BOOTSTRAP P-VALUES FOR TESTS OF NONPARAMETRIC HYPOTHESES

Dennis D. Boos and Cavell Brownie

Institute of Statistics Mimeo Series No. 1919

June, 1988

ABSTRACT

In this paper we consider totally nonparametric tests for location such as tests of H_O: equal means, or H_O: equal medians, without making any assumptions concerning nuisance parameters or distribution shapes. For a wide class of such problems and test statistics, we show how to obtain critical values using a new bootstrap approach. Asymptotic justification for the bootstrap procedures is given as well as Monte Carlo results for evidence of small sample properties. The empirical results show that the bootstrap methods compare favorably with a variety of procedures that have been proposed for use when homogeneity of variance cannot be assumed.

Key Words: Bootstrap, nonparametric, ANOVA, hypothesis testing, variance heterogeneity, Wilcoxon rank sum, median test.

1. Introduction

One of the most frequently encountered statistical procedures in the Psychology and Education literature is the analysis of variance (ANOVA) procedure for testing equality of means assuming common variances (and in small samples, normality). In studies with a simple design (e.g., the one-way classification), concern about validity of the common variance or normality assumptions may lead to use of a procedure more robust than ANOVA. Popular alternatives to ANOVA include, e.g., the Welch t and its k-sample extensions or the Mann-Whitney-Wilcoxon test and the k-sample Kruskall-Wallis procedure. It is our opinion, however, that often the practitioner does not fully understand the properties of these alternatives, or the H_o and H_a implied by their use. For more complex designs such as 2-way or 3-way factorial designs, alternatives that are robust to variance heterogeneity are not readily available, and ANOVA may be used whether appropriate or not. Thus, in spite of the large literature on the properties of ANOVA and more robust alternatives (e.g., see Tomarken and Serlin, 1986), appropriate methods are not always employed to test for location effects. This leads to the following objectives for this paper.

(i) We consider testing equality of means, or other location measures, in a totally nonparametric manner; i.e., testing H_o: equal means, or H_a: equal medians, without making any assumptions about nuisance parameters or distribution shapes.

(ii) For testing these "totally nonparametric" hypotheses, we introduce a bootstrap approach to estimate critical values. We show that this bootstrap approach works with a variety of test statistics in a wide class of such problems, including testing for main and interaction effects on location in factorial designs with unequal replication and variance heterogeneity.

Before discussing the nonparametric hypotheses, we review a number of null hypotheses for equality of location. For tests on means the notation is as follows. Given (X_{ij}, j=1,...,n_i, iid F_i, i=1,..., k), let $\bar{X}_i = \Sigma_{j=1}^{n_i} X_{ij}/n_i$ and $s_i^2 = \Sigma_{j=1}^{n_i}(X_{ij}-\bar{X}_i)^2/(n_i-1)$ denote the sample means and variances, and $\bar{\bar{X}} = \Sigma\Sigma X_{ij}/N$ the grand mean, where $N=\Sigma_{i=1}^{k} n_i$. Also let $\mu_i = \int x dF_i(x)$ and $\sigma_i^2 = \int (x-\mu_i)^2 dF_i(x)$ represent the population means and variances. We now review a range of hypotheses concerning the μ_i.

The usual parametric hypothesis for equality of the μ_i, for k=2, is

$$H_o: \mu_1=\mu_2 \text{ vs } H_a: \mu_1\neq\mu_2 \text{ ; with } F_i \text{ normal } (\mu_i, \sigma_i^2),\ \sigma_1^2=\sigma_2^2=\sigma^2, \sigma^2 \text{ unknown.} \quad (1)$$

The appropriate test statistic is the pooled t or $t_p=(\bar{X}_1-\bar{X}_2)/[s_p^2(1/n_1+1/n_2)]^{1/2}$, where $s_p^2=\Sigma_{i=1}^{2}(n_i-1)s_i^2/(N-2)$, with null distribution Student's t with N−2 degrees of freedom (df). For k≥2, the one-way ANOVA F-statistic

$$F_p = \left[\Sigma_{i=1}^{k}\, n_i\left(\bar{X}_i-\bar{\bar{X}}\right)^2/(k-1)\right] \Big/ \left[\Sigma_{i=1}^{k}(n_i-1)s_i^2/(N-k)\right]$$

is compared to the F distribution with k−1 and N−k df.

Removing the equal variance assumption produces the Behrens-Fisher model

$$H_o: \mu_1=\mu_2 \text{ vs } H_a\ \mu_1\neq\mu_2 \text{; with } F_i \text{ normal } (\mu_i, \sigma_i^2),\ \sigma_1^2, \sigma_2^2 \text{ possibly unequal.} \quad (2)$$

A useful procedure for (2), which we shall refer to here as the Welch t (e.g., Best and Rayner, 1987), is $t_w=\left(\bar{X}_1-\bar{X}_2\right)\Big/\left(s_1^2/n_1+s_2^2/n_2\right)^{1/2}$ compared to the t distribution with estimated Satterthwaite (1941) df.

Relaxing the normality assumption in (1) and (2) but assuming a location scale family results in the following semiparametric analogues of (1) and (2). For parameter identification let $\int x dF_o(x)=0$ and $\int x^2 dF_o(x)=1$ if the latter integral exists. Then with

$$F_i(x)=F_o\left((x-\mu_i)/\sigma_i\right),\ F_o \text{ unknown},$$

we have

$$H_o\colon \mu_1 = \cdots = \mu_k \text{ vs } H_a\colon \text{some } \neq;\ \sigma_1 = \cdots = \sigma_k = \sigma,\ \sigma \text{ unknown} \quad (3)$$

and

$$H_o\colon \mu_1 = \cdots = \mu_k \text{ vs } H_a\colon \text{some } \neq;\ \sigma_1,\ldots,\sigma_k \text{ possibly unequal.} \quad (4)$$

Note that H_o in (3) is equivalent to $H_o\colon F_1 = \ldots = F_k$ which is crucial for the validity of permutation tests. In particular the semiparametric (3) is the standard model for classical rank tests for shift, such as the Mann-Whitney-Wilcoxon and Kruskal-Wallis tests, which can be more efficient than the t or F tests for long-tailed distributions. Less obvious to practitioners is that validity of these rank tests cannot be assumed under (4) (e.g., Pratt, 1964, van der Vaart, 1961, Sen, 1962).

Going one step further, the assumption of a common parametric form F_o may also be removed, yielding (for $k \geq 2$)

$$H_o\colon \mu_1 = \cdots = \mu_k \text{ vs } H_a\colon \text{some } \neq;\ F_i \text{ unknown.} \quad (5)$$

More generally, for any location measure $\theta(F_i)$, we have

$$H_o\colon \theta(F_1)=\cdots=\theta(F_k) \text{ vs } H_a\colon \text{some } \neq;\ F_i \text{ unknown.} \quad (6)$$

We refer to hypotheses (5) and (6) as totally nonparametric because equality of location is tested without any parametric assumptions on the F_i. Since restrictive assumptions about the data are not required, these nonparametric hypotheses have practical utility and are the focus of this article.

Careful consideration should be given to data type and study design and objectives when choosing an appropriate model from (1) to (6) and the corresponding test procedure. With respect to data type, approximate normality usually suggests tests on means μ_i, whereas $\theta(F_i)$=median(F_i) is more appropriate with skewed or outlier-prone distributions. Experimental setting and objectives both affect assumptions concerning commonality of distribution scales or shapes under H_o and H_a. In studies comparing attributes of, or responses for, essentially different types of units (e.g., individuals of different sex, race or cultural group), equality of scales cannot be assumed under the null. Appropriate nulls are (6), or if the location-scale assumption can be made, nulls of the form (2) or (4).

For experiments where homogeneous units (e.g., rats of the same strain) are selected at random to receive one of k treatments, the null hypothesis of no treatment effect corresponds to H_o: $F_1=\cdots=F_k$ $\left(H_o \text{ for (3)}\right)$. The pooled t and ANOVA F are always valid under normality (and approximately valid without normality) in this experimental situation. Study objectives may, however, suggest posing a different null. If interest is in detecting location changes in the presence of (or instead of) scale changes, one might prefer a test procedure which is valid under a Behrens-Fisher type null (2) or (4) or the nonparametric (6) and which is also sensitive to location differences for any scale configuration. Thus, the pooled t or ANOVA F, may not be as desirable as popular alternatives such as the Welch t that provide a p-value which reflects "distance" from the null (4) rather than from H_o: $F_1=\cdots=F_k$. In other words, given random assignment of like units to treatment groups, variance

heterogeneity in the resulting data does not invalidate use of the ANOVA F, but does suggest use of an alternative procedure to detect changes in location rather than scale.

In Section 2 we introduce our bootstrap method for estimating critical values in testing the nonparametric hypotheses (5) and (6). In Sections 3-5 asymptotic results are given for the bootstrap procedures, together with results from Monte Carlo simulation as evidence of small sample behavior. Section 3 focuses on comparisons between means, Section 4 deals with a generalization of (6) involving the Mann-Whitney-Wilcoxon parameter P(Y>X), and Section 5 discusses the k-sample median test. On the whole, the Monte Carlo results show that bootstrap methods can be used to improve validity of levels but not without small losses in power. On the other hand, our theorems and results for quite different statistics illustrate that bootstrapping provides a comparatively easy and general method for achieving validity. That is, for nulls (6), our bootstrap approach can provide a reference distribution for simple and intuitive statistics, whereas the analytical construction of correctly studentized statistics and corresponding reference distributions can be difficult. We conclude with examples in Section 6.

