Bootstrapped Confidence Intervals as an Approach to Statistical Inference

Michael Wood

Organizational Research Methods 2005 8: 454
DOI: 10.1177/1094428105280059

Bootstrapped Confidence Intervals as an Approach to Statistical Inference

MICHAEL WOOD
University of Portsmouth

Confidence intervals are in many ways a more satisfactory basis for statistical inference than hypothesis tests. This article explains a simple method for using bootstrap resampling to derive confidence intervals. This method can be used for a wide variety of statistics—including the mean and median, the difference of two means or proportions, and correlation and regression coefficients. It can be implemented by an Excel spreadsheet, which is available to readers on the Web. The rationale behind the method is transparent, and it relies on almost no sophisticated statistical concepts.

Keywords: *bootstrap; confidence interval; hypothesis test; resampling; statistics*

Research in the organizational sciences makes frequent use of statistical inferences. These inferences often take the form of null hypothesis tests based on mathematical probability theory. We want to know whether the mean of one group differs from the mean of another group, or whether two variables are correlated, or if a parameter in a regression model differs from zero.

This article explains and discusses the idea of a bootstrap percentile confidence interval—which is an alternative to many of these methods. Furthermore, it is an alternative with substantial advantages over the usual methods. These advantages fall into two separate categories.

The first stems from the use of confidence intervals in place of significance levels (p values) and null hypotheses. Instead of, for example, testing the hypothesis that there is no difference between the means of two populations, and citing the resulting significance level, we can set up an interval estimate for the difference of the two means. This has a number of advantages, which are briefly explained in the next section.

The second category of advantages are those due to using the method of bootstrapping in place of conventional methods to derive confidence intervals. This is the main subject of this article. This second category of advantages divides into three subcategories:

Organizational Research Methods, Vol. 8 No. 4, October 2005 454-470
DOI: 10.1177/1094428105280059

1. The method is more transparent, simpler, and more general than conventional approaches. Understanding the rationale behind it requires very little knowledge of mathematics or probability theory. This is less true of some of the more advanced bootstrap approaches (e.g., the bias-corrected bootstrap), but these are not the subject of this article.
2. The assumptions on which the method depends are less restrictive, and more easily checked, than the assumptions on which conventional methods depend.
3. The method can be applied to situations where conventional methods may be difficult or impossible to find.

It is difficult to overstate the potential importance of these advantages. Statistics is a difficult area—even for trained researchers—that is dependent on subtle concepts. Approaches to statistics that are accessible without a lengthy specialist training have obvious benefits. And even for those with a specialist statistical training, the second and third of the three advantages listed above are powerful arguments in favor of bootstrapping. Sometimes there may be no alternative.

Bootstrap methods—including the percentile interval discussed here—are well established in the technical literature (e.g., Davison & Hinkley, 1997; Efron, 1979; Efron & Tibshirani, 1993; Hall, 1992; Lunneborg, 2000; Mooney & Duval, 1993). Inevitably, there are a variety of methods and rationales behind them, some simple and others more subtle. My aim in this article is *not* to review all these possibilities but to explain how to derive bootstrap confidence intervals as simply as possible, while making the underlying rationale, and the assumptions that need to be checked, clear. Most of the approaches to bootstrapping in the literature are not discussed in this article because my aim is focus on the simplest methods—the concept of the standard error, for example, is irrelevant to the methods and their rationale, and so readers do not need to be familiar with it. Obviously, readers should bear in mind that when the method discussed here is not adequate, there are further possibilities in the literature.

Despite the three potential advantages listed above, bootstrapping, and other similar methods, are not widely used in research. Studies evaluating their use (e.g., Russell & Dean, 2000) tend to focus on the second and third of the three potential advantages listed above: the fact that they do not require restrictive, and often unrealistic, assumptions (e.g., about measurements being normally distributed); or the fact that there may be no alternative. These are important, but my aim here is broader, encompassing the first potential advantage as well.

Bootstrapping requires computer software. The examples in this article are based on two simple programs—an Excel workbook and a small stand-alone program—available to readers on the Web at http://userweb.port.ac.uk/~woodm/nms/resample.xls and http://userweb.port.ac.uk/~woodm/nms/resample.exe.

Hypothesis Testing Versus Interval Estimates

It is not the purpose of this article to review this issue in detail but simply to point out some of advantages of the confidence interval approach to inference.

The concepts behind the testing of null hypotheses and the derivation of significance levels (p values) are potentially very confused and confusing. Misinterpretation and misuse are widespread, and there are strong arguments that null hypothesis testing is often not a sensible approach to statistical inference (e.g., Gardner & Altman, 1986; Kirk, 1996; Lindsay, 1995; Morrison & Henkel, 1970; Royall, 1997; Wood, 2003).

To take one example, more or less at random, McGoldrick and Greenland (1992) found that the mean rating for "helpful/friendly staff" of a sample of bank customers was 6.495 (on a 1 to 9 scale), whereas the equivalent figure from a sample of building society customers was 6.978. The significance level (p value) cited for this is *.000*—that is, it is highly significant. The fact that the size of the difference is less than 0.5 is not mentioned, but the implication is that because it is statistically significant it must be important. This, of course, is unlikely to be the case given the small size of the difference. Significance tests can produce highly significant results (in the statistical sense) even when the size of the effect is too small to be of any real significance (in the nonstatistical sense). This problem is exacerbated if readers of research reports misinterpret "significant" as meaning "important." There are other objections to significance tests—see the references cited above—but this is perhaps the most important in practice.

