TWO-SAMPLE BOOTSTRAP TESTS: WHEN TO MIX?

Subhash Lele
Department of Biostatistics
Johns Hopkins University
Baltimore, Maryland 21205

and

Ed Carlstein
Department of Statistics
University of North Carolina
Chapel Hill, North Carolina 27599

SUMMARY

We consider two-sample bootstrap tests for testing equality of a given functional in two populations. The usual bootstrap analogue of the permutation test "mixes" the two samples prior to resampling. It is shown that in the presence of nuisance parameters this test can be invalid or, even if valid, can have bad power characteristics. An alternative bootstrap test procedure is suggested which "mixes" only after resampling from the two separate samples. The new procedure is valid in the presence of nuisance parameters and has better power than the usual procedure under fairly general conditions.

Key words: Bootstrap; Permutation tests; Nuisance parameters; Efficiency.

1. INTRODUCTION

The two-sample hypothesis testing problem is as follows: Let $X_1, X_2,\ldots,X_n$ be a random sample from a distribution F and let $Y_1, Y_2,\ldots,Y_{m_n}$ be another random sample from a distribution G. The experimenter is interested in testing whether a particular characteristic θ, say the mean or the median, is the same for both the distributions. Let the test statistic used for testing this hypothesis be $T(X_1, X_2,\ldots,X_n; Y_1, Y_2,\ldots,Y_{m_n})$. In practice, even with knowledge of the parametric forms for F and G the calculation of the null distribution may be extremely difficult, or, moreo ver, one may not want to assume any particular parametric forms for F and G. One way out of this situation is to use the bootstrap or permutation test procedures (Edgington, 1980; Lambert, 1985; Beran, 1986; Romano, 1988, 1989). These test procedures essentially follow the following steps:

Step 1: Mix the two samples $\underset{\sim}{X}$ and $\underset{\sim}{Y}$. Consider this pooled sample $(X_1,\ldots,X_n, Y_1,\ldots,Y_{m_n})$ and calculate its empirical distribution function H_n say.

Step 2: Generate a random sample (with replacement for bootstrap or without replacement for permutation) of size $(n+m_n)$ from H_n, say $(Z_1^*,\ldots,Z_n^*, Z_{n+1}^*,\ldots,Z_{n+m_n}^*)$. Split this sample in two parts $(Z_1^*,\ldots,Z_n^*)$ and $(Z_{n+1}^*,\ldots,Z_{n+m_n}^*)$.

Step 3: Calculate the test statistic $T^* = T(Z_1^*,\ldots,Z_n^*; Z_{n+1}^*,\ldots,Z_{n+m_n}^*)$.

Step 4: Repeat steps 2 and 3 a large number of times, say B, to get T_j^*, $j = 1,2,\ldots,B$.

We obtain the cutoff points on the basis of the distribution of T_j^*.

If F and G are identical under H_0, this procedure would be valid; Beran (1986) and Romano (1988, 1989) make this statement precise. However, when F and G are in fact different, it may be because the parameter of interest is different (i.e., the alternative hypothesis is true) or just because the nuisance parameters are different (i.e., the null hypothesis is true but F and G are still not identical); thus it is not clear if the above procedure (using "prior mixing") is sensible. The difficulty is that Step 2 implicitly assumes that F = G under H_0.

Our motivation to study this problem came from the morphometric problems discussed in Lele and Richtsmeier (1990). In these situations we are more interested in the equality of the mean shapes than the equality of distributions. Similar situations are abundant in practice where one is interested in the equality of a particular parameter, in the presence of nuisance parameters which may or may not be equal in the two distributions. The purpose of this paper is to study the performance of bootstrap tests in such situations. Note that Beran (1986), Lambert (1985) and Romano (1988, 1989) assume that the nuisance parameters are equal.

In the next Section we discuss two examples where "prior mixing" leads to an invalid test or a test with miserably low power. Section 3 presents a general formulation of the two-sample hypothesis testing problem, and Section 4 proposes a bootstrap test procedure based on "posterior mixing". The properties of prior versus posterior mixing procedures are compared in Section 5, and several examples are analyzed in Section 6.

