BJP British Journal of Pharmacology

DOI:10.1111/j.1476-5381.2012.02101.x
www.brjpharmacol.org

EDITORIAL

Making do with what we have: use your bootstraps

Guillaume Calmettes[1], Gordon B Drummond[2] and Sarah L Vowler[3]

[1]Department of Medicine, University of California, Los Angeles, Cardiovascular Research Laboratory at UCLA, Los Angeles, CA, USA, [2]Department of Anaesthesia and Pain Medicine, University of Edinburgh, Royal Infirmary, Edinburgh, UK, and [3]Cancer Research UK, Cambridge Research Institute, Li Ka Shing Centre, Cambridge, UK

Correspondence
Dr Gordon B Drummond, Department of Anaesthesia and Pain Medicine, University of Edinburgh, Royal Infirmary, Edinburgh, 51 Little France Crescent, Edinburgh EH16 4HA, UK. E-mail: g.b.drummond@ed.ac.uk

This article is being published in *The Journal of Physiology*, *Experimental Physiology*, the *British Journal of Pharmacology*, *Advances in Physiology Education*, *Microcirculation*, and *Clinical and Experimental Pharmacology and Physiology*.

Guillaume Calmettes is a Postdoctoral Research Fellow in the Cardiovascular Research Laboratory at UCLA.

Gordon Drummond is Senior Statistics Editor for *The Journal of Physiology*.

Sarah Vowler is Senior Statistician in the Bioinformatics Core at Cancer Research UK's Cambridge Research Institute.

This article is the 10th in a series of articles on best practice in statistical reporting. All the articles can be found at http://onlinelibrary.wiley.com/journal/10.1111/(ISSN)1476-5381/homepage/statistical_reporting.htm.

Key points

- Bootstrap methods are a recent innovation.
- They depend on adequate computing power.
- Bootstraps are more empirical than other statistical processes.
- Given a suitable sample, many population features can be derived from the sample alone.
- The process takes repeated random samples from the original data.
- The process is flexible and has many applications.

A jackknife is a pocket knife that is put to many tasks, because it is ready to hand. Often there could be a better tool for the job, such as a screwdriver, a scraper or a can opener, but these are not usually pocket items. In statistical terms, the expression implies making do with what is available. Another simile, of an extreme situation, is the bootstrap: extricating oneself from a predicament by the only means available (Curran-Everett, 2009). These everyday terms have been applied to statistical methods that allow us to work with limited data and draw robust conclusions. We have already discussed permutation tests, which are similar in that they use only the data collected, and make no assumptions about the population (Drummond and Vowler, 2012). We noted that permutation tests become laborious with larger samples, and the same is true of the bootstrap tests, where the data sample is repeatedly *resampled*. These tests are quite recent developments, and require computers, and sadly have not yet been provided in many of the more general statistics software packages.

Obtaining robust conclusions from the data alone, without requiring assumptions about the population sampled, is an attractive proposition if we are unsure about the features of the population. It is all very well assuming that the data have been taken from a population with defined characteristics, but how can we be sure that such an assumption is correct? Think of the entire population of Californian frogs. In reality, we cannot verify our assumption that these values will have a normal distribution. The first step of the classic *t*-test involves exactly this assumption. When we report such results, we expect the reader to agree that the assumption is correct. We then proceed to postulate the

