12 The Bootstrap and why it works

For a review of many applications of bootstrap see Efron and Tibshirani (1994). For the theory behind the bootstrap see the books by Hall (1992), van der Waart (2000), Lahiri (2003) and Politis and Romano (1999).

12.1 The Bootstrap methodology

The heuristics above give us an explanation as to why the asymptotic normality assumption may not be particularly good for small samples. Bootstrap is a form of sampling from the data, which tries to ‘capture’ features in the distribution which the over simplified normal approximation cannot do. Resampling methods have been in the statistical literature for over 50 years. However, it was Efron who proposed the bootstrap as it is today, and really brought to attention the its importance in solving various statistical problems.

The bootstrap is a tool, which allows us to obtain better finite sample approximation of estimators. The bootstrap is used all over the place to estimate the variance, correct bias and construct CIs etc. There are many, many different types of bootstraps.

Here we describe two simple versions of the bootstrap for constructing CIs. They can be roughly described as the nonparametric bootstrap and the parametric bootstrap (in my opinion the nonparametric bootstrap is more flexible).

12.1.1 The nonparametric bootstrap confidence interval for the mean

We will assume that \{X_t\} are iid random variables with mean \(\mu\), variance \(\sigma^2\) and the fourth moment exists. To simplify the explanation we will assume the variance of \{X_t\} is known. All the sampling properties of the bootstrap procedure that we describe also hold when the the variance is unknown (in which case we need to use what is called the studentised bootstrap), however more sophisticated techiques and greater care has to be used to prove the results.

Let us consider the sample mean \(\bar{X} = \frac{1}{T} \sum_{t=1}^{T} X_t\). As we mentioned above asymptotically the distribution of \(\sqrt{T}\frac{\bar{X} - \mu}{\sigma}\) is normal, and the asymptotic \((1 - \alpha)100\%\) confidence interval of the mean \(\mu\) is

\[
\left[ \bar{X} + z_{\alpha/2}\sigma/\sqrt{T}, \bar{X} + z_{1-\alpha/2}\sigma/\sqrt{T} \right].
\]

But we want to obtain a better approximation of the true confidence interval

\[
\left[ \bar{X} + \xi_{\alpha/2}\sigma/\sqrt{T}, \bar{X} + \xi_{1-\alpha/2}\sigma/\sqrt{T} \right],
\]

where \(\xi_{\alpha}\) is the \(\alpha\) percent quantile of the distribution \(G_T\) which is the actually distribution of \(\sqrt{T}\left(\frac{\bar{X} - \mu}{\sigma}\right)\). However, we can obtain an estimator of \(G_T\). We recall that we observe the iid
random variables \( \{X_t\}_{t=1}^{T} \), the distribution function of \( X_t \) is \( F(\cdot) \). In nonparametric bootstrap we do not know the distribution of \( F(\cdot) \), but we can estimate it with the empirical distribution function

\[
F_T(x) = \frac{1}{T} \sum_{t=1}^{T} I(X_t \leq x),
\]

we note that though \( F_T(x) \) is random (since it depends on a sample), it is a proper distribution function and has a mean which is \( \bar{X} \) and variance \( \hat{\sigma} = \frac{1}{T} \sum_{t=1}^{T} (x_t - \bar{X})^2 \). Now we recall that \( G_T(x) \) is basically the distribution of \( \sqrt{T} \frac{\sum_{t=1}^{T} X_t - \mu}{\sigma} \), where \( X_t \) are independent draws from the unknown distribution \( F \). Hence, if we wanted to estimate \( G_T(x) \) and did not have \( F(\cdot) \) to sample, then it would be natural to sample from the the distribution that we do have available which is \( F_T(\cdot) \). We use the following algorithm:

(i) We could sample \( T \) independent times from \( F_T(\cdot) \) to obtain the bootstrap sample \( \{X_{T,1}^*, \ldots, X_{T,T}^*\} \). Using this we obtain the bootstrap estimator of the mean, \( \bar{X}_{T,1}^* = \frac{1}{T} \sum_{t=1}^{T} X_{t,1}^* \). As this is a sample from the empirical distribution function \( F_T \), the mean of \( \bar{X}_{T,1}^* \) is the mean of \( F_T \), which is \( \bar{X} \) (recall the mean of \( \bar{X} \) is the mean \( \mu \), which is the mean of the distribution \( F \)). We note this is equivalent to drawing from \( \{X_1, \ldots, X_T\} \) \( T \)-times with replacement.

