Link to Matlab inbuilt function for bootstrap.

https://in.mathworks.com/help/stats/bootstrp.html

Efron's bootstrap

The bootstrap was introduced by Brad Efron in the late 1970s. It is a computer-intensive method for approximating the sampling distribution of any statistic derived from a random sample. Here **Dennis Boos** and **Leonard Stefanski** give simple examples to show how the bootstrap is used and help to explain its enormous success as a tool of statistical inference.

Bootstrap basics

A fundamental problem in statistics is assessing the variability of an estimate derived from sample data. Consider, for example, a simple survey in which a newspaper with a circulation of 300 000 (the *population*) randomly samples 100 of its subscribers (the *sample*) and asks their preference as to whether front-page stories should continue on the second page or on the back page of the section. Suppose that in the sample of 100 readers 64% favoured the back page. If this study were repeated with a new random sample of 100 readers, then the results would be unlikely to be 64% again, but would probably be something else, say 59%. And if the study were repeated over and over, the results would be a large set of percentages, say {64, 59, 65, 70, 52, ...}. This hypothetical set of possible study results represents the *sampling distribution* of the sample proportion statistic. With it one can assess the variability in the real-sample estimate (e.g., attach a margin of error to it, say 64% ± 9%), and rigorously address questions such as whether more than half the readers prefer stories to continue on the back page.

The catch is, of course, that it is impractical to repeat studies, and thus the set of possible percentages described above is never more than hypothetical. The solution to this dilemma, before the widespread availability of fast computing, was to derive the sampling distribution mathematically. This is easy to do for simple estimates such as the sample proportion, but not so easy for more complicated statistics.

Fast computing opened a new door to the problem of determining the sampling distribution of a statistic. On the other side of that door was Efron's bootstrap, or what is now known simply as the bootstrap. In broad strokes, the bootstrap substitutes computing power for mathematical prowess in determining the sampling distribution of a statistic.

In practice, the bootstrap is a computer-based technique that mimics the core concept of random sampling from a set of numbers and thereby estimates the sampling distribution of virtually any statistic computed from the sample. The only way it differs from the hypothetical resampling described above is that the repeated samples are not drawn from the population, but rather from the sample itself because the population is not accessible.

Examples

To illustrate these ideas we use two simple examples where the statistics are the sample mean and median. Consider the data set in Table 1 of $n = 25$ adult male yearly incomes (in thousands of dollars) collected from a fictitious county in North Carolina.

The sample mean

The sample mean of the Table 1 data is $\bar{Y} = 47.76$. Statistical theory tells us that if these values were independently drawn from a population of incomes having mean μ and variance σ^2, then the sampling distribution of $\bar{Y}$ has mean μ, variance σ^2/n (here $n = 25$), and standard deviation $\sigma/\sqrt{n}$.

Table 1. Random sample of 25 yearly incomes in thousands of dollars (ordered from lowest to highest)

1	4	6	12	13	14	18	19	20	22	23	24	26
31	34	37	46	47	56	61	63	65	70	97	385	

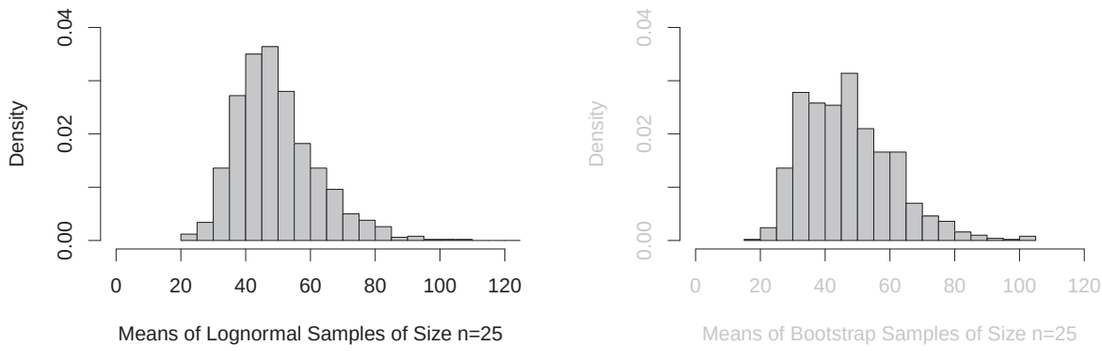

Figure 1. (Left) Histogram of 1000 sample means from repeated sampling of a theoretical lognormal population. (Right) Histogram of 1000 bootstrap sample means from randomly sampling with replacement from Table 1 data

The sampling distribution of a statistic computed from a random sample is the distribution of the statistic in repeated sampling from that population. Usually we do not know the population and cannot repeatedly sample, and thus we estimate μ with $\bar{Y}$ and also estimate the sampling standard deviation of $\bar{Y}$ (often called the *standard error*) by $s_{n-1}/\sqrt{n}$, where $s^2_{n-1} = (n-1)^{-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$ is the unbiased version of the sample variance. Statistical inference proceeds by relying on the fact that $\bar{Y}$ is approximately normally distributed due to the *central limit theorem*.