2. The Bootstrap

The bootstrap technique was introduced by Efron (1979) as a nonparametric technique for estimating unknown quantities of sampling distributions such as variance, bias, percentiles, etc. The statistics literature on the bootstrap has been rapidly expanding with most of the emphasis on parameter estimation and confidence interval construction. See Efron and Tibshirani (1986) for an overview. Less attention has been given to the bootstrap in testing situations, and there are features which make its application to testing somewhat different from the construction of confidence intervals.

Consider a statistic T which is useful to distinguish between a given H_o and H_a in (6). The idea behind the bootstrap in testing situations is to estimate the null percentiles of T regardless of whether the true state is H_o or H_a. Otherwise the test may not have good power under H_a. For purposes of analogy note that when using the t statistic and the t table, the critical values of course do not depend on the sample or on the truth of H_o or H_a. Bootstrap critical values depend on the sample but we would like to force them to be as constant as possible from sample to sample under H_o or H_a. Given "data" $\left(X_{ij}, j=1,\ldots,n_i, i=1,\ldots,k\right)$, this is achieved by "shifting" the data to produce $\left(\hat{X}_{ij}, j=1,\ldots,n_i, i=1,\ldots,k\right)$ such that whether null or alternative holds, the empirical distribution functions $\hat{F}_i$ of the $\hat{X}$ satisfy H_o. For example, consider testing (6) with $\theta(F_i)=\text{median}(F_i)$. Let $\hat{\theta}_i$ represent the i^{th} sample median, and set $\hat{X}_{ij}=X_{ij}-\hat{\theta}_i$, $j=1,\ldots, n_i$, $i=1,\ldots,k$. Then the null hypothesis of equal medians holds for the $\hat{F}_i(x) = \Sigma_j I(\hat{X}_{ij}\leq x)/n_i$, $i=1,\ldots,k$, whatever the true F_i. A bootstrap sample is then obtained by iid sampling, separately for $i=1,\ldots,k$, from each $\hat{F}_i$. The value of T for such a sample is labeled T^* to denote that it came from a bootstrap sample. Separate sampling from each $\hat{F}_i$ is necessary to produce a null situation because the nulls we consider allow different scales or shapes for the F_i. As is customary, B bootstrap samples and T^* values are obtained by resampling from the $\hat{F}_i$, $i=1,\ldots,k$, and the bootstrap p-value for a test where H_o is rejected for large T is $\hat{p}_B=\#\left\{T^*\geq T_o\right\}/B$, where T_o is T evaluated for the original data.

This approach has the following asymptotic justification. For the k-sample location problem, let $\mathcal{F}_o$ consist of all sets of k distribution functions $(F_1,\ldots, F_k)$ with equal locations $\theta(F_i)=\theta_i$ (and finite second moments if $\theta(F_i)=$mean μ_i of F_i). Hypothesis testing when the null set is so large raises philosophical questions. If H_o is rejected for T large, some might require that a critical value c_α satisfy

$$\sup_{(F_1,\ldots,F_k)\epsilon\mathcal{F}_o} P(T\geq c_\alpha)=\alpha\ .$$

Since $\mathcal{F}_o$ is so large, obtaining such a c_α seems like a hopeless task and could provide a fairly conservative procedure. Instead our bootstrap approach provides an estimated $\hat{c}_\alpha$ (and also a p-value) such that $P(T \geq \hat{c}_\alpha) \longrightarrow \alpha$ as $\min(n_1,\ldots,n_k) \longrightarrow \infty$ if the underlying true set $(F_1,\ldots,F_k)$ actually belongs to $\mathcal{F}_o$. We do this by using the $\hat{c}_\alpha$ which is appropriate for the set $\left(\hat{F}_1(x),\ldots,\hat{F}_k(x)\right)=\left(F_{n_1}(x+\hat{\theta}_1),\ldots,F_{n_k}(x+\hat{\theta}_k)\right)$, where $F_{n_i}(x)$ is the empirical distribution of the i^{th} sample and $\hat{\theta}_i$ is a consistent estimator of θ_i. If the true set $(F_1,\ldots,F_k) \in \mathcal{F}_o$, then clearly $(\hat{F}_1(x),\ldots,\hat{F}_k(x)) \xrightarrow{wp1} \left(F_1(x+\theta_1),\ldots,F_k(x+\theta_k)\right)$ and we can show (Corollary 1) that $P(T \geq \hat{c}_\alpha) \longrightarrow \alpha$. Usually we suggest studentized statistics T so that this convergence is faster than Central Limit Theorem rates, but we give no proofs of this faster convergence.

It is also necessary to consider interpretation of the estimated p-values when the true set $(F_1,\ldots,F_k)$ does not belong to $\mathcal{F}_o$. In some sense the bootstrap p-values attempt to provide a distance via T in probability units from the true set $(F_1,\ldots,F_k)$ to that member of $\mathcal{F}_o$, say $(G_1,\ldots,G_k)$, which has locations all equal and for which $G_i(x+\theta_i)=F_i(x)$. That is, we measure distance to that member of $\mathcal{F}_o$ whose elements have exactly the same shape as $(F_1,\ldots,F_k)$ but are shifted by $\theta_1,\ldots,\theta_k$. Of course we are using shifted empirical distribution functions as surrogates for $(G_1,\ldots,G_k)$ which seems intuitively reasonable, but there can be other options. For example, if all the distributions are supported on $[0,\infty)$ and we are interested in means, we could use $\left(F_{n_1}(x\,\bar{X}_1),\ldots,F_{n_k}(x\,\bar{X}_k)\right)$ in place of $\left(F_{n_1}(x+\bar{X}),\ldots,F_{n_k}(x+\bar{X})\right)$. If the true set $(F_1,\ldots,F_k)$ belongs to $\mathcal{F}_o$, then the two approaches should give similar results. However, under an alternative the two methods will differ so that we suggest choosing first the scale to work in which is most meaningful for thinking about location differences. In other words, we would suggest transformations such as log X in certain situations before bootstrapping from the aligned empirical distributions $\left(F_{n_1}(x+\bar{X}_1),\ldots,F_{n_k}(x+\bar{X}_k)\right)$.

3. Tests Based on Means

In this section we consider test procedures based on sample means $\bar{X}_i$ and sample variances s_i^2 for the nonparametric hypotheses (5), which include the Behrens-Fisher type problems (2) and (4).

It is well known (e.g. Scheffe, 1970) that the pooled t is not valid in small samples, or asymptotically, for Behrens-Fisher nulls (2), (4) or for the wider class (5). The Fisher permutation t is valid for $F_1=F_2$ nulls but not over the class (5) because for $F_1 \neq F_2$ the $\binom{n_1+n_2}{n_1}$ partitions of the data are not equally likely (the actual probabilities depending on F_1, F_2). Similarly, for $k>2$, comparison of the one-way ANOVA F_p statistic to either the $F(k-1, N-k)$ distribution or to a randomization distribution assuming all $\binom{N}{n_1,\ldots,n_k}$ partitions equally likely, is not correct over all nulls in (5).

In contrast, the Welch t is asymptotically correct for (5) and has been shown empirically to perform well (both in validity and power) in small samples under normality (Best and Rayner, 1987). Brown and Forsythe (1974a) proposed a k-sample procedure (which for $k=2$ is equivalent to the Welch t) based on comparison of

$$F_{BF} = \frac{\sum_i n_i\left(\bar{X}_i-\bar{\bar{X}}\right)^2}{\Sigma\left(1-\frac{n_i}{N}\right)s_i^2}$$

to the F distribution with $k-1$ numerator df and estimated Satterthwaite denominator df f where

$$\frac{1}{f} = \sum_i \frac{c_i^2}{(n_i-1)}, \quad c_i = \frac{\left(1-\frac{n_i}{N}\right)s_i^2}{\Sigma\left(1-\frac{n_i}{N}\right)s_i^2}.$$

F_{BF} was denoted F^* by Brown and Forsythe but we reserve the $*$ notation here to denote a bootstrap quantity. On the basis of a Monte Carlo study Brown and Forsythe concluded that

their procedure yielded approximately correct levels for (2) in small samples with performance comparable to Welch's (1951) k-sample procedure. In contrast, Tomarken and Serlin (1986) found the Brown and Forsythe procedure was liberal, and the Welch procedure more reliable, for moderately large samples (average $n_i \geq 10$). These empirical results are in agreement with our analysis below which shows that the Brown and Forsythe procedure, though undeniably useful in small samples, is not asymptotically correct.

