THE TEACHING AND PRACTICAL IMPLEMENTATION OF THE NON-PARAMETRIC BOOTSTRAP

Cornelia J. Swanepoel
Potchefstroom University for CHE
South Africa

The bootstrap is a general resampling procedure which can be applied to estimate the sampling distribution of a statistic. From the statistical practitioner's point of view it has attractive properties because it requires few assumptions, little modeling or analysis, and can be applied in an automatic way in a wide variety of situations regardless of their theoretical complexity. The bootstrap can provide answers to questions that are too complicated for traditional statistical analyses, which are usually based on asymptotic normal approximations. A brief discussion of the non-parametric bootstrap is presented, followed by examples and illustrations. Possible suggestions regarding the teaching of these concepts at various levels are made. The key requirements for computer implementation of the bootstrap method include a flexible programming language with a collection of reliable quasi-random number generators, a wide range of built-in statistical bootstrap procedures and a reasonably fast processor. The use of the statistical languages S and Fortran, using the current commercial versions S-Plus 4.5 and Digital Fortran 6.0, are illustrated.

INTRODUCTION

Central to statistical science stands the explicit recognition of uncertainty. Educators traditionally prepared students to deal with uncertainty by training them in probability and statistical theory. Shaughnessy (1977) describes most courses offered at university level to be rule-bound recipe-type courses or overly mathematized traditional instructions, delaying the student's development of statistical intuition and skilful application of modern technology. Biehler, et al. (1988) identified both the lack of experience in data manipulation and the tendency of teachers to avoid problems that depend on students' computational skills, as two *targets of difficulty* in statistical education. Efron and Tibshirani (1993) stated that the traditional road to statistical knowledge is blocked by a formidable wall of mathematics.

Although Efron (1979) introduced the bootstrap methodology more than two decades ago, it is still a relatively new statistical tool to the practitioner. *The bootstrap is a newly developed technique, based on modern computer power and technology for making certain kinds of statistical inferences. It provides a way of escaping from some mathematical handicaps and answers many real statistical questions without formulas, such as assigning measures of accuracy to statistical estimates.* However, knowledge and mastering of basic traditional statistical concepts are essential for the bootstrap to be applied correctly. Rice (1995) advises the inclusion of traditional topics, e.g. methods based on likelihood, topics in descriptive statistics and data analysis, interpretation of graphical displays, aspects of experimental design, and realistic applications of some complexity. But these statistical topics are generally not taught in secondary schools which causes the bootstrap course to be presented either as a final section in the under-graduate statistics course, or as a thoroughly defined honours statistical course, where a well written book such as Efron and Tibshirani (1993) can be useful. A possible text book for a follow-up bootstrap course, i.e. for an M.Sc. course for example, is Davison and Hinkley (1997). However, the bootstrap method can be introduced gradually in layman's terms wherever an appropriate opportunity arises in the under-graduate years, as it is done by Rice (1995), where the parametric bootstrap is introduced in Chapter 8 in an informal, intuitive way to a natural situation. The *simulation* concept is simultaneously introduced as being an interesting frequentistic way of imitating the truth contained in a random sample.

Having completed basic statistical training, students attending a bootstrap course will possess enriched instincts about statistical concepts such as *random sample, parameter, estimator, confidence interval, empirical distribution function, stepfunction* and especially *sampling distribution of a statistic,* on which modern inference techniques can be built. Due to the influential development of bootstrap methods during the past decade, bootstrap courses should also be included as an essential component of any statistical training program for statistical practitioners working in the fields of biomedical sciences, psychology, education, economics, communications theory, sociology, genetic studies, epidemiology, geology, physics, astronomy, financial mathematics and other practical areas.