An alternative approach would have been to derive confidence intervals for the difference between the means of the two groups of customers (Gardner & Altman, 1986; Kirk, 1996; Wood, 2003). It is normally possible to use confidence intervals instead of hypothesis tests: The confidence interval version of the above result might be that we can be 99.9% confident that the true difference between the building society and the bank rating is in the range 0.1 to 0.9. The fact that this interval does not include zero means that we can be almost completely confident that the building society rating is more than the bank rating but that the difference is less than one point on the 1 to 9 scale. This is less liable to misinterpretation than the p value, and provides information about the size of the effect, which the p value does not.

The same principle applies whenever we have a measure of the relationship between two variables. This might be a difference between two proportions, or a correlation coefficient, or a regression coefficient. As an alternative to testing the null hypothesis that the population value of the measure is zero, we can derive a confidence interval for the measure.

Intuitively, confidence intervals are an attractive idea. However, the way they are defined formally tends to be more convoluted than might be expected. For example, according to one introductory text, "The confidence level refers to the expected percentage of times that a confidence interval would contain the population value of the statistic being estimated, under repeated random sampling" (Smithson, 2000, p. 146). Alternatively, 95% (to take a specific figure) confidence intervals for the population parameter, θ, are sets of possible values, v, which are consistent with the data in the sense that a test of the null hypothesis that $\theta = v$ yields a significance level of more than 5%—that is, the result is not significant at the 5% level (Royall, 1997). This is a potentially confusing notion, which is also subject to logical problems (Royall, 1997, pp. 78-79). The rationale behind the bootstrap approach below is, I hope, simpler.

The Percentile Bootstrap Interval for Finite Populations

Bootstrapped confidence intervals can be derived for any numerical statistic based on a random sample. I will start by assuming that the sample is drawn from a finite population. I could present the argument in general terms, but it is easier to take a specific example. This uses data from a questionnaire sent to 650 of the 4,500 members on the mailing list of a society of accountants: 105 of them (16%) responded. An edited

Mean of resamples of 98 (no replacement, N = 4,508).
1st resample: 3 3 4 1 4 1 4 1 0 4 1 0 1 1 3 5 2 4 1 1 1 1 0 1 1 2 0 0 4 2 3
1 0 0 1 1 2 1 1 0 0 4 4 1 2 4 1 1 0 1 4 0 2 3 3 1 1 1 5 0 5 1 1 0 2 0 1 0
2 0 2 1 1 2 0 1 5 1 5 0 0 0 0 0 1 1 1 0 2 1 4 0 0 0 0 0 0 1 1 1
Mean of 1st resample is 1.5

```
1.7 |
1.6 |
1.5 |O
1.4 |
```

Figure 1: First Resample (Socializing Question)

version of some of this data is in the file, accquest.xls, which is on the Web at http://userweb.port.ac.uk/~woodm/nms/ (all other software mentioned is on this Web site).

One of the questions required respondents to indicate their agreement on a 6-point scale, ranging from *not at all* to *strongly*, with the statement "I'd welcome the opportunity to socialise more with other accountants."

The responses were coded using 0 to represent *not at all* and 5 to represent *strongly*. Seven of the 105 respondents to the questionnaire failed to answer this question, leaving 98 responses. The responses ranged from 0 to 5, with a mean of 1.3.

These responses are from 98 members of a population of size 4,500. We can use the mean of this sample (1.3) as an estimate for the mean of the whole population, but the problem is to assess its accuracy.

If we had the answer to the question from all 4,500 members, we could assess the size of the error due to using samples of only 98 simply by taking a few hundred such samples from this population. In fact, of course, the whole problem is that we only have one sample of 98, so the next best thing is to use this sample to guess what the population would be like. The sample of 98 makes up about 1 in 46 of the whole population of 4,500 members, so the obvious way to do this is to make 46 copies of the sample, making a total of 4,508. For example, 4 respondents answered 5 to the question, so the *guessed population* contains 46 times as many 5s—that is, 184 of them. The guessed population will not, of course, be identical to the real population, but it is the best we can do with the data available.

We can now draw lots of random samples of 98 responses from this guessed population of 4,508 and see how much they vary. These samples are called *resamples* because they are drawn from the original sample (or more precisely, from 46 copies of the original sample).

In fact, of course, this is only a practical proposition with a computer. The process is shown in Figures 1 and 2, which are two screens from the output from the program, resample.exe. (This is available on the Web at the address above, with some notes in resample.htm. This program needs data as a text file—this is accsoc.txt.)

Figure 1 shows the first resample and is designed to show what is going on. The first randomly chosen response from the guessed population is 3, the second is also 3, the third is 4, and so on. Users can run off as many individual resamples as they want to ensure they understand what the program is doing. Each resample mean will be represented by a new *O* on the tally chart.

