

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/233777359>

Resampling methods

Chapter · August 2009

CITATIONS

2

READS

1,049

2 authors:



[William Howard Beasley](#)

University of Oklahoma Health Sciences Center

35 PUBLICATIONS 646 CITATIONS

[SEE PROFILE](#)



[Joe Rodgers](#)

University of Oklahoma

194 PUBLICATIONS 6,507 CITATIONS

[SEE PROFILE](#)

Some of the authors of this publication are also working on these related projects:



Behavior Genetics [View project](#)



Anaerobic hydrocarbon and fatty acid metabolism by syntrophic bacteria and their impact on carbon steel corrosion [View project](#)

Re-Sampling Methods

William H. Beasley and Joseph L. Rodgers

INTRODUCTION

Re-sampling is a statistical approach that relies on empirical analysis, based on the observed data, instead of asymptotic and parametric theory. The goal of re-sampling is to make an inferential decision, which is the same goal as that of a parametric statistical test such as the conventional t or analysis of variance (ANOVA). The difference is in how the goal is achieved.

In this chapter, we will define and describe three re-sampling procedures: the permutation test, the jackknife and the bootstrap. We place a strong emphasis on the bootstrap because it is the most flexible and most frequently used. We will describe both the concepts and the mechanisms that underlie re-sampling theory. In the course of this development, we hope that readers new to this area will begin to see ways of incorporating re-sampling methods into various aspects of their applied research, ways that allow them to address novel questions that traditional parametric approaches cannot easily address. We also hope that practicing methodologists as well will find new applications for re-sampling methods, and appropriate appreciation for their flexibility and overall value (as well as their limitations).

PROTOTYPES OF RE-SAMPLING METHODS

Both the classical parametric methods and the re-sampling methods infer characteristics of a larger abstract population distribution from a smaller observed distribution of scores in a sample. The mean, median (MD), standard deviation and 95th percentile are some of the useful distribution characteristics. Parametric and re-sampling methods use different approaches in pursuit of the same goal; the defining difference is the type of sampling distribution used in relation to the relevant test statistic. For completeness, we note that a statistic's *sampling distribution* is defined as the distribution of the statistic across all possible samples of the same size from a specified population of scores.

A parametric method employs a *theoretical sampling distribution* to model sampling error probability. These distributions, such as the t or χ^2 , are mathematically derived, and are based on a set of specified assumptions. In contrast, a re-sampling method employs an *empirical sampling distribution* to model sampling error probability. These distributions are created by the researcher from the particular unique set of observed data.

Consider an experiment in which the researcher compares the means of two observed independent samples. Regardless of any hypothesis, the sampling distribution (either theoretical or empirical) ideally should represent the values of t that would be observed if a very large number of samples of the same size were repeatedly drawn from the population. Of course the population is rarely known and has to be approximated, and the theoretical and empirical sampling distributions are different approaches to this problem.

Although the parametric approach models the sampling distribution that would be obtained from the population with assumptions and asymptotic theory, the empirical approach models this same conceptual sampling distribution of the population by repeatedly recombining the observed scores in various ways to form many *re-sampled samples*. The statistic of interest, such as the *MD* or t , is then calculated for each of these re-sampled samples; the empirical sampling distribution is the collection of these calculated statistics. The ‘statistic of interest’ is typically called the *plug-in statistic*, and this is defined more formally later in the chapter. What follows describes how the prototypical forms of the permutation test, jackknife and bootstrap separately construct their empirical sampling distribution and concludes with a discussion of the relationship between them.

The permutation test

R. A. Fisher (1935) first described the permutation test, sometimes called a randomization test, to test null hypotheses. Consider a two-tailed hypothesis for an independent sample experiment where the group 1 scores are 18, and 19, and the group 2 scores are 20, 21, and 22.

Example 1

The five procedural stages of the permutation test flow naturally from the definition of the p -value. The definition for an independent samples t -test is, ‘given the null hypothesis that no group differences exist in the (abstract)

population of scores, the p -value is the probability of obtaining a t -value equal or more extreme than the one actually observed, t_{obs} ’:

- Stage 1. Collect the sample and calculate t_{obs} in the same manner as if a parametric inferential test were being used.
- Stage 2. Prepare the *sampling frame*, which is the pool of scores from which random samples are drawn. In this example, all 6 observed scores are placed in the sampling frame without regard to group membership (which under the null hypothesis is irrelevant).
- Stage 3. Create one re-sampled sample with two groups of $n_1 = n_2 = 3$. This is achieved by drawing 3 scores *without replacement* from the sampling frame and placing them in one group and the remaining 3 scores in the other. Repeat the process to form B re-sampled samples. For the permutation test, the goal is to recreate every possible sample that could occur if the null hypothesis were true. In this case, $B = N!/(n_1!n_2!) = \binom{6}{3} = 20$, which is the number of different ways to assign the 6 scores to two groups². Every iteration should produce a new recombination of scores, as shown in Figure 16.1.
- Stage 4. Calculate the plug-in statistic (an independent samples t in this example) for each re-sampled sample drawn in Stage 3, just as if the 6 re-sampled scores belonged to an observed sample. The conventional notation for a test statistic calculated from a re-sampled sample includes a trailing asterisk, such as t^* . In this example, there will be $B = 20$ values of t^* : t_1^*, \dots, t_{20}^* .
- Stage 5. Compare the absolute value of t_{obs} to the B absolute values of t^* calculated in Stage 4. The two-tailed p -value is simply the proportion of t^* s that is equal or more extreme than t_{obs} . The equation³ can be written as

$$p = \frac{\#\{|t^*| \geq |t_{\text{obs}}|\}}{\binom{N}{n_2}} = \frac{\#\{|t^*| \geq |t_{\text{obs}}|\}}{B}.$$

As Figure 16.1 shows, two of the 20 absolute values of t^* are equal or more extreme than the observed statistic, therefore $p = 2/20 = 0.1$.

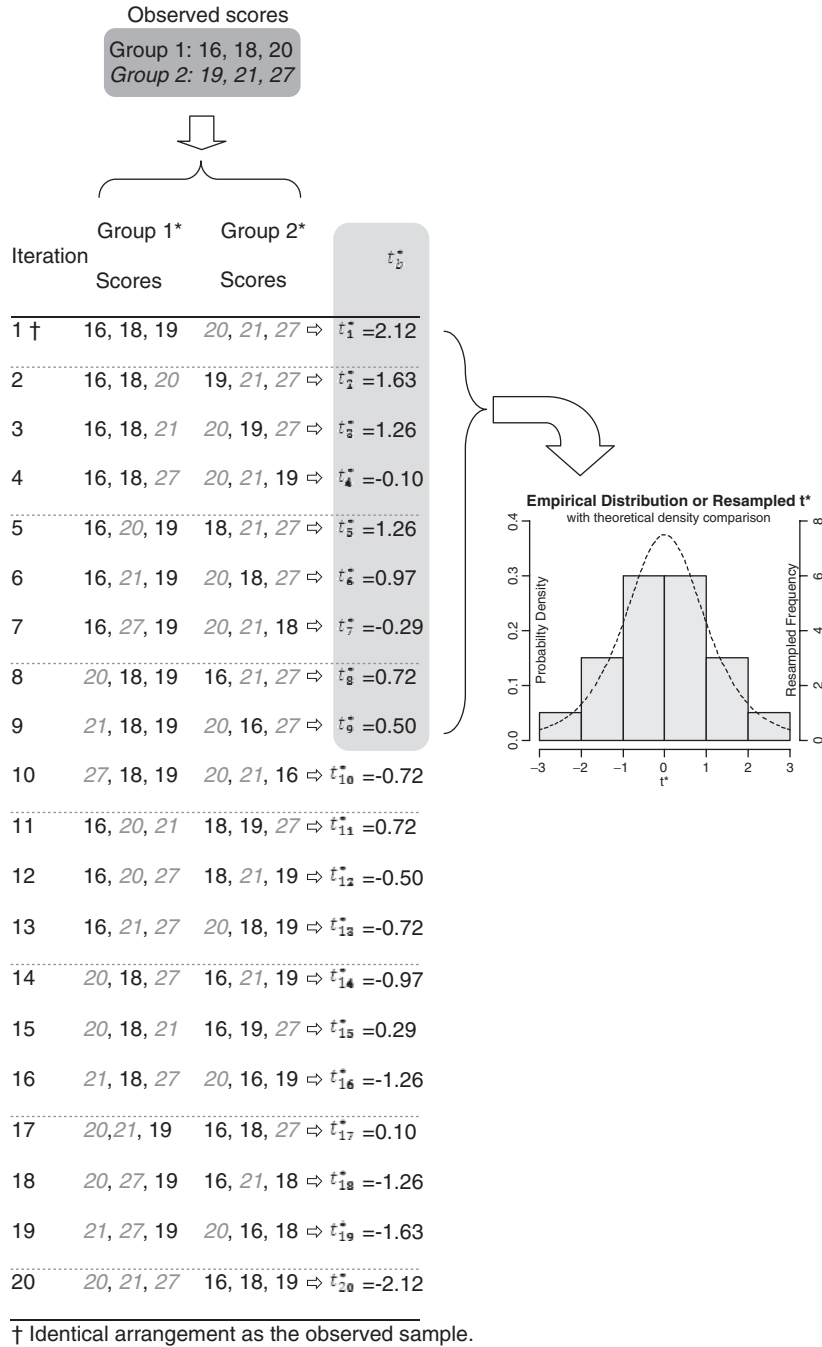


Figure 16.1 The observed sample and all possible re-sampled samples and the resulting plug-in statistic.

For comparison, the p -value of the observed t in relation to a theoretical t distribution with $df = 4$ is 0.101.

There is a tight conceptual connection between the sampling frame and the null hypothesis. In order to represent ‘no group differences in the population’, all six observed scores should have an equal probability of being placed in the re-sampled group 1 (and therefore of being placed in group 2). Fisher (1935: 30–54; see Fisher, 1966: 27–49 for a slight reformulation) formulated the statistical test to mimic random assignment to groups. His illustration revisited an analysis of Francis Galton and provided a more sound conclusion using a permutation test of dependent pairs.

Jackknife

The jackknife was developed by Quenouille (1949) and Tukey (1958, 1968/1986) to estimate bias and standard error (Miller, 1964). The distinctive feature of the jackknife is that a different observation is excluded in every jackknifed sample. While the permutation test draws samples of N scores, the jackknife draws smaller samples of $M = N - 1$ scores.

A jackknifed sample is denoted by $\mathbf{x}_{(\neq j)} = (x_1, x_2, \dots, x_{j-1}, x_{j+1}, \dots, x_N)$. If the four observed scores were 11, 12, 13, and 14, then the four jackknifed samples would be $\mathbf{x}_{(\neq 1)} = (12, 13, 14)$; $\mathbf{x}_{(\neq 2)} = (11, 13, 14)$; $\mathbf{x}_{(\neq 3)} = (11, 12, 14)$; $\mathbf{x}_{(\neq 4)} = (11, 12, 13)$.

