ON TEACHING BOOTSTRAP CONFIDENCE INTERVALS

Joachim Engel
Ludwigsburg University of Education, Germany
engel@ph-ludwigsburg.de

The basis for most inferential procedures is the idea of a sampling distribution. Computer simulation lets students gain experience with and intuition for this concept. The bootstrap can reinforce that learning. While today the bootstrap belongs to the toolkit of any professional statistician, its vast pedagogical potential is still in the process of being discovered. We discuss several bootstrap methods to compute confidence intervals from a teaching perspective.

INTRODUCTION

Confidence intervals are a tough concept for an introductory statistics course. The frequentist interpretation states that (100-α)% confidence interval characterizes a process for generating intervals with the following property: when the process is repeated several times, (100-α)% of the generated intervals contain the true but unknown parameter. In most introductory courses the usual treatment of confidence intervals often begins with a sampling distribution of a statistic, like the sample mean. From the sampling distribution, one argues that 95% of the time the sample mean $\overline{X}$ will fall between $\mu \pm 1.96\,\mathrm{SE}$ where SE is the standard error of $\overline{X}$ as the estimator of the population mean μ. Through some simple algebraic manipulations, we obtain the random interval $\overline{X} \pm 1.96\,\mathrm{SE}$.

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 approximately through simulations. While only providing rough solutions, simulation methods can be applied even in highly complex situations that may be too intricate for an analytical treatment. Thus simulations in statistics allow for testing the understanding, for confronting misconceptions, and for supporting valid intuitions in an experimental and exploratory approach to learning. In simulations we substitute a real random situation by a hypothetical experiment, which is a model for the original one but can easily be manipulated and analyzed. Computer-supported interactive simulation helps to build simplified models, where irrelevant features are disregarded, and the phenomena are condensed in time and akin to students' experience. Formal mathematics is reduced to a minimum, allowing students to explore underlying concepts and experiment with varying set-ups (Batanero, Biehler, Engel, Maxara & Vogel, 2004).

Figure 1 shows 50 simulated 90% confidence intervals for a proportion. Based on a sample of B(1, p) distributed data $x_1, \ldots, x_n$, we computed 50 confidence intervals. The population parameter p is represented through the vertical line. As we can see, in 5 of the 50 simulations the calculated interval misses the true parameter. This is an example of a "probabilistic" simulation, with samples generated from a known underlying distribution.

RESAMPLING AND THE BOOTSTRAP

Contrary to above example, in statistical practice as well as in data-oriented introductory statistics courses, the underlying distribution is unknown and all that is available are the data. If the population, represented through a probability distribution and its parameters is no longer available, what can we do? The basic idea is simple: if we can't sample from the population, we sample from the best approximation to this population we got: the observed data itself. The bootstrap idea is to take (re-)samples $x_1^*,\ldots,x_n^*$, drawn with replacement from the original sample $x_1, \ldots, x_n$. In other words: If we don't know the distribution function F our original sample was drawn from, then we generate new samples from the empirical distribution function $\hat{F}_n$.

In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS8, July, 2010), Ljubljana, Slovenia*. Voorburg, The Netherlands: International Statistical Institute. www.stat.auckland.ac.nz/~iase/publications.php [© 2010 ISI/IASE]

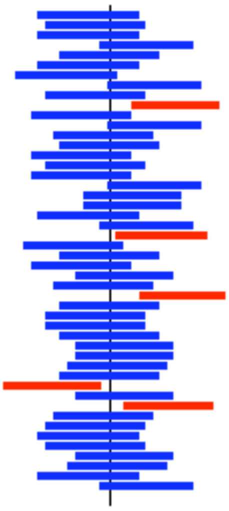


Figure 1. 50 simulated confidence intervals for a proportion, true value represented by straight line

The bootstrap is a prime example of synergy between technology and content that can be taught naturally within the paradigms of modern theories of learning. It has the potential to make central ideas of inferential statistics like the concept of a confidence interval more accessible to learners than the classical approach based on theoretical distributions. Furthermore, it provides a way to estimate parameters, approximate a sampling distribution or derive confidence intervals when we have data but do not know the underlying distribution. The bootstrap also frees us from the requirement to teach inference only for statistics for which simple formulas are available or accessible to our students.

Once students are familiar with simulation, especially with simulated sampling from a known population, it is fairly easy to introduce the idea of resampling. It is important to differentiate between simulated sampling from a population and resampling from a sample. These are different operations. However, what one gets out is a distribution, either a simulated sampling distribution or a resampling distribution, of possible values for a sample statistic.

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

Besides enabling 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 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 just a tool useful in sophisticated situations. It is an instrument to visualize and explore basic concepts of statistical inference. 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 for teaching statistics and probability by using the computer in order to gain direct experience and direct understanding through graphics. While the concept of an an underlying distribution is abstract, even beginners use histograms to visualize the distribution of a data set (Hesterberg, 1998).

