Resampling Methods: Randomization Tests, Jackknife and Bootstrap Estimators

Lecture Notes for EEB 596z, ©B. Walsh 2000

Resampling methods are becoming increasingly popular as statistical tools, as they are (generally) very robust, their simplicity is compelling, and their computational demands are (largely) no longer an issue to their widespread implementation. These methods involve either sampling or scrambling the original data numerous times, and we consider three general approaches here. **Randomization tests** involve taking the original data and either scrambling the order or the association of the original data. **Jackknife estimates** involve computing the statistic of interest for all combinations of the data where one (or more) of the original data points are removed. **Bootstrap** approaches attempt to estimate the sampling distribution of a population by generating new samples by drawing (with replacement) from the original data. Our treatment here largely follows the excellent text by Manly (1997). More advanced treatments can be found by Miller (1974), Efron and Gong (1983), Hinkley (1983), Hinkley (1988), Efron and Tibshirani (1986), Good (1994), and Edgington (1995).

RANDOMIZATION TESTS

Suppose we have a complicated data set and believe it shows a nonrandom pattern. Randomization tests, introduced by R. A. Fisher (1935), provide a very general and robust approach for obtaining the probability of the observed pattern under the null hypothesis of randomness.

To be more specific, suppose some feature of the *order* of the data is of interest. For example, suppose we have a DNA sequence, and we think the number of repeated sequences (e.g., AGTAGTAGT) in the sample is greater than expected by chance. Obviously, the frequencies of the four bases in the sample influences this probability, so we have to take this into account when computing the probability under the null hypothesis (random order of the bases). We can construct an empirical distribution under the null hypothesis by taking the original sample and scrambling the bases at random (shuffling them like a deck of cards). This creates a sample with the same base frequencies as the original sample, but with the order of bases assigned at random. Repeating this process (say) a thousand times generates a sample of sequences under the null hypothesis. Suppose that the original data had 17 repeats, but only 12 of the 1000 randomized samples had

this many (or more) repeats. This gives the probability of seeing the observed number of repeats under the null hypothesis as $12/1000 = 0.012$.

The DNA example involves randomization tests that **scramble the order** of the data. Suppose the vector of the original data ($n = 200$) is

$$(x_1, x_2, \cdots, x_{200})$$

and we compute some statistic Λ that is a function of the order. An empirical distribution using this data under the null hypothesis is computed by scrambling (or **shuffling**) the n elements at random to give a sequence with a random order, e.g.

$$(x_{11}, x_{103}, \cdots, x_{37})$$

This is done a large number of time (hundreds to thousands), computing the test statistic Λ for each new sample, and hence generating a distribution of Λ under the null hypothesis (the **resampling distribution**). Suppose the value of Λ for the original data is only exceeded by k of the N values of Λ from the resampling distribution. In this case, the probability of observing the original value of Λ under the null hypothesis is $p = k/N$. An accurate estimate for the 5% critical value requires N on the order of a few hundred, while N needs to be on the order of a few thousand for an accurate estimate of the 1% critical level.

The other class of problems where randomization tests are appropriate involving **scrambling of the association** between data from individuals. Suppose each individual has a vector of data with two components, so that the data for the ith individual is $\mathbf{x}_i = (\mathbf{x}_{i,1}, \mathbf{x}_{i,2})$. If we wish to test whether the values of $\mathbf{x}_{i,1}$ influence the values of $\mathbf{x}_{i,2}$, we first compute some association statistic on the original data. The significance of this statistic is obtained by generating the distribution of this statistic under the null hypothesis of no association. This is obtained by computing the statistic on samples formed by shuffling the $\mathbf{x}_{i,2}$ over the $\mathbf{x}_{i,1}$. Thus, the original data and a randomize sample would be

$$\text{original data} = \begin{pmatrix} \mathbf{x}_{1,1}, \mathbf{x}_{1,2} \\ \mathbf{x}_{2,1}, \mathbf{x}_{2,2} \\ \vdots \\ \mathbf{x}_{100,1}, \mathbf{x}_{100,2} \end{pmatrix}, \quad \text{randomized sample} = \begin{pmatrix} \mathbf{x}_{1,1}, \mathbf{x}_{25,2} \\ \mathbf{x}_{2,1}, \mathbf{x}_{94,2} \\ \vdots \\ \mathbf{x}_{100,1}, \mathbf{x}_{6,2} \end{pmatrix}$$