Example 2

In a jackknife example, suppose you calculated some statistic T , which is an estimator of some parameter θ , associated with 60 observed scores, and the goal is to estimate its standard error in the population of samples of size 60:

- Stage 1. Collect the sample and calculate T in the same manner as if a parametric inferential test were being used.
- Stage 2. Prepare the sampling frame. With single group samples for the jackknife, this involves simply using the observed sample.

- Stage 3. Create one jackknife sample of $M = 59$ by excluding one score in the sampling frame. Repeat the process to form B jackknife samples. As before, the goal is to mimic possible samples that could be drawn from the population. In this case,

$B = \binom{60}{59} = 60$, which is every possible sample when an observation is excluded once. Of course B and N will be equal when $M = N - 1$.

- Stage 4. Calculate the plug-in statistic, T^* , for every re-sampled sample drawn in Stage 3.

- Stage 5. Calculate the standard deviation of the B values:

$$\hat{se}_{\text{jackknife}} = \sqrt{\frac{B-1}{B} \sum_{j=1}^B (T_{(\neq j)}^* - \bar{T}^*)^2}, \text{ where}$$

$$\bar{T}^* = \frac{1}{B} \sum_{j=1}^B T_{(\neq j)}^* \quad (1)$$


Notice that the variability of the population is estimated from the spread of the B number of T^* s, not the spread of the N number of X values. This feature will be important with the bootstrap as well.

The prototypical jackknife works well with smooth statistics like the mean or sample correlation, but it can fail with unsmooth statistics such as the MD (Efron and Tibshirani, 1993, section 11.6). Two solutions exist: either use a delete- d jackknife (which excludes more than one score from each jackknife sample) or (preferably) use a bootstrap, which we will discuss now.

Bootstrap

Efron developed the bootstrap in the mid-1970s (Holmes et al., 2003). After his first article (Efron 1979), people soon realized the bootstrap was more efficient statistically and was able to address more experimental questions than the jackknife or the permutation test. Its flexibility allows it to estimate standard errors, p -values, confidence intervals (CI) and many other statistics

without the necessity of extensive theoretical mathematics for each unique scenario.

The bootstrap is distinguished from the permutation test because it is based on sampling with replacement, increasing the number of possible re-sampled samples. In the two group example above (used to illustrate the permutation test), the bootstrap approach would potentially form two groups in $N^N = 46,656$ ways⁴, whereas the permutation test had only 720. With larger observed samples, complete enumeration of all possible re-samples produces too many bootstrap samples to calculate realistically. A pragmatic (and probabilistically satisfactory) compromise is to choose scores *randomly* from the sampling frame to form B bootstrap samples, where B is typically between 499 and 9,999. The p -value is calculated slightly differently because of the additional stochastic noise introduced. We  discuss this and the choice of B later. Here we provide two bootstrap examples for comparison with the two previous re-sampling methods.

Example 3

For our first bootstrap example, we return to the analysis of group differences introduced in the permutation test example, using exactly the same data:

- Stage 1. Collect the sample and calculate t_{obs} .
- Stage 2. Prepare the sampling frame. All six observed scores are placed in the sampling frame without regard for group membership.
- Stage 3. Create one bootstrap sample with two groups of $n_1 = n_2 = 3$. This is achieved by drawing 3 scores *with replacement* from the sampling frame and placing them in one group and an additional 3 scores in the other. Repeat this stage to form B bootstrap samples, say 9,999.
- Stage 4. Calculate the plug-in statistic, t_b^* , for each of the B bootstrap samples drawn in Stage 3.
- Stage 5. Compare the absolute value of t_{obs} to the B absolute values of t^* . The p -value is:

$$p = \frac{1 + \# \{ |t^*| > |t_{\text{obs}}| \}}{B + 1} \quad (2)$$

Table 16.1 Complete enumeration of the bootstrap


Iteration	Group 1 scores	Group 2 scores		t_b^*
1	16, 16, 16	16, 16, 16	$\Rightarrow t_1^* =$	–
2	16, 16, 16	16, 16, 18	$\Rightarrow t_2^* =$	1.0
3	16, 16, 16	16, 16, 19	$\Rightarrow t_2^* =$	1.0
⋮				
1,942*	16, 18, 19	20, 21, 27	$\Rightarrow t_{1,942}^* =$	2.12
⋮				
7,776	16, 27, 27	27, 27, 27	$\Rightarrow t_{7,776}^* =$	1.0
7,777	18, 16, 16	16, 16, 16	$\Rightarrow t_{7,777}^* =$	–1.0
7,778	18, 16, 16	16, 16, 18	$\Rightarrow t_{7,778}^* =$	0.0
⋮				
46,655	27, 27, 27	27, 27, 21	$\Rightarrow t_{46,655}^* =$	–1.0
46,656	27, 27, 27	27, 27, 27	$\Rightarrow t_{46,656}^* =$	–

* Identical arrangement as the observed sample.

If the N^N bootstrap samples had been completely enumerated, the equation for p would have remained $\# \{ |t^*| > |t_{\text{obs}}| \} / B$. If this had occurred, the only difference between the permutation test and bootstrap algorithms is sampling with vs. without replacement in Stage 3. A complete enumeration of the scores is illustrated in Table 16.1. The two-tailed p -value was .0849 on our first run; this number will vary very little when $B = 9,999$.

A CI can be estimated by reusing the empirical sampling distribution created in Stage 4. Like the equation for p , the operational definition for the CI's lower and upper bound comes naturally from its conceptual definition. To estimate a $(1 - \alpha)$ CI where scores are likely to fall if the null is true, we begin by ordering all B values and then identify the two scores at the $\alpha/2$ and $1 - \alpha/2$ quantiles. For instance, in the previous example where $B = 999$, the CI would be $[t_{(25)}^*, t_{(975)}^*]$. CIs are discussed in more depth in the section 'Recognized variations of the bootstrap', below.

Example 4

In a second bootstrap example, suppose a researcher collects N scores and want to calculate the $(M$  and estimate its standard error. Because no closed form parametric equation exists for the MD 's standard

error, the bootstrap provides a very useful solution.

- Stage 1. Collect the sample and calculate MD_{obs} from the N scores.
- Stage 2. Prepare the sampling frame. In this situation, simply use the observed sample.
- Stage 3. Randomly draw N scores with replacement from the sampling frame. This will be the b th bootstrap sample, where $b = 1, \dots, B$. Repeat this stage to form B bootstrap samples.
- Stage 4. Calculate the plug-in statistic, MD_b^* , for each bootstrap sample drawn in Stage 3.
- Stage 5. Calculate the standard deviation of the B values:

$$\widehat{se}_{\text{bootstrap}} = \sqrt{\frac{B}{B-1} \sum_{b=1}^B (MD_b^* - \overline{MD^*})^2}, \text{ where}$$



$$\overline{MD^*} = \frac{1}{B} \sum_{b=1}^B MD_b^* \quad (3)$$

Notice that this is simply the equation for standard deviation, if the B bootstrapped statistics are treated as observed scores in a sample. Compare this to standard error equation for the jackknife. They both estimate the population's variability from the spread in the distribution of re-sampled statistics, not the spread in the observed scores. However the bootstrap needs a stronger scaling factor in Stage 5 (which is $1/(B-1)$ instead of $(B-1)/B$) because bootstrap samples wander from the observed sample more than jackknife samples wander. For illustration, consider that any prototypical jackknife sample will differ from the observed sample by one score only.

In contrast, a bootstrap sample conceivably could have *only one score in common* with the observed sample. This would occur when one observed score is drawn N times (due to sampling with replacement). This is a rare instance, occurring only N times in N^N re-samples.

Comparison of the permutation test, jackknife and bootstrap

The differences between the three prototypical re-sampling methods depends on whether they sample with or without replacement and whether each re-sampled sample has a full sample size of N scores, or is a sub-sample of fewer than N scores (see Rodgers, 1999). These distinctions are shown in Figure 16.2.

The remainder of this chapter focuses on the bootstrap. We believe the permutation test⁵ is a good starting point for a pedagogical and historical perspective, but it is used in applied research less often because its capabilities have been surpassed by the bootstrap⁶. Additional insight into the re-sampling lineage comes from Efron and Tibshirani (1993: 218): 'The bootstrap distribution was originally called the  combination distribution.' It was designed to extend the virtues of permutation testing to the great majority of statistical problems where there is nothing to permute.  hough the jackknife is occasionally used in contemporary statistics, we will give it less attention because it is only a linear approximation to the bootstrap. When estimating the standard error of a

		Re-sampled Sample Size	
		Subsample	Full Sample
Sampling method	Without replacement	Jackknife	Permutation test
	With replacement	m out of n bootstrap	Prototypical bootstrap

Figure 16.2 Classification of re-sampling methods. The m out of n bootstrap is mentioned in Section 5.

complex statistic such a correlation, it always will be less efficient than the bootstrap (Efron and Tibshirani, 1993: 146; Efron and Gong, 1983).

We conclude this section with one additional historical detail: the permutation test published by R.A. Fisher, and the jackknife-like test published by Gossett (Student, 1908), calculated the mean for each re-sampled sample, instead of t as we have here. Using the mean as the plug-in statistic allowed them to take enormous computational shortcuts, an important consideration before computers. But we chose to use t because it is the accepted practice of contemporary statistics in two-group settings; advanced readers are referred to the explanation of pivotal statistics and hypothesis testing presented by N.I. Fisher and Hall (1990).

THE PLUG-IN PRINCIPLE AND THE SAMPLING FRAME IN THE BOOTSTRAP

Now that we have described the basic re-sampling ideas and methods, we describe and develop two additional bootstrap concepts. We begin this development by introducing notation to support concepts we previously introduced. To restate slightly, an inferential procedure makes a decision about a population distribution of single scores (F) from an observed sample. An empirical distribution of single scores (\hat{F}) is the inferential procedure's best guess about F .

Although they are closely related, the sampling frame should not be confused with \hat{F} . As defined in Example 1, above, the *sampling frame* is the pool of scores from which bootstrapped samples' scores are drawn. Therefore, the construction of the sampling frame directly influences \hat{F} . Here's another way to think about it: in order to have \hat{F} match F as closely as possible, work backwards and construct the sampling frame in such a way that drawing single scores from it mimics how single scores are drawn from the population. In many elementary scenarios like Example 4, above, the observed sample

itself is a good sampling frame. Strategies for designing different sampling frames in more complicated scenarios are discussed below.

Finally, the bootstrap distribution should not be confused with the sampling frame or with \hat{F} . A bootstrap distribution is a type of empirical *sampling* distribution – each of its values represents a statistic calculated from one bootstrap sample. In other words, a bootstrap distribution is a function of which N scores \hat{F} from are drawn to calculate each of the B bootstrapped statistics.

Plug-in principle

The plug-in principle simply states that a function of a population, defined as $g(F)$, can be estimated by calculating a comparable function, $g(\hat{F})$, on the sample. These functions can be a familiar parameter like the *MD* or mean, or something less conceptually obvious, like the t statistic. The plug-in principle gives flexibility to the bootstrap. Davison et al. (2003: 142), stated that, subject to mild conditions, the function g can be algorithm of almost arbitrary complexity, shattering the naive notion that a parameter is a Greek letter appearing in the probability distribution and showing the possibilities for uncertainty analysis for the complex procedures now in daily use, but at the frontiers of the imagination a quarter of a century ago.'

