightforward.

One-sample problem. A nonparametric null model based on an exponentially tilted version of $F_{n,y}$ is obviously determined according to the procedure described in Section 3.1; thus, $\hat{F}_0 = F_{\tilde{p}}$.

For a parametric model, the condition $\mathcal{H}_0 : \hat{\mu}(F_\theta) = \mu_0$ is equivalent to $\mathcal{K}_0 : \mu(\theta) = \mu_0$, when $\hat{\theta}$ is Fisher consistent. We may, therefore, use the estimates $\check{\theta}_c$ (Criterion C) or $\acute{\theta}_c$ (Criterion Q) defined in Section 4.1 and set $\hat{F}_0 = F_{\check{\theta}_c}$ or $\hat{F}_0 = F_{\acute{\theta}_c}$. Note that the choice of the null model assumes that the possible distributions of Y are just rescaled versions of one another (i.e., $Y \sim F_{(\alpha, \mu_0/\xi(\alpha))}$ for some α) and that μ_0 is a scale factor. Thus, the test statistic $t(F_{n,y})$ with $h(\cdot) = \ln(\cdot)$, as well as its null distribution, do not depend on μ_0.

In order to reduce the model dependency of $\hat{F}_0$, one may also consider a semiparametric approach based on the assumption that $Y \sim G(\cdot/\sigma)$ for some unspecified distribution G and set $\hat{F}_0 = F_{n,\hat{y}}$, where $\hat{y} = (\mu_0/\hat{\mu})y$ and $\hat{\mu} = \hat{\mu}(F_{n,y})$. $\hat{F}_0$ satisfies $\mathcal{H}_0$ because $\mu_0/\hat{\mu}$ is a scale factor and $\hat{\mu}$ is scale equivariant. This procedure is a simple reformulation for scale and shape models of the semiparametric approach described in Efron and Tibshirani (1993, pp.224) for location and scale models.

Two-sample problem. A nonparametric robust null model based on exponential tilts of $F_{n,y}$ can be determined according to the methods of Section 3.2: we set $\hat{F}_{j0} = F_{\tilde{p}_j}$.

The parametric approach sets $\hat{F}_{j0} = F_{\check{\theta}_{jc}}$ or $\hat{F}_{j0} = F_{\acute{\theta}_{jc}}$, the parameter estimates being defined in Section 4.2. Note that, when using the Q-estimator, the common mean

μ_0 (Section 4) is a scale factor; therefore, if $h(\cdot) = \ln(\cdot)$, the test statistic t does not depend on μ_0 which can be arbitrarily chosen.

If we just assume that $Y_j \sim G_j(\cdot/\sigma_j)$ $(j = 1, 2)$ for some unspecified distributions G_1 and G_2, then we can use a semiparametric approach and set $\hat{F}_{j0} = F_{n_j, \hat{y}_j}$, where $\hat{y}_j = [\mu_0/\hat{\mu}_j]y_j$ and μ_0 is arbitrary.

Confidence intervals. In the one-sample case, confidence intervals for $\hat{\mu}(F)$ are based on simulated values of the unconstrained estimator $\hat{\mu}(\hat{F}_{n,y^*})$, where y^* is drawn from $F_{n,y}$ (nonparametric case) or from $F_{\hat{\theta}}$ (parametric case). Similarly, in the two sample case, confidence intervals for $d(F)$ are based on simulated values $d(F_{n,y^*})$, where y^* is drawn from (parametric or nonparametric) unconstrained estimates of F. However, if the hypothesis is accepted, confidence intervals for a common mean have to be computed on the ground of the constrained two-sample estimates defined by (3.3) (nonparametric case) and (4.3) (parametric case). (The two-sample Q-estimate is not well defined for this purpose because μ_0 needs to be specified). If a bootstrap procedure is used, y^* is drawn from $F_{\tilde{p}_1} \cdot F_{\tilde{p}_2}$ (nonparametric case) or from $F_{\check{\theta}_c}$ (parametric case). Unfortunately the computation of the constrained nonparametric estimator (Section 3) is rather complex and requires some manual tuning of the smooth parameters δ_i (Section 6). This makes the bootstrap unrealistic. On the contrary, the parametric case is straightforward; in addition, asymptotic approximations based on influence function estimates of variance are also available (Appendix 5).

REMARK. The parametric procedures described above do not require the same model to be assumed for all samples.

6 Examples and empirical results

In this section, we give examples of two-samples tests, the most frequent case found in practice. We llustrate major differences between robust and non-robust tests, as well as the computation of confidence intervals when the hypothesis is accepted (i.e., under the null model). We start with an example using real data; then, we explore the most important findings suggested by this example using simulated data.

Estimators and tests. The truncated mean (TM; Marazzi and Ruffieux, 1999) is a very robust and efficient estimator. Therefore, we expect that the bootstrap tests based on TM are robust and powerful. Since this estimator depends only slightly on the model, we use it to illustrate both the parametric and the nonparametric simulation schemes.

We compare bootstrap tests based on TM with three classical tests: the pooled t-test, the Cressie and Whitford (1986) test based on a skewness adjusted statistic T_2^*, and the mean ratio test for Gamma distributions of Shiue et al. (1988). In addition we consider simple tests based on the trimmed mean, i.e., the test of Guo and Luh (2000; abbreviated 7.5GL) that uses an invertible transformation to correct skewness and then employs the symmetric β-trimmed mean to correct heavy tails and variance heterogeneity (we set $\beta = 7.5\%$), as well as a bootstrap test based on the asymmetric 7.5%-trimmed mean (7.5T) and semiparametric resampling (7.5% trimming is nearly optimal for the data sets described below). For all bootstrap tests, influence function estimates of variance are used to compute the statistic t and the transformation $h(\cdot) = \log(\cdot)$ is used to stabilize the variance of $d(F_{n,y^*})$; simulations are based on 1000 replications.

The notation TM/C will be used to indicate a parametric null model based on TM and Criterion C, ML/Q for a null model based on ML and Criterion Q, etc.

Data sets. Three data sets, "Real data", A, and B are provided in Appendix 6. The real data are described below. Data set A is made of two samples of size 50: the first 47 Y_1-values are drawn from a Gamma distribution with $\sigma = 1$ and $\alpha = 5$, and the last 3 values from a contaminating uniform distribution on $[20, 70]$. The 50 Y_2-values are Gamma distributed with $\sigma = 5$ and $\alpha = 1$. Thus, the means of the uncontaminated populations are both equal to 5. Data set B consists of two samples of size 50. The first 47 Y_1-values are Gamma distributed with $\sigma = 1$ and $\alpha = 5$, and the last 3 are outliers from a uniform distribution on $[20, 70]$. The 50 Y_2-values are Gamma distributed with $\sigma = 7.5$ and $\alpha = 1$. Thus, the uncontaminated means are 5 and 7.5. The contamination fraction is 6% for both A and B.

Tuning. We use a special kind of truncated mean defined as the "TD-estimator" in Marazzi and Ruffieux (1999). The initial model fit is based on the symmetric β-trimmed mean and the γ-trimmed absolute deviation with $\beta = \gamma = 0.4$; thus, the breakdown point is 40%. The trimming points are chosen with the help of this model so that the TM is Fisher consistent and has an asymptotic relative efficiency of 0.8 with respect to the ML-estimator.