Example 1. Suppose the height and weight are measured on a sample of males and females, and we wish to test whether a height-weight score, say

$$W = \text{height}/(\text{weight})^{1/3}$$

is different in males vs. females. Further suppose that from a plot of W, we clearly see that the data are highly non-normal and also find that no data transformations appear to cleanly normalized the data. Thus, $t-$ or normal-based tests are inappropriate to assess significance. However, we can easily used a randomization test, by computing W for each individual and then shuffling the W values random over gender. Suppose that there are n_m males and n_f females. Drawing (without replacement) n_m W values from the original sample and assigning them as males, and the remainder as females, generates a randomized sample. Note that this creates a random association of sex with (height, weight), while still retaining the covariance between height and weight inherent in the original sample.

Suppose that in the original sample the mean value of W in males was 10.2 units larger than the mean value for females. In a resampling distribution with 2000 values, 12 show male minus female differences of 10.2 or greater and 17 show male minus females differences of -10.2 or less. Thus under the one-sided tests that males are larger than females, randomization estimates the probability under the null hypothesis as 12/2000 = 0.006. Under a two-sided test, the probability is 0.006+ 17/2000 = 0.0145.

A critical question is how many randomized samples one should generate. The general consensus is around 1000 samples for tests at the 5% level and 5000 for tests at the 1% level (reviewed by Manly 1997)

Generating the Null Distribution via Monte Carlo Simulation

In some cases, the data cannot be cleanly randomized, but it many be possible to generate a sample from the null distribution by simulation. For example, suppose we have two species distributed in space, and we are interested in some measure of spatial association between the two. We can have a computer randomly place individuals spatially, generating a random sample under the null distribution. Computing our spatial statistic on this sample generates a value under the null distribution. Generating a large number of such samples gives us an estimate of the distribution of our statistic under the null hypothesis.

Constructing Approximate Confidence Intervals

While randomization tests are generally used only for tests of significance, a little thought shows that we can also use them to construct approximate confidence intervals. The basic approach can be illustrated as follows. Let $D = \overline{x}_1 - \overline{x}_2$ be the observed difference between two groups. The lower end of (say) the 95% confidence interval for D is obtained by finding the value of $L_{0.025}$ such that when we subtract $L_{0.025}$ from each member of group 1 than the randomization test using this adjusted data gives a new D value corresponding to the upper 2.5% point of the randomized distribution. Likewise, the upper end of the 95%

confidence interval is giving by finding the value $U_{0.025}$ that we add to the values of group 1 to give a new D corresponding to the lower 2.5% point of the new randomized distribution. The result 95% confidence interval for D becomes $(D - L_{0.025}, D + U_{0.025})$. More generally a $100(1-\alpha)\%$ confidence interval is given by

$$(D - L_{\alpha/2}, D + U_{\alpha/2})$$

Example 2. Consider the height-weight vs. gender study from Example 1. In the original sample, the mean of W in males was 10.2 units large than the mean value of W for females. Denote this difference by $D = 10.2$ What is the 95% randomization confidence interval for D? By trait and error, we use a number of different L values. For each value, we substract it from the W values of all males, compute a new D and also recompute the resampling distribution using the new values $(W_i - L)$ for each male. For each trail L value, obtained the following values for the upper percentage points of the resampling distribution:

L	6	5.5	5	4.5	4
Upper % point	20.4	10	4.2	1.9	0.6

Linear interpolation gives $L = 4.6$ as the upper 2.5% point on the randomized distribution. Likewise, adding U to all the male W values gives

U	6	5.5	5	4.5
Lower % point	8.4	4.2	1.5	0.9

Interpolation gives $U = 5.2$ as the lower 2.5 % point on the randomized distribution. Thus, an approximate 95% confidence interval on the difference in W scores for males and females is $(10.2 - 4.6, 10.2 + 5.2) = (5.6, 15.4)$.

NONPARAMETIC TESTS AND RANDOMIZATION APPROACHES

A variety of **nonparametic** tests appear in the statistical literature, so-called because they make few (if any) assumptions about the underlying distributions. Many of these are based on simple randomization tests, and we review a few of these here.

