Available online at www.sciencedirect.com
ScienceDirect

Environmental Modelling & Software 22 (2007) 84–96

Environmental Modelling & Software

www.elsevier.com/locate/envsoft

Quantifying effects in two-sample environmental experiments using bootstrap confidence intervals

Manfred Mudelsee [a,b,*], Merianne Alkio [c]

[a] Climate Risk Analysis, Wasserweg 2, 06114 Halle (S), Germany
[b] Institute of Meteorology, University of Leipzig, Stephanstrasse 3, 04103 Leipzig, Germany
[c] Department of Biology, University of Massachusetts, Boston, 100 Morrissey Boulevard, Boston, MA 02125, USA

Received 1 August 2005; received in revised form 28 November 2005; accepted 10 December 2005
Available online 13 February 2006

Abstract

Two-sample experiments (paired or unpaired) are often used to analyze treatment effects in life and environmental sciences. Quantifying an effect can be achieved by estimating the difference in center of location between a treated and a control sample. In unpaired experiments, a shift in scale is also of interest. Non-normal data distributions can thereby impose a serious challenge for obtaining accurate confidence intervals for treatment effects. To study the effects of non-normality we analyzed robust and non-robust measures of treatment effects: differences of averages, medians, standard deviations, and normalized median absolute deviations in case of unpaired experiments, and average of differences and median of differences in case of paired experiments. A Monte Carlo study using bivariate lognormal distributions was carried out to evaluate coverage performances and lengths of four types of nonparametric bootstrap confidence intervals, namely normal, Student's *t*, percentile, and BCa for the estimated measures. The robust measures produced smaller coverage errors than their non-robust counterparts. On the other hand, the robust versions gave average confidence interval lengths approximately 1.5 times larger. In unpaired experiments, BCa confidence intervals performed best, while in paired experiments, Student's *t* was as good as BCa intervals. Monte Carlo results are discussed and recommendations on data sizes are presented. In an application to physiological source–sink manipulation experiments with sunflower, we quantify the effect of an increased or decreased source–sink ratio on the percentage of unfilled grains and the dry mass of a grain. In an application to laboratory experiments with wastewater, we quantify the disinfection effect of predatory microorganisms. The presented bootstrap method to compare two samples is broadly applicable to measured or modeled data from the entire range of environmental research and beyond.

Keywords: Agriculture; Bootstrap confidence interval; Monte Carlo simulation; Robust estimation; Seed filling; Two-sample problem; Wastewater

1. Introduction

The size of an effect of an experimental treatment is an important quantity not only in environmental sciences. It facilitates prediction and allows comparison of different treatments. It is therefore of central interest to estimate the effect of a treatment accurately. For this purpose, routine statistical methods like the Pitman test (Gibbons, 1985) or the *t*-test, of a hypothesis "treatment and control have identical means" are not sufficient. Instead, the present paper advocates usage of the difference of means between treatment (X) and control (Y) populations, abbreviated as d_{AVE} adopting the notation in Table 1. A confidence interval for d_{AVE} provides quantitative information, which also includes a statistical test (by looking whether it contains zero) but is not restricted to it.

Another objective of the present study is robustness of results against variations in distributional shape of the data. The experimental data analyzed in this paper show considerable amounts of skewness and deviation from Gaussian shape, which is a common feature of morphometric data. Therefore, a measure such as the difference of medians (d_{MED}) might be more suited than d_{AVE} because its accuracy is less influenced by variations in distributional shape. Furthermore, it is thought

* Corresponding author. Climate Risk Analysis, Wasserweg 2, 06114 Halle (S), Germany. Tel.: +49 345 532 3860; fax: +49 341 973 2899.
E-mail address: mudelsee@climate-risk-analysis.com (M. Mudelsee).

1364-8152/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.envsoft.2005.12.001

Table 1
Analyzed difference measures, degrees-of-freedom

Measure	ν
Unpaired	
$\widehat{d_{\mathrm{AVE}}} = \mathrm{AVE}(x) - \mathrm{AVE}(y)$	$n_x + n_y - 2$
$\widehat{d_{\mathrm{MED}}} = \mathrm{MED}(x) - \mathrm{MED}(y)$	$n_x + n_y - 2$
$\widehat{d_{\mathrm{STD}}} = \mathrm{STD}(x) - \mathrm{STD}(y)$	$n_x + n_y - 4$
$\widehat{d_{\mathrm{MAD}}} = \mathrm{MAD}'(x) - \mathrm{MAD}'(y)$	$n_x + n_y - 4$
Paired	
$\widehat{\mathrm{AVE}}_{\mathrm{d}} = \mathrm{AVE}(x - y)$	$n - 1$
$\widehat{\mathrm{MED}}_{\mathrm{d}} = \mathrm{MED}(x - y)$	$n - 1$

Note: $\{x(i), \quad i = 1, \ldots, n_x\}$ are treatment data, $\{y(i), \quad i = 1, \ldots, n_y\}$ are control data. In paired experiments, $n_x = n_y = n$. $\mathrm{AVE}(x) = \sum_{i=1}^{n_x} x(i)/n_x$ is sample mean, $\mathrm{STD}(x) = \sum_{i=1}^{n_x} [x(i) - \mathrm{AVE}(x)]^2 (n_x - 1)$ is sample standard deviation, $\mathrm{MED}(x)$ is sample median, and $\mathrm{MAD}'(x) = 1.4826 \ \mathrm{MAD}(x)$ where $\mathrm{MAD}(x) = \text{median } [|x(i) - \mathrm{MED}(x)|]$ is sample MAD (normalizing is used because a normal distribution has standard deviation MAD′); analogously for y; $\mathrm{AVE}(x - y) = \sum_{i=1}^{n} [x(i) - y(i)]/n$ and $\mathrm{MED}(x - y) = \text{median of } [x(i) - y(i)], \quad i = 1, \ldots, n$.

to measure the effects of a treatment on the variability. For example, a difference of standard deviations (d_{STD}) may indicate other influences than intended. Therefore, d_{STD} and, as a robust measure of the difference in scale, the difference in normalized median absolute deviations (d_{MAD}) (Tukey, 1977) are investigated.

The mentioned measures of difference are used in unpaired experiments, that means, where no dependence is assumed between treatment and control groups. We explore also paired experiments, in which the same element of a group is subjected to treatment and control. A paired experiment may be analyzed as a single sample of the differences (Moses, 1985), leading to the mean of differences ($\mathrm{AVE}_{\mathrm{d}}$), and its robust counterpart, the median of differences ($\mathrm{MED}_{\mathrm{d}}$) as measures. Note that the variation of a variable (say, X) is influenced by intra-group variation. Therefore, differences in scale are not defined for paired experiments.

The bootstrap (Efron and Tibshirani, 1993) is the prime source of confidence intervals for the variables measuring differences between treatment and control. This is because (1) it can be used without specifying the data distributions and (2) also properties of rather complex statistical variables which defy theoretical analysis (such as $\mathrm{MED}_{\mathrm{d}}$ in case of two lognormal distributions) can be evaluated by numerical simulation. Bootstrap resampling has been successfully used for constructing confidence intervals for location parameters, with the mean as the most often studied statistic (e.g., Polansky and Schucany, 1997). Thomas (2000) analyzed α-trimmed mean and Huber's proposal 2 as robust location measures. An α-trimmed mean from a sample of size n ignores the extreme $\mathrm{INT}(n\alpha)$ values in each tail, with α chosen between 0 and 0.5 and INT being the integer function, and takes the mean of the remainder. (Note that in Section 3 we use α instead to denote coverage.) Huber's proposal 2 employs a more flexible weighting scheme to reduce the influence of the extreme values in the tails and achieve robustness. The variance/standard deviation is the most often studied measure of scale in bootstrap confidence interval construction for one sample (e.g., Frangos and Schucany, 1990). For the purpose of comparing two samples, Zhou et al. (2001) have found reasonably accurate results using the bootstrap when compared with other statistical tests of equality of means, similarly did Thorpe and Holland (2000) in case of testing equality of variances. Tu and Zhou (2000) analyzed d_{AVE} and bootstrap-t confidence intervals in the presence of skewed data and obtained accurate results in a Monte Carlo study. The other difference measures studied in the present paper have, to our knowledge, not been examined previously.

Bootstrap theory has developed several types of confidence intervals (Efron and Tibshirani, 1993), differing in complexity and computational effort, and it is not clear which type is most accurate (i.e., has a coverage which comes closest to the nominal level) or has shortest length for the difference measures. Sample sizes in experiments can be 30 and even lower, which, together with high skewness, can have a considerable effect on coverage accuracy as found for one-sample experiments (Porter et al., 1997). Therefore, after explaining the data used in the applications (Section 2) and the types of bootstrap confidence intervals (Section 3) in a manner aimed at non-statisticians, we describe a Monte Carlo study (Section 4) conducted to estimate coverage accuracy and length of the intervals in dependence on data sizes, skewnesses, and other distribution parameters. Thereby, lognormal distributions for treatment and control were employed. Logarithmic transformations of treatment (x) and control (y) data prior to the estimations were avoided for the following reasons: (1) difference measures carry other information. For example, $\widehat{d_{\mathrm{MED}}}$, calculated from the transformed data, measures the ratio, not the difference of medians of untransformed distributions. (2) In the general case, the distributional shape is a priori unknown. The lognormal is used here only as an example of a right-skewed distribution. (3) Different types of transformations might be necessary for data from different experiments, which complicate interpretation of results. The practical issue of the number of bootstrap resamplings necessary for suppressing resampling noise is examined (Section 4). Results of applications are presented in Section 5.