The next two examples use the bivariate correlation coefficient for the plug-in statistic, one of the traditional 'Greek letters' defined on the abstract population, the plug-in function is $\rho = g(F) = \text{Covariance}(X, Y) / (\sigma_X \sigma_Y)$, where σ_X and σ_Y are the variables' standard deviations. Defined on the observed sample (and bootstrap samples), the plug-in statistic is built from observed sample values instead of abstract population parameters:

$$\begin{aligned} r &= g(\hat{F}) = \frac{\text{Covariance}(X, Y)}{S_X S_Y} \\ &= \frac{\sum_i (x_i - \bar{X})(y_i - \bar{Y})}{\sqrt{\sum_i (x_i - \bar{X})^2 \sum_i (y_i - \bar{Y})^2}} \end{aligned} \quad (4)$$

Sampling frame representation: null hypothesis versus observed scores

To this point the sampling frame in the descriptions in the first section has been straightforward and consistent. Stage 2, the description of the sampling frame, was included as a place holder to accommodate the following important distinction. The sampling frames (and thus the resulting bootstrap distributions) of the two previous bootstrap examples have a subtle difference. In Example 3, above, the sampling frame produced scores expected if the specified null hypothesis were true. However, in Example 4, the sampling frame came from the observed scores and was not connected to any hypothesis. The next two examples should make the distinction more clear. In both cases we would like to infer something about a correlation; the first example uses \hat{F}_{obs} , whereas the second uses \hat{F}_{null} , which we will define after this example.

Example 5

Because bivariate data are involved, we need to expand our definition of an observation to a vector of two scores, where $u_i = (x_i, y_i)$ represents the observed x and y scores of the i th subject. This example uses a procedure described in Diaconis and Efron (1983) which analyzes continuous, bivariate data. Suppose you collect bivariate data points from 40 subjects with $r_{\text{obs}} = 0.34$ and wish to estimate a confidence interval for correlation values likely to be observed across replications if 40 subjects were repeatedly sampled from the same population:

- Stage 1. Collect the sample and calculate r_{obs} from the N data points (pairs of X, Y values).
- Stage 2. Prepare the sampling frame. To produce \hat{F}_{obs} in this case, simply use the observed sample.
- Stage 3. Randomly draw N pairs of scores with replacement. In this specific approach, it is important to keep the pairs intact. For instance, if x_4 is selected, the accompanying value must be y_4 (i.e., the x and y scores for the fourth subject). Repeat this stage to form B bootstrap samples.

- Stage 4. Calculate the plug-in statistic, r_b^* , for each bootstrap sample drawn in Stage 3.
- Stage 5. Calculate the CI $[r_{(25)}^*, r_{(975)}^*]$. If a hypothesis test is desired, the null hypothesis can be rejected if ρ_{null} falls outside of the CI.

The bootstrap in the example above is based on \hat{F}_{obs} , which is our best guess of the population distribution that the observed scores were drawn from. Furthermore, it employed a bivariate sampling approach: a subject's x and y scores are always drawn together. A related, *but not equivalent*, approach is to base the bootstrap on \hat{F}_{null} , the population distribution that represents the null hypothesis (see Davison and Hinkley, 1997: 138, 161). To re-analyze the same 40 subjects using \hat{F}_{null} , a little more effort is needed to prepare the sampling frame, which we explain now.

To represent the null hypothesis that X and Y are linearly unrelated in the population, every subject's x must have an equal probability of being selected with every subject's y . This implies a univariate sampling approach (Lee and Rodgers, 1998). Previous bootstrap examples had sampling frames containing N possible scores, whereas a univariate sampling frame (of a bivariate sample) has N^2 . Figure 16.3 depicts the difference between this sampling frame and the one used previously.

Example 6

For this example our hypothesis is $\rho_{\text{null}} = 0$, i.e., no correlation exists in the population (testing non-zero correlation values are described in Beasley et al., 2007).

- Stage 1. Collect the sample and calculate r_{obs} from the N data points (pairs of X, Y values).
- Stage 2. Prepare the univariate sampling frame by pairing every x with every y value as shown in Figure 16.3.
- Stage 3. Randomly draw N pairs of scores with replacement from the N^2 possible points in the sampling frame. Repeat this Stage to form B bootstrap samples.
- Stage 4. Calculate the plug-in statistic, r_b^* , for each bootstrap sample drawn in Stage 3.

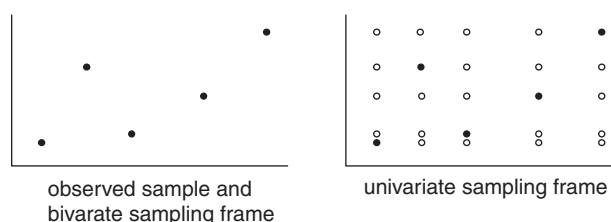


Figure 16.3 Scatter plots of a bivariate sampling frame based on \hat{F}_{obs} and a univariate sampling frame based on \hat{F}_{null} . Each point on the left has a 1/5 chance of being selected on each draw, whereas each point on the right has a 1/25 chance.

Stage 5. Calculate the CI $[r_{(25)}^*, r_{(975)}^*]$. If a hypothesis test is desired, the null hypothesis can be rejected if r_{obs} falls outside of the CI.

This CI (derived from \hat{F}_{null}) represents the variability around ρ_{null} , whereas the previous CI (derived from \hat{F}_{obs}) represents the variability around r_{obs} . The two contrasting equations for the one-tailed p -value for $H_0 : \rho > \rho_{\text{null}}$ are $p_{\hat{F}_{\text{obs}}} = \frac{(1 + \#\{r_b^* < \rho_{\text{null}}\})}{B+1}$ and $p_{\hat{F}_{\text{null}}} = \frac{1 + \#\{r_b^* > r_{\text{obs}}\}}{B+1}$. Notice that the value of ρ_{null} isn't present in the second p -value equation because it is reflected within the sampling frame, which is constrained by its construction to have a correlation of zero.

RECOGNIZED VARIATIONS OF THE BOOTSTRAP

The past three decades of bootstrap innovations have been both extensive and creative, producing many modifications, some of which we will survey in this third section. The majority of the innovations have modified either the sampling frame (Stage 3) or the CI definition (Stage 5 in Examples 5 and 6, above).

Sampling frame modifications

We already have discussed one sampling frame modification, the choice of \hat{F}_{null} or \hat{F}_{obs} , which can be closely tied to the underlying theoretical question. The following sampling frame modifications are more mechanical, and motivated to increase the reliability of

the inference by shaping and improving the bootstrap distribution so that it replicates the population more closely.

To this point, we have described a *non-parametric bootstrap*. Perhaps the earliest and most widely used sampling frame modification is the *parametric bootstrap*. The sampling frame scores do not come from the observed sample. Instead they are randomly generated to match the desired \hat{F} , which means that assumptions must be made about the population distribution. This contrasts with the prototypical non-parametric bootstrap, which assumes only that the observed scores are independently and identically distributed (iid).

Returning to the first correlation example, a parametric \hat{F}_{obs} can be created fairly easily if the scores are assumed to have a linear relationship in a bivariate normal population. The generated \hat{F}_{obs} needs to mimic the population correlation (e.g., using $r_{\text{obs}} = 0.34$ to estimate ρ), which can be achieved by expanding Stage 2 into two steps. In Stage 2a, generate an $N \times 2$ matrix of random scores $\text{iid} \sim N(0, 1)$. In Stage 2b, multiply this matrix by a decomposed-observed correlation matrix (e.g., Cholesky(\mathbf{R}) = Cholesky($\begin{bmatrix} 1 & .34 \\ .34 & 1 \end{bmatrix}$) = $\begin{bmatrix} 1 & .34 \\ 0 & .94 \end{bmatrix}$; see Kaiser and Dickman, 1962). The other stages are implemented as before, except that now Stage 2 will be repeated for each bootstrap sample. Notice that one parametric bootstrap ultimately can generate $2 \times N \times B$ unique scores.

The parametric bootstrap mixes the advantages and limitations of traditional parametric

theory and more recent bootstrap procedures. An ideal niche appears to be situations in which the population distributions are safely assumed (either through a large observed sample or prior knowledge), but the statistic does not have a known distribution or an accessible standard error formula. Because we can generate scores beyond the finite number observed, many options are available to create \hat{F} , where the (non-normal) marginal and (non-linear) conditional distributions can be specified (one of the many examples is Headrick and Kowalchuck, 2007).

A *Monte Carlo test* can be viewed as a parametric bootstrap that uses \hat{F}_{null} instead of \hat{F}_{obs} . Its history began more than three decades before the bootstrap to address issues in physics and chemistry (Metropolis and Ulam, 1949; also see the brief summary in Hall, 1992: 35). Scores are randomly generated from the null hypothesis, potentially without information estimated from any observed sample. To illustrate, we return to the second correlation example and make some unusual assumptions such as an exponential distributed X variable and a beta distributed Y variable that are slightly negatively correlated in the population, say $H_0: \rho = -0.2$. With a moderate amount of effort, scores with this distributional structure can be generated in Stage 2 with techniques referenced in the previous paragraph. As before, conclusions can be drawn from the bootstrap distribution in Stage 5. However a distinctive feature of the Monte Carlo test is that, depending on the questions being asked of it, an observed sample doesn't even need to be collected. Further, it obviously can account for substantial departures from normality, through its broad distributional flexibility.

A *semi-parametric bootstrap* re-samples the *residual errors* of a specified model, as opposed to re-sampling the raw observed values. This technique is applied in many bootstraps involving multiple predictor variables that are not independently distributed.

Example 7

Here we present a semi-parametric bootstrap that estimates the standard error of a linear

model of two continuous variables. The criterion score for the i th subject is $y_i = b_0 + b_1x_{1,i} + b_2x_{2,i} + e_i$, where the residual is defined as e_i .

- Stage 1. Collect the sample and calculate the sample coefficients (b_0, b_1, b_2) that estimate the population parameters ($\beta_0, \beta_1, \beta_2$).
- Stage 2. The sampling frame is formed from the N residuals (e_1, \dots, e_N).
- Stage 3a. Randomly draw N residuals with replacement ($e_1^*, e_2^*, \dots, e_N^*$).
- Stage 3b. If the covariates (X s) are considered fixed, then the each bootstrap sample is:

$$\begin{aligned} y_1^* &= b_0 + b_1x_{1,1} + b_2x_{2,1} + e_1^* \\ y_2^* &= b_0 + b_1x_{1,2} + b_2x_{2,2} + e_2^* \\ &\dots \\ y_N^* &= b_0 + b_1x_{1,N} + b_2x_{2,N} + e_N^* \end{aligned} \quad (5)$$

This creates a bootstrap sample of N values: ($y_1^*, y_2^*, \dots, y_N^*$). Repeat this stage to form B bootstrap samples.