The *non-parametric bootstrap* belongs to the general subfield *Non-parametric Statistics* which is defined by Dudewicz (1976) as the subfield of statistics *that provides statistical inference procedures which rely on weak assumptions* (or no assumptions at all) *about the underlying distribution of the population*. Statistical practitioners should use non-parametric procedures only in so far as the assumptions about the underlying distribution are seriously doubtful in their validity. Efron (1979) states that the bootstrap is a way to pull oneself up (from an unfavourable situation) by one's bootstrap, to provide trustworthy answers despite of unfavourable circumstances. In ideal parametric situations traditional ways or parametric methods such as the parametric bootstrap, may be more applicable, due to the fact that the more information is known and used about the underlying distributions, the more accurate statistical inference procedures will be. When assumptions are *not* violated, non-parametric procedures will usually have greater variance (in point estimation), less power (in hypothesis testing), wider intervals (in confidence interval estimation), lower probability of correct selection (in ranking and selection) and higher risk (in decision theory) when compared to a corresponding parametric procedure.

ESSENTIAL CORNERSTONES

The bootstrap methodology depends on the following concepts:

- *An original random data sample of size n, denoted by $x_1,\cdots,x_n$*: It is thought of as the outcomes of independent and identically distributed (i.i.d.) random variables $X_1,\cdots,X_n$ whose probability density function (PDF) and cumulative distribution function (CDF) will be denoted by f and F respectively. The sample is used for inference purposes regarding a population characteristic, generally denoted by θ, using a statistic $T_n = T_n(X_1,\cdots,X_n)$ whose value for the sample is t_n. If it is assumed that T_n is an estimate for θ, the attention is focused on the sampling distribution of T_n, to answer questions about, for example, the standard error, bias and quantiles of this distribution. The quantiles are needed for determining, among others, bootstrap confidence intervals for θ from probabilities such as $P(a_1 \le T_n(X_1,\cdots,X_n) \le a_2)$.
- *The empirical distribution function* (EDF): Having observed a random sample of size n from the CDF F, an estimate of F, say $\hat{F}$, can be constructed from this sample. The EDF $\hat{F} = F_n$ is defined to be the discrete distribution that puts probability $1/n$ on each value x_i, $i = 1,2,\cdots,n.$ The EDF can also be written as $F_n(x) = n^{-1}\sum_{j=1}^{n} I(x_j \le x),$ where $I(\cdot)$ is the indicator function. The values of the EDF are fixed $(0,1/n,2/n,\cdots,n/n)$. Efron and Tibshirani (1993) showed that all the information about F contained in $x_1,\cdots,x_n$ is also contained in the EDF, so that the bootstrap observations are often generated from the EDF by computer methods when F is unknown.
- *Alternative estimates of F*: Bootstrap observations are often generated from s*moother estimates of F,* such as kernel estimates. This procedure is referred to as the *smoothed bootstrap procedure*. The kernel estimate of F which is defined by $\hat{F}(x) = F_{n,c}(x) = n^{-1}\sum_{i=1}^{n} K\left(\left(x - X_i\right)/c\right)$, where $c = c_n$ is a sequence of smoothing parameters such that $c_n \to 0$ as $n \to \infty$, and K a known continuous CDF symmetric around zero, produces a smooth version of F. Azzalini (1981) obtained second order results showing asymptotic improvement in estimating F by $F_{n,c}$ instead of F_n, provided certain regularity conditions on F are met and that $\{c_n\}$ converges at a specific rate to zero. However, questions of whether the

smoothed bootstrap is superior to the classical bootstrap, and which smoothing parameter c is to be used, are important research topics. Silverman and Young (1987), DiCiccio and Romano (1989), Hall (1990), De Angelis and Young (1992), El-Nouty and Guillou (2000), and several others have produced valuable work in this field.