2. TWO EXAMPLES

Example 1: Invalid bootstrap test

In this example we consider a situation where F and G are not identical, but the null hypothesis of equality regarding a particular parameter, the median, is true. We show that in such a situation, the "prior mixing" bootstrap procedure can lead to a sampling distribution other than the true null distribution, thus not even getting the correct level of significance.

Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed F and $Y_1, Y_2, \ldots, Y_n$ be independent and identically distributed G. Let F and G be absolutely continuous with densities f and g respectively. We are interested in testing the null hypothesis of equality of medians. Let the test statistic be:

$$T_n = \sqrt{n}\,[\underset{1\le i\le n}{\text{median}}\{X_i\} - \underset{1\le i\le n}{\text{median}}\{Y_i\}] .$$

Let zero be the median for both the distributions but let $f(0) \neq g(0)$, with f(0) and g(0) positive. Then the true asymptotic distribution of T_n is

$$T_n \approx N\left(0, \frac{1}{4}\left(\frac{1}{f^2(0)} + \frac{1}{g^2(0)}\right)\right)$$

(Serfling, 1980) where $\approx$ means "approximately distributed."

The pooled sample empirical distribution function H_n approximates the distribution $\frac{1}{2}$ (F+G) with density $\frac{1}{2}$ (f+g). Let $\hat{T}_n^*$ denote the resampled version obtained by the "prior mixing" bootstrap procedure. Then

$$\hat{T}_n^* \approx N\left(0, 2\left(\frac{1}{f(0) + g(0)}\right)^2\right)$$

(see Bickel and Freedman (1981)). However, note that

$$\frac{1}{4}\left(\frac{1}{f^2(0)} + \frac{1}{g^2(0)}\right) \neq \frac{2}{[f(0) + g(0)]^2} ,$$

and in fact the discrepancy can be arbitrarily large as f(o) (say) approaches zero.

Thus the "prior mixing" bootstrap procedure fails to yield the correct cutoff points when $F \neq G$, even though the null hypothesis holds.

Example 2: Valid but inefficient bootstrap test

In this example, the "prior mixing" bootstrap procedure does lead to a correct null distribution when the null hypothesis is true. However, if the alternative hypothesis holds, the bootstrap procedure almost always accepts the null hypothesis.

Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed double-exponential (Laplace) random variables with parameters θ_F and 1. Let $Y_1, Y_2, \ldots, Y_n$ be independent and identically distributed double-exponential random variables with parameters θ_G and 1. We are interested in testing the equality of the medians, i.e., H_0: $\theta_F = \theta_G$ versus H_1: $\theta_F \neq \theta_G$. Let the test statistic be

$$T_n = \sqrt{n}\,[\underset{1\le i\le n}{\text{median}}\{X_i\} - \underset{1\le i\le n}{\text{median}}\{Y_i\}].$$

The asymptotic distribution of T_n is:

$$T_n \approx N(\sqrt{n}(\theta_F - \theta_G),\ 2).$$

The distribution obtained from "prior mixing" bootstrap procedure is

$$\hat{T}_n^* \approx N\left(0,\ \frac{1}{2}\left(\frac{1}{h^2\left(\frac{\theta_F + \theta_G}{2}\right)}\right)\right)$$

where $h(x) = \frac{1}{2} f(x) + \frac{1}{2} g(x)$. Now suppose $\theta_F = 10$ and $\theta_G = 0$, and $n = 100$ (which should be large enough for approximate normality). Then $T_n \approx N(100,2)$ and $\hat{T}_n^* \approx N(0, 44053)$. Therefore, T_n almost always lies in the interval (95.76, 104.24), while the critical regions obtained via $\hat{T}_n^*$ at various levels are:

$$CR(\alpha = 0.01) = \{(-\infty, -541.5) \cup (541.5, \infty)\},$$
$$CR(\alpha = 0.05) = \{(-\infty, -411.4) \cup (411.4, \infty)\}.$$

Thus we never reject the null hypothesis.

The moral of the two examples is the following:

1: If there are nuisance parameters which are unequal in F and G even under H_0, the "prior mixing" bootstrap test procedure may not be valid.

2: Since we are resampling from the mixed distribution, under H_1 the power of the "prior mixing" bootstrap test procedure can be near zero.

In the following sections we develop a general theory for an alternative bootstrap test procedure, based on "posterior mixing", which corrects these difficulties.

