Bootstrap Tilting Confidence Intervals and Hypothesis Tests

Tim C. Hesterberg
MathSoft, Inc.
1700 Westlake Ave N., Suite 500
Seattle, WA 98109-3044

Abstract
Bootstrap tilting confidence intervals could be the method of choice in many applications for reasons of both speed and accuracy. With the right implementation, tilting intervals are 37 times as fast as bootstrap BC-a limits, in terms of the number of bootstrap samples needed for comparable simulation accuracy. Thus 100 bootstrap samples might suffice instead of 3700.

Tilting limits have other desirable properties — second-order accuracy, transformation invariance, and better finite-sample coverage and/or shorter intervals on average than competing procedures.

Key Words: bootstrap, importance sampling.

1 Introduction

We begin with a short introduction to the bootstrap; for a more complete introduction see [9].

The original data is $\mathcal{X} = (x_1, x_2, \ldots, x_n)$, a sample from an unknown distribution F, which may be multivariate. Let $\theta = \theta(F)$ be a real-valued functional parameter of the distribution, such as its mean or a regression coefficient, and $\hat{\theta} = \theta(\hat{F})$ the value estimated from the data. We require that θ be a functional statistic, i.e. it depends on the data only through the empirical distribution, with no dependence on sample size or order of the observations. The sampling distribution of $\hat{\theta}$

$$G(a) = P_F(\hat{\theta} \le a) \tag{1}$$

is needed for statistical inference. In simple problems the sampling distribution can be approximated using methods such as the central limit theorem and the substitution of sample moments such as $\overline{x}$ and s into formulas obtained by probability theory. This may not be sufficiently accurate or even possible in many real, complex situations.

The bootstrap principle is to estimate some aspect of G, such as its standard deviation, by replacing F by an estimate $\hat{F}$; then the sampling distribution can be estimated easily by Monte Carlo simulation.

In this report we consider nonparametric problems for which $\hat{F}$ is the empirical distribution. Let $\mathcal{X}^* = (X_1^*, X_2^*, \ldots, X_n^*)$ be a "resample" (a bootstrap sample) of size n from $\hat{F}$, denote the corresponding empirical distribution $\hat{F}^*$, and write $\hat{\theta}^* = \theta(\hat{F}^*)$. For some number B of resamples (typically between 100 and 2000), sample $\mathcal{X}_b^*$ for $b = 1, \ldots, B$ with replacement from $\mathcal{X}$, then let

$$\hat{G}(a) = B^{-1} \sum_{b=1}^{B} I(\hat{\theta}_b^* \le a). \tag{2}$$

This involves two levels of approximation — approximating (1) by $P_{\hat{F}}(\hat{\theta} \le a)$, and estimating the latter by Monte Carlo simulation. Bootstrap tilting has advantages at both levels.

We restrict consideration to distributions with support on the observed data. Then we may describe a distribution in terms of the probabilities $\mathbf{p} = (p_1, \ldots p_n)$ assigned to the original observations; $\hat{F}$ corresponds to $\mathbf{p}_0 = (1/n, \ldots, 1/n)$. Let $\theta(\mathbf{p})$ be the corresponding parameter estimate (which depends implicitly on $\mathcal{X}$). Also write $\mathbf{p}^* = (M_1^*/n, \ldots, M_n^*/n)$ for the vector corresponding to resample $\mathcal{X}^*$, where M_i^* is the number of times x_i is included in $\mathcal{X}^*$. For later use, let

$$U_i(\mathbf{p}) = \lim_{\epsilon \to 0} \epsilon^{-1}(\theta(\mathbf{p} + \epsilon(\delta_i - \mathbf{p})) - \theta(\mathbf{p})) \tag{3}$$

where δ_i is the vector with 1 in position i and 0 elsewhere. When evaluated at $\mathbf{p}_0$ these derivatives are known as the empirical influence function, or infinitesimal jackknife [7].

A fundamental assumption in the application of the bootstrap is that the theoretical bootstrap distribution $P_{\hat{F}}(\hat{\theta}^* \le a)$ accurately approximates (1); in other words that $\hat{F}$ can substitute for the unknown F. Theoretical treatments of some aspects of the assumption are summarized in [11], using Edgeworth expansions, and [21], using functional analysis. We weaken the assumption by using the sampling distributions of $\hat{\theta}^*$ under certain distributions other than $\hat{F}$ which belong to "least-favorable" families (described below). These fam-

ilies play a major role in other bootstrap procedures [6, 8, 4, 5].