The potential candidates for a statistic to be used in our bootstrap approach to testing hypotheses (5) include $\bar{X}_1-\bar{X}_2$, t_w, t_p (for k=2), F_{BF} and the one-way ANOVA F_p. As outlined in Section 2, for T corresponding to one of the above statistics a bootstrap resample and T^* are generated by drawing an iid sample of size n_i from $\hat{F}_i(x)=F_{n_i}(x+\bar{X}_i)=$ $n_i^{-1}\Sigma_j I(X_{ij}-\bar{X}_i \leq x)$, separately for i=1,...,k. More intuitively, this corresponds to resampling from centered values $X_{ij}-\bar{X}_i$, so that $\hat{F}_1,\ldots,\hat{F}_k$ each have mean 0 and constitute a particular null in (5). For k=2, a two-tailed H_a: $\mu_1 \neq \mu_2$, and B resamples, p-values are obtained as $\hat{p}_B=\#\left\{|T^*|>|T_o|\right\}/B$ (c.f. Pratt and Gibbons, 1981, Sec. 4.5). Otherwise $\hat{p}_B$ is the appropriate one-tailed proportion.

Theorem 1 and Corollaries 1 and 2 provide asymptotic justification for the bootstrap p-values. Let $\bar{X}_i^*$ and s_i^{2*} be defined as $\bar{X}_i$ and s_i^2 respectively, but based on bootstrap samples, and let $Y_N = \left(n_1^{1/2}\bar{X}_1,\ldots,n_k^{1/2}\bar{X}_k\right)^T$. Note that Y_N^*, the bootstrap version of Y_N, will always have mean 0 since Y_N^* is drawn from centered distribution functions $(\hat{F}_1,\ldots,\hat{F}_k)$. Also note that we use $\xrightarrow{d^*}$ to mean covergence in distribution in the bootstrap world for an infinite set of k sequences of the original data samples. Since the bootstrap distribution is itself random depending on these samples, the "$\xrightarrow{d^*}$" must be accompanied by "almost surely" (a.s.) to reflect this randomness.

Theorem 1. For i=1,...,k let $X_{i1},\ldots,X_{in_i}$ be iid with distribution function $F_i(x)$, mean $E(X_{i1})=\mu_i$, and Var $X_{i1}=\sigma_i^2 < \infty$. Similarly, for i=1,...,k let $X_{i1}^*,\ldots,X_{in_i}^*$ be iid bootstrap samples with distribution functions $\hat{F}_i(x) = F_{n_i}(x+\bar{X}_i)$, where $F_{n_i}(x)$ is the empirical distribution function of $X_{i1},\ldots,X_{in_i}$. Then, as $\min(n_1,\ldots,n_k) \to \infty$

i) $Y_N^* = \left(n_1^{1/2}\,\bar{X}_1^*,\ldots,n_k^{1/2}\,\bar{X}_k^*\right)^T \xrightarrow{d^*}$ Mult. Normal $\left(0, \text{diag}\,(\sigma_1^2,\ldots,\sigma_k^2)\right)$ a.s.

and

ii) $\left(s_1^{2*},\ldots,s_k^{2*}\right) \xrightarrow{p^*} \left(\sigma_1^2,\ldots,\sigma_k^2\right)$ a.s.

Proof. Both i) and ii) may be proved by following closely the proof of Theorem 2.1 of Bickel and Freedman (1981) and noting that $\hat{F}_i(x) \xrightarrow{a.s.} F_i(x+\mu_i)$ wherever $F_i(x+\mu_i)$ is continuous. □

We give two corollaries which relate to many statistics of interest for a fixed-effects ANOVA situation. For constants $a_N=(a_{1N},\ldots,a_{kN})^T$ and $b_N=(b_{1N},\ldots,b_{kN})^T$ define $V_1 = \Sigma_{i=k}^k\, a_{iN}\, n_i^{1/2}\,\bar{X}_i = a_N^T Y_N$ and $V_2 = V_1/\left(\Sigma_{i=1}^k b_{iN}^2 s_i^2\right)$. V_1 is an arbitrary linear combination and V_2 is a studentized version where $|a_{iN}|=b_{iN}$ would be the "correct" studentization. We allow $|a_{iN}| \neq b_{iN}$ so that statistics designed for the equal variance case may be analyzed in the unequal variance case. Let $V_3 = Y_N^T C_N Y_N / \left(\Sigma_{i=1}^k d_{iN} s_i^2\right)$ where C_N is a k×k matrix of constants and $d_N=(d_{1N},\ldots,d_{kN})^T$ is a further vector of constants. Finally, let V_i^* be like V_i, but based on the $\bar{X}_i^*$, and recall that $\bar{X}_i^*$ has expectation 0 with respect to $\hat{F}_i$.

Corollary 1. If the assumptions of Theorem 1 hold with $n_i/N \to \lambda_i \in (0,1)$ $i=1,\ldots,k$, and $a_N \to a$, $b_N \to b$, $d_N \to d$, $C_N \to C$, where b, d, and C each have at least one nonzero element, then a.s.

$$V_1^* \xrightarrow{d^*} N\Big(0, \sum_{i=1}^{k} a_i^2 \sigma_i^2\Big)$$

$$V_2^* \xrightarrow{d^*} N\Big(0, \sum_{i=1}^{k} a_i^2\sigma_i^2 / \sum_{i=1}^{k} b_i^2\sigma_i^2\Big)$$

$$V_3^* \xrightarrow{d^*} Z^\top C Z \Big/ \Big(\sum_{i=1}^{k} d_i^2 \sigma_i^2\Big),$$

where Z is the multivariate normal random variable with mean 0 and covariance matrix diag $\left(\sigma_1^2,\ldots,\sigma_k^2\right)$.

Proof. These convergences follow directly from Theorem 1 and standard results found in Chapter 1 of Serfling (1980). □

We now give a second corollary which relates to the p-values obtained from bootstrapping under H_o and H_a. Here $\hat{p}_i = P^*\left(V_i^* \geq V_i\right)$ is the p-value appropriate for rejecting for large values of V_i when $B=\infty$. The case where $B \to \infty$ as $N \to \infty$ can be handled similarly, but requires more notation. Also, the two-sided case is similar but not presented.

Corollary 2. If the assumptions of Corollary 1 hold, then

i) Under H_o: $\Sigma\, a_i \lambda_i^{1/2} \mu_i = 0$ and assuming that $\Sigma\, a_{iN} n_i^{1/2} \mu_i \to 0$, $\hat{p}_i \xrightarrow{d} U(0,1)$, $i=1,2$.

ii) Under H_o: $E(Z^\top CZ)=0$ and assuming that $E(Y_N^\top C_N Y_N) \to 0$, $\hat{p}_3 \xrightarrow{d} U(0,1)$.

iii) Under H_a: $\Sigma\, a_i \lambda_i^{1/2} \mu_i > 0$, $\hat{p}_i \xrightarrow{p} 0$, $i=1,2$.

iv) Under H_a: $E(Z^T C Z) > 0$, $\hat{p}_3 \xrightarrow{p} 0$.

Proof. i) and ii) Under H_o both V_i^* and V_i converge to the same distribution so that an asymptotic version of the probability integral transformation gives the results. iii) Under an alternative V_1 and V_2 tend to ∞ but V_1^* and V_2^* converge to normal random variables because of the centering. iv) is similar. □

Examples. k=2

1) For bootstrapping $\bar{X}_2 - \bar{X}_1$ Corollary 2 applies with $V_1 = \sqrt{N}\left(\bar{X}_2 - \bar{X}_1\right)$ and

$$a_{1N} = -\sqrt{\frac{N}{n_1}} = -1/\lambda_{1N}^{1/2} \quad , \qquad a_{2N} = \sqrt{\frac{N}{n_2}} = 1/\lambda_{2N}^{1/2} .$$

2) $t_p = \dfrac{\bar{X}_2 - \bar{X}_1}{\sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = V_2$ with $a_{1N} = -\sqrt{\frac{n_2}{N}}$, $a_{2N} = \sqrt{\frac{n_1}{N}}$, $b_{1N}^2 = \frac{(n_1-1)}{N-2}$,

and $b_{2N}^2 = \frac{n_2-1}{N-2}$.

Note that $|a_{1N}| = b_{1N}$ iff $n_1 = n_2$.

3) $t_w = \dfrac{\bar{X}_2 - \bar{X}_1}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = V_2$ with $a_{1N} = -\sqrt{\frac{n_2}{N}}$, $a_{2N} = \sqrt{\frac{n_1}{N}}$, $b_{1N}^2 = \frac{n_2}{N}$ and

$b_{2N}^2 = \frac{n_1}{N}$.

Note that t_w is correctly studentized (i.e. $|a_{iN}| = b_{iN}$).

