The Number of Samples for Resampling Algorithms

David J. Olive*

Southern Illinois University

June 23, 2011

Abstract

Resampling algorithms have many applications, including the bootstrap, the delete-d jackknife and randomization tests. Often theoretical results are given for the impractical algorithm that uses all possible samples, but the impractical algorithm is replaced by a practical algorithm that uses $B = 1000$ samples. This paper shows that using $m = \max(B, [n \log(n)])$ samples results in a practical algorithm that has good theoretical properties.

KEY WORDS: Bootstrap, Jackknife; Randomization Tests.

*David J. Olive is Associate Professor, Department of Mathematics, Southern Illinois University, Carbondale, IL 62901 (e-mail: dolive@math.siu.edu). This work was supported by the National Science Foundation under grant DMS 0600933.

1 INTRODUCTION

Resampling algorithms have many applications. See, for example, Chernick (2008), Davison and Hinkley (1997), Efron and Tibshirani (1993), Good (2005) and Polansky (2008).

Although theory for resampling algorithms given in Lehmann (1999, p. 425) and Sen and Singer (1993, p. 365) has the number of samples $m \to \infty$ as the sample size $n \to \infty$, much of the literature suggests using $m = B$ between 999 and 10000. This choice is often justified using simulations and binomial approximations. An exception is Shao (1989) where $n/m \to 0$ as $n \to \infty$. Let $[x]$ be the integer part of x, so $[7.7] = 7$. Then $m = [n^{1.01}]$ may give poor results for $n < 900$. To combine theory with empirical results, we suggest using $m = \max(B, [n\log(n)])$. Justification for this choice is given below and in the following section.

Often resampling algorithms are used to provide information about the sampling distribution of a statistic $T_n \equiv T_n(F) \equiv T_n(\boldsymbol{Y}_n)$ where $\boldsymbol{Y}_n = (Y_1, ..., Y_n)^T$ and the Y_i are iid from a distribution with cumulative distribution function (cdf) $F(y) = P(Y \leq y)$. Then T_n has a cdf $H_n(y) = P(T_n \leq y)$. If $F(y)$ is known, then m independent samples $\boldsymbol{Y}^*_{j,n} = (Y^*_{j,1}, ..., Y^*_{j,n})^T$ of size n could be generated, where the $Y^*_{j,k}$ are iid from a distribution with cdf F and $j = 1, ..., m$. Then the statistic T_n is computed for each sample, resulting in m statistics $T_{1,n}(F), ..., T_{m,n}(F)$ which are iid from a distribution with cdf $H_n(y)$. Equivalent notation $T_{i,n}(F) \equiv T^*_{i,n}(\boldsymbol{Y}^*_{i,n})$ is often used, where $i = 1, ..., m$.

If $W_1, ..., W_m$ are iid from a distribution with cdf F_W, then the empirical cdf F_m corresponding to F_W is given by

$$F_m(y) = \frac{1}{m}\sum_{i=1}^{m} I(W_i \leq y)$$

where the indicator $I(W_i \leq y) = 1$ if $W_i \leq y$ and $I(W_i \leq y) = 0$ if $W_i > y$. Fix m and y. Then $mF_m(y) \sim$ binomial $(m, F_W(y))$. Thus $E[F_m(y)] = F_W(y)$ and $V[F_m(y)] = F_W(y)[1 - F_W(y)]/m$. By the central limit theorem,

$$\sqrt{m}(F_m(y) - F_W(y)) \xrightarrow{D} N(0, F_W(y)[1 - F_W(y)]).$$

Thus $F_m(y) - F_W(y) = O_P(m^{-1/2})$, and F_m is a reasonable estimator of F_W if the number of samples m is large.