2. Data

2.1. *Physiological source–sink manipulation experiments with sunflower*

Sunflower is an important crop and a renewable resource grown for edible oil and increasingly for technical purposes. The oil yield depends on several characteristics, among them the number of filled grains per unit area (Cantagallo et al., 1997) and the dry mass of grains. Unfilled grains, which commonly occur in mature plants, diminish the yield. Several reasons for disturbed grain filling have been discussed, for example, shortage of water and mineral nutrients (for review, see Connor and Hall, 1997). Source limitation (Patrick, 1988) is a further possible cause for reduced grain filling. It means that the amount of photoassimilates exported by the green leaves (source) is not sufficient to fill all grains (sink).

We analyze data from two of our experiments in the field, carried out to study grain filling in sunflower in relation to source–sink ratio. In the first experiment, source–sink ratio was increased by shading plants for 10 days during the stage of floret initiation. Shading reduced the number of florets but did not influence the number of leaves. This treated group of plants was compared with the control (unshaded) group, with possible effects of inter-plant variability (unpaired experiment). In the second experiment, defoliation, source–sink ratio was decreased by excising all leaves on one side of each plant within a group 3–4 days before the inflorescence (capitulum, ''head'') opened. Thereby we drew advantage of the fact that, in an intact sunflower plant, a leaf supplies photoassimilates to a defined sector of the capitulum (Alkio et al., 2002). This allowed the use of a single plant for treatment (defoliation) and control, avoiding effects of inter-plant variability (paired experiment).

2.2. Disinfection effect of predatory microorganisms in wastewater

Many large cities in coastal areas practice ocean disposal of their sewage through marine outfall systems (Yang, 1995). At those places, disinfection of the wastewater is of environmental relevance. An important chemical disinfectant is chlorine, which, however, itself poses a risk to the environment. The capability of other potential, natural ''disinfectants'' like predatory microorganisms in the water is therefore studied.

Yang et al. (2000) conducted a series of experiments with mixtures of artificial seawater and wastewater taken from a Taiwan sewage treatment plant. We analyze their data on the influence of predatory microorganisms. The samples for this paired experiment were prepared as follows. The seawater–wastewater mixture was sterilized/not sterilized to produce control/treatment samples without/with predators. As an indicator of the disinfection strength, Yang et al. (2000) employed the die-off rate of the test organism *Escherichia coli*. These bacteria were added to the flasks, in which their initial concentrations were controlled to lie in a range from 10^7 to 10^8 cfu per 100 ml. The die-off rate is the inverse of the mean lifetime in the exponential formula describing the decay of the number of bacteria with time; it was determined (Yang et al., 2000) using bacterial counts performed over time. Each pair of samples (treatment, control) had the same values in other variables such as salinity, mixing ratio or temperature. Table 2 in the paper by Yang et al. (2000) contains the data.

The authors employed a paired t-test under the Gaussian assumption and found a disinfecting effect of predatory microorganisms. In other words, they concluded that $\mathrm{AVE_d}$, the average of the difference in die-off rate (''with predators'' minus ''without predators'') is significantly greater than zero.

3. Bootstrap confidence intervals

The nonparametric bootstrap (Efron, 1979) is used to estimate standard errors of the difference measures. In case of an unpaired experiment, draw with replacement a bootstrap sample of same size $\{x^*(i),\quad i=1,\ldots,n_x\}$ from the treatment data set $\{x(i),\quad i=1,\ldots,n_x\}$, analogously draw a bootstrap control sample $\{y^*(i),\quad i=1,\ldots,n_y\}$ from the control data set. In case of a paired experiment, draw pairs $\{x^*(i),\ y^*(i),\quad i=1,\ldots,n\}$. Calculate the bootstrap replication of a difference measure from the bootstrap samples. For example, $\widehat{d_{\mathrm{AVE}}}^* = \mathrm{AVE}(x^*) - \mathrm{AVE}(y^*)$ where $\mathrm{AVE}(x^*)$ is the sample mean of x^*. Bootstrap replicates of other measures are calculated in a similar manner. Repeat the procedure by resampling and calculating until B bootstrap replications exist for each measure. The bootstrap estimate of standard error, $\widehat{\mathrm{se}}$, is calculated as the sample standard deviation of the bootstrap replications. For example,

$$\widehat{\mathrm{se}}\left(\widehat{d_{\mathrm{AVE}}}\right) = \left\{\sum_{b=1}^{B}\left(\widehat{d_{\mathrm{AVE}}}^{*b} - \left\langle\widehat{d_{\mathrm{AVE}}}^{*}\right\rangle\right)^2 / (B-1)\right\}^{1/2},$$

where $\left\langle\widehat{d_{\mathrm{AVE}}}^{*}\right\rangle = \sum_{b=1}^{B}\widehat{d_{\mathrm{AVE}}}^{*b}/B$ and $\widehat{d_{\mathrm{AVE}}}^{*b}$ denotes the bth bootstrap replication of $\widehat{d_{\mathrm{AVE}}}$. Bootstrap estimates of standard errors of other difference measures, $\widehat{d_{\mathrm{MED}}}$, $\widehat{d_{\mathrm{STD}}}$, $\widehat{d_{\mathrm{MAD}}}$, $\widehat{\mathrm{AVE_d}}$, and $\widehat{\mathrm{MED_d}}$, are calculated similarly.

The bootstrap replications are used to construct equi-tailed $(1-2\alpha)$ confidence intervals for the estimated difference measures. Two approaches, standard error based and percentile based, dominate theory and practice. The accuracy of the bootstrap method depends critically on the similarity (in terms of standard errors or percentiles) of the distribution of the bootstrap replication and the true distribution. Various concepts exist for accounting the deviations between the two distributions (Efron and Tibshirani, 1993; DiCiccio and Efron, 1996; Davison and Hinkley, 1997; Carpenter and Bithell, 2000). Since it is not clear which type of confidence interval is appropriate for the difference measures we analyze four types: normal, Student's t, percentile, and BCa.

Coverage error is helpful to compare different confidence intervals. Let $\widehat{\Theta}(\alpha)$ be the single endpoint of confidence interval for a quantity of interest, Θ, with nominal one-sided coverage α: $\mathrm{Prob}\left\{\Theta \le \widehat{\Theta}(\alpha)\right\} = \alpha + C$ for all α. If the error, C, is of $\mathcal{O}\left(n^{-1/2}\right)$ where n is the sample size, $\widehat{\Theta}(\alpha)$ is first-order accurate; if C is of $\mathcal{O}(n^{-1})$ then $\widehat{\Theta}(\alpha)$ is second-order accurate. Hall (1988) determined coverage accuracies of bootstrap confidence intervals, which are reported in the following subsections. However, his results apply only to ''smooth functions'' like d_{AVE}, $\mathrm{AVE_d}$, or d_{STD} of a vector mean. Performances of order statistics used here (d_{MED}, $\mathrm{MED_d}$, and d_{MAD}) have to be assessed on the basis of the results of the Monte Carlo study.

3.1. Normal confidence interval

The bootstrap normal confidence interval, in case of difference measure $\widehat{d_{\mathrm{AVE}}}$, is

$$\left[\widehat{d_{\mathrm{AVE}}} - z^{(1-\alpha)}\,\widehat{\mathrm{se}}\left(\widehat{d_{\mathrm{AVE}}}\right), \widehat{d_{\mathrm{AVE}}} + z^{(1-\alpha)}\,\widehat{\mathrm{se}}\left(\widehat{d_{\mathrm{AVE}}}\right)\right],$$

where $z^{(1-\alpha)}$ is the $100(1-\alpha)$th percentile point of the standard normal distribution. For example, $z^{(0.95)} = 1.645$.

3.2. Student's t confidence interval

The bootstrap Student's t confidence interval, in the case of $\widehat{d_{\mathrm{AVE}}}$, is

$$\left[\widehat{d_{\mathrm{AVE}}} - t_{\nu}^{(1-\alpha)}\widehat{\mathrm{se}}\left(\widehat{d_{\mathrm{AVE}}}\right), \widehat{d_{\mathrm{AVE}}} + t_{\nu}^{(1-\alpha)}\widehat{\mathrm{se}}\left(\widehat{d_{\mathrm{AVE}}}\right)\right],$$

where $t_{\nu}^{(1-\alpha)}$ is the $100(1-\alpha)$th percentile point of Student's t-distribution with ν degrees-of-freedom (Table 1). Unless the true distribution for the used measure is normal, bootstrap normal and bootstrap Student's t confidence intervals are first-order accurate.

3.3. Percentile confidence interval

The bootstrap percentile confidence interval, in the case of $\widehat{d_{\mathrm{AVE}}}$, is

$$\left[\widehat{d_{\mathrm{AVE}}}^{*(\alpha)}, \widehat{d_{\mathrm{AVE}}}^{*(1-\alpha)}\right],$$

that is, the interval between the 100αth percentile point and the $100(1-\alpha)$th percentile point of the bootstrap distribution of $\widehat{d_{\mathrm{AVE}}}^{*}$. Since an infinite number, B, of replications is necessary to obtain an exact bootstrap percentile confidence interval, in practice (finite B) an approximate interval is employed. In Section 4, we evaluate the appropriate value of B in case of the difference measures using Monte Carlo simulations.

