

Robust parametric means of asymmetric distributions: estimation and testing

A. Marazzi and Giulia Barbati

Institut de médecine sociale et préventive, Université de Lausanne

[This article is published in *Estadística*, Vol 54, pp 47-72 (2002)]

Abstract. A robust parametric mean is the mean of a robustly estimated parametric model. Here we review a family of robust means for location-scale and shape-scale univariate models such as Lognormal, Weibull, Gamma, and Pareto distributions. The procedures are based on three steps: first the model is adjusted using a high breakdown-point estimator; second, outliers with respect to the initial estimate are rejected; third, an efficient estimate of mean is computed with the remaining data. In addition, the final estimate is corrected to make it Fisher consistent. The procedures include the truncated mean and the truncated maximum likelihood estimate. We also discuss the use of the bootstrap for the computation of the finite sample distribution of a robust mean and consider the “robust bootstrap” as a tool for improving the robustness of the approximation. For the problem of testing hypotheses concerning robust means we describe methods to estimate the models under the null hypotheses. We detail the two-sample case and provide examples with real data.

Keywords: robust estimates, robust tests, tests of means, bootstrap tests.

1 Introduction

Asymmetric distributions of positive random variables occur in many statistical applications. Often the population mean (i.e., the expected value of the random variable) is the main parameter and the problem of comparing means of one or several populations is very common. Unfortunately, the mean is a difficult parameter to estimate and test well: the sample mean, which is the natural estimate is very non-robust. For example, the mean costs of medically homogeneous groups of patients are used for hospital budgeting and it is common to compare cost means among different hospitals or over different periods of time. It is easy to give examples where a few atypical stays drastically change the mean estimate and where common tests of means (e.g. the t-test and its variants) lead to a different decision when these outliers are removed from the data set.

In recent years procedures for automatic outlier detection, robust mean estimation, and comparison of robust means of asymmetric data have been studied. In particular, procedures based on robust fitting of parametric models have been shown to be useful in applications (e.g., Victoria Feser and Ronchetti,

1994, 1997; Victoria-Feser, 2000; Marazzi et al., 1998). Within this framework a *robust parametric mean* is defined as the mean of the estimated model.

Both location-scale and shape-scale models were considered and several estimates were investigated. For example, various kinds of M-estimates were described by Hampel et al. (1986). Their definition and computation is however rather intricate because several implicit equations have to be solved. In addition, certain types of M-estimates were implemented (Marazzi and Ruffieux, 1996) but turned out to be too sensitive to contamination in many applications (e.g., Marazzi and Ruffieux, 1999). Simpler and more robust procedures were therefore investigated.

The truncated mean (Marazzi and Ruffieux, 1999) and the truncated maximum likelihood estimate (Marazzi and Yohai, 2002) are highly robust procedures for trimming atypical observations and computing a mean estimate. These procedures are based on three steps: firstly, a parametric model is fitted to the data using a high breakdown-point – but inefficient – estimator such as an S-estimator; secondly, upper and lower trimming limits are defined on the ground of the fitted model and observations beyond these limits are rejected; thirdly, the arithmetic mean or the maximum likelihood estimate is computed using the remaining observations. In addition, the final estimate is corrected to make it asymptotically unbiased at the model (i.e., Fisher consistent). Appropriate choice of the cut-off limits can make the final mean estimates very efficient. Some of these procedures have been extended to regression with asymmetric error models (Marazzi and Yohai, 2002).

The sample distribution of these estimators can be approximated using asymptotic results, but finite sample inference – e.g., confidence intervals and tests – is usually obtained with the help of the bootstrap. Three kinds of difficulties must however be faced. Firstly, non-parametric bootstrap of robust estimators is not robust: an exceeding frequency of outliers in several simulated samples may drastically affect the bootstrap distribution (e.g., Gosh et al., 1984; Singh, 1998). Secondly, the computation of robust estimates is usually slow. Thirdly, in testing hypotheses, the bootstrap samples have to be generated from a null model (i.e., a model that satisfies the null hypothesis) whose computation may be awkward. A remedy to the first two shortcomings is the “robust bootstrap” proposed by Salibian-Barrera (2000) and Salibian-Barrera and Zamar (2002). The computation of null models for hypotheses concerning robust means of asymmetric distributions has been discussed in Marazzi (2002).

This paper reviews in a unified way the most common estimation and testing procedures mentioned above. For simplicity, only univariate data are considered (i.e., we do not consider regression). We discuss the two-sample testing problem in detail, the generalization to the one- and multi-sample cases being straightforward. Section 2 introduces the models. Section 3 defines and discusses the estimation procedures. Section 4 examines parametric and non-parametric procedures for the computation of null models. Section 5 describes the robust bootstrap and Section 6 details an algorithm to compute a robust bootstrap null distribution in a two-sample problem. Examples with real data are provided in Section 7.

2 Models

Let $X > 0$ be a random variable with unknown cumulative distribution function G and (asymmetric) density g . We write $X \sim G$. We are interested in estimating $\mu = \mu(G)$, the expected value of X , using a sample $\mathbf{x} = (x_1, \dots, x_n)$ of independent observations. We suppose that we are willing to use a two-parameter model $F_{\tau, \sigma}$ for the distribution $F = G \circ h^{-1}$ of some monotone increasing transformation $Y = h(X)$. The corresponding model for G is $G_{\tau, \sigma}$. We assume that σ is a scale parameter of $F_{\tau, \sigma}$ and that, either τ is a location parameter or τ is a shape parameter of $F_{\tau, \sigma}$. Unlike the symmetric case, the scale parameter is usually a main component of the mean of an asymmetric distribution and cannot be considered as a nuisance parameter. Let G_n and F_n be the empirical cdf's of \mathbf{x} and $\mathbf{y} = (h(x_1), \dots, h(x_n))$.

Common examples are Lognormal, Weibull, Gamma, and Pareto distribution models. In the Lognormal and Weibull cases, the transformation $h(X) = \ln(X)$ provides a location-scale parametrization of $F_{\tau, \sigma}$ (Gauss or Log-Weibull) for $\ln(X)$; in the Gamma case, $h(X) = X$ provides the usual shape-scale parametrization of $F_{\tau, \sigma}$.

Location-scale families of distributions can be derived from a standard parameter free model $G_{0,1}$ (e.g., standard Gauss or standard Log-Weibull) and usual estimators are location-scale invariant. Unfortunately, there is no standard case for genuine shape-scale models and most estimators (see Section 3) are scale equivariant but not shape equivariant.

3 Estimates

We first consider some high breakdown point estimators that can be used in the initial step. Then, we define two high breakdown-point and high efficiency estimators: the truncated mean and the truncated maximum likelihood estimate.