One-Way ANOVA

The treatment sum of squares in one-way ANOVA may be expressed as $\Sigma\, n_i\left(\bar{X}_i-\bar{\bar{X}}\right)^2=Y_N^T\, C_N Y_N$, where $C_N = I_k - e_N e_N^T$ with $e_N^T=\left((n_1/N)^{1/2},\ldots,(n_k/N)^{1/2}\right)$. Since $n_i/N \to \lambda_i$ we have $C_N \to C$ and C is symmetric and idempotent with rank $k-1$. The ANOVA F-statistic can thus be written

$$F_p= \frac{1}{k-1}\, Y_N^T C_N Y_N \Big/ \sum_{i=1}^{k} d_{iN} s_i^2, \text{ where } d_{iN} = (n_i-1)\Big/(N-k).$$

By Corollary 1 the bootstrap version of $(k-1)F_p$ converges in distribution a.s. to $Z^T C\, Z/\ \Sigma\, \lambda_i \sigma_i^2$. In the proof of Corollary 2(ii) we use the fact that $(k-1)F_p$ converges to this same distribution under H_o. Note that if the variances σ_i^2 are all equal, then this limiting distribution is χ^2_{k-1}. The Brown and Forsythe (1974a) statistic, suggested for use when the variances are not equal, can be written as

$$F_{BF} = Y_N^T C_N Y_N \Big/ \sum_{i=1}^{k} \left(1-\frac{n_i}{N}\right) s_i^2 .$$

Note that the limiting null distribution of F_{BF} is $Z^T CZ/\Sigma(1-\lambda_i)\sigma_i^2$ which has mean 1 but is not a $\chi^2_{k-1}\Big/(k-1)$ distribution unless the variances σ_i^2 are all equal. Thus, in large samples, where the limit of the Brown and Forsythe procedure is comparison of $(k-1)F_{BF}$ to the χ^2_{k-1} distribution, levels will not be quite correct. Turning to F_p, the null limiting distribution has mean $(k-1)^{-1}\ \Sigma\left(1-\lambda_i\right)\sigma_i^2\Big/\Sigma\lambda_i\sigma_i^2$, which is dependent on the σ_i^2 unless either the sample sizes are equal (and thus $F_p=F_{BF}$) or the σ_i^2 are equal.

Corollary 2 gives results which justify using bootstrap p-values in large samples for a variety of statistics. Corollary 2 tells us nothing, however, about the relative power of bootstrapping different test statistics or how fast the convergence takes place. Past experience

and second order expansions (e.g., Babu and Singh, 1983) suggest that convergence will be faster for "correctly" studentized statistics, so that bootstrapping should work better with F_{BF} than with F_p. To obtain empirical evidence concerning these issues, and to compare the bootstrap procedures with obvious competitors, limited Monte Carlo work was carried out.

Results are described separately for the 2-sample and k-sample statistics, but first we list several features that were common to the Monte Carlo studies referred to here and in Sections 4 and 5.

1. In every situation NMC =1000 independent sets of Monte Carlo replications were generated. Thus, empirical test rejection rates follow the binomial (NMC=1000, p=probability of rejection) distribution.
2. P-values were computed for each test statistic and rejection of H_o at α=.05 means p-value $\leq$ .05.
3. Recall that a bootstrap p-value $\hat{p}_B$ is based on B bootstrap replications within each of the NMC=1000 Monte Carlo replications. For tests based on k=2 means, B=500 was used. To reduce costs for the k>2 means test and for rank based procedures we used a two-stage sequential procedure:
   a) start with B=100; b) if $\hat{p}_B > .20$, stop; c) if $\hat{p}_B \leq .20$, take 400 more replications and use all B=500 replications to compute $\hat{p}_B$.
4. For each statistic we counted the number of p-values falling in the intervals (0,.01), (.01, .02),..., (.09, .10), (.10, 1.0). These counts were used in two ways. For continuous statistics, a correct p-value should have a uniform (0,1) distribution under H_o. To check this a chi squared goodness-of-fit test of uniformity was computed using the counts for the 11 intervals. This approach conveys more information (concerning the range of interest $0<p<.10$) than just reporting empirical rejection rates for a level .05 test.

5. In non-null situations it can be useless to compare empirical rejection rates ("observed power") if the null levels are much different than the nominal levels. Therefore, in addition to "observed power" we calculated "adjusted power" estimates using the cell counts described above. This was done by adding the counts (or an appropriate fraction thereof) for those cells for which counts summed to α under H_o. For example, if the first 5 cells had counts (19, 9, 13, 12, 10) under H_o, and (648, 114, 51, 34, 25) under H_a, then the estimated level under H_o for nominal α=.05 is 63/1000= .063, the observed power under H_a is .872 and the adjusted power is $\left(648+114+51+\ 34\ \times\ 9/12\right)/1000\ =\ .839$. In these calculations, the counts used were for the H_o which, in the appropriate class of nonparametric nulls, was in some sense closest to the given power situation. For example in Table 1, adjusted power estimates, in parentheses, for H_a: $\mu_2=\mu_1+2\sigma_1$ and extreme value data, were calculated using the null case $\mu_1=\mu_2$, and extreme value data, with the same scales and sample sizes. Adjusted powers in parentheses in Tables 1-4 thus provide rough comparisons of test power assuming critical values could be correctly specified for each procedure.

Two sample results

Monte Carlo results are given in Table 1 for the pooled t (t_p compared to the $t(n_1+n_2-2)$ distribution), the Welch t, t_p and t_w with bootstrap p-values, and $\bar{X}_2-\bar{X}_1$ with bootstrap p-values (included for comparison with the studentized statistics). Two distributions were used, one symmetric (the normal) and one asymmetric (extreme value with distribution function $F(x)=\exp(-\exp(-x))$), with unequal scales ($\sigma_2=2\sigma_1$) and unequal sample sizes ($n_1=16$, $n_2=8$). Results for null performance are given in separate columns for the left and right tailed tests (H_a: $\mu_2<\mu_1$ and H_a: $\mu_2>\mu_1$, respectively) and for the two tailed test (H_a: $\mu_1\neq\mu_2$) because these can be quite different.

----Insert Table 1----

Under normality, null performance for both the Welch t and the bootstrapped t_w is satisfactory. The pooled t is liberal (this is expected when the larger n_i is coupled with the smaller σ_i^2) but bootstrapping t_p produces marked improvement. Smaller χ^2_{10} values for the bootstrapped t_w, as compared to the bootstrapped t_p, illustrate the slower convergence for the incorrectly studentized t_p. Bootstrapping $\bar{X}_2 - \bar{X}_1$ for these sample sizes is clearly inferior to bootstrapping t_p or t_w, and is only slightly better than the usual pooled t (row 1). For the extreme value samples, similar conclusions hold concerning null performance except that none of the procedures does well on the left tailed test. For the alternative $\mu_2 = \mu_1 + 2\sigma_1$, results for adjusted power (in parentheses) show there is a small cost in power for using the bootstrap. Overall, the Welch t seems the best procedure, but bootstrapping the appropriately studentized t_w does almost as well.

k-sample Results

Table 2 contains empirical rejection rates for four k-sample procedures. These are the usual F-statistic F_p with p-values obtained from the $F(k-1, N-k)$ distribution (row 1) or from bootstrapping (row 2), and the Brown and Forsythe (1974a) F_{BF} with p-values from the F distribution with estimated Satterthwaite denominator df (row 3) or from bootstrapping (row 4). Results in Table 2 cover a limited number of situations, each for $k=6$ samples and standard deviations (3,3,2,2,1,1) (cf., Brown and Forsythe, 1974a, Table 1). However, comparing results for the balanced cases ($n=4$ and $n=8$) gives an idea of the effect of increasing sample sizes, and the unbalanced case (with small n_i paired with large σ_i and vice versa) is one where the usual F-test is expected to do poorly. Results are displayed for extreme value data but very similar patterns were seen with normals.

----Insert Table 2----

Even in the balanced cases, the usual F-test is too liberal under H_o (row 1) but the Brown and Forsythe procedure (row 3) is, at worst, only slightly liberal. The bootstrap of F_p (row 2) and F_{BF} (row 4) are similar and generally conservative, but show improvement with increased samples sizes. Under the alternative H_a $(\mu_1, \mu_2, \mu_3, \mu_4, \mu_5, \mu_6)=(4,2,1,1,0,0)$ the Brown and Forsythe procedure has better power than the bootstrap procedures, but adjusted power (in parentheses) indicates the difference in power is due mainly to the more conservative levels of the bootstrap.

Extensions

Having concentrated on the one-way ANOVA problem, we now indicate briefly ways to generalize the bootstrap approach to more complex fixed effects designs when variance homogeneity cannot be assumed. For the 2-way classification with replication in each cell, and balanced data ($n_{ij}=n$, $i=1,\ldots,a$, $j=1,\ldots,b$), Brown and Forsythe (1974b) suggested the usual F-statistics, but with denominator df estimated by the Satterthwaite method, to test for main effects and interaction. Given balanced data this procedure should be satisfactory, but when there is unequal replication, constructing statistics to test meaningful hypotheses is not straightforward even with $\sigma_{ij}^2 \equiv \sigma^2$ (e.g., Searle, 1987, Chapter 4). Brown and Forsythe (1974b, p. 723) outline an approach that allows for variance heterogeneity and imbalance but is not easily implemented for testing meaningful main effect and interaction hypotheses. An alternative is to bootstrap a more intuitive sum of squares, appropriately studentized. For example, given (X_{ijk}, $k=1,\ldots,n_{ij}$, iid with mean μ_{ij}, variance σ_{ij}^2, $i=1,\ldots,a$, $j=1,\ldots,b$), to test for no row effects H_o: $b^{-1}\Sigma_j \mu_{1j} = \ldots = b^{-1}\Sigma_j \mu_{aj}$, Searle (1987, p. 90) gives

$$SSA_w = \sum_{i=1}^{a} w_i(\tilde{X}_{i..} - \Sigma w_i \tilde{X}_{i..}/w.)^2,$$

with $\tilde{X}_{i..} = b^{-1}\Sigma_j \bar{X}_{ij.}$, $\bar{X}_{ij.} = n_{ij}^{-1} \Sigma_k X_{ijk}$, $w_i = b^{-2} \Sigma_j n_{ij}^{-1}$, and $w. = \Sigma_i w_i$.