Fisher's One-sample Randomization Test

Suppose we have a series of observations on a single variable $(x_1, \cdots, x_n)$, and we are interested in testing whether the mean equals a particular value, μ_0. Under the assumption that the distribution of x_i is symmetric, we can apply a randomization

approach to test this hypothesis. Rescaling the data as $z_i = x_i - \mu_0$, we expect the z_i to be distributed symmetrically about zero, and hence we expect an equal number of positive and negative values.

This observation for the basis for **Fisher's one-sample randomization test**, where the null hypothesis is reject is the number of positive values is significantly different from the number of negative values. Suppose we observe k positive and $n - k$ negative values in the sample. Under the null hypothesis, the expected number of positives (or negatives) follows a binomial distribution with success parameter $p = 1/2$. The mean and variance (under the null hypothesis) of the number k of positive values are $n/2$ and $n(1/2)(1 - 1/2) = n/4$. For small sample sizes, tables of binomial probabilities can be used to compute an exact p value for the sample. For n moderate to large, we can use a normal approximation for the binomial,

$$\frac{k - n/2}{\sqrt{n/4}} \sim N(0, 1) \tag{1}$$

Runs Test

A related test where we have dichotomized data (such as male/female, positive/negative) is the **runs test**, which looks at the order of the data and sees if the number of runs is different from that expected by chance alone. For example, suppose we are looking at birth order in a large family, with the children being B B B B G G B. Here, there are three runs (BBBB, GG, B). Likewise, we may be looking at the residuals from a regression, scoring these as simply positive or negative. If there is nonlinearity in the regression, we can see over clustering of the residuals (Figure 1).

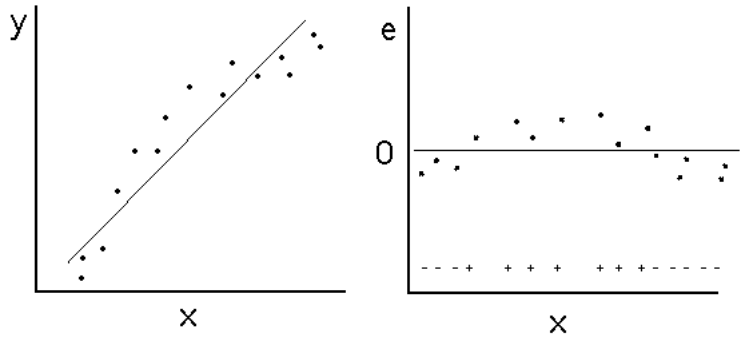

Figure 1. Nonlinearity can result in a reduction in the number of runs for residuals when a linear regression is fit to the data. On the left, a nonlinear function is fitted using the best linear regression. The graph on the right shows the resulting residuals (e), which shows just three runs in this case.

Under the null hypothesis that the two binary values are ordered at random, the expected number of runs is completely characterized by the numbers of the two classes n_1 and n_2. The expected number of runs is

$$\mu_r = \begin{cases} \dfrac{2n_1n_2}{n_1+n_2} + 1 & n_1 \neq n_2 \\ \\ n+1 & n_1 = n_2 = n \end{cases} \tag{2a}$$

with variance

$$\sigma_r^2 = \begin{cases} \dfrac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1+n_2)^2(n_1+n_2-1)} & n_1 \neq n_2 \\ \\ n+1 & n_1 = n_2 = n \end{cases} \tag{2b}$$

If both n_1 and n_2 are small, exact tables of the p value for the observed number r of runs can be found is most standard statistical tables. If one of the sample sizes exceeds 20, then we can a normal approximation,

$$\frac{r - \mu_r}{\sigma_r} \sim N(0,1) \tag{2c}$$

Rank-based Tests: Wilcoxon and Mann-Whitney U Tests

The final nonparametric test we discuss are those that use the **ranks** of the data. Rank-based approaches proceed by transforming the raw data into ranks, 1 for the smallest value up to n (the sample size) for the largest. We will assume no **ties** in the data (these can be handled, but we ignore this complication here). Consider the following data measured on males and females.

$$\text{Raw data} = \begin{pmatrix} 10 & M \\ 20 & M \\ 15 & M \\ 12 & M \\ 9 & F \\ 11 & F \\ 8 & F \end{pmatrix}, \quad \text{Ranks} = \begin{pmatrix} 3 & M \\ 7 & M \\ 6 & M \\ 5 & M \\ 2 & F \\ 4 & F \\ 1 & F \end{pmatrix}$$

Thus, the males have ranks of 3, 7, 6, 5 while the females have ranks of 1, 2, 4. We could use a standard randomization tests on the raw values (shuffling the M/F labels over the data values). Alternatively, we may simply ignore the actual values and use the ranks instead. This procedure discounts unusually large or small values. While we could generate a resampling distribution using the ranks, there is no need for this, as the distribution of ranks under the hull hypothesis has been tabulated.