Figure 2 shows a similar tally chart with 10,000 similar resamples. The 2.5 and 97.5 percentiles of this distribution, as worked out by the program, are 1.02 and 1.59, which

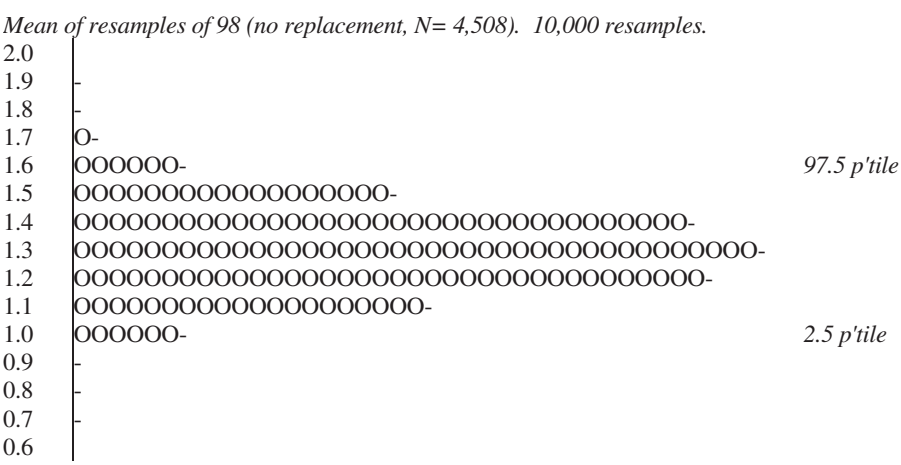


Figure 2: Resample Distribution (Socializing Question)

Note. O represents 60 resamples. - represents fewer than 60 resamples.

we can round off to 1.0 and 1.6. This means that 2.5% of the resample means are below 1.0, and a further 2.5% are above 1.6. The majority, 95%, are between these two limits. The conclusion is that there is 95% chance that the mean of samples of 98 drawn from this guessed population will lie between 1.0 and 1.6. Few were outside this range, and none, for example, was as high as 2.

However, the difficulty with this, of course, is that it is derived from a *guessed* population, not the real population. The trick now is to think in terms of the *errors* we would make in using individual resamples to find out about the guessed population. Then we can use this to make a probabilistic estimate of the error in using the real sample.

The mean of this guessed population is obviously the same as the mean of the original sample, that is, 1.3. The first resample shown in Figure 1 has a mean of 1.5. If we were to use this to estimate the mean of the whole guessed population the error would be 0.2: The estimate would be 0.2 too high. Similarly, each of the resamples with a mean of 1.0 in Figure 2 has an error of –0.3, that is, the estimate is 0.3 less than the true mean of the guessed population, 1.3.

This means that we can associate an error with each resample mean: +0.2, –0.3, or whatever. These errors are shown in Figure 3, which is identical to Figure 2 except that the numbers have been rescaled to represent the errors rather than the means themselves.

Figure 3 shows that 95% of the errors are in the range –0.3 to +0.3: We can be 95% sure that the error in the mean of a resample of 98 responses will be no more than 0.3: We will refer to this maximum likely error as the *error limit*. Errors of 0.5 or 0.6 were very rare, errors of 0.7 and above did not happen in any of the 10,000 resamples.

This is all based on 10,000 resamples from the guessed population, which are, in effect, some experiments on an imaginary population. We do not know what the real population is like, but by experimenting on an imaginary one, it is possible to infer the size of error we may make when using a single, real sample. In this case, the 95% con-

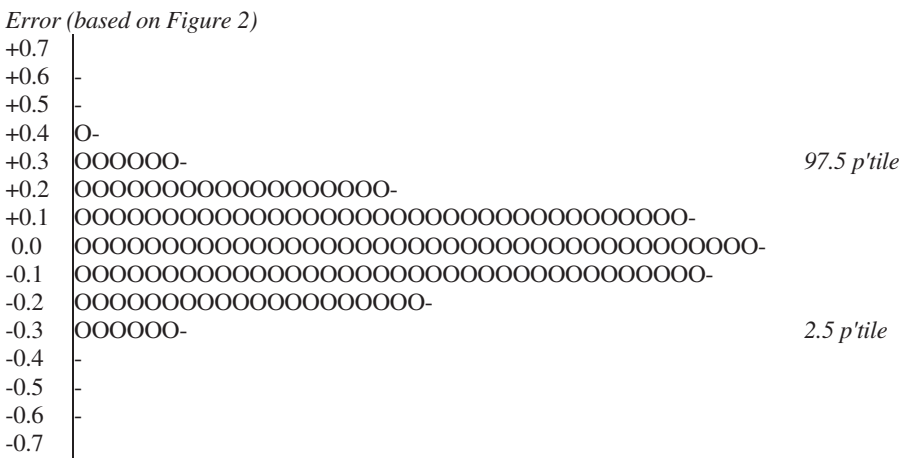


Figure 3: Resample Error Distribution (Socializing Question)
Note. O represents 60 resamples. - represents fewer than 60 resamples.

fidence bound for the error is 0.3 units. The assumption is that the guessed population is sufficiently like the real population for this estimate to be realistic. Experience shows that with reasonably large samples, this assumption is justified (see below for a more systematic analysis of the assumptions underlying the method).

The actual sample mean (which is the basis of all the resamples) is, of course, 1.3, and if the error range is +/–0.3, this means the confidence interval goes from 1.0 to 1.6. We can be reasonably (95%) certain that the truth about the average response to this question from all 4,500 members would lie within this range. We can be confident of this, despite the fact that most members have not answered the question. The range 1.0 to 1.6 corresponds to the central 95% of the resample distribution (Figure 2), so we can use the resample distribution directly as a *confidence distribution* and use it to read off confidence intervals.

In rough terms, the argument is that the resample tally chart (Figure 2), which gives a picture of the likely extent of sampling error, can also be interpreted as a confidence distribution for the population value, because the best guess is the center of the distribution, and errors are reflected by discrepancies from this central value. A standard 95% confidence interval goes from the 2.5th to the 97.5th percentile of this resample distribution.

