Bootstrap Confidence Intervals for the Correlation Coefficient

Homer S. White

12 November, 1993

1 Introduction

We will discuss the general principles of the bootstrap, and describe two applications of the bootstrap to constructing approximate confidence intervals for an unknown parameter. We will illustrate these methods, together with the well-known Fisher z-transform, to find confidence intervals for the correlation coefficient of some bivariate normal data taken from our text (Anderson [1]). Finally, we will give the results of some simulations designed to compare the coverage properties of the bootstrap confidence intervals with those obtained by the Fisher transform.

2 The Idea of the Bootstrap

Say that we have a sample $\vec{x} = (x_1, \ldots, x_n)$, where the x_i are drawn from some unknown distribution F_X.[1] We want to estimate an unknown parameter θ of F_X (hereafter abbreviated F). For example, θ could be the mean or the median of F, or the correlation coefficient of F if F is bivariate. If we wish, we may write $\theta = \theta(F)$ to indicate that θ depends on F.

We estimate θ by means of a statistic $\hat{\theta}$, the value of which depends on our sample $\vec{x}$. Thus we may write $\hat{\theta} = \hat{\theta}(\vec{x})$. For example, if θ is the mean of F, then $\hat{\theta}$ may be the average of the sample values $x_1, \ldots, x_n$.

Part of our job is to get some idea of the accuracy of θ. We would be helped greatly in this if we knew the true distribution of $\hat{\theta}$, denoted $F_{\hat{\theta}}$. Of course, $F_{\hat{\theta}}$ depends on the original distribution F, so we can (and indeed will) write it as $F_{\hat{\theta}}(F)$, even though the notation is starting to look awful. If we know $F_{\hat{\theta}}$ exactly, then we can go on to compute the variance of $\hat{\theta}$, construct exact confidence intervals for θ, and so on. In some cases, as when we are estimating the mean of a distribution that we know is normal (or know is exponential, or some such thing), we can compute $F_{\hat{\theta}}$. Quite often, however, we do not know F and are not willing to assume much about it. It is also not usually feasible to get more samples from F in order to see how $\hat{\theta}$ changes from sample to sample. How, then, are we to approximate $F_{\hat{\theta}}$?

Well, we do at least have our sample $\vec{x}$. This sample determines the *empirical distribution* $\hat{F}$ defined by

$$\hat{F}_{\vec{x}}(t) \stackrel{\text{def}}{=} \frac{\#\{x_i \mid x_i \leq t\}}{n},$$

[1] The vector notation is not meant to imply that we are doing multivariate analysis, since the x_i themselves may be either univariate or multivariate.

where $\#A$ denotes the number of elements in the set a. We know that the empirical distribution $\hat{F}_{\vec{x}}$ (written $\hat{F}_n$ when we want to emphasize dependence on the sample size, or just $\hat{F}$ for short) is usually a pretty decent approximation of F, especially as the sample size n goes up. In fact, we have:

Theorem 1 *With probability 1,*

$$\lim_{n\to\infty} \sup_{-\infty<t<\infty} \mid \hat{F}_n(t) - F(t) \mid = 0.$$

Much stronger results are known. See Serfling ([4], Section 2.1) for more information.

Since $\hat{F}$ is known to us, the distribution of $\hat{\theta}$ when we consider $\hat{F}$ to be the original distribution from which samples are drawn, can in principle be computed exactly. We will write this distribution as $F_{\hat{\theta}}(\hat{F})$, without apology.

We know from Theorem 1 that $\hat{F}$ is liable to be close to F. It seems plausible, therefore, that $F_{\hat{\theta}}(\hat{F})$ is liable to be close to $F_{\hat{\theta}}(F)$, the true distribution of $\hat{\theta}$ which we seek. We can, in principle, get hold of $F_{\hat{\theta}}(\hat{F})$. The foregoing suggests the

Bootstrap Principle: *Approximate* $F_{\hat{\theta}}(F)$ *by* $F_{\hat{\theta}}(\hat{F})$.