3.4. BCa confidence interval

The bootstrap bias-corrected and accelerated (BCa) confidence interval, in the case of $\widehat{d_{\mathrm{AVE}}}$, is

$$\left[\widehat{d_{\mathrm{AVE}}}^{*(\alpha 1)}, \widehat{d_{\mathrm{AVE}}}^{*(\alpha 2)}\right],$$

where

$$\alpha 1 = \Phi\left[\widehat{z_0} + \frac{\widehat{z_0} + z^{(\alpha)}}{1 - \widehat{a}\{\widehat{z_0} + z^{(\alpha)}\}}\right],$$

$$\alpha 2 = \Phi\left[\widehat{z_0} + \frac{\widehat{z_0} + z^{(1-\alpha)}}{1 - \widehat{a}\{\widehat{z_0} + z^{(1-\alpha)}\}}\right]. \qquad (1)$$

Bias correction, $\widehat{z_0}$, is computed as

$$\widehat{z_0} = \Phi^{-1}\left(\frac{\text{number of replications where } \widehat{d_{\mathrm{AVE}}}^{*b} < \widehat{d_{\mathrm{AVE}}}}{B}\right).$$

Acceleration, $\widehat{a}$, can be computed (Efron and Tibshirani, 1993) as

$$\widehat{a} = \frac{\sum_{i=1}^{n_x}\sum_{j=1}^{n_y}\left\{\left\langle \widehat{d_{\mathrm{AVE}}}_{(i,j)}\right\rangle - \widehat{d_{\mathrm{AVE}}}_{(i,j)}\right\}^3}{6\left[\sum_{i=1}^{n_x}\sum_{j=1}^{n_y}\left\{\left\langle \widehat{d_{\mathrm{AVE}}}_{(i,j)}\right\rangle - \widehat{d_{\mathrm{AVE}}}_{(i,j)}\right\}^2\right]^{3/2}}, \qquad (2)$$

where $\widehat{d_{\mathrm{AVE}}}_{(i,j)}$ is the jackknife value of $\widehat{d_{\mathrm{AVE}}}$. That is, let $x_{(i)}$ denote the original treatment sample with point $x(i)$ removed and $y_{(j)}$ the control sample without $y(j)$, then $\widehat{d_{\mathrm{AVE}}}_{(i,j)} = \mathrm{AVE}\{x_{(i)}\} - \mathrm{AVE}\{y_{(j)}\}$ where $\mathrm{AVE}\{x_{(i)}\}$ is the sample mean calculated without point $x(i)$. The mean, $\langle \widehat{d_{\mathrm{AVE}}}_{(i,j)}\rangle$, is given by $\{\sum_{i=1}^{n_x}\sum_{j=1}^{n_y}\widehat{d_{\mathrm{AVE}}}_{(i,j)}\}/(n_x n_y)$.

We have extended Efron and Tibshirani's (1993) one-sample recipe for computing $\widehat{a}$ to two-sample experiments. Note that other extension methods to compute the acceleration exist, such as using separate sums over i and j instead of the nested sums used in Eq. (2). A Monte Carlo experiment (results not shown), employing separate sums and otherwise unchanged conditions in comparison with the experiment shown in Fig. 3, yielded only very small differences (vs. nested sums) in case of robust measures and small differences in case of non-robust measures, not affecting the main conclusions of this paper.

Whereas bootstrap percentile confidence intervals are first-order accurate, bias correction (i.e., shifting) and acceleration (i.e., scaling) make BCa intervals second-order accurate.

3.5. Remarks

Bootstrap normal as well as bootstrap Student's t confidence intervals are symmetric about the estimate which may produce unrealistic results, if the estimate sits at the edge of its permissible range. Further, they can exhibit substantial coverage errors for non-normally distributed replications. In particular, Hall (1988) demonstrated that the skewness of the sampling distribution of "smooth functions" has a major effect on the coverage accuracy.

Bootstrap percentile confidence intervals, while avoiding such deficiencies, can have profound coverage errors when the distribution of replications differs considerably from the (in practice unknown) distribution of the estimate. The BCa method accounts for such differences by adjusting for bias and scaling (see Efron and Tibshirani, 1993). However, Davison and Hinkley (1997) pointed out that BCa confidence intervals produce larger coverage errors for $\alpha \rightarrow 0$ (for which the right-hand side of Eq. (1) → $\Phi(\widehat{z_0} - 1/\widehat{a}) \neq 0$) for $\alpha \rightarrow 1$. Finally, Polansky (1999) found that percentile-based bootstrap confidence intervals have intrinsic finite sample bounds on coverage.

Other, computing-intensive methods (Efron and Tibshirani, 1993; Davison and Hinkley, 1997) are briefly mentioned. Boot strap-t confidence intervals are formed using the standard error, $\widehat{\mathrm{se}}^{*}$, of a single bootstrap replication. For simple quantities like the mean, plug-in estimates can be used for $\widehat{\mathrm{se}}^{*}$. However, for complicated quantities (as in Table 1), no plug-in estimates are available. A second bootstrap loop (bootstrapping from bootstrap samples) had to be invoked which would mean a marked increase in computing time. Similarly, bootstrap calibration or double bootstrap methods invoke a second loop.

Tu and Zhou (2000) analyzed d_{AVE} for lognormally distributed data. In that case, a plug-in estimate for $\widehat{\mathrm{se}}^*$ is available which allowed the use of bootstrap-t intervals without a second loop of calculation. They further employed parametric resampling (from a lognormal with estimated parameters) and found good coverage performance in a Monte Carlo study. The emphasis of the present paper is to study how various robust/non-robust measures of differences in location and scale perform in dependence on data sizes. The data are assumed to exhibit skewness, but not to be distributed in an a priori known way. Thus we use nonparametric resampling.

We also avoid data transformations. In the Monte Carlo study, x and y are taken from lognormally distributed populations. Since the distribution function of the difference between two lognormal distributions cannot be obtained analytically, it is not possible to give a transformation of $\widehat{\mathrm{AVE}_{\mathrm{d}}}$ or $\widehat{\mathrm{MED}_{\mathrm{d}}}$ which would result in advantageous, normally distributed measures.

4. Monte Carlo study

Monte Carlo simulations were carried out for studying coverage performances and lengths of the bootstrap confidence interval types described in Section 3. Simulated treatment and control data were taken from lognormal distributions: $X \sim \mathrm{LN}(a_1, b_1, s_1)$ where a_1, b_1, and s_1 are location, scale, and shape parameters, respectively; analogously, $Y \sim \mathrm{LN}(a_2, b_2, s_2)$; the correlation between X and Y is denoted as ρ_{LN} (see Appendix 1). The lognormal distribution is commonly found in natural sciences (Aitchison and Brown, 1957).

Table 2 lists the Monte Carlo designs. Lognormal parameters s_1 and s_2 are restricted to two values ("small", "large"), thereby encompassing the values found in the applications (Section 5). Each design was studied for various combinations of n_x and $n_y \in \{5, 10, 20, 50, 100\}$. Other lognormal designs (same as designs 1 and 2, with b_1 and/or b_2 set to 25.3) were also studied.

Table 2
Monte Carlo designs: lognormal parameters and theoretical difference measures

Design	a_1	b_1	s_1	a_2	b_2	s_2	d_{AVE}	d_{MED}	d_{STD}	d_{MAD}	$\mathrm{MED}_{\mathrm{d}}$
1	0	4	0.2	0	4	0.2	0	0	0	0	0
2	0	4	0.8	0	4	0.2	1.43	0	4.39	2.17	0

Note: $\mathrm{AVE}_{\mathrm{d}} = d_{\mathrm{AVE}}$.

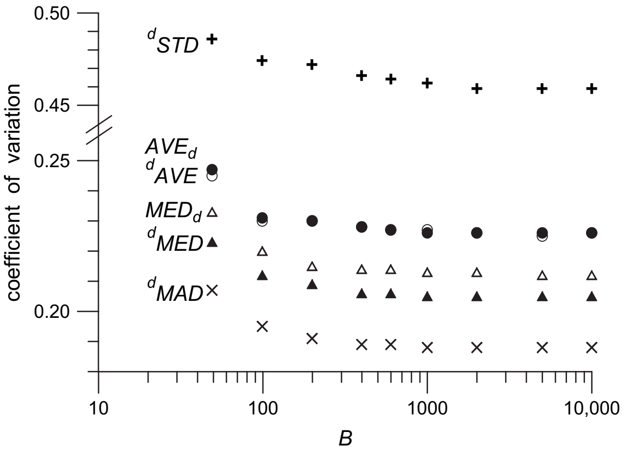

Fig. 1. Coefficient of variation (standard deviation divided by the mean) of the 97.5% quantile (upper bootstrap percentile confidence bound) in dependence on the number of bootstrap resamplings, B, for $a_1 = 0$, $b_1 = 1.0$, $s_1 = 1.0$, $a_2 = 0$, $b_2 = 0.5$, $s_2 = 0.5$, $n_x = 100$, $n_y = 100$, $\rho_{\mathrm{LN}} = 0$, and $n_{\mathrm{sim}} = 10{,}000$.

Values for α were set to 0.025, 0.05, and 0.1 for which $n_{\mathrm{sim}} = 10{,}000$ simulations (simulated pairs of treatment and control data sets) yield a reasonably small standard error, σ, of nominal coverages $\left(\sigma = \sqrt{\alpha(1-\alpha)/n_{\mathrm{sim}}} \leq 0.003\right)$. To assess the bootstrap resampling error, due to finite B, a preliminary Monte Carlo study was carried out. Fig. 1 shows the coefficient of variation of the 97.5% quantile (upper bootstrap percentile confidence bound) approaching saturation for $B \geq B_{\mathrm{sat}} \approx 2000$ in case of $\widehat{d_{\mathrm{STD}}}$ and $B \geq B_{\mathrm{sat}} \approx 1000$ in case of the other difference measures. Design and choice of n_x and n_y in Fig. 1 constitute a conservative approach as regards choice of B: using (1) smaller values for n_x and n_y, (2) smaller scale and shape parameters b_1, b_2, s_1, and s_2, (3) larger α, or (4) bootstrap normal confidence intervals gave lower B_{sat} values while using (5) $a_1 = 1.0$, or (6) $\rho_{\mathrm{LN}} = 0.6$ produced similar B_{sat} values (not shown). We thus conclude that using $B = 1999$ suppressed reasonably well resampling noise in the Monte Carlo study, in agreement with the general recommendation of Efron and Tibshirani (1993).

