Bootstrap

Vera

1 Introduction

The estimation of parameters in probability distributions is a basic problem in statistics that one tends to encounter already during the very first course on the subject. Along with the estimate itself we are usually interested in its accuracy, which is can be described in terms of the bias and the variance of the estimate, as well as confidence intervals around it. Sometimes, such measures of accuracy can be derived analytically. Often, they can not. Bootstrap is a technique that can be used to try and estimate them numerically from a single data set.

2 The general idea

Let $X_1, ..., X_n$ be a i.i.d. sample from distribution F, that is $\mathbf{P}(X_i < x) = F(x)$, and let $X_{(1)}, ..., X_{(n)}$ be the corresponding ordered sample. We are interested in some parameter θ which is associated with the distribution (mean, median, variance etc). There is also an estimator $\hat{\theta} = t(\{X_1, ..., X_n\})$, with t denoting some function, that we can use to estimate θ from data. It is the deviation of $\hat{\theta}$ from θ that is of primarily interest, or, more precisely, the distribution of $\hat{\theta}$.

Ideally, to get an approximation of the estimator distribution we would like to repeat the data-generating experiment, say, N times, calculating $\hat{\theta}$ for each of the N data sets. That is, we would draw N samples of size n from the true distribution F. In practice, this is impossible.

The bootstrap method is based on the following simple idea: *Even if we do not know F we can approximate it from data and use this approximation, $\hat{F}$, instead of F itself.*

This idea leads to several flavours of bootstrap that deviate in how, exactly, the approximation $\hat{F}$ is obtained. Two broad areas are the *parametric* and *non-parametric* bootstrap.

The non-parametric estimate is the so called empirical distribution (you will see the corresponding pdf if you simply do a histogram of the data), that can be formally defined as follows:

With # denoting the number of members of a set, the *empirical distribution function* $\tilde{F}$ is given by

$$\tilde{F}(x) = \frac{\#\{i : X_i \leq x\}}{n} = \begin{cases} 0 & \text{if} \quad x \in (-\infty, X_{(1)}), \\ i/n & \text{if} \quad x \in [X_{(i)}, X_{(i+1)}) \quad \text{for} \quad i \in \{1, \ldots, n-1\}, \\ 1 & \text{if} \quad x \in [X_{(n)}, \infty). \end{cases}$$

That is, it is a discrete distribution that puts mass $1/n$ on each data point in your sample. It can be shown that $\tilde{F}$ converges to F as $n \to \infty$.

The parametric estimate assumes that the data comes from a certain distribution family (Normal, Gamma etc). That is, we say that we know the general functional form of the pdf, but not the exact parameters. Those parameters can then be estimated from the data (typically with Maximum Likelihood) and plugged in the pdf to get $\hat{F}$. This estimation method leads to more accurate inference if we guessed the distribution family correctly but, on the other hand, $\hat{F}$ will not approach F if the family assumption is wrong.

2.1 Algorithms

The two algorithms below implement the above. To concretize, let us say that $X_i \sim N(0,1)$, is the median and it is estimated by $\hat{\theta} = X_{(n/2)}$.

Non-parametric bootstrap
Assuming a data set $x = (x_1, ..., x_n)$ is available.

1. Fix the number of bootstrap re-samples N. Often $N \in [1000, 2000]$.
2. Sample a new data set x^* set of size n from x *with replacement* (this is equivalent to sampling from the empirical cdf $\hat{F}$).
3. Estimate θ from x^*. Call the estimate it $\hat{\theta}^*_i$. Store.
4. Repeat 2 and 3 N times.
5. Consider the empirical distribution of $(\hat{\theta}^*_1, ..., \hat{\theta}^*_N)$ as an approximate of the true cdf of $\hat{\theta}$.

Parametric bootstrap
Assuming a data set $x = (x_1, ..., x_n)$ is available.