Under the **Wilcoxon two-sample test**, n_1 and n_2 denote the two sample sizes, and R is the sum of ranks for the smaller sample size (n_2). In our example above, $R = 1 + 2 + 4 = 7$. The Wilcoxon test statistic C is computed as

$$C = n_1 n_2 + \frac{n_2(n_2+1)}{2} - \sum_{i=1}^{n_2} R_i \tag{3a}$$

where R_i is the rank of the i individual from the class with the smaller sample size. For small values of n_2, one can look up the exact p values under the null hypothesis (ranks are randomized over samples) in standard statistical tables. The mean and variance for this test statistic are

$$\mu_C = \frac{n_1 n_2}{2}, \qquad \sigma_C^2 = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12} \tag{3b}$$

For the above data, $n_1 = 4$, $n_2 = 3$, and $\sum R_i = 7$, giving $C = 12 + 6 - 7 = 11$, while (for these sample sizes), $\mu_C = 6$ and $\sigma_C^2 = 6 \cdot 8/12 = 4$. If $n_1 > 20$, C statistical is approximately normally distributed, with

$$\frac{C - \mu_C}{\sigma_C} \sim N(0,1) \tag{3c}$$

A very closely related two-sample rank-based test is the **Mann-Whitney U Test**. When there are more than two samples, the **Kruskal-Wallis test** refers to the extension of this approach to k samples (the nonparametric equivalent of the one-way ANOVA with k factors).

THE JACKKNIFE

Before the days of duct tape and the Swiss army knife, the lowly jackknife ruled as an all-in-one, fix everything tool. In this spirit, Tukey (1958) suggested a simple approach, **jackknife estimates**, based on removing data and then recalculating the estimator provides a general purpose statistical tool that is both easy to implement and solves a number of problems.

The motivation behind the jackknife is as follows. If $\overline{x}$ denotes the mean for a sample of size n, we can also compute the sample mean when the jth data point is removed (or **jackknifed**),

$$\overline{x}_{-j} = \frac{1}{n-1} \sum_{i \neq j}^{n} x_i \tag{4a}$$

Observe that if we know both $\overline{x}$ and $\overline{x}_{-j}$ we can compute the value of the jth data point as

$$x_j = n\,\overline{x} - (n-1)\,\overline{x}_{-j} \tag{4b}$$

Suppose we wish to estimate some parameter θ as a (potentially very complex) statistic of the n data points,

$$\widehat{\theta} = \phi(x_1, x_2, \cdots, x_{i-1}, x_i, x_{i+1}, \cdots, x_n) \tag{5a}$$

Motivated by Equation 4a, let the jth **partial estimate** of θ be given by the estimate computed with data point x_j removed,

$$\widehat{\theta}_j = \phi(x_1, x_2, \cdots, x_{i-1}, x_{i+1}, \cdots, x_n) \tag{5b}$$

By analogy with Equation 4b, define the jth **pseudovalue** by

$$\widehat{\theta}_j^* = n\,\widehat{\theta} - (n-1)\,\widehat{\theta}_j \tag{5c}$$

These pseudovalues assume that same role as the x_j in estimating the mean, hence the **jackknife estimate** of θ is given by the average of the pseudovalues,

$$\widehat{\theta^*} = \frac{1}{n}\sum_{i=1}^{n} \widehat{\theta}_i^* \tag{6a}$$

An approximate sampling error for $\widehat{\theta^*}$ can be obtained from the sample variance of the pseudovalues,

$$\text{Var}\left(\widehat{\theta^*}\right) = \frac{\text{Var}\left(\widehat{\theta}_j^*\right)}{n} = \frac{\sum_{j=1}^{n}\left(\widehat{\theta}_j^* - \widehat{\theta^*}\right)^2}{n(n-1)} \tag{6b}$$

Likewise, an approximate $(1-\alpha)\%$ confidence interval is given by

$$\widehat{\theta^*} \pm t_{\alpha/2,n-1}\sqrt{\frac{\sum_{j=1}^{n}\left(\widehat{\theta}_j^* - \widehat{\theta^*}\right)^2}{n(n-1)}} \tag{6c}$$

Where $t_{\alpha,n}$ satisfies $\Pr(t_n \geq t_{\alpha/2,n-1}) = \alpha$, with t_n denoting a t-distributed random variable with n degrees of freedom. Jackknife estimates of confidence intervals must be used with great caution, as they can either over- or under-estimate the true confidence interval.