2 Hypothesis Tests

Consider testing H_0: $\theta = \theta_0$. In a one-parameter parametric problem one would compare the observed $\hat{\theta}$ with a critical value of its null distribution, obtained by sampling from the parametric distribution F_{θ_0} rather than $F_{\hat{\theta}}$.

Similarly, bootstrap sampling for a hypothesis test should be from a distribution consistent with the null distribution. This seems to conflict with the common bootstrap practice of sampling from the observed distribution, but in fact the bootstrap principle is to sample from the best estimate of the underlying distribution, given the information available, which may include the constraint implied by the null hypothesis. For example [19, 20] sample in this way, for testing independence, rotational invariance, symmetry, and similar problems. Others (e.g. [2]) sample in various ways consistent with the null hypothesis in two-sample and multi-sample problems. Bootstrap tilting hypothesis tests also sample this way, and were used by [23] for a one-sample mean and suggested by [16] for comparing two means.

In bootstrap tilting we hold the observed values fixed, but allow unequal probabilities $\mathbf{p} = (p_1, \ldots p_n)$ on the observations. Then the maximum likelihood estimate of the distribution maximizes $\prod p_i$ subject to $p_i \geq 0$, $\sum p_i = 1$, and $\theta(\mathbf{p}) = \theta_0$. In the case of a mean, $\theta(\mathbf{p}) = \sum p_i x_i$, $U_i(\mathbf{p}) = x_i - \overline{x}$, and the solution can be written in the form

$$p_i = c(1 - \tau(x_i - \overline{x}))^{-1}, \tag{4}$$

where τ is a "tilting" parameter and c normalizes the probabilities to sum to 1. The value of τ that satisfies the last constraint is found numerically. These probabilities are a special case of what we call "maximum likelihood tilting" (ML tilting), and are shown in Figure 1. Here the unweighted sample mean is less than the null hypothesis value, so tilting places higher probabilities on the larger values of x to make the weighted mean match θ_0.

In bootstrap tilting hypothesis testing, the null distribution of $\hat{\theta}$ is estimated by resampling from the weighted empirical distribution, e.g. the p-value against the alternative H_a: $\theta > \theta_0$ is

$$P_{F_\tau}(\hat{\theta}^* \geq \hat{\theta}), \tag{5}$$

where F_τ is the weighted empirical distribution induced by tilting with parameter τ.

The procedure can be generalized to nonlinear statistics by substituting another single-parameter family for (4). The family should be least-favorable, i.e. inference

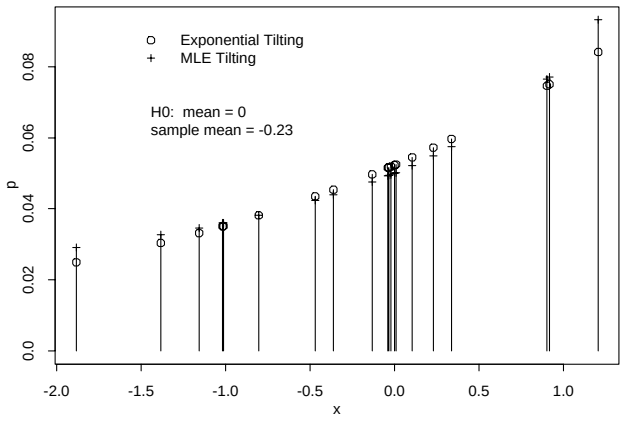

Figure 1: Exponential and Maximum Likelihood Tilting for a mean.

within a family is not easier, asymptotically, than in the full $(n-1)$-dimensional family. Two notable families are:

$$\begin{aligned} \mathcal{F}_{\text{exp}} &: \quad p_i = c\exp(\tau U_i(\mathbf{p}_0)) \\ \mathcal{F}_{\text{ml}} &: \quad p_i = c(1 - \tau U_i(\mathbf{p}_0))^{-1} \end{aligned} \tag{6}$$

each indexed by a tilting parameter τ, where each c normalizes the corresponding vector to add to 1. $\mathcal{F}_{\text{exp}}$ is known as "exponential tilting", these weights are also shown in Figure 1. Similarly $\mathcal{F}_{\text{ml}}$ is ML tilting and is the same as (4) for a mean. In the sequel we write $\mathbf{p}_\tau$ and F_τ for the corresponding probability vector and weighted empirical distribution, respectively. Note that $\tau = 0$ corresponds to $\mathbf{p}_0$ and $\hat{F}$.

