On Teaching the Bootstrap

Engel, Joachim
University of Education
Department of Mathematics and Computer Sciences
Reuteallee 46
71634 Ludwigsburg, Germany
E-mail: engel@ph-ludwigsburg.de

While the bootstrap in its first two decades after its inception by Brad Efron revolutionized the methods of the expert data analysts, its vast pedagogical potential is still in the process of being discovered. The bootstrap is a prime example of synergy between technology and content that can naturally be taught within the paradigms of new pedagogy. It has the potential to make core concepts of inferential statistics like the sampling distribution more accessible to learners. After analyzing key ingredients of the bootstrap method from a teaching perspective, we discuss methods of teaching the bootstrap in an activity-based learning environment with and without didactical software, including potential pitfalls and fallacies.

1 Simulation and Resampling

With the availability of cheap and flexible computing power simulation based Monte-Carlo methods have gained increasing importance in probability and statistics. They are a very powerful tool for problem solving as well as an excellent device for visualization to support the learning process of students. Mathematically based on the law of large number and conceptually on a frequentist notion of probability virtually any probability problem can be solved approximatively through simulations. While only providing approximate solutions, simulation methods can be applied even in highly complex situations that maybe too intricate for an analytic treatment. For statistics and probability education simulations allow to test understanding, confront misconceptions, support valid intuitions about concepts of chance and probability and encourage an experimental and exploratory approach to learning. In simulations we substitute a real random situation by a different experiment, which is a model for the original but can easily be manipulated and analyzed. Computer-supported interactive simulation helps to build a simplified model, where irrelevant features are disregarded, and the phenomena is condensed in time and available to the students work. Formal mathematics is reduced to a minimum allowing students to explore underlying concepts and experiment with varying set-ups (Batanero et al., 2004).

Despite an acknowledged lack of empirically based studies (Mills, 2002) there is a broad consensus among statistics educators that new computing technologies offer exciting advantages over the lecture- and book-reading methods of instructing in enhancing student understanding of abstract concepts. Computer simulation methods allow students to be actively involved in producing and analyzing data and experiment with random samples from a population with known parameters for the purpose of exploring and clarifying difficult concepts.

While simulations in probabiity usually start with a given probability law from which to generate random data, the direction of inference in statistics is vice versa: the genuine statistics problem starts with data and aims at inferring about a model for the data or an hypothesis related to the law generating the data. Resampling methods have been around now quite some time. Almost 50 years ago Tukey (1958) introduced a universal method to estimate bias and approximate confidence intervals. Tukey's jackknife is based on previous ideas by Quenouille (1949) to compute estimates by leaving out some observations in order to use the omitted data for an evaluation of the estimate. Twenty years later

Efron (1979) generalized the jackknife to the bootstrap. The idea is to base inference on the empirical distribution or some other data-based estimate of the unavailable population. Today the bootstrap and other resampling methods are common tools of the professional statisticans, yet the methods are barely taught in introductory level courses. However, resampling methods can be made accessible on almost any level. Furthermore, the bootstrap frees us from the requirement to teach inferential statistics only for statistics for which simple formulas are available. For example, introductory statistics classes usually discuss various concepts of location parameters. When discussing statistical inference only the mean is considered because it can be dealt with mathematically. More robust alternatives like the median or the trimmed mean – robust and important concepts in data oriented exploratory teaching – are ignored because they are mathematically intractable for the novice. The bootstrap, however, makes it just as easy to do inference with mathematically more intricate sample statistics as with the classical mean.

Besides enabling even the introductory student to use modern methods of statistics and adding a flexible and versatile instrument to his or her data analysis toolkit, the bootstrap is also very instructive from a learner's perspective. The core concept of inferential statistics is the the idea of a sampling distribution. Resampling – like any other simulation method – allows the learner to experience how a statistic varies from sample to sample and – with increasing number of resamples – how an (empirical) sampling distribution evolves. Therefore, from a perspective of statistics education, the bootstrap is much more than a tool useful in some sophisticated situations. It is an instrument to visualize and explore basic concepts of statistical inference. Once students are familiar with simulation, especially simulated sampling from a known population, it is straightforward to introduce the idea of resampling. It is important to point out that now we are resampling from a sample while before we simulated sampling from a population. Yet, in both situations one obtains a distribution, either a simulated sampling distribution or a resampling distribution for a sample statistic. When applying simulation, the novice experiences how the values of the sample statistic vary from sample to sample and how the sampling distribution evolves over time with increasing number of sample or resample observations. Simulation and bootstrapping offer a way to teach statistics and probability by using the computer to gain direct experience and direct understanding trough graphics. While an underlying distribution is an abstract concept, even beginners use histograms to visualize the distriubution of a data set (Hesterberg, 1998).