Let $W_i = T_{i,n}(F)$. Then $F_m \equiv \tilde{H}_{m,n}$ is an empirical cdf corresponding to H_n. Let $W_i = Y_i$ and $m = n$. Then F_n is the empirical cdf corresponding to F. Let $\boldsymbol{y}_n = (y_1, ..., y_n)^T$ be the observed data. Now F_n is the cdf of the population that consists of $y_1, ..., y_n$ where the probability of selecting y_i is $1/n$. Hence an iid sample of size d from F_n is obtained by drawing a sample of size d with replacement from $y_1, ..., y_n$. If $d = n$, let $\boldsymbol{Y}^*_{j,n} = (Y^*_{j,1}, ..., Y^*_{j,n})$ be an iid sample of size n from the empirical cdf F_n. Hence each $Y^*_{j,k}$ is one of the $y_1, ..., y_n$ where repetition is allowed. Take m independent samples from F_n and compute the statistic T_n for each sample, resulting in m statistics $T_{1,n}(F_n), ..., T_{m,n}(F_n)$ where $T_{i,n}(F_n) \equiv T^*_{i,n}(\boldsymbol{Y}^*_{i,n})$ for $i = 1, ..., m$. This type of sampling can be done even if F is unknown, and if $T_n(F_n) \approx T_n(F)$, then the empirical cdf based on the $T_{i,n}(F_n)$ may be a useful approximation for H_n.

For general resampling algorithms let $T^*_{i,n}(\boldsymbol{Y}^*_{i,n})$ be the statistic based on a randomly chosen sample $\boldsymbol{Y}^*_{i,n}$ used by the resampling algorithm. Let $H_{A,n}$ be the cdf of the $T^*_{i,n}$ based on all J_n possible samples, and let $H_{m,n}$ be the cdf of the $T^*_{i,n}$ based on m randomly chosen samples. Often theoretical results are given for $H_{A,n}$ but are not known for $H_{m,n}$. Let $G_{N,n}$ be a cdf based on a normal approximation for H_n. Central limit type theorems are used and $G_{N,n}$ is often first order accurate: $H_n(y) - G_{N,n}(y) = O_P(n^{-1/2})$. Approximations $G_{E,n}$ based on the Edgeworth expansion (which is not a cdf) and $H_{A,n}$ are sometimes second order accurate: $H_n(y) - H_{A,n}(y) = O_P(n^{-1})$.

Theory for resampling algorithms such as first order accuracy of the bootstrap and the power of randomization tests is usually for the impractical algorithm that uses all J_n samples. See Hall (1988), Hoeffding (1952), Robinson (1973) and Romano (1989). Practical algorithms use B randomly drawn samples where B is chosen to give good performance when n is small. The following section shows that using $m = \max(B, [n\log(n)])$ randomly drawn samples results in a practical algorithm that is asymptotically equivalent to the impractical algorithm up to terms of order $n^{-1/2}$ while also having good small sample performance. The following two examples follow DasGupta (2008, pp. 462, 469, 513).

Example 1. Let $Y_1, ..., Y_n$ be iid with cdf F. Then the *ordinary bootstrap distribution* of T_n is $H_{A,n}(y) = P_{F_n}(T_n(\boldsymbol{Y}^*_{i,n}) \leq y)$ where $\boldsymbol{Y}^*_{i,n} = (Y^*_{i,1}, ..., Y^*_{i,n})$ is an iid sample of size n from the empirical cdf F_n obtained by selecting with replacement from $Y_1, ..., Y_n$. Here $T^*_{i,n}(\boldsymbol{Y}^*_{i,n}) = T_n(\boldsymbol{Y}^*_{i,n})$. Note that there are $J_n = n^n$ ordered samples and $n^n/n!$ unordered samples from F_n. The bootstrap distribution $H_{m,n}$ typically used in practice is based on m samples randomly selected with replacement. Both $H_{A,n}$ and $H_{m,n}$ are estimators of H_n, the cdf of T_n.