Jackknife Estimates Reduce Bias

The major motivation for many jackknife estimates is that they reduce bias. In particular, Quenouille (1956) showed that using a jackknife estimate removes

bias of order $1/n$. To see this, suppose our original estimator $\widehat{\theta} = \phi(x_1, \cdots, x_n)$ of θ is biased, with

$$E(\,\widehat{\theta}\,) = \theta\left(1 + \frac{A}{n}\right) \tag{7a}$$

Thus

$$E(\,\widehat{\theta}_j\,) = \theta\left(1 + \frac{A}{n-1}\right) \tag{7b}$$

as this estimate is based on only $n-1$ data points. Now observe that the expected value of each pseudovalue becomes

$$\begin{aligned} E(\,\widehat{\theta}_j^*\,) &= n\,E[\,\widehat{\theta}\,] - (n-1)\,E[\,\widehat{\theta}_j\,] \\ &= \theta\left(n\left[1 + \frac{A}{n}\right] - (n-1)\left[1 + \frac{A}{n-1}\right]\right) = \theta \end{aligned} \tag{7c}$$

More generally, bias of order $1/n^p$ is removed by using the pth order jackknife (each partial estimate excludes p data points)

Example 3. Suppose we are trying to estimate the species diversity (total number of species) based on n samples. While the total number S of species in our sample is not an unreasonable estimate, it is clearly a lower bound and hence downward-biased. Burnhan and Overton (1978, 1979) suggested using a jackknifed estimator of S to reduce this bias. If we have species over n time points, an improved estimator is given by

$$\widehat{S^*} = S + \left(\frac{n-1}{n}\right) f_1 \simeq S + f_1 \quad \text{for } n >> 1$$

where f_1 are the number of species recorded from only one sample. Suppose a total of 120 species are observed during 10 collecting seasons, with 25 of these species being seen in only a single season. The jackknife estimate for total species number becomes

$$\widehat{S^*} = 120 + \left(\frac{9}{10}\right) 25 = 147.5$$

BOOTSTRAP RESAMPLING

For many problems in statistics, we are interested in the distribution of values from a random sample of the population. If the underlying distribution from

which the values are drawn is known, we can use developed theory to generate the sampling distribution. What can be done in the absence of any significant information about the underlying distribution? Efron (1979) suggested the use of **bootstrapping**, motivated by notion pulling oneself out of the mud by their bootstraps. The idea behind the bootstrap is very simple, namely that (in the absence of any other information), the sample itself offers the best guide of the sampling distribution. By resampling *with replacement* from the original sample, we can create a **bootstrap sample**, and use the empirical distribution of our estimator in a large number of such bootstrapped samples to construct confidence intervals and tests for significance.

Note the important difference between a randomized and a bootstrap sample. A randomized sample is generated by scrambling the existing data (sampling *without* replacement), while a bootstrap sample is generated by sampling *with* replacement from the original sample. Thus, in any particular bootstrap sample, we expect some data points for the original sample to be present two or more times, while others are absent. Randomization tests are appropriate when the order or association between parts of the data is assumed to be important (i.e., we are testing the null hypothesis that the order/association is random). On the other hand, where the order makes no difference in the statistic (such as the mean), every randomized sample returns the same value of the statistics. By contrast, each bootstrap sample is highly likely to return a (slightly) different value from the original sample.

Bootstrap and Balanced Bootstrap Sampling

Bootstrap samples are typically generated by sampling (with replacement) from the original data — for a set of n points, a particular point has probability $1/n$ of being chosen on each draw. Hence, from the binomial the probability a particular point is chosen exactly k times is

$$\Pr(k) = \frac{n!}{k!(n-k)!} \left(\frac{1}{n}\right)^k \left(\frac{n-1}{n}\right)^{n-k} \quad \text{for} \quad 0 \leq k \leq n$$

While the investigator may generate as many bootstrap samples as desired, note that if the sample size is small, eventually the entire universe of samples will be explored, and adding additional samples has no effect. In practice, this is very unlikely to occur. This does, however, point out that a major limitation with bootstrapping is the original sample size n. Clearly, the larger n, the more robust the bootstrap.