1. Assume that the data comes from a known distribution family F_ψ described by a set of parameters ψ (for a Normal distribution $\psi = (\mu, \sigma)$ with μ being the expected value and σ the standard deviation).
2. Estimate ψ with, for example, Maximum likelihood, obtaining the estimate $\hat{\psi}$.
3. Fix the number of bootstrap re-samples N. Often $N \in [1000, 2000]$.
4. Sample a new data set x^* set of size n from $F_{\hat{\psi}}$.
5. Estimate θ from x^*. Call the estimate it $\hat{\theta}^*_i$. Store.
6. Repeat 4 and 5 N times.
7. Consider the empirical distribution of $(\hat{\theta}^*_1, ..., \hat{\theta}^*_N)$ as an approximate of the true cdf of $\hat{\theta}$.

In the Figure 1 the distribution of $\hat{\theta}$ (the median) approximated with the non-parametric and parametric bootstrap is plotted. Note that the parametric distribution is smoother than the non-parametric one, since the samples were drawn from a continuous distribution. As stated earlier, using the parametric version could be more effective, but also potentially more difficult to set up. In particular, we have to sample from $F_{\hat{\psi}}$ which is not easily done if F is not one of the common distributions (such as Normal) for which a pre-programmed sample function is usually available.

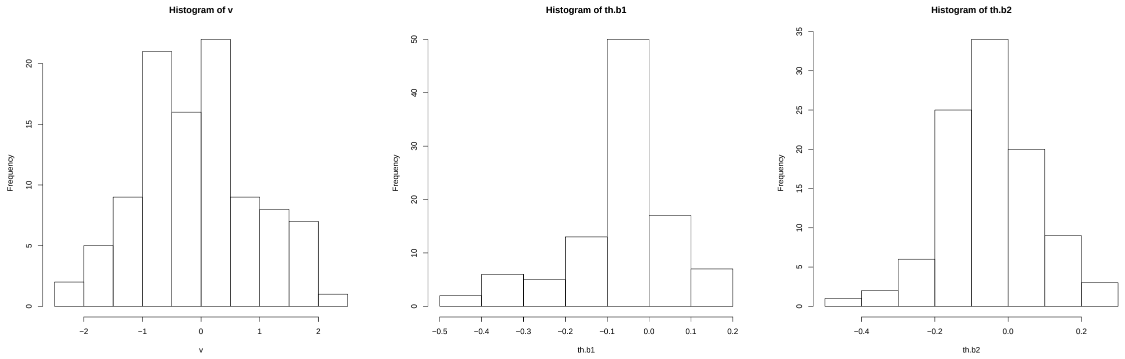

Figure 1: The data distribution. The non-parametric sample. The parametric sample.

3 Bias and variance estimation

The theoretical bias and variance of an estimator $\hat{\theta}$ are defined as

$$\mathbf{Bias}(\hat{\theta}) = \mathbf{E}[\hat{\theta} - \theta] = \mathbf{E}[\hat{\theta}] - \theta$$

$$\mathbf{Var}(\hat{\theta}) = \mathbf{E}\left[(\hat{\theta} - \mathbf{E}[\hat{\theta}])^2\right] \approx \mathbf{E}[(\hat{\theta} - \theta)^2]$$

with the approximation in the variance definition being valid if the bias is small. In words, the bias is a measure of a systematic error ($\hat{\theta}$ tends to be either smaller of larger than θ) while the variance is a measure of random error.

In order to obtain the bootstrap estimates of bias and variance we plug in the original estimate $\widehat{theta}$ (which is a constant given data) in place of θ and $\hat{\theta}^*$ (the distribution of which we get from bootstrap) in place of $\hat{\theta}$. This leads us to the following approximations:

$$\mathbf{Bias}(\hat{\theta}) \approx \frac{1}{N}\sum_i \hat{\theta}^*_i - \hat{\theta} = \hat{\theta}^*_{\cdot} - \hat{\theta}$$

$$\mathbf{Var}(\hat{\theta}) \approx \frac{1}{N-1}\sum_i (\hat{\theta}^*_i - \hat{\theta}^*_{\cdot})^2$$

That is, the variance is, as usual, estimated by the sample variance (only for the bootstrap sample of $\hat{\theta}$) and bias is estimated by how much the original $\hat{\theta}$ deviates from the average of the bootstrap sample.