3. STATEMENT OF THE PROBLEM

Consider the general two-sample problem: We observe two independent samples $\underset{\sim}{X}_n := (X_1, X_2, \ldots, X_n)$ and $\underset{\sim}{Y}_n := (Y_1, Y_2, \ldots, Y_{m_n})$, where the X_i's are independent identically distributed random variables from F, and the Y_i's are independent identically distributed random variables from G.

We assume that (F, G) belongs to *H*, a class of distribution pairs

which may be, for example: all distributions with finite variance, or all continuous distributions, or a parametric family of distributions such as Poisson distributions.

We are interested in a functional $\theta(\cdot)$ and the hypotheses H_0: $\theta(F)=\theta(G)$ versus H_A: $\theta(F) \neq \theta(G)$. In terms of the class H, this can be stated as:

$$H_0\text{: } (F, G) \in H_0\text{: } = \{(F, G) \in H\text{: } \theta(F) = \theta(G)\}$$
$$H_A\text{: } (F, G) \in H_A\text{: } = \{(F, G) \in H\text{: } \theta(F) \neq \theta(G)\}.$$

Our test statistic is T_n: $= T_n(\underset{\sim}{X}_n, \underset{\sim}{Y}_n)$ and the rejection rule is of the form: " Reject H_0 if and only if $|T_n| > C_n^\alpha$ " where α is fixed.

The subject of our study is: how to select C_n^α via a bootstrap procedure.

3.1 Validity of a procedure

Suppose H_0 is true, i.e., $(F, G) \in H_0$, and that there exists $\xi_n^\alpha(F, G)$ such that

$$P_{F,G}\{|T_n| > \xi_n^\alpha(F, G)\} = \alpha.$$

In general, $\xi_n^\alpha(F, G)$ may not be available explicitly, either because it is analytically intractable to determine, or because it depends on the unknown F and G nontrivially.

If C_n^α is approximately equal to $\xi_n^\alpha(F, G)$ then we say that our rejection rule is approximately valid, i.e., $P_{F,G}\{\text{Reject } H_0\} = P_{F,G}\{|T_n| > C_n^\alpha\} \approx \alpha$.

3.2 Power of a procedure

Suppose H_A is true, i.e., $(F, G) \in H_A$. Then the power of the procedure is given by $\beta_n^\alpha(F, G)$: $= P_{F,G}\{|T_n| > C_n^\alpha\}$. Moreover if, as $n \to \infty$, $\beta_n^\alpha(F, G) \to 1$ then our test procedure is consistent.

3.3 "Validity" of a procedure under H_A

In general, $\xi_n^\alpha(F, G)$ depends on the particular common value of $\theta(F) = \theta(G)$ and the nuisance parameters of F and G. However there are some situations where $\xi_n^\alpha(F, G)$ is invariant for all $(F, G) \in H_0$. In this case, if, even under H_A, C_n^α is approximately equal to ξ_n^α, then our test procedure is approximately "valid" under H_A. (See Section 6, Example D and Section 5.1.2.)

Our aim is to select C_n^α such that the resulting test procedure is at least approximately valid and has good power.

4. METHODS FOR SELECTING C_n^α

Let us denote by:

$F_n(\cdot)$: the e.c.d.f. generated by $\underset{\sim}{X}_n$, i.e., a distribution which puts mass $\frac{1}{n}$ at each X_i, $1 \le i \le n$;

$G_n(\cdot)$: the e.c.d.f. generated by $\underset{\sim}{Y}_n$, i.e., a distribution which puts mass $\frac{1}{m_n}$ at each Y_i, $1 \le i \le m_n$;

$H_n(\cdot)$: the e.c.d.f. generated by $(\underset{\sim}{X}_n, \underset{\sim}{Y}_n)$, i.e., a distribution which puts mass $\frac{1}{n+m_n}$ at each of X_i, $1 \le i \le n$, and Y_j, $1 \le j \le m_n$.

Let us assume that T_n is of the form:

$$T_n := T_n(\underset{\sim}{X}_n, \underset{\sim}{Y}_n) = \sqrt{n}\,[t_n(F_n) - t_{m_n}(G_n)]$$

where $\{t_n(\cdot), n \ge 1\}$ are functionals which estimate $\theta(\cdot)$.