3.1 Initial estimators. We assume that τ and σ can be characterized by means of high breakdownpoint location-scale measures $m(F)$ and $s(F)$, i.e., for given values of $m(F)$ and $s(F)$, τ and σ can be obtained by solving

$$m(F_{\tau, \sigma}) = m(F) \quad \text{and} \quad s(F_{\tau, \sigma}) = s(F). \quad (1)$$

We assume that $\bar{m} = m(F_n)$ and $\bar{s} = s(F_n)$ are consistent estimates of $m(F)$ and $s(F)$. We define $(\bar{\tau}, \bar{\sigma})$ as a solution of

$$m(\tau, \sigma) = \bar{m}, \quad s(\tau, \sigma) = \bar{s}, \quad (2)$$

where $m(\tau, \sigma)$ and $s(\tau, \sigma)$ are abbreviations for $m(F_{\tau, \sigma})$ and $s(F_{\tau, \sigma})$, respectively. The estimate $(\bar{\tau}, \bar{\sigma})$ has been called a *location-dispersion estimator* in Marazzi and Ruffieux (1999). A simple choice for m and s is the median and the median absolute deviation; for simplicity, the estimate $(\bar{\tau}, \bar{\sigma})$ based on this choice will be referred to as a *D-estimate* in the following.

A more sophisticated initial estimator is the S-estimator (Rousseeuw and Yohai, 1984) defined as follows. Let

$$\chi_k(z) = \begin{cases} 3(z/k)^2 - 3(z/k)^4 + (z/k)^6 & \text{if } |z| \leq k, \\ 1 & \text{if } |z| > k. \end{cases} \quad (3)$$

For a given value of k (e.g., $k = 1$) and $a = 0$, the S -estimate of $m(F)$ is

$$\bar{m} = \arg \min_M S(M), \quad (4)$$

where $S(M)$ is the solution of

$$0.5 = \frac{1}{n} \sum_{i=1}^n \chi_k((y_i - M)/S - a) \quad (5)$$

with respect to S , for given M ; $\bar{s} = S(\bar{m})$ is the corresponding estimate of $s(F)$. Here, $m(F)$ and $s(F)$ are simply defined as the asymptotic values of \bar{m} and \bar{s} . We may then use $m(F_{\tau,\sigma})$, $s(F_{\tau,\sigma})$, \bar{m} , and \bar{s} in (2) to obtain a location-dispersion estimate of (τ, σ) . Alternatively, we may firstly specify the functionals $m(F)$ and $s(F)$, and then compute a *corrected S-estimate* (\bar{m}, \bar{s}) according to (4)-(5), where the constants k and a are chosen so that (\bar{m}, \bar{s}) is consistent for $(m(F), s(F))$ when $F = F_{\tau,\sigma}$, i.e.,

$$a = \arg \min_a \int \chi_k((y - m(\tau, \sigma))/s(\tau, \sigma) - a) dF_{\tau,\sigma}(y), \quad (6)$$

$$0.5 = \int \chi_k((y - m(\tau, \sigma))/s(\tau, \sigma) - a) dF_{\tau,\sigma}(y). \quad (7)$$

In the location-scale case, we may choose m and s such that $m(\tau, \sigma) = \tau$, $s(\tau, \sigma) = \sigma$ and (6)-(7) can just be solved for the case $\tau = 0$, $\sigma = 1$. For example, in the Gaussian case, one obtains $a = 0$ and $k = 1.548$; in the Log-Weibull case, $a = -0.135$ and $k = 1.718$. In the shape-scale case, we may take $m(\tau, \sigma)$ as the model mean and s such that $s(\tau, \sigma) = \sigma$. Unfortunately, k and a will depend on τ and equations (6)-(7) must be solved together with (4)-(5) and (2) with $\tau = \bar{\tau}$.

3.2 The truncated mean. Assume that $(\bar{\tau}, \bar{\sigma})$ is a given high breakdown point and Fisher consistent estimate of (τ, σ) . For example, $(\bar{\tau}, \bar{\sigma})$ is a location-dispersion- or a corrected S-estimate. Let $u \in (0.5, 1)$ be a user chosen number (e.g., 0.99; see Section 3.4) and

$$T_u = q_u(G_{\bar{\tau}, \bar{\sigma}}) \quad (8)$$

where $q_u(G)$ denotes the u -quantile of G . Define T_l so that

$$\frac{1}{u - G_{\bar{\tau}, \bar{\sigma}}(T_l)} \int_{T_l}^{T_u} x dG_{\bar{\tau}, \bar{\sigma}}(x) = \int x dG_{\bar{\tau}, \bar{\sigma}}(x). \quad (9)$$

In other words, the mean of the truncated model equals the mean of the entire model. The *truncated mean* or *TM-estimate* of $\hat{\mu}$ is then defined as

$$\hat{\mu} = \text{ave}\{x_i \mid T_l < x_i \leq T_u\}, \quad (10)$$

i.e., the arithmetic mean of the x_i such that $T_l < x_i \leq T_u$.

3.3 The truncated maximum likelihood estimate. Again, we assume that $(\bar{\tau}, \bar{\sigma})$ is a given high breakdown point and Fisher consistent estimate. Let $u \in (0.5, 1)$ be a user chosen number and let

$$R_u = q_u(F_{\bar{\tau},1}). \quad (11)$$

Define R_l and b so that

$$\frac{1}{u - F_{\bar{\tau},1}(R_l)} \int_{R_l}^{R_u} s_j\left(\frac{y}{b}; \bar{\tau}, 1\right) dF_{\bar{\tau},1}(y) = 0, \quad j = 1, 2 \quad (12)$$

where $s_1(y; \tau, \sigma)$ and $s_2(y; \tau, \sigma)$ denote the score functions of $F_{\tau, \sigma}$ with respect to τ and σ , respectively. The *truncated maximum likelihood (TML-) estimate* $(\hat{\tau}, \hat{\sigma})$ of (τ, σ) is defined as a solution of the system of two equations

$$\text{ave}\{s_j(y_i/b; \tau, \sigma) \mid T_l < y_i \leq T_u\} = 0, \quad j = 1, 2, \quad (13)$$

where $T_l = \bar{\sigma}R_l$ and $T_u = \bar{\sigma}R_u$. Finally, the *truncated maximum likelihood- or TML-mean* is defined as $\hat{\mu} = \mu(G_{\hat{\tau}, \hat{\sigma}})$.

3.4 Discussion. Both the truncated mean and the maximum likelihood mean are examples of robust parametric means. Their asymptotic values are interpreted as approximations of the population mean after removal of the extreme values. Both these estimates maintain the breakdown-point of the initial estimate (e.g., 50%) and are asymptotically normal (under mild regularity assumptions). Expressions for their influence functions and variance estimates are available. The TML-estimate can be made as efficient as desired with a convenient choice of u . It is even possible to use adaptive cut-off limits T_l and T_u that, asymptotically, do not reject any observation when the data are generated according to the model (Gervini and Yohai, 2002; Marazzi and Yohai, 2002). The *adaptively truncated maximum likelihood estimates* obtained in this way are fully efficient at the model. Except for extremely skewed distributions (e.g., Lognormal with normal scale greater than 1), the TM-estimate is also very efficient if u is sufficiently large (Marazzi and Ruffieux, 1999, give values of u for which the asymptotic relative efficiency of the TM-estimate with respect to the maximum likelihood estimator is 0.8, 0.85, and 0.9). We note, however, that the TM-estimate does not fully make use of the model; this usually prevents full efficiency. On the other hand, the fact that the estimator does not strongly depend on the model may be a desirable feature in many situations.