We can see now the reason for the name "bootstrap". One gets an idea of the distribution of $\hat{\theta}$ by using only the original sample from which $\hat{\theta}$ was computed in the first place!

Of course, it is not usually feasible to compute $F_{\hat{\theta}}(\hat{F})$ precisely. In so doing we would have to compute $\hat{\theta}(\vec{y})$ for each sample $\vec{y}$ of size n that could be drawn from the original data $x_1, \ldots, x_n$. Unfortunately, the number of possible samples $\vec{y}$ is C_n^{2n}, the number of ways of choosing n objects from $2n$ objects. This latter quantity grows rapidly with n.

Since we do not have the time to compute the value of $\hat{\theta}$ for all possible samples from $\hat{F}$, we usually perform Monte-Carlo simulation. That is, we choose some large number B ($B = 2000$ usually suffices), and we repeat the following process B times: draw at random, with replacement, a sample $\vec{y}$ from $x_1, \ldots, x_n$, and compute $\hat{\theta}^\star = \hat{\theta}(\vec{y})$. The values $\hat{\theta}^\star$ thus generated determine an empirical distribution called $\hat{F}_{\hat{\theta}}(\hat{F})$, in order to emphasize the dependence on the original empirical distribution $\hat{F}$ from which the $\vec{y}$ were chosen.[2]

By Theorem 1, we know that $\hat{F}_{\hat{\theta}}(\hat{F})$ gets closer to $F_{\hat{\theta}}(\hat{F})$, as $B \to \infty$. Hence for large enough B we may take $\hat{F}_{\hat{\theta}}(\hat{F})$ as an approximation of the desired true distribution of $\hat{\theta}$.

Figure 1 (appended) illustrates these ideas. Arrows in the figure represent approximations.

The advantage of the Bootstrap Principle is that it applies to a wide variety of statistical problems, with no assumptions on the underlying distribution. Its main disadvantage is the non-trivial amount of computation involved.

3 Bootstrap Confidence Intervals

Numerous methods have been suggested for constructing bootstrap confidence intervals. Many of them are described in Efron and Tibshirani [2]. We will discuss the most naive of these, the so-called *percentile* method, and one of the most advanced: *bias-corrected and accelerated* (or BCA for short).

[2] OK, we apologize for *this* notation.

3.1 The Percentile Method

Let's say that we want a $1-\alpha$ confidence interval for θ. (For example, $\alpha = .10$ denotes 90% confidence.) We can take the middle $100(1-\alpha)$ percent of $\hat{F}_{\hat{\theta}}(\hat{F})$. That is, we sort the B values of $\hat{\theta}(\vec{y})$ that were obtained during the Monte-Carlo simulation, and throw out lowest $B\alpha/2$ values, as well as the highest $B\alpha/2$ values. Our confidence interval will have as its lower (respectively, upper) limit the lowest (respectively, the highest) remaining value. More precisely, the interval we construct is

$$(\hat{\theta}^{\star}_{\alpha/2}, \hat{\theta}^{\star}_{1-\alpha/2})$$

where $\hat{\theta}^{\star}_{\gamma}$ denotes the 100γ percentile point of $\hat{F}_{\hat{\theta}}(\hat{F})$.

The reader should bear in mind that other, more sophisticated, methods go by the same name. See Hall [3] for more details.

3.2 Bias-Corrected and Accelerated

It may be helpful to begin by reviewing the standard approaches to constructing confidence intervals:

- by a direct application of the central limit theorem, or
- by applying appropriate transformations and then appealing to the CLT.

When the distribution of $\hat{\theta}$ is asymptotically normal (as is the case when $\hat{\theta}$ is the sample average and the underlying distribution has finite second moment, or when $\hat{\theta}$ is the sample correlation coefficient and the underlying distribution is multivariate normal), one may form an approximate $1-\alpha$ confidence interval for θ as:

$$(\hat{\theta} - z_{\alpha/2}\hat{\sigma_{\hat{\theta}}}, \hat{\theta} + z_{1-\alpha/2}\hat{\sigma_{\hat{\theta}}})$$

where $\hat{\sigma_{\hat{\theta}}}$ denotes the square root of an estimate of the variance $\sigma^2_{\hat{\theta}}$ of $\hat{\theta}$, and z_γ denotes the 100γ percentile point of the standard normal distribution. Of course, such confidence intervals are not exact since the distribution of $\hat{\theta}$ is not exactly normal.