This is the percentile bootstrap interval (Davison & Hinkly, 1997; Efron & Tibshirani, 1993). We have illustrated it for the mean, but it works for any other statistic that is calculated from a random sample of data, provided a number of assumptions—to be discussed below—are met. Both resample.exe and resample.xls will derive bootstrap confidence intervals for many other statistics including proportions (these are means if we use a code of 0 for the absence of a characteristic and 1 for its presence), the median, standard deviation and interquartile range. Resample.xls will also derive confidence intervals for statistics relating two variables—this is an important possibility that is explored in a later section.

Different Sized Populations and Infinite Populations

The above argument brings in population size in a very direct way: by constructing a guessed population of roughly the right size. It is easy to experiment with different population sizes and to show, empirically, how little difference this makes to the width of the confidence intervals.

If the population is infinite, the guessed population would also need to be infinite—which is obviously impossible to simulate on a computer. However, infinite guessed populations can be simulated by *resampling with replacement*. This means that we resample from the sample itself (without making extra copies), but each member of the resample *is replaced* in the sample before drawing the next member of the resample. Every member of the resample is drawn from the same sample—which is used to represent the guessed population. (When resampling without replacement from a finite population, as in the section above, the composition of the remaining population changes as the sample is drawn.) Resampling with replacement is a useful device for modeling the sampling process from infinite, or very large, populations.

In the case of the society of accountants, the population is not infinite, but it is large compared with the sample, so resampling with replacement is a viable method. (It is also the method to use if we are want to generalize the result to the potentially infinite population of *all* current and future accountants.) To do this, we select one value at random from the sample of 98. Now replace this in the sample, and choose another. Repeat 98 times for a resample of 98, and work out the mean of these 98 responses. Repeat this whole process a few thousand times.

There are likely to be more of some values in the resample than there are in the original sample, and fewer of others. (Otherwise, each resample would be identical to the original sample, which would be a pointless exercise.) One resample with replacement, for example, was the following:

0 0 0 0 2 5 0 2 0 0 1 2 1 2 2 2 1 0 4 2 3 5 1 2 3 2 2 1 5 1 0 0 5 2 4 4 3 0 0 1 0 0 2 4 0 0 1 4 0
1 2 0 1 4 0 0 0 0 0 0 0 1 0 0 2 1 0 1 1 0 4 1 5 4 4 3 0 0 1 4 0 0 4 1 0 2 5 3 0 0 0 1 0 1 1 0 0 3.

This contains 6 fives, whereas there were only four in the original sample. Another resample contained only 3. This procedure leads to almost the same confidence interval as for the finite population of 4,508: 1.01 to 1.59.

This illustrates one of the important, but for many people counterintuitive, results of statistics: namely, that the size of the population has little impact on the magnitude of sampling error. The width of the confidence intervals shows that a random sample of 98 provides very nearly as accurate a guide to a population of 98 million as it does to a population of 4,500. Even with a population of 196, the confidence interval was only slightly narrower (1.1 to 1.5, instead of 1.0 to 1.6) than it was for the larger populations.

In practice, resampling with replacement is easier than constructing a finite guessed population, so this is the best method to use for all but small populations. It is the method used in the rest of this article.

Table 1
Estimated 95% Confidence Intervals and Error Limits Based on Resampling From a Pilot Sample of 10

Sample Size	*Mean*	*Estimated Confidence Interval*	*Estimated Error Limit*
20	2.30	1.55 to 3.05	0.75
98	2.30	1.96 to 2.64	0.34
500	2.30	2.15 to 2.45	0.15

The Effect of Sample Size

Similarly, it is trivial to experiment to assess what would happen with different sample sizes. This is likely to be of particular interest in the design phase of a research study before serious data collection has started. Table 1 was constructed from a pilot sample of 10 responses with a mean of 2.3 (they were the first 10 responses in the data set analyzed above). However, even this small sample can be used to gauge the likely error in using different size samples. The first row, for example, used this sample of 10 to experiment with resamples of size 20 (as each item is replaced before drawing the next, the resample can be as large as we want). The error limit here is 0.75: If this is considered acceptable, then a sample of 20 may be sufficient.

How Many Resamples Do We Need?

Figure 2 is based on 10,000 resamples that yielded a 95% confidence interval extending from 1.01 to 1.59. One issue we have not considered so far is whether 10,000 is a suitable number of resamples to generate.

A second set of 10,000 resamples gave 1.02 to 1.60, and a third run of 100,000 resamples gave 1.01 to 1.60. This suggests that with these large numbers of resamples the results are consistent. From this perspective, 10,000 would seem sufficient. Furthermore, as 10,000 resamples took only 20 seconds to generate, there seems little point in using a smaller number of iterations. The spreadsheet software discussed in the next section is slower, but even with this, 1,000 resamples is reasonably easy.

The most helpful practical advice is to experiment with relatively small numbers of resamples. Then, for the final analysis, generate as many resamples as is practical. Then repeat the analysis to check that the results are reasonably stable.

Bootstrapping by Spreadsheet and Other Software Options

Lunneborg (2000, pp. xv-xvi, 556) gave a list of some of the software options for resampling. It is possible to use standard statistical packages such as SPSS for resampling, and there are also specialist tools available. However, they all require some programming by the user.

In this article, I have used two much simpler programs—both available at http://userweb.port.ac.uk/~woodm/nms/. The first, resample.exe, is a menu-driven, stand-alone program. It requires no programming, but the range of statistics it will deal with is limited: It cannot, for example, cope with any of the examples in the next section. On the other hand, it is easy to use and fast—I have just run off 100,000 resamples of size

12 in about 1 minute. There is some further information in resample.htm at the same Web site.