We will consider bootstrap methods for determining C_n^α. In general, we will generate bootstrap replicates T_n^* of T_n, and use the percentiles of T_n^*'s distribution (conditional on $(\underset{\sim}{X}_n, \underset{\sim}{Y}_n)$) for calculating C_n^α. That

is, C_n^α is the solution of:

$$P[|T_n^*| > C_n^\alpha | \underset{\sim}{X}_n, \underset{\sim}{Y}_n] = \alpha \ , \qquad (*)$$

assuming for simplicity that an exact solution exists. The question is: how to generate T_n^* from $(\underset{\sim}{X}_n, \underset{\sim}{Y}_n)$, in order to obtain a reasonable C_n^α. We will consider two bootstrap algorithms: $\hat{T}_n^*$ denoting the bootstrap version of T_n using PRIOR mixing, and $\tilde{T}_n^*$ denoting the bootstrap version of T_n using POSTERIOR mixing.

4.1 Prior mixing algorithm

This algorithm is analogous to the two-sample permutation test procedure (Romano 1988, 1989).

(1) Generate $\underset{\sim}{X}_n^* := (X_{n1}^*, \ldots, X_{nn}^*)$ with X_{ni}^* (conditionally) i.i.d. r.v.s from $H_n(\cdot)$.

Generate $\underset{\sim}{Y}_n^* := (Y_{n1}^*, \ldots, Y_{nm_n}^*)$ with Y_{ni}^* (conditionally) i.i.d. r.v.s from $H_n(\cdot)$.

(2) Define F_n^* as the e.c.d.f. of $\underset{\sim}{X}_n^*$, and G_n^* as the e.c.d.f. of $\underset{\sim}{Y}_n^*$.

(3) Calculate $\hat{T}_n^* := T_n(\underset{\sim}{X}_n^*, \underset{\sim}{Y}_n^*) = \sqrt{n}[t_n(F_n^*) - t_{m_n}(G_n^*)]$. The distribution of $\hat{T}_n^*$ yields $\hat{C}_n^\alpha$ by solving equation (*).

It seems that the above algorithm should yield a good approximation to the null distribution of T_n, since here our X_{ni}^*'s and Y_{ni}^*'s both come from distributions with the same value of $\theta(\cdot)$, viz. $\theta(H_n)$. But in fact this algorithm goes too far in matching up the distributions of X_{ni}^* and Y_{ni}^*. It actually makes the entire distributions equal, instead of just

matching their $\theta(\cdot)$ values. The next algorithm retains information about how F and G may differ (in features other than $\theta(\cdot)$), by resampling separately from F_n and G_n and mixing afterwards.

4.2 Posterior mixing algorithm

Part A:

Generate ${}_F\underset{\sim}{X}^*_n := ({}_FX^*_{n1}, {}_FX^*_{n2}, \ldots, {}_FX^*_{nn})$ with ${}_FX^*_{ni}$ (conditionally) i.i.d. from F_n.

Generate ${}_F\underset{\sim}{Y}^*_n := ({}_FY^*_{n1}, {}_FY^*_{n2}, \ldots, {}_FY^*_{nm_n})$ with ${}_FY^*_{ni}$ (conditionally) i.i.d. from F_n.

Let ${}_FF^*_n$ be the e.c.d.f. of ${}_F\underset{\sim}{X}^*_n$ and ${}_FG^*_n$ be the e.c.d.f. of ${}_F\underset{\sim}{Y}^*_n$.

Calculate ${}_FT^*_n := T_n({}_F\underset{\sim}{X}^*_n, {}_F\underset{\sim}{Y}^*_n) = \sqrt{n}\,[t_n({}_FF^*_n) - t_{m_n}({}_FG^*_n)]$.

Part B: (Similar to part A.)

Generate ${}_G\underset{\sim}{X}^*_n := ({}_GX^*_{n1}, {}_GX^*_{n2}, \ldots, {}_GX^*_{nn})$ with ${}_GX^*_{ni}$ (conditionally) i.i.d. from G_n.

Generate ${}_G\underset{\sim}{Y}^*_n := ({}_GY^*_{n1}, {}_GY^*_{n2}, \ldots, {}_GY^*_{nm_n})$ with ${}_GY^*_{ni}$ (conditionally) i.i.d. from G_n.