In some cases we are able to apply a transformation ϕ to the estimator $\hat{\theta}$ which is chosen so that the distribution of the new random variable, $\phi(\hat{\theta})$, is closer to normal than is the distribution of $\hat{\theta}$. We then form a confidence interval for $\phi(\theta)$ in the usual way, and apply ϕ^{-1} to the endpoints of this interval in order to obtain a confidence interval for θ. Such intervals are more accurate than those formed by applying the CLT directly.

In the case of sample correlation, we have the well-known z-transform, first suggested by R.A. Fisher. This transformation is defined by:

$$z = \phi(r) = \frac{1}{2}\log\left(\frac{1+r}{1-r}\right),$$

where r is the sample correlation coefficient. Not only is F_z closer to normal than is F_r, but also the variance of z is essentially independent of the unknown parameter ρ, the true correlation. Such transformations are said to be "variance-stabilizing".

Of course, the transformation z is not perfect. F_z is still a bit skewed. Also, recall ([1], p. 123) that

$$\mathcal{E}(z) = \mathcal{E}(\phi(r)) \approx \phi(\rho) + \frac{\rho}{2n},$$

so that z is somewhat biased. (A transformation ϕ is said to be *unbiased* if $\mathcal{E}(\phi(\hat{\theta})) = \phi(\theta)$.)

The idea for the BCA method begins with a fantasy: suppose we knew of a transformation ϕ that was ideal, in the sense that for any value of the parameter θ, $F_{\hat{\theta}}$ was normal. We would certainly be interested in making use of such a transformation. However, ϕ could well be biased, and both the bias and the variance $\sigma^2_{\phi(\hat{\theta})}$ of $\phi(\hat{\theta})$ probably depend on the unknown parameter θ (through the distribution of $\hat{\theta}$, which depends on θ).

In order to make good use of ϕ, we would need to know something about $\sigma^2_{\phi(\hat{\theta})}$ so that we could construct a confidence interval for $\phi(\theta)$. We would also need to know something about the bias of ϕ in order to correct for it, before we apply ϕ^{-1} to obtain a confidence interval for θ itself. All of this seems pretty much a pipe-dream, since we don't know ϕ at all.

UN-deterred, the bootstrapper fantasizes further, and supposes that this ideal ϕ also has the property that $\sigma_{\phi(\hat{\theta})}$ (which, as we saw, depends on θ), is actually a linear function of $\phi(\theta)$. It then becomes possible for the bootstrapper to obtain estimates of the bias and of $\sigma_{\hat{\theta}}$ *without knowing* ϕ *at all*! Details of this process appear in Chapter 22 of [2]. The reader should note that the assumption of linearity is reasonable, because a smooth function may always be approximated locally by a linear one.

4 An Illustration

In these days of powerful machines and slick software, the computational aspects of the bootstrap are not overly burdensome. The appended program, written in the SAS/IML language, constructs confidence intervals for the correlation of given bivariate data, using all three of the methods outlined so far:

1. normal-theory enhanced by the z-transform;
2. the bootstrap percentile;
3. Bias-Corrected and Accelerated.

In our program, the **boot** module constitutes the "engine" for the bootstrap approach. It is quite general, and can be used for any data, univariate or multivariate, provided that it calls the appropriate function to compute $\hat{\theta}$ for the sample drawn from $\hat{F}$. In our situation, the appropriate call is to the **corr** function module, which computes the sample correlation. If you were interested in estimating the mean of a distribution, you would want to replace the call to **corr** with a call to a function module that computes the sample average. Similar remarks apply to the **bca** module, which actually constructs the BCA confidence interval. The **bper** module, which constructs the bootstrap percentile interval, can be used without modification in other programs (of course, you have to feed it the sorted $\hat{\theta}^\star$ values obtained from **boot**. Feel free to use any of these routines in your own work.

