PDFlib PLOP: PDF Linearization, Optimization, Protection

Page inserted by evaluation version
www.pdflib.com – sales@pdflib.com

Advanced Statistics: Bootstrapping Confidence Intervals for Statistics with "Difficult" Distributions

Jason S. Haukoos, MD, MS, Roger J. Lewis, MD, PhD

Abstract

The use of confidence intervals in reporting results of research has increased dramatically and is now required or highly recommended by editors of many scientific journals. Many resources describe methods for computing confidence intervals for statistics with mathematically simple distributions. Computing confidence intervals for descriptive statistics with distributions that are difficult to represent mathematically is more challenging. The bootstrap is a computationally intensive statistical technique that allows the researcher to make inferences from data without making strong distributional assumptions about the data or the statistic being calculated. This allows the researcher to estimate confidence intervals for statistics that do not have simple sampling distributions (e.g., the median). The purposes of this article are to describe the concept of bootstrapping, to demonstrate how to estimate confidence intervals for the median and the Spearman rank correlation coefficient for non-normally-distributed data from a recent clinical study using two commonly used statistical software packages (SAS and Stata), and to discuss specific limitations of the bootstrap. **Key words:** bootstrap; resampling; median; Spearman rank correlation; SAS; Stata; NOSIC Score; confidence intervals. ACADEMIC EMERGENCY MEDICINE 2005; 12:360–365.

The use of confidence intervals in reporting the results of biomedical research has increased dramatically over the past several years. It is well known that confidence intervals provide more information than p-values, and editors of many scientific journals are now requiring or highly recommending their use.[1,2] While a number of articles report methods by which to calculate confidence intervals , they primarily focus on estimating confidence intervals for statistics with mathematically simple distributions, at least when the data themselves have a straightforward sampling distribution (e.g., normal or binomial distribution).[3–6]

In a recent publication, Okada et al. reported confidence intervals around Spearman rank correlation coefficients.[7] The primary objective of their study was to develop and evaluate a neurologic outcome measure, called the Neurologic Outcome Scale for Infants and Children (NOSIC), for pediatric research subjects with neurologic deficits. The NOSIC scale ranges from 3 to 100 and was applied independently by two clinical investigators to a cohort of patients in order to assess its reliability. The first rater (rater 1) applied the NOSIC to 157 patients and the second rater (rater 2) applied it to 84 of the 157 patients. These data are shown in Figures 1–3. It is evident from Figures 1 and 2 that the distributions are highly skewed, making reporting of the medians and Spearman rank correlation coefficient more valid than reporting the means and Pearson correlation coefficient for characterizing each rater's scores and the interrater reliability.

The confidence intervals for the Spearman rank correlation coefficients were estimated using the bootstrap, a statistical method based on resampling that can be used to perform statistical inference.[8] The purpose of this article is to describe the steps in bootstrapping, to demonstrate how to estimate confidence intervals using two commonly used statistical software packages (SAS[9] and Stata[10]) using the data from the Okada study, and to briefly discuss some limitations of the technique.

BOOTSTRAPPING

Bootstrapping was introduced in 1979 as a computationally intensive statistical technique that allows the researcher to make inferences from data without making strong distributional assumptions.[8,11] There are two distributions to consider. The first is the