The second program is an Excel spreadsheet, resample.xls. This requires the user to paste or key in some data (one or two variables) and then enter a resample size and an Excel formula for the resample statistic (e.g., *=average(F7:F104)* for the example above). The Read this worksheet explains how to use the spreadsheet—the second example described is a standard bootstrap confidence interval. It is reasonably easy to modify: Nothing is hidden or locked, and there are no macros.

Both programs are adequate for simple analyses. They are intended to demonstrate the potential of resampling, not for extensive data analysis (they have no facilities for dealing with missing data, for example). As well as deriving confidence intervals, they also show exactly how resampling works. In each case, you will need to edit out missing data, then analyze each variable or pair of variables (see next section) separately.

Resample.xls also has the advantage of the flexibility of a spreadsheet: The resample statistic does not have to be a standard one. I have extended the capability of Excel slightly by defining several user-defined functions: These are in the Add-in nms.xla (at the same Web site). To install this, download it and then, in Excel, click Tools—Add-ins and Browse to find the file.

Bootstrap Confidence Intervals for Measures of the Relationship Between Two Variables

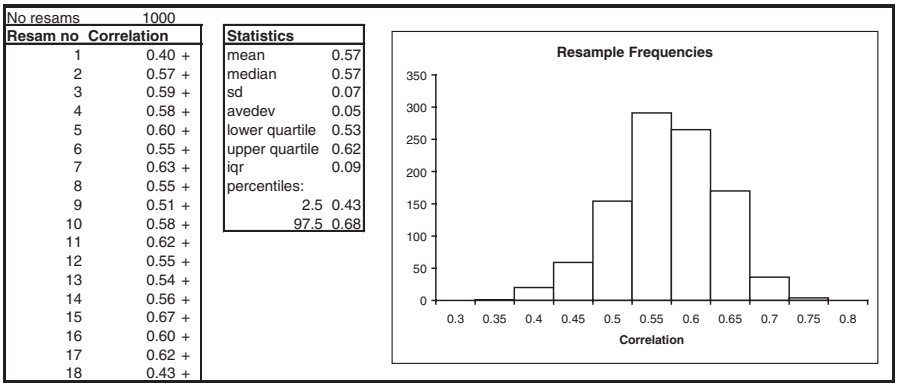
In practice, researchers are usually interested not so much in a single variable as in the relationship between two or more variables. Mean responses to questionnaire scales, for example, are generally only useful if there is something to compare them with. In terms of the hypothesis testing paradigm, we would set up a null hypothesis of no difference, or no relationship, and then derive a *p* value to see how consistent the data is with this hypothesis.

The equivalent in the confidence interval paradigm is to set up a confidence interval for a measure of this difference or relationship. The power of the bootstrap approach is that we can use exactly the same method as we used for the mean of a single variable.

I will give two examples, then list the general possibilities for two variables. First, the relationship between age, and the response to the question about socializing with other accountants (in the data from the society of accountants described above), can be analyzed in various ways, but one obvious approach is to compare the mean response to the question from young members (up to 35) and old members (55 and older). The mean from the young members was 2.0, and from the old members was 1.0: the difference of 1.0 indicating a difference in attitude between the two groups. The difficulty, of course, is that this conclusion is only based on a small sample, so we need to know whether it can reasonably be assumed to apply to members in general. The conventional method of doing this is to set up the null hypothesis of no difference and then derive a *p* value: the *Compare means* procedure in SPSS gives a *p* value (based on ANOVA) of 4%.

To bootstrap a confidence interval here, we use the difference of the two means as the statistic we calculate from each resample. To use resample.xls,

1. Starting from the data in accquest.xls, after eliminating any individuals who have not answered the questions about socialising and age, we are left with 46 individuals in age



No resams	1000
Resam no	**Correlation**
1	0.40 +
2	0.57 +
3	0.59 +
4	0.58 +
5	0.60 +
6	0.55 +
7	0.63 +
8	0.55 +
9	0.51 +
10	0.58 +
11	0.62 +
12	0.55 +
13	0.54 +
14	0.56 +
15	0.67 +
16	0.60 +
17	0.62 +
18	0.43 +

Statistics	
mean	0.57
median	0.57
sd	0.07
avedev	0.05
lower quartile	0.53
upper quartile	0.62
iqr	0.09
percentiles:	
2.5	0.43
97.5	0.68

Figure 4: Top of Lots of Resamples Sheet Showing Resample Results

category 1 (35 and younger) or 4 (55 and older). Paste this data into the Sample sheet with responses to the socializing question as Variable 1 and age (1 or 4) as Variable 2.

2. Now go to the Single resample sheet and tell the program that the resample size is 46 and the resample statistic is

   =diffofmeans(F7:F52,G7:G52)

   This is one of the functions in the Add-in nms.xla, so you must have this installed. It calculates the mean response from (simulated) individuals with Age = 1, and the mean from those with Age = 4, and subtracts the second from the first. Pressing F9 will recalculate the spreadsheet and produce another resample.
3. Finally, go to the Lots of resamples sheet to see the results. These include a graph and the 2.5th and 97.5th percentiles, which give the standard 95% confidence interval.

Alternatively, the workbook diffofmeansconfidence.xls at the same Web site is set up specifically to work out a confidence interval for the difference of two means—go to the Read this sheet for instructions.

Both workbooks are set to work with 200 resamples. To work out this result, I changed this to 1000 (see the Read this worksheet): The resulting 95% confidence interval extended from

0.1 to 1.9.

This appears on the Lots of resamples sheet—Figure 4, based on the next example, illustrates the format. Pressing F9 to recalculate the spreadsheet produced only small changes in this interval: Neither end changed by more than 0.1.