The smooth parameters δ_i (Section 3) were first set to 0.2 in all examples. However, some irregularities in the influences of $\bar{\hat{\mu}}$ were observed with the simulated data from Gamma($\alpha = 5, \sigma = 1$). In those cases, δ_i was set to 0.4 for all i. For the data set A, the choice was assessed as follows. First, the empirical cdf-s F_{n_j,y_j} ($j = 1, 2$) were compared to their smooths $\hat{F}_{n_j,y_j} = \sum_i (1/n_{ji}) G^{\delta_i}_{y_{ji}}$. Figure 1, panels (a) and (b) show an excellent agreement. Then the estimates $\hat{\mu}(F_{n_1,y_1}) = 4.87$, $\hat{\mu}(F_{n_2,y_2}) = 4.61$, $\bar{\hat{\mu}}(F_{n_1,y_1}) = 4.93$, and $\bar{\hat{\mu}}(F_{n_2,y_2}) = 4.61$ were obtained, showing that the agreement between the original and the smoothed estimates was satisfactory. Finally, the influence functions $IF(y, \bar{\hat{\mu}}, F_{n_j,y_j})$ and $IF(y, \hat{\mu}, \bar{F}_{n_j,y_j})$ were compared (Figure 1, panels (c) and (d)). Again, the agreement was excellent. A similar assessment was done for all data sets.

Figure 1

Two-sample problem with real data. Figure 2, panel (a), shows the histogram of the lengths of stay (LOS, in days) for 315 patients hospitalized in Belgium (BE) during 1988 for certain "disorders of the nervous system"; panel (b) shows the LOS histogram for 32 patients hospitalized during the same year in Switzerland (CH) for the same kind of illness.

Figure 2

LOS is an important indicator of (and substitute for) hospital costs that are not yet easily available (e.g., Lave and Leinhardt, 1976). LOS "averages" of medically homogeneous groups of patients are routinely used as a basis for hospital resource allocation. (Around 500 groups and several hundreds of patients are common.) Unfortunately, LOS distributions are asymmetric and contain outliers whose value and frequency fluctuate from sample to sample, e.g., from year to year; this makes the most common measure of average, the arithmetic mean, very unstable and inappropriate. Therefore, outliers are usually removed according to various rules (Beguin et al., 1991) and the means of the remaining stays are computed. On the basis of these robust means, resources are allocated prospectively to the expected regular cases in each group; exceptionally extreme cases are carefully inspected and reimboursed retrospectively. Clearly, comparisons of robust means among different hospitals or over different periods of time are of great interest.

For the data of this example, the arithmetic means are 7.9 (BE) and 25.5 (CH) days. Both distributions contain outliers (not shown on the histograms, which are truncated at LOS=50 days); two very extreme stays - 374 and 198 days - belong to the Swiss sample. Removing these two stays reduces the Swiss mean to 8.1 days.

Table 1 gives one-sided P-values of the tests mentioned above for testing equality of the robust population means $\hat{\mu}_{BE}$ and $\hat{\mu}_{CH}$ against the alternative $\hat{\mu}_{BE} < \hat{\mu}_{CH}$. All the classical tests strongly reject the hypothesis but accept it when the two most extreme outliers are removed. The test decision is therefore determined by just two observations.

Shape and scale models are particularly appropriate for this kind of data: cost variations are commonly modeled with the help of scale transformations and it is reasonable not to assume identical shapes for hospitals with different management or medical policies. The two densities drawn with solid lines are Gamma distributions provided by the fitting procedure of the TM-estimator. Although stays are rounded to integer values, the two models approximately describe the majority of the data; the TM-estimates are 4.97 (BE) and 4.00 (CH) days, both with the complete and the reduced data set. Similar results were obtained using Weibull models. (For the purpose of comparison, the Gamma densities fitted by maximum likelihood are drawn with thin lines.) Both these models were used to compute tests based on TM and the P-values reported in Table 1 were automatically non-significant (at the usual 5% level) with the complete data set. They remained virtually unchanged when outliers were removed.

The tests based on trimming were also clearly non significant with the full data set; however, their P-values greatly changed when outliers were removed.

Table 1

Two-sample problems with simulated data. The previous example showed that the P-values of classical and trimmed mean based tests lack of robustness. We now illustrate the consequences of this deficiency with the help of simulated data with known "true" parameters.

Table 2 gives the P-values of the tests mentioned above for Data set A. With the full data set, all classical tests wrongly reject the hypothesis of identical means but they accept it when outliers are removed. All the TM tests correctly accept the hypothesis with large P-values, with and without outliers. Note that the same conclusion is reached even if the

Weibull model is erroneously assumed to compute TM. Finally, 7.5T and 7.5GL attain the correct decisions, but their P-values are disappointly low.

Table 2

Table 3 gives the P-values for Data set B. The classical tests do not detect the mean difference unless the outliers are removed, whereas the TM tests clearly reject the hypothesis of identical means, both with the full data set and when outliers are removed. Finally, the power of the tests based on trimmed means (that do not use any information about the shapes) is too low: 7.5T never detect the mean difference and 7.5GL is inconclusive.

Table 3

Confidence intervals. Table 4 reports symmetric 90%-percentile confidence intervals, as well as the asymptotic normal approximations (based on Appendix 5), for the common average LOS obtained with 1000 simulated values of the two-sample constrained estimates CML, ML/C, and TM/C. Note the excellent agreement between the CML and ML/C procedures – both in the Gamma and the Weibull cases – and the dramatic disagreement between classical and robust intervals.

Table 4

Table 5 reports similar intervals for the common mean ($\mu = 5$) of the two samples in data set A. As expected, the confidence intervals based on the ML estimator are slightly shorter than those based on TM in the absence of outliers. With the full data set, the procedures based on ML fail to cover the true common mean.

Table 5

Power. The power p_{TM} of the two-sample test based on TM was compared to the power p_R of the Shiue et al. (2000) ratio test and to the power p_{GL} of 7.5GL. In a first step, an approximation to the null distribution of the test statistic based on 1000 samples of size $n_1 = n_2 = 50$ — such that $Y_1 \sim \text{Gamma}(\alpha = 5, \sigma = 1)$, $Y_2 \sim \text{Gamma}(\alpha = 1, \sigma = 5)$ — were obtained; critical values $k_{.90}$ and $k_{.95}$ were determined as the quantiles $q = .90$ and $q = .95$ of this distribution. In a second step, 1000 values t^* based on samples of size $n_1 = n_2 = 50$ — such that $Y_1 \sim (1 - \epsilon)\text{Gamma}(\alpha = 5, \sigma = 1) + \epsilon\text{Uniform}[0, 50]$, $Y_2 \sim \text{Gamma}(\alpha = 1, \sigma)$ — were generated for $\sigma = 5, 6, 7, 8, 9$, $\epsilon = 0.00$, and $\epsilon = 0.06$, and the approximation $p_{TM}(\sigma) = \#\{t^* \geq k_q\}/1000$ was computed. On the basis of the same simulated samples we also computed $p_{GL}(\sigma)$ and $p_R(\sigma)$ (the critical values for the ratio test were taken from Table 1 of Shiue et al., 2000). Finally, we computed the asymptotic approximation $ap_{TM}(\sigma)$ to $p_{TM}(\sigma)$ given in Appendix 1. The results are summarized in Table 6 and point out that p_{TM} is very close to p_R in the absence of outliers but much better under contamination. The quality of the asymptotic approximation is fair. The power of 7.5GL is poor both with $\epsilon = 0$ and $\epsilon = 0.06$.

Table 6

REMARK 1. Figure 3 shows the bootstrap null distributions of the tests based on TM and the Weibull model; each panel is a qq-plot comparing the distribution obtained with the complete and the reduced data sets. We note that the parametric procedures were more stable than the resampling procedures. This "lack of robustness" of the classical bootstrap can be explained by the "exceeding" frequency of outliers in many simulated samples, a feature that affects the tails of the null distributions, which has also been noted by other authors, e.g., Salibian (2000) who proposes a remedy.