- Stage 4. Calculate new values of b_0^*, b_1^* and b_2^* with the same two parameter linear model for each bootstrap sample created in Stage 3.
- Stage 5. Three different bootstrap distributions can be formed – one for each estimated coefficient (b_0, b_1, b_2). Estimate their standard errors by calculating the standard deviations of the B values as described in Stage 5 of Example 4.

Notice that in Stage 3b only the residuals (and the resulting y^* s) differ from the observed sample; the x values are not shuffled because these variables were considered fixed, not random in this specific example. Additional bootstrap distributions for plug-in statistics such as R^2 and F can be used for other research goals, like testing a hypothesis (e.g., Manly, 2007, chapter 7). The coefficients can be estimated a variety of ways, such as parameter values that minimize the sum of the squared residuals (i.e., least squares) or values that minimize the median of absolute values of deviations (MAD) in the

Author: pls
provide
definition

sample. Furthermore, the stages easily can accommodate many LM variations as well as more traditionally exploratory models, like loess curves and smoothing splines (Hastie et al., 2001).

At the cost of more restrictive assumptions, bootstrapping residuals can accommodate more types of models. Bootstrapping observed cases (described in Examples 5 and 6) does not assume the errors are homogeneously distributed. However the semi-parametric approach estimates and re-samples the residuals as if they were interchangeable, which requires the assumption of homogenous variance. To address this weakness, adjustments such as standardizing the residuals may improve the robustness of semi-parametric approaches. Davison and Hinkley (1997; sections 3.3, 6.2 and 6.3) directly address these issues. Situations with dependent data can be difficult to handle with re-sampling, and are discussed further in the section 'Resampling dependent data', below.

Confidence interval modifications

The prototypical bootstraps described above use the *percentile* CI method: the quantile of the bootstrap distribution maps directly to the quantile of the inferred population distribution. For example, the 3,482th ordered statistic in a bootstrap distribution of $B = 9,999$ would represent our best guess of the 34.82th quantile of the population, given the null hypothesis. Of course ordered scores such as the 250th, 9,500th, 9,750th, and 9,900th typically are more relevant to CIs in applied settings. If the bootstrap distribution is perfectly normal, the result from the percentile CI is identical to the standard normal Z .

Unfortunately, the simple mapping of the percentile method does not produce unbiased population inferences in most conditions. Several alternative CI methods often create inferences with less bias and greater efficiency. Here we describe only basic aspects of five different CI methods to help applied researchers understand the

broader distinctions. Readers interested in further explanation of the available adjustments to CIs should refer to Efron (1982, 1987 and comments). Less technical coverage can be found in the comprehensive reviews by DiCiccio and Efron (1996) and Efron and Tibshirani (1993: chapters 12, 14 and 22).

The bias-corrected (BC) method approximates the *MD* bias of the bootstrapped statistics and consequently adjusts the endpoint of the CI. If the *MD*-bootstrapped statistic equals the observed statistic (e.g., $t_{(500)}^* = t_{\text{obs}}$ when $B = 1,000$), the BC is exactly equal to the percentile CI. If the *MD* is greater than the observed statistic, the bootstrapped statistics are considered biased upwards relative to the observed statistic, so the CI endpoints are shifted left on the number line.

The bias-corrected and accelerated method (BC_a) (Efron, 1987) builds on the BC by approximating how much the statistic's variance changes as its value changes. If the acceleration is zero (meaning that the approximated variance is constant), the BC_a will have identical endpoints to the BC. If the approximated variance increases as the statistic's value increases, the acceleration is positive and the CI is shifted right.

As a comparison, consider a bootstrap distribution ($B = 999$) that has a positive bias and positive acceleration⁷. The endpoints for percentile CI are the 25th and 975th ordered bootstrap statistics. Using this same specific fictional distribution, the BC endpoints would be the 15th and 961st values. Notice how the upper point was shifted left 14 scores relative to the percentile, while the lower point was shifted only 10. Finally accounting for acceleration, the BC_a endpoints are the 19th and 967th values.

The approximate-bootstrap CI method (ABC) is an interesting analytic approximation to the BC_a . Strictly speaking, it does not belong in our definition of a CI modification because no bootstrap distribution is even produced. Consequently it is much less computationally expensive than the other three CI methods. However the time saved is unlikely to be noticed by an applied researcher.

Author: pls provide
definition

The *bootstrap-t*, or *percentile-t*, involves more than a simple re-mapping of order-bootstrap statistics to population quantiles. It is important to realize that a bootstrap-*t* is not the prototypical bootstrap of a *t* statistic, as shown in Example 3. The most direct explanation involves a scenario where a CI is sought for the difference between two group means, $d = \bar{X}_{\text{Group 1}} - \bar{X}_{\text{Group 2}}$. Before we describe the bootstrap-*t*, we first review how the prototypical bootstrap and the traditional parametric approach would be used. With the prototypical bootstrap, the first four stages outlined in Example 3 would remain the same, except that d_{obs} and d_b^* replace t_{obs} and t_b^* . Stage 5 would define a 95% CI as $[d_{(25)}^*, d_{(975)}^*]$. Alternatively with the traditional parametric approach, the 95% CI is $[d_{\text{obs}} - (t_{0.975})se_{\text{obs}}, d_{\text{obs}} + (t_{0.025})se_{\text{obs}}]$, where se_{obs} is the statistic's standard error and the critical values of *t* are determined by the theoretical student *t* distribution with $N - 1$ df (Hays, 1994, section 8.6).

With the bootstrap-*t*, the plug-in statistic (d in this scenario) is *studentized* in each bootstrap sample: $Z_b^* = (d_b^* - d_{\text{obs}})/se_b^*$. The se_b^* is the estimated standard error of d_b^* (calculated from the *b*th bootstrap). Authors commonly use the notation Z^* instead of t^* , to emphasize that it is not a conventional *t*, among other reasons. The 95% CI of the bootstrap-*t* is defined as $[d_{\text{obs}} - (Z_{(975)}^*)se_{\text{obs}}, d_{\text{obs}} + (Z_{(25)}^*)se_{\text{obs}}]$. Accounting for differences between the nominal α and the actual α_{obs} , the coverage probability for the bootstrap-*t* CI can be stated as:

$$\Pr \left(Z_{\left(\frac{\alpha}{2}\right)} \leq \frac{d_{\text{obs}} - d_{\text{hyp}}}{se_{\text{obs}}} \leq Z_{1-\frac{\alpha}{2}}^* \right) = \Pr \left(d_{\text{obs}} + \left(Z_{1-\frac{\alpha}{2}}^* \right) se_{\text{obs}} \leq d_{\text{hyp}} \leq d_{\text{obs}} + \left(Z_{\frac{\alpha}{2}}^* \right) se_{\text{obs}} \right) = 1 - \alpha_{\text{obs}} \quad (6)$$

Notice three things about the bootstrap-*t*. First, the critical values (i.e., $Z_{1-\frac{\alpha}{2}}^*$ and $Z_{\frac{\alpha}{2}}^*$) are the only change from the parametric CI; both approaches use the observed d to center the CI and the observed standard error to estimate the variability. Second, unlike the critical values

from a theoretical distribution, the bootstrap-*t* critical values won't necessarily be symmetric around zero. The algebraic re-arrangement in the previous equations demonstrates why $Z_{(975)}^*$ (which is actually larger than $Z_{(25)}^*$) is used for the lower endpoint of the CI. Third, it is the only bootstrap variation we discuss that needs a closed form equation for the statistic's standard error. Other bootstrap methods have estimated standard error by the standard deviation of the bootstrap distribution, as in Stage 5 in Example 4.

Two primary issues should be considered when choosing a CI method. Is it transformation invariant? How accurate is it? Both issues favor the BC_a . The bootstrap-*t* is very sensitive to the scale of the plug-in statistic. In contrast, the percentile, BC, BC_a , and ABC methods⁸ are unaffected by a monotonic transformation of the plug-in statistic. The same reject/retain decision would be made if the Fisher *r*-to-*z* transformation ($z_r = 0.5 \ln[(1+r)/(1-r)]$) is used in place of *r* for Examples 5 and 6.

The benefits of the newer CI methods are illustrated in another scenario, when the researcher is deciding between using a biased or unbiased estimator (i.e., $\Sigma(x_i - \bar{X})^2/N$ vs. $\Sigma(x_i - \bar{X})^2/(N-1)$) for the plug-in statistic. The BC, BC_a , and ABC approximate and attempt to correct for the downward bias, so the decision between the two variance estimators makes little difference (Efron and Tibshirani, 1993: 170). In the real world, choosing the preferred variance is trivial, because the properties of the two variance estimators have been studied extensively by statisticians. However if a different scenario involves unstudied estimators, the automatic bias correction is uniquely advantageous. In a sense, the choice between monotonic transformations doesn't matter because the corrective features of the CI method will pick the best one.

The accuracy of the CI should be considered as well, and the BC_a is the most accurate of the bootstrap CI methods that are transformation respecting. In fact, under a large class of problems, including non-parametric situations' the BC_a is more

accurate than even the traditional parametric approach (Efron, 1987: 199; Hall, 1992: 136). Because statistical inference is not perfect, the true coverage is the sum of the nominal coverage, α , and coverage error. If $B = 9,999$, the percentile 95% CI coverage area is actually $\Pr(r_{(250)}^* \leq \rho \leq r_{(9,750)}^*) = (1 - \alpha) + \text{CoverageError}$, where the coverage error shrinks to 0 as the sample size approaches the population size.

Percentile and parametric methods are *first-order accurate*, meaning the coverage errors in the CI are proportional to $1/\sqrt{N}$. However the BC_a is *second-order accurate*, having smaller errors which are proportional to $1/N$. It is notable that a robust statistical procedure with fewer assumptions can still outperform the traditional parametric procedures in many situations. For a thorough explanation of accuracy (and correctness), we recommend Hall (1992)⁹ and the other references mentioned at the beginning of this CI discussion.

Although the bootstrap- t is second-order accurate, it performs poorly when re-sampling statistics that are not variance stabilized, such as the sample correlation. Efron and Tibshirani (1993: 160) state that it ‘can give somewhat erratic results, and can be heavily influenced by a few outlying data points’ and recommend the BC_a instead. Additional

discussion can be found in DiCiccio and Efron (1996: 199) in the paragraph which begins, ‘More seriously, the bootstrap- t can be numerically unstable’

Although the BC_a is believed to be the best general CI method available now, it isn’t necessarily the best in all situations. Before choosing a CI for an applied analysis, we recommend searching for simulations of comparable conditions, to see which exhibited the most desirable Type I Error control and power. This strategy is further discussed later.