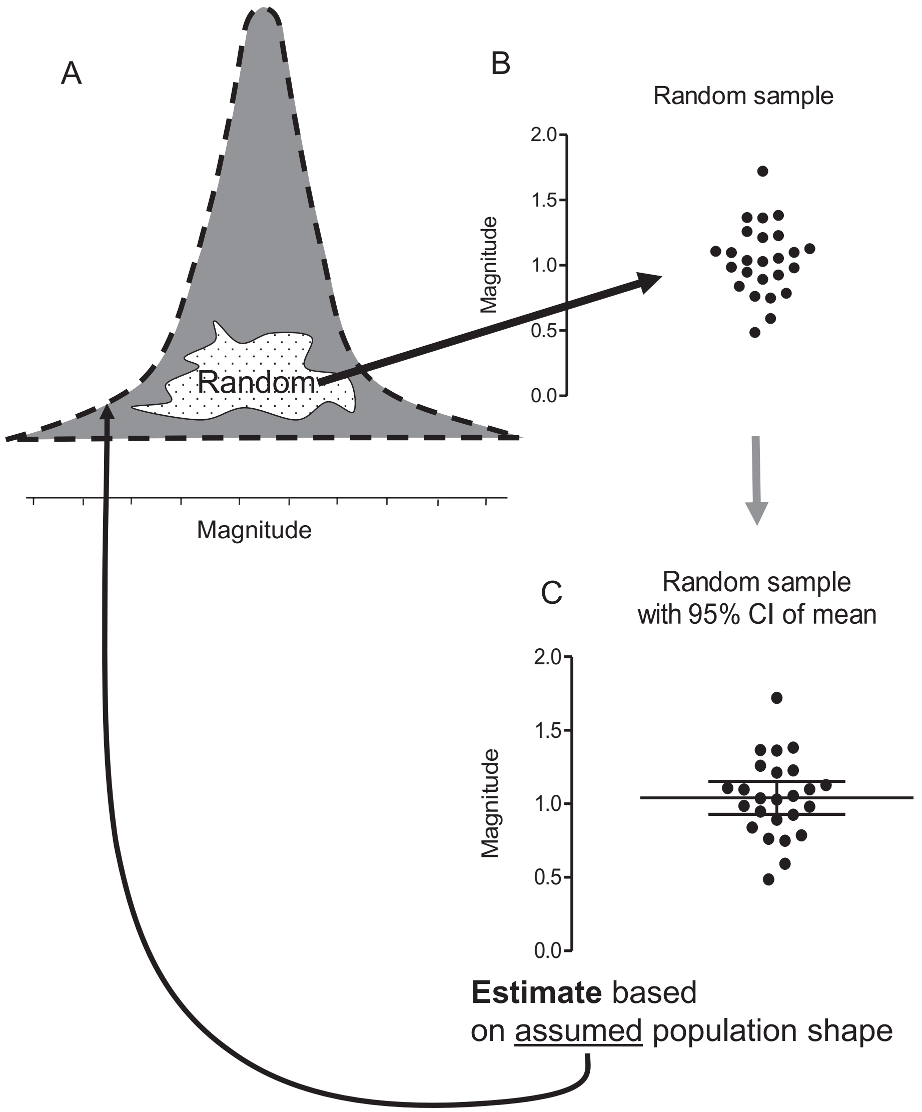

Figure 1

The classic procedure for estimation of population parameters. The population characteristics are derived from the sample on the basis of assumptions concerning the characteristics of the parent population. CI, confidence interval.

population parameters (such as mean, SD). After taking a sample, we can calculate a mean, and by assuming a specific form of population distribution we can estimate confidence intervals for our sample (Figure 1).

In practice, most researchers would be more reassured to know that the mean and its confidence interval could be reliably estimated directly from a sample that they had taken, and not wish to make assumptions about the characteristics of the population. Often, an initial test to assess a sample for 'normal distribution' can be unhelpful or misleading, often because there may be insufficient data to provide a convincing test result – another occasion when absence of proof is not proof of absence.

The bootstrap process is a way of working with the data we have. As long as we have a sample of sufficient size, we can use the sample to determine the probability distribution, and hence other population measures such as confidence intervals. We do not need to assume anything else about the population from which we have taken the sample, other than the fact that it has been randomly sampled. The saying 'a bird on the hand is worth two in the bush' summarizes this philosophy: although the data in the sample we have taken may be insufficient to provide an adequate probability distribution directly, we now have the opportunity to take repeated further random samples from the sample that we already have. The sample we have taken contains values that reflect the original population, and random sampling from these values allows characteristics of the original population to be inferred. The principle of the bootstrap is that we use our sample, which can be repeatedly, randomly sampled, to estimate features of a source population that is inaccessible.