2 An Activity Based Example

An elementary, yet instructive and not trivial example suitable to demonstrate many facets of important statistical concepts is the estimation of animal abundance based on the capture-recapture method: Ideas on how to implement this experiment in an activity-based learning environment can be found, e.g., in Scheaffer et al. (1996) and Engel (2000). In its paradigmatic version it is about estimating the number of fish in a lake. A number m of fish is caught, marked (the capture), released back into the lake and after a while a sample (the recapture) of n fish is taken. The number of marked fish in the recapture – a random quantity – be denoted by k. A reasonable estimate for the population size N in the lake is

$$\widehat{N} = \frac{m \cdot n}{k}.$$

Based on a one-time data collection the quality of this estimate is difficult to evaluate. A repeat of the experiment will most likely lead to a different estimate for the population size. In a probability-based simulation set-up we could repeat this experiment many times over to obtain an (approximate) sampling distribution for $\widehat{N}$. As a statistics problem the "lake" (or in a more abstract notion the theoretical distribution on the random variable $\widehat{N}$) is not available. All we have to rely on are the

data at hand, i.e. the sample or recapture of size n. These data – if drawn by some random mechanism – may well be taken as a good representation of the total fish population.

Implementing the bootstrap idea, we therefore consider samples (=resamples) drawn from the best approximation to the the population we may have: the sample.

1. Draw a sample (re-sample or bootstrap sample) of size n with replacement from the original sample and count the number $k^\star$ of marked elements in that resample.
2. Compute $\widehat{N}^\star = \frac{mn}{k^\star}$ (bootstrap estimator)
3. Repeat step 1 and 2 many times over to obtain the (empirical) bootstrap distribution of $\widehat{N}^\star$, represented e.g. as a histogram. This distribution – the bootstrap distribution – is the proxy to the unknown sampling distribution of $\widehat{N}$.

Figure 1 illustrates the procedure of the bootstrap to obtain the sampling distribution for a sample statistic $\widehat{\theta}$:

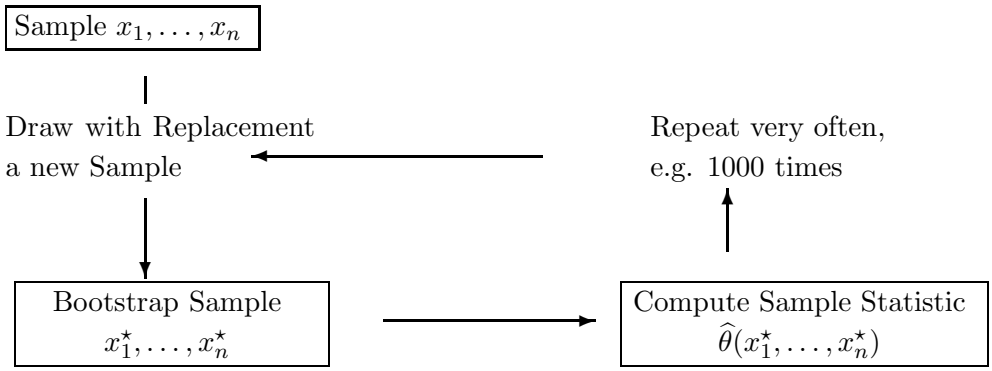

Figure 1: Illustration of the bootstrap procedure for obtaining an approximation to the sampling distribution for $\hat{\theta}(x_1, \ldots, x_n)$

In light of the unavailable true sample distribution inference now is based on the bootstrap distribution. Any quantity of interest, that is not available because of the unknown sample distribution of $\hat{\theta}$, can be computed from the just generated bootstrap distribution, e.g. bias, standard error or confidence intervals for $\hat{\theta}$.

