Tutorial -Bootstrapping

By Kota Ando

1 Introduction

This is a tutorial about the basics of bootstrapping. Bootstrapping, like many other statistical methods, is a method used to infer the characteristics of a population from the samples collected from it. The difference is, that instead of only using the samples at hand to calculate those characteristics, bootstrapping will create more sets of data by simulating the population from the original sample, and then conducting the calculations over all those newly created samples.

This tutorial is designed to a give brief overview of bootstrapping to people who have not studied much statistics in the past. Therefore, it will cover the very basics of statistics first, then it will show how bootstrapping is done in practice with some examples.

2 Very basics of Statistics

The basic idea of statistics is to provide some way to measure and analyse the qualitative or quantitative data of a population without actually measuring the population. The reason for not measuring the whole population is quite obvious. If you want to know the average height of New Zealand residents, do you want to measure more than 3 million people?
Also what about those who are usually New Zealand residents, but just away overseas for a week?
So this is where the idea of sampling from a population comes in, and this is where the subject of statistics comes in. Statistics is the subject where they infer a population's parameters from the sample.

2.1 Sample and Population

The population represents the mini-world which you want to find out something about. The sample is a subset of the population from which the data is obtained. So the question is how they relate.

2.1.1 How they relate?

In most of the cases, if the sample is taken from a population reasonably well, the sample should have similar characteristics to the population. For example, if you take a sample of 10,000 from the population which has a mean of 10, your sample should also have a mean of around 10. (note: it will most likely not be exactly 10, but very close to it) But if you have a sample size of 1, then the mean of the sample, which is equal to the value of the sample, may be quite a bit off, like 6 or 15. This is one of the reasons that the sample has to be taken from a population well, and a reasonable sample size must be taken.

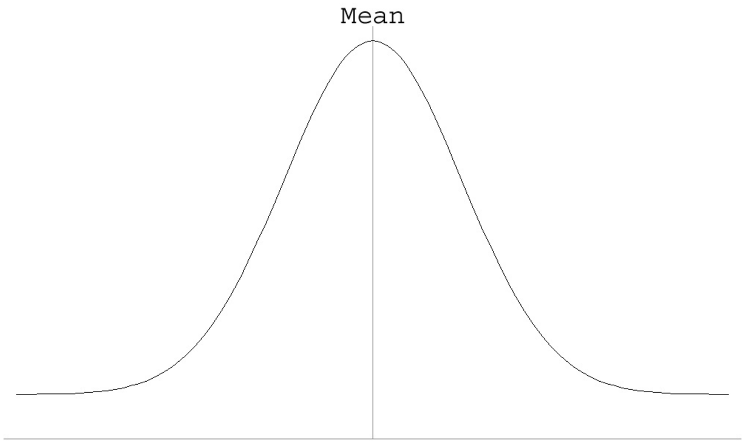
Another question is, how much error for the sample mean can be acceptable? I have mentioned the value 6 or 15 for the mean of the sample, where the population has a mean of 10, but if the data spans plus or minus 1 million or so, it may be acceptable. But if you have a population for which its values are all between 9 and 11, those sample values are not acceptable. This is the reason why a large number of samples are taken from population.

Now, if enough samples are taken from a population, the sample mean reflects the population well. So, if we only have samples without knowledge of the population, as long as we have enough of them, it should reflect the population.

2.1.2 Infer population parameters from samples

To infer population parameters from samples, you must have some way to relate the sample to the population. One such way is to assume the statistic from the sample is reflecting its population parameter. For example, if sample has a mean of 9, then you assume that it is also the mean of the population. This can be done if there are enough sampling units, and you assume that the sample has the same characteristics as the population.

One example of such an occasion would be when dealing with the natural world, when samples are taken from a population, sometimes the sample data is normally distributed. (See the figure)



Normal distribution looks like a bell shape, where peak of the line is at the mean, and it goes downhill as a value gets away from the mean. What it means is that if a sample were drawn from a population which is normally distributed, it is more common to get the sample data which has a value closer to population mean, rather than those with a value significantly different from the population mean.

For example, if we use the example about the people's height again, it is more likely to find people close to the average height, than someone who is extremely short, or extremely tall if we pick a person randomly from whole New Zealand population.

Now, there are many kinds of distribution other than the normal distribution. And each distribution has a different set of parameters which are used to describe its distribution. For example, the normal distribution can be described by the mean and standard deviation of the population. But binomial distribution is described by the probability of success of an event and the number of trials. In this tutorial, we will only mention the normal distribution to simplify matters.