Example 2. Let $X_1, ..., X_{k_1}$ be iid with probability density function (pdf) $f(y)$ while $Y_1, ..., Y_{k_2}$ are iid with pdf $f(y - \mu)$. Let $n = k_1 + k_2$ and consider testing $H_0 : \mu = 0$. Let $T_n \equiv T_{k_1,k_2}$ be the two sample t-statistic. Under H_0, the random variables in the combined sample $X_1, ..., X_{k_1}, Y_1, ..., Y_{k_2}$ are iid with pdf $f(y)$. Let $\boldsymbol{Z}_n$ be any permutation of $(X_1, ..., X_{k_1}, Y_1, ..., Y_{k_2})$ and compute $T_n(\boldsymbol{Z}_n)$ for each permutation. Then $H_{A,n}$ is the cdf based on all of the $T_n(\boldsymbol{Z}_n)$. H_0 is rejected if T_n is in the extreme tails of $H_{A,n}$. The number of ordered samples is $J_n = n!$ while the number of unordered samples is $\binom{n}{k_1}$. Such numbers get enormous quickly. Usually m randomly drawn permutations are selected with replacement, resulting in a cdf $H_{m,n}$ used to choose the appropriate cutoffs c_L and c_U.

For randomization tests that used a fixed number $m = B$ of permutations, calculations using binomial approximations suggest that $B = 999$ to 5000 will give a test similar to those based on using all permutations. See Dwass (1957), Edgington (1995, p. 50), Efron and Tibshirani (1993, pp. 208-210), Manly (1997, pp. 81-83) and Marriott (1979). Jöckel (1986) shows, under regularity conditions, that the power of a randomization test is increasing and converges as $m \to \infty$. It is suggested that the tests have good power if $m = 999$, but the pvalue of such a test is bounded below by 0.001 since the pvalue = (1 + the number of the m test statistics at least as extreme as the observed statistic)/$(m + 1)$. Buckland (1984) shows that the expected coverage of the nominal $100(1-\alpha)\%$ percentile method confidence interval is approximately correct, but the standard deviation of the coverage is proportional to $1/\sqrt{m}$. Hence the percentile method is a large sample confidence interval, in that the true coverage converges in probability to the nominal coverage, only if $m \to \infty$ as $n \to \infty$. These results are good reasons for using $m = \max(B, [n \log(n)])$ samples, and this choice is explored further in the following section.

2 THEORY FOR RESAMPLING ALGORITHMS

The key observation for theory is that $H_{m,n}$ is an empirical cdf. To see this claim, recall that $H_{A,n}(y) \equiv H_{A,n}(y|\boldsymbol{Y}_n)$ is a random cdf: it depends on the data $\boldsymbol{Y}_n$. Hence $H_{A,n}(y) \equiv H_{A,n}(y|\boldsymbol{y}_n)$ is the observed cdf based on the observed data. $H_{A,n}(y|\boldsymbol{y}_n)$ can be computed by finding $T^*_{i,n}(\boldsymbol{Y}^*_{i,n})$ for all J_n possible samples $\boldsymbol{Y}^*_{i,n}$. If m samples are selected with replacement from all possible samples, then the samples are iid and $T^*_{1,n}, ..., T^*_{m,n}$ are iid with cdf $H_{A,n}(y|\boldsymbol{y}_n)$. Hence $F_m \equiv H_{m,n}$ is an empirical cdf corresponding to $F \equiv H_{A,n}(y|\boldsymbol{y}_n)$.

Thus empirical cdf theory can be applied to $H_{m,n}$. Fix n and y. Then $mH_{m,n}(y) \sim$ binomial $(m, H_{A,n}(y|\boldsymbol{y}_n))$. Thus $E[H_{m,n}(y)] = H_{A,n}(y|\boldsymbol{y}_n)$ and $V[H_{m,n}(y)] = H_{A,n}(y|\boldsymbol{y}_n)[1 - H_{A,n}(y|\boldsymbol{y}_n)]/m$. Also

$$\sqrt{m}(H_{m,n}(y) - H_{A,n}(y|\boldsymbol{y}_n)) \xrightarrow{D} N(0, H_{A,n}(y|\boldsymbol{y}_n)[1 - H_{A,n}(y|\boldsymbol{y}_n)]).$$

Thus $H_{m,n}(y) - H_{A,n}(y|\boldsymbol{y}_n) = O_P(m^{-1/2})$. Note that the probabilities and expectations depend on m and on the observed data $\boldsymbol{y}_n$.