Let ${}_GF^*_n$ and ${}_GG^*_n$ be the e.c.d.f.s of ${}_G\underset{\sim}{X}^*_n$ and ${}_G\underset{\sim}{Y}^*_n$ respectively.

Calculate ${}_GT^*_n := T_n({}_G\underset{\sim}{X}^*_n, {}_G\underset{\sim}{Y}^*_n) = \sqrt{n}\,[t_n({}_GF^*_n) - t_{m_n}({}_GG^*_n)]$.

Part C: Calculate $\tilde{T}_n^* := \sqrt{\pi_n}\ {}_F T_n^* + \sqrt{(1-\pi_n)}\,{}_G T_n^*$ where $\pi_n := \frac{m_n}{n + m_n}$.

The distribution of $\tilde{T}_n^*$ yields the cutoff value $\tilde{C}_n^{\alpha}$, by solving equation (*).

5. COMPARISON OF PRIOR AND POSTERIOR MIXING SCHEMES

In this Section, we compare the two algorithms on the basis of validity and power.

Assumptions:

A1: $\pi_n \to \pi \in (0, 1)$ as $n \to \infty$; denote $\bar{\pi} := 1-\pi$.

A2: We assume that $t_n(\cdot)$ estimates $\theta(\cdot)$ in the following sense.

If Q_n is the e.c.d.f. based on n i.i.d. observations from a distribution Q, then

$$\sqrt{n}[t_n(Q_n) - \theta(Q)] \to N(0, \sigma^2(Q))$$

in distribution as $n \to \infty$. Here $\sigma^2(Q)$ is not necessarily the variance of Q; the form of the functional $\sigma^2(\cdot)$ depends on $t_n(\cdot)$ and $\theta(\cdot)$.

Examples of such $t_n(\cdot)$ and $\theta(\cdot)$ (under appropriate regularity conditions) include:

(1) *Differentiable Statistical Functionals* $\theta(\cdot)$: Take $t_n(Q_n)$ as $\theta(Q_n)$.

(2) *M-Estimators*: Here $\theta(Q)$ is a solution of the equation $\int \psi(x, \theta)dQ(x) = 0$ for some function ψ. We take $t_n(Q_n)$ as $\theta(Q_n)$.

(3) *V-Statistics*: Here $\theta(Q) = \int\int\ldots\int \phi(x_1, x_2,\ldots,x_k) \prod_{i=1}^{k} dQ(x_i)$ and $t_n(Q_n) = \theta(Q_n)$.

(4) *U-Statistics*: Here $\theta(Q) = \int\int\ldots\int \phi(x_1,\ldots,x_k) \prod_{i=1}^{k} dQ(x_i)$ and

$t_n(Q_n) = \frac{1}{\binom{n}{k}} \sum_n^* \phi(x_1, x_2, \ldots, x_k)$ where $\sum_n^*$ is the sum over all $\binom{n}{k}$ possible samples of size k obtained without replacement from Q_n. Under A1 and A2, we find that for $(F, G) \in H$, T_n is approximately normal with mean $\sqrt{n}(\theta(F) - \theta(G))$ and variance $\sigma^2(F) + \frac{\bar{\pi}}{\pi} \sigma^2(G)$. Thus under H_o, this yields

$$\xi_n^\alpha(F, G) \approx Z_{\alpha/2} \sqrt{\sigma^2(F) + \frac{\bar{\pi}}{\pi} \sigma^2(G)} =: \xi_\pi^\alpha (F, G),$$

where $Z_{\alpha/2}$ is the $(1 - \alpha/2)$ percentile for a standard normal distribution.

5.1 Bootstrap procedures

To evaluate the bootstrap procedures for selecting C_n^α, we need an assumption analogous to A2:

A3: Let Q_n' be an estimate of distribution Q, based on O(n) observations. If Q_n^* is the e.c.d.f. based on n (conditionally) i.i.d. observations from Q_n', then:

$$\sqrt{n}[t_n(Q_n^*) - \theta(Q_n')] \to N(0, \sigma^2(Q))$$

in (conditional) distribution as $n \to \infty$, a.s. Here $\sigma^2(\cdot)$ is the same functional as in A2.