The problem is that in some situations, there may be some indication of what the population is like, but there may also be situations where you have absolutely no idea about the population.

2.1.3 Parametric and Non-parametric Methods

Now, when assumptions about the population can be made safely, it means we have got some population parameters that describe its distribution. In such a situation, we can use the "Parametric Method" to infer some statistic. As name suggest, parametric methods are the ones which can be used if we can assume some population parameters.

An example of such situation could be that when a sample of people's height is taken from the whole New Zealand population. Since we can generally assume that the height of people is normally distributed, we can use parametric methods.

Otherwise, if we can not assume anything about the population parameters or its distribution, you can not use the above method, and hence we must use the "Non-Parametric Method". Also such methods could be used if there are some doubts about the sample data. An example of such a case is, say when a sample of people's heights is taken from the people in some region where they happened to have more children compared to the whole New Zealand, then it is more likely that the data would be skewed towards shorter people. This will break the assumption of the normal distribution, since if a diagram of the distribution is drawn, it will not be symmetrical.

This is where we have to be careful, since those methods are completely separate from each other.

3 Bootstrapping

3.1 Basic Idea

This is the method of sampling suggested in 1979 by B. Efron [1], and there are two methods or types of bootstrapping, namely parametric and non-parametric method. In general, the non-parametric method is used more often, but a parametric method also exists.

In either case, the basic idea is that by simulating the population from the sample, and creating more sets of samples from the simulated population many times and conducting the same statistical inference on each set of samples and averaging the result, the result should reflect the characteristics of population relatively well.

For example, if you are unsure of reliability of the statistic inferred from current sample set, one of the ways to find the error of sample is to simply find the error of the sample assuming the formula exists for such a calculation. Alternatively, you can conduct many simulations and recreate the sample itself, and then find the error of each newly created sample and average the it over the total number of simulations. If the number of simulations is large enough, this average should converge to the value such it should represent the actual error value reasonably well.

3.1.1 Origin of the name

Since it sounds like the sample is doing something to itself, the term "bootstrapping" is used, as it sounds similar to the case where a man tries to pull himself up from the ground using his own bootstrap.

3.2 When to use it

Bootstrapping can be used in many occasions, since both parametric and non-parametric methods exist, and there are not too many restrictions about when it can be used. Of course, the method (parametric or non-parametric) will have to be determined depending on the type of data, and the assumption of the distribution.

Now, one situation which should be mentioned specifically regarding the time when bootstrapping can be used, is when you want to calculate the standard error of a certain statistic.

This is because some statistics do not have a standard error associated with their estimate from the sample, because formula for standard error often assumes something about the sampling distribution. For example, the standard error for the mean assumes the sample is normally distributed. Now, when we need to find the standard error of some statistic which does not have a formula for the standard error associated with it, you must take the bootstrap samples and find the error of the statistic and use them. For example, the median of the sample has no associated formula to calculate the standard error directly. That is where bootstrapping can be used to simulate many sets of samples and work out the error from the each sets of samples, and find the standard error. See the examples which are on later on in the tutorial.

3.3 Method

3.3.1 Non-parametric Method

Supposed we have sample of n, and we can not assume the distribution of the population from which the samples were taken from. Then the normal parametric method can not be applied.

Suppose we want to calculate the mean and its error. First, from the original samples we calculate the mean in the usual way. Then, to calculate its standard error, we take a bootstrap sample of size n from the original sample with replacement. Another word, we pick the bootstrap samples of size n from the original sample of size n with replacement, so some sample values may appear multiple times, while some other values may not appear at all in a given sample. Repeat this k times, where k is the number of bootstrap samples we want, and call each set of samples B_i, where i is an integer which represents the number of samples i.e. i is between 1 and k. (e.g. if first bootstrap sample is B_1, then the 100^{th} sample will be denoted by B_{100}) For each set of samples, find the mean $\bar{x}_i$.

Next, find the mean of those sample means:

$$\bar{x} = \frac{\sum_{j=1}^{k} \bar{x}_j}{k} \tag{1}$$

And then, to find the error of those means, apply the following equation:

$$Std\_Err_{\bar{x}} = \sqrt{\frac{\sum_{j=1}^{k} (\bar{x}_j - \bar{x})^2}{k-1}} \tag{2}$$

This is the standard deviation of the bootstrap replicate of the mean.