This is roughly consistent with the p value of 4% because zero is just outside this interval. (We should not expect exact correspondence between the two approaches because they are based on different models.) However, the confidence interval gives you more information, and its interpretation is more straightforward: We are not sure of the exact population difference, but the sample data suggests we can be 95% confident that it is somewhere between 0.1 and 1.9.

As a second example, I have taken the correlation between the question on socializing and a question on the importance of networking. We would expect a strong positive correlation between these two responses, and this is in fact what we find: The correlation is 0.57.

Working out a bootstrap confidence interval for this correlation with resample.xls is now straightforward. The resample size is now 93 (i.e., the sample size after eliminating missing data), and the resample statistic is

=correl(F7:F99,G7:G99).

Figure 4 shows the results: the 95% confidence interval extends from

0.43 to 0.68.

In this example, with a largish sample and a fairly symmetrical resampling distribution, the bootstrap percentile interval is likely to give reasonable results (the underlying assumptions are discussed in the next section): This is the approach used in Diaconis and Efron (1983), although there are more advanced methods that are likely to give better results (Lunneborg, 2000).

As SPSS, or statistical tables, will confirm, the *p* value for the null hypothesis of no correlation is very low. However, given the nature of the questions, this null hypothesis is too silly to be worth taking seriously; the confidence interval gives a much more useful conclusion.

As an alternative to the standard (Pearson) correlation coefficient, we could use the nonparametric Kendall coefficient. Again the same bootstrap method is feasible: There is a function, *kendall*, in the Add-in nms.xla (although it is slow and may take a long time with anything but small data sets).

Some other possibilities for analyzing the relation between two variables are shown in Table 2.

Resample.xls can be used to derive bootstrap confidence intervals for each of the five model types in this table. We have just seen examples of Model Types 2 and 3. If the two values of the category variables are coded as 1 and 0, the difference of the proportions (Model Type 1) is simply the difference of the means, and we can use the formula above. For Model Type 4, we need to create a new variable—the difference of the two original variables: then we simply derive a confidence interval for this. For Model Type 5, we can use the Excel formula for the regression coefficient—*slope* (check Help for details). In principle, multiple regression models could be analyzed in the same way, but there are no built in Excel functions for the coefficients in multiple regression models. (Excel will perform multiple regression, but not by means of functions in worksheet cells.)

Confidence intervals can be derived by methods based on probability theory. SPSS, for example, will produce confidence intervals for means, correlations, regression coefficients, and so on. However, the bootstrap approach has the three advantages outlined in the introduction. The one general method will fit a wide range of problems—all those in Table 2, for example. Furthermore, understanding the rationale behind this one method depends on a bare minimum of technical expertise—there are no standard deviations, standard errors, variances, or statistical tables in the explanation of the percentile interval. There are also no appeals to the central limit theorem or other theo-

Table 2
Methods of Analyzing Relationships Between Two Variables

Model Type	*Variable 1*	*Variable 2*	*Standard Hypothesis Test*	*Relationship Statistic*
1	Category (2 values)	Category (2 values)	Fisher exact/Chi square	Difference of proportions
2	Number	Category (2 values)	Unpaired *t* test/ANOVA	Difference of means
3	Number	Number	Test of correlation = 0	Correlation
4	Number	Number	Paired *t* test	Mean difference between variables
5	Number	Number	Test of regression = 0	Regression coefficient

rems of mathematical statistics, and there are no checks for assumptions such as normality of the data. Instead, there is a simple process, implemented by computer, which allows the direct construction of the confidence intervals from simulation results. In addition to this, the generality of the bootstrap approach means it can be applied to problems where conventional solutions would be difficult to find (e.g., analyzing the difference of two correlations).

Assumptions Underlying the Percentile Interval

The percentile method is the most straightforward approach to bootstrapping, but there are sometimes problems. A variety of more sophisticated approaches have been devised (see, for example, Davison & Hinkley, 1997; Efron & Tibshirani, 1993; Lunneborg, 2000) to overcome some of these problems, but the difficulty with many of these is that the simplicity of the method is lost. Many of these methods do depend on far more sophisticated statistical concepts than the percentile interval does. This raises the question of when it is reasonable to use the percentile method and when it is necessary to make use of more advanced methods.

This section outlines the assumptions on which my derivation of percentile confidence intervals depends. This should help the reader to appreciate how reasonable it is to use the percentile interval in a given situation. This section also avoids mathematical notation—although some of the assumptions are subtle, in Efron and Tibshirani's (1993) words, "At least the difficulties are the appropriate conceptual ones, and not mathematical muddles" (p. xiv). The aim is to clarify the circumstances in which it is reasonable to use the percentile interval.

All of these assumptions include the word *reasonable*, implying that some judgment is required. Such judgments are almost inevitable in modeling real-world phenomena—for example, judgments about whether distributions are reasonably normal in conventional statistics. It is possible to research the accuracy of these judgments in various ways, but then new judgments may be needed to gauge whether the results apply to a new real-world situation.

Assumption 1: Sampling process modeled reasonably accurately—normally the assumption is that the sample is a random sample from a wider population.

Obviously random resampling simulates a random sampling process.

Another possibility is that the sample may be stratified: For example, the samples of old and young members of the accounting society may have been sampled separately so that the number of each was known in advance. In this case, obviously, the resampling method needs to be mimic this. It is fairly easy to modify resample.xls to do this (by changing the formulae in cells E7:E52 in the Single resample sheet): The change, in terms of the results, is slight: The 95% confidence interval, expressed to one decimal place, was identical to the one given above.