For the sake of illustration, the program applies these methods to the data on head-length and head-breadth that is given in Table 3.3 of [1] (first son only). Results are given following the program listing.

5 Performance of the Bootstrap

Consider a method for generating a $1-\alpha$ confidence interval $I_n^\alpha(\vec{x})$ from a sample $\vec{x}$ of size n. The method may be regarded as an interval-valued random variable, through the dependence of the interval on the sample $\vec{x}$. Such a method is said to be *r-order accurate* if

$$P(\theta \in I_n^\alpha) = 1 - \alpha + O(n^{-\frac{r}{2}}),$$

for all values of α. The BCA method produces confidence intervals that are second-order accurate no matter what the initial distribution may be. (A proof of this fact, based on the Edgeworth expansion, may be found in [3].) The method based directly on the CLT is usually only first-order accurate (although one may do better if the underlying distribution is symmetric and close to normal). The bootstrap percentile method cannot lay claims to the same level of performance as the BCA, but it still performs acceptably in simulations, and is easier to explain to the boss than is the BCA.

The BCA performs better asymptotically. However, this improved performance may only be apparent for larger values of n. What about the smaller sample sizes that are encountered sufficiently often in practice as to rise to the demand for better confidence intervals? To compare the two bootstrap methods with Fisher's z for the case of sample correlation, I wrote a simulation program in SAS/IML. (By the way, now that I've run these programs, I cannot recommend SAS as an environment for statistical simulation. Try MatLab.) Results are shown in the appended tables.

Here is how to interpret these tables.

The top table refers to results obtained from a bivariate distribution of the form

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{bmatrix} 1 & 2 \\ -3 & 1 \end{bmatrix} \begin{pmatrix} W_1 \\ W_2 \end{pmatrix},$$

where W_1 and W_2 are $iid \sim N(0,1)$. The true correlation ρ of X_1 and X_2 is about -.14 — not much at all.

For each of the given sample sizes n, I did the following 500 times: I got a sample of size n from the bivariate distribution, and constructed 90% confidence intervals for ρ using each of the three methods we have discussed. For each interval, I checked to see whether or not it contained ρ. If it did not contain ρ, I called it a *lemon*, and updated a counter which recorded the total number of lemons seen so far in confidence intervals generated by that method.

Lemon-counts for each of the methods are given in the final three columns of the table. A method that produces perfect 90% confidence intervals should produce about 50 lemons, give or take 7 or so, in 500 tries. The BCA performs better than I would have expected at smaller sample sizes, but it cannot be said to have beat out Fisher's z, which is playing on its home turf, and is a mighty good method anyway.

The second table works just as above except that the matrix used to produce the bivariate distribution is

$$\begin{bmatrix} 1 & -1 \\ -.8 & 1.2 \end{bmatrix}$$

which gives a value for ρ that is very nearly -1.

References

[1] T. W. Anderson. *An Introduction to Multivariate Statistical Analysis.* 2nd. Edition. John Wiley & Sons, New York, 1984.

[2] B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap.* Chapman & Hall, New York, 1993.

[3] P. Hall. *The Bootstrap and Edgeworth Expansion.* Springer-Verlag, New York, 1992.

[4] R. J. Serfling. *Approximation Theorems of Mathematical Statistics.* John Wiley & Sons, New York, 1980.

6 A Conversation About The Bootstrap by Homer White

The so-called *bootstrap* is a powerful tool for statistical inference that was proposed in 1978 by Bradley Efron at Stanford. Since then many versions of the bootstrap have been developed to tackle a wide array of problems in statistics. Here, in a nutshell, is the core idea of the bootstrap.

Suppose that you are studying a new drug that is supposed to prolong the survival time of certain patients with a serious form of cancer. You give the new drug to 125 cancer patients, and then wait around like a vulture until they all die. Now you might want to estimate the average amount of time that a patient, given this drug, will survive. A very reasonable estimate of this would be the average of the survival times of your 125 patients (this is called the *sample average* or the *sample mean*. This sample mean is intended to approximate the true mean survival time. (The true mean survival time is the average survival time for all possible cancer patients when they are given this drug. Clearly we cannot compute the true mean because we can't give this drug to every past, present, and future cancer patient in the world.)