Example 1 *In above simulation experiment we "marked"* $m = 80$ *fishes and caught a sample of* $n = 60$ *animals of which* $k = 13$ *had a marker leading to an estimate for the population size of* $80 \cdot 60/13 \approx 369$. *Relying on the available sample of size 60, we resampled repeatedly 500 times to obtain the bootstrap distribution of* $\widehat{N}$ *as displayed in Figure 2 obtaining a 95% confidence interval of* $[250; 600]$ *by cutting off 2.5% from both tails.* [1] *Furthermore, it is easy to compute the standard error as the standard deviation of the bootstrap distribution to obtain a value of 78.68. Also, an estimate for the bias of* $\widehat{N}$ *can be computed as difference between the average of the bootstrap distribution and the estimate of* $\widehat{N}$ *in the original sample:* $321.058 - 240 = 81.058$.[2]

It is instructive to consider how modern software allows implementing the bootstrap. Above histogram in Figure 2 has been created with the educational statistics package Fathom. In Fathom it

[1] Based on the hypergeometric distribution for the number k of marked elements it is here possible to compute the exact 95% confidence intervals of $[218; 480]$ while a probability simulation (based on the total population) resulted in a 95% confidence interval of [228; 485].

[2] The true population size in the simulation example is $N = 310$.

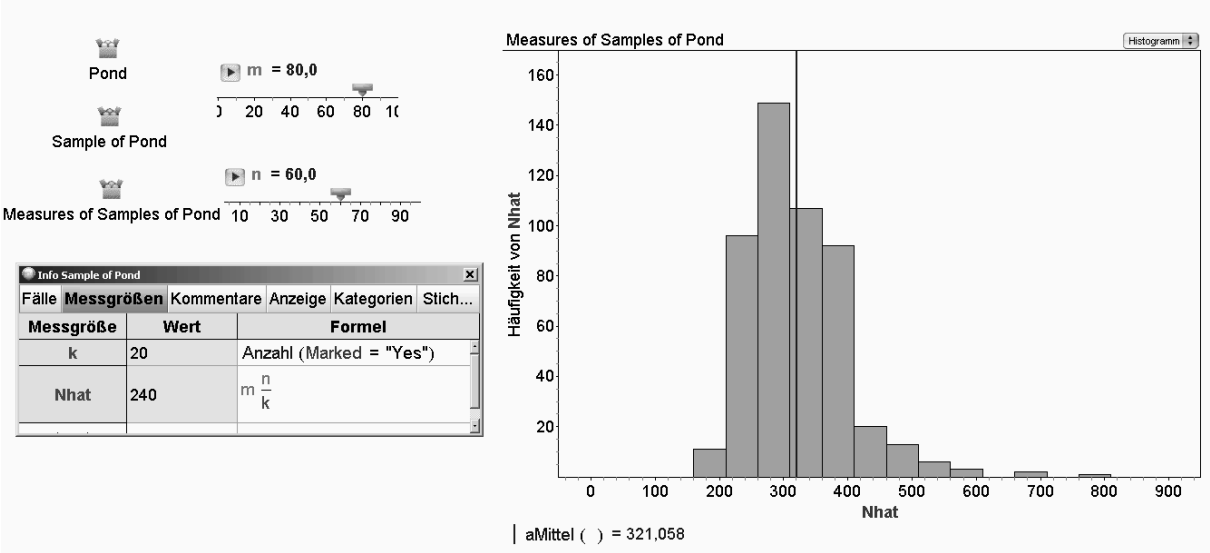

Figure 2: Bootstrap distribution of the estimated population size: Implementation in Fathom (left) and histogram (right)

is very natural to implement a resampling procedure like the bootstrap. After defining a collection or data set (the "pond") containing some pre-specified number of fishes of which m elements are marked, we draw the recapture sample. The bootstrap here is a simulation based on the 'sample of pond' and not on the original collection. The number k of marked elements in the 'sample of pond' as well as the quantity 'Nhat' is computed as what is called in Fathom a measure. Then 500 of these measures are collected to be represented in a histogram (see Figure 2).