This result suggests that if $H_{A,n}$ is a first order accurate estimator of H_n, then $H_{m,n}$ can not be a first order accurate estimator of H_n unless m is proportional to n. If $m = \max(1000, [n\log(n)])$, then $H_{m,n}$ is asymptotically equivalent to $H_{A,n}$ up to terms of order $n^{-1/2}$. Using $m = \max(1000, [0.1n^2\log(n)])$ makes $H_{m,n}$ asymptotically equivalent to $H_{A,n}$ up to terms of order n^{-1}.

As an application, Efron and Tibshirani (1993, pp. 187, 275) state that percentile method for bootstrap confidence intervals is first order accurate and that the coefficient of variation of a bootstrap percentile is proportional to $\sqrt{\frac{1}{n} + \frac{1}{m}}$. If $m = 1000$, then the percentile bootstrap is not first order accurate. If $m = \max(1000, [n \log(n)])$, then the percentile bootstrap is first order accurate. Similarly, claims that a bootstrap method is second order accurate are false unless m is proportional to n^2. See a similar result in Robinson (1988).

Practical resampling algorithms often use $m = B = 1000$, 5000 or 10000. The choice of $m = 10000$ works well for small n and for simulation studies since the cutoffs based on $H_{m,n}$ will be close to those based on $H_{A,n}$ with high probability since $V[H_{10000,n}(y)] \leq 1/40000$. For the following theorem, also see DasGupta (2008, p. 6) and Serfling (1981, pp. 59-61).

Theorem 1: Let $\boldsymbol{Y}_1, ..., \boldsymbol{Y}_n$ be iid $k \times 1$ random vectors from a distribution with cdf $F(\boldsymbol{y}) = P(Y_1 \leq y_1, ..., Y_k \leq y_k)$. Let

$$D_n = \sup_{\boldsymbol{y} \in \Re^k} |F_n(\boldsymbol{y}) - F(\boldsymbol{y})|.$$

a) Massart(1990) $k = 1$: $P(D_n > d) \leq 2\exp(-2nd^2)$ if $nd^2 \geq 0.5 \log(2)$.

b) Kiefer (1961) $k \geq 2$: $P(D_n > d) \leq C \exp(-(2 - \epsilon)nd^2)$ where $\epsilon > 0$ is fixed and the positive constant C depends on ϵ and k but not on F.

To use Theorem 1a, fix n (and suppressing the dependence on $\boldsymbol{y}_n$), take $F = H_{A,n}$ computed from the observed data and take $F_m = H_{m,n}$. Then

$$D_m = \sup_{y \in \Re} |H_{m,n}(y) - H_{A,n}(y)|.$$

Recalling that the probability is with respect to the observed data, consider the following choices of m.

i) If $m = 10000$, then $P(D_m > 0.01) \leq 2e^{-2} \approx 0.271$.

ii) If $m = \max(10000, [0.25n \log(n)])$, then for $n > 5000$

$P\left(D_m > \frac{1}{\sqrt{n}}\right) \leq 2 \exp(-2[0.25n \log(n)]/n) \approx 2/\sqrt{n}.$

iii) If $m = \max(10000, [0.5n^2 \log(n)])$, then for $n > 70$

$P\left(D_m > \frac{1}{n}\right) \leq 2 \exp(-2[0.5n^2 \log(n)]/n^2) \approx 2/n.$

Example 3. Suppose F is the cdf of the $N(\mu, \sigma^2)$ distribution and $T_n(F) = \overline{Y}_n \sim N(\mu, \sigma^2/n)$. Suppose m independent samples $(Y^*_{j,1}, ..., Y^*_{j,n}) = \boldsymbol{Y}^*_{j,n}$ of size n are generated, where the $Y^*_{j,k}$ are iid $N(\mu, \sigma^2)$ and $j = 1, ..., m$. Then let the sample mean $T^*_{j,n} = \overline{Y}^*_{j,n} \sim N(\mu, \sigma^2/n)$ for $j = 1, ..., m$.