Naturally we cannot expect that the sample mean is exactly equal to the true mean. After all, our sample of 125 patients, even if it is chosen completely at random, is not likely to be perfectly representative of the population of all cancer patients. By the luck of the draw we could have, say, a higher proportion of really sick people in our sample than occurs in the general population of people with cancer: then the sample mean survival time would be lower than the true mean. The reverse could be the case as well: maybe our sample contains an unusually large number of people with strong constitutions and strong wills to survive. In this case our sample mean would be somewhat bigger than the true mean.

Therefore, when we present the results of our study to the boss, (or the doctors or whoever) we can't just report the sample mean. People will also want us to give some indication of how close the sample mean is likely to be to the true mean. (For those that know the variance of the mean is 1/n of the variance of the population, you may be thinking that this is easy. Try median or trimmed mean!)

One way to give people an idea of the accuracy of the sample mean is to construct a *confidence interval* for the true mean. A confidence interval is just an interval of numbers (a, b) that you hope will contain the true mean. Now any confidence interval has a certain level of confidence associated with it. For example, let's say that the boss would like to be 90% confident that the true mean lies within the interval. Then we would construct a 90% confidence interval for the mean. There are many methods for constructing such a confidence interval. If such a method is any good, then it should "work" about 90% of the time. That is, if we used our method over and over again to construct confidence intervals in hundreds or thousands of independent situations (say in other studies of cancer patients) then about 90% of the confidence intervals that we construct in each situations should contain the true mean that we are trying to estimate in that situation.

In our specific case, suppose that the sample mean survival time turned out to be 2.5 years. Maybe our 90% confidence interval is computed to be (2.1,2.9). (I am just making up these numbers for the sake of illustration— we haven't yet discussed any method for constructing confidence intervals.) Given this interval, the boss could say: "Well, the true mean survival time of any patients who could be given this drug is around 2.5 years, give or take some. I can be about 90% sure that the true mean is more than 2.1 years and less than 2.9 years, but I have to accept the possibility that the true mean could be less than 2.1 years or more than 2.9 years. I guess I can live with that uncertainty."

How are such intervals constructed? If you consult the elementary statistics textbooks, you will find chapters devoted to this subject. The methods in these chapters all rely on some underlying assumptions about the type of *distribution* of the survival time over all possible cancer patients. A distribution is a statistical function that, roughly speaking, tells you what proportion of all cancer patients would survive for a given period of time (i.e., maybe 30% will live less than one year, 10% will survive more than a year but less than 1.5 years, and so on). The methods given in the elementary texts will give reasonably accurate confidence intervals so long as this distribution is close to *normal* type. (Or the only measurment you are interested is the mean, not median etc.) A distribution is said to be normal if, when you make a histogram of all possible data (not just the data you have in your sample), you get a histogram with a bell-shape (see the appended figures). There is a specific mathematical formula for a normal distribution, but this need not concern us now. The further away from normal type the underlying distribution is, the less accurate our confidence intervals will be.

In many cases it is quite all right to assume that the underlying distribution is normal type. For example, if we are trying to estimate the average height of female students at Pikeville college by measuring 20 of them at random, then the methods in the textbooks will do fine, because the distribution of heights is usually quite close to normal (most people are near the average height, and the proportion of really tall or really short people drops off quickly). However, if we were trying to estimate the average income of U.S. residents, based on a random sample of 20 residents, then the methods in the text, while still OK, don't work quite as well. That's because the underlying distribution of income for U.S. residents is not very close to normal — in fact, it is rather skewed to the right (see the figures). In our situation, with the cancer patients and the new drug, we have virtually no information about the underlying distribution. The methods in the text could be used to give the boss an "answer" but we would know in our hearts that the answer might not be a very good one.