The median and median absolute deviation are very simple initial estimates. However, the final estimate generally performs better if it starts with a S-estimate. For a numerical comparison, consider the TM-estimate starting with

median and median deviation (TM/D), the TM-estimate starting with the corrected S-estimate (TM/S), the TML-estimates starting with D (TML/D) and S (TML/S) as well as the initial D- and S-estimates ($u = 0.99$ in all computations). Figure 1 shows the asymptotic bias of the corresponding estimates of $\log(\mu)$ (i.e., $\log(\hat{\mu})$) at a contaminated Weibull model with shape 1 and scale 1 ($\log(\mu) = 0$). A 10% point mass is placed at a varying value y_{out} . As expected, the bias of the final estimates is higher than the maximum bias of the corresponding initial estimates over a limited range of y_{out} values. However, the bias of the final estimates jumps down to zero as soon as y_{out} exceeds the rejection point. Despite the maximum bias of D being lower than the maximum bias of S, the bias curves of TM/S and TML/S look better than those of TM/D and TML/D: they have a lower maximum and jump down earlier for increasing y_{out} .

Figure 1. Asymptotic bias of various estimates of $\log(\mu)$ ($= 0$) at a contaminated Weibull model (a 10% point mass is placed at a varying value y_{out}). Panel (a): D-estimate (thin full line), S-estimate (thin broken line), TML/D-estimate (thick full line), TML/S (thick broken line). The plots for TM/S and TM/D are almost identical to those of TML/D and TML/S.

Table 1 reports the results of a small simulation study comparing the same estimators. Data were generated according to models of the form $G = (1 - \varepsilon)G_{\tau,\sigma} + \varepsilon\tilde{G}$, where the contamination \tilde{G} is a uniform distribution over the interval $[0, a]$, $\varepsilon \in [0, 1]$ is the contamination proportion, and the central model $G_{\tau,\sigma}$ is a Weibull or a Lognormal distribution. The mean of both the central distributions is 0.5, but the skewnesses are very different. The mean squared errors (multiplied by n) for the estimates of $\log(\mu)$ over 2000 samples of size $n = 100$ were computed. Beside the obvious observation that all robust estimators offer protection against contamination, one can note that the estimators starting with S are better than those starting with D when the contamination is longtailed (but they are slightly worse in case of no or very short contamination).

	$G_{\tau,\sigma} : \text{Lognormal}$				$G_{\tau,\sigma} : \text{Weibull}$			
	ε	0.0	0.1	0.1	0.1	0.0	0.1	0.1
a	–	10	50	100	–	10	50	100
M	1.57	4.44	78.59	187.84	0.26	14.85	172.21	350.95
ML	1.42	5.70	47.89	93.18	0.26	15.27	128.40	235.99
D	2.92	8.16	12.29	12.94	0.41	1.11	1.37	1.40
S	4.31	7.75	7.38	7.39	0.74	0.70	0.66	0.65
TMD	1.99	7.27	9.49	5.83	0.31	0.36	0.32	0.32
TMS	2.12	6.96	6.40	4.10	0.29	0.35	0.30	0.31
TMLD	2.00	7.97	7.61	4.47	0.32	0.36	0.32	0.32
TMLS	1.98	7.91	5.91	3.76	0.32	0.33	0.32	0.33

Table 1. Simulated mean squared errors (multiplied by $n = 100$) of various estimators when $G = (1 - \varepsilon)G_{\tau,\sigma} + \varepsilon\tilde{G}$, $G_{\tau,\sigma}$ is Weibull (shape=2, scale=1.860) or Lognormal (normal mean=0; normal scale=1) and \tilde{G} is uniform over $[0, a]$.

4 Bootstrap tests: setting up the null models

Marazzi (2002) considers the problem of testing the hypothesis that the robust mean of a single population equals a given value (one-sample problem) or that the robust means of several populations are identical (multi-sample problem). For simplicity, we summarize here only the main methods for the two-sample problem, the most frequent case in practice.

We suppose that X_j ($j = 1, 2$) are random variables with unknown cdf's G_j and that G_j can be approximately described with the help of some, possibly contaminated model G_{τ_j, σ_j} , with shape τ_j and scale σ_j . In some cases (e.g., Lognormal and Weibull), it is convenient to use a transformations $Y_j = h(X_j)$ (e.g., $h(\cdot) = \ln(\cdot)$) in such a way that the transformed model F_{τ_j, σ_j} belongs to a location-scale family. We assume the same family for both samples, although one could give up this restriction. Let $\mathbf{x}_j = (x_{j1}, \dots, x_{jn_j})$ ($j = 1, 2$) be samples of iid observations from G_j and let G_{jn_j} be the empirical cdf of X_j . We use the abbreviation $\mu(\tau_j, \sigma_j)$ in place of $\mu(G_{\tau_j, \sigma_j})$ and assume that some procedure for the computations of a robust Fisher consistent and asymptotically normal estimator $(\hat{\tau}(G_{jn_j}), \hat{\sigma}(G_{jn_j}))$ of (τ_j, σ_j) with asymptotic value $(\hat{\tau}(G_j), \hat{\sigma}(G_j))$ is available (Section 3), that $\hat{\mu}$ is scale equivariant, and define $\hat{\mu}(G_j) = \hat{\mu}(G_{\hat{\tau}(G_j), \hat{\sigma}(G_j)})$. We use the abbreviations $\hat{\tau}_j$, $\hat{\sigma}_j$, and $\hat{\mu}_j$ in place of $\hat{\tau}(G_{jn_j})$, $\hat{\sigma}(G_{jn_j})$ and $\hat{\mu}(G_{jn_j})$.

Like a sample mean, the population mean can be strongly affected by a small proportion of extreme data – often members of a foreign population (see examples in Section 7). A robust mean can then be interpreted as an approximation of the population mean after removal of the extreme values and is, therefore, a very interesting parameter. In this section we consider inference techniques for the robust mean $\hat{\mu}(G_j)$ (in place of $\mu(G_j)$) and use the bootstrap as a general tool to compute the finite sample distribution of the test statistic.

