Statistical inference using bootstrap confidence intervals

This article was published a few years ago. I have now improved the spreadsheet, and extended the method beyond confidence intervals: see http://woodm.myweb.port.ac.uk/SL/resample.xlsx, and https://arxiv.org/abs/1702.03129v2

Bootstrap confidence intervals provide a way of quantifying the uncertainties in the inferences that can be drawn from a sample of data. The idea is to use a simulation, based on the actual data, to estimate the likely extent of sampling error. **Michael Wood** explains how simple bootstrapping works and explores some of its advantages.

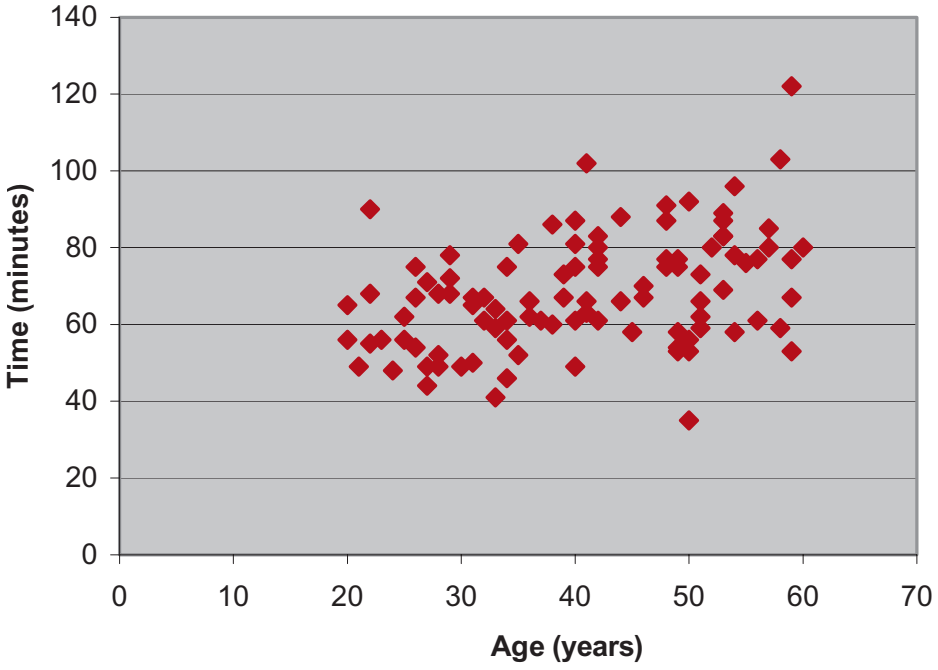
An important task of statistics is to establish the degree of trust that can be placed in a result based on a limited sample of data. One way to do this is to set up a null hypothesis and to estimate a significance level—see "Does significance matter?" by R. Allan Reese in the first issue of *Significance*.

Another way of approaching the problem is to set up a confidence interval—this often makes more sense than testing a null hypothesis (as we shall see below). Confidence intervals are normally derived using probability theory, but they can be estimated by an alternative method known as *bootstrapping*, which has a number of further advantages over conventional methods. Bootstrap confidence intervals thus have a double potential advantage over most hypothesis tests—due to the fact that they are confidence intervals, and due to the bootstrapping method.

Bootstrap confidence intervals

Let's imagine we have got data on the age and time taken for a 10 km run (in minutes, over a hilly course) from a random sample of 100 people aged between 20 and 60 from a large population. (I've used a fabricated example to avoid getting distracted by questions about this population and the sampling process—these are obviously important but irrelevant for my purposes here.) These data are shown in Figure 1. The (Pearson) correlation between the variables, based on the whole sample, is 0.40—indicating the unfortunate tendency for people to slow down as they get older.

Figure 1. 10 km race times for a sample of 100



Obviously, this can only give a rough guide to the correlation in the wider population. To establish just how rough this guide is, the bootstrap method makes a guess about what the population might look like, and then runs some computer simulation experiments on this guessed population. In this case, the obvious guess is that the distribution of the population is just the same as that of the sample—i.e. 1% like the first person in the sample (aged 56 with a time of 61 minutes), 1% like the second person, and so on for all 100 people in the sample. It is now easy to draw *resamples* of 100 "people" from this guessed population: all we do is choose a member of the sample at random, replace it and choose another one, and so on until we have a simulated sample (a resample) of 100 people from this guessed population. (Replacing each sample member before drawing the next member of the resample ensures that the probability of each member of the sample being drawn remains at 1%.) Now we can work out the correlation from this resample.