Figs. 2 and 3 show the results (empirical vs. nominal coverages) of the Monte Carlo study for unpaired and correctly specified ($\rho_{\mathrm{LN}} = 0$) experiments. In general, the empirical coverages approach their nominal levels as the sample sizes (n_x, n_y) increase, as expected. However, the decrease of coverage error is not monotonic, as the "jumps" between $n_x = n_y$ and $n_x \neq n_y$ reveal.

The $\widehat{d_{\mathrm{AVE}}}$ measure of difference in location is worse (in terms of coverage error) than $\widehat{d_{\mathrm{MED}}}$ in nearly all cases, especially when treatment and control populations differ in the lognormal parameters shape (s_1 vs. s_2, Fig. 3), scale (b_1 vs. b_2, results not shown), or both (results not shown). Even more pronounced are the deficiencies of $\widehat{d_{\mathrm{STD}}}$ as measure of difference in scale in comparison with $\widehat{d_{\mathrm{MAD}}}$, with similar dependences on the lognormal parameters as in the case of $\widehat{d_{\mathrm{AVE}}}$.

The influence of lognormal parameters b_1, b_2, s_1, and s_2 on the coverage error of $\widehat{d_{\mathrm{MED}}}$ seems to be relatively weak. The coverage error of $\widehat{d_{\mathrm{MAD}}}$ is more strongly affected, especially when treatment and control differ in one of the parameters (or in both). In such cases (design 2), normal or Student's t confidence intervals exhibit higher coverage errors than BCa intervals. This seems to be because of the high degree of skewness (data not shown) of the bootstrap distribution of $\widehat{d_{\mathrm{MAD}}}$ and $\widehat{d_{\mathrm{MED}}}$. Additional Monte Carlo experiments with $a_1 = 1.0$, $b_1 = 4.0$, $s_1 = 0.2$, $a_2 = 0$, $b_2 = 4.0$, $s_2 = 0.2$,

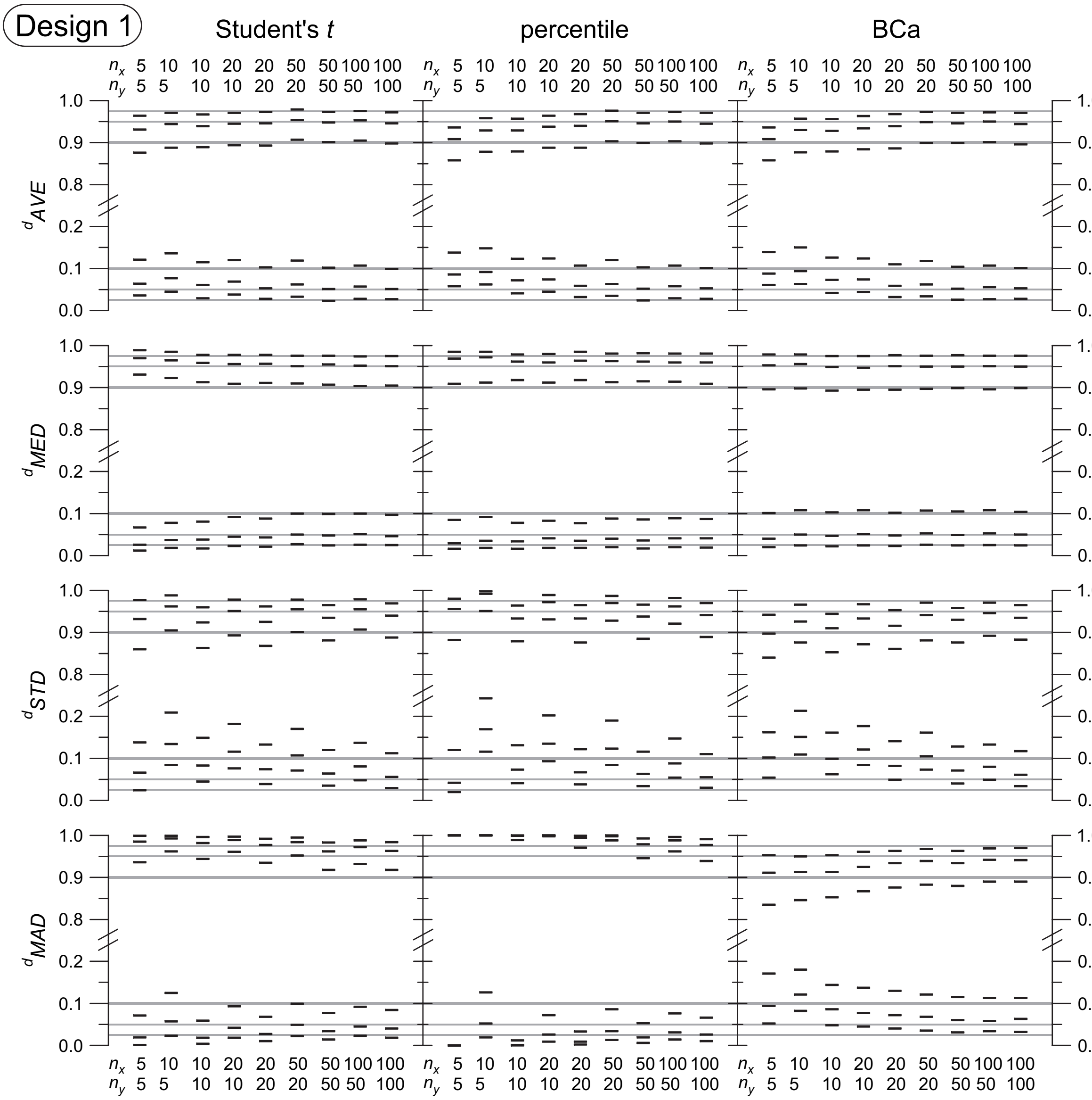

Fig. 2. Monte Carlo results, unpaired experiment ($\rho_{LN} = 0$), design 1: empirical coverages ($\alpha = 0.025$, 0.05, and 0.1; upper and lower limits) in dependence on sample sizes; nominal levels as horizontal lines (thickness corresponds to nominal standard error). Note: *y*-axis measures coverages, *x*-axis data sizes; axes titles denote columns and rows of panels.

$\rho_{LN} = 0$, and 0.6, produced nearly identical coverage errors (results not shown) as the experiments with design 1.

In general, differences in empirical coverage between Student's *t* and normal confidence intervals were negligible in the Monte Carlo study (therefore, results of normal intervals are not shown). On the other hand, BCa intervals yielded considerably lower coverage errors than percentile intervals, especially for measure $\widehat{d_{MAD}}$ (for data sizes $\gtrsim 20$).

Of practical importance is to decide where to use BCa and where to use Student's *t* intervals. In case of difference in location, we favor $\widehat{d_{MED}}$ as measure and the BCa confidence interval. For $n_x \gtrsim 20$ and $n_y \gtrsim 20$, this choice provides coverages which seem to be reasonably low in error and also robust against variations in lognormal parameters (at least over the ranges investigated here, see Table 2).

In the case of difference in scale, we favor $\widehat{d_{MAD}}$ as measure and the BCa confidence interval. For data sizes $\gtrsim 50$, this provides acceptable coverage errors and clearly better robustness against variations in lognormal parameters than Student's *t* intervals.

Fig. 4 shows the result for unpaired ($\widehat{d_{AVE}}$, $\widehat{d_{MED}}$) misspecified ($\rho_{LN} = 0.6$, 0.9) experiments. The coverage error is serious already for $\rho_{LN} = 0.6$ in all cases. The coverage error does not approach zero with sample size. Similar behavior was found for other Monte Carlo designs (not shown). Fig. 4 further shows the result for paired ($\widehat{\mathrm{AVE}}_d$, $\widehat{\mathrm{MED}}_d$) experiments with $\rho_{LN} = 0.6$ and 0.9. $\widehat{\mathrm{MED}}_d$ as a robust measure of location yields smaller coverage errors than $\widehat{\mathrm{AVE}}_d$ for all investigated interval types, sample sizes $n \geq 10$, and also for other analyzed Monte Carlo designs (not shown here).