So that we know what the bootstrap should be estimating, we generated the data in Table 1 as $Y_i = 30\exp(Z_i)$ (×$1000), $i = 1, \ldots, 25$, where $Z_1, \ldots, Z_{25}$ are independently distributed standard normal random variables. Thus our fictitious sample is known to come from a lognormal population or distribution. Since we know the population distribution, we can also generate the true sampling distribution of $\bar{Y}$ by creating independent random samples in the same manner, and then computing $\bar{Y}$ for each one. We did this for 1000 random samples and plotted a histogram of the $\bar{Y}$ values in the left panel of Figure 1.

The bootstrap can be used to *approximate* the sampling distribution of $\bar{Y}$ when we do not know the population from which the sample was obtained (always the case with real data). The *nonparametric bootstrap* proceeds by treating the data in Table 1 as a population and drawing random samples from it. A bootstrap random sample (also called a *resample*) is drawn from the Table 1 pseudo-population by randomly choosing 25 values with replacement from the values in Table 1. Table 2 displays two such samples.

Note that repeated values of the original data appear within each resample because the sampling is *with replacement* (as opposed to *without replacement*). The only sample of size $n = 25$ that could be drawn without replacement is the original sample itself. The right panel in Figure 1 is a histogram of the 1000 sample means computed from 1000 resamples. It is the bootstrap estimate of the distribution in the left panel. Remember that we have the left panel in this case only because we generated the sample from a known probability distribution. In any real application we cannot produce the left panel, but the bootstrap can always produce the right panel. The two panels are similar, but there are differences resulting from the bootstrap step that uses the sample as if it were the population.

An important use of the bootstrap is calculation of the *standard error* of an estimate (the essential component of the margin of error associated with a statistical estimate). For our toy example, the bootstrap standard error of the mean estimate, 47.76, is

$$\left\{\frac{1}{1000-1}\sum_{i=1}^{1000}(\bar{Y}_i - \bar{\bar{Y}})^2\right\}^{1/2} = 13.8$$

In this case we can also use the theoretically derived formula to get the non-bootstrap standard error estimate $s_{n-1}/\sqrt{n} = 14.8$ for the Table 1 data. The difference between the two estimated standard errors (13.8 *versus* 14.8) has two components. The random component is due to the fact that the bootstrap estimate is based on 1000 resamples. Had we used a much larger number of resamples, then the bootstrap standard error would approximate $s_n/\sqrt{n} = 14.5$. The second component is due to the difference in the denominators between s_n and s_{n-1}. These are relatively minor discrepancies, and most analysts are usually willing to accept a small amount of variation in bootstrap standard errors due to the Monte Carlo simulation, that is, using 1000 resamples rather than say 1 million resamples. (And of course the fact that means from even 1000 resamples are calculated implies the bootstrap's practical need for a computer. Happily, it was developed just as computing power became widely available.)

Bootstrapping needs computing power. Happily it was devised just as computers became common

In some situations, we might feel comfortable making a guess at the type of distribution that the data came from, that is, the basic shape of the underlying population. For example, the data in Table 1 is actually from a normal distribution and then exponentiated to get lognormal data. Another way to do bootstrap sampling is to estimate the parameters of the assumed distribution and then generate bootstrap samples from the estimated population. This is called *parametric bootstrapping*, and is best used when the distribution type is reasonably well known.

Table 2. Bootstrap resamples from Table 1

Sample 1	1	4	4	6	18	22	22	23	23	23	24	26	31
	37	46	47	47	56	56	61	61	63	65	65	65	
Sample 2	1	4	6	13	14	14	18	19	22	23	23	23	24
	26	26	37	46	46	47	47	63	63	70	70	97	

The sample median

A histogram of the Table 1 data (not displayed) reveals that it is quite skewed to the right. This skewness is also clear from the fact that the sample mean 47.8 is much larger than the sample median, 26. In situations with such skewness it is typical to use the median to measure central tendency instead of the mean. Not

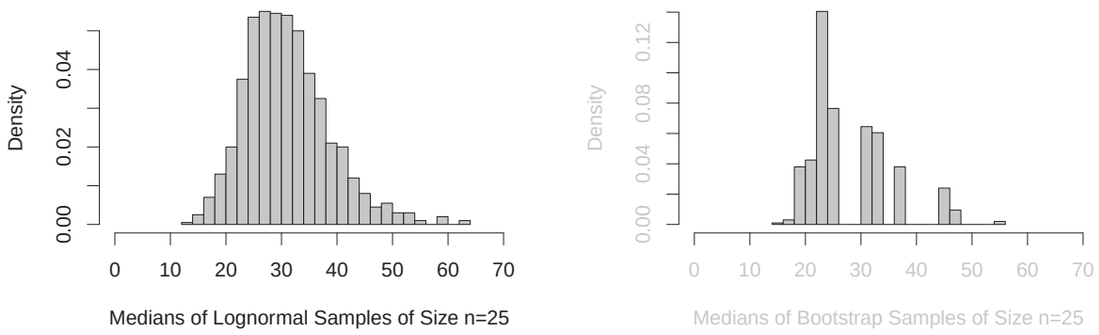

Figure 2. (Left) Histogram of 1000 sample medians from repeated sampling of a theoretical lognormal population. (Right) Histogram of 1000 bootstrap sample medians from Table 1 data

only is the median more representative of typical data values, but the sampling distribution standard deviation is much smaller for the sample median than for the sample mean (small is good for standard deviations of estimators!). Thus, for example, the US Census Bureau routinely uses medians to summarize income data.