The professional statistician prefers a more flexible and powerful software, in particular when considering a complex sample statistic. To implement the bootstrap in R for the Capture-Recapture bootstrap requires a few lines of code only. The following R commands produce a histogram similar to Figure 2.

```
CapRecapBoot<-function(n,m,B){
pond=append(rep(1,m),rep(0,(N-m)))    # Define pond
samplepond=sample(pond,n,replace=FALSE) # Sample or recapture
k=sum(samplepond)
Nhat=n*m/k # Estimator for  N
Nhatstar = matrix(0,B,1)  # for estimates based on resamples

for (b in 1:B){  # main loop for bootstrap
resampleindices = ceiling(n * runif(n))
 Nhatstar[b]= n*m/sum(samplepond[resampleindices])
}
hist(Nhatstar)
}
```

Note, that the distribution of $\widehat{N}$ is biased and skewed to the right. In fact, if $n + m < N$, there is a positive probability of having no fish at all in the recaptue implying an estimate of infinity. Therefore, to avoid an infinite bias of the population estimate, Seber (1973) suggests to estimate N by

$$\widehat{N}_1 = \frac{m \cdot (n+1)}{k+1} - 1.$$

3 Content Analysis: Plug-in and Monte Carlo

In a more symbolic notation, the bootstrap can be characterized as follows: Given an unknown distribution F, we are interested in some parameter $\theta = \theta(F)$ depending on the distribution F. Based on n observations $x_1, \ldots, x_n$ let $\widehat{\theta} = \widehat{\theta}(x_1, \ldots, x_n)$ denote an of estimate θ. Observe that $\widehat{\theta}$ is a random quantity whose distribution depends on the unknown F. To be precise, we denote this dependence by $\widehat{\theta}(x_1, \ldots, x_n|F)$. With F being unknown the distribution of the random variable $\widehat{\theta}$ is not available which is the source of the trouble. To resolve the situation, the bootstrap consists of two steps: a statistical part based on a plug-in idea and a numerical part based on a Monte-Carlo simulation:

1. Plug-In Step
   The unknown distribution $\widehat{\theta}(x_1, \ldots, x_n|F)$ is estimated by simply replacing F with the empirical distribution $\widehat{F_n}$ given as

   $$\widehat{F_n}(x) = \frac{1}{n}\#\{x_i|x_i \leq x\}.$$

   Therefore, the distribution of $\widehat{\theta}$ ist estimated by $\widehat{\theta}(x_1, \ldots, x_n|\widehat{F_n})$: the *Bootstrap distribution.*

2. Monte-Carlo Step
   With n original observations the sample space for the bootstrap replicates is finite but generally very large, consisting of n^n elements. This is the number of different resamples (with replacement) of a sample of size n. While this sample space has a simple Laplacean uniform distribution, the distribution of the sample statistics $\widehat{\theta}(x_1, \ldots, x_n)$ is – except in some very special situations, see below – very difficult if not impossible to compute analytically.

   To obtain the distribution of $\widehat{\theta}(x_1, \ldots, x_n|\widehat{F})$ we resort to simulations: We simulate by drawing a sample of size n from $\widehat{F}_n$, denoting it by $x_1^\star, \ldots, x_n^\star$ (the bootstrap sample), which is nothing but a resample from the original data. Then we compute the corresponding estimate $\widehat{\theta}(x_1^\star, \ldots, x_n^\star)$. Repeating this many times over and over again, say B times, results in an empirical approximation to the bootstrap distribution.

Remarks

1. Instead of the empirical distribution $\widehat{F}_n$ we could use as well any other estimate of F like a smoother kernel estimate resulting in the *smoothed bootstrap.* If we have reasons to believe that F belongs to a certain parametric family of distributions, a parametric estimate of F may be used leading to the *parametric bootstrap.* However, one of the great advantages of the bootstrap over classical methods of statistical inference is its applicability for modern robust and nonparametric statistical methods avoiding unrealistic assumptions.