4.1 The test statistic. In order to test the hypothesis

$$\mathcal{H}_0 : \hat{\mu}(G_1) = \hat{\mu}(G_2), \quad (14)$$

various test statistics could be considered. For the purpose of illustration, we restrict our attention to

$$t(\mathbf{x}_1, \mathbf{x}_2) = \frac{z(\hat{\mu}_1) - z(\hat{\mu}_2)}{\sqrt{z'(\hat{\mu}_1)^2 \hat{V}_1 + z'(\hat{\mu}_2)^2 \hat{V}_2}}, \quad (15)$$

where z is a variance stabilizing transformation, usually $z(\cdot) = \ln(\cdot)$, and \hat{V}_j is an estimate of the variance of $\hat{\mu}_j$, for example, the influence function estimate $\hat{V}_j = (1/n_j) \int IF(x; \hat{\mu}_j, G_{\hat{\tau}_j, \hat{\sigma}_j})^2 dG_{\hat{\tau}_j, \hat{\sigma}_j}$, where $IF(x; \hat{\mu}_j, G_j)$ denotes the influence function of $\hat{\mu}_j$ evaluated at a point x , when $X \sim G_j$ (Hampel et al., 1996)

4.2 The null models. In order to compute the null distribution of the test statistic, simulated samples are generated from estimates $\hat{G}_1^{(0)} \cdot \hat{G}_2^{(0)}$ of a joint

cdf $G_1 \cdot G_2$ that satisfy the null hypothesis, i.e., from a *null model*. Three approaches to the computation of the null model can be considered.

A *non-parametric null model* supported on the sample values can be constructed as follows. Let \mathcal{P} be the set of discrete cdf's $G_{1p_1} \cdot G_{2p_2}$ such that $G_{jp_j} = \sum p_{ji} \Delta_{x_{ji}}$, where $\Delta_{x_{ji}}$ denotes the cdf of a point mass at x_{ji} , $p_{ji} \geq 0$ for $i = 1, \dots, n_j$, $p_j = (p_{j1}, \dots, p_{jn_j})$, and let $\mathcal{P}_0 = \{G_{1p_1} \cdot G_{2p_2} \in \mathcal{P} \mid \hat{\mu}(G_{1p_1}) = \hat{\mu}(G_{2p_2})\}$. We look for a distribution $G_{1\tilde{p}_1} \cdot G_{2\tilde{p}_2} \in \mathcal{P}_0$ such that

$$G_{1\tilde{p}_1} \cdot G_{2\tilde{p}_2} = \arg \min_{\mathcal{P}_0} [d_{KL}(G_{1n_1}, G_{1p_1}) + d_{KL}(G_{2n_2}, G_{2p_2})], \quad (16)$$

where $d_{KL}(G_{jn_j}, G_{jp_j}) = \sum p_{ji} \ln(n_j p_{ji})$ denotes the Kullback-Leibler disparity between G_{jn_j} and G_{jp_j} . In other words, $G_{1\tilde{p}_1} \cdot G_{2\tilde{p}_2}$ is the non-parametric maximum likelihood estimate of $G_1 \cdot G_2$ under the constraint \mathcal{H}_0 . This problem is considered in Efron and Tibshirani (1993) and Davison and Hinkley (1997) for the case where $\hat{\mu}(G_j)$ is the mean of G_j . Unlike that case, a robust mean $\hat{\mu}(G_{jp_j})$ is usually nonlinear in p_j and its computation is not straightforward. One can show, however, that the solution is a discrete cdf $G_{1\tilde{p}_1} \cdot G_{2\tilde{p}_2}$ that satisfies the implicit equations

$$\tilde{p}_{1i} = \frac{\exp(\lambda IF(x_{1i}; \hat{\mu}_1, G_{1\tilde{p}_1}))}{\sum_i \exp(\lambda IF(x_{1i}; \hat{\mu}_1, G_{1\tilde{p}_1})}, \quad i = 1, \dots, n_1, \quad (17)$$

$$\tilde{p}_{2k} = \frac{\exp(-\lambda IF(x_{2k}; \hat{\mu}_2, G_{2\tilde{p}_2}))}{\sum_k \exp(-\lambda IF(x_{2k}; \hat{\mu}_2, G_{2\tilde{p}_2})}, \quad k = 1, \dots, n_2, \quad (18)$$

where the scalar λ is determined so that $\mu(G_{1\tilde{p}_1}) = \mu(G_{2\tilde{p}_2})$. One says that $G_{j\tilde{p}_j}$ is an *exponentially tilted version* of G_{jn_j} . The computation of \tilde{p}_{ji} is quite cumbersome but some algorithms are given in Marazzi (2002). For the non-parametric bootstrap simulation, the null model is therefore $\hat{G}_1^{(0)} \cdot \hat{G}_2^{(0)} = G_{1\tilde{p}_1} \cdot G_{2\tilde{p}_2}$.

A *parametric null model* is obtained by fitting a model $G_{\tau_1, \sigma_1} \cdot G_{\tau_2, \sigma_2}$ to the empirical cdf $G_{1n_1} \cdot G_{2n_2}$ under the constraint

$$\mathcal{H}_0 : \hat{\mu}(G_{\tau_1, \sigma_1}) = \hat{\mu}(G_{\tau_2, \sigma_2}). \quad (19)$$

This condition is equivalent to $\mu(G_{\tau_1, \sigma_1}) = \mu(G_{\tau_2, \sigma_2})$ when $(\hat{\tau}_j, \hat{\sigma}_j)$ is Fisher consistent. A convenient (easily computable) robust estimate of $(\tau_1, \sigma_1, \tau_2, \sigma_2)$ satisfying this constraint is the *C-estimator based on* $(\hat{\tau}_1, \hat{\sigma}_1)$ and $(\hat{\tau}_2, \hat{\sigma}_2)$ defined as

$$(\tilde{\tau}_1, \tilde{\sigma}_1, \tilde{\tau}_2, \tilde{\sigma}_2) = \arg \min_{\Theta_0} [d_{KL}(G_{\tau_1, \sigma_1}, G_{\hat{\tau}_1, \hat{\sigma}_1}) + d_{KL}(G_{\tau_2, \sigma_2}, G_{\hat{\tau}_2, \hat{\sigma}_2})], \quad (20)$$

where $\Theta_0 = \{(\tau_1, \sigma_1, \tau_2, \sigma_2) \mid \mu(G_{\tau_1, \sigma_1}) = \mu(G_{\tau_2, \sigma_2})\}$ is the parameter space under \mathcal{H}_0 and $d_{KL}(H, K)$ denotes the Kullback-Leibler disparity between H and K . Here, $(\hat{\tau}_1, \hat{\sigma}_1)$ and $(\hat{\tau}_2, \hat{\sigma}_2)$ are available unconstrained estimates such as those described in Section 3. For the non-parametric bootstrap simulation, the null model is therefore $\hat{G}_1^{(0)} \cdot \hat{G}_2^{(0)} = G_{\tilde{\tau}_1, \tilde{\sigma}_1} \cdot G_{\tilde{\tau}_2, \tilde{\sigma}_2}$.