Figure 3

REMARK 2. Figure 4, panel (a), shows the influences of the two samples in the data set A on the smoothed TM-estimator ($\bar{\hat{\mu}}$, Section 3); these influences determine the exponentially tilted null model probabilities displayed in panel (b). Panel (c) shows smooth densities corresponding to the two null probability distributions and, for comparison, the parametric null model is represented in panel (d). Due to the redescending shape of the IF and the negative value of λ, extreme observations in sample 1 are more likely resampled than many smaller observations. This surprising feature of the tilted null model is a source of instability in the classical bootstrap distribution.

Figure 4

REMARK 3. Figure 4, panel (c) shows smooth densities corresponding to the exponentially tilted distributions $\tilde{p}_{1i}$ and $\tilde{p}_{2i}$ for data set A. Dotted densities have been obtained using the iterative algorithm derived from Proposition 1 for both samples and solving the constraint $\hat{\mu}(F_{\tilde{p}_1(\lambda)}) = \hat{\mu}(F_{\tilde{p}_2(\lambda)})$. Solid lines have been obtained using the linearized constraint. The approximation is very satisfactory.

REMARK 4. The Wilcoxon test has not been included in the examples because it is not appropriate when the shapes are different. For the real data set, it accepts the hypothesis of identical populations with $P = 0.68$ (full data) and $P = 0.45$ (outliers removed). For data set A, it wrongly rejects this hypothesis in favor of the alternative that Y_1 is stochastically larger that Y_2 with $P = 0.01$ (full data) and $P = 0.03$ (outliers removed). For data set B, it wrongly accepts the hypothesis both with the full ($P = 0.50$) and the cleaned ($P = 0.81$) data sets.

7 Conclusions

Available robust estimators for the means of popular asymmetric models have been used to develop bootstrap tests for comparing robust means of asymmetric distributions with unequal shapes. In addition, four bootstrap schemes were explored, with the help of real and simulated data. We summarize our main conclusions.

1. In the presence of outliers, the robust tests can reach the correct decision, whereas the usual tests are likely to result in both kinds of testing errors. Unfortunately, tests based on simple robust estimators - such as the trimmed mean - can also fail due to insufficient robustness of the type I error probability (when the trimming proportion is low) or insufficient power (when the trimming proportion is high). On the contrary, robust parametric estimators with a high degree of robustness - such as the truncated mean - recognize different shapes and provide reliable tests under realistic contamination. In the absence of outliers, the parametric bootstrap tests based on efficient robust estimators can be almost as powerful as their classical counterparts.

2. Minimization of the Kullback-Leibler disparity between the constrained model and the available unconstrained estimates provides numerically satisfactory and computationally convenient constrained estimates that can be used to define parametric null models and to compute confidence intervals for common means when the hypothesis is accepted. Programming of more conventional but computationally awkward constrained robust estimates (e.g., constrained M-estimates) does not appear to be necessary.

3. Nonparametric simulation from (modified versions of) the empirical distribution provides less stable null distributions (with respect to the presence of outliers) than simulation from parametric models. On the other hand, it can be argued that parametric simulation tends to underestimate real fluctuations, since it ignores contamination. Both parametric and nonparametric simulation schemes should therefore be used as complementary, rather than exclusive, tools of inference.

The procedures described in this paper have been illustrated using the truncated mean for the Gamma and the Weibull models but extensions to more complex models, where shape is described with the help of several parameters (e.g. the generalized Gamma and F distributions; Kalbfleisch and Prentice, 1980), will be straightforward as soon as the corresponding robust estimates are available. S-plus functions for the truncated mean and for optimally constrained one- and two-sample ML-, TM-, and M-estimators for the Log-normal, Gamma, and Weibull models can be found at www.hospvd.ch/iumsp (download buttom).

Appendix

1. Asymptotic power

In the one-sample case, let θ_0 be the null model parameter vector such that $\mu(\theta_0) = \mu_0$. Let η be a parameter change, $\theta_n = \theta_0 + \eta/n$, and $c_{n,1-\gamma}$ be the critical value for testing θ_0 against θ_n at the γ level. Denote by $P_\theta(A)$ the probability of A when $Y \sim F_\theta$. A standard derivation (e.g., Staudte and Sheather, 1990, p.159) shows that the asymptotic power function is $\pi(\eta) = \lim_{n\to\infty} P_{\theta_n}(t > c_{n,1-\gamma}) = \Phi(z_\gamma - e_t^{\mathrm{T}}\eta)$, where z_γ denotes the standard normal γ-quantile, $e_t = V(d, F_{\theta_0})^{-1/2} \cdot (\partial d(F_\theta)/\partial\theta)_{\theta=\theta_0}$ is the test efficacy, $d = h(\hat{\mu}(F)) - h(\mu_0)$, and $V(d, F) = h'(\hat{\mu}(F))^2 V(\hat{\mu}, F)$. Note that e_t and $\pi(\eta)$ do not depend on h.

In the two-sample case, the null model parameter vector is $\theta_0 = (\alpha_{10}, \sigma_{10}, \alpha_{20}, \sigma_{20})$ and the parameter change is $\eta = (\eta_1, \eta_2, \eta_3, \eta_4)$. The power and efficacy formulae still hold with $\theta_n = \theta_0 + (\eta_1/n_1, \eta_2/n_1, \eta_3/n_2, \eta_4/n_2)$, defining d as in Section 2.2, and letting n_1 and n_2 tend to ∞.

2. Proof of Proposition 1

We consider the objective functional $L(F_p) = D(F_p) - \lambda(\hat{\mu}(F_p) - \mu_0)$ defined over $\mathcal{P}$, where $D(F_p) = \sum p_i \ln(p_i)$ and λ is a Lagrange multiplier. According to a variational argument, if a distribution $F_{\tilde{p}}$ is locally optimal over $\mathcal{P}$ (for a fixed λ and $0 < \tilde{p}_i < 1$), then $IF(y_j, L, F_{\tilde{p}}) = 0$ for $j = 1, \ldots, n$. We obtain $IF(y_j, D, F_{\tilde{p}}) = \ln(\tilde{p}_j) - D(F_{\tilde{p}})$ and, therefore, the conditions

$$\ln(\tilde{p}_j) = \lambda I_j(F_{\tilde{p}}) + D(F_{\tilde{p}}), \quad j = 1, \ldots, n,$$

where $I_j(F) = IF(y_j, \hat{\mu}, F)$. Thus, $\tilde{p}_j = \exp(\lambda I_j(F_{\tilde{p}}))\exp(D(F_{\tilde{p}}))$ and, since $\sum \tilde{p}_i = 1$, $\exp(D(F_{\tilde{p}})) = [\sum \exp(\lambda I_i(F_{\tilde{p}}))]^{-1}$. The multiplier λ must obviously satisfy $\hat{\mu}(F_{\tilde{p}}) = \mu_0$. The two- and multi-sample cases can be treated in similar ways.