From the Department of Emergency Medicine, Denver Health Medical Center (JSH), Denver, CO; the Department of Preventive Medicine and Biometrics, University of Colorado Health Sciences Center (JSH), Denver, CO; the Department of Emergency Medicine, Harbor–UCLA Medical Center (RJL), Torrance, CA; the Los Angeles Biomedical Research Institute at Harbor–UCLA Medical Center (RJL), Torrance, CA; and the David Geffen School of Medicine at UCLA (RJL), Los Angeles, CA.
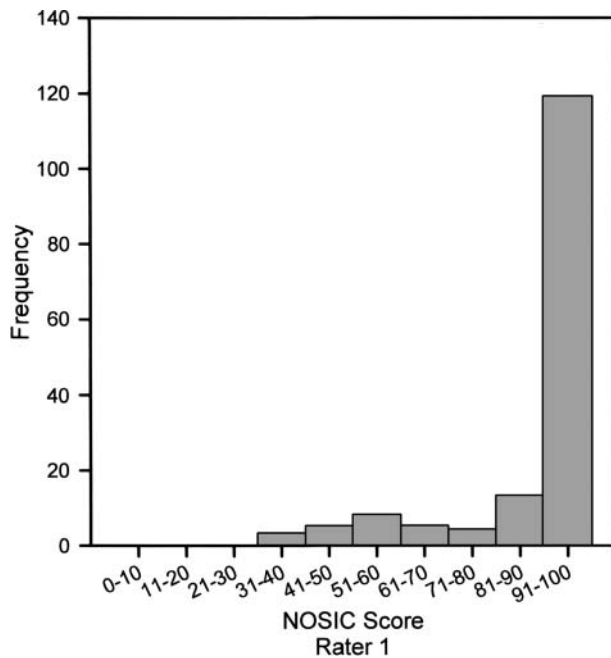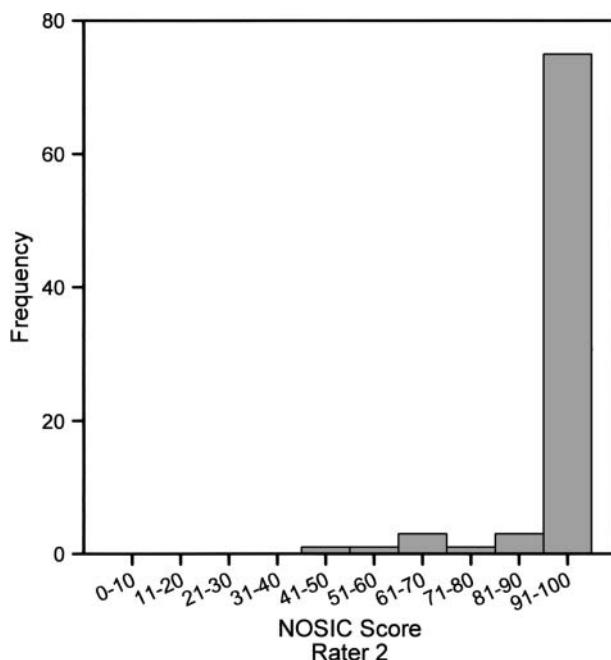
Address for correspondence and reprints: Jason S. Haukoos, MD, MS, Department of Emergency Medicine, Denver Health Medical Center, 777 Bannock Street, Mail Code 0108, Denver, CO 80204. Fax: 303-436-7541; e-mail: jason.haukoos@dhha.org.

Figure 1. Frequency histogram of scores using the Neurologic Outcome Scale for Infants and Children (NOSIC) for rater 1 (*n* = 157). The mean value is 90 (standard deviation = 16) and the median value is 97 (interquartile range: 92–100).

Figure 3. Neurologic Outcome Scale for Infants and Children (NOSIC) scores for rater 1 and rater 2 (*n* = 84). Data points are ``smeared'' using a normally-distributed random number generator to improve the representation of exactly overlapping data. As a result, some data points exceed 100. The Pearson correlation coefficient is 0.97 and the Spearman correlation coefficient is 0.77.

underlying distribution of the data themselves, which is frequently described as a probability function (e.g., normal, binomial, or Poisson) that shows all the values that the variables can have and the likelihood, or probability, that each will occur. The second is the
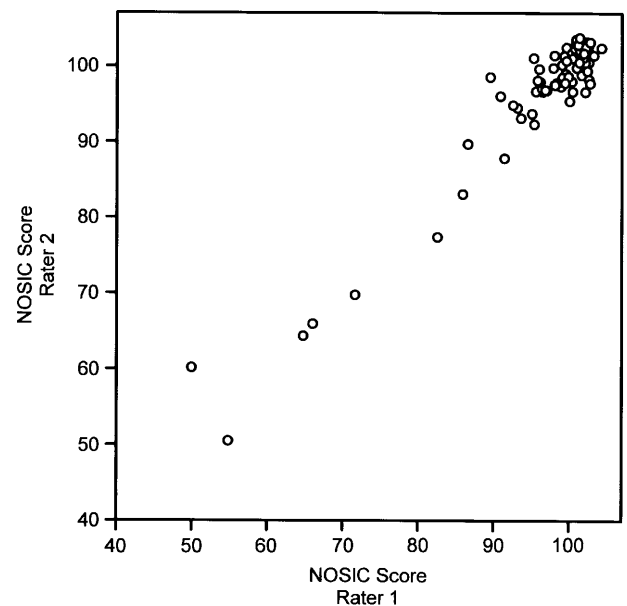