BOOTSTRAP CONFIDENCE INTERVALS

Once students are introduced to simulation and resampling, it is a straightforward step to estimate quantities of interest of the not available true sampling distribution by the corresponding quantity of the bootstrap distribution, e.g., standard error, bias, standard deviation. Concerning confidence intervals, the situation is a bit more complicated. A variety of different concepts have been proposed (see, e.g. Davison and Hinkeley 1997 or DiCiccio and Efron, 1996) for producing approximate confidence intervals. With q_c^* denoting the *c* percentile of the bootstrap distribution three rather elementary concepts are:

- *bootstrap percentile interval*: In order to construct a confidence interval for a parameter θ we need to know how $\hat{\theta}$ as an estimator of θ varies in repeated sampling from the population. Since all the information we have about the population is contained in the sample, bootstrap methods treat the sample as if it were the population. The (100- α)% bootstrap percentile interval contains the middle (100- α)% of the bootstrap distribution taking the α/2 and (100-α/2) percentile of the bootstrap distribution as interval endpoints: $[q^*_{\alpha/2}, q^*_{100-\alpha/2}]$.
- *basic bootstrap interval*: To create a 100- α confidence interval for a parameter θ based on a sample estimate $\hat{\theta}$, we determine the distance that we plausible expect $\hat{\theta}$ to fall from θ at the α% level. If we knew critical values c_1 and c_2 such that $P(c_1 \le \hat{\theta} - \theta \le c_2) = 1 - \alpha/100,$ then we could rearrange the inequalities as follows $P(\hat{\theta} - c_2 \le \theta \le \hat{\theta} - c_1) = 1 - \alpha/100$ to produce a (100-α)% confidence interval for θ. As we don't know the distribution of $\hat{\theta}$ (or of $\hat{\theta} - \theta$) and their percentiles, we take the corresponding percentiles of the bootstrap distribution. Under the assumption that the percentiles of the sampling distribution and the bootstrap distribution are very close, we obtain
$$(100-\alpha)\% = P(q^*_{\alpha/2} - \hat{\theta} \le \hat{\theta}^* - \hat{\theta} \le q^*_{100-\alpha/2} - \hat{\theta})$$
$$\approx P(q^*_{\alpha/2} - \hat{\theta} \le \hat{\theta} - \theta \le q^*_{100-\alpha/2} - \hat{\theta}) = P(2\hat{\theta} - q^*_{100-\alpha/2} \le \theta \le 2\hat{\theta} - q^*_{\alpha/2})$$
resulting in $[2\hat{\theta} - q^*_{100-\alpha/2}, 2\hat{\theta} - q^*_{\alpha/2}]$
- *bootstrap-t interval*: similar to the standard normal case the interval is constructed according to $\hat{\theta} \pm q^*_\alpha \cdot SE_{Boot}$ where $\mathrm{SE}_{\mathrm{Boot}}$ is the standard error of the estimator $\hat{\theta}$, estimated from the bootstrap distribution.

More sophisticated approaches to obtain bootstrap confidence interval include the bias-corrected and accelerated method (*BCa intervals*), the approximate bootstrap confidence intervals (*ABC intervals*) and *bootstrap tilting*. All these methods are asymptotically superior to those mentioned above and are implemented in the bootstrap package of the software R. Detailed description and precise comparison of the various concepts can be found in the literature, e.g., DiCiccio and Efron (1996) or Efron.

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. Let the number of marked fish in the recapture–a random quantity–be denoted by *k*. Then a reasonable estimate for the population size N in the lake is

$$\hat{N} = \frac{m \cdot n}{k}$$

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 $\hat{N}$. As a statistics problem the "lake" (or in a more abstract notion the theoretical distribution on the random variable $\hat{N}$) 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 population we may have: the sample.