3. Proof of Proposition 2

Let $f_\theta(y)$ denote the density of a member of the exponential family of distributions (e.g., Lindsey, 1996, pp. 26–28) in the natural parametrization. The log-likelihood function is $\ell(F_\theta; \Delta_y) = \ln[f_\theta(y)] = \theta^{\mathrm{T}} t(y) - \kappa(\theta) - t_0(y)$, where θ is a k-dimensional parameter vector, $t(y)$ is the sufficient statistic for θ, $\kappa(\theta)$ is the negative logarithm of the normalizing constant, and $t_0(y)$ a given function. The score function is $s(\theta; y) = t(y) - \dot{\kappa}(\theta)$, where $\dot{\kappa}(\theta) = \partial\kappa/\partial\theta(\theta)$ denotes the gradient of κ with respect to θ. The ML-estimate $\hat{\theta}$ satisfies $\mathrm{ave}\{t(y)\} = \dot{\kappa}(\theta)$, where $\mathrm{ave}\{t(y)\}$ is an abbreviation for $\int t(y) dF_{n,y}(y)$. Therefore, $\int \mathrm{ave}\{t(y)\} dF_{\hat{\theta}}(y) = \dot{\kappa}(\hat{\theta}) = \mathrm{ave}\{t(y)\}$ and $\ell(F_\theta, F_{\hat{\theta}}) = \theta^{\mathrm{T}} \mathrm{ave}\{t(y)\} - \kappa(\theta) - \int t_0(y) dF_{\hat{\theta}}(y)$. It follows that the functions $\ell(F_\theta, F_{\hat{\theta}})$ and $\ell(F_\theta, F_{n,y}) = \theta^{\mathrm{T}} \mathrm{ave}\{t(y)\} - \kappa(\theta) - \mathrm{ave}\{t_0\}$, differing only in the constant term, have the same extremes under any constraint.

4. C-estimators of some common models

We consider the Weibull, the Gamma, and the first and second kind Pareto distributions. For each one of these models we give the log-likelihoods $\ell(F_\theta, \Delta_y)$ and $\ell(F_\theta, F_{\hat{\theta}})$ and indicate how the C-estimator can be computed. In the Weibull case, α and σ are replaced by more convenient parameters.

The Weibull distribution. We consider the distribution $G_{(\tau,\varsigma)}$ of $X = \ln(Y)$ with location τ and scale ς. The shape parameter of the Weibull distribution of Y is $\alpha = 1/\varsigma$ and the scale is $\sigma = \exp(\tau)$. With $z = (x-\tau)/\varsigma$, we obtain

$$\ell(G_{(\tau,\varsigma)}, \Delta_x) = z - \exp(z) - \ln(\varsigma),$$

$$\ell(G_{(\tau,\varsigma)}, G_{(\hat{\tau},\hat{\varsigma})}) = (-\hat{\varsigma}C_e + \hat{\tau})/\varsigma - \tau/\varsigma - \exp((\hat{\tau}-\tau)/\varsigma)\Gamma(\hat{\varsigma}/\varsigma + 1) - \ln(\varsigma),$$

where C_e denotes the Euler constant ($C_e = 0.577215665\ldots$). The function $\ell(G_{(\tau,\varsigma)}, G_{(\hat{\tau},\hat{\varsigma})})$ has to be numerically maximized under the constraint $\exp(\tau)\Gamma(1+\varsigma) = \mu_0$.

The Gamma distribution. For the Gamma distribution $F_{(\alpha,\tau)}(y)$ with shape α and scale $\sigma = \exp(\tau)$ we obtain

$$\ell(F_{(\alpha,\tau)}, \Delta_y) = (\alpha-1)\ln(z) - z - \tau - \ln\Gamma(\alpha),$$

$$\ell(F_{(\alpha,\tau)}, F_{(\hat{\alpha},\hat{\tau})}) = (\alpha-1)[\dot{\Gamma}(\hat{\alpha}) + \hat{\tau}] - \hat{\alpha}\exp(\hat{\tau}-\tau) - \alpha\tau - \ln(\Gamma(\alpha)),$$

with $z = y/\sigma$ and $\dot{\Gamma}(\alpha) = (d/d\alpha)\ln\Gamma(\alpha)$. The function $\ell(F_{(\alpha,\tau)}, F_{(\hat{\alpha},\hat{\tau})})$ must be numerically maximized subject to $\exp(\tau)\alpha = \mu_0$.

The Pareto distributions. For the Pareto distribution of the first kind, with density $f_{(\alpha,\sigma)}(y) = \alpha\sigma^\alpha/y^{\alpha+1}$ ($\sigma > 0$, $\alpha > 0$, $y \geq \sigma$, see Johnson et al. (1994, p.574)). We have

$$\ell(F_{(\alpha,\sigma)}, \Delta_y) = \alpha\ln(\sigma) + \ln(\alpha) - (\alpha+1)\ln(y),$$

$$\ell(F_{(\alpha,\sigma)}, F_{(\hat{\alpha},\hat{\sigma})}) = \alpha\ln(\sigma) + \ln(\alpha) - (\alpha+1)[\ln(\hat{\sigma}) + 1/\hat{\alpha}].$$

The constraint is $\alpha\sigma/(\alpha-1) = \mu_0$ (provided that $\alpha > 1$). For the Pareto distribution of the second kind, with density $f_{(\alpha,\sigma)}(y) = (\alpha/\sigma)(y/\sigma+1)^{-(\alpha+1)}$ ($\sigma > 0$, $\alpha > 0$, $y \geq 0$, see Johnson et al. (1994, p.575)), we obtain

$$\ell(F_{(\alpha,\sigma)}, \Delta_y) = \ln(\alpha) - \ln(\sigma) - (\alpha+1)\ln(x/\sigma+1),$$

$$\ell(F_{(\alpha,\sigma)}, F_{(\hat{\alpha},\hat{\sigma})}) = \ln(\alpha/\sigma) - (\alpha+1)\hat{\alpha}\cdot I,$$

where

$$I = \int_1^\infty \gamma\ln(z)[\gamma(z-1)+1]^{-\hat{\alpha}-1}dz,$$

and $\gamma = \sigma/\hat{\sigma}$. The constraint for the C-estimator is $\sigma/(\alpha-1) = \mu_0$ (provided that $\alpha > 1$).

5. Influence function of the C-estimator

We consider the two-sample C-estimator as an example, the extention to the multi-sample and the restriction to the one-sample cases being straightforward. By definition, the IF of $\check{\theta}_c = (\check{\theta}_{1c}, \check{\theta}_{2c})$ is given by

$$IF(z; \check{\theta}_c, F) = \lim_{t \downarrow 0} \frac{\check{\theta}_c((1-t)F + t\Delta_z) - \check{\theta}_c(F)}{t}$$

where $z = (y_1, y_2)$ denotes a value of $Z = (Y_1, Y_2)$, F the cdf of Z and Δ_z the cdf of a point mass at z. To find a more explicit expression, it is convenient to parametrize the models with $\vartheta = (\mu_0, \alpha_1, \alpha_2)$, where $\mu_0 = \mu((\alpha_1, \sigma_1)) = \mu((\alpha_2, \sigma_2))$ is the common mean and to assume that $n_j/(n_1+n_2)$ tends to a constant w_j when both sample sizes increase. As a functional, the C-estimator $\check{\vartheta}_c = (\check{\mu}_{0c}, \check{\alpha}_{1c}, \check{\alpha}_{2c})$ based on unconstrained estimators $\hat{\theta}_1 = (\hat{\alpha}_1, \hat{\sigma}_1)$, $\hat{\theta}_2 = (\hat{\alpha}_2, \hat{\sigma}_2)$ of θ_1 and θ_2 is then defined by

$$\check{\vartheta}_c(F) = \text{argmin}_{\mu_0, \alpha_1, \alpha_2} D_\ell(\mu_0, \alpha_1, \alpha_2; \hat{\theta}_1(F), \hat{\theta}_2(F)),$$

where

$$D_\ell(\mu_0, \alpha_1, \alpha_2; \hat{\theta}_1, \hat{\theta}_2) = w_1 \cdot d_\ell(F_{\alpha_1, \mu_0/\xi(\alpha_1)}, F_{\hat{\alpha}_1, \hat{\sigma}_1}) + w_2 \cdot d_\ell(F_{\alpha_2, \mu_0/\xi(\alpha_2)}, F_{\hat{\alpha}_2, \hat{\sigma}_2}),$$