For any family, τ is found numerically to satisfy the null hypothesis,

$$\theta(\mathbf{p}_\tau) = \theta_0 \tag{7}$$

and the decision to reject is based on the estimated p-value under weighted bootstrap sampling (5).

3 Confidence Intervals

Bootstrap tilting hypothesis tests are consistent with the bootstrap tilting confidence intervals defined by [6], in that the test rejects H_0 iff the confidence interval excludes θ_0. After choosing a least-favorable family, the lower limit of a one-sided $(1-\alpha)$ interval is found by solving

$$P_{F_\tau}(\hat{\theta}^* \geq \hat{\theta}) = \alpha \tag{8}$$

in τ, then defining the lower limit as

$$\theta_\alpha = \theta(F_\tau).$$

Upper limits are found similarly. [5] show that bootstrap tilting intervals are second-order correct under general

assumptions, i.e. that the one-sided coverage errors are $O(n^{-1})$. This is the same rate as for better-known procedures such as the bootstrap-t [6] and BC-a [8] intervals.

Bootstrap tilting corresponds to an exact method in single-parameter parametric problems, where the lower limit of the confidence interval is defined to be that value θ_α for which $P_{\theta_\alpha}(\hat{\theta}^* > \hat{\theta})$, where $\hat{\theta}$ is the estimate from the observed data and $\hat{\theta}^*$ is the random estimate obtained from a new sample. Here, by restricting to a least-favorable family, the problem is reduced to a single-parameter parametric family.

4 Implementation using Importance Sampling Reweighting

The most difficult step in implementing bootstrap tilting intervals is solving (8). This involves finding the value of τ for which resampling from F_τ yields a tail probability of α.

One approach is to sample from the weighted empirical distribution F_τ for different values of τ, estimate the tail probabilities for each τ, smooth the estimated probabilities, and numerically find the τ for which the value of the smooth curve is α. Because tail probabilities are relatively difficult to estimate using Monte Carlo simulation, this requires a large number of resamples (typically 1000) for each candidate value of τ. This can be expensive. [10] use the Robbins-Monroe algorithm, which would be similar in results and number of resamples required. [4] suggest an alternative, the "automatic percentile method", which requires bootstrap sampling only from one candidate F_τ (in each tail for two-sided intervals) in addition to sampling from $\hat{F}$; this would typically require 3000 resamples. The automatic percentile method may also be used as an iterative process, whose fixed point is the bootstrap tilting endpoint; iterating more than once should give more accurate endpoints, but requires more resamples.

A much more efficient approach [6] uses importance sampling reweighting (ISR), a non-traditional application of importance sampling. We review this method here before turning to its application in bootstrap tilting inference. Variations have appeared under other names, e.g. likelihood ratio sensitivity analysis, likelihood ratio gradient estimation, the score function method, polysampling, likelihood ratio reweighting, importance sampling sensitivity analysis, importance reweighting, and recycling [1, 18, 22, 13, 15, 3, 17].

Importance sampling is traditionally used to obtain more accurate answers in Monte Carlo simulation by concentrating effort on important regions in the sample space. In order to estimate an integral $\int Y(\mathcal{X}) f(\mathcal{X}) d\mathcal{X}$, one could generate B observations from density f and compute the average observed value of Y, $B^{-1}\sum_{b=1}^{B} Y_b$. Alternately, by rewriting the integral as $\int (Y(\mathcal{X})f(\mathcal{X})/g(\mathcal{X}))g(\mathcal{X})d\mathcal{X}$, where g dominates f, one could generate observations from g, and report the average observed value of (Yf/g). If g is well chosen, so that g is larger than f in "important" regions where Y is relatively large, then (Yf/g) has smaller variance (under g) than does Y (under f) [12].

The name "importance sampling" and the association with estimating integrals obscure the more general utility of the procedure. The procedure utilizes samples from a "design distribution" g in order to estimate the distribution for Y that would be obtained under sampling from the "target distribution" f. It need not be the case that f is fixed and g is chosen for variance reduction; in bootstrap tilting g is chosen for convenience, and a single set of observations (resamples) from g is used for estimation under an infinite number of target distributions.