In order to reduce the parametric dependency of the null model, one may also consider a *semi-parametric null model* based on the assumption that $X_j \sim G_j(\cdot/\sigma_j)$ for some unspecified distributions G_1 and G_2 and set $\hat{G}_j^{(0)}$ equal to the empirical cdf of the rescaled sample $\mathbf{x}_j^{(0)} = (\mu_0/\hat{\mu}_j)\mathbf{x}_j$ and μ_0 is arbitrary. $\hat{G}_1^{(0)} \cdot \hat{G}_2^{(0)}$ satisfies \mathcal{H}_0 because $\mu_0/\hat{\mu}_j$ is a scale factor and $\hat{\mu}_j$ is scale equivariant. When $h(\cdot) = \log(\cdot)$, the test statistic (15) does not depend on μ_0 .

4.3 Discussion. The C-estimator possesses the following property: if F_{τ_j, σ_j} belongs to the exponential family, the C-estimate based on unconstrained maximum likelihood estimates coincides with the constrained maximum likelihood estimate (and is a good approximation in other cases). This makes the C-estimator a simple and attractive way to obtain constrained robust estimates when unconstrained ones are already available. The asymptotic normality of the C-estimator readily follows from the assumed properties of the unconstrained estimators and the C-estimator clearly inherits the breakdown-point of the given unconstrained estimators.

Several numerical examples of two-sample tests with real and simulated data are discussed in Marazzi (2002). Bootstrap tests based on the truncated mean are compared with classical tests (including the pooled t-test and the test of Cressie and Whitford, 1986) as well as simple robust tests based on the trimmed mean (including the test of Guo and Luh, 2000). The main conclusion is that tests based on high breakdownpoint parametric estimates perform much better than their competitors – with respect to both type I and II error probabilities – both when parametric and non-parametric simulation is used.

Non-parametric simulation from (modified versions of) the empirical distribution provides, however, less stable null distributions than simulation from the parametric models. This is due to the “exceeding frequency” of outliers in many simulated samples, a feature that affects the tails of the null distribution. This “lack of robustness” of non-parametric bootstrap has been noted by several authors (e.g., Gosh et al., 1984; Singh, 1998). On the other hand, parametric simulation tends to underestimate real fluctuations, since it ignores contamination. A procedure to reduce the outlier effect in non-parametric simulation is the “robust bootstrap” proposed by Salibian-Barrera (2000) and Salibian-Barrera and Zamar (2002).

5 Robust bootstrap simulation

In this section, the notations $\hat{\mu}, \hat{\tau}, \hat{\sigma}$ will be used as abbreviations of $\hat{\mu}(F)$, $\hat{\tau}(F)$, and $\hat{\sigma}(F)$. The robust bootstrap is based on the observation that many robust estimates can be expressed as weighted averages of the observations. We consider – as in Section 3 – a random variable X with cdf G , a sample $\mathbf{x} = (x_1, \dots, x_n)$ of X , and the transformed random variable $Y = h(X)$ with cdf F . As a robust estimate we take, for example, the truncated mean $\hat{\mu}$ based on

the S-estimates $\bar{\tau}$, $\bar{\sigma}$. The S-estimate satisfies

$$\sum \psi_k((y_i - \bar{\tau})/\bar{\sigma}) = 0, \quad \sum \chi_k((y_i - \bar{\tau})/\bar{\sigma}) = n/2, \quad (21)$$

where $\psi_k(z) = (d/dz) \chi_k(z)$. Therefore, we may write

$$\hat{\mu} = \frac{\sum x_i u_i}{\sum u_i}, \quad \bar{\tau} = \frac{\sum y_i v_i}{\sum v_i}, \quad \bar{\sigma} = \sum (y_i - \bar{\tau}) w_i, \quad (22)$$

where

$$u_i = I(T_l < x_i \leq T_u), \quad (23)$$

$$v_i = \psi_k((y_i - \bar{\tau})/\bar{\sigma})/(y_i - \bar{\tau}), \quad (24)$$

$$w_i = \frac{2\bar{\sigma} \chi_k((y_i - \bar{\tau})/\bar{\sigma})}{n (y_i - \bar{\tau})}. \quad (25)$$

Here, $T_l = T_l(\bar{\tau}, \bar{\sigma})$ and $T_u = T_u(\bar{\tau}, \bar{\sigma})$ are defined by (8) and (9) and $I(\cdot)$ denotes the indicator function. Thus, u_i , v_i , and w_i are all functions of $(\bar{\tau}, \bar{\sigma})$ and \mathbf{x} and we define the vector function

$$\mathbf{k}(\hat{\mu}, \bar{\tau}, \bar{\sigma}; \mathbf{x}) = \left(\frac{\sum x_i u_i}{\sum u_i}, \frac{\sum y_i v_i}{\sum v_i}, \sum (y_i - \bar{\tau}) w_i \right)^T. \quad (26)$$

Note that

$$(\hat{\mu}, \bar{\tau}, \bar{\sigma})^T = \mathbf{k}(\hat{\mu}, \bar{\tau}, \bar{\sigma}; \mathbf{x}). \quad (27)$$

In order to satisfy some required regularity assumptions it is necessary to replace I in (23) with a smoothed version \hat{I} (see Remark, below).

We now suppose that $\mathbf{x}^* = (x_1^*, \dots, x_n^*)$ is a non-parametric bootstrap sample from G_n , that $y_i^* = h(x_i^*)$, $u_i^* = \hat{I}(T_l < x_i^* \leq T_u)$, $v_i^* = \psi_k((y_i^* - \bar{\tau})/\bar{\sigma})/(y_i^* - \bar{\tau})$, $w_i^* = (2\bar{\sigma}/n)[\chi_k((y_i^* - \bar{\tau})/\bar{\sigma})]/(y_i^* - \bar{\tau})$, and that

$$(\hat{\mu}^*, \bar{\tau}^*, \bar{\sigma}^*)^T = \left(\frac{\sum x_i^* u_i^*}{\sum u_i^*}, \frac{\sum y_i^* v_i^*}{\sum v_i^*}, \sum (y_i^* - \bar{\tau}) w_i^* \right)^T. \quad (28)$$