or by the system

$$\mathcal{D}(\check{\mu}_{0c}, \check{\alpha}_{1c}, \check{\alpha}_{2c}; \hat{\theta}_1, \hat{\theta}_2) = 0,$$

where $\mathcal{D} = (\partial D_\ell/\partial \mu_0, \partial D_\ell/\partial \alpha_1, \partial D_\ell/\partial \alpha_2,)^{\mathrm{T}}$ (a column vector). Differentiating the system, one obtains

$$IF(z; \check{\vartheta}_c, F) = -\left(\frac{\partial}{\partial \mu_0}\mathcal{D}, \frac{\partial}{\partial \alpha_1}\mathcal{D}, \frac{\partial}{\partial \alpha_2}\mathcal{D}\right)^{-1} \left(\frac{\partial}{\partial \hat{\alpha}_1}\mathcal{D}, \frac{\partial}{\partial \hat{\sigma}_1}\mathcal{D}, \frac{\partial}{\partial \hat{\alpha}_2}\mathcal{D}, \frac{\partial}{\partial \hat{\sigma}_2}\mathcal{D}\right) IF(z; (\hat{\theta}_1, \hat{\theta}_2), F),$$

where

$$IF(z; (\hat{\theta}_1, \hat{\theta}_2), F) = (IF(y_1; \hat{\alpha}_1, F), IF(y_1; \hat{\sigma}_1, F), IF(y_2; \hat{\alpha}_2, F), IF(y_2; \hat{\sigma}_2, F))^{\mathrm{T}},$$

is the vector influence function of the unconstrained estimator. Thus, the asymptotic covariance matrix of $\check{\vartheta}_c$, $V(\check{\vartheta}_c, F) = \int IF(z; \check{\vartheta}_c, F) IF(z; \check{\vartheta}_c, F)^{\mathrm{T}} dF(z)$, is directly obtained from the available variances and covariances of $(\hat{\alpha}_1, \hat{\sigma}_1)$ and $(\hat{\alpha}_2, \hat{\sigma}_2)$.

6. Data sets

Real data

Frequency distribution (Freq.) of lengths of stay (LOS) in days of patients hospitalized in Belgium and Switzerland during 1988 for certain "disorders of the nervous system". Source: a data base described in Marazzi et al., 1998.

	Belgium																		
LOS	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	19	21
Freq.	51	59	34	32	32	9	6	12	9	11	11	8	4	3	0	4	6	2	1
LOS	22	26	28	29	32	33	34	35	36	37	40	43	44	49	60	68	81	96	134
Freq.	1	2	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	1
	Switzerland																		
LOS	1	2	3	4	5	6	7	8	9	16	115	198	374						
Freq.	2	6	5	5	4	2	2	1	1	1	1	1	1						

Data set A

$$(y_{1i}) = (6.40, 4.08, 8.12, 2.30, 2.09, 5.32, 3.43, 4.60, 5.43, 3.20, 4.99, 5.69, 4.46, 9.96, 4.67, 4.39, 2.88, 3.49, 2.23, 3.14, 6.15, 7.82, 4.95, 6.12, 8.04, 3.34, 4.20, 5.67, 8.52, 3.16, 3.72, 2.65, 4.75, 4.47, 4.74, 4.71, 8.80, 4.18, 5.79, 2.95, 11.76, 3.05, 4.59, 2.46, 6.69, 4.53, 7.20, 39.10, 65.88, 43.78),$$

$$(y_{2i}) = (6.14, 5.76, 7.16, 6.52, 0.52, 3.92, 5.01, 3.92, 0.68, 12.60, 4.65, 0.91, 1.43, 1.52, 1.44, 1.55, 2.25, 0.22, 1.08, 0.08, 5.08, 8.92, 3.67, 4.65, 1.38, 5.05, 4.28, 3.66, 7.47, 14.65, 1.03, 1.65, 4.75, 15.06, 21.15, 4.05, 11.15, 2.57, 2.66, 2.62, 4.10, 9.61, 0.30, 6.96, 3.34, 0.63, 2.73, 3.26, 3.35, 6.51).$$

Data set B

$$(y_{1i}) = (2.97, 7.08, 2.92, 5.49, 4.41, 1.45, 3.47, 6.28, 6.69, 7.31, 4.07, 6.04, 4.35, 3.67, 3.43, 5.98, 10.67, 4.44, 6.65, 10.41, 6.41, 5.66, 4.36, 2.93, 6.38, 9.20, 6.97, 8.37, 3.81, 3.55, 5.42, 3.29, 5.18, 4.20, 6.02, 3.39, 4.37, 5.76, 3.53, 2.97, 5.52, 4.40, 6.21, 3.88, 10.14, 8.35, 1.38, 63.95, 32.74, 36.86),$$

$$(y_{2i}) = (7.44, 13.79, 1.80, 0.58, 2.52, 3.76, 2.64, 14.82, 6.21, 5.13, 5.01, 1.23, 18.91, 0.90, 18.77, 9.44, 1.95, 3.53, 31.36, 3.16, 0.96, 5.17, 6.30, 3.55, 11.40, 8.84, 1.23, 2.18, 3.64, 7.87, 14.63, 10.80, 4.88, 19.66, 3.67, 18.72, 18.35, 1.76, 1.48, 1.41, 10.90, 9.23, 29.54, 4.80, 4.13, 17.04, 6.52, 5.07, 7.35, 1.78).$$

References

Beguin, C., Closon, M.C., Roger, F.H., 1991. Advances in DRGs data pooling in Europe: results from the Hoscom project in relation to outliers. Paper presented at the second EURODRG Workshop, Dublin, 24-25 April 1991.

Cressie N.A., Whitford H.J., 1986. How to Use the Two Sample t-Test. Biometric. Journal, 28(2) 131-148.

Davison A.C., Hinkley D.V., 1997. Bootstrap Methods and their Applications. Cambridge University Press, Cambridge.

Efron B., Tibshirani R.J., 1993. An Introduction to the Bootstrap. Chapman & Hall, New York.

Guo J-H., Luh W-M., 2000. An invertible transformation two-sample trimmed t-statistic under heterogeneity and nonnormality. Statistics & Probability Letters , 49 1-7.

Hampel F.R., Ronchetti E.M., Rousseeuw P.J., Stahel W.A., 1986. Robust Statistics: An Approach Based on the Influence Function. Wiley, New York.

Hettmansperger T.P., McKean J.W., 1998. Robust Nonparametric Statistical Methods. Arnold, London.

Huber, P.J., 1981. Robust Statistics. Wiley, New York.

Kalbfleisch J.D., Prentice R.L., 1980. The statistical analysis of failure time data. Wiley, New York.

Johnson L., Kotz S., Balakrishnan N., 1994. Continuous Univariate Distributions, Volume 1. Wiley, New York.

Lave and Leinhardt, 1976. The cost and length of a hospital stay. Inquiry, 13 327-343.

Lindsey, J.K., 1996. Parametric Statistical Inference. Oxford Science Publications, Clarendon Press, Oxford.

Liu R.Y., Singh K., 1997. Notions of Limiting P Values Based on Data Depth and Bootstrap. Journal of the American Statistical Association, 92(437) 266-277.

Marazzi A., Ruffieux C., 1999. The Truncated Mean of an Asymmetric Distribution. Computational Statistics and Data Analysis, 32(1) 79-100.

Marazzi A., Paccaud F., Ruffieux C., Beguin C., 1998. Fitting the Distributions of Length of Stay by Parametric Models. Medical Care 36(6) 915-927.

Salibian-Barrera, M., 2000. Contributions to the Theory of Robust Inference. Unpublished Ph.D. Thesis. (University of British Columbia, Department of Statistics, Vancouver, BC.
Available at ftp://ftp.stat.ubc.ca/pub/matias/Thesis/)

Shiue W.K., Bain L.J., Engelhardt M., 1988. Test of Equal Gamma-Distribution Means With Unknown and Unequal Shape Parameters. Technometrics 30(2) 169-174.