In terms of our society of accountants, we do not, of course know if we can reasonably regard the sample as random: There are a number of obvious checks we could run. However this is not just an issue for bootstrap confidence intervals: The issue arises in just the same manner in a conventional analysis. In either case, the question is whether an analysis based on the assumption of random sampling is good enough to be useful.

Assumption 2: Guessed population reasonably "similar" to the real population.

The idea of bootstrapping is to use estimates of sampling error derived from the guessed population to make inferences about the real population, so "similarity" has to be judged from this point of view: This is obviously a subtle judgment! (If the two populations were known to be identical, the inference problem would be solved and confidence intervals would be unnecessary.) Any confidence interval clearly needs some model of the underlying population distribution: Conventional parametric statistics achieves this by making (often questionable) assumptions about the form of the distribution of the population parameter.

Almost by definition, it is very difficult to assess the extent to which this assumption is satisfied: The whole problem of confidence interval estimation is that the parent population is unknown. Dissimilarities between the two populations may mean that the accuracy of the resulting confidence intervals is disappointing. For example, Efron and Tibshirani (1993, p. 175) drew 300 samples of 10 from a standard normal distribution and worked out 95% bootstrap percentile intervals for each of them. They found that 10% of the intervals failed to include the population mean—twice the intended value of 5%. They explain this by pointing out that "the percentile interval has no knowledge of the underlying normal distribution and uses the empirical distribution in its place. In this case, it underestimates the tails of the distribution."

Using more sophisticated methods to derive the guessed population may help to obtain reasonable results from small samples: For example, we may use the sample to generate a normally distributed guessed population or to smooth the distribution in some way (Lunneborg, 2000, pp. 90-96). In practice, these more subtle methods of deriving a guessed population are unlikely to be necessary if the statistic we are analyzing is the mean, or the sample is reasonably large.

Assumption 3: Unbiased estimate—the value of the statistic derived from the sample data is reasonably similar to the corresponding statistic derived from the guessed population.

The argument above that allowed us to go from Figure 2 to Figure 3, and then to argue that Figure 3 can be regarded as a confidence distribution, rested on the assumption that the mean of the sample (1.3) is the same as the mean of the guessed population. In this case, and in many other cases, this is obviously true. To take another example, the correlation from the data in the example illustrated by Figure 4 was .57, which was the same as the correlation from a resample of size 50,000—which we can take as a surrogate for an infinite population.

There are circumstances in which this assumption is not satisfied. For example, if we wanted to estimate a bootstrap confidence interval for a population standard deviation from a small sample (say: 3, 3, 4, 10), the obvious statistic to use is the unbiased estimator, *s* or *stdev* in Excel (the formula with $n - 1$ in the denominator): The sample value of *s* is 3.4. If the guessed population comprises a very large, or infinite, number of copies of the original sample, the standard deviation of this population can be worked out by applying the formula *stdevp* (the formula with *n* in the denominator) to the sample: This comes to 2.9. This is less than the value of *s* for the sample: This difference means that the simple argument above for using the percentile interval does not quite work. However, this is a decidedly artificial example: In real research, we would always take a bigger sample to estimate *s*, and with larger samples the differ-

ence between the two formulae is too small to worry about, and so we can assume this assumption is satisfied.

Assumption 4: Symmetry—Assumption 3 holds, and the distribution of resample statistics is reasonably symmetrical about the value of the statistic derived from the data.

If the distribution were not symmetrical about the value derived from the data, it would not be reasonable to ignore the distinction between positive and negative errors as we did above in the section on the percentile bootstrap interval for finite populations. This is easily checked. Figures 2 and 4 look roughly symmetrical about the sample values (1.3 and 0.57).

The problem that arises if the distribution is not symmetrical can be seen by analyzing the relationship between Figures 2 and 3. A resample with a mean of 1.0, for example, is 0.3 too *low* as an estimate of the mean of the guessed population, and the tally chart gives an indication of how many resamples suffer from this error. Now, if the *real* sample mean were 0.3 too low (and remember we are using the resamples to experiment to see what the likely pattern among real samples is), this suggests that the real population mean is 0.3 *more* than the sample mean (i.e., 1.6), so the probability that the real population mean is 1.6 (0.3 *above* the mean) is estimated by the probability that the resample mean is 1.0 (0.3 *below* the mean). As the distribution is symmetrical this reversal makes no difference, but if it were asymmetrical the reversal would have to be taken into account, and we could not use the resample distribution as the confidence distribution.

Even if the resample distribution is not symmetrical, there may still be a good case for using it to derive confidence intervals (Efron and Tibshirani, 1993, chap. 13). However, this is using a different, more subtle, argument from the argument presented above.

Assumption 5: Error distribution independent of the true parameter value—it is reasonable to talk of the error distribution independently of the true value of the population parameter.

To see the difficulty here, consider the confidence interval derived from the proportion of respondents in the sample of accountants who attended the annual dinner—this was 12 out of 82 (15%). One of the resample proportions was 30.5%. Running through the argument for the percentile interval above, this resample has an error, compared with the true proportion in the guessed population (15%), of +15.5%: It is 15.5% too high. This suggests that the actual sample value, 15%, may suffer from a similar error, in which case the true population proportion would be *minus* 0.5%. This is obviously impossible. Similarly, a resample proportion of 30% would imply a true proportion of 0%—which is also impossible, because there would be no 1s in the population so the observed 15% could not have happened. Another of the resample proportions was 29%: By a similar argument, this corresponds to a true population proportion of 1%. But if the population proportion were 1%, the resample distribution would be far more skewed than the actual resample distribution based on the sample. The bootstrap idea is to assume the same error distribution (like Figure 3) applies to all possible values—in the case of Figure 3, this is reasonable between the 2.5 and 97.5 percentiles, but not at the extremes of the distribution.