Note that Q_n' is not necessarily the e.c.d.f. of n observations that are i.i.d. from Q. We will only need to invoke A3 for reasonable choices of Q_n' (See Sections 5.1.1 and 5.1.2).

5.1.1 Prior mixing

In this situation, the roles of Q_n' and Q are played by H_n and $H_\pi := \bar{\pi}\ F + \pi\ G$, respectively. Note that $H_n(\cdot) \to H_\pi(\cdot)$ a.s. as $n \to \infty$, although H_n is based on $n+m_n$ observations which are not i.i.d. from H_π.

Using A1 and A3, we find that for $(F, G) \in H$, $\hat{T}_n^*$ is approximately $N(0, \sigma^2(H_\pi)/\pi)$ (conditionally), yielding

$$\hat{C}_n^\alpha \approx \hat{C}_\pi^\alpha(F, G) := \sqrt{\frac{\sigma^2(H_\pi)}{\pi}} \times Z_{\alpha/2} \ .$$

Validity: Suppose H_0 is true, i.e., $(F, G) \in H_0$. For validity of the prior mixing algorithm, we need $\hat{C}_\pi^\alpha(F, G) = \xi_\pi^\alpha(F, G)$, or equivalently

$$\sigma^2(H_\pi) = \pi\ \sigma^2(F) + \bar{\pi}\ \sigma^2(G).$$

This, in general, does not hold (recall Example 1). We can obtain validity in the following two special cases:

(i) $F = G$.

(ii) Suppose $\sigma^2(\cdot)$ is a linear functional, i.e., $\sigma^2(H_\pi) = \bar{\pi}\ \sigma^2(F) + \pi\ \sigma^2(G)$.

Then prior mixing is valid if and only if $\sigma^2(F) = \sigma^2(G)$ or $\pi = \frac{1}{2}$.

Notice that both cases involve restriction of the nuisance parameters.

Power: For $(F, G) \in H_A$, the power of the prior mixing procedure is

$$\hat{\beta}_n^\alpha(F, G) \approx P_{F,G}\{|T_n| > \hat{C}_\pi^\alpha(F, G)\} \approx \Phi\left(\frac{-\sqrt{n}[\theta(F) - \theta(G)] - \hat{C}_\pi^\alpha(F, G)}{\sqrt{\sigma^2(F) + \frac{\bar{\pi}}{\pi}\sigma^2(G)}}\right)$$

$$+ \Phi\left(\frac{\sqrt{n}\ [\theta(F) - \theta(G)] - \hat{C}_\pi^\alpha\ (F, G)}{\sqrt{\sigma^2(F) + \frac{\bar{\pi}}{\pi}\sigma^2(G)}}\right) .$$

Since $\hat{\beta}_n^{\alpha}(F, G) \to 1$ as $n \to \infty$, this procedure is consistent.

5.1.2 Posterior mixing

Using A3 and A1, we find that for $(F, G) \in H$, $\tilde{T}_n^* \approx N\left(0, \sigma^2(F) + \frac{\bar{\pi}}{\pi}\sigma^2(G)\right)$ (conditionally), yielding $\tilde{C}_n^{\alpha} \approx \tilde{C}_{\pi}^{\alpha}(F, G) := Z_{\alpha/2}\sqrt{\sigma^2(F) + \frac{\bar{\pi}}{\pi}\sigma^2(G)}$.

Validity: Suppose H_0 is true, i.e., $(F, G) \in H_0$; then $\tilde{C}_{\pi}^{\alpha}(F, G) = \xi_{\pi}^{\alpha}(F, G)$. Thus the posterior mixing algorithm always yields a valid procedure, even in the presence of unequal nuisance parameters.

Now suppose that the nuisance parameters are fixed (but possibly unequal), i.e., $\sigma^2(F) = \sigma_1^2$ and $\sigma^2(G) = \sigma_2^2$ for all $(F, G) \in H$. Then $\xi_{\pi}^{\alpha}(F, G) \equiv \xi_{\pi}^{\alpha}$ is invariant for all $(F, G) \in H_0$. In this case we have $\tilde{C}_{\pi}^{\alpha}(F, G) \equiv \xi_{\pi}^{\alpha}$ for all $(F, G) \in H_A$, and thus posterior mixing yields a "valid" test procedure even when the alternative holds. (See Section 6, Example D.)