Under **balanced bootstrap sampling**, samples are generated in such a a way that each original data point is present exactly P times in the entire collection of bootstrap samples. This is most easily done by first constructing a vector of length Pn which contains P elements equal to one, P elements equal to two, and so on up to P elements of n. We then randomly scramble the elements. The first n elements of the scrambled vector corresponds to the indices of the data points

included in the first bootstrap sample. For example, if the first four elements are (13, 16, 9, 13), the bootstrap contains data point 13 (twice), and points 9 and 16. Likewise, the second bootstrap sample is given by the data points whose indices correspond to the elements $n+1, \cdots 2n$, and so on up to the last sample (elements $(P-1)n+1, \cdots Pn$).

Bootstrap Estimates of Bias

Consider some estimator $\widehat{\theta}$ of the unknown parameter θ, where

$$\widehat{\theta} = \phi(x_1, \cdots, x_n)$$

Let $\widehat{\theta}_0$ denote the estimate using the original data, while $\widehat{\theta}_i$ is the estimate using the ith bootstrap sample. The mean $\widehat{\theta}_B$ of all the bootstrap estimators is thus

$$\widehat{\theta}_B = \frac{1}{P} \sum_{i=1}^{P} \widehat{\theta}_i \tag{8a}$$

$\widehat{\theta}_B$ provides an estimate of the bias b of the estimator $\widehat{\theta}$. Noting that $b = E[\widehat{\theta}] - \theta$, and that $\widehat{\theta}_B$ is the bootstrap estimate of $E[\widehat{\theta}]$, we have

$$\widehat{b} = \widehat{\theta}_B - \widehat{\theta}_0 \tag{8b}$$

Noting that $E[\widehat{\theta} - b] = \theta$, a **bootstrap bias correction** for the original estimate is given by

$$\widehat{\theta}_0 - \widehat{b} = 2\widehat{\theta}_0 - \widehat{\theta}_B \tag{8c}$$

Example 3. Suppose our goal is to estimate the average skewness of a sample, using $S = (n-1)^{-1} \sum (x_i - \overline{x})^3$, and the estimate from the original population is $\widehat{S}_0 = -12.15$. Over 2000 bootstrap samples, the average value of this statistic is found to be $\overline{\widehat{S}} = -11.56$. Hence, the estimated bias of our original estimate for skew is $\widehat{b} = \overline{\widehat{S}} - \widehat{S}_0 = -11.56 - (-12.15) = 0.59$. Since $E[\widehat{S}] = b + \mu_3$, our estimate overestimates the true skew (μ_3). The bootstrap bias corrected estimate of the skew becomes $\widehat{S}_0 - \widehat{b} = -12.15 - 0.59 = -12.74$.

Bootstrap Confidence Intervals

A number of approaches for the construction of confidence intervals using bootstrap samples have been suggested (reviewed by Manly 1997), and we discuss just a few of these here.

Standard bootstrap confidence limits are based on the assumption that the estimator $\widehat{\theta}$ is normally distributed with mean θ (i.e., $\widehat{\theta}$ is an unbiased estimator) and variance σ^2. Assuming the sample variance

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{p} \left(\widehat{\theta}_i - \widehat{\theta}_B \right)^2 \tag{9}$$

of the bootstrap samples about their mean values provides a good estimate of σ^2, then an approximate $100(1-\alpha)\%$ confidence interval is given by

$$\widehat{\theta} \pm z_{\alpha/2}\, S \quad = \quad \widehat{\theta} \pm z_{\alpha/2} \sqrt{\frac{1}{n-1} \sum_{i=1}^{p} \left(\widehat{\theta}_i - \widehat{\theta}_B \right)^2} \tag{10a}$$

where z_α satisfies $\Pr(z > z_\alpha) = \Pr(z < -z_\alpha) = \alpha$ for a unit normal random variable z. For example, $z_{0.025} = 1.96$ is used for a 95% confidence interval. An improved interval is given by using the bootstrap correction for bias (Equation 8c),

$$2\widehat{\theta} - \widehat{\theta}_B \pm z_{\alpha/2} \sqrt{\frac{1}{n-1} \sum_{i=1}^{p} \left(\widehat{\theta}_i - \widehat{\theta}_B \right)^2} \tag{10b}$$

Around 100 to 200 bootstrap samples are generally sufficient for estimating the bootstrap standard deviation (Manly 1997).