For example, in a typical resample drawn from this sample, the labels (corresponding to the sequence in the original sample) of the points drawn were 84, 26, 55, 95, 73, 93, 58, 20, 26.... Note that point 26 has already appeared twice—some points will arise several times, others not at all, in any resample. The idea is to simulate the process of taking a random sample from the guessed population, which we are assuming is similar, in some sense, to the real population. The correlation between the two variables from

the whole of this resample comes to 0.37, which is close, but not identical, to the value from the original sample (0.40).

Now we just do this lots of times. The whole process can be set up on a spreadsheet (see the section on Software below for details): Figure 2 shows the correlations from 2000 resamples. The mean of all these correlations is 0.40, the 2.5 percentile is 0.22 and the 97.5 percentile is 0.57. 2000 resamples are sufficient to give fairly stable results—running the same simulation again gave a mean of 0.40 (as before), and the two percentiles were 0.23 and 0.57.

"At first sight, the process seems to be simply a matter of recycling the original data; it is not obvious how any new information can be obtained"

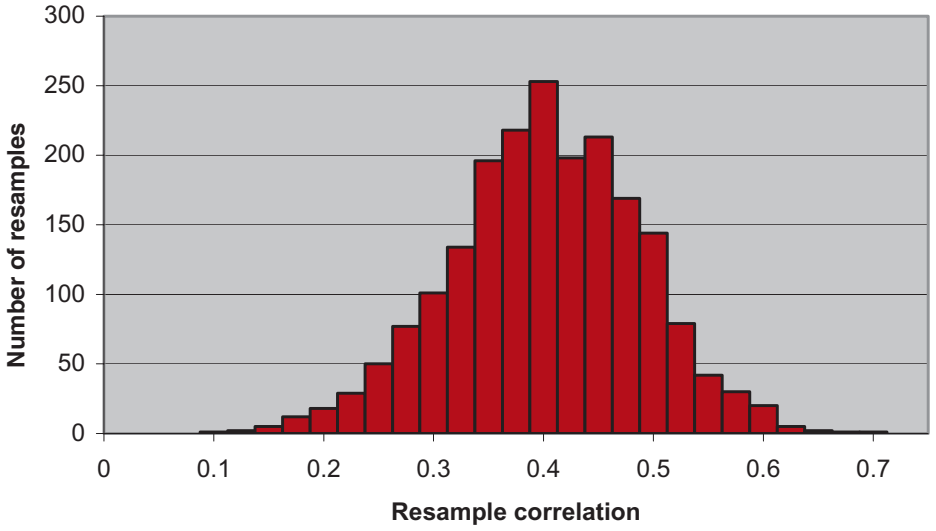


Figure 2. Resample frequencies. Correlation from 2000 resamples

As the mean of this distribution represents the sample value of the correlation (0.40), and the distribution is designed to tell us about sampling error, the obvious thing to do is to use Figure 2 as a sort of confidence distribution: the conventional 95% confidence interval, for example, extends from 2.5 to the 97.5 percentiles—from 0.22 to 0.57 (roughly obvious from the graph; the exact values can be obtained from the spreadsheet). This interval is known as the *percentile* bootstrap interval because it follows the percentiles of the resampling distribution.

It is important to distinguish between the three different types of "sample" in this process. First there is a sample of 100 people. Second, there are resamples based on the original sample—intended to represent hypothetical samples drawn from a similar population to the actual sample. And third, there is the collection of 2000 of these resamples.

At first sight, the process seems to be simply a matter of recycling the original data; it is not obvious how any new information can be obtained. This is the origin of the term "bootstrapping": the process is analogous to the idea of pulling oneself up by one's bootstraps[1] without anything else to help (like some mathematical probability theory), and it may seem impossible for similar reasons. The reason why statistical bootstrapping works, and why the argument is not circular, is that the procedure of resampling with replacement mimics the process of taking samples from a large population, so the simulation can provide us with new information about the variation in these samples, and so about sampling error.

It may, or may not, seem obvious to you that this is a reasonable way to derive a confidence interval, but it is worth being a little careful. Remember that we are trying to infer a general truth (about a population parameter) that goes *beyond the data we actually have*. In my view, it would be odd if there were a way of doing this that was unambiguously and uncontroversially right. One argument is outlined in the next paragraph, but inevitably it has limitations.