[6] lets the design distribution be $\hat{F}$, and generates a single set of B resamples by simple bootstrap sampling (with equal probabilities). Let $M^*_{b,i}$ be the number of times x_i is included in $\mathcal{X}^*_b$. Then for any target distribution F_τ, with probabilities $\mathbf{p}_\tau$ on the observed data, the likelihood ratio $W = f/g$ for $\mathcal{X}^*_b$ is

$$W_b = \sum_{i=1}^{n} (np_i)^{M^*_{b,i}}. \tag{9}$$

Tail probability estimates

$$\begin{aligned} \hat{P}_{F_\tau}(\hat{\theta}^* \geq \hat{\theta}) &= B^{-1}\sum_{b=1}^{B} W_b I(\hat{\theta}^* \geq \hat{\theta}) \\ \hat{P}_{F_\tau}(\hat{\theta}^* \leq \hat{\theta}) &= B^{-1}\sum_{b=1}^{B} W_b I(\hat{\theta}^* \leq \hat{\theta}) \end{aligned} \tag{10}$$

are used for $\tau < 0$ and $\tau > 1$, respectively (the two probabilities do not add to 1 because $\sum_{b=1}^{B} W_b \neq B$).

This procedure has a number of advantages. Sampling with equal probabilities is simpler, and a single set of resamples is used for both sides in a two-sided confidence interval, for every statistic if confidence intervals are required for multiple statistics (e.g. multiple regression coefficients), and for every α. The estimated tail probabilities are a smooth function of τ, simplifying root-finding and eliminating the need for smoothing. Finally, by a fortunate coincidence, the unweighted empirical distribution is a nearly optimal design distribution for the traditional role of importance sampling as a variance reduction technique, at least for the mean and exponential tilting, or for other statistics if sample sizes are large.

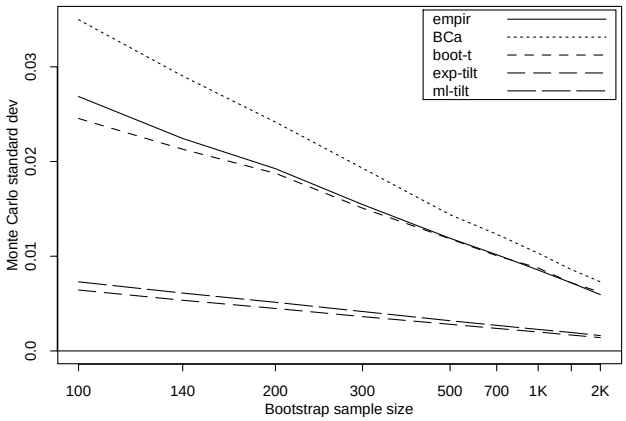

Figure 2: Estimated Monte Carlo variability (due to finite B) for confidence intervals for the mean, one-sided with $\alpha = 0.025$. There are 2000 datasets of normal data, $n = 40$; for each dataset and each value of B two bootstrap samples are created and the sample variance of the two interval endpoints is calculated. Numbers shown are the square roots of the averages of the 2000 sample variances.

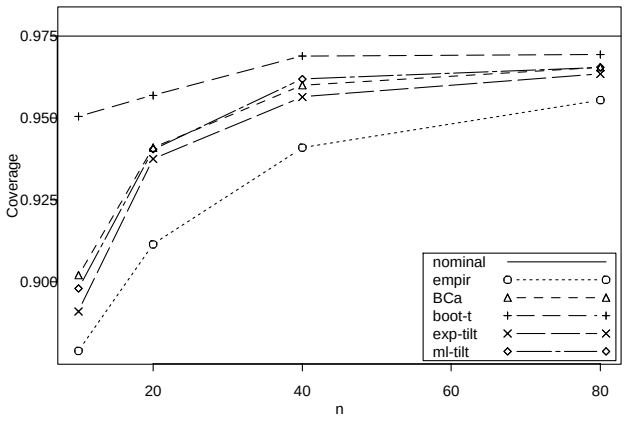

Figure 3: Coverage accuracy, for upper 97.5% intervals for the sample mean of exponential data.