4 Confidence intervals

There are several methods for CI construction with Bootstrap, the most popular being ”normal”, ”basic” and ”percentile”. Let $\hat{\theta}^*_{(i)}$ be the ordered bootstrap estimates, with $i = 1, ..., N$ indicating the different samples. Then the two-sided CIs at the significance level α are defined as

Basic	$[2\hat{\theta} - \hat{\theta}^*_{(N+1)(1-\alpha/2)}, 2\hat{\theta} - \hat{\theta}^*_{(N+1)\alpha/2}]$
Normal	$[\hat{\theta} - z_{\alpha/2}\hat{se}, \hat{\theta} + z_{\alpha/2}\hat{se}]$
Percentile	$[\hat{\theta}^*_{(N+1)\alpha/2}, \hat{\theta}^*_{(N+1)(1-\alpha/2)}]$

with z_α denoting an α quantile from a Normal distribution and $\hat{se}$ the estimated standard deviation of $\hat{\theta}$ calculated from the bootstrap sample.

Basic CI

To obtain this confidence interval we start with $\hat{\theta} - \theta$ and, like in the variance and bias calculations, approximate with $\hat{\theta}^* - \hat{\theta}$. Let us call the p-quantile of the distribution of $\hat{\theta}^* - \hat{\theta}$ for q_p. We then have that an α-level CI for $\hat{\theta} - \theta$ can be approximated with

$$q_{\alpha/2} \leq \hat{\theta} - \theta \leq q_{1-\alpha/2} \Leftrightarrow \hat{\theta} - q_{1-\alpha/2} \leq \theta \leq \hat{\theta} - q_{\alpha/2}$$

We will get the formula for the Basic CI by noting that $q_p = \hat{\theta}^*_p - \hat{\theta}$ and substituting it in the expression above.

This interval construction relies on an assumption about the distribution of $\hat{\theta} - \theta$, namely that it is independent of θ. This assumption, called pivotality, does not necessarily hold in most cases. However, the interval gives acceptable results even if $\hat{\theta} - \theta$ is close to pivotal, that is its dependence on θ is weak.

Normal CI

This confidence interval probably looks familiar since it is almost an exact replica of the commonly used confidence interval for a population mean. Indeed, similarly to that familiar CI, we can get it by assuming that $Z = (\hat{\theta} - \theta)/\hat{se} \sim N(0, 1)$. Observe that this also implicitly assumes pivotality.

Percentile CI

May seem natural and obvious but actually requires a quite convoluted argument to motivate. Without going into details, you get this CI by assuming that there exist a transformation h such that the distribution of $h(\hat{\theta}) - h(\theta)$ is pivotal, symmetric and centered around 0. You then construct a Basic CI for $h(\theta)$ rather than θ itself, get the α quantiles and transform those back to the original scale by applying h^{-1}. Observe that although we do not heed to know h explicitly, the existence of such a transformation is a must, which is not always the case.

5 The restrictions of bootstrap

Yes, those do exist. Bootstrap tends to give results easily (too much so, in fact), but it is possible that those results are completely wrong. More than that, they can be completely wrong without being obvious about it. Let take a look at a few classics.

5.1 Infinite variance

In the "general idea" of bootstrapping we plugged in an estimate of the density F, $\hat{F}$, and then sampled from it. This works only if $\hat{F}$ actually is a good estimate of F, that is it captures the essential features of F despite being based on a finite sample. This may not work well if F is very heavy tailed, e. g. has infinite variance. The intuition is that in this case extremely large or small values can occur, and when they do they have a great effect on the bootstrap estimate of the distribution of $\hat{\theta}$, making it unstable. As a consequence, the measures of accuracy of $\hat{\theta}$, such as CI, will be unreliable.