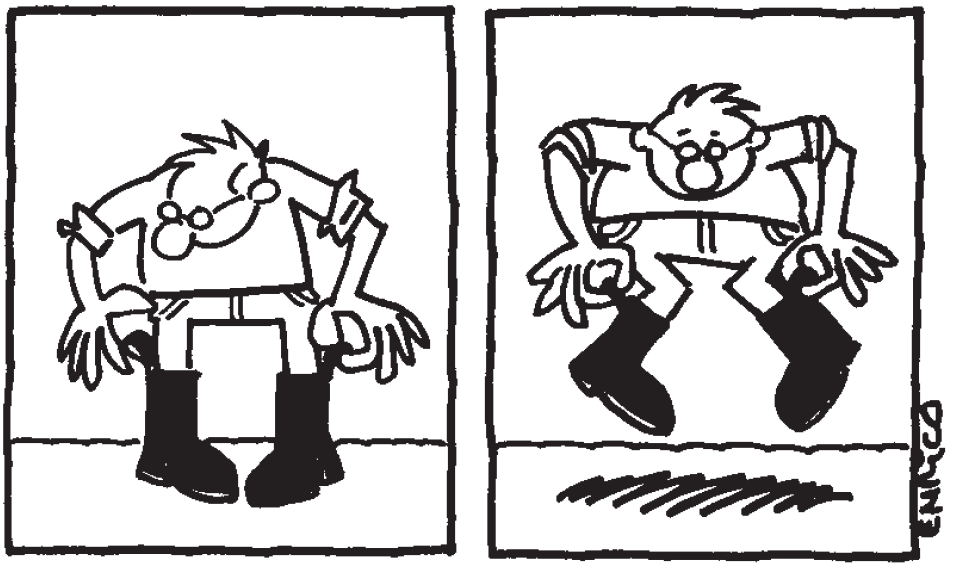
Figure 2 (and the exact percentiles from the spreadsheet) suggests that a correlation worked out from a random sample of 100 from this guessed population has a 95% chance of being within about 0.17 of the population correlation of 0.40. Now remember that we have one random sample from the *real* population with a correlation of 0.40. Taken together, these two pieces of information suggest that a sensible 95% confidence interval for the *real* population correlation is the sample value, 0.40, plus or minus about 0.17: i.e. 0.23 to 0.57, which, of course, is almost the same as the confidence interval we read off from Figure 2.

Assumptions and difficulties

Like many arguments in statistics, this is a little less secure than it may at first appear. I have prefaced the figure 0.17 with the qualifier "about"—this is because it comes from the top end of Figure 2 (the 97.5 percentile, 0.57, minus the mean, 0.40), whereas if I had used the bottom it would have been 0.18 (0.40 – 0.22). The argument assumes that the distribution is reasonably symmetrical; if the sample had been much smaller or the correlation much closer to 1, Figure 2 might have been noticeably asymmetrical, and the argument above gets much more complicated. Similarly, if the sample had been very small—say 5 people instead of 100—it would not have been reasonable to form the guessed population by extrapolating the pattern in the way we did above. The argument depends on a number of assumptions[2] which may not be exactly met in practice.

There are a number of more elaborate methods designed to overcome some of the deficiencies of the simple percentile interval[3]. Some of these are fairly obvious (for example, for a small sample it may make sense to use a mathematically defined probability distribution to derive a guessed population); others less so. However, I think it is possible to be too much of a purist here. Provided the resample distribution is reasonably symmetrical, and the sample reasonably large (otherwise the guessed population may be too crude for making inferences about the real population), I think it is reasonable to use the percentile interval. It is easy to forget that the idea of confidence intervals is to assess uncertainty, so it is almost a contradiction in terms to insist on too great a level of certainty in their definition. Many standard techniques have errors that are conventionally ignored (think of the commonly used normal approximation to the binomial distribution, or the various unchecked assumptions in many analyses of variance). If the resample distribution is slightly asymmetric, I would be inclined to use the more spread-out side of the distribution to define the width of a symmetric confidence interval—so the above confidence interval would be 0.40 ± 0.18. This could perhaps be described as an *at least 95%* confidence interval, to indicate that 95% is likely to be an underestimate of the true confidence level.

The method can be used for many statistics—mean, median, standard deviation, etc. The only difference is that, instead of working out a correlation from each resample, we work out another statistic. However, there are obviously circumstances where the simple bootstrap method makes little sense. Suppose we were trying to use the sample in Figure 1 to derive an estimate for the best (minimum) race time in the whole population. The simple method above obviously needs adjusting—the main problem here being that the population minimum time must be less than, or the same as, the sample minimum time,



so an interval which is symmetrical about the sample minimum time is clearly stupid.

Software

The analysis above was carried out with the spreadsheet at http://userweb.port.ac.uk/~woodm/nms/resample.xls. (This is set up for only 200 resamples to reduce the file size and to improve the recalculation time: the "Read this" sheet explains how to increase this.) An even simpler possibility for the analysis of a single variable only is the stand-alone program at http://userweb.port.ac.uk/~woodm/nms/resample.exe.