Figure 2 shows the relative computational efficiency of bootstrap confidence interval procedures. The tilting intervals suffer less variability with $B = 100$ (bootstrap samples) than the BC-a and bootstrap-t intervals do with $B = 2000$. Similar results hold for other examples.

The asymptotic relative efficiency compared to either sampling with probabilities $\mathbf{p}_\tau$ or to the bootstrap percentile interval is

$$\frac{\mathrm{Var}(\hat{\theta}_{\alpha,\mathrm{MC}})}{\mathrm{Var}(\hat{\theta}_{\alpha,\mathrm{IS}})} = \frac{\mathrm{Var}(\hat{\theta}_{\alpha,\mathrm{perc}})}{\mathrm{Var}(\hat{\theta}_{\alpha,\mathrm{IS}})} = \frac{\alpha(1-\alpha)}{\exp(z_\alpha^2)\Phi(2 * z_\alpha) - \alpha^2} \tag{11}$$

where Φ is the standard normal distribution function, $\Phi(z_\alpha) = \alpha$, and $\hat{\theta}_{\alpha,\mathrm{MC}}$, $\hat{\theta}_{\alpha,\mathrm{IS}}$ and $\hat{\theta}_{\alpha,\mathrm{perc}}$ are the lower endpoints of one-sided $(1-\alpha)$ confidence intervals — $\hat{\theta}_{\alpha,\mathrm{IS}}$ is the exponential tilting interval using importance sampling, $\hat{\theta}_{\alpha,\mathrm{MC}}$ is the exponential tilting interval estimated using weighted Monte Carlo sampling, and $\hat{\theta}_{\alpha,\mathrm{perc}}$ is the bootstrap percentile interval estimated using simple Monte Carlo sampling. The variances are conditional on the observed data, the result is asymptotic as both $n \to \infty$ and $B \to \infty$, and depends on certain regularity conditions, that the statistic being bootstrapped is asymptotically linear and normal. The relative efficiency is about 17 for a two-sided 95% interval ($\alpha = 0.025$).

The advantage is greater when α is smaller, e.g. when Bonferroni is used to adjust individual α values in a multiple-testing procedure. For $\alpha = .005$ the relative efficiency is approximately 68.

The computational advantage is even greater relative to the bootstrap BC-a interval [8], probably the most common second-order-correct bootstrap interval, if z_0 is estimated from the data by the usual procedure $\hat{z}_0 = \Phi^{-1}(\hat{G}(\hat{\theta}))$. Then the Monte Carlo variability of the BC-a interval is greater than that of the bootstrap percentile interval by a factor of 2.18 (asymptotically) if $a = z_0 = 0$; if $a = z_0 = 0.1$ the factor is about 7.77.

Other design distributions are possible, in particular (defensive) mixture designs [14] of the form $\sum_{k=1}^{K} \lambda_k \hat{F}_{\tau_k}$. These mixtures are more robust for statistics other than a sample mean, but are beyond the scope if this article.

5 Statistical Properties

Preliminary results in a variety of applications are that the coverage accuracy of exponential tilting confidence intervals is roughly comparable to the BC-a intervals, and the ML tilting intervals are slightly more accurate. The accuracy of bootstrap-t intervals depends on whether a statistic is transformed; e.g. they are not accurate when the statistic is the variance, but are accurate when the statistic is the log of the variance. All are more accurate than the bootstrap percentile interval, which is only first-order accurate, with one-sided coverage errors of order $O(n^{-1/2})$. Figure 3 shows typical results; the actual coverage probabilities approach the nominal value as n increases, but for fixed n the bootstrap-t interval is most accurate and the bootstrap percentile interval the least accurate.

The ML tilting intervals are slightly wider on average than exponential tilting and BC-a intervals, but substantially shorter than bootstrap-t intervals. See e.g. Figure 4.