Unfortunately the sampling distribution of the sample median is difficult to analyse theoretically. In fact, there is no simple expression for the standard deviation of the sample median like the expression $\sigma/\sqrt{n}$ for the sample mean. We can of course study the distribution by Monte Carlo sampling from a true population when it is known.

The left panel of Figure 2 gives a histogram of sample median values from the same 1000 lognormal samples as in the left panel of Figure 1. This histogram of medians approximates the true sampling distribution of the sample median. However, in real life we only know the sample, not the population. Thus the right panel of Figure 2 gives the histogram of 1000 sample medians computed from the same resamples as used in the right panel of Figure 1.

Note that the vertical scales are different in Figure 2. Because of the discreteness of the bootstrap pseudo-population and the nature of the median, the estimated sampling distribution is very discrete, with most of the sample medians concentrated on the Table 1 central values 22, 23, 24, 26, 31 and 34. For most purposes this discreteness is not a problem. Comparing Figures 1 and 2 visually suggests that the bootstrap distribution for the sample mean is a better estimate of the true sampling distribution of the sample mean than it is for the sample median. This reflects the fact that the sampling distribution of the median is more difficult to estimate than the sampling distribution of the mean. However, the bootstrap still estimates the sampling distribution well enough, and in particular provides a valid standard error estimate for the median, whereas the best-known other computationally-based method for estimating standard errors, the jackknife, does not.

Using the 1000 bootstrap medians depicted in the right panel of Figure 2, the bootstrap standard error (of the median estimate 26) is

$$\left\{\frac{1}{1000-1}\sum_{i=1}^{1000}(\widehat{M}_i-\overline{\widehat{M}})^2\right\}^{1/2}=7.3$$

Comparing this bootstrap standard error, 7.3, to that for the sample mean, 13.8, empirically supports the claim that the median is a less variable statistic than the mean for skewed data.

The 1000 bootstrap median values are commonly used for other purposes as well. The plots in Figure 2 suggest that the sampling distribution of the median is mildly skewed to the right. (In large samples the sampling distribution approximates a normal distribution.) Thus, we might be interested in the bias in the sample median, that is, the difference between the mean of the sampling distribution and the true population median. The bootstrap estimate of that bias is $(1000)^{-1}\sum_{i=1}^{1000}\widehat{M}_i - 26 = 28.4 - 26 = 2.4$, leading to the bootstrap bias-adjusted median estimate $26 - 2.4 = 23.6$.

The only requirement for a bootstrap is a computer program and a method to draw resamples

Another important statistical technique amenable to the bootstrap is confidence interval construction. The simplest bootstrap approach to confidence intervals is first to order the 1000 bootstrap medians displayed in the right panel of Figure 2, say $\widehat{M}_{(1)} \le \widehat{M}_{(2)} \ldots \le \widehat{M}_{(1000)}$. Then $(\widehat{M}_{(25)}, \widehat{M}_{(975)}) = (19, 46)$ is called the 95% bootstrap *percentile interval*. In this case, an exact nonparametric confidence interval for the median is available, given by (19, 47) with exact coverage probability 0.957. Efron[1] pointed out the close similarity between the bootstrap percentile interval and this nonparametric confidence interval.

Conclusion

The power of the bootstrap lies in the fact that the method applies to (almost) any estimator, no matter how complicated. The only requirement is a computer program to calculate the estimator from a sample and a method to draw resamples. We have described only the case of simple random sampling. However, the bootstrap method applies to any type of probability-based data collection, provided that it can be imitated via a computer program to generate resamples that relate statistically to the real sample in the same way that the real sample relates to the population from which it was selected. For example, economic data is often in the form of time series where all the sample data are correlated. A parametric bootstrap would assume a specific model such as a normal autoregressive process. After estimating the unknown parameters of the model, many independent bootstrap time series would be generated from the estimated autoregressive process.

There are literally thousands of articles on the bootstrap and many expository reviews. For starters, though, the book by Efron and Tibshirani[2] is a good introduction, and those by Efron[1] and Shao and Tu[3] can be consulted for more technical accounts.

References

1. Efron, B. (1982) *The Jackknife, the Bootstrap, and Other Resampling Plans*. Philadelphia: Society for Industrial and Applied Mathematics.
2. Efron, B. and Tibshirani, R. J. (1993) *An Introduction to the Bootstrap*. New York: Chapman and Hall.
3. Shao, J., and Tu, D. (1996) *The Jackknife and Bootstrap*. New York: Springer.

Dennis Boos and Leonard Stefanski are at the Department of Statistics, North Carolina State University.