Staudte R.G., Sheather S.J., 1990. Robust estimation and testing, Wiley, New York.

Tiku, M.L., 1982. Robust statistics for testing equality of means or variances. Comm. Statist. 22 2543-2558.

Victoria-Feser M.P., Ronchetti E., 1994. Robust Methods for Personal-Income Distribution Models. The Canadian Journal of Statistics, 22 247-258.

Victoria-Feser M.P., Ronchetti E., 1997. Robust Estimation for Grouped Data. Journal of the American Statistical Association, 92(437) 333-340.

Welch, B.L., 1951. On the comparison of several mean values: an alternative apprach. Biometrika, 38 330-336.

Wilcox R.R., 1990. Comparing the means of two independent groups. Biometric. Journal, 32(7) 771-780.

Wilcox R.R., Keselman H.J., Kowalchuck R.K., 1998. Can tests for treatment group equality be improved ? The bootstrap and trimmed means conjecture. British Journal of Mathematical and Statistical Psychology, 51 123-134.

Table 1. One-sided P-values of various two sample tests.
Real data: $\mathcal{H}_0 : \hat{\mu}_{BE} = \hat{\mu}_{CH}$ against $\mathcal{H}_1 : \hat{\mu}_{BE} < \hat{\mu}_{CH}$.

	Full real data set			2 outliers removed		
Test	Gamma	Weibull	Other	Gamma	Weibull	Other
pooled t			$< .001$			0.47
Cressie & Whitfort T_2^*			0.04			0.42
Shiue et al. ratio	$< .005$			$> .25$		
7.5GL			0.27			0.95
7.5T/Semipar.			0.66			0.96
TM/Q	0.87	0.87		0.90	0.88	
TM/C	0.88	0.87		0.91	0.88	
TM/Semipar.	0.95	0.94		0.95	0.92	
TM/Exp. tilting	0.94	0.95		0.94	0.93	

Table 2. One-sided P-values of various two sample tests.
Data set A: $\mathcal{H}_0 : \hat{\mu}_{y_1} = \hat{\mu}_{y_2}$ against $\mathcal{H}_1 : \hat{\mu}_{y_1} > \hat{\mu}_{y_2}$.

	Full data set A			3 outliers removed		
Test	Gamma	Weibull	Other	Gamma	Weibull	Other
pooled t			0.04			0.31
Cressie & Whitfort T_2^*			0.01			0.32
Shiue et al. ratio	$< .005$			$> .25$		
7.5GL			0.07			0.15
7.5T/Semipar.			0.06			0.09
TM/Q	0.42	0.38		0.39	0.42	
TM/C	0.41	0.38		0.39	0.42	
TM/Semipar.	0.37	0.43		0.39	0.50	
TM/Exp. tilting	0.39	0.43		0.37	0.50	

Table 3. One-sided P-values of various two sample tests.
Data set B: $\mathcal{H}_0 : \hat{\mu}_{y_1} = \hat{\mu}_{y_2}$ against $\mathcal{H}_1 : \hat{\mu}_{y_1} < \hat{\mu}_{y_2}$.

	Full data set B			3 outliers removed		
Test	Gamma	Weibull	Other	Gamma	Weibull	Other
pooled t			0.44			0.02
Cressie & Whitfort T_2^*			0.46			0.01
Shiue et al. ratio	$> .25$			$< .01$		
7.5GL			0.06			0.05
7.5T/Semipar.			0.10			0.14
TM/Q	0.01	0.01		0.01	0.01	
TM/C	0.01	0.01		0.02	0.01	
TM/Semipar.	0.01	0.01		0.02	0.01	
TM/Exp. tilting	0.02	0.02		0.03	0.01	

Table 4. Percentile 90%-confidence intervals for a common average length of stay based on 1000 simulated samples and asymptotic normal approximation.

Constrained estimate	Method	Model	
		Gamma	Weibull
CML	param. boot.	(7.63, 9.23)	(7.26, 8.91)
ML/C	param. boot.	(7.67, 9.23)	(7.28, 8.92)
TM/C	param. boot.	(4.54, 5.45)	(4.57, 5.57)
ML/C	normal appr.	(7.67, 9.13)	(7.30, 8.86)
TM/C	normal appr.	(4.59, 5.43)	(4.62, 5.48)

Table 5. Percentile 90%-confidence intervals for a common mean $\mu = 5$ in data set A based on 1000 simulated samples and asymptotic normal approximations.

Constrained estimate	Method	Data set	
		Full	3 outliers removed
ML/C	param. boot.	(5.42, 7.27)	(4.50, 5.39)
TM/C	param. boot.	(4.46, 5.50)	(4.36, 5.44)
ML/C	normal appr.	(5.66, 7.06)	(4.63, 5.24)
TM/C	normal appr.	(4.58, 5.36)	(4.49, 5.24)

Table 6. $p_{TM}(\sigma)$ is the simulated power function (for the levels 0.05 and 0.10) of the two-sample tests based on TM for $Y_1 \sim (1-\epsilon)\text{Gamma}(\alpha = 5, \sigma = 1) + \epsilon\text{Uniform}[0, 50]$, $Y_2 \sim \text{Gamma}(\alpha = 1, \sigma)$, $n_1 = n_2 = 50$, and $\sigma = 5, 6, 7, 8, 9$. $p_R(\sigma)$ and p_{GL} are the corresponding power functions of the test of Shiue et al. (1988) and the test of Guo and Luh (2000). $ap_{TM}(\sigma)$ is the asymptotic approximation for $p_{TM}(\sigma)$.

		$\varepsilon = 0$			$\varepsilon = 0.06$			
σ	level	$p_R(\sigma)$	$p_{GL}(\sigma)$	$p_{TM}(\sigma)$	$p_R(\sigma)$	$p_{GL}(\sigma)$	$p_{TM}(\sigma)$	$ap_{TM}(\sigma)$
5	0.10	0.10	0.02	0.09	0.01	0.01	0.11	0.10
	0.05	0.05	0.01	0.04	0.01	0.00	0.03	0.05
6	0.10	0.45	0.20	0.39	0.14	0.12	0.35	0.44
	0.05	0.31	0.11	0.29	0.08	0.05	0.23	0.30
7	0.10	0.78	0.53	0.76	0.37	0.38	0.62	0.83
	0.05	0.66	0.38	0.63	0.26	0.23	0.52	0.73
8	0.10	0.95	0.79	0.92	0.62	0.65	0.88	0.98
	0.05	0.91	0.69	0.85	0.50	0.52	0.82	0.96
9	0.10	0.99	0.95	0.97	0.84	0.86	0.98	1.00
	0.05	0.98	0.90	0.96	0.75	0.76	0.96	1.00