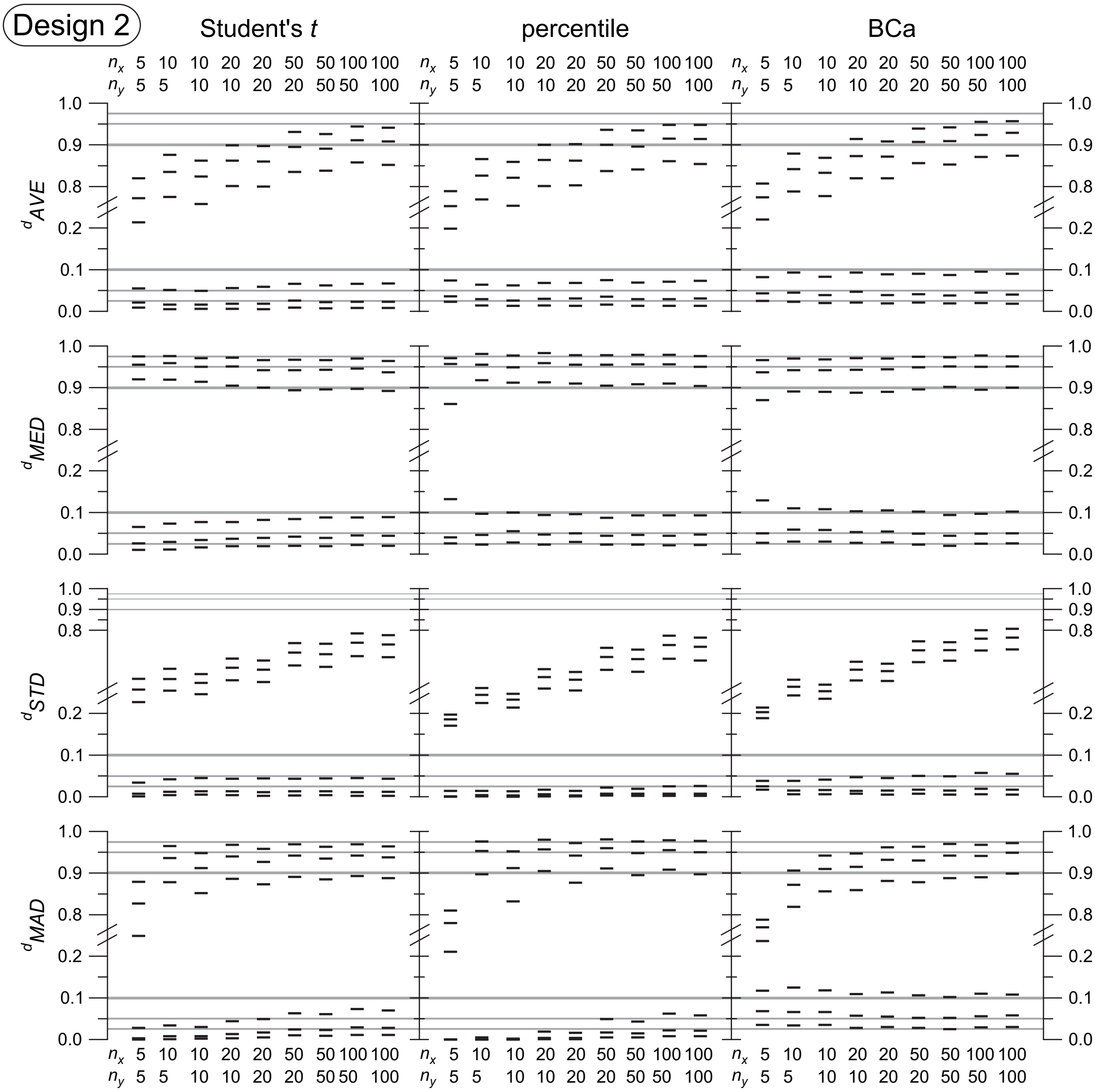

Fig. 3. Monte Carlo results, unpaired experiment ($\rho_{LN} = 0$), design 2 (cf. Fig. 2).

For $\widehat{MED}_d$, dependence on n is rather weak; already 10 treatment/control data pairs produce empirical coverages close to the nominal levels (Student's t and BCa confidence intervals). Nearly identical results are obtained either for $\rho_{LN} = 0.6$ or 0.9 (Fig. 4) or for $\rho_{LN} = 0$ (results not shown).

Fig. 5 shows the resulting average confidence interval lengths for unpaired experiments (designs 1 and 2). As expected, the length decreases with the sample sizes. This decrease has noticeable ''jumps'' between $n_x = n_y$ and $n_x \neq n_y$ in the case of design 2. The robust measures produce clearly wider (by a factor of approximately 1.5) intervals than their non-robust counterparts. In cases of only minor deviations of the data distributions from the normal shape, the non-robust measures might therefore be preferred. For the analysis of agricultural data here, we use the robust versions because these data exhibit considerable amounts of skewness (Fig. 7). The average interval length depends only weakly on interval type, especially for data sizes above 10 (Fig. 5). Similar findings were obtained in the Monte Carlo simulations for paired experiment (Fig. 6).

5. Applications

5.1. Physiological source–sink manipulation experiments with sunflower

In both experiments (shading and defoliation), sunflower hybrid ''Rigasol'' was used at a low density of 1.33 plants/m^2 to minimize inter-plant competition. The overall effect of shading on the source–sink ratio was quantified by dividing the total green leaf area at the end of flowering by the total number of florets. Treated and control plants had ratios of 12.5 and 9.0 cm^2 per floret (medians), respectively. After physiological

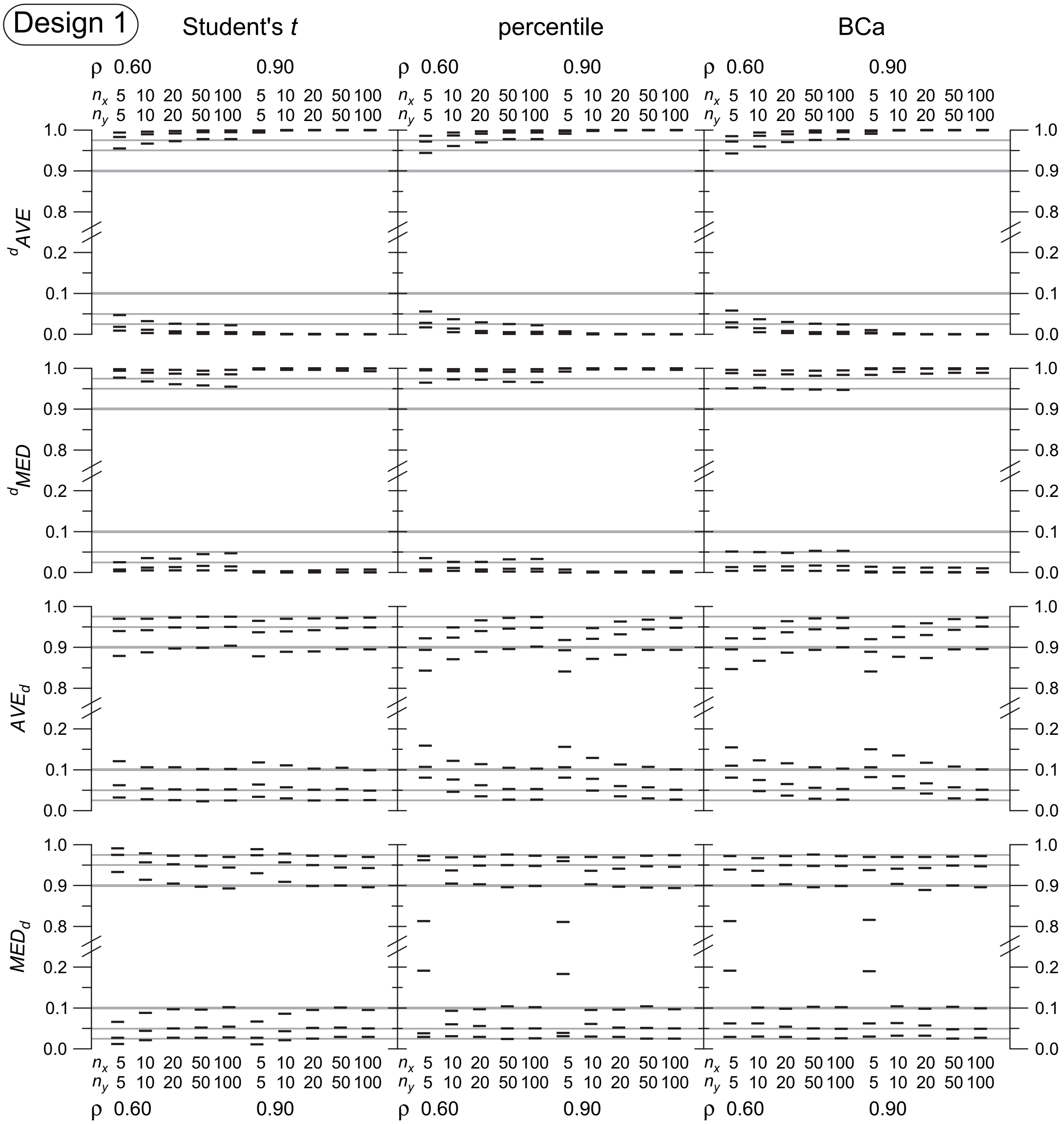

Fig. 4. Monte Carlo results: empirical coverages, mis-specified unpaired and paired experiments, design 1 (cf. Fig. 2).

maturity, all grains from each plant were sorted (filled, unfilled), counted, and average dry mass per grain was determined. In the defoliation experiment, mature sunflower capitulae were divided into a treated 1/4-sector (corresponding to the defoliated stem side, see Alkio et al., 2002) and a control 1/4-sector (located opposite of the treated sector). Grains from both sectors were sorted, counted, and average dry mass per grain was determined.

Fig. 7 reveals that the data distribution for percentages of unfilled grains is roughly lognormal in shape, similarly for dry mass per grain data (not shown). Estimated lognormal parameters lie in the ranges for the Monte Carlo designs (Section 4). Data sizes are large enough (see Section 4) to allow estimation of reliable confidence intervals for difference measures in location, and acceptably accurate confidence intervals for differences in scale.

Fig. 8 shows the results, that is, the estimated measures of difference in location and scale, for the sunflower data (percentages of unfilled grains). There is overlap between Student's *t* and BCa ($\alpha = 0.025$) confidence intervals, as well as some amount of agreement between the robust and non-robust measures ($\widehat{d_{\text{MED}}}$ vs. $\widehat{d_{\text{AVE}}}$, $\widehat{d_{\text{MAD}}}$ vs. $\widehat{d_{\text{STD}}}$, and $\widehat{\text{MED}_{\text{d}}}$ vs. $\widehat{\text{AVE}_{\text{d}}}$).