Note that the estimates $\bar{\tau}$ and $\bar{\sigma}$ involved in u_i^* , v_i^* , and w_i^* are based on the original data, not on the bootstrap sample. From (27) we have:

$$(\hat{\mu}, \bar{\tau}, \bar{\sigma})^T \approx \mathbf{k}(\hat{\mu}, \bar{\tau}, \bar{\sigma}; \mathbf{x}) + \nabla \mathbf{k}(\hat{\mu}, \bar{\tau}, \bar{\sigma}; \mathbf{x}) [(\hat{\mu}, \bar{\tau}, \bar{\sigma})^T - (\hat{\mu}^*, \bar{\tau}^*, \bar{\sigma}^*)^T], \quad (29)$$

where $\nabla \mathbf{k}(\hat{\mu}, \bar{\tau}, \bar{\sigma}; \mathbf{x})$ is the matrix of first derivatives of \mathbf{k} with respect to $(\hat{\mu}, \bar{\tau}, \bar{\sigma})$. Salibián-Barrera (2000) uses this linear approximation to show that the bootstrap distribution of $(\hat{\mu}, \bar{\tau}, \bar{\sigma})^T$ can be approximated with the help of the ‘‘robust bootstrap’’ simulated values $(\hat{\mu}^{R*}, \bar{\tau}^{R*}, \bar{\sigma}^{R*})$ defined by

$$\begin{bmatrix} \hat{\mu}^{R*} \\ \bar{\tau}^{R*} \\ \bar{\sigma}^{R*} \end{bmatrix} = \begin{bmatrix} \hat{\mu} \\ \bar{\tau} \\ \bar{\sigma} \end{bmatrix} + \mathbf{A} \begin{bmatrix} \hat{\mu}^* - \hat{\mu} \\ \bar{\tau}^* - \bar{\tau} \\ \bar{\sigma}^* - \bar{\sigma} \end{bmatrix}, \quad (30)$$

where $\mathbf{A} = [\mathbf{I} - \nabla \mathbf{k}(\hat{\mu}, \bar{\tau}, \bar{\sigma}; \mathbf{x})]^{-1}$. Thus, there is no need to recompute the original estimates for each bootstrap sample. Moreover, outlying points are typically associated with small weights in equation (28), and hence have little effect on the re-calculated estimates.

The ‘‘robust bootstrap approximation’’ can be used to compute bootstrap confidence intervals for $\hat{\mu}$, $\hat{\tau}$ and $\hat{\sigma}$. In the next section, for the purpose of testing an hypothesis \mathcal{H}_0 , we consider a robust bootstrap simulation from a null model. Resampling from a null model is also useful for computing confidence intervals when \mathcal{H}_0 is accepted.

Remark. An example of smooth function \check{I} is:

$$\check{I}(x; L, U) = I(L < x < U) + I(L - \gamma \leq x \leq L) \left[1 - ((x - L)/\gamma)^2\right]^2 + I(U \leq x \leq U + \gamma) \left[1 - ((x - U)/\gamma)^2\right]^2,$$

with $L = T_l - \gamma$, $U = T_u + \gamma$ and γ is small.

6 Algorithm for the robust bootstrap two-sample test

Here, we detail an algorithm for a bootstrap test of $\mathcal{H}_0 : \hat{\mu}(G_1) = \hat{\mu}(G_2)$ based on the truncated mean and the test statistic (15). The test uses a semi-parametric null model and the robust bootstrap. The approximation to the null distribution of the test statistic is based on a given number B of bootstrap samples.

Step 1. Set the simulation counter $K = 1$ and compute $t(\mathbf{x}_1, \mathbf{x}_2)$.

Step 2. For $j = 1, 2$, compute the rescaled samples $\mathbf{x}_j^{(0)} = (\mu_0/\hat{\mu}_j)\mathbf{x}_j$ and the transformed rescaled samples $\mathbf{y}_j^{(0)} = (h(x_{j1}^{(0)}), \dots, h(x_{jn_j}^{(0)}))$ (i.e., the semi-parametric null models $\hat{G}_1^{(0)} \cdot \hat{G}_2^{(0)}$ and $\hat{F}_1^{(0)} \cdot \hat{F}_2^{(0)}$). Compute $\bar{\tau}_j^{(0)} = \bar{\tau}(\hat{F}_j^{(0)})$, $\bar{\sigma}_j^{(0)} = \bar{\sigma}(\hat{F}_j^{(0)})$, $\mu_j^{(0)} = \hat{\mu}(\hat{G}_j^{(0)})$ (and note that $\hat{\mu}_1^{(0)} = \hat{\mu}_2^{(0)}$). Compute the weights u_{ji} , v_{ji} , w_{ji} associated with $x_{ji}^{(0)}$, for $i = 1, \dots, n_j$.

Step 3. For $j = 1, 2$, draw a sample $\mathbf{x}_j^* = (x_{j1}^*, \dots, x_{jn_j}^*)$ from $\hat{G}_j^{(0)}$, let $(u_{ji}^*, v_{ji}^*, w_{ji}^*)$ be the weights associated with x_{ji}^* , and compute a robust bootstrap simulated value of $(\hat{\mu}_j, \bar{\tau}_j, \bar{\sigma}_j)$ according to

$$\begin{bmatrix} \hat{\mu}_j^{R*} \\ \bar{\tau}_j^{R*} \\ \bar{\sigma}_j^{R*} \end{bmatrix} = \begin{bmatrix} \hat{\mu}_j^{(0)} \\ \bar{\tau}_j^{(0)} \\ \bar{\sigma}_j^{(0)} \end{bmatrix} + \mathbf{A}_j \begin{bmatrix} \hat{\mu}_j^* - \hat{\mu}_j^{(0)} \\ \bar{\tau}_j^* - \bar{\tau}_j^{(0)} \\ \bar{\sigma}_j^* - \bar{\sigma}_j^{(0)} \end{bmatrix}, \quad (31)$$

where $\mathbf{A}_j = [\mathbf{I} - \nabla \mathbf{k}(\hat{\mu}_j^{(0)}, \bar{\tau}_j^{(0)}, \bar{\sigma}_j^{(0)}; \mathbf{x}_j)]^{-1}$ and

$$(\hat{\mu}_j^*, \bar{\tau}_j^*, \bar{\sigma}_j^*)^\top = \left(\frac{\sum x_{ji}^* u_{ji}^*}{\sum u_{ji}^*}, \frac{\sum y_{ji}^* v_{ji}^*}{\sum v_{ji}^*}, \sum (y_{ji}^* - \bar{\tau}_j^{(0)}) w_{ji}^* \right)^\top. \quad (32)$$

Step 4. Compute a simulated value of the test statistic (15),

$$t^{R*} = \frac{z(\hat{\mu}_1^{R*}) - z(\hat{\mu}_2^{R*})}{\sqrt{z'(\hat{\mu}_1^{R*})^2 \hat{V}_1^{R*} + z'(\hat{\mu}_2^{R*})^2 \hat{V}_2^{R*}}}, \quad (33)$$

where \hat{V}_j^{R*} is an estimate of the variance of $\hat{\mu}_j^{R*}$ (see Remark, below).

Step 5. If $K < B$, increase K by 1 and go to step 3.

Step 6. Compute the achieved significance level as the proportion of simulated values t^{R*} that are greater than $t(\mathbf{x}_1, \mathbf{x}_2)$.