Power: For $(F, G) \in H_A$, the power for posterior mixing is

$$\tilde{\beta}_n^{\alpha}(F, G) \approx P_{F,G}\{|T_n| > \tilde{C}_{\pi}^{\alpha}(F, G)\} \approx \Phi\left(\frac{-\sqrt{n}[\theta(F) - \theta(G)] - \tilde{C}_{\pi}^{\alpha}(F, G)}{\sqrt{\sigma^2(F) + \frac{\bar{\pi}}{\pi}\sigma^2(G)}}\right)$$

$$+ \Phi\left(\frac{\sqrt{n}[\theta(F) - \theta(G)] - \tilde{C}_{\pi}^{\alpha}(F, G)}{\sqrt{\sigma^2(F) + \frac{\bar{\pi}}{\pi}\sigma^2(G)}}\right).$$

Since $\tilde{\beta}_n^{\alpha}(F, G) \to 1$ as $n \to \infty$, the posterior mixing procedure is consistent.

5.1.3 Power comparison between prior and posterior mixing

For $(F, G) \in H_A$, our asymptotic approximations (above) show that posterior mixing will have better power than prior mixing if $\tilde{C}_\pi^\alpha(F, G) \leq \hat{C}_\pi^\alpha(F, G)$, or equivalently:

$$\pi\, \sigma^2(F) + \bar{\pi}\, \sigma^2(G) \leq \sigma^2(H_\pi) \ . \qquad (\ddagger)$$

Suppose $\sigma^2(\cdot)$ is a concave functional, i.e., $\sigma^2(H_\pi) \geq \bar{\pi}\, \sigma^2(F) + \pi\, \sigma^2(G)$; then to obtain $(\ddagger)$ it suffices to have either

(a) $\sigma^2(F) \geq \sigma^2(G)$ and $\pi \leq \frac{1}{2}$,

or (b) $\sigma^2(F) \leq \sigma^2(G)$ and $\pi \geq \frac{1}{2}$.

The relative sampling proportions in (a) and (b) correspond to a logical sampling design: a larger sample size is drawn from the population where the statistic has larger variance.

6. EXAMPLES

In this Section we apply the analysis from Section 5 to several examples. We assume A1 throughout.

Example A: The sample mean (general case)

Let $H = \{(F, G)$: F and G have finite variances$\}$, $\theta(Q) = \int x dQ(x)$, $t_n(Q) = \theta(Q)$ for all n. In this scenario, A2 holds with $\sigma^2(Q) = \int (x - \theta(Q))^2\, dQ(x)$. The approximations from Sections 5.1.1 and 5.1.2 can be formalized by the following theorems.

Theorem 1: (Prior Mixing)

$$\sup_{u \in R} \left| P\{\hat{T}_n^* \le u | \underset{\sim}{X}_n, \underset{\sim}{Y}_n\} - P\left(N\left(0, \frac{\sigma^2(H_\pi)}{\pi}\right) \le u\right)\right| \longrightarrow 0 \quad \text{a.s. as } n \to \infty$$

where $\sigma^2(H_\pi) = \bar{\pi}\, \sigma^2(F) + \pi\, \sigma^2(G) + \pi\, \bar{\pi}(\theta(F) - \theta(G))^2$.

Theorem 2: (Posterior Mixing)

$$\sup_{u \in R} \left| P\{\tilde{T}_n^* \le u | \underset{\sim}{X}_n, \underset{\sim}{Y}_n\} - P\left(N\left(0, \sigma^2(F) + \frac{\bar{\pi}}{\pi}\, \sigma^2(G)\right) \le u\right)\right| \longrightarrow 0 \quad \text{a.s. as } n \to \infty .$$

Theorem 1 is proved by a straightforward modification of the argument used in Theorem 1.A of Singh (1981). Theorem 2 follows directly from Theorem 2.1 of Bickel and Freedman (1981).

Observe that $\sigma^2(\cdot)$ is a linear functional for $(F, G) \in H_0$ and it is a concave functional for $(F, G) \in H_A$. So, for $\pi \neq \frac{1}{2}$, prior mixing is valid only if the nuisance parameters are equal, i.e., $\sigma^2(F) = \sigma^2(G)$; posterior mixing is always valid. And, for a logical sampling proportion π, posterior mixing will have better power than prior mixing. Moreover, for fixed $(\pi, \sigma^2(F), \sigma^2(G))$ we have (‡) for all $|\theta(F) - \theta(G)|$ sufficiently large.