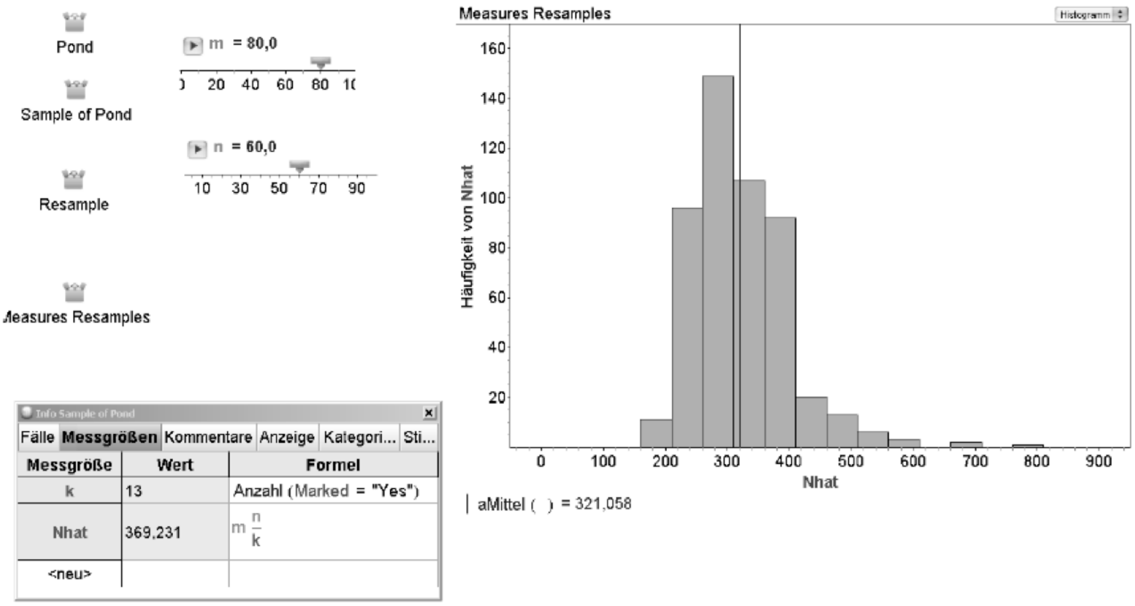
1. Draw a sample (re-sample or bootstrap sample) of size n with replacement from the original sample and count the number k^* of marked elements in that resample.

2. Compute $\hat{N}^* = \dfrac{m \cdot n}{k^*}$ (bootstrap estimator)

3. Repeat step 1 and 2 many times over to obtain the (empirical) bootstrap distribution of $\hat{N}^*$, represented e.g. as a histogram. This distribution – the bootstrap distribution – is the proxy to the unknown sampling distribution of $\hat{N}$.

Example 1: *In the simulation experiment described above 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· 60/13 ≈ 369. Relying on the available sample of size 60, we resampled repeatedly 500 times to obtain the bootstrap distribution of* $\hat{N}$ *as displayed in Figure 2 obtaining a 95% percentile confidence interval of [240; 600] by cutting off 2.5% from both tails.*[1] *For the basic bootstrap interval we obtain [138, 498] while the bootstrap-t interval is [211, 522] 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* $\hat{N}$ *can be computed as difference between the average of the bootstrap distribution and the estimate of* $\hat{N}$ *in the original sample: 321.058 − 240 = 81.058*[2].

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



Messgröße	Wert	Formel
k	13	Anzahl (Marked = "Yes")
Nhat	369,231	$m \frac{n}{k}$
<neu>		

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

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? Part of the enthusiasm 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 $\hat{\theta}(x_1,...,x_n \mid F)$, but with the bootstrap we infer from $\hat{\theta}(x_1,...,x_n \mid \hat{F}_n)$. 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 can be proven mathematically, but requires highly advanced methods whose foundation is the theory of empirical processes and their convergence (see, e.g., Hall, 1992; Mammen, 1992).

In order to exhibit the required convergence, we resort to the previous example of the capture-recapture estimate. The distribution of marked elements in the resample is a typical case of a hypergeometric distribution

$k \sim H(N,m,n)$, i.e.

$$P(k=i) = \frac{\binom{m}{i}\binom{N-m}{n-i}}{\binom{N}{n}}$$

while 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)

$k^* \sim B(n,p)$, i.e.

$$P(k^*=i) = \binom{n}{i} p(1-p)^{n-i} \text{ with } p = \frac{k}{n}.$$

Now it is straightforward that both distributions, appropriately normalized, converge to the same limit as $N,m,n \to \infty$ with $\frac{m}{N} \to p$.

It is instructive to consider elementary examples where the bootstrap method 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 $x_1,...,x_n$. The maximum likelihood estimate for the total number of taxis N is $\hat{\theta}_{\text{MLE}} = \max(x_1,...,x_n)$. It is straightforward to show that the distribution of $n\left(\theta - \hat{\theta}_{\text{MLE}}\right)$ 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 $n\left(\theta - \hat{\theta}_{\text{MLE}}\right)$ is the distribution of $n\left(\hat{\theta} - \max\{x_1^*,...,x_n^*\}\right)$, where $\hat{\theta} = x_i$ for some $i \in \{1,...,n\}$. 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 concept of confidence intervals is 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 $\hat{F}_n$ in place of the distribution F 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 $\hat{F}_n$ 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.

NOTES

(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.

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.

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

DiCiccio, T. J., & Efron, B. (1996). Bootstrap Confidence Intervals (with discussion). *Statistical Science*, *11*(3), 189-228.

Efron, B., & Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. London: Chapman & Hall

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

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

Engel, J. (2007). On Teaching the Bootstrap. *Bulletin of the International Statistical Institute 56th Session*, Lisbon.

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.

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

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