We want to examine, for a given m and n, how well do the sample quantiles $T^*_{(\lceil m\ \rho \rceil)} = \overline{Y}^*_{(\lceil m\ \rho \rceil),n}$ of the $\overline{Y}^*_{j,n}$ estimate the quantiles $\xi_{\rho,n}$ of the $N(\mu, \sigma^2/n)$ distribution and how well does $(T^*_{(\lceil m\ 0.025 \rceil)}, T^*_{(\lceil m\ 0.975 \rceil)})$ perform as a 95% CI for μ. Here $P(X \leq \xi_{\rho,n}) = \rho$ if $X \sim N(\mu, \sigma^2/n)$. Note that $\xi_{\rho,n} = \mu + z_\rho \sigma/\sqrt{n}$ where $P(Z \leq z_\rho) = \rho$ if $Z \sim N(0, 1)$.

Fix n and let f_n be the pdf of the $N(\mu, \sigma^2/n)$ distribution. By theory for quantiles such as Serfling (1980, p.80), as $m \to \infty$

$$\sqrt{m}(\overline{Y}^*_{(\lceil m\ \rho \rceil),n} - \xi_{\rho,n}) \xrightarrow{D} N(0, \tau^2_n)$$

where

$$\tau^2_n \equiv \tau^2_n(\rho) = \frac{\rho(1-\rho)}{[f_n(\xi_\rho)]^2} = \frac{\rho(1-\rho)2\pi\sigma^2}{n \exp(-z^2_\rho)}.$$

Since the quantile $\xi_{\rho,n} = \mu + z_\rho \sigma/\sqrt{n}$, need m fairly large for the estimated quantile to be good. To see this claim, suppose we want m so that

$$P(\xi_{0.975,n} - 0.04\sigma/\sqrt{n} < \overline{Y}^*_{(\lceil m\ 0.975 \rceil),n} < \xi_{0.975,n} - 0.04\sigma/\sqrt{n}) > 0.9.$$

(For $N(0, 1)$ data, this would be similar to wanting the estimated 0.975 quantile to be between 1.92 and 2.00 with high probability.) Then $0.9 \approx$

$$P(\frac{-0.04\sigma\sqrt{m}}{\tau_n\sqrt{n}} < Z < \frac{0.04\sigma\sqrt{m}}{\tau_n\sqrt{n}}) \approx P(-0.01497\sqrt{m} < Z < 0.01497\sqrt{m}\ )$$

or

$$m \approx \left(\frac{z_{0.05}}{-0.01497}\right)^2 \approx 12076.$$

With $m = 1000$, the above probability is only about 0.36. To have the probability go to one, need $m \to \infty$ as $n \to \infty$.

Note that if $m = B = 1000$, say, then the sample quantile is not a consistent estimator of the population quantile $\xi_{\rho,n}$. Also, $(\overline{Y}^*_{(\lceil m\ \rho\rceil),n} - \xi_{\rho,n}) = O_P(n^{-\delta})$ needs $m \propto n^{2\delta}$ where $\delta = 1/2$ or 1 are the most interesting cases. For good simulation results, typically need m larger than a few hundred, eg $B = 1000$, for small n. Hence $m = \max(B, n\log(n))$ combines theory with good simulation results.