Example B: The sample mean (Poisson case). Let $H = \{(F, G)$: F and G are Poisson$\}$, and let $\theta(\cdot)$, $t_n(\cdot)$ and $\sigma^2(\cdot)$ be as in Example A. Since $F = G$ under H_0, we find that both the prior as well as posterior mixing procedures are valid. However under H_A, we find that the posterior mixing procedure has better power than prior mixing, for any fixed π, whenever $|\theta(F) - \theta(G)| > |\pi - \bar{\pi}|/\pi\, \bar{\pi}$.

Example C: The sample median (general case) Consider the situation in Example 1, i.e., $H = \{(F, G)$: F and G are absolutely continuous with densities f and g that are positive at their respective medians$\}$, $\theta(\cdot)$ and $t_n(\cdot)$ are medians, and $\pi_n \equiv \frac{1}{2}$. Here A2 holds with $\sigma^2(Q) = \frac{1}{4q^2(\theta(Q))}$, so that under H_0, $\sigma^2(H_\pi) = \frac{1}{[f(\theta(F)) + g(\theta(F))]^2}$ while

$\pi\sigma^2(F) + \bar{\pi}\sigma^2(G) = \frac{1}{8}\left[\frac{1}{f^2(\theta(F))} + \frac{1}{g^2(\theta(F))}\right]$. Therefore, prior mixing is valid if and only if $f(\theta(F)) = g(\theta(F))$. Observe that $\sigma^2(\cdot)$ is not a linear functional.

Example D: The sample median (Double-exponential case) Consider the situation in Example 2, i.e., $H = \{(F, G)$: F and G are double-exponential with scale 1$\}$; $\theta(\cdot)$, $t_n(\cdot)$, π_n, and $\sigma^2(\cdot)$ are as in Example C. Under H_0 we have $F = G$, hence the prior as well as posterior mixing procedures are valid. Note that for all $(F, G) \in H$, $\sigma^2(F) \equiv \sigma^2(G) \equiv 1$; thus the posterior mixing procedure is "valid" even when the alternative is true. Moreover, for $\theta(F) = 10$, $\theta(G) = 0$, and $n = m_n = 100$, the power approximations for the two procedures yield:

	$\alpha = 0.05$	$\alpha = 0.01$
Prior mixing	0.00	0.00
Posterior mixing	1.00	1.00

In fact, for all $(F, G) \in H_A$, we get $\sigma^2(H_\pi) = \exp\{|\theta(F) - \theta(G)|\} > 1 = \pi\,\sigma^2(F) + \bar{\pi}\,\sigma^2(G)$ and hence a strict increase in power by using the posterior mixing algorithm.

ACKNOWLEDGEMENTS

This work was initiated during the first author's visit to the Department of Statistics, University of North Carolina at Chapel Hill. We thank Professor P. K. Sen for helpful discussions. The second author's work was supported by N.S.F. Grant #DMS-8902973.

REFERENCES

BERAN, R. (1986). Simulated power functions. Ann. Statist. 14, 151-173.

BICKEL, P. and FREEDMAN, D. (1981). Some asymptotic theory for the bootstrap. Ann. Statist. 9, 1196-1217.

EDGINGTON, E. S. (1980). Randomization Tests. New York: Dekker.

LAMBERT, D. (1985). Robust two sample permutation tests. Ann. Statist. 13, 606-25.

LELE, S. and RICHTSMEIER, J. T. (1990). A coordinate free approach for comparing biological shapes I. Landmark data. Am. J. Phys. Anthropol. (to appear).

ROMANO, J. P. (1988). A bootstrap revival of some nonparametric distance tests. J. Am. Statist. Assoc. 83, 698-708.

ROMANO, J. P. (1989). Bootstrap and randomization tests of some non-parametric hypothesis. Ann. Statist. 17, 141-59.

SERFLING, R. J. (1980). Approximation Theorems of Mathematical Statistics. New York: John Wiley and Sons.

SINGH, K. (1981). On the asymptotic accuracy of Efron's bootstrap. Ann. Statist. 9, 1187-1195.