In the shading experiment (unpaired), the confidence intervals indicate that the differences in location between treatment and sample are significant in the sense that the intervals do not contain zero. One-sided Wilcoxon tests conclude the same. Increasing the source–sink ratio (by shading during floret initiation) reduced the percentage of unfilled grains in sunflower.

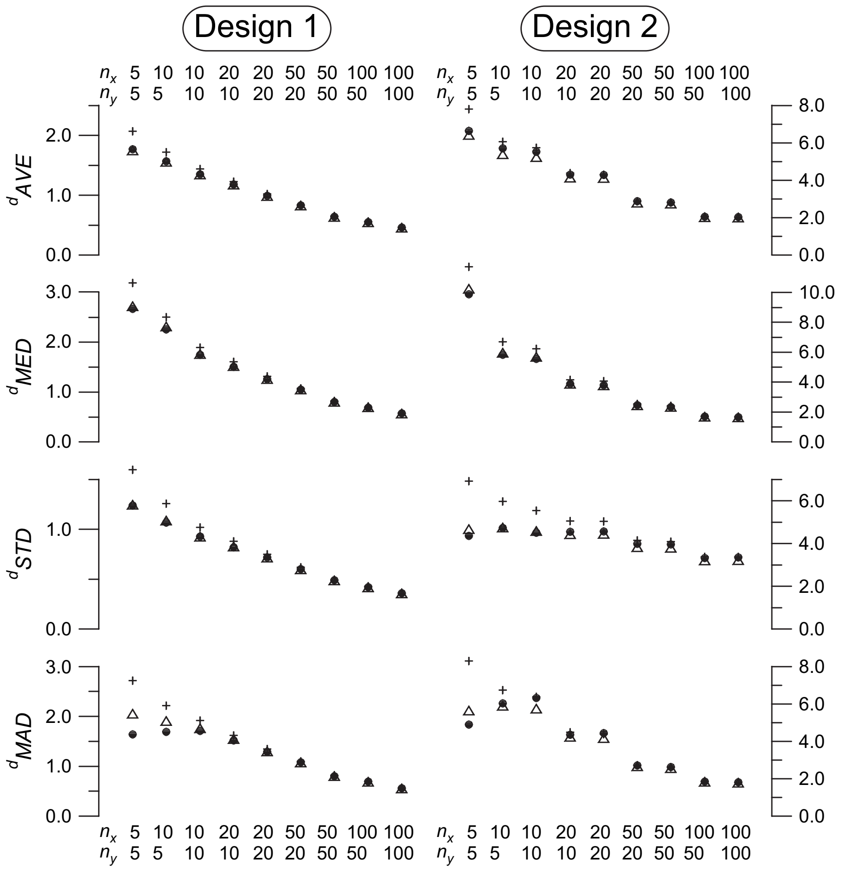

Fig. 5. Monte Carlo results, unpaired experiment ($\rho_{\mathrm{LN}} = 0$), designs 1 and 2: average confidence interval lengths in dependence on sample sizes ($\alpha = 0.025$; Student's t, crosses; percentile, triangles; and BCa, circles). Note: x-axis shows data sizes.

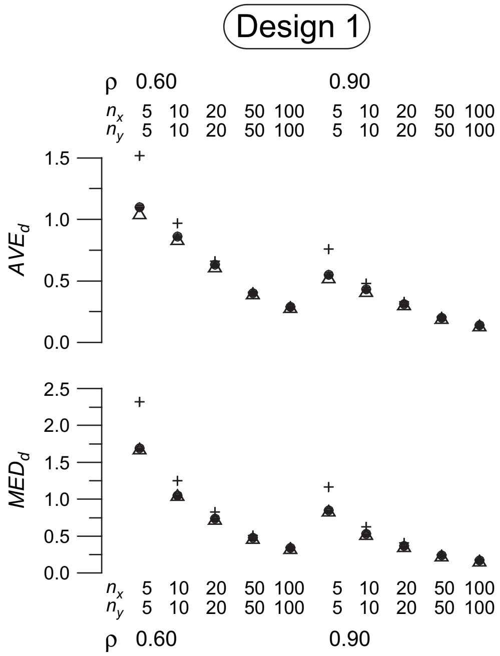

Fig. 6. Monte Carlo results: average confidence interval lengths, mis-specified unpaired and paired experiments, design 1 (cf. Fig. 5).

The further information provided by using $\widehat{d_{\mathrm{MED}}}$ is that this reduction equals −2.7% points with 95% BCa confidence interval (−4.1%, −0.9%). The shading-induced increase in dry mass per single, filled grain is 16.8 mg ($\widehat{d_{\mathrm{MED}}}$) with 95% BCa confidence interval (12.1 mg, 20.7 mg). The difference in scale (variability) in the shading experiment, however, is not significantly different from zero, suggesting that shading influenced plant growth only via increasing the source–sink ratio.

In the defoliation experiment (paired), the difference (treatment vs. control) in location is significantly greater than zero (confirmed by Wilcoxon tests). Decreasing the source–sink ratio (by removing supplying leaves) increased the percentage of unfilled grains in sunflower by 6.4% ($\widehat{\mathrm{MED}_{\mathrm{d}}}$) with 95% BCa confidence interval (5.1%, 9.0%). The defoliation-induced decrease in dry mass per single, filled grain is −6.9 mg ($\widehat{\mathrm{MED}_{\mathrm{d}}}$) with 95% BCa confidence interval (−9.6 mg, −4.6 mg).

These results indicate that grain filling in sunflower is controlled by the amount of photoassimilates available and the number of grains, that means, the source–sink ratio. See Alkio et al. (2003), where further experimental data are analyzed.

5.2. Disinfection effect of predatory microorganisms in wastewater

Yang et al. (2000) determined die-off rates of *E. coli* for 20 pairs of water samples. The authors removed two pairs,

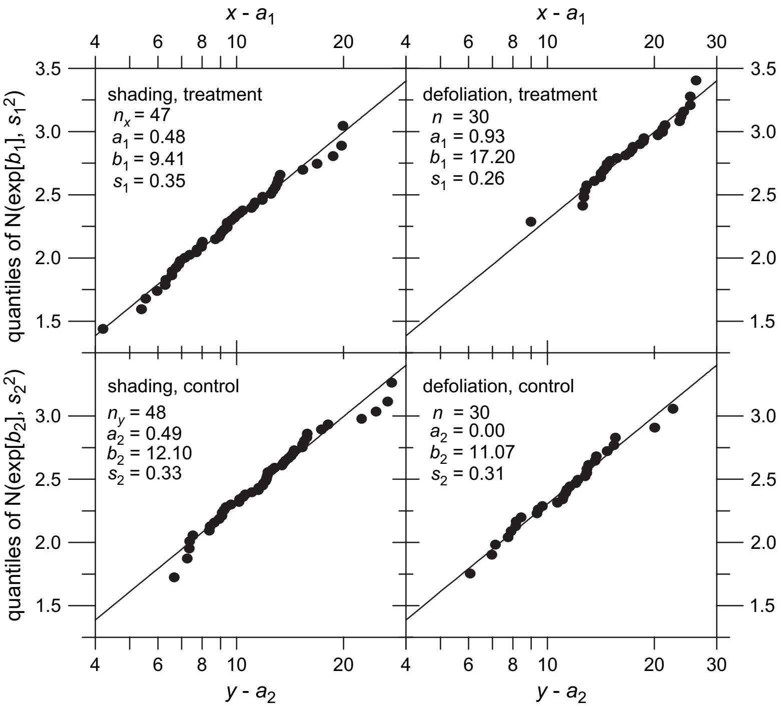

Fig. 7. Application to sunflower experiments (x, y = percentage of unfilled grains); shading (unpaired) and defoliation (paired); data and theoretical lognormal curves (straight lines) (note logarithmic x axes); correlation coefficient (treatment vs. control) for defoliation: 0.26; estimation of a_1 by brute force minimization of $\left[\text{median}(x) - \left(\widehat{a_1} + \widehat{b_1}\right)\right]$, analogously for $\widehat{a_2}$. Deviations from normal shape are as follows. Shading, treatment: skewness = 1.03, kurtosis = 0.66; shading, control: skewness = 1.40, kurtosis = 1.87; defoliation, treatment: skewness = 0.28, kurtosis = −1.05; defoliation, control: skewness = 0.99, kurtosis = 1.07.

for which they could not exclude experimental errors, as outliers. The remaining $n = 18$ pairs yield $\widehat{d_{\text{AVE}}} = 0.0126\ \text{min}^{-1}$, that means, the die-off rate of *E. coli* is larger for water samples with existing predatory microorganisms than for samples without. Yang et al. (2000) analyzed $\widehat{d_{\text{AVE}}}$ with a paired t-test and found that the difference is significant at the 98.15% level. They further gave a 95% confidence interval for $\widehat{d_{\text{AVE}}}$ of (0.0024 min^{-1}; 0.0228 min^{-1}).

Our analysis using 2SAMPLES (Appendix 2) basically confirmed Yang et al.'s (2000) analysis of $\widehat{d_{\text{AVE}}}$. The bootstrap Student's t 95% confidence interval is (0.0027 min^{-1}; 0.0225 min^{-1}), but this may be slightly down-biased, as indicated by the bootstrap BCa interval of (0.0060 min^{-1}; 0.0269 min^{-1}). We also confirmed the significance level of 98.15% by employing a version of 2SAMPLES with $\alpha = 0.00925$, 0.00900, and 0.00875. The correlation coefficient (treatment vs. control) is 0.22. However, usage of $\widehat{d_{\text{AVE}}}$ might be inappropriate for these data because they exhibit considerable amounts of skewness (treatment, 1.8; control, 1.2). Taking $\widehat{d_{\text{MED}}}$ instead of $\widehat{d_{\text{AVE}}}$ gave another result (Fig. 9). This robust measure of difference in location had the following confidence intervals: bootstrap Student's t, (−0.0051 min^{-1}; 0.0150 min^{-1}) and bootstrap BCa, (−0.0008 min^{-1}; 0.0166 min^{-1}). Both types of intervals for $\widehat{d_{\text{MED}}}$ indicate a non-significant effect. We do not want to dispute the finding of Yang et al. (2000) that predatory microorganisms in wastewater have a disinfecting effect, but we think that more data and experiments are required to evaluate whether this is a valid conclusion.

