amc technical brief

Analytical Methods Committee No.8 Aug 2001

The Bootstrap: A Simple Approach to Estimating Standard Errors and Confidence Intervals when Theory Fails

Standard errors and confidence intervals for a simple statistic like the mean can be calculated by the use of an algebraic formula derived from familiar assumptions about the data, such as the normal distribution. For a more complex type of statistic (like the familiar relative standard deviation), or where the standard assumptions do not apply, we often find that an algebraic formula cannot be derived. In such instances, a simple alternative method based on re-sampling the data is becoming increasingly popular. This computer-intensive method, known as the bootstrap[1], is widely applicable and is introduced here by two straightforward examples.

Basics

The bootstrap can be used to estimate the standard error of the estimate of a parameter q calculated from a dataset **x** consisting of n individual values, *i.e.*, $\mathbf{x} = (x_1, x_2,...,x_n)$. q could be, for example, a simple mean or a more complex entity calculated from the data. We generate a large number B of new data sets, each of the same size as the original, by sampling **x** at random with replacement. Each resampled data set **x*** is known as a bootstrap sample.

Sampling with replacement means that, if any member x_i of the original set is chosen as the first value of the bootstrap sample, it could also be chosen as any of the successive values. In principle, therefore, a bootstrap sample could consist of the same value repeated n times. In practice, however, such an occurrence would be unlikely, because the number of different bootstrap samples available would be n^n. Even for a dataset of size $n = 5$ there would be 3125 distinct possible bootstrap samples.

For each of these bootstrap samples $\mathbf{x}_b^* \left(b = 1, \mathsf{K}, B\right)$ we calculate $\hat{q}_b^*$ (a bootstrap replication), which is the estimate of the parameter q obtained from the b-th bootstrap sample. We obtain the bootstrap estimate of the standard error of q simply by calculating the standard deviation of the $\hat{q}_b^*$ values. The confidence intervals could be estimated from the usual formula $\hat{q} \pm z\hat{s}_b$ where $\hat{q}$ is the ordinary mean, $\hat{s}_b$ is standard deviation of the $\hat{q}_b^*$ values and z represents the critical value on the N(0,1) distribution, and takes the value of 1.96 for the 95% confidence level. This latter operation depends on the assumption that $\hat{q}_b^*$ is normally distributed. We could inspect a histogram of $\hat{q}_b^*$ to see whether that assumption was plausible.

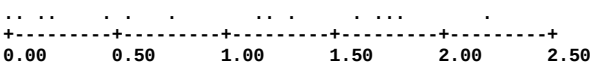
Where $\hat{q}_b^*$ seems to differ from the normal, confidence intervals can be estimated by sorting the values of $\hat{q}_b^*$ into ascending order. If we wanted (say) a 95% confidence interval and we had $B = 1000$ bootstrap samples, the empirical lower and upper limits would be the 25th ($0.025B$) and 975th ($0.975B$) values in the sorted data. In practice the distribution of $\hat{q}_b^*$ is often found to be skewed (because we are usually dealing with a complex type of statistic), so these empirical confidence intervals are probably safer.

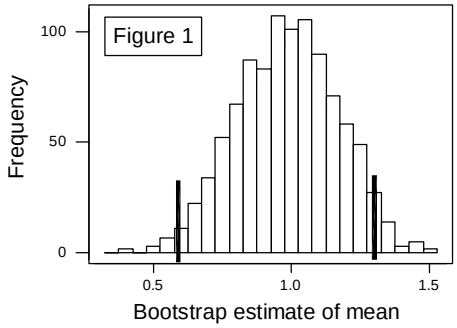
A Simple Example

For demonstration purposes, we use the bootstrap method here to calculate a standard error (SEM) and 95% confidence interval for a ordinary mean. In this instance, of course, the two results can be obtained by statistical theory under the normal assumption, so we can compare the bootstrap result with a classical *t*-interval. The data used are shown below.

0.003 0.070 0.164 0.195 0.441 0.566 0.742 1.136 1.160
1.312 1.623 1.684 1.750 1.803 2.180

Examination of a dotplot of the data shows no obviously suspect data, although we might reasonably entertain doubts that the parent distribution was normal. (There is, in fact, a significant deviation from normality.)





A More Complex Example

The real benefit of the bootstrap is that it can be used on very complex statistics where statistical theory does not provide an answer. Here we use the bootstrap to look at a moderately complex example, the results of a collaborative trial. In this trial, twelve laboratories have independently analysed portions of a homogeneous test material, in duplicate, by a specified method. The results (in ppm) are as follows.

Lab. No.	1st result	2nd result
1	63	61
2	64	62
3	70	68
4	64	60
5	76	75
6	71	71
7	64	65
8	61	64
9	50	53
10	65	70
11	73	74
12	76	72

The most important statistic derived from a collaborative trial is that describing the reproducibility (between-laboratory) precision,