Hall (1992; appendix III) convincingly argues that a *confidence picture*, which he defines as a smoothed histogram of the bootstrap distribution, provides more useful information than a simple CI. Examples are shown in Figure 16.4, which clearly portrays the asymmetry involved. The left pane displays the 95% CI endpoints for some of the different methods that we have discussed. The right pane focuses on the BC_a and marks the endpoints of the CI with different coverage areas (80%, 90%, 95%, 99%). Note that it was not necessary to smooth these histograms because we used $B = 1,000,000$. The large bootstrap consumed 300 seconds, which was quicker than learning the syntax of a smoothing routine. Readers who

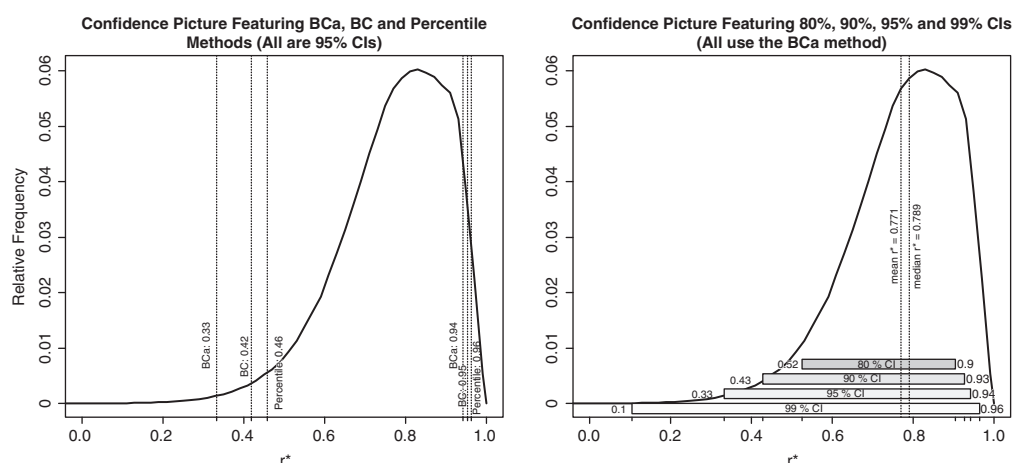


Figure 16.4 Histograms of the bootstrap distribution communicate many aspects of the inference beyond the locations of the two confidence interval (CI) endpoints. For comparison, the 95% parametric CI is [.44, .92].

appreciate Hall's confidence pictures may be interested in the graphical bootstrapping approaches described in Hall (1992, section 4.3.7) and especially in Davison and Hinkley (1997, section 4.2.4)¹⁰.

The final CI modification we discuss is the *double bootstrap*, in which each bootstrap sample is then bootstrapped (Efron, 1983, section 5). Based on the second layer's bootstrap distribution, quantiles in the initial bootstrap layer are adjusted up or down. The double bootstrap is at least second-order accurate, but is much more expensive than the procedures previously discussed (Martin, 1990). If 9,999 replications are used in the first layer and 4,999 in the second, the double bootstrap will calculate $(9,999 + 4,999 \times 9,999) = 49,995,000$ statistics.

Example 8

Here we revisit the scenario described in Example 3, which used a percentile CI. In exchange for the added complexity, this double bootstrap should be more accurate. A new bootstrap distribution is introduced in which B values of u are created in Stage B5:

- Stage A1. Collect the sample and calculate t_{obs} from the N scores.
- Stage A2. Prepare the sampling frame, which again is simply the observed sample in this scenario.
- Stage A3. Randomly draw N scores with replacement from the sampling frame. This will be the b th bootstrap sample, where $b = 1, \dots, B$. We will call this first level of bootstrap samples $X_1^*, X_2^*, \dots, X_b^*, \dots, X_S^*$. Repeat this Stage to form B first level bootstrap samples.
- Stage A4. Calculate the plug-in statistic, t_b^* , for each first-level bootstrap sample drawn in Stage A3.

Stages B1–B5 are repeated for each of the B bootstrap samples:

- Stage B1. The b th first-level bootstrap sample, X_b^* , now becomes the 'observed' sample for a second level of bootstrapping.
- Stage B2. Prepare the second-level sampling frames, which is simply X_b^* .

Stage B3. Randomly draw N scores with replacement from the first-level sampling frame. This will be the d th bootstrap sample, where $d = 1, \dots, D$. Second-level bootstrap samples will be denoted with two trailing asterisks: $X_{b,1}^{**}, X_{b,2}^{**}, \dots, X_{b,d}^{**}, \dots, X_{b,D}^{**}$. Repeat this Stage to form D second-level bootstrap samples.

Stage B4. Calculate the plug-in statistic, $t_{b,d}^{**}$, for each second-level bootstrap sample drawn in Stage B3. (Note that there will be $B \times D$ of these bootstrapped statistics at the second level by the conclusion of Stage A5.)

Stage B5. Calculate the b th value of u by counting how many times the second-level bootstrap t^{**} is less than the observed t :

$$u_b = \frac{1}{D} \sum_{d=1}^D (t_{b,d}^{**} < t_{\text{obs}}) \quad (7)$$

Stage A5. Calculate the positions and values for the double bootstrap CI. The adjusted quantile is the $[\alpha / 2 \times (B + 1)]$ th order value of u for the lower endpoint and the $[(1 - \alpha / 2) \times (B + 1)]$ th order value of u for the upper endpoint. When $\alpha = 0.05$ and $B = 999$, the adjusted quantiles are $u_{(25)}$ and $u_{(975)}$. For illustration, assume $u_{(25)} = 0.018$ and $u_{(975)} = 0.962$.

The positions of double bootstrap CI have been shifted left relative to the single bootstrap CI. The lower boundary is $0.018 \times (B + 1) = 18$ and the upper boundary is $0.962 \times (B + 1) = 962$ and therefore the CI is $[t_{(18)}^*, t_{(962)}^*]$. For comparison, the prototypical percentile CI will always be $[t_{(25)}^*, t_{(975)}^*]$.

A *nested bootstrap* can have even deeper nested loops (i.e., Stages C1–C5 can be inserted after Stage B4); this generalizes to the larger concept of *bootstrap iteration* (Chermick, 1999, section 3.1.4; Davison and Hinkley, 1997). A double bootstrap is a nested bootstrap with a single iteration. When bootstrap iteration is applied specifically to improve a CI, it is sometimes called *bootstrap calibration* (Efron and Tibshirani, 1993, section 18.1).

RE-SAMPLING DEPENDENT DATA

Although re-sampling usually is easy to conceptualize for iid data, dependent relationships between variables or observations can be difficult to model. The difficulty is building the appropriate type of dependence into \hat{F} . As mentioned previously, more assumptions are required when raw scores cannot be simply drawn from the sampling frame with equal probability. We would like to draw attention to three applications: (1) multiple regression; (2) time series, and; (3) complex sampling design.

Many behavior research studies involve two or more continuous-explanatory quantitative variables. Most re-sampling approaches to *multiple regression* resemble the previous semi-parametric bootstrap example, in which we estimate a model and its residuals, and then build bootstrap samples by combining the predicted values and bootstrapped residuals. The foundational concepts are explained concisely in Efron and Tibshirani (1993, chapter 9) and additional practical information is found in Davison and Hinkley (1997, chapters 6 and 7) and Chernick (1999, chapter 4).

When analyzing a *time series*, the relationship between consecutive observations is modeled. If scores were naively sampled in any order, the resulting empirical distribution would be an \hat{F}_{null} in which no relationship existed between the observations. An inferential conclusion from this nil hypothesis rarely will be useful. Although time series based on \hat{F}_{null} with a specified non-zero relationship are possible, in practice most are based on \hat{F}_{obs} .

In a time series, one way to construct \hat{F}_{obs} is by estimating the overall regression coefficients as well as the disturbances between successive terms. The sampling frame is formed from the estimated disturbance terms. Then the procedure progresses much like a semi-parametric bootstrap: the values for the T predicted time points ($\hat{y}_1, \hat{y}_2, \dots, \hat{y}_T, \dots, \hat{y}_T$) are added to randomly drawn T disturbances ($e_1^*, e_2^*, \dots, e_t^*, \dots, e_T^*$) to form one bootstrap sample of T values, and the dependencies are modeled through the es . This of course is

repeated B times to form B bootstrap samples. Typically in a time series, a bootstrap distribution contains the regression coefficients between the t th and t th $- 1$ observation. Additional bootstrap distributions of statistics such as the second-order coefficients (i.e., between the t th and t th $- 2$ observation) can be constructed as well.

The *moving-blocks bootstrap* is a second way to use \hat{F}_{obs} with a times series. The observed points are divided into overlapping blocks or windows of length L , which form the sampling frame. The blocks then are drawn randomly and spliced together to form one bootstrap sample. This process is repeated to form B bootstrap samples, and the desired sample statistics are collected to construct bootstrap distributions like before. The width of L should be related to the window of dependency around each observation. For example, an L of 4 implies that the t th time point is influenced by the previous three time points. Alternatively, a *stationary bootstrap's* value of L is not fixed, but instead is randomly determined.

Introductions to time series re-sampling can be found in Mammen and Nandi (2003), Efron and Tibshirani (1993, sections 8.5 and 8.6), Davison and Hinkley (1997, section 8.2), and Chernick (1999, chapter 5). Extensive coverage is provided in Lahri (2003) and Politis (2003). Econometrics has a large interest in time-series re-sampling, and behavioral researchers using these methods should survey their developments as well. Those who analyze spatial data may benefit from Lahiri's (2003, chapter 12) adaptation of the moving blocks bootstrap to this related problem.

Re-sampling approaches to *stratified samples* and *clustered data* apparently have become popular in many survey agencies such as the US Census Bureau and the Bureau of Labor Statistics (Shao, 2003: 193). Some modifications to the prototypical bootstrap are still required, particularly if the population size is finite. These concepts are discussed in Field and Welsh (2007), Kovar et al. (1988), and Davison and Hinkley (1997, section 3.7).

ADDITIONAL TECHNIQUES AND APPLICATIONS

A number of re-sampling methods and variants will not be treated within this chapter. Some are theoretically interesting, but not practical to implement in an applied research setting. Some are highly specific to the problem that is solved, which precludes general treatment. In either case, these variations are better grasped with a strong understanding of the foundational concepts we have discussed.

We mention a number of such specific methods in this section, with little elaboration. The section is not an exhaustive treatment, but rather a collection of re-sampling techniques potentially useful in behavioral research. Many have attracted attention primarily from theoretical statisticians, but hold promise for applied statistical settings. For more information, see the dedicated issue of *Statistical Science*, in which many leading bootstrap developers describe its outlook and impact on a number of fields (e.g., sociology, biostatistics, and econometrics; Casella, 2003).

- The *balanced bootstrap* has a sampling frame variation that controls for the frequency of observed statistics in all B bootstrap samples combined (Davison et al., 1986; Gleason, 1988), which may be useful if B must be small. Because the re-sampled scores are more evenly distributed across bootstrap samples, the balanced bootstrap resembles a Latin hypercube design (e.g., Gigli, 1996).
- *Pre-pivoting* a bootstrap statistic through bootstrap iteration helps increase accuracy. This process and its advantages are discussed in Beran (1987, 1988, 2003). The distribution of a *pivotal* statistic is independent of the parameter values (Hall 1992: 14; Efron and Tibshirani, 1993: 161).
- The *m out of n* bootstrap chooses a bootstrap sample size (M) that is smaller than the observed sample size (N ; see Bickel et al., 1997; Politis et al., 1999). When applied to complex survey data (Rao and Wu, 1988), it has been called the *rescaling bootstrap*.
- The *multiple-deletion jackknife*, or *delete-d jackknife*, excludes more than one observation from each jackknife sample. This improves problems

with unsmooth statistics like the *Median* (Efron and Tibshirani, 1993, section 11.7).