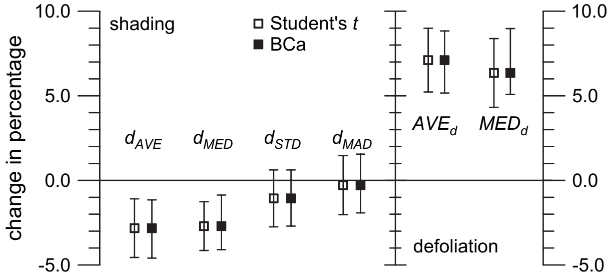

Fig. 8. Sunflower experiments; resulting estimates and 95% confidence intervals for difference measures (treatment vs. control) in percentage of unfilled grains.

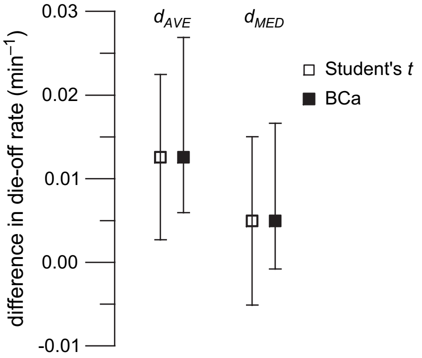

Fig. 9. Wastewater experiments; resulting estimates and 95% confidence intervals for difference measures (treatment vs. control) in die-off rate of *E. coli*.

6. Summary and conclusions

The following measures for difference in location between treatment and control were analyzed:

Unpaired experiment:
- $\widehat{d_{\mathrm{AVE}}}$ (difference of averages),
- $\widehat{d_{\mathrm{MED}}}$ (difference of medians).

Paired experiment:
- $\widehat{\mathrm{AVE}_{\mathrm{d}}}$ (average of differences),
- $\widehat{\mathrm{MED}_{\mathrm{d}}}$ (median of differences).

Likewise the following measures for difference in scale:

Unpaired experiment:
- $\widehat{d_{\mathrm{STD}}}$ (difference of standard deviations),
- $\widehat{d_{\mathrm{MAD}}}$ (difference of MADs).

A Monte Carlo study using bivariate lognormal distributions was carried out to evaluate coverage performances of four types of nonparametric bootstrap confidence intervals for the estimated measures: normal, Student's t, percentile, and BCa.

The robust measures ($\widehat{d_{\mathrm{MED}}}$, $\widehat{\mathrm{MED}_{\mathrm{d}}}$, $\widehat{d_{\mathrm{MAD}}}$) performed better (i.e., had smaller coverage errors) than their non-robust counterparts over the entire investigated space of lognormal parameters, data sizes, and correlation coefficients. On the other hand, the robust measures are less efficient: they produced average confidence interval lengths, which were approximately 1.5 times larger than those of the non-robust measures. The practical implications of the Monte Carlo study are as follows.

- BCa and Student's t confidence intervals of $\widehat{\mathrm{MED}_{\mathrm{d}}}$ as measure of location of difference offer good coverage performance in paired experiments for $n \gtrsim 10$.
- BCa confidence intervals of $\widehat{d_{\mathrm{MED}}}$ as measure of difference in location offer good coverage performance in unpaired experiments for $n_x \gtrsim 20$ and $n_y \gtrsim 20$.
- BCa confidence intervals of $\widehat{d_{\mathrm{MAD}}}$ as measure of difference in scale offer acceptably coverage performance in unpaired experiments for $n_x \gtrsim 50$ and $n_y \gtrsim 50$.
- For applications where data distributions exhibit considerable deviations from the normal shape, it is advised to use the robust measures to achieve a good coverage accuracy. For only minor deviations from the normal shape, the more efficient non-robust measures can be used.

Reliable confidence intervals for smaller data sizes require more complex types of bootstrap confidence intervals, involving a second bootstrap loop (e.g., Efron and Tibshirani, 1993). The Monte Carlo study further revealed that mis-specification (i.e., use of unpaired method for correlated data) leads to rather large coverage errors.

In the analysis of own field experiments with sunflower, we quantified the influence of an altered source–sink ratio on the percentage of unfilled grains and the dry mass per grain. There is evidence that grain filling is significantly controlled by the source–sink ratio. In the analysis of published data from experiments on the disinfection effect of predatory microorganisms in wastewater, we found that the claimed significant effect, based on using $\widehat{d_{\mathrm{AVE}}}$, becomes non-significant when using the more appropriate robust measure, $\widehat{d_{\mathrm{MED}}}$.

In general, the method presented here for quantifying the differences in location and scale between two samples is broadly applicable – to data from fields as biology, agriculture, medicine, and econometrics. It avoids transformations and appears to be robust with respect to the distributional shape.

In particular, this approach to compare two samples using bootstrap confidence intervals can be applied to a range of topics from environmental research raised in recent issues of this journal. For example, Goyal and Sidharta (2004) compared measured and modeled distributions of suspended particulate matter in the area around a thermal power station in India. The demonstration of non-significant differences would lend further credence to their modeling effort. Also, the effects of different implementations of environmental models can be quantified. Giannakopoulos et al. (2004) used a chemical transport model with/without a scheme of mixing processes in the planetary boundary layer (PBL) to study global ozone distributions. The PBL scheme seems to have a significant influence, which could be quantified using bootstrap confidence intervals. Peel et al. (2005) found that the selection of biophysical parameters for eucalypts had negligible influence on the simulation of the January climate of Australia. Such a conclusion could be strengthened by evaluating the differences using bootstrap confidence intervals. Finally, also real-time monitoring systems could benefit from incorporating a "bootstrap tool" into their data analysis. For example, Chrysoulakis et al. (2005) developed a software for low-resolution image analysis to detect the occurrence of major industrial accidents using satellite imagery data. In this case, the control sample would come from the data before the accident and the treatment sample would come from thereafter. The hypothetical time boundary between control and treatment would be varied and significant changes in location tested. The new data coming in would require to update this test permanently. Quinn et al. (2005) designed a system for real-time management of dissolved oxygen in a ship channel in California, which might be further enhanced using data analysis with bootstrap confidence intervals.

Acknowledgements

The Editor-in-Chief and the reviewers made constructive and helpful comments. We thank E. Grimm, M. Knoche (both at Institute of Agronomy and Crop Science, University Halle–Wittenberg, Germany), and Q. Yao (Department of Statistics, London School of Economics, UK) for their comments and advices on aspects of plant physiology and statistics. The experimental data analyzed here and the 2SAMPLES software can be downloaded from http://www.climate-risk-analysis.com.

Appendix 1. Technical details

The Monte Carlo samples x and y (treatment $X \sim \mathrm{LN}(a_1, b_1, s_1)$, control $Y \sim \mathrm{LN}(a_2, b_2, s_2)$) were generated from a binormal distribution $(U_1, U_2) \sim N(\mu_1, s_1, \mu_2, s_2$, correlation $\rho_N)$ by the transformations $x = \exp(U_1 + a_1)$, $y = \exp(U_2 + a_2)$ with $b_1 = \exp(\mu_1)$, $b_2 = \exp(\mu_2)$, using the pseudorandom number generator RAN and routine GASDEV from Press et al. (1996). The correlation coefficient between X and Y is

$$\rho_{LN} = [\exp(\rho_N s_1 s_2) - 1] / \left\{ \left[\exp\left(s_1^2\right) - 1\right]^{1/2} \left[\exp\left(s_2^2\right) - 1\right]^{1/2} \right\}.$$

For lognormally distributed X and Y, theoretical difference measures d_{AVE}, d_{MED}, d_{STD}, and AVE_d are easily obtained from the properties of the lognormal (e.g., Johnson et al., 1994). Calculation of d_{MAD} requires the MAD value for a lognormal distribution, which is given by the minimum of the two solutions of the equation (here for X):

$$\text{erf}\left[\frac{\log(1 + \text{MAD}/b_1)}{\sqrt{2}s_1}\right] - \text{erf}\left[\frac{\log(1 - \text{MAD}/b_1)}{\sqrt{2}s_1}\right] = 1$$

where erf is the error function. For the Monte Carlo simulations, this equation was solved by numerical integration of the normal densities. Calculation of MED_d is hindered by the fact that the density function of the difference of two lognormally distributed variables is in the general case analytically unknown (e.g., Johnson et al., 1994). MED_d was, therefore, determined by numerical simulation which outperformed numerical integration of the density in terms of precision and computing costs.

The normal distribution Φ was approximated using routine ERFCC of Press et al. (1996); the inverse normal distribution was approximated using the routine of Odeh and Evans (1974); the inverse Student's t-distribution was approximated using the formula in Abramowitz and Stegun (1965). Numerical accuracy of theoretical difference measures is evaluated as $<10^{-3}$.