A more direct approach for constructing a $100(1-\alpha)\%$ confidence limits is to use the upper and lower $\alpha/2$ values of the bootstrap distribution. Such approaches using the full bootstrap distribution are often referred to as **percentile confidence limits**. If $\widehat{\theta}_{L,\alpha/2}$ denotes the estimate of θ from the bootstrap distribution such that only a fraction $\alpha/2$ of all bootstrap estimates are less than this value, and likewise $\widehat{\theta}_{H,\alpha/2}$ is the estimate exceeded by only $\alpha/2$ of all bootstrap estimates, then an approximate confidence interval is given by

$$(\widehat{\theta}_{L,\alpha/2},\ \ \widehat{\theta}_{H,\alpha/2}) \tag{11}$$

This is **Efron's percentile confidence limit**. As with randomization tests, around 1000 bootstrap samples are required for 95% limits and 5000 samples for 99% limits. An alternative is **Hall's percentile confidence limit** (1992),

$$(2\widehat{\theta} - \widehat{\theta}_{H,\alpha/2},\ \ 2\widehat{\theta} - \widehat{\theta}_{L,\alpha/2}) \tag{12}$$

Hall's rational is that we are interested in the error ϵ between the estimate and true value, $\epsilon = \widehat{\theta} - \theta$, and that the distribution of $\epsilon_b = \widehat{\theta}_b - \widehat{\theta}$ approximates this

distribution, where $widehat\theta_b$ denotes the estimate for a random sinlge bootstrap sample. Hence,

$$\Pr(\epsilon_L < \widehat{\theta} - \theta < \epsilon_h) \approx \Pr(\epsilon_L < \widehat{\theta}_B - \widehat{\theta} < \epsilon_h)$$

Thus a$100(1-\alpha)$ confidence interval becomes

$$\Pr(\epsilon_{\alpha/2} < \widehat{\theta}_B - \widehat{\theta} < \epsilon_{(1-\alpha/2)}) = 1 - \alpha$$

Rearranging shows that

$$\Pr(\epsilon_{\alpha/2} < \widehat{\theta} - \theta < \epsilon_{(1-\alpha/2)}) = \Pr(\widehat{\theta} - \epsilon_{(1-\alpha/2)} < \theta < \widehat{\theta} - \epsilon_{\alpha/2})$$

giving Hall's interval.

A number of increasingly complicated procedures for using bootstrap samples for constructing percentage confidence intervals have been developed, and these are reviewed by Manly (1997). At present, there is (unfortunately) no clear reason for favoring one of these limits over the other.

References

Burnhan, K. P. and W. S. Overton. 1978. Estimation of the size of a closed population when capture probabilities vary among animals. *Biometrika* 65: 625–633.

Burnhan, K. P. and W. S. Overton. 1979. Robust estimation of population size when capture probabilities vary among animals. *Ecology* 60: 927–936.

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

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

Efron, B., and G. Gong. 1983. A leisurely look at the bootstrap, the jackknife, and cross validation. *American Statistician* 37: 36–48.

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

Fisher, R. A. F. 1935. *The Design of Experiments*, Oliver and Boyd, Edinburgh.

Good, P. 1994. *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*, Spring-Verlag, New York.

Hall, P. 1992. *The Bootstrap and Edgeworth Expansions*, Spring-Verlag, New York.

Hinkley, D. V. 1983. Jackknife methods. *Encyclopedia of Statistical Sciences* 4: 280-287.

Hinkley, D. V. 1988. Bootstrap methods. *Journal of the Royal Statistical Society*, B 50: 321–337.

Manly, B. F. J. 1997. *Randomization, Bootstrap and Monte Carlo Methods in Biology*, Chapman and Hall.

Miller, R. G. 1974. The jackknife – a review. *Biometrika* 61: 1–15.

Quenouille, M. H. 1956. Notes on bias in estimation. *Biometrika* 43: 353–360.

Tukey, J. W. 1958. Bias and confidence in not quite large samples (Abstract). *Annals of Mathematical Statistics* 29: 614.