Remark. The covariance matrix of $(\hat{\mu}_j^{R*}, \bar{\tau}_j^{R*}, \bar{\sigma}_j^{R*})^T$ can be estimated by $(1/n_j)\mathbf{A}_j^{-1}\hat{\Sigma}_j^{R*}\mathbf{A}_j^{-T}$, where $\hat{\Sigma}_j^{R*}$ is an estimate of the asymptotic covariance matrix of $(\hat{\mu}_j, \bar{\tau}_j, \bar{\sigma}_j)^T$, for example

$$\hat{\Sigma}_j^{R*} = \frac{1}{n_j} \sum IF(y_{ji}^*; (\hat{\mu}_j, \bar{\tau}_j, \bar{\sigma}_j), F_{\hat{\tau}_j^{R*}, \hat{\sigma}_j^{R*}})^T IF(y_{ji}^*; (\hat{\mu}_j, \bar{\tau}_j, \bar{\sigma}_j), F_{\hat{\tau}_j^{R*}, \hat{\sigma}_j^{R*}}).$$

7 Examples

The average cost of “medically homogeneous groups of patients” (or Diagnosis Related Groups; Freeman et al., 1995) is routinely used as a basis for hospital resource allocation and budgeting. Around 500 groups and several hundred patients are common for any single hospital. Unfortunately, cost distributions are asymmetric and contain outliers whose value and frequency fluctuate from sample to sample (e.g., from year to year) and this makes the most common measure of average, the arithmetic mean, very unstable and inappropriate. Very often, outliers are badly classified stays. Therefore, outliers are usually removed according to various rules (Beguin et al., 1991) and the means of the remaining data are computed. The “robust means” are used to determine the reimbursent of the regular stays; extreme cases are carefully inspected and reimbursed with a different rule. Clearly, automatic estimation and comparison of robust means among different hospitals or over different periods of time are of great interest.

We first consider two samples of 74 and 77 patients hospitalized in a Swiss hospital during 1999 and 2000 for “cardiovascular surgery”. Their mean costs were 11'211 (1999) and 12'303 (2000) Swiss francs. Table 2 gives one sided P-values of two classical tests for comparing means, the usual t-test and the Cressie-Whitford (1986) test (based on a skewness adjusted statistic) as well as the P-value of a moderately robust test of Guo and Luh (2000) (based on a transformation for skewness and the symmetric 5%-trimmed mean). All these tests accept the null hypothesis of identical means (at the usual 5% level) when the entire data set is used. Figure 2, panels (a) and (b) shows normal probability plots of Lognormal models (the most common description of cost distributions) that have been fitted with the help of S-estimates. The models fit well the majority of the most recent (2000) costs, but there is a very strong contamination for 1999 (a thorough analysis shows that the contamination is substantially due to emergency cases). The truncated means (TM/S, $u = 0.95$) are 8'273

(1999) and 10'163 (2000) Swiss francs. Despite the poor fit for 1999, the robust bootstrap test described in Section 6 (with $h(\cdot) = \log(\cdot)$) is very significant in favour of a cost increase both with the complete data set and when outliers are removed and, in this case, all tests become strongly significant

Test	Full real data set	Outliers removed
Pooled t	0.33	0.00
Cressie -Whitford	0.15	0.00
Guo-Luh	0.20	0.00
TM/Robust bootstrap	0.001	0.00

Table 2. P-values of fours two sample test on “cardiovascular surgery” costs.

We now consider two samples of 83 and 66 patients hospitalized in 1999 and 2000 for “heart surgery”. The mean costs were 25'508 (1999) and 37'115 (2000) Swiss francs, whereas the truncated means are 25'694 and 27'706. Panels (c) and (d) in Figure 2 show the normal probability plots of Lognormal models based on S-estimates (again, the models provide reasonable descriptions of the cost majority and point out five very expensive cases in 2000). The results reported in Table 3 show that, according to the robust bootstrap test, there is no significant evidence in favour of a cost increase, both with the entire and the cleaned data set. The other tests are significant with the full data set but non-significant when outliers are removed.

Test	Full real data set	Outliers removed
pooled t	0.0030	0.14
Cressie -Whitford	0.0003	0.07
Guo-Luh	0.0020	0.07
TM/Robust bootstrap	0.0800	0.09

Table 3. P-values of fours two sample test on “heart surgery” costs.

Figure 2. Normal probability plot (expected/observed) of “foot surgery” costs for years 1999 (panel a) and 2000 (panel b) and normal probability plots of “heart surgery” costs for years 1999 (panel c) and 2000 (panel d).

Aknowledgement.

This work was supported by grant 21-54146.98 from the Swiss National Science Foundation.

References

- Beguín C., Closon M.C., Roger F.H., 1991. Advances in DRGs data pooling in Europe: Results from the Hoscom project in relation with 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.
- Freeman J.L., Fetter R.B., Park H., et al., 1995. Diagnosis Related Groups refinement with diagnosis- and procedure specific comorbidities and complications. *Medical Care* (33) 765.
- Gervini D., Yohai V.J., 2002. A class of robust and fully efficient regression estimates. *The Annals of Mathematical Statistics* (In press).
- Ghosh M., Parr W. C., Singh K. and Babu G.J., 1984. A note on bootstrapping the sample median. *The Annals of Statistics*, 12, 1130-1135.
- Guo J-H., Luh W-M., 2000. An invertible transformation two-sample trimmed t-statistics 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.
- Marazzi A., Ruffieux C., 1996. Implementing M-estimators of the Gamma distribution. In: H.Rieder (Ed.), *Robust Statistics, Data Analysis, and Computer Intensive Methods, In Honor of Peter Huber's 60th Birthday*, Lecture Notes in Statistics, 109, Springer Verlag, Heidelberg.
- Marazzi A., Ruffieux C., 1999. The truncated mean of an asymmetric distribution. *Computational Statistics & 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.
- Marazzi A., Yohai V.J., 2002. Adaptively truncated maximum likelihood regression with asymmetric errors. Draft.
- Marazzi A., 2002. Bootstrap tests for robust means of asymmetric distributions with unequal shapes. *Computational Statistics & Data Analysis*. In press.
- Rousseeuw P.Y., Yohai V., 1984. *Robust regression by means of S-estimators. Robust and Nonlinear Time Series Analysis*. Lecture Notes in Statist (26) 256-272. Springer, New York.
- Salibian-Barrera M., 2000. *Contributions to the theory of robust inference*. Ph.D. thesis, Dept. Statist., Univ. British Columbia, Vancouver.
- Salibian-Barrera M., Zamar R. H., 2002. Bootstrapping robust estimates of regression. *The Annals of Statistics*, 30 (2) 556-582.

Singh, K., 1998. Breakdown theory for bootstrap quantiles. *The Annals of Statistics*, 26, 1719-1732.

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.

Victoria-Feser M.P., 2000. Robust methods for the analysis of income distribution, inequality and poverty. *International Statistical Review*, 68, 3 277-293.