The classic example of this is the mean estimator $\hat{\theta} = \bar{X}$ with the data generated from a Cauchy distribution. For it, both the first and the second moments (e.g. mean and variance) are infinite, leading to nonsensical confidence intervals even for large sample sizes. A less obvious example is a non-central Student t distribution with 2 degrees of freedom. This distribution has pdf

$$f_X(x) = \frac{1}{2\,(1 + (x - \mu)^2)^{3/2}} \quad \text{for} \quad x \in \mathbb{R}.$$

where μ is the location parameter, defined as $\mathbf{E}[X]$. So, the first moment is finite and its estimator $\bar{X}$ is consistent. The second moment, however, is infinite, and the right tail of the distribution grows heavier with increasing μ. This leads to the approximate 95% CI coverage probabilities displayed in Figure 2 when the "basic" and "percentile" methods are used on a non-parametric bootstrap sample (the "Normal" method is, of course, inapplicable due to the infinite variance). The actual coverage is quite far from the the supposed 95% even when the sample size is as large as 500.

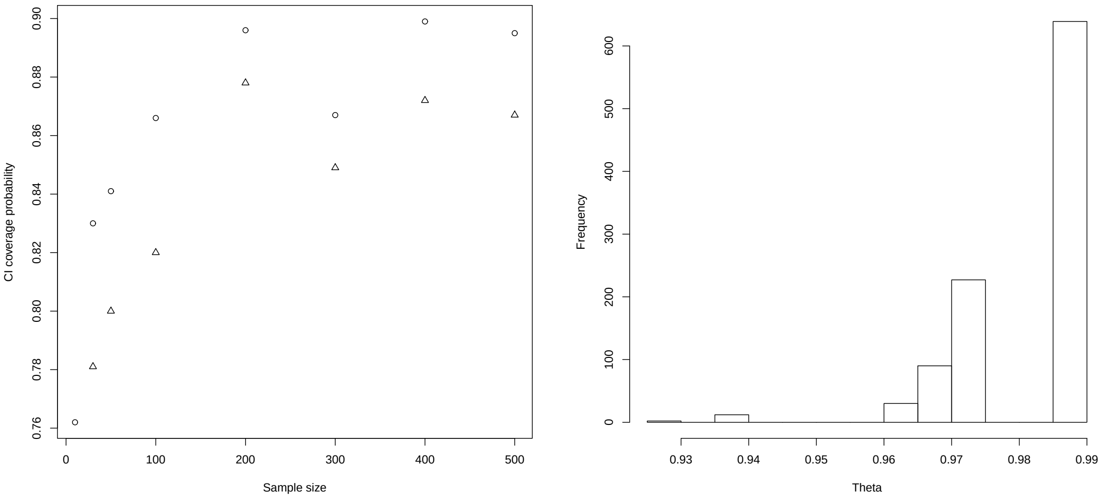

Figure 2: Left: The approximate coverage probabilities for the "basic" (triangles) and "percentile" (circles) CI when $\mu = 10$. Right: histogram of 1000 non-parametric bootstrap samples of $\hat{\theta}^*$, the data from $Uni(0, \theta)$ distribution. $\theta = 1$.

This potential problem can be circumvented with the sub-sampling methods, such as Jack-Knife. This method, however, is not without its own issues, especially clear if $\hat{\theta}$ is not affected much by sub-sampling, for example if it is a median.

5.2 Parameter on the boundary

The classical example here is the $Uni(0, \theta)$ distribution. The maximum likelihood estimate $\hat{\theta}$ is simply $\max(x_1, ..., x_n)$, which will always be biased ($\hat{\theta} < \theta$). In this case the non-parametric bootstrap leads to a very discrete distribution, with more than half of the bootstrap estimates $\hat{\theta}^*$ equal to the original $\hat{\theta}$ (Figure 2). Clearly, if the quantiles used in CIs are taken from this distribution the results will be far from accurate. However, the parametric bootstrap will give a much smoother distribution and more reliable results.

5.3 Lack of pivotality

In all the CI descriptions above the word "pivotality" shows up. So we can guess that it is a bad thing not to have (see Lab, task 3). To circumvent this, something called "studentized bootstrap" can be used.

The idea behind the method is simple and can be seen as an extrapolation of the Normal bootstrap CI. There, we assumed that the distribution of the standardized test statistic Z is known, namely that it is $N(0, 1)$. In studentized bootstrap we instead say that the distribution of Z is not known, but can be approximated from data. Through, again, bootstrap.