- A *jackknife-after-bootstrap* estimates the variability of an estimate made by a bootstrap. For instance, Efron and Tibshirani (1993, section 19.4) estimates the error of a standard error estimate.
- With the non-parametric bootstraps discussed above, each element in the sampling frame has a $1/M$ chance of being selected on any single draw, where M is the number of elements in the sampling frame. A *weighted bootstrap* alters these probabilities so that they are no longer uniform (Barbe and Bertail, 1995; Davison et al., 2003).
- A *Bayesian bootstrap* is a type of weighted bootstrap. The $1/M$ probability of each element is altered to correspond with its prior probability (Shao and Tu, 1995, chapter 10; Chernick, 1999, section 6.2.1). Boos and Monahan (1986) incorporate prior information differently; instead of placing a prior on each observation, they place a prior on the distribution of the plug-in statistic.
- Bootstrap aggregation, or *bagging*, increases the accuracy of a predicted value by averaging the predictions made by all bootstrap samples. It can be advantageous with non-linear statistics as well as discrete structures, such as classification trees (Hastie et al., 2001).
- Not to be confused with the double bootstrap, the *two-step* bootstrap is used as a multiple comparison procedure to contrast the outcomes of different groups (Beran, 2003).

Finally, there are variants with strange names and relatively exotic applications. Descriptive treatment of these is beyond the scope of this chapter, though we name several to pique the interest of technically-oriented readers. They include the *boosted bootstrap*, the *weird bootstrap*, the *wild bootstrap*, the *multinomial bootstrap*, *bootstrap tilting*, the *infinitesimal jackknife* and the *sandwich estimator*.

It should not be surprising that re-sampling can be used to build sampling distributions for many more established statistical techniques than we have discussed. We list some common applications that have been addressed with bootstrapping. The following can be found in Davison and Hinkley (1997): survival-analysis, hierarchical-data, logistic-regression, cross-validation, generalized-linear and generalized-additive models, and imputation of missing data. Some of these

methods are also covered in Manly (2007) at a more basic and readable level, including discriminant-function analysis and repeated-measures ANOVA.

Bollen and Stine (1992) describe how bootstrapping should be modified when testing fit statistics of structural equation models (SEM). Estimating of principal components and eigenvalues are demonstrated in Diaconis and Efron (1983) and Efron and Tibshirani (1993, chapter 24), respectively. Bootstrapping has been applied even to multidimensional scaling (Weinberg et al., 1984; Kiers and Groenen, 2006).

Analysis of directional data (e.g., circular and spherical) has been studied by Fisher et al. (1996). Evolutionary psychologists with a strong biological interest may benefit from bootstrapping methods for phylogenetic trees (Holmes, 2003; Soltis and Soltis, 2003). Finally, multivariate techniques are described in many places, including Pesarin (2001), Shao and Tu (1995, section 8.6) and Srivastava (2002, chapter 17).

PRAGMATIC RE-SAMPLING ADVICE

Converting to a re-sampling orientation isn't necessarily an all-or-nothing proposition. An overall statistical analysis can incorporate re-sampling methods in some areas and parametric in others. Consider a researcher fitting an SEM who is content with the parametric χ^2 distribution, but is reluctant to trust symmetric CIs around the means and covariances. A reasonable solution is to use the bootstrap only for the standard errors. If a later project is more suited for re-sampling the fit statistic as well, a Bollen-Stine (1992) bootstrap procedure could be incorporated. Another illustration of a heterogeneous strategy involves a researcher who uses a parametric estimation for standard error of the arithmetic mean with a bootstrapped standard error of the trimmed mean.

We regret not having the space (or knowledge) necessary to communicate the empirical performance and robustness of all the techniques described in this chapter.

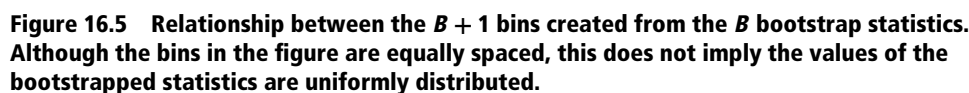
We advise that applied researchers read Monte Carlo experiments that simulated conditions similar to their own observed conditions (regardless of whether the statistical techniques incorporate re-sampling). These specialized articles are likely to discuss additional concepts and procedures that we did not cover. Many limitations of the bootstrap have been identified in this way, and some of these are discussed in the limitation section below.

Even if no relevant simulations have been published, we encourage creative researchers to use bootstrapping in novel ways. After all, one of the bootstrap's advantages is that new statistical approaches can be developed without complicated mathematical derivations. However, be cautious and understand that simply because a procedure can be conceived does not guarantee that it will behave desirably (though re-sampling methods are less suspect in principle than parametric methods). As in any unusual or unfamiliar statistical setting, flaws and assumptions can be hidden.

To protect against misuse, we advise that a 'proactive Monte Carlo analysis' be performed before the collected data set is analyzed, and ideally before the data are collected (Steiger, 2006). This precaution will help identify if the new procedure has: (1) a liberal Type I error rate; (2) inadequate power for the sample size; or (hopefully) (3) a robust nature and a promising chance to produce a reliable inference. The proactive analysis should be run with a variety of likely sample sizes and non-normal populations. Bootstrappers have an undeniable advantage over other practitioners, because bootstrapping is itself a type of Monte Carlo analysis. Any data generation routine used by the parametric bootstrap is a good candidate for the proactive analysis as well.

Bootstrap distribution size

As mentioned earlier, a subset is drawn of all possible bootstrap samples because complete enumeration is not practical with moderate and large sample sizes. For an applied researcher, choosing the size of B is



The size of B should be chosen so that $(B + 1)\alpha$ is an integer [Figure 16.5 may

help explain why it is not the more intuitive quantity $(B)\alpha$]. Each of the B bootstrapped statistics forms a boundary between the $B + 1$ quantiles of the bootstrap distribution. It may be convenient to think of these distinct quantiles as bins. Assume a two-tailed hypothesis, $\alpha = .05$, $B = 200$ and the sampling frame mimics \hat{F}_{obs} . If the hypothesized value (e.g., t_{null}) falls within the smallest five bins or the largest five bins, the p -value is less than α . The endpoints are defined as $t_{(0)}^* = -\infty$ and $t_{(B+1)}^* = \infty$. Boos (2003) provides a more mathematical explanation of the ‘99 Rule’ using an \hat{F}_{null} example.

Success of the bootstrap, in the sense of doing what is expected under a probability model for data, is not universal. Modifications to Efron's (1979) definition of the bootstrap are needed to make the idea work for estimators that are not classically regular. (Beran, 2003: 176)

The bootstrap will perform poorly when the population distribution is not accurately

reproduced in the empirical distribution (\hat{F}_{null} or \hat{F}_{obs}). Several problems that can confront behavioral research follow:

- Like all statistics that estimate population parameters, an inference will be compromised when the sampling process is flawed. This includes the presence of missing data or outliers, or a sample that is otherwise unrepresentative of the population.
- When dependence between variables is not modeled correctly, re-sampling (and parametric) procedures can be very misleading.
- When using sample values to estimate the maximum value of a population, the bootstrap inference is significantly biased downward. For example, it will fail when estimating θ on a uniform distribution with boundaries $(0, \theta)$ (Bickel and Freedman, 1981). Admittedly, this scenario does not arise in behavioral research frequently, but it is related to the next scenario.
- When the estimated parameter is close to a boundary, the bootstrap estimation is not consistent (Andrews, 2000). This could occur when a subject's minimum reaction time is being modeled. The problem is related to the fact that reaction time is restricted to be non-negative.
- Procedures that are heavily reliant on the asymptotic characteristics can perform poorly when small sample sizes are used. Schenker (1985) illustrates a scenario where the CIs are too narrow, such as when a (nominal) 90% CI has only a 78% coverage.
- Additional problems that are less likely to occur in behavioral research can be found in Andrews (2000; section 2), Mammen (1992) and LePage and Billiard (1992).

Although we have mentioned the important assumptions of the bootstrap throughout this chapter, we have not provided an exhaustive, authoritative list for two reasons. First, each bootstrap procedure carries different assumptions. For example, a semi-parametric correlational procedure assumes the residuals are iid, while a non-parametric correlational procedure does not even consider the distinction between model and residuals. Second, the mathematical proofs of the asymptotic assumptions (e.g., how \hat{F} approaches F as the sample size increases) are well beyond the scope of introductory bootstrap material. Thorough details of the assumptions common

to all bootstrap procedures can be found in the article by Bickel and Freedman (1981), and the books by Mammen (1992) and LePage and Billiard (1992). A more accessible summary is given by Young (1994).

Unique advantages of the bootstrap

Despite the limitations listed above, re-sampling methods nevertheless have a broad, powerful, and highly useful role in behavioral statistics. Caution should be applied, of course, in the use of both bootstrap and parametric procedures alike. Strategies that have been supported by positive results from proactive Monte Carlo analyses should be considered more trustworthy than strategies that have not; software and methods have been developed that support implementation of this type of research.


From the perspective of the research practitioner, the bootstrap's two strongest advantages arise from its non-parametric nature. One advantage is that inferences about non-normal populations are typically more reliable with re-sampling procedures than with parametric procedures. Taking the correlation coefficient as an example, bootstrap procedures can perform well with skewed populations (Beasley et al., 2007), composite populations (Lee and Rodgers, 1998) or populations with a restricted range (Mendoza et al 1991; Chan and Chan, 2004). Frequently with small samples, a researcher may have limited or no knowledge about a population distribution's shape; in these situations, a bootstrap can provide better protection against liberal p -values and misleading standard errors.

A second advantage (discussed previously in the 'plug-in principle' section) is that a practitioner can create an entirely new statistic without deriving the standard error formula and sampling distribution. For instance, we recently encountered a longitudinal dataset of 40 cases where each time point was a ratio of two scores. Re-sampling was the ideal inferential tool because we were unaware of an appropriate standard error equation, and thus a parametric method would require

considerable mathematical derivation or the parametric delta method [see Examples 1 and 2 in Boos (2003) for a similar perspective]. Furthermore, the delta method is based on large-sample approximation and can have trouble estimating parameters from a small sample (Bollen and Stine, 1990: 133, 137).

Software and programming

User-friendly bootstrapping software is certainly not as accessible as its parametric counterparts, and this status impedes the adoption of re-sampling techniques. Ironically, the flexibility that makes re-sampling theoretically attractive creates difficulty in programming a re-usable generalizable bootstrap routine. For a researcher interested in CIs and p -values, for example, a generalizable parametric software is easier to develop because only a handful of theoretical distribution functions are required; routines for distributions such as the incomplete beta and the non-central chi-square are widely available (e.g., Benton and Krishnamoorthy, 2003; Press et al., 2003).