These procedures have only become popular with the use of computers to do statistical tests (Efron and Tibshirani, 1991), because the calculations are tedious and have to be

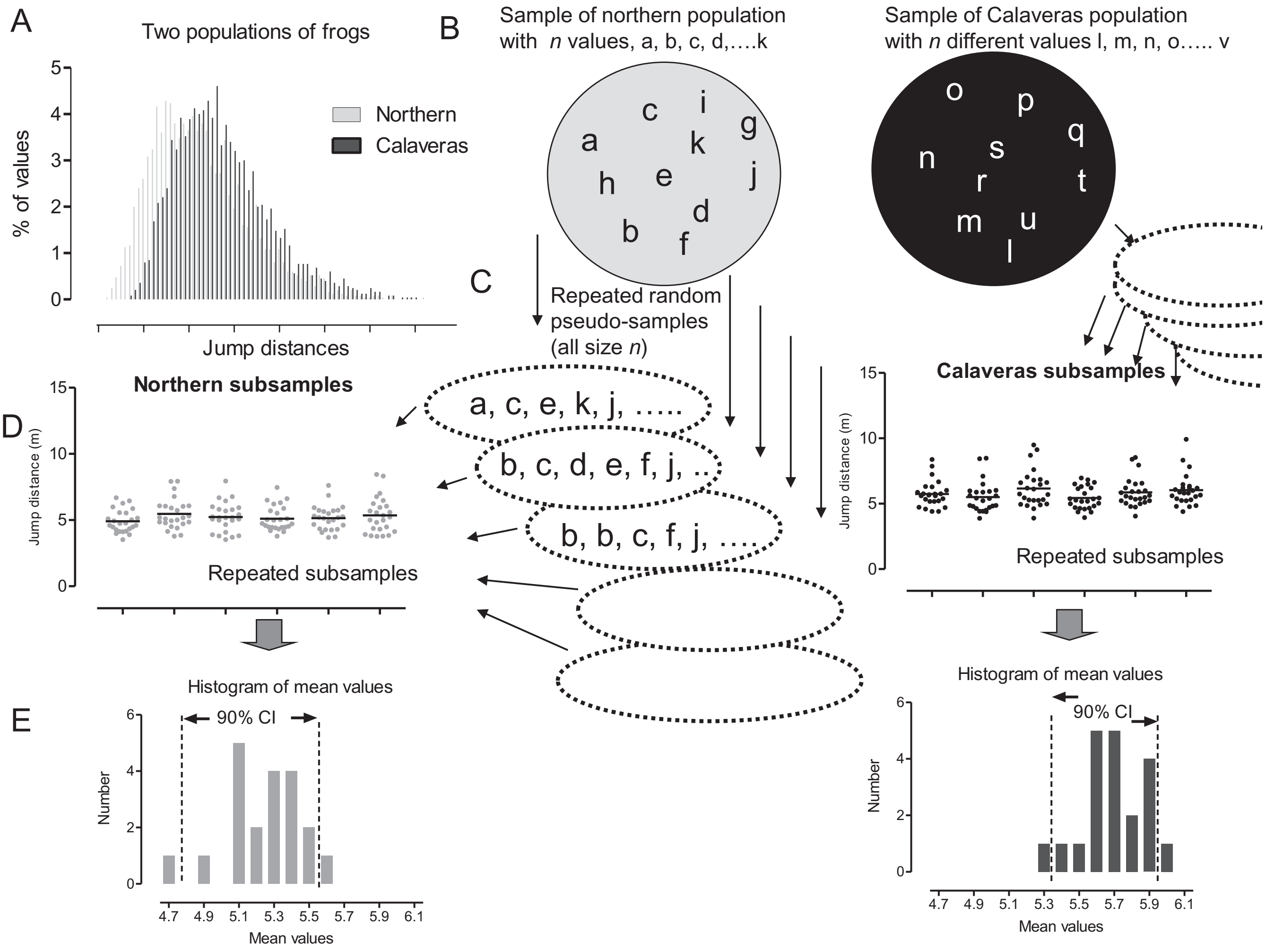

Figure 2

The bootstrap process. (A) The population distributions that are to be sampled. (B) Samples are obtained randomly from each population. (C) From each sample, repeated pseudo-samples of same size as the original sample are drawn randomly, choosing from all the values. Each observation remains in the original sample after the value is noted, so that values in the original sample can be chosen more than once to make up the pseudo-sample. Thus, each bootstrap sample will usually contain duplicate observations from the original sample. (D) The mean of each pseudo-sample is calculated for the many repeats. (E) The distribution of these mean values can provide a measure of the confidence limits of the mean of each original sample. CI, confidence interval.

done many times. A famous photograph of R. A. Fisher (he devised a permutation test that also required repeated calculations) shows him keying data into a mechanical calculator. Any moderate bootstrap test can involve 1000s of 'resamples' and would probably wear out both calculator and statistician long before the test was completed!