To allow for variance heterogeneity, we bootstrap the studentized

$$FA_W = SSA_W \Big/ \left[\sum_i \left(1 - \frac{w_i}{w_\cdot}\right) \frac{w_i}{b^2} \sum_j \frac{s_{ij}^2}{n_{ij}} \right],$$

which can be shown to satisfy the conditions of Theorem 1. This is illustrated in Section 6 where a test for interaction is also described.

Another approach, which is illustrated in Section 6, is to bootstrap appropriately studentized analogs of the F statistics used in the "unweighted means" analysis (e.g. Glass, 1970, p. 441). Following Welch (1951), yet another statistic which could be bootstrapped is SSA_V, like SSA_W, but with weights $v_i = b^{-2} \Sigma_j s_{ij}^2/n_{ij}$. Analogues to Theorem 1 can be developed for these Welch-type statistics but are not presented here. Note, however, that these statistics arise from the estimated generalized least squares approach.

4. The Mann-Whitney-Wilcoxon Parameter P(Y>X).

In the two-sample problem, an appealing alternative to the parametric and semiparametric Behrens-Fisher models (2) and (4) is to focus on the parameter $\theta = P(Y>X)$ (e.g. Wolfe and Hogg, 1971). Using X_j, F and Y_j, G instead of X_{ij}, F_i, i=1 and 2 in this section only, let $R_1,\ldots,R_{n_2}$ be the ranks of $Y_1,\ldots,Y_{n_2}$ among the combined samples $X_1,\ldots,X_{n_1}$, $Y_1,\ldots,Y_{n_2}$. Then, for F and G continuous, $\theta = P(Y_1>X_1)$ arises naturally from consideration of the Wilcoxon rank sum statistic $W = \Sigma_{i=1}^{n_2} R_i$, or the Mann-Whitney statistic $U = \Sigma_{i=1}^{n_1} \Sigma_{j=1}^{n_2} I(Y_j>X_i) = W - n_2(n_2+1)/2$, U/n_1n_2 being the U-statistic estimator of θ.

The properties of U are well known and much work has been devoted to finding confidence intervals for θ (e.g., see Halperin, Gilbert and Lachin, 1987). For hypothesis testing U has typically been used to test H_o: F=G, but in some experimental settings it may be more

reasonable to test $\theta=1/2$ without assuming F=G. We are therefore concerned here with using U (or W) for testing

$$H_o: \theta = \int F(x)dG(x)=1/2 \ , \ F, G \text{ unknown and possibly unequal.} \quad (7)$$

Under H_o: $\theta=1/2$, $F \neq G$, the distribution of U(or W) depends on F and G and does not coincide with the tabled F=G null distribution or the usual asymptotic normal approximation. Van der Vaart (1961) and Pratt (1964) show clearly the effect of unequal variances on null levels when W is compared to the F=G distribution to test H_o: $\theta=1/2$. For example, given normality and equal sample sizes the true level of a nominal $\alpha=.05$ test can be .09. For unequal sample sizes it can range up to .17. These are asymptotic results, but true levels about twice the nominal level are not unlikely in realistic Behrens-Fisher situations where $\theta=1/2$.

One solution to this problem is to estimate the variance of U without assuming F=G and use a studentized version of U. Sen (1962) and Fligner and Policello (1981) proposed similar rank-based variance estimators each yielding a studentized statistic which is distribution-free if F=G and is asymptotically correct for (7). These procedures have undoubted merit but we feel there are also advantages to the bootstrap approach which we now present.

To generate a bootstrap null distribution for U (or W) we first align the Y's by the Hodges-Lehmann shift estimator $\hat{\Delta}$ = median $\left\{Y_j - X_i,\ 1 \leq i \leq n_1, 1 \leq j \leq n_2\right\}$ giving $Y_1 - \hat{\Delta}, \ldots, Y_{n_2} - \hat{\Delta}$ with empirical distribution function $G_{n_2}(x+\hat{\Delta})$. Then we generate

$$X_1^*, \ldots, X_{n_1}^* \text{ iid from } \hat{F}(x)=F_{n_1}(x) \text{ and } Y_1^*, \ldots, Y_{n_2}^* \text{ iid from } \hat{G}(x)=G_{n_2}(x+\hat{\Delta}). \quad (8)$$

This creates a null situation in the bootstrap world since $\theta^*=P^*(Y_1^*>X_1^*)=\Sigma_i\Sigma_j I(Y_i-\hat{\Delta}>X_j)/(n_1n_2)$ is 1/2 when n_1n_2 is even and is very close to 1/2 otherwise. For each of B bootstrap X^* and Y^* samples, U^* is calculated, and if U_o is the value of U for the original X and Y samples, the bootstrap p-value is $\hat{p}_B=\#\left\{U_i^*\geq U_o\right\}/B$ for H_a: $\theta > 1/2$, or $\hat{p}_B= \#\left\{|U_i^*|\geq|U_o|\right\}/B$ for H_a: $\theta \neq 1/2$. We use $\hat{p}$ to denote the p-value when $B=\infty$.

The following theorem justifies the use of $\hat{p}$ in large samples although we could state it for $\hat{p}_B$ (with more cumbersome notation) with $B \to \infty$ as $\min(n_1,n_2) \to \infty$.

Theorem 2. Suppose that $X_1,\ldots,X_{n_1}$ are iid with continuous distribution function F(x) and independent of $Y_1,\ldots,Y_{n_2}$ which are iid with continuous distribution function G(x). Suppose that Δ is the unique solution of $\int F(x)dG(x+\Delta)=1/2$. Let U^* be based on bootstrap sampling from $\hat{F}(x)$ and $\hat{G}(x)$ in (8). Then as $\min(n_1,n_2) \to \infty$ with $n_1/N \to \lambda \in (0,1)$

i) $N^{1/2}\left(\frac{U^*}{n_1n_2} - \frac{1}{2}\right) \xrightarrow{d^*} N(0, \sigma^2)$ a.s.,

where $\sigma^2 = \frac{1}{\lambda}\int\left[G(x+\Delta)-\frac{1}{2}\right]^2dF(x) + \frac{1}{1-\lambda}\int\left[F(x)-\frac{1}{2}\right]^2 dG(x+\Delta)$.

ii) Under H_o: $\theta=\int F(x)\, dG(x) = \frac{1}{2}$, $\hat{p} \xrightarrow{d} U(0,1)$.

iii) Under H_a: $\theta=\int F(x)\, dG(x) > \frac{1}{2}$, $\hat{p} \xrightarrow{p} 0$.

Proof. For i) and ii) we use Theorem 5.3.20 of Randles and Wolfe (1979, p. 160) and note that the conditions of that theorem can be verified using the fact that the kernel $h(x,y)=I(y>x)$ is bounded and that $\sup_x|G_{n_2}(x+\hat{\Delta})-G(x+\Delta)| \xrightarrow{a.s.} 0$ since $\hat{\Delta} \xrightarrow{a.s.} 0$ and G

is continuous. Part iii) follows since $N^{1/2}(U/n_1n_2-\theta)$ converges to a normal distribution and i) continues to hold under H_a. □

To provide evidence to support use of the bootstrap p-values in small samples, limited Monte Carlo work was carried out. Results are given in Table 3 for four procedures based on the Mann-Whitney U (each of which could equivalently be described in terms of the Wilcoxon W). Both U and the studentized $\hat{U}$ (Fligner and Policello, 1981, equation 3.2) were bootstrapped (rows 2 and 4). For comparison row 1 contains results for U with p-values obtained using a continuity correction and the Edgeworth approximation to the asymptotic distribution assuming F=G, and row 3 represents $\hat{U}$ with a continuity correction and asymptotic normal approximation. Again two distribution types were used, the normal and extreme value with $F(x)=\exp(-\exp(-(x-\omega)/\tau)$. Scales were $\tau_1=1$, $\tau_2=4$ for the normal samples, with $\omega_1=\omega_2=0$ for the null ($\theta=.5$) situation and $\omega_1=0$, $\omega_2=3.47$ for an alternative ($\theta=.8$). For the extreme value data, we used $\tau_1=1$, $\tau_2=2$, with $\omega_1=0$, $\omega_2= -.289$ for the null situation and $\omega_1=0$, $\omega_2 = 1.709$ for $\theta= .8$.

- - - -Insert Table 3- - - -

Results for the two distribution types show similar patterns. In the null situation, the Mann-Whitney-Wilcoxon test (row 1) is liberal because of the different scales (small τ_i with large n_i). Bootstrapping U (row 2) produces levels that are much more acceptable. Fligner and Policello's $\hat{U}$ seems somewhat liberal (row 3), but it too is improved by bootstrapping (row 4). Comparison of row 3 with Table 2 of Fligner and Policello suggests that results there for $\hat{U}$ are a little optimistic because they do not reflect the effect on performance of different and small sample sizes coupled with different scales. This is borne out by our results for two-tailed tests (not shown) where the bootstrap procedures compare perhaps more favorably

with $\hat{U}$. These conclusions are the same whether based on observed (nominal $\alpha = .05$) rejection rates or on the χ^2_{10} values, though allowance must be made for discreteness of the statistics (particularly U) when comparing χ^2_{10} values. An advantage of $\hat{U}$ relative to U, noted by Fligner and Policello and reflected to some extent by the χ^2_{10} values, is the greater range of achievable natural levels provided by $\hat{U}$. The bootstrap procedures share this advantage.