- *The plug-in principle:* This concept, discussed in Efron and Tibshirani (1993), is a simple method of estimating parameters from samples and a handy tool in teaching the bootstrap method. It enhances that, if some aspect of a probability distribution F is to be determined from a random sample drawn from F, and if, for example, the EDF $\hat{F} = F_n$ is used to estimate F, any interesting aspect of F such as its mean, median or correlation is estimated by using the corresponding aspect of F_n. For example, the plug-in estimate of a parameter $\theta = t(F)$ is defined to be $\hat{\theta} = t(\hat{F})$. The bootstrap can then be used in an automatic way to study the bias and standard error, for example, of $\hat{\theta} = t(\hat{F})$, no matter how complicated the functional mapping $\theta = t(F)$ may be.
- *Bootstrap sample:* A bootstrap sample is defined to be a random *i.i.d.* sample of size m drawn from the EDF F_n, *with replacement* from the population of n objects $\boldsymbol{x} = (x_1, x_2, \cdots, x_n)$, and consists of members of the original data set $x_1, x_2, \cdots, x_n$, some appearing zero times, some appearing once, twice, etc., in which case we denote the bootstrap sample by $\boldsymbol{x}^*$ = $(x_1^*, x_2^*, \cdots, x_m^*)$. The *-notation indicates that $\boldsymbol{x}^*$ is not the actual data set $\boldsymbol{x}$, but rather a *resampled* version of $\boldsymbol{x}$. A large number of independent bootstrap samples are usually needed, which are easily generated repeatedly by computer methods, using a random number device to select integers $i_1, i_2, \cdots, i_m$, each of which equals any value between 1 and n with probability $1/n$. The bootstrap sample then consists of the corresponding members of $x_1, x_2, \cdots, x_n$, i.e. $x_1^* = x_{i_1}, x_2^* = x_{i_2}, \cdots, x_m^* = x_{i_m}$. Usually $m = n$ (classical bootstrap). However, Swanepoel (1986) defined the *modified bootstrap* procedure ($m \neq n$) and recommends this method for cases where the classical bootstrap fails.
- *Computer skills*: To apply the bootstrap practically, students should have achieved at least a user knowledge of computer packages such as Fortran or S-Plus. Shaughnessy (1977) remarks that the use of computers to conduct Exploratory Data Analysis (EDA) is the direction for statistical education to proceed on, and during the past decade skills involving the use of systems such as SAS, Statistica, BMDP, S-Plus and several others, became part of the statistics curriculum. *Monte Carlo* studies include the application of simulation methods according to bootstrap algorithms and rules, utilising the speed and availability of modern high speed computers. Most questions involving the size of simulation studies are answered in Efron and Tibshirani (1993).
- *The place of the non-parametric bootstrap in statistical inference*: It should be realised that *if no particular mathematical model with adjustable constants and parameters that fully determines F exists, i.e. when aspects of F are unknown, the use of non-parametric analysis, such as the non-parametric bootstrap, is inevitable. Also, when F is known but T_n is a statistic of complexity, explicit probabilistic calculations are often impossible to conduct, and the application of non-parametric methods is essential to estimate the desired quantities.*

THE NON-PARAMETRIC BOOTSTRAP ESTIMATION TECHNIQUE

Let $\hat{\theta} = T_n(\boldsymbol{x})$ be a calculated estimate of a parameter $\theta = t(F)$. Corresponding to a bootstrap data set $\boldsymbol{x}^*$, $\hat{\theta}^* = T_n(\boldsymbol{x}^*)$ then denotes one bootstrap replication of $\hat{\theta}$, i.e. the quantity $T_n(\boldsymbol{x}^*)$ resulted from applying the same function $T_n(\cdot)$ to $\boldsymbol{x}^*$ as was applied to $\boldsymbol{x}$. This concept is applied repeatedly in the following examples, illustrating the implementation of the bootstrap method in assessing the accuracy of an estimator $\hat{\theta}$ of a parameter θ.

Example 1: Estimating the *standard error* of the statistic $\hat{\theta}$: It is assumed that $\hat{\theta}$ is a statistic of *complexity,* in which case the computation of an exact numerical value of $\mathrm{se}_{\mathrm{F}}(\hat{\theta})$, the standard error of $\hat{\theta}$, will be complicated or impossible. *The bootstrap estimate of* $\mathrm{se}_{\mathrm{F}}(\hat{\theta})$ *is then a plug-in estimate that uses an estimate* $\hat{F}$ *in place of the unknown distribution* F*, and is defined by* $\mathrm{se}_{\hat{\mathrm{F}}}(\hat{\theta}^{*})$. The bootstrap algorithm which follows below is always applicable and is an easily computational way of obtaining a good approximation to the numerical value of $\mathrm{se}_{\hat{\mathrm{F}}}(\hat{\theta}^{*})$. Briefly stated, the bootstrap algorithm in this case is implemented by drawing many independent bootstrap samples, evaluating the corresponding bootstrap replication for each bootstrap sample, and estimating the standard error of $\hat{\theta}$ by the empirical standard error of the replications, and the result is referred to as *the bootstrap estimate of standard error*. More explicitly:

1. Select, by using computer methods, B independent bootstrap samples $\boldsymbol{x}^{*1}, \boldsymbol{x}^{*2}, \cdots, \boldsymbol{x}^{*B}$, each consisting of n data values (classical bootstrap) drawn with replacement from $\boldsymbol{x}$.
2. Evaluate the bootstrap replication of $\hat{\theta}$ corresponding to each bootstrap sample, i.e. $\hat{\theta}^{*}(b) = T_n(\boldsymbol{x}^{*b})$, $b = 1,2,\cdots,B.$
3. Approximate the standard error $\mathrm{se}_{\mathrm{F}}(\hat{\theta})$ by $\widehat{se}_B = \left\{(B-1)^{-1}\sum_{b=1}^{B}\left[\hat{\theta}^{*}(b)-\hat{\theta}^{*}(\cdot)\right]^2\right\}^{1/2}$, where $\hat{\theta}^{*}(\cdot) = \sum_{b=1}^{B}\hat{\theta}^{*}(b)/B.$

The limit of $\widehat{se}_B$ as B goes to infinity, is the bootstrap estimate of $\mathrm{se}_{\mathrm{F}}(\hat{\theta})$, i.e. $\mathrm{se}_{\hat{\mathrm{F}}}(\hat{\theta}^{*})$, which follows from the strong law of large numbers. How large B should be depends on factors such as affordability of computer time and the complexity of $\hat{\theta} = T_n(\boldsymbol{x})$. Efron and Tibshirani (1993) deduced a rule of thumb from examining the coefficient of variation of $\widehat{se}_B$, i.e. the ratio of $\widehat{se}_B$'s standard deviation to its expectation, that $B = 200$ replications are adequate for estimating the standard error. Larger values of B are needed for other estimation problems.

Example 2: Estimating the *bias* of $\hat{\theta} = T_n(\boldsymbol{x})$: The same algorithm is used, where the bias of $\hat{\theta}$ is defined by $\mathrm{bias}_F\left(\hat{\theta}\right) = E_F\left(\hat{\theta}\right) - t\left(F\right)$. *The bootstrap estimate of bias* is defined to be the estimate $\mathrm{bias}_{\hat{F}}\left(\hat{\theta}^{*}\right)$ we obtain by substituting $\hat{F}$ for F in the previous expression, i.e. $\mathrm{bias}_{\hat{F}}\left(\hat{\theta}^{*}\right) = E_{\hat{F}}\left(\hat{\theta}^{*}\right) - t\left(\hat{F}\right)$. Here $t\left(\hat{F}\right)$, the plug-in estimate of θ, may differ from $\hat{\theta} = T_n(\boldsymbol{x})$. Having followed the first and second steps in the above algorithm, the bootstrap expectation $E_{\hat{F}}\left(\hat{\theta}^{*}\right)$ is approximated by the average $\hat{\theta}^{*}(\cdot) = \sum_{b=1}^{B}\hat{\theta}^{*}(b)/B$. The bootstrap estimate of bias based on the B replications is then $\widehat{bias}_B = \hat{\theta}^{*}(\cdot) - t\left(\hat{F}\right)$. Sometimes $\widehat{se}_B$ and $\widehat{bias}_B$ can be calculated simultaneously from the same set of bootstrap replications, but it was found that more often the number of bootstrap replications B must be much larger to produce a trustworthy estimation of bias and an improved estimation method for bias have been suggested by Efron and Tibshirani (1993, chapter 10).