Figure 2. Frequency histogram of scores using the Neurologic Outcome Scale for Infants and Children (NOSIC) for rater 2 (*n* = 84). The mean value is 95 (standard deviation = 10) and the median value is 98 (interquartile range: 95–100).

distribution of the statistic (e.g., the median) calculated from the data. Both the items of data and the calculated statistic will vary in ways that can be described mathematically under the assumption that new sets of data were obtained or "sampled" and, for each set of data, a new statistic was calculated. More precisely, the statistic's sampling distribution is the probability of all possible values of the estimated statistic calculated from a sample of size *n* drawn from a given population.[12] Bootstrapping uses resampling with replacement (also known as Monte Carlo resampling) to estimate the statistic's sampling distribution. The sampling distribution, if it can be determined, may then be used to estimate standard errors and confidence intervals for that particular statistic.

The steps for estimating confidence intervals using the bootstrap are as follows (Figure 4): First, one uses resampling with replacement to create *m* resampled data sets (also known as bootstrap samples) that contain the same number of observations (*n*) as the original data set. To perform resampling with replacement, an observation or data point is randomly selected from the original data set and copied into the resampled data set being created. Although that data point has been "used," it is not deleted from the original data set or, using the usual terminology, is "replaced." Another data point is then randomly selected, and the process is repeated until a resampled data set of size *n* is created. As a result, the same observation may be included in the resampled data set one, two, or more
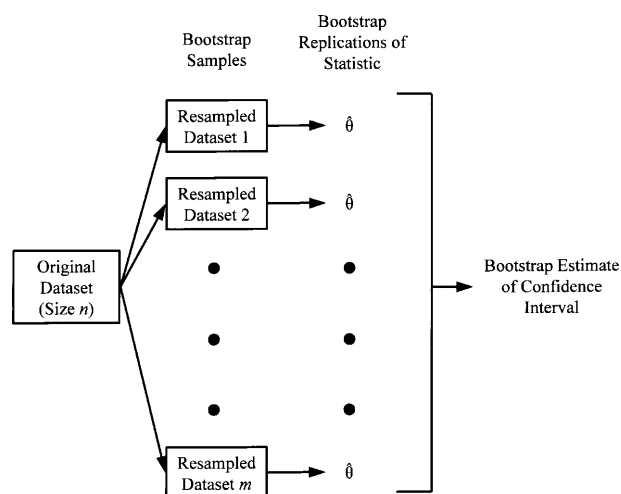
Figure 4. Schematic depiction of the steps in the bootstrap.

times, or not at all. Second, the descriptive statistic of choice is computed for each resampled data set. Third, a confidence interval for the statistic is calculated from the collection of values obtained for the statistic. At this point in the analysis, there are several options for computing confidence intervals, including the normal approximation method, the percentile method, the bias-corrected (BC) method, the bias-corrected and accelerated (BC_a) method, and the approximate bootstrap confidence (ABC) method.[8]

Each bootstrap sample should have the same sample size as the original data set. If the bootstrap sample sizes differ from the sample size of the original data set, the calculated estimation for the confidence interval may be biased.[13] A correction for this bias has been described, although there seems to be no practical advantage gained by performing the analysis in this manner.[14]