So, we proceed in the similar way as before and approximate Z with $Z^* = (\hat{\theta}^* - \hat{\theta})/\hat{se}^*$, where $\hat{se}^*$ is the estimate of the standard deviation of $\hat{\theta}^*$. This quantity, Z^*, is calculated for each bootstrap sample, so that N values are obtained. Those values can then be viewed as a sample from the distribution of Z. They are ordered to get the ordered sample $(Z^*_{(1)}, ..., Z^*_{(N)})$, and we use $Z^*_{(N+1)\alpha/2}$ and $Z^*_{(N+1)(1-\alpha/2)}$ as quantiles to put in the formula for the Normal CI instead of $z_{\alpha/2}$. That is, the studentized CI is defined as

$$\hat{\theta} - Z^*_{(N+1)\alpha/2}\hat{se} \leq \theta \leq \hat{\theta} - Z^*_{(N+1)(1-\alpha/2)}\hat{se}$$

This CI, unlike the three previous ones, is generally considered to be very reliable. However, there is a catch, and this catch is the estimate of the standard deviation of $\hat{\theta}^*$, $\hat{se}^*$. This estimate is not easily obtained. You can get it either parametrically (but then you need a model which you probably don't have) or through re-sampling. This means that you have to do an extra bootstrap for each of the original bootstrap samples. That is, you will have a loop within a loop, and it can become very heavy computationally.

6 software

Will be done in R. Functions you may find useful:
sample, replicate, rgamma, boot, boot.ci, dgamma, nlm

7 Laboration: Bootstrap re-sampling.

The aim of this laboration is to study the performance of bootstrap estimates, as well as corresponding variance, bias and CI, using different bootstrap techniques.

7.1 Tasks

1. **Non-parametric bootstrap.** We start with the easier task of setting up a non-parametric bootstrap. First, generate some data to work with. Let us say that the data comes from a definitely non-normal, highly skewed distribution. Such as, for example, Gamma.

    - Generate a data set of size $n = 100$ from a $Gamma(k, 1/\gamma)$ distribution, where $k = 1$ is the shape parameter and γ the scale parameter. Do a histogram of the data. What are the theoretical mean and variance of a *Gamma* distribution? Observe that *Gamma* can be defined both through γ and $1/\gamma$. Look out for which definition R uses.
    - Let us say that our parameter of interest θ is the mean. Write a function that approximates the distribution of θ with non-parametric bootstrap. Present the distribution in a histogram. What is the estimated bias of $\hat{\theta}$? Variance?
    - As with many other things nowadays, there already exist a package that does bootstrap for you. In R this package is **boot**. Use it to get the distribution of the mean and the estimates of the bias and variance.
    - Construct CI for θ (normal, basic, percentile). Check your results with the pre-programmed function **boot.ci**.
    - Bootstrap is known to give less than reliable results if the sample size n is small. Study this by doing the following. Fix an $n = 10$. Draw 1000 samples of this size. For each sample, calculate bootstrap CI, recording whether they cover the true value of θ (the theoretical mean, that is). Repeat this for $n = 20, 30, ..., 100$. Do the CI have the correct coverage probabilities? Present results in a graph. You can use either your own code or **boot.ci**.

2. **Parametric bootstrap.** (2 p) Repeat the previous task using parametric bootstrap instead of non-parametric. In order to do this you have to estimate the k and γ parameters in the Gamma distribution through, for example, Maximum Likelihood. A nice, closed form ML solution for the parameters does not exist, so you will have to write a function that finds the maximum of the Likelihood function L numerically.

    Tip 1: most optimization packages look for minimum and not maximum, so work with -log(L) instead of L itself.

    Tip 2: Look up **nlm** function.

3. **Studentized CI.** (2 p) Our parameter of interest θ is no longer the mean, it is the variance.

    - Repeat the last question in task 1. What can be said about the performance of the standard confidence intervals?

- Construct Studentized confidence intervals for the variance. Check the coverage as in the above. Any improvement?

You can construct the studentized CI both from scratch with your own code and using **boot** and **boot.ci**. If you choose to do the latter, make sure that the "statistic" function that you have to specify in **boot** returns two values: $\hat{\theta}^*$ and $\mathrm{Var}(\hat{\theta}^*)$.

In order to pass the lab task 1 has to be completed.