3.3.2 Parametric Method

In the case of the parametric method, we do similar sampling to the non-parametric method, but we take the bootstrap samples differently. That is, in this method, we assume some distribution, say the normal distribution, and randomly sample from the assumed normal distribution. An example would be, suppose we have a sample size of 50. Find its mean and standard deviation, which are two parameters for the normal distribution, and use those as a population parameter to generate a specific normal distribution curve. Then using those distribution parameters, generate data sets which fits the distribution using some algorithm like the one suggested in the book by Knuth [2], and create a number of bootstrap samples. (in this case, each sample would have size of 50) Then conduct the same calculations as mentioned above in Non-Parametric method section to find the standard error of the sample.

One may ask, what right do we have to assume things about the populations distribution. Answer is, none. But parametric methods are often used because the mathematics is available to calculate parameter estimates.

3.4 Examples

To show how it actually works, here are some examples.

Firstly, 1000 numbers were randomly chosen such that they are normally distributed, and have a mean of 0 and standard deviation of 1. (actual data I have used has a mean of 0.010247) Let this be our population . Then a sample of 50 is drawn from those 1000 numbers.

Now, we will only use the sample information to see whether we can infer the population parameters, say mean. In this example, we will use both parametric and non-parametric methods, and see how both will perform. Since we know that the population is normally distributed, we can use the parametric method, but it is also a good chance to test whether the non-parametric method can also be used even when we can assume population parameters. The first thing you can do, is to find whatever statistic you want from the sample. If you want the mean of the given sample, then find the mean of the original sample first. This will be used later.

Now, we want to find the standard error of the mean of the sample we have got, so we know the error. To do so using bootstrapping, we have to create some of boot samples. Say we make 1000 boot samples. We will first take a look at the non-parametric method.

3.4.1 Non-Parametric Method Example:

For this , we will have to pick boot samples of size 50 from the original sample with replacement from the sample. So if you have say data labelled 1 to 50, we could have boot samples like the following:

B_1	28	27	10	49	8	16	11	49	8	28	16	4	...
B_2	7	9	24	19	42	33	49	1	27	25	17	27	...
...	...	...	...	...	...	...	...	...	...	...	...	...	...
B_n	...	...	...	...	...	...	...	...	...	...	...	...	...

After the bootsamples have been chosen, you calculate the statistic of you choice for each set of boot samples namely B_1 to B_{1000}. So for our example, we want to work out the mean of each sets. Lets call them $\bar{x}_{Bi}$ where i represents the bootsample number (between 1 and

1000). Then we find the mean of those means by applying the equation (1):

$$\bar{x} = \frac{\sum_{i=1}^{1000} \bar{x}_{Bi}}{1000} \tag{3}$$

Then we apply the standard error equation (2) for the bootstrap samples:

$$Std\_Err_{\bar{x}} = \sqrt{\frac{\sum_{i=1}^{1000} (\bar{x}_i - \bar{x})^2}{1000 - 1}} \tag{4}$$

Note, this result will vary slightly every time you conduct the calculations, since all the samples are chosen randomly, and hence some difference in the final result is unavoidable. You can however, make this error very small by creating larger sets of bootsamples. So instead of 1000 samples, you can take 10000, 100000 or more sets at a time.

Now, this standard error, and the original sample's mean can be used to calculate things like the confidence interval, which is perhaps the one we want.

Here are the results from the program which does bootstrapping and the calculation of the mean, its standard error and the 95% confidence interval. Now, one of the result is:

```
Type of method: Non-parametric Method
Mean = 0.066550
Standard Error of Mean = 0.153649
95% Confidence interval of mean: -0.234601 < x < 0.367702
```

If I run it again:

```
Type of method: Non-parametric Method
Mean = 0.066550
Standard Error of Mean = 0.170672
95% Confidence interval of mean: -0.267967 < x < 0.401068
```

What the confidence interval means, is that we are at least 95% sure that the population mean is in this given range. This is correct since our population we have generated has a mean of 0.010247. This may not be the best way to show that it actually works for many things, but it should be reasonably clear that it does work.

3.4.2 Parametric Method Example:

For this, instead of picking the bootstrap samples from the original samples, we simulate the population, and create a new sample from that population. Now, there are number of ways to simulate population from the sample. One way, is to have an educational guess as to what the population parameters should be. In some cases, this can be done if you know enough about the population. The other and perhaps more sensible way to simulate a population is to assume that the sample is normally distributed and it has the same or similar distribution parameters to the population. If we use this second assumption and also assume that the sample is normally distributed, we can use the equation associated with the normal distribution to find the mean and the standard deviation of the sample first, and use those as if they are the population parameters. Then, since we know the population parameters, by using a suitable algorithm, we can generate an entirely new set of samples.