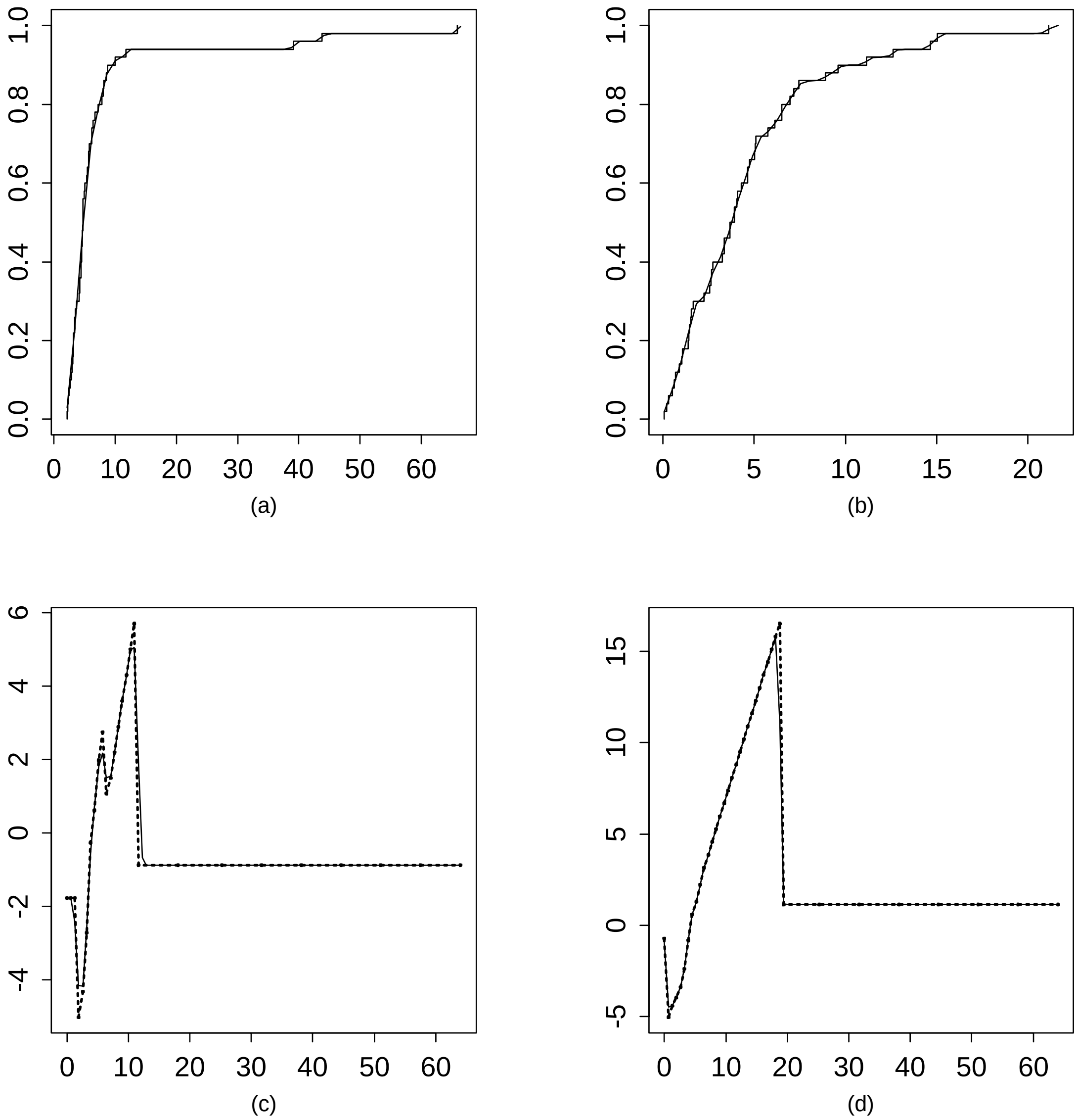

Figure 1. Smooths for Data set A.
Panel (a): Empirical cdf F_{n_1,y_1} and smooth $\bar{F}_{n_1,y_1}$.
Panel (b): Empirical cdf F_{n_2,y_2} and smooth $\bar{F}_{n_2,y_2}$.
Panel (c): Influence functions $IF(y, \bar{\hat{\mu}}, F_{n_1,y_1})$ (solid) and $IF(y, \hat{\mu}, \bar{F}_{n_1,y_1})$ (dotted).
Panel (d): Influence functions $IF(y, \bar{\hat{\mu}}, F_{n_2,y_2})$ (solid) and $IF(y, \hat{\mu}, \bar{F}_{n_2,y_2})$ (dotted).

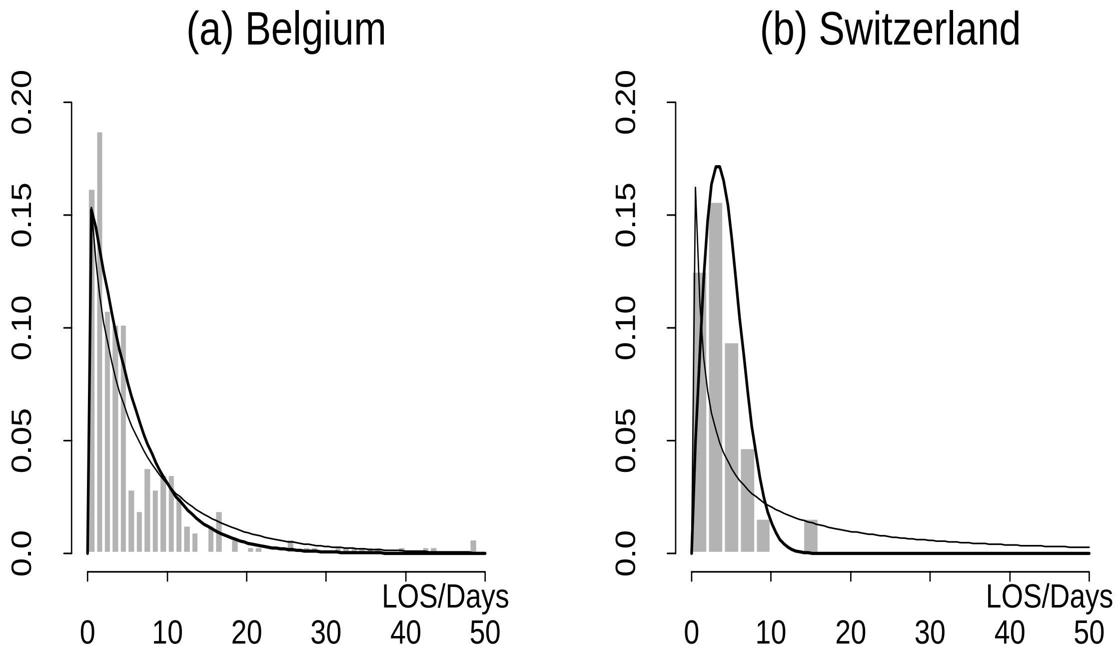

Figure 2. (a) Histogram of 315 lengths of stay (LOS) in days of patients hospitalized in Belgium during 1988 for certain "disorders of the nervous system". (b) Histogram of 32 LOS of patients hospitalized during the same year in Switzerland for the same kind of illness. Both histograms are truncated at LOS = 50. The densities of the Gamma distributions have been determined by means of the maximum likelihood (thin lines) and a robust (TM) estimates (bold lines).