Comparison of observed power is a little unfair to the less liberal bootstrap procedures, but the adjusted power (in parentheses) does not differ much among procedures. To summarize, for n_1, n_2 large (say > 20), $\hat{U}$ may be compared to the asymptotic normal approximation, and for n_1, n_2 both ≤ 12, $\hat{U}$ with Table 1 of Fligner and Policello seems to give satisfactory results. However for n_1, n_2 between these ranges and unequal, and particularly if validity is important or a p-value is desired, the bootstrap procedures can be recommended.

We conclude this section with a comment on the null hypotheses associated with k-sample linear rank statistics based on the Wilcoxon and other score functions when the assumption of a location model is <u>not</u> made. For k=2 the Chernoff-Savage asymptotic means $\int J(H(x))dF(x)$, with $H(x) = \lambda F(x) + (1-\lambda)G(x)$ were examined for score functions $J(\cdot)$ corresponding to Wilcoxon scores, median scores, normal scores and log-rank scores. We see that $\int J(H(x))dF(x)=0$ implies $\int G(x)\, dF(x)=1/2$ for Wilcoxon scores, and implies median(F) − median(G)=0 for median scores, leading to the useful hypotheses (7) and (6). No such simple, meaningful and sample size independent "parameters" are obtained with normal or log-rank scores, however. Thus in a totally nonparametric testing situation (i.e., in the absence of a location or location scale model) the only meaningful "parameters" from linear rank statistics on which to base a null hypothesis seem to be P(Y>X) and median(F) − median(G).

For k>2, the nonparametric null for a rank statistic based on Wilcoxon scores would appear to be

$$H_o: \ P(X_{i1} > X_{j1})=\tfrac{1}{2} \text{ for all i, j pairs.} \tag{9}$$

Without restricting consideration to symmetric F_i, (or to $F_1=\cdots=F_k$ nulls) the null class of distributions $(F_1,\ldots,F_k)$ corresponding to (9) is unappealing. For example, given a set of nonsymmetric distributions $(F_1,\ldots,F_k)$ it is not necessarily possible to align the F_i to produce a set which satisfies (9). That is, it is not generally the case that $\Delta_1,\ldots,\Delta_k$ exist such that $P(X_{i1}-\Delta_i > X_{j1} - \Delta_j)=1/2$ for all (i, j) pairs. With median scores there is no such difficulty with defining the null class for k>2. Thus for k-sample tests based on linear rank statistics the only truly nonparametric null hypothesis seems to be that of equal medians, or (6) with $\theta(F_i)$ = median(F_i). This null is considered in Section 5.

5. K-Sample Median Test

Hajek and Sidak (1967, p. 105) give the following modification of Mood's k-sample median test. For continuous data, let R_{ij} be the rank of X_{ij} in the combined samples $(X_{11},\ldots,X_{1n_1},\ldots,X_{k1},\ldots,X_{kn_k})$. The null hypothesis of equal medians is rejected for large values of the rank statistic

$$Q=4\ \Sigma_{i=1}^{k}(A_i-n_i/2)^2/n_i,$$

where $A_i=\Sigma_j\left[\text{sign}(R_{ij}-(N+1)/2)+1\right]/2$, i=1,...,k, and sign (x)=1,0,−1 if x >, =, < 0 respectively. Critical values are obtained from the χ^2_{k-1} distribution.

This procedure is asymptotically valid if attention is restricted to nulls of the form H_o: $F_1=\cdots=F_k$, but the limiting distribution of Q is not χ^2_{k-1} for all $(F_1,\ldots,F_k)$ in the larger null

class corresponding to the nonparametric H_o: median(F_1)=$\cdots$=median(F_k). For this nonparametric null, an asymptotically valid procedure is obtained by applying our bootstrap approach to the statistic Q.

For i=1,...,k, let $\hat{\theta}_i$ be the i^{th} sample median and $\hat{X}_{ij}=X_{ij}-\hat{\theta}_i$, j=1,...,$n_i$. A bootstrap resample is obtained by drawing $X^*_{i1},\ldots,X^*_{in_i}$ iid from $\hat{F}_i(x)=F_{n_i}(x+\hat{\theta}_i)$, i=1,...,k. Let Q_o be Q evaluated for the original sample and Q^*_i, i=1,...,B be Q evaluated for each of B bootstrap resamples. Then the bootstrap p-value is $\hat{p}_B=\#\left\{Q^*_i \geq Q_o\right\}/B$.

In the interest of brevity theorems on the asymptotic validity of $\hat{p}_B$ are not given here. Information concerning small sample behavior was obtained via Monte Carlo with the results reported in Table 4. Two procedures were compared, Q with p-values from the χ^2_{k-1} distribution (row 1) and from our bootstrap approach (row 2). For k=4 samples, we studied three extreme value cases and only one with normal samples because the median test is not recommended for near-normal data. Larger scales were paired with larger sample sizes (see Pratt, 1964) resulting in liberal performance of Q when p-values are based on the χ^2_3 distribution (row 1). In contrast, the bootstrap p-values were conservative (row 2). Increasing sample sizes from (16, 16, 8, 8) to (24, 24, 16, 16) appeared to cause some improvement in the bootstrap method but gave even more liberal results using the χ^2_3 approximation. As with the Mann-Whitney U, applying the bootstrap to Q produces a greater range of naturally achievable levels. This is reflected in the moderate χ^2_{10} values for goodness-of-fit to uniformity in 2 of 3 null cases in row 2. Observed power is of course higher for the liberal χ^2_3 approximation, but adjusted power differs little between the two procedures. Once again, the bootstrap approach can be used to achieve approximate validity in a Behrens-Fisher type situation with a statistic designed for the location problem assuming equal scales.

----Insert Table 4----

6. Example

Glass (1970, Sec. 19.7) notes that studies in Psychology and Education are often "natural" rather than "controlled" experiments, because there is no random assignment of factor-level combinations to experimental units. Instead units are selected randomly from existing populations that can be classified according to levels of one or more factors such as anxiety level, whether or not a student elects to take Latin, socioeconomic status, marital status, etc. One of the points we have tried to make here is that homogeneity of variances (or common distribution shapes) cannot necessarily be assumed for the populations sampled, so that procedures robust to variance heterogeneity should be used to compare population means. To emphasize this, we illustrate the use of the bootstrap approach to test for main effects and interaction for data from a 2 x 3 classification without assuming equal variances or equal replication in each cell.

The following notation is needed. For a 2-way classification with factor A at a levels and B at b levels, the data are given by X_{ijk}, $k = 1,\ldots,n_{ij}$, $i = 1,\ldots,a$, $j=1,\ldots,b$. Let

$$\bar{X}_{ij.} = \frac{1}{n_{ij}} \sum_k X_{ijk},\ s_{ij}^2 = \frac{1}{n_{ij}-1} \sum_k (X_{ijk} - \bar{X}_{ij.})^2,\ i = 1,\ldots,a,\ j = 1,\ldots,b;$$

$$\tilde{X}_{i..} = \frac{1}{b} \sum_j \bar{X}_{ij.},\ i = 1,\ldots,a;\ \tilde{X}_{.j.} = \frac{1}{a} \sum_i \bar{X}_{ij.},\ j = 1,\ldots,b;$$

$$n_{i.} = \sum_j n_{ij},\ w_i = \frac{1}{b^2} \sum_j \frac{1}{n_{ij}},\ i = 1,\ldots,a;$$

$$n_{.j} = \sum_i n_{ij},\ v_j = \frac{1}{a^2} \sum_i \frac{1}{n_{ij}},\ j = 1,\ldots,b\ ;$$

$$w_. = \sum_i w_i,\ v_. = \sum_j v_j,\ n_{..} = \sum_{i,j} n_{ij},\ \text{and}\ \tilde{X}_{...} = \frac{1}{ab} \sum_{i,j} \bar{X}_{ij.}\ .$$

Two types of statistics were bootstrapped, one with unweighted, the other with weighted numerator sums of squares. To test for main effects for A and B, and for an A*B interaction, these are, respectively,

$$FA_u = \frac{\frac{b}{a\text{-}1}\sum_i (\tilde{X}_{i..}\text{-}\tilde{X}_{...})^2}{\frac{1}{ab}\sum_{i,j}\frac{s_{ij}^2}{n_{ij}}} ,$$

$$FB_u = \frac{\frac{a}{b\text{-}1}\sum_j (\tilde{X}_{.j.}\text{-}\tilde{X}_{...})^2}{\frac{1}{ab}\sum_{i,j}\frac{s_{ij}^2}{n_{ij}}} ,$$

$$FAB_u = \frac{\frac{1}{(a\text{-}1)(b\text{-}1)}\sum_{i,j}\hat{\gamma}_{ij}^2}{\frac{1}{ab}\sum_{i,j}\frac{s_{ij}^2}{n_{ij}}} ,$$

$$FA_w = \frac{\sum_i w_i\left(\tilde{X}_{i..} - \frac{\Sigma w_i \tilde{X}_{i..}}{w_.}\right)^2}{\sum_i\left(1 - \frac{w_i}{w_.}\right)\frac{w_i}{b^2}\sum_j\frac{s_{ij}^2}{n_{ij}}} ,$$

$$FB_w = \frac{\sum_j v_j\left(\tilde{X}_{.j.} - \frac{\Sigma v_j \tilde{X}_{.j.}}{v_.}\right)^2}{\sum_j\left(1 - \frac{v_j}{v_.}\right)\frac{v_j}{a^2}\sum_i\frac{s_{ij}^2}{n_{ij}}} ,$$

$$FAB_W = \frac{\sum_{i,j} n_{ij}\,\hat{\gamma}_{ij}^2}{\sum_{i,j} \frac{s_{ij}^2}{ab}\left[(a-2)(b-2) + \frac{(a-2)}{b}\,n_{i\cdot} + \frac{(b-2)}{a}\,n_{\cdot j} + \frac{n_{\cdot\cdot}}{ab}\right]},$$

with $\hat{\gamma}_{ij} = \bar{X}_{ij\cdot} - \tilde{X}_{i\cdot\cdot} - \tilde{X}_{\cdot j\cdot} + \tilde{X}_{\cdots}$, i = 1,...,a, j = l,...,b. For comparison with the bootstrap p-values, the unweighted means analysis, which assumes equal variances, was carried out (see Glass, 1979, p. 441 for formulas).