Comparable bootstrap software requires the same routines for traditional statistics (such as correlation and ANOVA calculations) in addition to the recombination algorithms needed to build all the possible empirical sampling distributions. Notice that if parametric inferences were used, all eight examples in this chapter would be accommodated by one routine for the  distribution. However the re-sampling routines are not as reusable and at least six different algorithms are required.

Our experiences are consistent with Fan's (2003) evaluation of available bootstrapping software. Many SEM programs have good graphical user interfaces for bootstrapping the correlations' standard errors and CIs. Furthermore, the SEM bootstrap procedure by Bollen and Stine (1992) is a frequently used application of re-sampling in behavioral research. Outside of SEM however, the researcher probably will have to write new code or modify code that has been written previously. R and S-PLUS have the most active bootstrapping community, and Stata¹¹

and SAS¹² have a smaller, but still useful presence. Manly (2007, appendix), Edgington and Onghena (2007, section 15.5) and Good (2006, chapter 1; 2005, chapter 14) briefly review additional software programs that we have not used ourselves.

The S-PLUS and R routines that accompany the books by Efron and Tibshirani (1993) and Davison and Hinkley (1997) are reusable to some extent, and these provide a good starting point for beginning bootstrappers¹³. Their base routines try to encapsulate common bootstrap mechanisms (such as the selection of bootstrap samples and CI construction). The user first creates a specialized function that defines the Stage 1 and 2 behavior of the specific statistical procedure. Next, that function is passed to the reusable base routine as a parameter. However, it can be tricky to define this specialized function, even for common analyses, such as those that incorporate multiple groups, \hat{F}_{null} , or sampling frames that do not have exactly N rows.

It is difficult to create a reusable base routine that accommodates all of these scenarios, and we don't have promising ideas for improving the existing ones. We advise that users develop their own routines if the required analysis does not fit cleanly with the existing base routines. This is not as challenging as it may appear, even for non-professional programmers. A good starting point is to choose code for a similar analysis and adapt it to fit specific needs¹⁴. Most of the code accompanying this chapter was based on the Efron and Tibshirani (1993) 'bcanon' routine (i.e., nonparametric BC_a).

If a researcher writes their own routine, it should gracefully handle bootstrap samples that are not mathematically defined. For example, a t -test that has no variation in the scores will have a zero in the denominator. If not handled properly, one mischievous bootstrap sample will ruin the whole routine. Therefore, the program should wrap that calculation with error-handling code if the language possesses that error-handling capability¹⁵. It is also necessary to decide whether to redraw scores for that bootstrap sample or more simply to treat the undefined

t^* as a zero. It will occur infrequently enough that this decision is not likely to affect the bootstrap inference, but frequently enough to corrupt the bootstraps.

Despite the obstacles mentioned above, bootstrapping is worth the effort in situations where it holds a clear statistical advantage. These include, in particular, settings in which the desired statistic does not have a closed-form standard error equation, ones where the population distribution does not meet the necessary parametric assumptions, or especially in settings with small sample sizes combined with the previous restrictions.

Conclusion

Re-sampling was conceived decades before it was practical. In 1908, Gosset used a predecessor of the jackknife to create an empirical sampling distribution. He wrote 3,000 observed biometrical measurements on individual pieces of cardboard. After shuffling, he arbitrarily drew samples of $M = 4$ and recorded the 750 re-sampled samples. He then calculated their sample standard deviations and plotted a histogram. This process likely took Gosset several days, while the corresponding parametric distribution may have taken him only an afternoon to construct. His 1908 article (Student, 1908, section VI) presented diagrams comparing empirical and theoretical sampling distributions on the same axes, much like we have in Figure 16.1, 100 years later.

Prefacing the description of his re-sampling procedure Gosset said, 'Before I had succeeded in solving my problem analytically, I had endeavored to do so empirically'. He wasn't advocating that all practitioners follow his example and use re-sampling procedures. Instead he published the re-sampling exercise in order to justify that his theoretical sampling distribution could be a valid approximation. Once the quantiles of the theoretical distribution were tabled and published, the remaining computation was greatly reduced for all subsequent researchers. If an experiment's error was appropriately

modeled by the t , several days of unnecessary hand computation had been eliminated.

Fisher (1936: 59) later described a similar hypothetical re-sampling scenario and presented a similar argument: 'Initially, the statistician does not carry out this very simple and very tedious process, but his conclusions have no justification beyond the fact they could have been arrived at by this very elementary method.' One could argue that statisticians like Gosset and Fisher believed that re-sampling was the ideal approach, but theoretical approximations such as the t distribution were the only practical solution at the time. It is interesting that parametric methods and re-sampling theory emerged from the same minds at almost the same moment in statistical history; the two men who developed F and t^{16} – two of the most influential theoretical sampling distributions – used empirical sampling distributions as a primary justification.

Parametric procedures were the only reasonable approach in the era of mechanical calculators. And for at least three reasons, we imagine parametric procedures will continue to be valuable for many years. First, the theoretical sampling distributions of traditional statistics frequently are justifiable if the observed sample is very large or is drawn from a known population distribution. Second, there always will be occasions where parametric statistics are more convenient, even if the only advantage involves software limitations. We paraphrase the third reason from Efron and Tibshirani (1993: 61). Processes like relaxing assumptions in performing the bootstrap 'are not all pure gain'; standard error equations such as $\hat{\sigma}_{\bar{X}} = \sigma/\sqrt{n}$ and $\hat{\sigma}_r = (1 - r^2)/\sqrt{n - 3}$ teach us something about *theoretical* patterns and relationships that completely analytical techniques cannot.

Re-sampling is a technique that can benefit many applied statisticians. Although empirical sampling distributions will not solve all inferential problems, they can aid the creation of new statistics and can add robustness to many traditional ones. We hope this chapter has demonstrated that the different

procedural stages of the bootstrap can nimbly adapt to many unconventional experimental designs. As Efron and Gong (1983: 43) stated, “bootstrap” is not a well-defined verb, and ... there may be more than one way to proceed in complicated situations.’ If the statistic can be written as a simple or complex computational formula, or even as a highly complex algorithm, a bootstrap distribution can be developed to support computing CIs, standard errors, effect sizes, and hypothesis testing.

NOTES

1 See Pesarin (2001) and Edgington and Onghena (2007) for a modern treatment.

2 There are actually $\binom{\text{Unique arrangements of first } n_1 \text{ scores}}{\text{of remaining } n_2 \text{ scores}} = \binom{(n_1+n_2)!}{n_2!} = N! = 720$ different arrangements with respect to order (i.e., permutations). As each of the $\binom{n_1+n_2}{n_1} = 20$ arrangements are equally likely, it is equivalent and simpler to use this smaller number; re-sampling 720 samples does not provide any information beyond the unique 20.

3 The p -value for a one sided hypothesis is $\frac{\#\{t^* \geq t_{\text{obs}}\}}{B}$ or $\frac{\#\{t^* \leq t_{\text{obs}}\}}{B}$ (Davison and Hinkley, 1997, eq. 4.21).

4 Actually there are only $\binom{2n_1+n_2-1}{n_1} = 3,136$ unique arrangements, but $\binom{2n_2+n_1-1}{n_2} = 3,136$ unique arrangements, but they are not all equally likely. For example, there are more recombinations if six different scores are drawn (i.e., 6! unique ways with respect to order), than if one score is drawn six times (i.e., 1 unique way). Using N^N to count total samples accounts for these different probabilities of occurrence.

5 Some authors (e.g., Edgington and Onghena, 2007, section 1.12) distinguish between a permutation test and a randomization test. In their terminology, a randomization test is applied to data that have been randomly assigned, while a permutation test is the same procedure applied to non-randomized data.

6 For a variant opinion, see Pesarin (2001, section 5.6, remark 4).

7 For the record, the bias and acceleration approximations for this 95% CI are $z_0 = -0.1$, $a = 0.02$.

8 Actually there are two versions of the ABC. One is transformation respecting and one is not (Efron and Tibshirani, 1993: 331).

9 Hall (1992) defines CI methods with different names than most of the literature. The BC_a is referred to as the ‘ABC’ and he has definitions for ‘the percentile method’ and ‘the other-percentile method’ (which we identified as the ‘basic’ and ‘percentile’ method, respectively); see Manly (2007, section 3.3) for a direct comparison of the two percentile methods. Hall’s book relies on more mathematical explanations than most. A useful metaphor involving nested Russian Matryoshka dolls describes how bootstrap samples (F_2) are derived from observed samples (F_1), which are derived from population distributions (F_0). The relationship between F_2 and F_1 is assessed and projected on to the relationship between F_1 and F_0 , in order to make an inference about the unobservable F_0 .

10 As well as their ‘plot.boot’ routine for S-PLUS and R. Their routines are discussed below in the section ‘Additional techniques and applications’.

11 <http://www.stata.com/help.cgi?bootstrap>.

12 <http://support.sas.com/faq/003/FAQ00350.html> (address is case-sensitive).

13 In R, these routines are members of the ‘bootstrap’ and ‘boot’ packages and are free, even to those who don’t own the books. Packages are discussed in Chapter 13 of the June 2007 version of *An Introduction to R*. The latest version of this document is accessible through the help menu of R. After loading the desired library, help files will appear after typing ‘?bootstrap’ or ‘?boot’, depending on which package was loaded. Both packages have good help files, with ‘boot’ holding a slight advantage here.

14 In R, the actual code is displayed by typing the name of the base bootstrap routine (e.g., ‘bcanon’ when the ‘bootstrap’ package has been installed and loaded). Most R users recommend copying and pasting this code into a new word processing document. This permits the user to modifying and save the code easily. To execute the modified code, copy and paste it back into R.

15 For an explanation in R, type ‘?try’.

16 See Eisenhart (1979) for Fisher’s reformulation of Gosset’s distribution.

REFERENCES

- Andrews, D.W.K. (2000) ‘Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space’, *Econometrica*, 68: 399–405.
- Barbe, P. and Bertail, P. (1995) *The Weighted Bootstrap*. New York: Springer-Verlag.
- Beasley, W.H., DeShea, L., Toothaker, L.E., Mendoza, J.L., Bard, D.E., and Rodgers, J.L. (2007) ‘Bootstrapping to test for nonzero population correlation coefficients using univariate sampling’, *Psychological Methods*, 12: 414–433.