In order to construct a good confidence interval, we would need to have some idea of how much the sample average would hop around, from sample to sample. In a perfect world, we could get an idea of this by taking another random sample of 125 patients, giving them the drug, and seeing how long they live on the average. Then we could take yet another sample of 125 patients, give them the drug, compute the sample mean survival time for this sample, and so on. If we grabbed many such samples (say, 1000 or so), then we could look at the 1000 sample means that we computed, and see how much they vary. If they vary a lot, we would have to construct a fairly wide 90% confidence interval. If they don't vary much at all, then we would construct a fairly narrow confidence interval (there are systematic methods for doing this—it's not just guesswork). Unfortunately, this is not a perfect world: these 125 cancer patients are the only ones we will ever get to look at.

In order to get an idea of how much the sample mean would vary from sample to sample, we will have to work only with the sample we have already collected. What can we do?

Here is where the bootstrap comes in. Since we cannot take more samples from the original population (all possible cancer patients), why not take more samples from the existing sample itself? After all, a histogram based on our 125 patients' survival times is probably not going to be too different from a histogram of the survival times of all possible cancer patients, because our 125 patients are a random sample from all cancer patients and are therefore fairly representative of all cancer patients.

To make this idea clearer, let's illustrate it with specific values. Suppose that the survival times for the original 125 patients are as given in the appendix.

From this list of survival times, choose one number at random. Record it, and put it back in the list. Then choose another number at random (it could be the same number as you drew the first time, but of course probably won't be). Record this number, and put it back again. Keep repeating this process until you have drawn 125 numbers (same as the size of the original sample). Now compute the average of the 125 numbers that you have drawn. Call this average A_1.

Now repeat the above process again, getting a new average A_2. Keep repeating this process many times (say, 2000 times). You will get a list of 2000 averages. These averages will hop around a little bit.

I actually forced my computer to do the above experiment. The first five re-samples, together with the five averages computed from them, are recorded on an attached sheet. The entire list of 2000 averages, sorted from smallest to highest, is printed on the next sheet. Look at the list to see how the averages vary.

The **Bootstrap Principle** says that *the re-sample averages will hop around (sample average) about as much as the sample averages would have hopped around (population average) if we could have taken our further samples from the original population.* This is very important, so think about it for a few minutes.

Here's how we can construct a 90% confidence interval for the true mean survival time. Take your list of 2000 re-samples averages. Hack off the lowest 5% of the list (that's the first 100 values in the sorted list). Then hack off the highest 5% of the list (the last 100 values in the sorted list). Your 90% confidence interval will be the interval from the lowest remaining number to the highest remaining number. Our 90% interval is given in one of the appended printouts. This method of constructing a 90% confidence interval is called the *bootstrap percentile* method. Notice that the interval spans the middle 90% of the re-sampled averages. Other, more sophisticated bootstrap methods for confidence intervals have been developed, but they are all variations of this basic method.

I should point out that the above method does not produce exact 90% confidence intervals. That is, if I could repeatedly draw samples of 125 cancer patients from the original population, give them the drug, and compute a 90% confidence interval using the bootstrap method, then it would not be true that exactly 90% of my confidence intervals would contain the true mean survival time. However, the percentage of confidence intervals that do contain the true mean survival time would be fairly close to 90%. That's good. No other method produces exact intervals, either.

Well, that's the idea of the bootstrap. People are always coming up with ways to use this idea in other situations. People have been trying to use this method to make estimates based on *multivariate* data. Data is multivariate when it comes at you with more than one number at a time. For example, in a study of the human body, you might need to make more than one measurement on each person in your study. For example, for each person you could record their weight, height, and nose length. If you wanted to estimate the true mean weight and true mean height and true mean nose length of all people in the population, based on your sample, you could just treat each of those variables separately by the bootstrap method above. However, if you are interested in certain other quantities (like the trimmed mean discussed in some elementary textbooks), you can't do the same thing. You have to deal with the data as multivariate data. In this case, when you come to the part of the bootstrap where you throw out the 5% highest and 5% lowest of your re-sampled values, you get stuck: how do you talk about "highest" and "lowest" in two or more dimensions?