2. There are situations where the (exact) bootstrap distribution can be computed analytically. Then the Monte Carlo Step is not necessary. Consider an odd number $n = 2m + 1$ of pairwise distinct data and look at the median as the quantity of interest. By increasing order of the data $x_{(1)} < x_{(2)} < \ldots < x_{(n)}$ we have that $\widehat{\theta} = x_{(m+1)}$, i.e. we estimate the population median by the sample median. Each resample $x_1^\star, \ldots, x_n^\star$ has as median $\widehat{\theta}^\star$ one of the resampled data (n still being odd), hence one of the original data points. Obviously, we have $\widehat{\theta}^\star \leq x_{(i)}$ exactly when at least $(k + 1)$-times one of the i data has been drawn that is less or equal $x_{(i)}$, i.e. we have for $i = 1, \ldots, n$

   $$P(\widehat{\theta}^\star \leq x_{(i)}|X_1 = x_1, \ldots, X_n = x_n) = \sum_{k=m+1}^{n} \binom{n}{k} \left(\frac{i}{n}\right)^k \left(1 - \frac{i}{n}\right)^{n-k}$$

   This example illustrates that the bootstrap distribution can be computed without simulations.

3. It is important to note that we are dealing with two distinct approximations: the (exact) bootstrap distribution is a proxy for the distribution of the original sample statistic. The quality of this approximation depends on the original sample size n which is usually very costly to increase (collecting more data). A diferent approximation is the empirical bootstrap distribution approaching the exact bootstrap distribution. This spproximation is governed by the number of replicates B. Increasing the number of bootstrap resamples is very cheap in times of plunging hardware prizes and ever more powerful software.

4 When does Bootstrap work: An Issue for Teaching?

The bootstrap may be an instructive, very useful and intuitively reasonable algorithm, but does the method really produce reasonable results? Some of the enthusiam about the bootstrap method was founded on the misunderstanding that mathematics (as the basis in classical inference) is replaced by mere computing power.

Sound statistical inference is based on the sampling distribution of $\widehat{\theta}(x_1, \ldots, x_n|F)$, but with the bootstrap we infer from $\widehat{\theta}(x_1, \ldots, x_n|\widehat{F_n})$. To achieve valid conclusions requires that these two distributions are close to each other, at least in some asymptotic sense. Hence, we need a continuity argument to guarantee that the bootstrap is more then a "stab in the dark" (Young 1994).

To formalize, we have to show that these two distributions – appropriately normalized – converge to the same limiting distribution. For broad classes of situations this is possible to prove mathematically, but requires highly advanced methods whose foundation is the theory of empirical processes and their convergence (see, e.g. Hall 1992, Mammen 1992).

To illustrate the required convergence, we resort to the previous example of the capture-recpature estimate. The distribution of marked elements in the resample is a typical case of a hypergeometric distribution

$$\begin{aligned} k &\sim H(N, m, n), \quad \text{i.e.} \\ P(k = i) &= \frac{\binom{m}{i}\binom{N-m}{n-i}}{\binom{N}{n}} \end{aligned}$$

while for for the Bootstrap distribution of marked elements in the resample we have (drawing with n elements from the resample, where k out of n elements are marked)

$$\begin{aligned} k^\star &\sim B(n, p) \quad \text{i.e.} \\ P(k^\star = i) &= \binom{n}{i} p(1-p)^{n-i} \quad \text{with } p = \frac{k}{n} \end{aligned}$$

Now it is straightforward that $|P(k = i) - P(k^\star = i)| \to 0$ as $N, m, n \to \infty$ with $m/N \to p$. As consequence the distributions of $\widehat{N} = \frac{mn}{k}$ and $\widehat{N}^\star = \frac{mn}{k^\star}$ convergerge to each other.

It is instructive to consider elementary examples where the bootstrap methods fails. Consider the "Frankfurt Taxi Problem" (A. Engel, 1987): in Frankfurt taxis are presumably numbered from 1 to N. After arriving at the train station you observe n taxi cabs numbered by $x_1, \ldots, x_n$. The maximum likelihood estimate for the total number of taxis N is $\widehat{\theta}_{\text{MLE}} = \max(x_1, \ldots, x_n)$. It is straightforward to show that the distribution of $n(\theta - \widehat{\theta}_{\text{MLE}})$ is exponential with parameter $1/\theta$ which implies in particular that – as for any continuous distribution – the value 0 is assumed with zero probability. The bootstrap distribution for $n(\theta - \widehat{\theta})$ is the distribution of $n(\widehat{\theta} - \max\{x_1^\star, \ldots, x_n^\star\})$, where $\widehat{\theta} = x_i$ for some $i \in \{1, \ldots .n\}$. We obtain a value of 0 with that probability for which the element x_i is being resampled. However, it is well known (compare the "rencontre problem") that this probability converges towards $1 - 1/e$, hence not towards the value 0 provided by the exponential distribution.