The normal approximation method computes an approximate standard error using the sampling distribution resulting from all the bootstrap resamples. The confidence interval is then computed using the z-distribution (original statistic ± 1.96 × standard error, for a 95% confidence interval). The percentile method uses the frequency histogram of the *m* statistics computed from the bootstrap samples. The 2.5 and 97.5 percentiles constitute the limits of the 95% confidence interval. The BC_a method adjusts for bias in the bootstrapped sampling distributions relative to the actual sampling distribution, and is thus considered a substantial improvement over the percentile method.[8] The BC_a confidence interval is an adjustment of the percentiles used in the percentile method based upon the calculation of two coefficients called "bias correction" and "acceleration." The bias correction coefficient adjusts for the skewness in the bootstrap sampling distribution. If the bootstrap sampling distribution is perfectly symmetric, then the bias correction will be zero.[8] The

acceleration coefficient adjusts for nonconstant variances within the resampled data sets.[8] The ABC method is an approximation of the BC_a method that requires fewer resampled data sets than the BC_a method.[8]

As a general guideline, 1,000 or more resampled data sets should be used when calculating a BC_a confidence interval.[11] As a result of not having to calculate bias correction, a smaller value, in the range of 250, can be used when using the percentile method for estimating a confidence interval.[13] As the number of resampled data sets decreases, more variability is introduced into the confidence interval estimation (i.e., the variability is inversely related to the number of resampled data sets).[8,13]

Example 1: Determining a Confidence Interval around a Median Value. A median value is defined as the observation at the 50th percentile in a set of data ordered from the lowest value to the highest value.[15] This measure of center for a set of values is commonly reported and is considered a more valid definition of center when the frequency distribution of the variable is skewed (i.e., not symmetric around its center). Unlike the mean, there is no simple method for calculating the 95% confidence interval (95% CI) for the median, and it is not valid to use a 95% CI calculated from the standard error to represent the 95% confidence for the median value, unless the distribution of the underlying data is normal. As a result, the bootstrap can be used to estimate the sampling distribution of the median. The central limit theorem states that as the number of resampled data sets increases, the distribution of the resulting statistic, in this case the median, will become approximately normal.[15] This subsequently allows for a relatively unbiased estimation of the confidence interval.

The steps required to bootstrap the 95% CI for a median value are: 1) to resample with replacement from the original data set, creating *m* bootstrapped data sets; 2) to independently compute the median value for each bootstrapped data set; and 3) to compute the 95% CI from the set of computed median values from the bootstrapped data sets using either the normal approximation method, the percentile method, the BC method, the BC_a method, or the ABC method.

These steps can be accomplished using the SAS software program (SAS Institute, Inc., Cary, NC) as follows. The SAS macro JACKBOOT, which can be obtained from the SAS Web site, must be invoked prior to performing a bootstrap analysis in SAS.[16] A "macro" is a program that can be executed by SAS and that may be modified by the user, while a SAS procedure is a "fixed" program that performs a specific statistical calculation or other task. The JACK-BOOT macro requires another macro (called ANALYZE) to be written that provides it with the procedure

whose result (e.g., the median of the original data set) requires bootstrapping. The univariate procedure (PROC UNIVARIATE) in SAS is used to compute the median value for a group of observations. The following is the ANALYZE macro, modified to bootstrap a 95% CI around a median value for the variable "normscr1" (NOSIC score for rater 1):

```
%macro analyze (data=, out=);
proc univariate noprint data=&data;
  output out=&out (drop=_freq_ _type_)
  median=median;
  var normscr1;
  %bystmt;
run;
%mend;
```

In SAS, the "%macro" term indicates the beginning of a macro, and is followed by its title (i.e., "analyze"). The "%mend" term indicates the end of a macro, and all text between "%macro" and "%mend" is called macro text. In this example, PROC UNIVARIATE is invoked with the "noprint" option. The "data=&data" term references the original data set through the JACKBOOT macro using the "%boot" term (see below). The "output" statement directs SAS to create a temporary output file for only the median values, as indicated by the term "median=median," for the variable "normscr1." The "%bystmt" term references a macro within the JACKBOOT macro that computes a statistic (in this case, the median) for the original data set and for each resampled data set.