The concept is that after a random sample has been taken, the values in this sample are repeatedly, randomly 'resampled' to generate a large series of new sets of values that we shall call 'pseudo-samples'. From each of these pseudo-samples, we can calculate values that characterize the source population. For example, in Figure 2 we derive mean values. In other words, we are using the original sample as a substitute for the original population to provide further samples. We use these further samples, in this case, to estimate the sampling distribution of the mean, but we can use the same process to obtain other features of the population from the data.

The original sample contains within it the features of the population it was drawn from. We then apply the same principles of inference and sampling to the sample we have taken as we did when we took our original sample from the population. *The bootstrap samples are to the original sample, as the original sample was to the population.* Although this process yields an approximate result, it is likely to give more accurate estimates of population parameters than if these values had been calculated on the basis of an initial incorrect assumption of a specific pattern of distribution. The estimates may be inaccurate if the first sample is limited, because limited data cannot provide a sufficiently representative sample. The

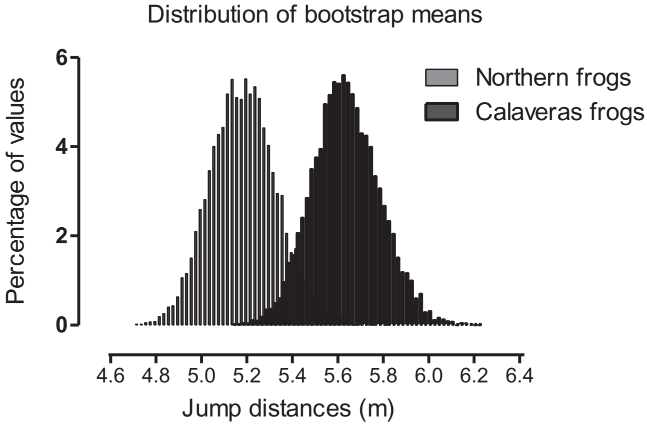

Figure 3

The distribution of the means derived from the repeated pseudo-samples.

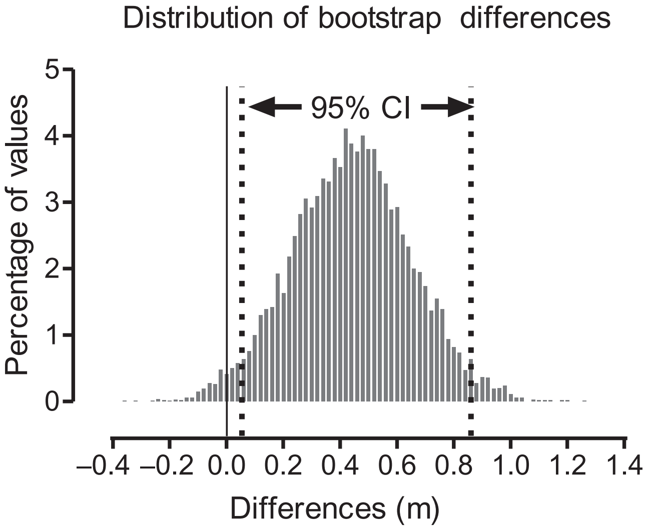

Figure 4

The distribution of the differences of the mean values obtained from pseudo-samples taken from each sample. The dashed lines indicate the central 95% of these values and the continuous line indicates zero difference. CI, confidence interval.

bootstrap process is particularly suited to describing populations; it can also be used for comparisons, but other tests such as permutation tests may be more appropriate here.