Problems are likely to arise with Assumption 5 if the resample distribution suggests that a boundary value—zero in this case—is a possibility. In these cases, the resample distribution is likely to be skewed by this boundary (30.5% is possible but –0.5% is not), so Assumption 4 will fail.

Assumptions 2 to 5 are often satisfied to a reasonable degree. Figures 2 and 4 are both based on a large enough sample to give a reasonable picture of the population, the sample estimates are unbiased, the distributions are roughly symmetrical, and there are no boundaries values that would make Assumption 5 problematic. Assumption 1—the randomness of the sample—is more problematic as the respondents to the survey may be a biased sample, but this problem arises whatever statistical methods are used.

If Assumptions 2 to 5 are not satisfied to a reasonable degree, there are many further avenues for making bootstrap estimates more realistic, although, as noted above, some of these are far more difficult to explain and justify than the percentile interval. These more advanced methods would normally be advocated for a correlation, for example. However, the analysis here suggests that a percentile interval based on Figure 4 is reasonable, although this may not be true if the sample were smaller, or the correlation was closer to one (which would result in the resample distribution being skewed). The percentile interval is often, but not always, adequate.

Conclusions

This article has shown how bootstrap confidence intervals can be used for statistical inference. They can be applied to estimates of a very wide variety of population parameters, including those that summarize the relationship between two variables—such as a difference of the means of two subgroups and correlation and regression coefficients. They can also be applied to finite populations and could easily be adapted to model stratified sampling schemes.

This article has also demonstrated the rationale behind the method and its dependence on five assumptions. It is worth pausing to note what has *not* been mentioned in the above derivations. There are no tabulated probability distributions, no central limit theorem, no standard deviations and variances. The arguments involve an understanding of frequency distributions and percentiles, averages and the other statistics analyzed, and that is about it. Furthermore, the one argument suits all statistics. The amount that users need to know to understand the rationale behind the methods is far less than with conventional methods.

When should these bootstrap confidence intervals be used? We have seen how they are applicable for a wide variety of statistics—including the mean, median, proportions, and the statistics used to summarize relationships between two variables in Table 2. But there are more conventional approaches to these problems, so why choose bootstrap confidence intervals?

The argument depends first on the case for using confidence intervals instead of adopting the hypothesis testing paradigm. This is outlined briefly in the second section, although it is not the focus of this article. If this case is not accepted, there are methods for testing null hypotheses based on resampling, which have very similar advantages in terms of conceptual economy to the bootstrap methods for deriving confidence intervals (Lunneborg, 2000; Noreen, 1989). These methods can be imple-

mented on a spreadsheet in a very similar manner to the bootstrap methods (Wood, 2003).

If the case for confidence intervals is accepted, these may be derived using the conventional formulae of probability theory or by bootstrapping. In many situations, either approach can be used and the answers will be similar. The main advantages of the bootstrap method are its transparency and conceptual economy, and its applicability in contexts where there is no convenient, conventional method.

Alternatively, bootstrapping can be used simply as a means of gaining insight into the meaning of answers obtained by conventional methods—which is important, both for producers, and readers, of research studies.

References

Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their application*. Cambridge: Cambridge University Press.

Diaconis, P., & Efron, B. (1983, May). Computer intensive methods in statistics. *Scientific American, 248*, 96-108.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, *7*(1), 1-26.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.

Gardner, M., & Altman, D. G. (1986, March 15). Confidence intervals rather than P values: Estimation rather than hypothesis testing. *British Medical Journal*, *292*, 746-750.

Hall, P. (1992). *The bootstrap and Edgeworth expansion*. New York: Springer-Verlag.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement*, *56*(5), 746-759.

Lindsay, R. M. (1995). Reconsidering the status of tests of significance: An alternative criterion of adequacy. *Accounting, Organizations and Society*, *20*(1), 35-53.

Lunneborg, C. E. (2000). *Data analysis by resampling: Concepts and applications*. Pacific Grove, CA: Duxbury.

McGoldrick, P. M., & Greenland, S. J. (1992). Competition between banks and building societies. *British Journal of Management*, *3*, 169-172.

Mooney, C. Z., & Duval, R. D. (1993). *Bootstrapping: A nonparametric approach to statistical inference*. London: Sage.

Morrison, D. E., & Henkel, R. E. (1970). *The significance test controversy*. London: Butterworths.

Noreen, E. W. (1989). *Computer intensive methods for testing hypotheses*. Chichester, UK: Wiley.

Royall, R. (1997). *Statistical evidence: A likelihood paradigm*. London: Chapman & Hall.

Russell, C. J., & Dean, M. A. (2000). To log or not to log: Bootstrap as an alternative to the parametric estimation of moderation effects in the presence of skewed dependent variables. *Organizational Research Methods*, *3*(2), 166-185.

Morrison, D. E., & Henkel, R. E. (1970). *The significance test controversy*. London: Butterworth's.

Smithson, M. (2000). *Statistics with confidence*. London: Sage.

Wood, M. (2003). *Making sense of statistics: A non-mathematical approach*. Basingstoke, UK: Palgrave.

Michael Wood has degrees in mathematics, philosophy of science, and education and is currently a teacher and researcher in the Business School at Portsmouth University, United Kingdom. His interest include quantitative and qualitative research methods and decision analysis.