For illustrative purposes, hypothetical data were used (but for a comparable situation see Glass, 1970, p. 451). The data were generated assuming normal distributions with means $\mu_{11} = 1$, $\mu_{12} = 0$, $\mu_{13} = -.5$, $\mu_{21} = -1.5$, $\mu_{22} = -.5$, $\mu_{23} = 0$, corresponding to an A*B interaction and a main effect for A ($\Sigma_j\mu_{1j}/3 = \frac{1}{6}$, $\Sigma_j\mu_{2j}/3 = -\frac{2}{3}$). Sample sizes used were $n_{11} = n_{12} = n_{13} = n_{21} = 8$, $n_{22} = n_{23} = 16$, and to emphasize the effect of sample size and variance configuration on the pooled ANOVA tests, two sets of standard deviations were used. That is, using exactly the same pseudorandom N(0,1) deviates, and μ_{ij} values, two data sets were generated, the first with $\sigma_{11} = \sigma_{12} = 3$, $\sigma_{13} = \sigma_{21} = 2$, $\sigma_{22} = \sigma_{23} = 1$ and the second with $\sigma_{11} = \sigma_{12} = 1$, $\sigma_{13} = \sigma_{21} = 2$, $\sigma_{22} = \sigma_{23} = 3$. Thus in generating data set 1, the largest sample sizes were paired with the smallest variances (inverse pairing) and for data set 2, the largest sample sizes were paired with the largest variances (direct pairing).

Summary statistics for the generated data, for data set 1, were

$$\begin{array}{lll} \bar{X}_{11\cdot} = 0.49 & \bar{X}_{12\cdot} = 0.60 & \bar{X}_{13\cdot} = -1.35 \\ s_{11}^2 = 13.78 & s_{12}^2 = 8.948 & s_{13}^2 = 3.924 \\ \bar{X}_{21\cdot} = -2.68 & \bar{X}_{22\cdot} = -0.18 & \bar{X}_{23\cdot} = 0.20 \\ s_{21}^2 = 4.611 & s_{22}^2 = 1.045 & s_{23}^2 = 0.823\,, \end{array}$$

and for data set 2, were

$$\bar{X}_{11\cdot} = 0.83 \qquad \bar{X}_{12\cdot} = 0.20 \qquad \bar{X}_{13\cdot} = -1.35$$

$$s^2_{11} = 1.531 \qquad s^2_{12} = 0.994 \qquad s^2_{13} = 3.924$$

$$\bar{X}_{21\cdot} = -2.68 \qquad \bar{X}_{22\cdot} = 0.47 \qquad \bar{X}_{23\cdot} = 0.59$$

$$s^2_{21} = 4.611 \qquad s^2_{22} = 9.404 \qquad s^2_{23} = 7.406 \ .$$

For data set 1, results for an overall test of H_o: $\mu_{11}=\cdots=\mu_{23}$ were $F_p = 3.27$ with $p = .011$ from the F(5,58) distribution, and $F_{BF} = 2.42$ with $p = .064$ obtained from the F distribution with 5 and 24.6 df or $p = .084$ obtained by bootstrapping. For data set 2, results were $F_p = 3.04$ with $p = .017$ from the F(5,58) distribution and $F_{BF} = 3.87$ with $p = .005$ from the F distribution with 5 and 51.0 df or $p = .011$ obtained from bootstrapping. Tests for main and interaction effects for the two data sets are summarized in Table 5 in terms of the calculated values of the test statistics and associated p-values. All bootstrap p-values were based on $B = 10000$ replications.

- - - -Insert Table 5- - - -

Comparing results for the two data sets, we see that for data set 1 p-values are smaller for the procedures that assume common variance than for the more robust alternatives, while for data set 2 the opposite is true. The smaller p-values for the pooled variance procedures in data set 1 are an indication of the greater power of these procedures under an alternative, but also of their liberal nature in null situations (H_o: $\mu_{11}=\cdots=\mu_{23}$), when there is inverse pairing of variances and sample sizes. Larger p-values in data set 2 reflect the conservative nature of the pooled variance procedures given direct pairing of variances and sample sizes.

The main effect for A was not apparent in the generated data but the procedures did detect the A*B interaction. Given such an outcome, a researcher might look at simple effects such as the comparison of A at each level of B, and so the two-sample procedures of Section 3 were used to test H_o: $\mu_{13} = \mu_{23}$. For illustrative purposes, only data from the A_1B_3 and A_2B_3 cells were used with the pooled t, though in practice the pooled ANOVA would usually be followed by t-tests based on the pooled error mean square with 58 df. Results for the 2-sided H_a: $\mu_{13} \neq \mu_{23}$ were as follows. For data set 1, $t_p = 2.66$ with $p = .014$ from the t(22) distribution, and $t_w = 2.11$ with $p = .066$ from the t(8.5) distribution or $p = .070$ via the bootstrap. For data set 2, $t_p = 1.79$, $p = .088$, and $t_w = 1.99$ with $p = .061$ from the t(18.7) distribution or $p = .064$ via the bootstrap. Thus, again we see that p-values are smaller for the pooled-variance t than for the alternative procedures in data set 1, while the reverse is true for data set 2. The bootstrap t_w and Welch t agree closely.

Following Brown and Forsythe (1974a), the Satterthwaite approach could be used to obtain approximate df and hence an approximate reference distribution for the statistics FA_w, FB_w, FAB_w and analogs for a 3-way or higher order classification. We have not applied this approach to the 2 x 3 example here, partly because we wanted to emphasize the flexibility of the bootstrap method of constructing a null reference distribution. It should be apparent from this example how to apply the bootstrap to statistics appropriate for higher order fixed-effects designs, provided there is replication within each cell.

TABLE 1

Rejection Rates for Tests Based on Means when k=2, $\sigma_2=2\sigma_1$, $n_1=16$, $n_2=8$, and Nominal $\alpha=.05$.

	Normal Samples, H_o: $\mu_1=\mu_2$						H_a: $\mu_2=\mu_1+2\sigma_1$
	Left-tailed		Right-tailed		Two-tailed		Two-tailed
	.05	χ^2_{10}	.05	χ^2_{10}	.05	χ^2_{10}	.05
Statistic (Ref. Dist.)							
$t_p\left(t_{n_1+n_2-2}\right)$	.088	62.2	.100	95.1	.126	155.8	.84(.71)
t_p (bootstrap)	.055	8.3	.065	20.6	.057	16.6	.67(.65)
t_w (t, est. d.f.)	.047	6.3	.053	9.1	.046	8.4	.64(.67)
t_w (bootstrap)	.043	8.8	.058	8.2	.040	9.9	.60(.63)
$\bar{X}_2-\bar{X}_1$ (bootstrap)	.087	43.3	.086	60.8	.101	119.5	.80(.66)
	Extreme Value Samples, H_o: $\mu_1=\mu_2$						H_a: $\mu_2=\mu_1+2\sigma_1$
$t_p\left(t_{n_1+n_2-2}\right)$	.118	173.8	.070	41.5	.124	162.6	.88(.73)
t_p (bootstrap)	.100	140.6	.038	7.4	.074	23.9	.70(.59)
t_w (t, est. d.f.)	.087	62.2	.030	21.9	.059	16.8	.69(.66)
t_w (bootstrap)	.082	65.7	.034	10.7	.058	21.9	.61(.59)
$\bar{X}_2-\bar{X}_1$ (bootstrap)	.120	281.6	.048	9.7	.111	190.3	.86(.66)

Note: Each rejection rate is the proportion of p-values $\leq .05$ in 1000 Monte Carlo replications. The χ^2_{10} entries are chi-squared goodness-of-fit test statistics for uniformity for p-values with 10 equal cells on (0, .10] and (.10, 1.0] for the 11^{th} cell.

TABLE 2

Rejection Rates under H_o: equal means for k=6 Extreme Value Samples with Standard Deviations (3, 3, 2, 2, 1, 1).