Example 3: Estimating *probabilities* of the form $P_F(a_1 \le \hat{\theta} \le a_2)$ where $\hat{\theta} = T_n(\boldsymbol{x})$: The bootstrap estimate of the probability is defined to be $P_{\hat{F}}(a_1 \le \hat{\theta}^{*} \le a_2)$ where $\hat{\theta}^{*} = T_n(\boldsymbol{x}^{*})$. As before, follow the first and second steps in the above algorithm. The bootstrap estimate of

$P_F(a_1 \leq \hat{\theta} \leq a_2)$ based on the B replications is $\hat{P}_B = B^{-1}\sum_{b=1}^{B} I\left(a_1 \leq \hat{\theta}_b^* \leq a_2\right)$, which is increasingly accurate for large values of n and B.

Example 4: Constructing a $100(1-\alpha)\%$ *confidence interval* for θ : Let $\hat{\theta} = T_n(\boldsymbol{x})$ be some (complex) estimator of θ. The so-called *bootstrap percentile* $100(1-\alpha)\%$ *confidence interval* for θ can be constructed as follows: Calculate, as before, bootstrap replications $\hat{\theta}^*(b)$, $b = 1,\cdots,B$, and then determine the order statistics $\hat{\theta}^*_{(1)} \leq \hat{\theta}^*_{(2)} \leq \ldots \leq \hat{\theta}^*_{(B)}$. The desired confidence interval is then $\left[\hat{\theta}^*_{(r)}, \hat{\theta}^*_{(s)}\right]$, where $r = \left[B\alpha/2\right]$ and $s = \left[B\left(1-\alpha/2\right)\right]$, and $\left[a\right]$ denotes the integerpart of a. Better bootstrap confidence intervals, like *bias-corrected percentile, accelerated bias-corrected percentile* and *percentile-t* are discussed in Chapters 12-14 of Efron and Tibshirani (1993).

REMARK

As far as the *smoothed bootstrap* is concerned, an algorithm for constructing a bootstrap sample from $\hat{F} = F_{n,c}\left(x\right) = n^{-1}\sum_{i=1}^{n} K\left(\left(x - X_i\right)/c\right)$, i.e. the kernel estimator of F, is the following: Generate $Y_1^*, Y_2^*, \cdots, Y_n^*$ which are independent random variables with common CDF F_n and also generate $Z_1, Z_2, \cdots, Z_n$ which are independent random variables with common CDF K. If these two samples are independent, then $X_j^* = Y_j^* + cZ_{j,}\ \ j = 1,2,\cdots,n,$ denotes a bootstrap sample from $F_{n,c}$.

THE NON-PARAMETRIC BOOTSTRAP APPLIED TO MORE COMPLEX DATA

Literature contains many non-parametric bootstrap techniques for analysing linear regression models, generalised linear models, non-linear models, survival models, time series models, and many more. The following example shows that the algorithms become slightly more complicated.

Example 5: Estimating the *sampling distribution* of the *estimated parameters* of a *regression model:* The probability structure of a regression model is usually expressed in the form $X_i = g(\boldsymbol{t}_i, \beta) + \varepsilon_i$, $i = 1,2,\cdots,n$, with $\boldsymbol{X}$ the response vector, g a known real-valued function of the covariates $\boldsymbol{t}$ and the parameters β. The errors ε_i are assumed to be a random sample from an unknown CDF F having expectation 0. Least-squares estimates for the parameters β are obtained from the definition $\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n}(X_i - g(\boldsymbol{t}_{\mathrm{i}}, \beta))^2$. The algorithm below provides a trustworthy approximation to the sampling distribution of $\hat{\beta}$, i.e. the bootstrap distribution of $\hat{\beta}^*$, from which accuracy and confidence measures can be deduced.