So, if we have a sample of size 50, we can first find its mean and standard deviation and use these to create new bootstrap samples of size 50 using an algorithm such as the one in Knuth's book. And then we repeat it say 1000 times to create bootstrap samples, then we can apply the equations **??** and **??** to calculate the statistic in the same way as before.

Now here is the result from the program using the parametric method:

```
Type of method: Parametric Method
Mean = 0.066550
Standard Error of Mean = 0.160856
95% Confidence interval of mean: -0.248727 < x < 0.381827
```

Run it again, and we get:

```
Type of method: Parametric Method
Mean = 0.066550
Standard Error of Mean = 0.162975
95% Confidence interval of mean: -0.252881 < x < 0.385982
```

Once again, the range of the confidence interval includes the population mean of 0.010247. This should give an indication that it does work.

4 Application

In the field of Computer vision, it seems its application is not directly related to the actual vision method, but it seems it is used more for the testing of its algorithm and to compare it with the other methods to see the overall result.

For example, when a new algorithm is invented for face recognition, by creating many bootstrap sample sets and conducting an experiment on each set separately, you can find the general performance of the each algorithm by averaging the score over the number of bootstrap sample sets.

5 Conclusion

Bootstrapping seems to be useful for many reasons. Perhaps one of the biggest reason is that in order to apply this method, one does not need any assumptions about the population parameters. This would be a critical point when applying bootstrapping in the field of computer vision for say analysing performance of the new algorithm, since there will be no indication of how the performance of algorithms are distributed. And since the non-parametric method can be applied even when you actually know the distribution, this is at least a safe method to apply in any situation. The only potential problem is that this process is rather computer intensive. Depending on the size of the sample and the accuracy required, it may take a while. Having said that, it will not be a major problem anymore, as the speed of computers is increasing rapidly.

6 Glossary

6.1 Definition

Statistics The science of recording, organizing, analysing and reporting quantitative information.

Population Every data in a certain mini-world at which you want to infer some information about

Sample Sample is a subset of a population, which are chosen to be representative of the population.

Parameter any characteristic of the population.

Statistic any characteristic of a sample.

Sampling Error The discrepancy, or amount of error that exists between a sample statistic and the corresponding population parameter.

Standard Error A measure of precision of the statistic. In simple words, its the error of the statistic infered from the sample

7 Appendix

7.1 Instruction for the use of the program:

```
Usage: bootstrap -i filename [arguments]
  Argument List:
    -p Use Parametric Bootstrapping
    -np Use Non-parametric Bootstrapping
    -n Number Number specifies how many bootstrap sample set should be taken.
    -mean Calculate statistic using mean
    -median Calculate statistic using median
    -ci Conduct Confidence interval calculation
    -s Specify seed for random number generator
    -h Display this help message
    -m Display more help message
```

7.2 Implementation, and notes about expanding the program in the future

For the details of implementation, there is a source code written in C attached at the end of this tutorial. In brief, it reads the numbers in the given file, and it creates bootstrap samples. The method used for it will be specified in the command line argument. Also, the statistics which we need to calculate can be specified in the command line argument. To be flexible in the size of the data it can read etc. nearly all the data structures are dynamic. It utilises the linked list for the input from the file, and for the bootstrap samples, it uses 2-dimensional arrays. The reason it uses the linked list, is so that we do not have to do 2 passes of the file to identify how many data we have to read. It does however, create some inefficiency later as a result, since we have to pick numbers randomly from that set. Since the linked list is not the best structure for it, it will copy the contents to the 1^{st} column of the array which later will store the bootstrap samples. The design of the program in general should be reasonably modularised so that it should be able to be expanded if more statistical methods should be added in the future. Following are the general ideas about how to expand the program:

1. Add required parameters flag to the parameters structure

2. Add required command line argument processor

3. Add section in main function which checks for the parameters, and does certain actions depending on it
4. Add functions which do the calculation

The equation for the standard error of the statistic remains the same for any other ones like correlation, difference of two means etc. The only difference is, that instead of having say the mean in place of $\bar{x}_{Bi}$, we will substitute the corresponding statistic of the bootstrap samples in place of $\bar{x}_{Bi}$.

References

[1] B Efron. Bootstrap methods: another look at the jackknife. *Annals of Statistics*, 1979.

[2] Ronald E. Knuth. *The Art of Computer Programming.* Addison-Wesley Publishing, 1969.