$(n_1,n_2,n_3,n_4,n_5,n_6)$ =	(4, 4, 4, 4, 4, 4)			(4, 6, 6, 8, 10, 12)			(8, 8, 8, 8, 8, 8)		
	H_o		H_a	H_o		H_a	H_o		H_a
	.05	χ^2_{10}	.05	.05	χ^2_{10}	.05	.05	χ^2_{10}	.05
Statistic (Ref. Dist.)									
$F_p\left(F_{k-1,N-k}\right)$	.077	28.9	.54(.43)	.168	503.0	.87(−)	.078	51.0	.90(.84)
F_p (bootstrap)	.030	17.9	.28(.35)	.049	20.1	.53(.53)	.043	4.9	.79(.81)
$F_{BF}\left(F_{k-1,\text{est. d.f.}}\right)$	.045	2.9	.38(.40)	.057	10.4	.58(.55)	.063	14.4	.87(.84)
F_{BF} (bootstrap)	.024	18.9	.28(.37)	.036	8.4	.45(.53)	.039	8.3	.78(.83)

Note: The alternative considered under headings labeled H_a is $(\mu_1, \mu_2, \mu_3, \mu_4, \mu_5, \mu_6)$= (4, 2, 1, 1, 0, 0). In parentheses are estimates of power for tests adjusted to have true level = .05. Adjusted power was not presented in one instance due to greater inaccuracy of the estimate when more than 5% of the null p-values fall in the first cell (0, .01]. Entries under H_o and .05 are the number of test rejections for nominal level α=.05 in 1000 Monte Carlo replications and have s.d. $\leq$ .01. Entries under χ^2_{10} are chi-squared goodness-of-fit statistics for p-values.

TABLE 3

Rejection Rates for Right-Tailed Mann-Whitney U and modified U Tests when $(n_1, n_2) = (16, 8)$ and Nominal level $\alpha = .05$.

	Normal[1]			Extreme Value[2]		
P(Y>X) =	.50		.80	.50		.80
Statistic (Ref. Dist.)	.05	χ^2_{10}	.05	.05	χ^2_{10}	.05
U (F=G Edgeworth approx.)	.106	87.4[3]	.75(.63)[4]	.081	51.3[3]	.81(.71)
U (bootstrap)	.068	16.1	.65(.60)	.057	12.5	.74(.70)
$\hat{U}$ (normal approx.)	.074	41.2	.69(.62)	.075	61.8	.81(.72)
$\hat{U}$ (bootstrap)	.063	11.5	.65(.62)	.064	17.5	.78(.72)

1 The distribution for the first sample is N(0,1) for both situations, and for the second sample is N(0,16) for P(Y>X)=.5 and N(3.47, 16) for P(Y>X)=.8.

2 Extreme value samples have distribution function $F(x)=\exp(-\exp(-(x-\omega)/\tau))$. For the first sample $\omega=0$ and $\tau=1$ for both situations. For the second sample $\omega= -.289$, $\tau=2$ and $\omega= 1.709$, $\tau=2$ correspond to P(Y>X)=.5 and .8 respectively.

3 The χ^2_{10} values are somewhat inflated by the discreteness of U.

4 Adjusted power values in columns 3 and 6 are appropriate only if the null situations are those reported in columns 1 and 4 respectively.

TABLE 4

Rejection Rates for Nominal level α=.05 Under H_o: equal medians for k=4 samples with scales (4, 4, 1, 1).

	Normal		Extreme Value[1]				
(n_1, n_2, n_3, n_4) =	(16, 16, 8, 8)		(16, 16, 8, 8)		(24, 24, 16, 16)		(16, 16, 8, 8)
	H_o		H_o		H_o		H_a
Statistic (Ref. Dist.)	.05	χ^2_{10}	.05	χ^2_{10}	.05	χ^2_{10}	.05
Q $(\chi^2_3)^2$	.074		.070		.091		.60(.52)[3]
Q (bootstrap)	.035	20.3	.039	7.9	.044	4.9	.44(.50)

1 Extreme value samples have distribution function $F(x)=\exp(-\exp(-(x-\omega)/\tau)$. For the null situation $(\omega_1, \omega_2, \omega_3, \omega_4)=(-1.10, -1.10, 0, 0)$ corresponding to scales (4, 4, 1, 1). For H_a the medians were (2.97, 2.97, 0.37, 0.37), obtained using $(\omega_1, \omega_2, \omega_3, \omega_4)=(1.5, 1.5, 0, 0)$ and the same scales.

2 Discreteness of p-values based on comparison of Q to the χ^2_3 distribution prohibits use of the goodness-of-fit χ^2_{10} value as a measure of performance in Row 1.

3 Adjusted power in column 7 is appropriate only if the null situation is that reported in column 3.

TABLE 5

Results for unweighted means analysis and bootstrap procedures for the hypothetical data sets.[1]

1. Inverse pairing $(\sigma_{11}, \sigma_{12}, \sigma_{13}, \sigma_{21}, \sigma_{22}, \sigma_{23}) = (3, 3, 2, 2, 1, 1)$.

	Unweighted means analysis		Bootstrap of studentized unweighted SS		Bootstrap of studentized weighted SS	
Effect	F-Value	p-Value	F-Value	p-Value	F-Value	p-Value
A	2.16	.147	1.43	.237[2]	1.43	.237
B	1.96	.150	1.29	.274	1.30	.273
A*B	6.28	.003	4.15	.034	4.07	.033

2. Direct pairing $(\sigma_{11}, \sigma_{12}, \sigma_{13}, \sigma_{21}, \sigma_{22}, \sigma_{23}) = (1, 1, 2, 2, 3, 3)$

	Unweighted means analysis		Bootstrap of studentized unweighted SS		Bootstrap of studentized weighted SS	
Effect	F-Value	p-Value	F-Value	p-Value	F-Value	p-Value
A	0.48	.493	0.70	.404	0.70	.404
B	1.35	.267	1.97	.150	1.90	.162
A*B	6.60	.003	9.63	.0005	8.61	.0007

1. Data were generated as independent random samples from normal distributions with $(\mu_{11}, \mu_{12}, \mu_{13}, \mu_{21}, \mu_{22}, \mu_{23}) = (1, 0, -.5, -1.5, -.5, 0)$ using sample sizes $(n_{11}, n_{12}, n_{13}, n_{21}, n_{22}, n_{23}) = (8, 8, 8, 8, 16, 16)$.

2. All bootstrap p-values were based on B = 10000 Monte Carlo replications.

References

Babu, G.J., and Singh, K. (1983). Inference on means using the bootstrap. *Annals of Statistics*, 11, 999-1003.

Best, D.J., and Rayner, J.C.W. (1987). On Welch's approximate solution for the Behrens-Fisher problem. *Technometrics*, 29, 205-210.

Bickel, P.J., and Freedman, D.A. (1981). Some asymptotic theory for the bootstrap. *Annals of Statistics*, 9, 1196-1217.

Brown, M.B., and Forsythe, A.B. (1974a). The small sample behavior of some statistics which test the equality of several means. *Technometrics*, 16, 129-132.

Brown, M.B., and Forsythe, A.B. (1974b). The ANOVA and multiple comparisons for data with heterogeneous variances. *Biometrics*, 30, 719-724.

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

Efron, B., and Tibshirani R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, 1, 54-77.

Fligner, M.A., and Policello, G.E. (1981). Robust rank procedures for the Behrens-Fisher problem. *Journal of the American Statistical Asssociation*, 76, 162-168.

Glass, G.V. (1970). *Statistical Methods in Education and Psychology.* New Jersey, Prentice Hall.

Hajek, J., and Sidak, Z. (1967). *Theory of Rank Tests.* New York, Academic Press.

Halperin, M., Gilbert, P.R., and Lachin, J.M. (1987). Distribution-free confidence intervals for Pr ($X_1 < X_2$). *Biometrics*, 43, 71-80.

Pratt, J.W. (1964). Robustness of some procedures for the two-sample location problem. *Journal of the American Statistical Association*, 59, 665-680.

Pratt, J.W., and Gibbons, J.D. (1981). *Concepts of Nonparametric Theory.* New York, Springer-Verlag.

Randles, R.H., and Wolfe, D.A. (1979). *Introduction to the Theory of Nonparametric Statistics.* New York, Wiley.

Satterthwaite, F. E. (1941). Synthesis of variance. *Psychometrika*, 6, 309-316.

Scheffé, H. (1970). Practical solutions of the Behrens-Fisher problem. *Journal of the American Statistical Association*, 65, 1501-1508.

Searle, S.R. (1987). *Linear Models for Unbalanced Data.* New York, Wiley.

Sen, P.K. (1962). On studentized non-parametric multi-sample location tests. *Annals of the Institute of Statistical Mathematics*, 15, 117-135.

Serfling, R.J. (1980). *Approximation Theorems of Mathematical Statistics.* New York, Wiley.

Tomarken, A.J., and Serlin, R.C. (1986). Comparison of ANOVA alternatives under variance heterogeneity and specific noncentrality structures. *Psychological Bulletin*, 99, 90-99.

Van der Vaart, H.R. (1961). On the robustness of Wilcoxon's two-sample test. In *Quantitative Methods in Pharmacology*, H. de Jonge, Ed. New York, Interscience, 140-158.

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

Wolfe, D.A., and Hogg, R.V. (1971). On constructing statistics and reporting data. *The American Statistician*, 25, 27-30.