The CI length behaves fairly well for large n. For example, the 95% CI length will be close to $3.92/\sqrt{n}$ since roughly 95% of the $\overline{Y}^*_{j,n}$ are between $\mu - 1.96\sigma/\sqrt{n}$ and $\mu + 1.96\sigma/\sqrt{n}$. The coverage is conservative (higher than 95%) for moderate m. To see this, note that the 95% CI contains μ if $T^*_{(\lceil m\ 0.025\rceil)} < \mu$ and $T^*_{(\lceil m\ 0.975\rceil)} > \mu$. Let $W \sim$ binomial $(m, 0.5)$. Then

$$P(T^*_{(\lceil m\ 0.975\rceil)}) > \mu) \approx P(W > 0.025m) \approx P(Z > \frac{0.025m - 0.5m}{0.5\sqrt{m}}) = P(Z > -0.95\sqrt{m})$$

$\to 1$ as $m \to \infty$. (Note that if $m = 1000$, then $T^*_{(\lceil m\ 0.975\rceil)} > \mu$ if 225 or more $\overline{Y}^*_{j,n} > \mu$ or if fewer than 975 $\overline{Y}^*_{j,n} < \mu$.)

Since F is not known, we can not sample from $T_n(F)$, but sampling from $T_n(F_n)$ can at least be roughly approximated using computer generated random numbers. The bootstrap replaces m samples from $T_n(F)$ by m samples from $T_n(F_n)$, that is, there is a single sample $Y_1, ..., Y_n$ of data. Take a sample of size n with replacement from $Y_1, ..., Y_n$ and compute the sample mean $\overline{Y}^*_{1,n}$. Repeat to obtain the bootstrap sample $\overline{Y}^*_{1,n}, ..., \overline{Y}^*_{m,n}$. Expect the bootstrap estimator of the quantile to perform less well than that based on samples from $T_n(F)$. So still need m large so that the estimated quantiles are near the population quantiles.

Simulated coverage for the bootstrap percentile 95% CI tends to be near 0.95 for moderate m, and we expect the length of the 95% CI to again be near $3.92/\sqrt{n}$. The bootstrap sample tends to be centered about the observed value of $\overline{Y}$. If there is a "bad sample" so that $\overline{Y}$ is in the left or right tail of the sampling distribution, say $\overline{Y} > \mu + 1.96\sigma/\sqrt{n}$ or $\overline{Y} < \mu - 1.96\sigma/\sqrt{n}$, then the coverage may be much less that 95%. But the probability of a "bad sample" is 0.05 for this example.

As a final remark, two tail tests with nominal level α and confidence intervals with nominal coverage $1 - \alpha$ tend to use the lower and upper $\alpha/2$ percentiles from $H_{m,n}$. This procedure corresponds to an interval covering $100(1-\alpha)\%$ of the mass. The interval is short if the distribution corresponding to $H_{m,n}$ is approximately symmetric. Shorter intervals can be found if the distribution is skewed by using the shorth(c) estimator where $c = \lceil m(1-\alpha) \rceil$ and $\lceil x \rceil$ is the smallest integer $\geq x$, e.g., $\lceil 7.7 \rceil = 8$. See Grübel (1988). That is, let $T^*_{(1)}, ..., T^*_{(m)}$ be the order statistics of the $T^*_{1,n}, ..., T^*_{m,n}$ computed by the resampling algorithm. Compute $T^*_{(c)} - T^*_{(1)}, T^*_{(c+1)} - T^*_{(2)}, ..., T^*_{(m)} - T^*_{(m-c+1)}$. Let $[T^*_{(s)}, T^*_{(s+c-1)}]$ correspond to the closed interval with the smallest distance. Then reject $H_0 : \theta = \theta_0$ if θ_0 is not in the interval.