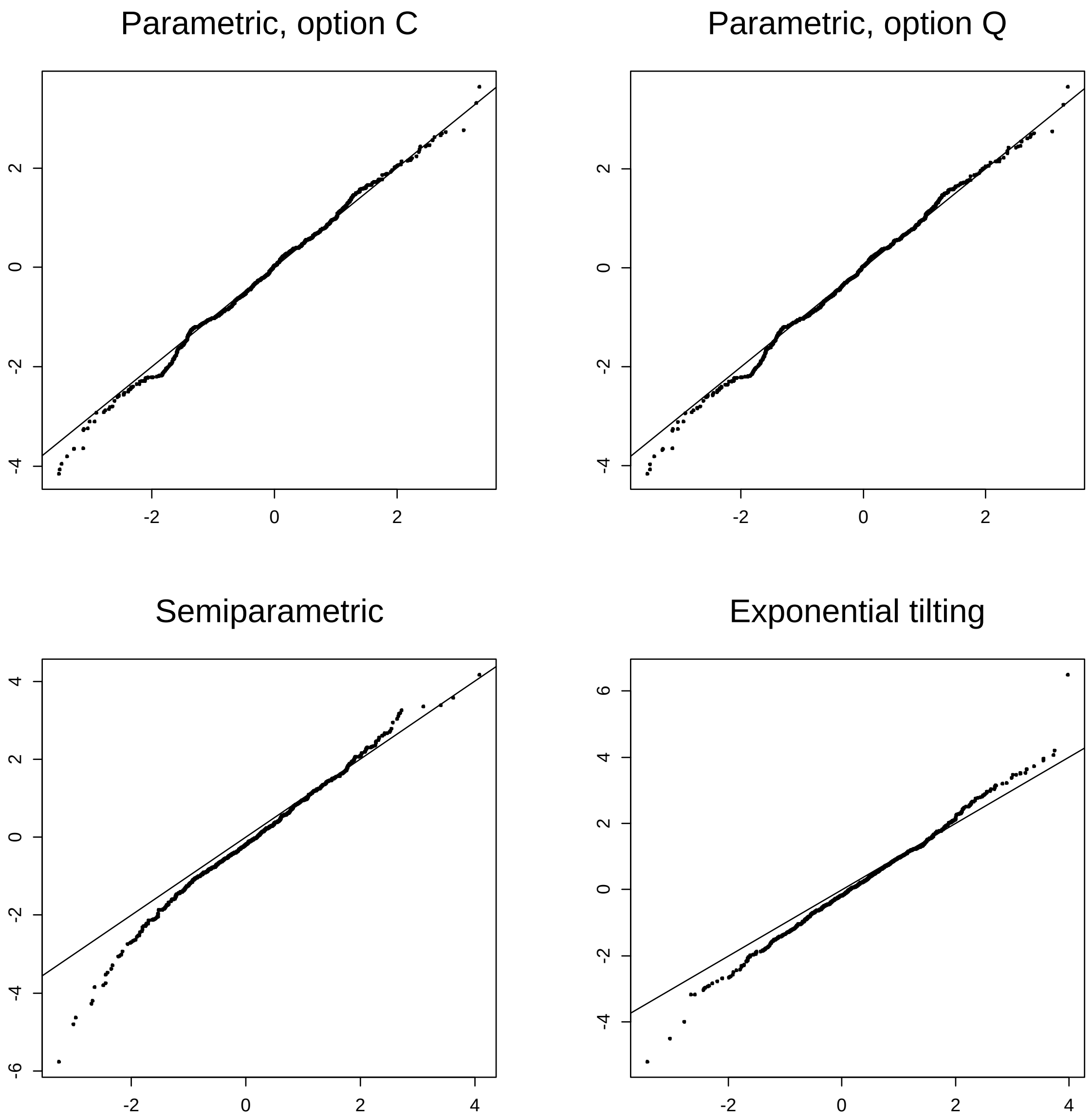

Figure 3. Bootstrap null distributions of robust tests based on TM (Weibull) and four null models. Each panel shows the qq-plot for comparing the null distributions obtained with the complete (horizontal axis) and the reduced (two outliers removed; vertical axis) real data sets.

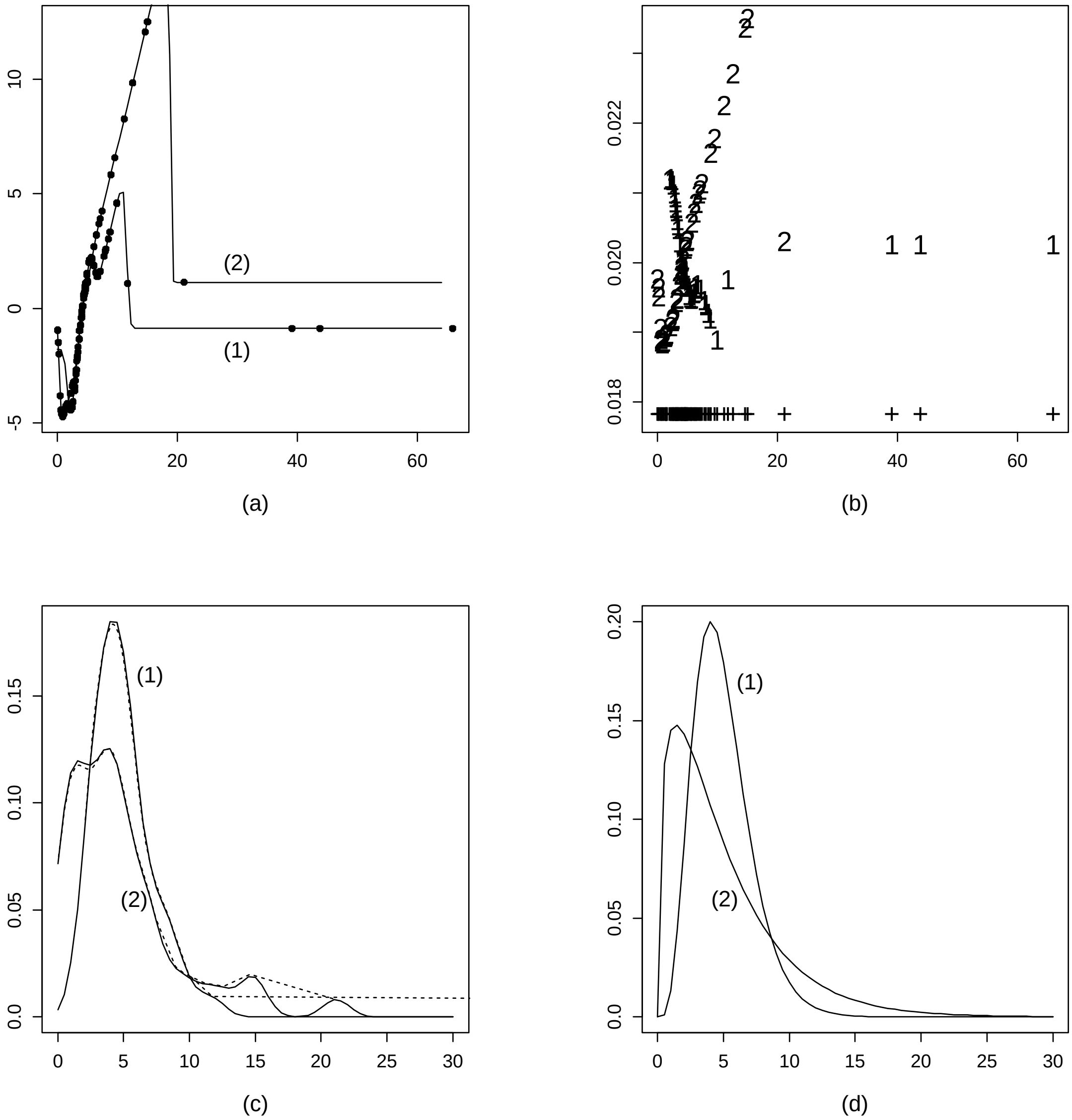

Figure 4. Results for Data set A (complete) and Gamma model. Panel (a): interpolated smoothed influences $IF(y_{1i}, \bar{\hat{\mu}}, F_{n_1,y_1})$ (1) and $IF(y_{2i}, \bar{\hat{\mu}}, F_{n_2,y_2})$ (2). Panel (b): exponentially tilted null probability distributions $\tilde{p}_{1i}$ (1) and $\tilde{p}_{2i}$ (2) with equal means ($\lambda = -0.0120$). Panel (c): smooth densities (truncated at 30) corresponding to null probability distributions for samples 1 and 2. Dotted lines are computed with the iterative algorithm derived from Proposition 2, solid lines using the linearized constraint. Panel (d): densities (truncated at 30) of the parametric null model estimated with TM/C.