Statistics Notes: Bootstrap resampling methods

J Martin Bland,[1] Douglas G Altman[2]

[1]Department of Health Sciences, University of York, York YO10 5DD, UK

[2]Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford OX3 7LD, UK;

Correspondence to: J M Bland martin.bland@york.ac.uk

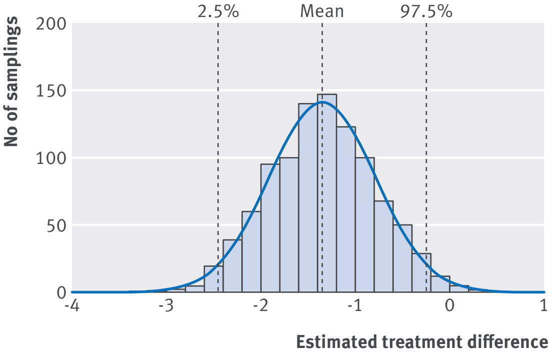
Cite this as: *BMJ* 2015;350:h2622
doi: 10.1136/bmj.h2622

In medical research we study a sample of individuals to make inferences about a target population. Estimates of interest, such as a mean or a difference in proportions, are calculated, usually accompanied by a confidence interval derived from the standard error. The data from a single sample are used here to quantify the variation in the estimate of interest across (hypothetical) multiple samples from the same population.[1] As we have only one sample we need to make assumptions about the data. Most methods of analysis are called parametric because they incorporate assumptions about the distribution of the data, such as that observations follow a normal distribution. Non-parametric methods avoid assumptions about distributions but generally provide only P values and not estimates of quantities of interest.[2]

For a given dataset the assumptions may not be met. In such cases there is an alternative way to estimate standard errors and confidence intervals without any reliance on assumed probability distributions. We use the sample dataset and apply a resampling procedure called the bootstrap. (In general language, a bootstrap method is a self sustaining process that needs no external input.)

The clever idea behind the bootstrap is to create multiple datasets from the real dataset without needing to make any assumptions. Our observed sample is representative of a population about which we wish to make inferences, so a set of randomly chosen observations from our sample will be equally representative of the original population. We can generate a sample of the same size as the original data set by randomly choosing real observations one at a time. Each observation has an equal chance of being chosen each time, so some observations will be picked more than once and some won't be picked at all. That doesn't matter; the new "bootstrap" sample is comparable to the original data set and is equally representative of the target population.

For an example, CADET[3] was a cluster randomised trial comparing collaborative care for depression detected in primary care with treatment as usual. The outcome measure was the PHQ-9 depression scale, and data were available for 505 participants. The estimated mean difference (collaborative care minus treatment as usual) was –1.33 points on the PHQ-9 scale (95% confidence interval –2.31 to –0.35) adjusted for baseline PHQ-9, age, the list size, index of multiple deprivation, city of the practice, and clustering.



Histogram of 1000 resampling estimates of the treatment difference from the CADET data, with corresponding normal distribution curve, mean, and 2.5 and 97.5 centiles

We created another sample of 505 by resampling as described above, the full original sample being available for each of the 505 choices. The resulting new sample of 505 observations included 313 of the original 505 participants, some once, some more than once, a maximum of five times. The same regression analysis which produced the original treatment effect estimate was repeated for this new sample resulting in a slightly different estimated treatment difference of –1.25 points.

Instead of resampling once, we should do it many times and use the variability of the results to obtain a confidence interval. The distribution of the estimated treatment effect from 1000 resamplings of the CADET data is shown in the figure. The mean and standard deviation of this distribution are –1.353 and 0.565. This standard deviation provides an alternative estimate of the standard error of the mean difference between the treatments, which does not make use of any theory about the distribution of the data. There are two ways to use the bootstrap estimates to find a confidence interval. If the resampling distribution is close to normal, as is the case here, the 95% confidence interval will be –1.353–(1.96×0.565) to –1.353+(1.96×0.565), or –2.46 to –0.25. This interval is similar to that obtained using the standard error from the least squares regression on the real data. The other approach is to take the 95% confidence interval directly from the 2.5th and 97.5th centiles of the distribution. For these data the bootstrap confidence interval calculated this way is –2.44 to –0.26. This second approach can be used regardless of the distribution of the bootstrap estimates.

Clearly we need enough repetitions so that the estimates are stable—usually thousands of bootstrap samples are used, especially when using the observed centiles of the distribution of estimates. A repetition of the whole bootstrap analysis for CADET produced almost identical values of the mean (–1.335) and standard deviation (0.567).

This note gives the general idea of the bootstrap; there are many variations.[4] We can get a bootstrap estimate for any quantity we can calculate from any sample. Bootstrap methods are particularly favoured by health economists, because cost data tend to be highly skewed and unsuited to conventional approaches.[5] They are also useful for complex datasets—for example, when the observations aren't independent.

Contributors: JMB and DGA jointly wrote and agreed the text.

Competing interests: We have read and understood the BMJ Group policy on declaration of interests and have no relevant interests to declare.

Provenance and peer review: Not commissioned; not externally peer reviewed.

1 Altman DG, Bland JM. Standard deviations and standard errors. *BMJ* 2005;331:903.

2 Altman DG, Bland JM. Parametric v nonparametric methods for data analysis. *BMJ* 2009;338:a3167.

3 Richards DA, Hill JJ, Gask L, et al. Clinical effectiveness of collaborative care for depression in UK primary care (CADET): cluster randomised controlled trial. *BMJ* 2013;347:f4913.

4 Carpenter J, Bithell J. Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. *Stat Med* 2000;19:1141-64.

5 Schroeder E, Petrou S, Patel N, et al. Cost effectiveness of alternative planned places of birth in woman at low risk of complications: evidence from the Birthplace in England national prospective cohort study. *BMJ* 2012;344:e2292.