Let us compare a sample of frogs from the north of California with a group sampled near Calaveras, where escapees from the jumping competition have interbred with native frogs. We wish to compare estimates of the two populations from which these samples have been drawn. We generate repeated random pseudo-samples from each sample. Each time a value is taken, we can choose any of the values that are in the original sample so there is the opportunity for each value to be chosen more than once, in each pseudo-sample, and from pseudo-sample to pseudo-sample. (This is sampling ‘with replacement’, i.e. after it has been chosen, the value is replaced in the original stock of values.) Figure 2 shows a simplified version of the process.

The mean of each pseudo-sample is calculated. Figure 2 shows the first six of these pseudo-samples as dot plots. If we continue taking pseudo-samples until we have 20, and calculate the mean of each of these, we can then arrange these mean values as a distribution histogram, as seen in the bottom panels of Figure 2. By taking the central 18 values, we define the 90% confidence limits for the mean of the original sample.

In practice, the process is repeated many more times than this. Typically, we could generate 10 000 pseudo-samples to generate a range of mean values. We use these mean values to generate the confidence limits. This is referred to as the bootstrap percentile method. The distribution of the means for this number of samples can be seen in Figure 3.

Further analysis of the data can be done with the same basic method. We can use the method to conduct comparisons. Thus, in our example we find that the mean jump distance of our Calaveras sample is 0.44 m greater: what are the 95% confidence intervals of this estimate? We approach this by independently drawing a random pseudo-sample, with replacement, from each group (as in Figure 2B and C), and calculate the mean value of each sample. The difference between these mean values is an estimate of the difference between the groups. We continue to repeat the process of randomly taking pseudo-samples from each group and calculating differences between the means. Figure 4 shows the distribution of differences obtained after repeating this process 10 000 times. If we count 250 from each extreme of these differences, we define the 95% confidence interval for the original observed difference: 0.44 m (0.039, 0.847). These values do not include zero, and thus we conclude that the observed difference in jump distance between the two groups of frogs is unlikely to be a result of chance, and gain an indication of the likely size of this difference.

The bootstrap method is flexible and robust, well suited for analysis of data whose population distribution is uncertain, as is often the case in biological studies (Efron and Tibshirani, 1993). For example, assumptions about distribution can be avoided when comparing quanta at synapses (Van der Kloot, 1996). There may be occasions when the bootstrap can fail: for example, it is not good with extreme distributions, or to estimate statistics – like the maximum – that depend on very small features of the data. Modern computers make the tedious procedure of repeated sampling straightforward. However, standard textbooks and the standard statistics packages have failed to acknowledge the value of the bootstrap approach. Curran-Everett describes how to use the statistical software package R for bootstrap methods (Curran-Everett, 2009), an add-on facility exists for SPSS and Excel, and Cole has described a macro to use with SAS (Cole, 1999). Most current packages lack standard facilities for these procedures. These useful and powerful methods should become gradually more common in standard statistical software.

Note: Those with some programming skills can set up basic approaches themselves. Useful insights on how to implement different bootstrap methods in various programming languages can be found in Good’s (2006) book. For users of the Python programming language, the data used in this paper and the code used for its analysis are available at http://bit.ly/KJ67RW (Calmettes, 2012).

References

Calmettes G (2012). Bootstrap-tools. http://bit.ly/KJ67RW accessed 26-5-2012.

Cole SR (1999). Simple bootstrap statistical inference using the SAS system. Comput Methods Programs Biomed 60: 79–82.

Curran-Everett D (2009). Explorations in statistics: the bootstrap. Adv Physiol Educ 33: 286–292.

Drummond GB, Vowler SL (2012). Different tests for a difference: how do we do research? Br J Pharmacol 165: 1217–1222.

Efron B, Tibshirani R (1991). Statistical data analysis in the computer age. Science 253: 390–395.

Efron B, Tibshirani RJ (1993). An Introduction to the Bootstrap. Chapman and Hall: New York.

Good PI (2006). Resampling Methods: A Practical Guide to Data Analysis. Birkhäuser: Boston, MA.

Van der Kloot W (1996). Statistics for studying quanta at synapses: resampling and confidence limits on histograms. J Neurosci Methods 65: 151–155.