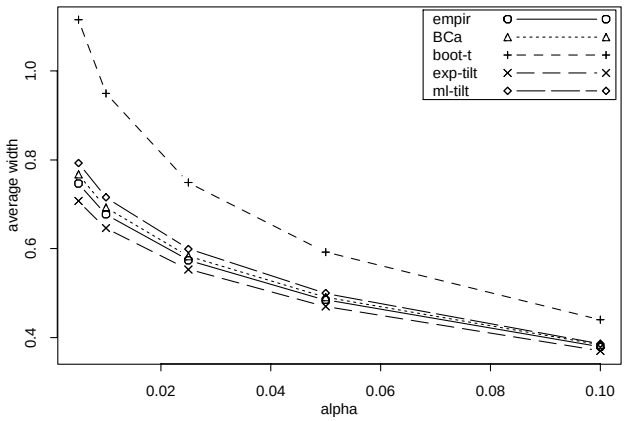

Figure 4: Confidence Interval Length. Average half-length of $(1 - 2\alpha)$ confidence intervals, for samples of size 10 from normal distributions.

In summary, bootstrap tilting confidence intervals and hypothesis tests are very computationally efficient, are shorter on average than bootstrap-t intervals, and have good coverage accuracy.

Acknowledgements

This work was supported by NSF Phase I SBIR Award No. DMI-9861360.

References

[1] R. J. Beckman and M. D. McKay. Monte Carlo estimation under different distributions using the same simulation. *Technometrics*, 29:153–160, 1987.

[2] D. D. Boos, P. Janssen, and N. Veraverbeke. Resampling from centered data in the two sample problem. *J. Statist. Plan. Inference*, 21:327–345, 1989.

[3] A. Davison and D. Hinkley. *Bootstrap Methods and their Applications*. Cambridge University Press, 1997.

[4] T. J. DiCiccio and J. P. Romano. The automatic percentile method: accurate confidence limits in parametric models. *The Canadian Journal of Statistics*, 17(2):155–169, 1989.

[5] T. J. DiCiccio and J. P. Romano. Nonparametric confidence limits by resampling methods and least favorable families. *International Statistical Review*, 58(1):59–76, 1990.

[6] B. Efron. Nonparametric standard errors and confidence intervals. *Canadian Journal of Statistics*, 9:139 – 172, 1981.

[7] B. Efron. *The Jackknife, the Bootstrap and Other Resampling Plans*. National Science Foundation – Conference Board of the Mathematical Sciences Monograph 38. Society for Industrial and Applied Mathematics, Philadelphia, 1982.

[8] B. Efron. Better bootstrap confidence intervals (with discussion). *Journal of the American Statistical Association*, 82:171 – 200, 1987.

[9] B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall, 1993.

[10] P. H. Garthwaite and S. T. Buckland. Generating Monte Carlo confidence intervals by the Robbins-Monro process. *Applied Statistics*, 41(1):159–171, 1992.

[11] P. Hall. *The Bootstrap and Edgeworth Expansion*. Springer, New York, 1992.

[12] J. M. Hammersley and D. C. Hanscomb. *Monte Carlo Methods*. Methuen, London, 1964.

[13] T. C. Hesterberg. *Advances in Importance Sampling*. PhD thesis, Statistics Department, Stanford University, 1988.

[14] T. C. Hesterberg. Weighted average importance sampling and defensive mixture distributions. *Technometrics*, 37(2):185–194, 1995.

[15] T. C. Hesterberg. Estimates and confidence intervals for importance sampling sensitivity analysis. *Mathematical and Computer Modeling*, 23(8/9):79–86, 1996.

[16] D. V. Hinkley. Bootstrap significance tests. *Bulletin of the International Statistical Institute*, pages 65–74, 1989.

[17] M. A. Newton and C. J. Geyer. Bootstrap recycling: A Monte Carlo alternative to the nested bootstrap. *Journal of the American Statistical Association*, 89(427):905–912, 1994.

[18] M. I. Reiman and A. Weiss. Sensitivity analysis via likelihood ratios. In *Proceedings of the 1986 Winter Simulation Conference*, pages 285–289, 1986.

[19] J. P. Romano. A bootstrap revival of some nonparametric distance tests. *Journal of the American Statistical Association*, 83(403):698–708, 1988.

[20] J. P. Romano. Bootstrap and randomization tests of some nonparametric hypotheses. *Annals of Statistics*, 17:141–159, 1989.

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

[22] J. W. Tukey. Configural polysampling. *SIAM REVIEW*, 29:1–20, 1987.

[23] G. A. Young. Resampling tests of statistical hypotheses. In D. Edwards and N. E. Raun, editors, *Compstat: Proceedings in Computational Statistics*, pages 233–238, Heidelberg, 1988. Physica-Verlag.