1. By using the original estimators $\hat{\beta}$, first calculate the centered residuals $\hat{\varepsilon}_i = X_i - g(\boldsymbol{t}_{\mathrm{i}}, \hat{\beta}) - \mathrm{n}^{-1}\sum_{i=1}^{n}(X_i - g(\boldsymbol{t}_{\mathrm{i}}, \hat{\beta}))$, $i = 1,\cdots,n$, from which $\varepsilon_1^*, \varepsilon_2^*, \cdots, \varepsilon_n^*$, an error bootstrap sample, is generated as before.
2. By using ε^* and $\hat{\beta}$, the first bootstrap sample $\boldsymbol{X}^*$ is created, with $X_i^* = g(\boldsymbol{t}_{\mathrm{i}}, \hat{\beta}) + \varepsilon_i^*$, $i = 1,2,\cdots,n$, and the first bootstrap replication of $\hat{\beta}$ is determined, i.e. the least-squares estimator $\hat{\beta}^*\left(1\right)$ from the definition $\hat{\beta}^*\left(1\right) = \arg\min_{\beta} \sum_{i=1}^{n}(X_i^* - g(\boldsymbol{t}_{\mathrm{i}}, \beta))^2$.

3. Repeat the previous steps B times independently to obtain $\hat{\beta}^*(1), \hat{\beta}^*(2), \cdots, \hat{\beta}^*(B)$, from which histograms, for example, can be constructed to approximate the sampling distribution of $\hat{\beta}$.

PRACTICAL IMPLEMENTATION

Several flexible computer programs are available for the implementation of resampling methods, for example S-Plus and Fortran. The free cloned version of S-Plus, namely R, of which the basic package can easily be downloaded and installed from the website *http://cran.r-project.org*, enables the application of resample methods free of cost in both UNIX and Windows environments. Several add-on bootstrap packages are available on this website, two of which are complementing the books of Efron and Tibshirani (1993), as well as Davison and Hinkley (1997). The bootstrap library for Davison and Hinkley can also be used, which is obtainable from the home page http://dmawww.epf1.ch/dvison.mosaic/BMA/ for UNIX users, or from a disk to be used with S-Plus for Windows, which accompanies Davison and Hinkley's book. The functions *boot, boot.array, print.boot, boot.ci* on the disk were found easy to use and time saving. Free S-Plus functions for Windows, complementing both the books mentioned above, are available at *http://lib.stat.cmu/edu/S/* under the names *bootstrap.funs* and *davison-hinkley,* and contains more than 50 bootstrap related functions as well as useful data files. Similar free add-on R functions are available on CRAN. Free help on learning S-Plus and/or R code is available at *http://www.math.montana.edu/stat/tutorials/S-Plus.html*.

REFERENCES

Biehler, R., Rach, W., & Winkelmann, B. (1988). *Computers and mathematics teaching: the German situation and reviews of international software*. Occasional paper #103. Institute für Didaktik der Mathematik. Bielefeld, FRG: University of Bielefeld.

Davison, A.C., & Hinkley, D.V. (1997). *Bootstrap methods and their application*. Cambridge University Press.

De Angelis, D., & Young, G.A. (1992). Smoothing the bootstrap. *International Statistical Review 60*, 45-56.

DiCiccio, T.J., & Romano, J.P. (1989). The automatic percentile method: Accurate confidence limits in parametric models. *Canadian Journal of Statistics, 17*, 155-169.

Dudewicz, E.J. (1976). *Introduction to statistics and probability*. Holt, Rinehart and Winston, New York.

Efron, B., & Tibshirani, R.J. (1993). *An introduction to the bootstrap*. Chapman & Hall, New York.

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

El-Nouty, C., & Guillou, A. (2000). On the smoothed bootstrap. *Journal of Statistical Planning and Inference, 83*, 203-220.

Hall, P. (1990). Performance of bootstrap balanced resampling in distribution function and quantile problems. *Prob. Th. Rel. Fields 85*, 239-267.

Rice, J.A. (1995). *Mathematical statistics and data analysis* (2nd edn.). Duxbury press, Belmont, California.

Shaughnessy, J.M. (1977). Misconceptions of probability: An experiment with a small-group, activity-based, model-building approach to introductory probability at the college level. *Educational Studies in Mathematics*, *8*, 285-316.

Swanepoel, J.W.H. (1986). A note on proving that the (modified) bootstrap works. *Community Statistics, 12*, 2059-2083.

Silverman, B.W., & Young, G.A. (1987). The bootstrap: to smooth or not to smooth? *Biometrika, 74*, 469-479.