For more advanced work there are many other possibilities[4]. These include packages designed for simulation methods such as Resampling Stats (available from http://www.resample.com) and major statistical packages such as SPSS.

Bootstrap *versus* conventional methods

How do bootstrap confidence intervals compare with more conventional methods of derivation? I tried to find the answer SPSS produced from the data above, but confidence intervals are not on the menu of offerings produced with correlations. Instead, I compared the bootstrap 95% interval for the mean run time (65.6–71.3 minutes, estimated in exactly the same way as the interval for the correlation, except that we work out the mean from each resample) with the corresponding result from SPSS (65.4–71.3 minutes). The two methods give very similar answers. Despite this similarity, bootstrapping has a number of features that differentiate it from conventional approaches to deriving confidence intervals.

The first is that there are no technical mathematical concepts involved except, inevitably, the correlation coefficient for which we want the confidence interval: no central limit theorem, no *t* tables, no standard errors, and so on. People without any knowledge of these concepts should be in a position to follow the argument, which is surely an advantage. Despite this, the argument is rigorous in the sense that the rationale behind each step is clear, and so it is possible to see potential problems. For example, the assumption that the sampling process is a random one is clearly implicit in the way random resamples are used in the simulation. Similarly, it should be clear that sampling errors are the only source of errors that are incorporated: measurement errors, for example, are ignored.

This is related to a second feature: when deriving a bootstrap percentile interval, no assumptions are made about such things as the data being normally distributed, so no checks are necessary and there are no difficulties with such assumptions being violated. The approach is more robust in this sense. (There are, of course, other assumptions involved which may mean that the results are not fully accurate—but this is true of any approach.)

Thirdly, exactly the same method will work for many other statistics—for example, a mean, median, standard deviation, proportion, regression coefficient and Kendall's correlation coefficient. (Excel does not have a built-in function for the Kendall correlation coefficient. I have included one in the add-in: http://userweb.port.ac.uk/~woodm/nms.xla; installing this will enable the spreadsheet described above to derive bootstrap confidence intervals for Kendall correlation coefficients in just the same way as Pearson coefficients—although it may take a long time because the function is rather slow.) It will also work for the difference of the means of two subgroups (http://userweb.port.ac.uk/~woodm/diffofmeansconfidence.xls), and other similar statistics. This flexibility is obviously a tremendous advantage, and means that, even in cases where no approach based on probability theory is available, it may still be possible to derive bootstrap confidence intervals. Bootstrapping may be the method of last resort when all else fails.

On the other hand, as the answer comes from a simulation, repeating the analysis means you may get a different answer each time you run the simulation. The obvious way of resolving this is to run the simulation with as many resamples as is practical, and then to run it again to check that the results are reasonably stable. (The fact that the results are stable does not, of course, mean they are right!)

Confidence intervals *versus* testing a null hypothesis

Another feature, which is perhaps still unconventional in some contexts, is the idea of setting up an interval for a parameter instead of testing a null hypothesis. This has a number of clear advantages. If you are using SPSS, for example, to analyse the data above, you won't be offered a confidence interval for the correlation; instead you will be given a significance level based on the null hypothesis of no correlation. This would be extremely small, which tells you that the null hypothesis of no correlation is untenable. But this is hardly news: the correlation was obviously going to be positive, the only question being how positive. The confidence interval gives you a realistic assessment of the likely inaccuracies in relying on a sample of 100, whereas the significance level just answers a silly question and gives no answers to the real question.

Similarly, many published regression studies include significance levels for regression coefficients but not confidence intervals. This is reasonable if all you want to do is to test whether there is significant evidence that the coefficients differ from zero, but not if you want some idea of how accurate they are likely to be if they are significant. And, if you are interested in the difference of two means, you can either test the null hypothesis that there is no difference, or you derive a confidence interval for the difference. The latter has the advantage that you get to know how large the difference is, as well as whether or not the difference exists. It is also more user friendly, and less liable to misinterpretation.

References

1. Efron, B. and Tibshirani, R. J. (1993) *An Introduction to the Bootstrap*, p. 5. New York: Chapman and Hall.
2. Wood, M. (2003) *Making Sense of Statistics: a Non-mathematical Approach*. Basingstoke: Palgrave.
3. See, for example, References 1 and 4.
4. Lunneborg, C. E. (2000) *Data Analysis by Resampling: Concepts and Applications*, p. 556. Pacific Grove: Duxbury.

Michael Wood teaches and researches in the Business School at Portsmouth University. As well as statistics, his interests include research methods, decision analysis and cycling.