The ANALYZE macro is followed immediately by the following bootstrap commands:

```
%boot (data=temp, samples=2500);
%bootci (percentile);
%bootci (bca);
```

In this example, the ANALYZE macro is used by the JACKBOOT macro to apply the statistical procedure (PROC UNIVARIATE) to the original data set (data=temp, referenced in the "%boot" statement). The "%boot" command invokes the bootstrap procedure, resulting in 2,500 bootstrapped samples, and the "%bootci" command invokes the bootstrap confidence interval procedure. The first "%bootci" command uses the percentile method to compute a 95% CI for the median and the second "%bootci" command uses the BC_a method to compute a 95% CI for the median of the variable "normscr1." The median value was 97 [interquartile range (IQR): 92–100, range 32–100], and the 95% CIs for the median were 96–98 (percentile) and 97 to 98 (BC_a).

Using Stata (Stata Corporation, College Station, TX) to perform the same calculations is substantially simpler. The following Stata commands compute the median value for the variable "normscr1" and bootstrap the 95% CIs using the normal, percentile, and BC_a methods using 2,500 resamples[17]:

```
centile normscr1
bs ''centile normscr1'' ''r(c_1)'',
rep(2500)
```

The "centile" command calculates the median value for the variable "normscr1." The "bs" command calculates a bootstrapped confidence interval for the median value for the variable "normscr1." The primary code appears in the first quotations, "'r(c_1)'" refers to the reference statistic for which the 95% CI will be calculated, and "rep(2500)" indicates the number of resampled data sets. After the primary command has been executed, the command "return list" can be used to display the codes for each of the resulting statistics for the primary command. In this example, "c_1" is the code that refers to the median value calculated by the "centile" command.

Example 2: Determining a Confidence Interval around a Spearman Rank Correlation Coefficient. The Spearman rank correlation coefficient is the nonparametric counterpart to the parametric Pearson correlation coefficient.[15] The Pearson correlation coefficient is a valid statistical technique for determining correlation between two normally-distributed continuous variables. On the other hand, the Spearman rank correlation coefficient is a valid statistical technique for determining correlation between two non-normally-distributed continuous variables.

The PROC CORR procedure in SAS is used to compute the Pearson correlation coefficient, and there are two methods for computing the Spearman rank correlation coefficient. The first method simply involves incorporating the option "spearman" into the PROC CORR statement. The second method involves ranking the data, using PROC RANK, prior to using PROC CORR.

The following illustrates the ANALYZE macro used by the JACKBOOT macro to perform the bootstrap in SAS:

```
%macro analyze (data=, out=);
proc rank data=&data out=tempdata;
  var normscr1 normscr2;
  %bystmt;
proc corr noprint
  data=tempdata
  out=&out (rename=(_type_=stat
  _name_=with));
  var normscr1 normscr2;
  %bystmt;
run;
%mend;
```

The macro text in this example includes the PROC RANK command for variables "normscr1" and "normscr2." This command is followed by the PROC CORR command, which performs correlation of the two ranked variables for each resampled data set. The "out=tempdata" term writes a temporary output file

of all ranked resampled data sets. This is read as an input file using the term "data=tempdata" in the PROC CORR command.

Again, the ANALYZE macro is followed immediately by the bootstrap commands:

```
%boot (data=temp, id=stat with,
samples=2500);
%bootci (percentile, id=stat with);
%bootci (bca, id=stat with);
```

In this example, 2,500 bootstrapped samples were created, and the percentile and BC_a methods were used to compute 95% CIs for the Spearman rank correlation coefficient between the variables "normscr1" and "normscr2" (NOSIC score for rater 2). The Spearman rank correlation coefficient was 0.77 for the original data set and the 95% CIs were 0.62–0.88 (percentile) and 0.62–0.87 (BC_a).

Again, it is simpler to perform this calculation using Stata. The following Stata commands compute the Spearman rank correlation coefficient between "normscr1" and "normscr2," and bootstrap the 95% confidence intervals using the normal, percentile, and BC_a methods using 2,500 resamples:

```
spearman normscr1 normscr2
bs ``spearman normscr1 normscr2''
``r(rho)'', rep(2500)
```

The "spearman" command calculates the Spearman rank correlation coefficient for "normscr1" and "normscr2." The primary code appears in the first quotations, "r(rho)" refers to the reference statistic for which the 95% CI will be calculated, and "rep(2500)" indicates the number of resampled data sets.