Appendix 2. Description of software

Program title: 2SAMPLES; developer: Manfred Mudelsee; year first available: 2004; hardware requirements: IBM-compatible computer system, 32 MB or more RAM, Pentium type CPU, VGA display; software requirements: Windows 98 system or higher, freeware graphics program Gnuplot (Version 3.6 or higher) residing as "gnuplot.exe" in path; program language: Fortran 90; program size: 1.1 MB; availability and cost: 2SAMPLES is free and, together with Gnuplot, available for download from http://www.climate-risk-analysis.com.

Data format: control and treatment data are in separate ASCII files, with decimal point and one value per line. Data size limitations: virtually none.

After starting the program, input the data file names. 2SAMPLES plots data as histograms, using the class number selector of Scott (1979). If $n_x = n_y$, you can choose between paired and unpaired experiment type, if $n_x \neq n_y$, the type is unpaired. Then input α (0.01, 0.025, 0.05, or 0.1). 2SAMPLES calculates difference measures (paired or unpaired) with $(1 - 2\alpha)$ bootstrap confidence intervals (Student's t, BCa) using $B = 1999$. The result is plotted on screen and also written to output file "2SAMPLES.DAT".

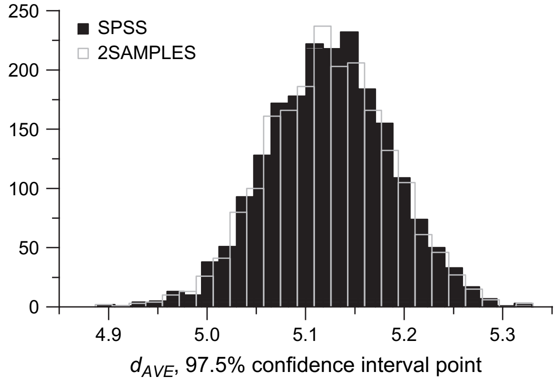

Fig. 10. Software validation; 97.5% confidence interval point for $\widehat{d_{AVE}}$; SPSS vs. 2SAMPLES (Student's t interval type). Unpaired experiment: $X \sim N(5, 2)$, $Y \sim N(0, 2)$, $n_x = n_y = 2000$. Number of comparisons: 2000.

Appendix 3. Software validation

2SAMPLES was subjected to a validation experiment with SPSS for Windows, Version 13.0 (SPSS, Inc., Chicago, IL 60606, USA). 2SAMPLES and SPSS calculate confidence interval points in a different manner: bootstrap vs. Student's t approximation (Norušis, 2004). The following setting was therefore employed to ensure "ideal" conditions for each software: unpaired experiment with X and Y normally distributed; large, equal data sizes; and $\widehat{d_{AVE}}$ as estimated measure of the difference in location. 2SAMPLES' Student's t confidence interval was selected because it corresponds closest to SPSS's Student's t approximation. Significant differences in confidence interval points would then likely indicate a programming error, meaning that the validation has failed.

Two thousands comparisons of the 97.5% confidence interval point were made. The interval points were calculated automatically using a Fortran 90 wrapper (2SAMPLES) and SPSS's Production Mode Facility. Evaluating visually the result of the comparison reveals only minor differences between 2SAMPLES and SPSS (Fig. 10). This was confirmed by subjecting the 2000 differences in confidence interval point to a paired t-test (under SPSS). The mean of the differences is (at five significant digits) equal to zero, the 95% confidence interval for the true difference, SPSS confidence interval point minus 2SAMPLES confidence interval point, is (−0.00009, 0.00008), and the test concludes that the difference is not significant at the 95% level. Using the BCa interval type instead of Student's t gave the same test result, although the 95% confidence interval for the true difference was wider, (−0.00025; 0.00017). We conclude that 2SAMPLES has passed the validation.

References

Abramowitz, M., Stegun, I. (Eds.), 1965. Handbook of Mathematical Functions. National Bureau of Standards, Washington, DC, 1046 pp.

Aitchison, J., Brown, J.A.C., 1957. The Lognormal Distribution. Cambridge University Press, Cambridge, 176 pp.

Alkio, M., Diepenbrock, W., Grimm, E., 2002. Evidence for sectorial photoassimilate supply in the capitulum of sunflower (*Helianthus annuus*). New Phytologist 156, 445–456.

Alkio, M., Schubert, A., Diepenbrock, W., Grimm, E., 2003. Effect of source–sink ratio on seed set and filling in sunflower (*Helianthus annuus* L.). Plant, Cell and Environment 26, 1609–1619.

Cantagallo, J.E., Chimenti, C.A., Hall, A.J., 1997. Number of seeds per unit area in sunflower correlates well with a photothermal quotient. Crop Science 37, 1780–1786.

Carpenter, J., Bithell, J., 2000. Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. Statistics in Medicine 19, 1141–1164.

Chrysoulakis, N., Adaktylou, N., Cartalis, C., 2005. Detecting and monitoring plumes caused by major industrial accidents with JPLUME, a new software tool for low-resolution image analysis. Environmental Modelling and Software 20, 1486–1494.

Connor, D.J., Hall, A.J., 1997. Sunflower physiology. In: Schneiter, A.A. (Ed.), Sunflower Technology and Production. American Society of Agronomy, Crop Science Society of America, Soil Science Society of America, Madison, WI, pp. 113–182.

Davison, A.C., Hinkley, D.V., 1997. Bootstrap Methods and Their Application. Cambridge University Press, Cambridge, 582 pp.

DiCiccio, T.J., Efron, B., 1996. Bootstrap confidence intervals (with discussion). Statistical Science 11, 189–228.

Efron, B., 1979. Bootstrap methods: another look at the jackknife. The Annals of Statistics 7, 1–26.

Efron, B., Tibshirani, R.J., 1993. An Introduction to the Bootstrap. Chapman and Hall, London, 436 pp.

Frangos, C.C., Schucany, W.R., 1990. Jackknife estimation of the bootstrap acceleration constant. Computational Statistics and Data Analysis 9, 271–281.

Giannakopoulos, C., Chipperfield, M.P., Law, K.S., Shallcross, D.E., Wang, K.-Y., Petrakis, M., 2004. Modelling the effects of mixing processes on the composition of the free troposphere using a three-dimensional chemical transport model. Environmental Modelling and Software 19, 391–399.

Gibbons, J.D., 1985. Pitman tests. In: Kotz, S., Johnson, N.L., Read, C.B. (Eds.), Encyclopedia of Statistical Sciences, vol. 6. Wiley, New York, pp. 740–743.

Goyal, P., Sidharta, 2004. Modeling and monitoring of suspended particulate matter from Badarpur thermal power station, Delhi. Environmental Modelling and Software 19, 383–390.

Hall, P., 1988. Theoretical comparison of bootstrap confidence intervals (with discussion). The Annals of Statistics 16, 927–985.

Johnson, N.L., Kotz, S., Balakrishnan, N., 1994. In: Continuous Univariate Distributions, second ed., vol. 1. Wiley, New York, 756 pp.

Moses, L.E., 1985. Matched pairs. In: Kotz, S., Johnson, N.L., Read, C.B. (Eds.), Encyclopedia of Statistical Sciences, vol. 5. Wiley, New York, pp. 287–289.

Norušis, M.J., 2004. SPSS 13.0 Statistical Procedures Companion. Prentice-Hall, Upper Saddle River, NJ, 614 pp.

Odeh, R.E., Evans, J.O., 1974. The percentage points of the normal distribution. Applied Statistics 23, 96–97.

Patrick, J.W., 1988. Assimilate partitioning in relation to crop productivity. HortScience 23, 33–40.

Peel, D.R., Pitman, A.J., Hughes, L.A., Narisma, G.T., Pielke Sr., R.A., 2005. The impact of realistic biophysical parameters for eucalypts on the simulation of the January climate of Australia. Environmental Modelling and Software 20, 595–612.

Polansky, A.M., 1999. Upper bounds on the true coverage of bootstrap percentile type confidence intervals. The American Statistician 53, 362–369.

Polansky, A.M., Schucany, W.R., 1997. Kernel smoothing to improve bootstrap confidence intervals. Journal of the Royal Statistical Society, Series B 59, 821–838.

Porter, P.S., Rao, S.T., Ku, J.-Y., Poirot, R.L., Dakins, M., 1997. Small sample properties of nonparametric bootstrap *t* confidence intervals. Journal of the Air & Waste Management Association 47, 1197–1203.

Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P., 1996. Numerical Recipes in Fortran 90. Cambridge University Press, Cambridge. pp. 935–1446.

Quinn, N.W.T., Jacobs, K., Chen, C.W., Stringfellow, W.T., 2005. Elements of a decision support system for real-time management of dissolved oxygen in the San Joaquin River Deep Water Ship Channel. Environmental Modelling and Software 20, 1495–1504.

Scott, D.W., 1979. On optimal and data-based histograms. Biometrika 66, 605–610.

Thomas, G.E., 2000. Use of the bootstrap in robust estimation of location. Journal of the Royal Statistical Society, Series D 49, 63–77.

Thorpe, D.P., Holland, B., 2000. Some multiple comparison procedures for variances from non-normal populations. Computational Statistics and Data Analysis 35, 171–199.

Tu, W., Zhou, X.-H., 2000. Pairwise comparisons of the means of skewed data. Journal of Statistical Planning and Inference 88, 59–74.

Tukey, J.W., 1977. Exploratory Data Analysis. Addison-Wesley, Reading, MA, 688 pp.

Yang, L., 1995. Review of marine outfall systems in Taiwan. Water Science and Technology 32, 257–264.

Yang, L., Chang, W.-S., Huang, M.-N., 2000. Natural disinfection of wastewater in marine outfall fields. Water Research 34, 743–750.

Zhou, X.-H., Li, C., Gao, S., Tierney, W.M., 2001. Methods for testing equality of means of health care costs in a paired design study. Statistics in Medicine 20, 1703–1720.