- Benton, D. and Krishnamoorthy, K. (2003) 'Computing discrete mixtures of continuous distributions: noncentral chisquare, noncentral t and the distribution of the square of the sample multiple correlation coefficient', *Computational Statistics and Data Analysis*, 49: 249–267.
- Beran, R. (1987) 'Pre-pivoting to reduce level error of confidence sets', *Biometrika*, 74: 457–468.
- Beran, R. (1988) 'Prepivoting test statistics: a bootstrap view of asymptotic refinements', *Journal of the American Statistical Association*, 83: 687–697.
- Beran, R. (2003) 'The impact of the bootstrap on statistical algorithms and theory', *Statistical Science*, 18: 175–184.
- Bickel, P.J. and Freedman, D.A. (1981) 'Some asymptotic theory for the bootstrap', *The Annals of Statistics*, 9: 1196–1217.
- Bickel, P.J., Götze, F. and van Zwet, W.R. (1997) 'Resampling fewer than n observations: gains, losses, and remedies for losses', *Statistica Sinica*, 7: 1–31.
- Bollen, K.A. and Stine, R.A. (1990) 'Direct and indirect effects: classical and bootstrap estimates of variability', *Sociological Methodology*, 20: 115–140.
- Bollen, K.A. and Stine, R.A. (1992) 'Bootstrapping goodness-of-fit measures in structural equation models', *Sociological Methods Research*, 21: 205–229.
- Boos, D.D. (2003) 'Introduction to the bootstrap world', *Statistical Science*, 18: 168–174.
- Boos, D.D. and Monahan, J.F. (1986) 'Bootstrap methods using prior information', *Biometrika*, 73: 77–83.
- Casella, G. (Ed.). (2003) 'Silver anniversary of the bootstrap', *Statistical Science* [Special issue], 18(2).
- Chan, W. and Chan, D.W.L. (2004) 'Bootstrap standard error and confidence intervals for the correlation corrected for range restriction: a simulation study', *Psychological Methods*, 9: 369–385.
- Chernick, M.R. (1999) *Bootstrap Methods: A Practitioner's Guide*. New York: Wiley.
- Davison, A.C. and Hinkley, D.V. (1997) *Bootstrap Methods and Their Application*. Cambridge, UK: Cambridge University Press.
- Davison, A.C., Hinkley, D.V. and Schechtman E. (1986) 'Efficient bootstrap simulation', *Biometrika*, 73: 555–566.
- Davison, A.C., Hinkley, D.V. and Young, G.A. (2003) 'Recent development in bootstrap methodology', *Statistical Science*, 18: 141–157.
- Diaconis, P. and Efron, B. (1983) 'Computer-intensive methods in statistics', *Scientific American*, May, 116–130.
- DiCiccio, T.J. and Efron, B. (1996) 'Bootstrap confidence intervals', *Statistical Science*, 11: 189–228.
- Edgington, E.S. and Onghena, P. (2007) *Randomization Tests* (4th edn.). Boca Raton, FL: Chapman and Hall/CRC.
- Eisenhart, C. (1979) 'On the transition from "Student's" z to "Student's" t ', *American Statistician*, 33: 6–10.
- Efron, B. (1979) 'Bootstrap methods: Another look at the jackknife', *Annals of Statistics*, 7: 1–26.
- Efron, B. (1982) *The Jackknife, the Bootstrap, and Other Resampling Plans*. Philadelphia: Society for Industrial and Applied Mathematics.
- Efron, B. (1983) 'Estimating the error rate of a prediction rule: Improvement on cross-validation', *Journal of the American Statistical Association*, 78: 316–331.
- Efron, B. (1987) 'Better bootstrap confidence intervals', *Journal of the American Statistical Association*, 82: 171–185.
- Efron, B. and Gong, G. (1983) 'A leisurely look at the bootstrap, the jackknife, and cross-validation', *The American Statistician*, 37: 36–48.
- Efron, B. and Tibshirani, R.J. (1993) *An Introduction to the Bootstrap*. Boca Raton, FL: Chapman and Hall/CRC.
- Fan, X. (2003) 'Using commonly available software for bootstrapping in both substantive and measurement analyses', *Educational and Psychological Measurement*, 63: 24–50.
- Fay, M.P. and Follmann, D.A. (2002) 'Designing Monte Carlo implementations of permutation or bootstrap hypothesis tests', *American Statistician*, 56: 63–70.
- Field, C.A. and Welsh, A.H. (2007) 'Bootstrapping clustered data', *Journal of the Royal Statistical Society Series B*, 69: 369–390.
- Fisher, N.I. and Hall, P. (1990) 'On bootstrap hypothesis testing', *Australian Journal of Statistics*, 32: 177–190.
- Fisher, N.I., Hall, P., Jing, B., and Wood, A.T.A. (1996) 'Improved pivotal methods for constructing confidence regions with directional data', *Journal of the American Statistical Association*, 91: 1062–1070.
- Fisher, R.A. (1935) *Design of Experiments* (1st edn.). Edinburgh: Oliver and Boyd.
- Fisher, R.A. (1936) 'The "index" coefficient of racial likeness" and the future of craniometry', *The Journal of the Royal Anthropological Institute of Great Britain and Ireland*, 66: 57–63.
- Fisher, R.A. (1966) *Design of Experiments* (8th edn.). New York: Hafner.
- Gigli, A. (1996) 'Efficient bootstrap methods: a review', *Statistical Methods and Applications*, 5: 99–127.
- Gleason, J.R. (1988) 'Algorithms for balanced bootstrap simulations', *The American Statistician*, 42: 263–266.
- Good, P.I. (2005) *Permutation, Parametric and Bootstrap Tests of Hypotheses* (3rd edn.). New York: Springer.

- Good, P.I. (2006) *Resampling Methods: A Practical Guide to Data Analysis* (3rd edn.). Boston: Birkhäuser.
- Hall, P. (1986) 'On the number of bootstrap simulations required to construct a confidence interval', *The Annals of Statistics*, 14: 1453–1462.
- Hall, P. (1992) *The Bootstrap and Edgeworth Expansion*. New York: Springer-Verlag.
- Hastie, T., Tibshirani, R. and Friedman, J. (2001) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer.
- Hays, W.L. (1994) *Statistics*. Belmont, CA: Wadsworth.
- Headrick, T.C. and Kowalchuk R. K. (2007) 'The power method transformation: its probability density function, distribution function, and its further use for fitting data', *Journal of Statistical Computation and Simulation*, 77: 229–249.
- Holmes, S. (2003) 'Bootstrapping phylogenetic trees: theory and methods', *Statistical Science*, 18: 241–255.
- Holmes, S., Morris, C., Tibshirani, R. and Efron, B. (2003) 'Bradley Efron: a conversation with good friends', *Statistical Science*, 18: 268–281.
- Kaiser, H.F. and Dickman, K. (1962) 'Sample and population score matrices and sample correlation matrices from an arbitrary population correlation matrix', *Psychometrika*, 27: 179–182.
- Kiers, H.A.L. and Groenen, P.J.F. (2006) 'Visualizing dependence of bootstrap confidence intervals for methods yielding spatial configurations', in Zani, S., Cerioli, A., Riani, M. and Vichi, M. (eds.), *Data Analysis, Classification and the Forward Search*. Berlin: Springer. pp. 119–126.
- Kovar, J.G., Rao, J.N.K. and Wu, C.F.J. (1988) 'Bootstrap and other methods to measure errors in survey estimates', *The Canadian Journal of Statistics*, 16: 25–45.
- Lahri, S. N. (2003) *Resampling Methods for Dependent Data*. New York: Springer.
- Lee, W. and Rodgers, J.L. (1998) 'Bootstrapping correlation coefficients using univariate and bivariate sampling', *Psychological Methods*, 3: 91–103.
- LePage, R. and Billard, L. (Eds.). (1992) *Exploring the Limits of Bootstrap*. New York: Wiley.
- Mammen, E. (1992) *When does bootstrap work?* New York: Springer-Verlag.
- Mammen, E. and Nandi, S. (2003) 'Bootstrap and resampling', in Gentle, J.E., Härdle, W. and Mori, Y. (Eds.), *Handbook of Computational Statistics*. Berlin: Springer-Verlag. pp. 467–495.
- Manly, B. (2007) *Randomization, Bootstrap and Monte Carlo Methods in Biology*. Boca Raton, FL: Chapman and Hall/CRC.
- Martin, M.A. (1990) 'On bootstrap iteration for coverage correction in confidence intervals', *Journal of the American Statistical Association*, 85: 1105–1118.
- Mendoza, J.L., Hart, D.E. and Powell, A. (1991) 'A bootstrap confidence interval based on a correlation corrected for range restriction', *Multivariate Behavioral Research*, 26: 255–269.
- Metropolis, N. and Ulam, S. (1949) 'The Monte Carlo method', *Journal of the American Statistical Association*, 44: 335–341.
- Miller, R.G. (1964) 'A trustworthy jackknife', *Annals of Mathematical Statistics*, 35: 1594–1605.
- Pesarin, F. (2001) *Multivariate Permutation Tests: With Applications in Biostatistics*. New York: Wiley.
- Politis, D.N. (2003) 'The impact of bootstrap methods on time series analysis', *Statistical Science*, 18: 219–230.
- Politis, D.N., Romano, J.P. and Wolf, M. (1999) *Subsampling*. New York: Springer.
- Press, W.H., Teukolsky, S.A., Vetterling, W.T. and Flannery, B.P. (2003) *Numerical recipes in C++* (2nd edn.). Cambridge, UK: Cambridge University Press.
- Quenouille, M.H. (1949) 'Approximate tests of correlation in time-series', *Journal of the Royal Statistical Society Series B*, 11: 68–84.
- Rao, J.N.K. and Wu, C.F.J. (1988) 'Resampling inference with complex survey data', *Journal of the American Statistical Association*, 83: 231–241.
- Rodgers, J.L. (1999) 'The bootstrap, the jackknife and the randomization tests: a sampling taxonomy', *Multivariate Behavioral Research*, 34: 441–456.
- Schenker, N. (1985) 'Qualms about bootstrap confidence intervals', *Journal of the American Statistical Association*, 80: 360–361.
- Shao, J. (2003) 'Impact of the bootstrap on sample surveys', *Statistical Science*, 18: 191–198.
- Shao, J. and Tu, D. (1995) *The Jackknife and Bootstrap*. New York: Springer.
- Soltis, P.S. and Soltis, D.E. (2003) 'Applying the bootstrap in phylogeny reconstruction', *Statistical Science*, 18: 256–267.
- Srivastava, M.S. (2002) *Methods of Multivariate Statistics*. New York: Wiley-Interscience.
- Steiger, J. H. (2006, October) 'Things we could have known: Some thoughts on seeing the future and rediscovering the past in data analysis and model selection'. Paper presented at the meeting of the Society of Multivariate Experimental Psychology, Lawrence, KA.
- Student (1908) 'The probable error of a mean', *Biometrika*, 6: 1–25.
- Tukey, J.W. (1958) 'Bias and confidence in not-quite large samples', [Abstract] *Annals of Mathematical Statistics*, 29: 614.
- Tukey, J.W. and Mosteller, F. (1986) 'Data analysis, including statistics', in Jones, L.V. (ed.). *The Collected*

Works of John W. Tukey. Volume 4. Monterey, CA: Wadsworth and Brooks/Cole. pp. 655–686. (Reprinted from Lindzey, G. and Aronson, E. (Eds.) (1968) *Handbook of Social Psychology* (2nd edn.). New York: Addison-Wesley. pp. 80–112 and 122–183.)

Weinberg, S.L., Carroll, J.D. and Cohen, H.S. (1984) 'Confidence regions for INDSCAL using the jack-knife and bootstrap techniques', *Psychometrika*, 49: 475–491.

Young, G.A. (1994) 'Bootstrap: more than a stab in the dark?' *Statistical Science*, 9: 382–395.