Limitations of the Bootstrap. Although the idea of the bootstrap has been around for nearly two centuries, theoretical work on the bootstrap is relatively recent and, therefore, the limitations of the bootstrap are not entirely understood.[11] The bootstrap is a tool used, in part, to calculate confidence intervals for point estimates of descriptive statistics. The bootstrap should not be used to compute point estimates themselves, however. The sampling distribution of the bootstrapped statistics is frequently not symmetric. Thus, computing point estimates in this manner may reflect, as opposed to alleviate, biased estimation from the samples.[11] The extent of bias can be estimated but is subject to high variability, making bias correction infeasible.[8]

The most important limitation of the bootstrap is the assumption that the distribution of the data represented by the sample is a reasonable estimate of the population distribution function from which the data are sampled. In other words, the sample must reflect the variety and range of possible values in the population from which it was sampled. If the distribution of data from the sample does not reflect

the population distribution function, then the random sampling performed in the bootstrap procedure may add another level of sampling error, resulting in invalid statistical estimations.[18] This emphasizes the importance of obtaining quality data that accurately reflect the characteristics of the population being sampled.

Additionally, the smaller the original sample, the less likely it is to represent the entire population. Thus, the smaller the sample, the more difficult it becomes to compute valid confidence intervals. The bootstrap relies heavily on the tails of the estimated sampling distribution when computing confidence intervals, and using small samples may jeopardize the validity of this computation.[18]

Random sampling performed in the bootstrap procedure also adds another level of potential sampling error. This, as mentioned previously, is reflected in the variation and bias estimates commonly performed during a bootstrap analysis.

CONCLUSIONS

The bootstrap is a relatively simple statistical concept that requires computationally intensive procedures to implement. Modern statistical software packages now allow researchers to employ relatively simple programming to compute confidence intervals for statistics with inconvenient or unknown sampling distributions.

References

1. Uniform requirements for manuscripts submitted to biomedical journals. International Committee of Medical Journal Editors. JAMA. 1997; 277:927–34.
2. Cooper RJ, Wears RL, Schriger DL. Reporting research results: recommendations for improving communication. Ann Emerg Med. 2003; 41:561–4.
3. Blyth CR. Approximate binomial confidence limits. J Am Stat Assoc. 1986; 81:843–55.
4. Troendle JF, Frank J. Unbiased confidence intervals for the odds ratio of two independent binomial samples with application to case–control data. Biometrics. 2001; 57:484–9.
5. Young KD, Lewis RJ. What is confidence? Part I: the use and interpretation of confidence intervals. Ann Emerg Med. 1997; 30:307–10.
6. Young KD, Lewis RJ. What is confidence? Part II: detailed definition and determination of confidence intervals. Ann Emerg Med. 1997; 30:311–8.
7. Okada PJ, Young KD, Baren JM, et al. Neurologic outcome score for infants and children. Acad Emerg Med. 2003; 10:1034–9.
8. Efron B, Tibshirani RJ. An Introduction to the Bootstrap. New York: Chapman & Hall/CRC, 1998.
9. SAS Version 8.2. SAS Institute, Inc., Cary, NC.
10. Stata Version 8. Stata Corporation, College Station, TX.

11. Mooney CZ, Duval RD. Bootstrapping: A Nonparametric Approach to Statistical Inference. Beverly Hills, CA: Sage Publications, 1993.

12. Levy PS, Lemeshow S. Sampling of Populations: Methods and Applications—3rd edition. New York: John Wiley & Sons, 1999.

13. Efron B, Tibshirani R. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. Stat Sci. 1986; 1:54–77.

14. Bickel PJ, Freedman DA. Some asymptotic theory for the bootstrap. Ann Stat. 1981; 9:1196–217.

15. Zar JH. Biostatistical Analysis—4th edition. Englewood Cliffs, NJ: Prentice Hall, 1999.

16. Jackknife and bootstrap analyses: SAS jackboot macro. Available at: http://ftp.sas.com/techsup/download/stat/jackboot.html. Accessed Apr 9, 2004.

17. Stata17ase Reference Manual: Volume 1, A-F, Release 8. College Station, TX: Stata Press, 2003.

18. Schenker N. Qualms about bootstrap confidence intervals. J Am Stat Assoc. 1985; 80:360–1.