5 Summary

It is well known that the meaning of a sampling distribution for statistical inference is very hard for students to grasp. Computer simulation lets students gain experience with and intuition for these concepts. The bootstrap provides a prominent opportunity to enhance that learning in view of genuine statistical reasoning, i.e. in situations where we have data but do not know the underlying distribution.

For the mathematical statistician the bootstrap is a highly advanced procedure whose consistency is based on the convergence of empirical processes, for users of statistics it is mainly a simulation method and an algorithm. For a sound understanding and appropriate handling any user should be aware that random and chance enter at two distinct points: in the Plug-In step by considering the distribution of the sampling statistics under $\widehat{F_n}$ resulting in the bootstrap distribution and in the Monte-Carlo step by obtaining an empirical approximation to the exact bootstrap distribution. There are situations that allow to compute the exact bootstrap distribution, i.e. then the Monte Carlo step is not needed. In the vast majority of situations the bootstrap distribution is not tractable analytically. Then simulation from the empirical distribution $\widehat{F}_n$ yields an empirical approximation to the bootstrap distribution. This approximation can be made arbitrarily close by increasing the bootstrap sample size, given the availability of sufficient computing power. In contrast, the asymptotic equivalence of the bootstrap distribution and original distribution of the sampling statistic is far from being trivial and is based on convergence results for empirical processes.

A great deal of research efforts on learning and instruction over the last decades focuses on how to take advantage of modern technology to support learning. The availability of modern technology also influences the content of teaching considered valuable and worthwhile. Moore (1997) speaks of synergy effects between technology, content and new pedagogy. Working with technology may influence qualitatively the thinking of learners about mathematics. New content reflects the computer-intensive practice of modern statistics. The bootstrap is a prime example for synergy between technology, content and new insights into the learning process. The method is based on a conceptually simple idea that is generally very useful and instructive. Without available cheap computing power the bootstrap is not feasible. As for any simulations in probability and statistics, the bootstrap can be implemented in an activity-based, exploratory and experimental learning environment. While the method in its first three decades has been mainly a very useful method for the expert data analyst, time has come to take advantage of its great potential to enhance learning of concepts in inferential statistics.

REFERENCES

Batanero, C. ; Biehler, R.; Engel, J.; Maxara, C. & Vogel, M. (2005): Using Simulation to Bridge Teachers Content and Pedagogical Knowledge in Probability. In: ICMI-Study 1, Online: `http://stwww.weizmann.ac.il/G-math/ICMI/log_in.html`

Efron, B. (1982): Bootstrap mehtods: another look at the jackknife. *Annals of Statistics*, 7, 1-26.

Engel, A. (1987): Stochastik. Klett: Stuttgart.

Engel, J. (2000): Markieren - Einfangen - Schtzen: Wie viele wilde Tiere? *Stochastik in der Schule*, 2, 17 - 24.

Hall, P. ((1992): *The Bootstrap and Edgeworth Expansion.* New York: Springer.

Hesterberg, T. (1998): Simulation and Bootstrapping for Teaching Statistics, in: American Statistical Association: *Proceedings of the Section on Statistical Education*, 44 - 52.

Mammen, E. (1992): *When does bootstrap work? Asymptotic results and simulations* Heidelberg: Springer

Mills, J. (2002): Using Computer Simulation Matheods to Teach Statistics: A Review of the Literature. *J. Statistics Education*, Vol. 10 (1). Online: http://www.amstat.org/publications/jse/

Moore, D. (1997): New pedagogy and new content: the case of statistics. *International Statistical Review* **65**, 123-166.

Seber, G. A. (1973): The Estimation of Animal Abundance and related parameters. London: Griffin.

Young, A.. (1994): More than a stab in the dark? *Statistical Science*, 9: 382-415