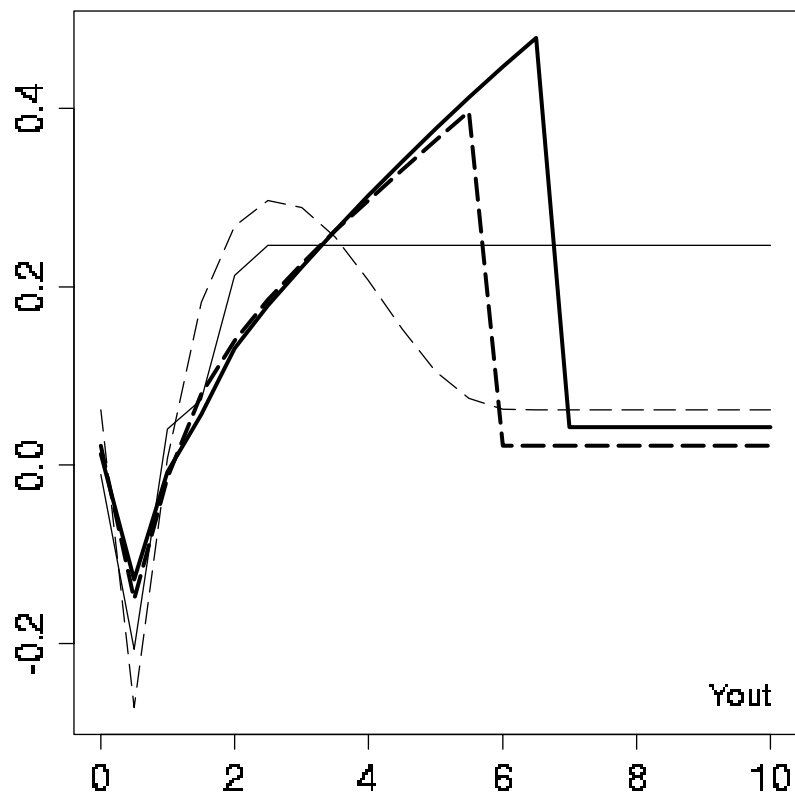


Figure 1

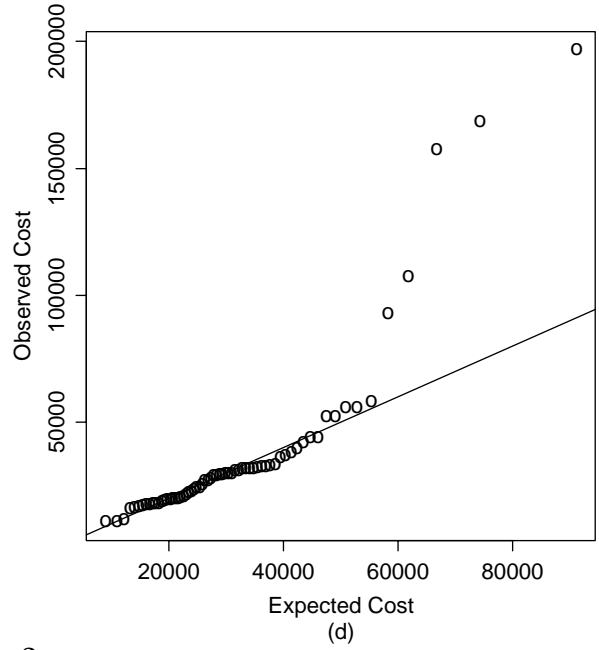
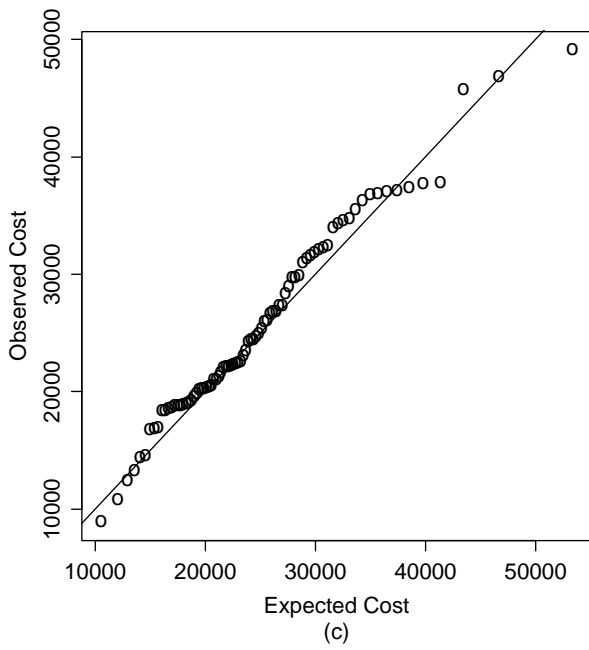
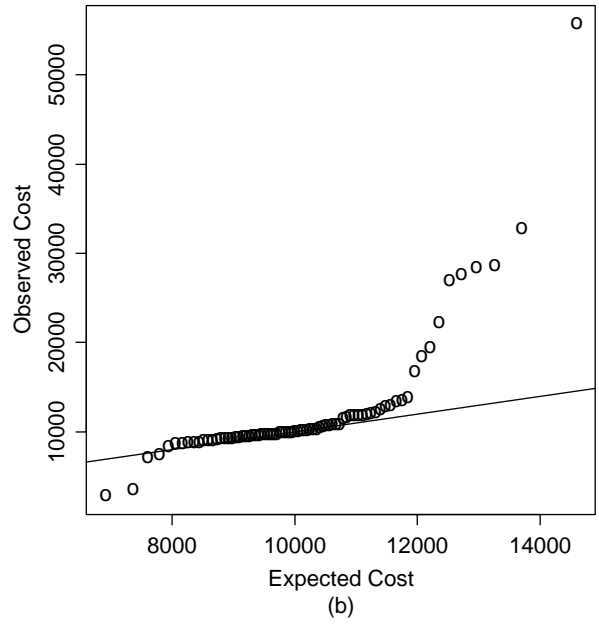
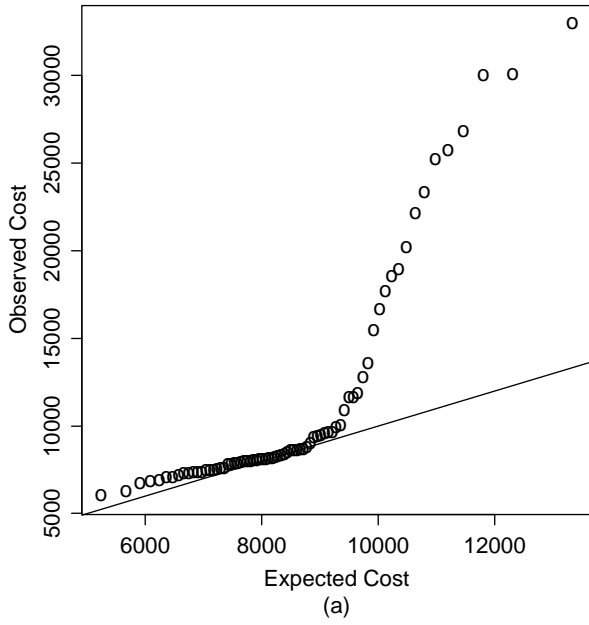


Figure 2