Resampling methods can be used in courses on resampling methods, nonparametric statistics, and experimental design. In such courses it can be stated that it is well known that $H_{m,n}$ has good statistical properties (under regularity conditions) if $m \to \infty$ as $n \to \infty$, but algorithms tend to use $m = B$ between 999 and 10000. Such algorithms may perform well in simulations, but lead to tests with pvalue bounded away from 0, confidence intervals with coverage that fails to converge to the nominal coverage, and fail to take advantage of the theory derived for the impractical all subset algorithms. Since $H_{m,n}$ is the empirical cdf corresponding to the all subset algorithm cdf $H_{A,n}$, taking $m = \max(B, [n \log(n)])$ leads to a practical algorithm with good theoretical properties (under regularity conditions) that performs well in simulations.

3 References

Buckland, S. T. (1984), "Monte Carlo Confidence Intervals," *Biometrics*, 40, 811-817.

Chernick, M. R. (2008), *Bootstrap Methods: A Guide for Practitioners and Researchers*, 2nd ed., Hoboken, NJ: Wiley.

DasGupta, A. (2008), *Asymptotic Theory of Statistics and Probability*, New York, NY: Springer.

Davison, A. C., and Hinkley, D. V. (1997), *Bootstrap Methods and Their Application*, Cambridge, UK: Cambridge University Press.

Dwass, M. (1957), "Modified Randomization Tests for Nonparametric Hypotheses," *The Annals of Mathematical Statistics*, 28, 181-187.

Edgington, E. S. (1995), *Randomization Tests,* 3rd ed., New York, NY: Marcel Dekker.

Efron, B., and Tibshirani, R. J. (1993), *An Introduction to the Bootstrap*, New York, NY: Chapman & Hall/CRC.

Good, P. I. (2005), *Permutation, Parametric, and Bootstrap Tests of Hypotheses*, 3rd ed., New York, NY: Springer.

Grübel, R. (1988), "The Length of the Shorth," *The Annals of Statistics*, 16, 619-628.

Hall, P. (1988), "Theoretical Comparisons of Bootstrap Confidence Intervals," (with discussion), *The Annals of Statistics*, 16, 927-985.

Hoeffding, W. (1952), "The Large Sample Power of Tests Based on Permutations of Observations," *The Annals of Mathematical Statistics,* 23, 169-192.

Jöckel, K. H. (1986), "Finite Sample Properties and Asymptotic Efficiency of Monte Carlo Tests," *The Annals of Statistics*, 14, 336-347.

Kiefer, J. (1961), "On Large Deviations of the Empiric D. F. of a Vector of Chance Variables and a Law of Iterated Logarithm," *Pacific Journal of Mathematics*, 11, 649-660.

Lehmann, E. L. (1999), *Elements of Large–Sample Theory*, New York, NY: Springer.

Manly, B. F. J. (1997), *Randomization, Bootstrap and Monte Carlo Methods in Biology,* 2nd ed., London, UK: Chapman & Hall/CRC.

Marriott, F. H. C. (1979), "Barnard's Monte Carlo Tests: How Many Simulations?" *Applied Statistics*, 28, 75-77.

Massart, P. (1990), "The Tight Constant in the Dvoretzky-Kiefer-Wolfowitcz Inequality," *The Annals of Probability*, 3, 1269-1283.

Polansky, A. M. (2008), *Observed Confidence Levels*, Boca Raton, FL: Chapman & Hall/CRC.

Robinson, J. (1973), "The Large Sample Power of Permutation Tests for Randomization Models," *The Annals of Statistics,* 1, 291-296.

Robinson, J. (1988), "Discussion of 'Theoretical Comparison of Bootstrap Confidence Intervals' by P. Hall," *The Annals of Statistics*, 16, 962-965.

Romano, J. P. (1989), "Bootstrap and Randomization Tests of Some Nonparametric Hypotheses," *The Annals of Statistics*, 17, 141-159.

Sen, P. K., and Singer, J. M. (1993), *Large Sample Methods in Statistics: An Introduction with Applications,* New York, NY: Chapman & Hall.

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

Shao, J. (1989), "The Efficiency and Consistency of Approximations to the Jackknife Variance Estimators," *Journal of the American Statistical Association*, 84, 114-119.