(ii) We do this multiple times. In fact one can draw \( T^T \) different samples. For each bootstrap sample we calculate the sample mean, so that we have \( \{\bar{X}_{T,1}^*, \ldots, \bar{X}_{T,n}^*\} \), where \( n = T^T \). Based on this we can construct the bootstrap estimator of the distribution \( G_T(x) \) which is

\[
\hat{G}_T(x) = \frac{1}{T^T} \sum_{k=1}^{T^T} I\left( \sqrt{T} \left[ \frac{\bar{X}_{T,k}^* - \bar{X}}{\hat{\sigma}} \right] \leq x \right),
\]

we use \( \bar{X} \) and \( \hat{\sigma}/\sqrt{T} \) in the definition of \( \hat{G}_T \) because this is the mean and variance of \( \bar{X}_{T}^* \) based on sampling from \( F_T \). Now if \( F_T(x) \) were the true distribution of \( X_t \), then \( \hat{G}_T(x) = G_T(\cdot) \). Of course it is not, so \( \hat{G}_T(x) \) is only an estimator of \( G_T(\cdot) \). I reality this may not be possible to obtain all \( T^T \) samples (this is a lot!), but we sample enough times and in a good way to obtain a good enough approximation of \( \hat{G}_T(x) \). We will assume that we can obtain \( \hat{G}_T(x) \).

(iii) Since \( \hat{G}_T(x) \) is an estimator of \( G_T(\cdot) \) we can obtain an estimator of the quantiles. Thus we can use this to obtain an estimator of the CIs and hope that it is more accurate than the standard normal approximation. Let \( \xi_\alpha \) be such that \( \hat{G}_T(\xi_\alpha) = \alpha \).
The 95% CI bootstrap CI of the mean $\mu$ is

$$[\bar{X} + \xi_{\alpha/2}\sigma/\sqrt{T}, \bar{X} + \xi_{1-\alpha/2}\sigma/\sqrt{T}]$$

12.1.2 The parametric bootstrap confidence interval of an estimator

Let us suppose $\{X_t\}$ are iid random variables with distribution $f(\cdot; \theta_0)$, where the parameter $\theta_0$ is unknown. Suppose we use mle to estimate the parameter $\theta_0$, which we denote as $\hat{\theta}_T$. We know that if all the regularity conditions are satisfied then we have

$$\sqrt{T}(\hat{\theta}_T - \theta_0) \xrightarrow{D} \mathcal{N}(0, I(\theta_0)^{-1}),$$

where $I(\theta_0) = \int \left( \frac{\partial f(x; \theta_0)}{\partial \theta} \right)^2 \theta=\theta_0 f(x; \theta_0) dx$. Of course this is an asymptotic result. If the sample size is small, we may want to obtain a better finite sample approximation of $\sqrt{T}(\hat{\theta}_T - \theta_0)$, to construct better CIs. Let $G_T$ denote the distribution of $\sqrt{T}(\hat{\theta}_T - \theta_0)$.

(i) We now sample $T$ independent times from the distribution $f(x; \hat{\theta}_T)$, and for each bootstrap sample $X^*_T, 1 = (X^*_1, \ldots, X^*_T)$ construct the bootstrap mle $\hat{\theta}^*_T, 1$. We do this many times. We denote the $k$th bootstrap estimator as $\hat{\theta}^*_T, k$.

(ii) Unlike the nonparametric bootstrap there is likely to be an infinite number of draws one can make. Hence we cannot construct an estimator of $G_T$ using all possible draws of $f(x, \hat{\theta}_T)$. But one can construct an estimate of the finite sample distribution of $\hat{\theta}_T - \theta_0$ using a large number of draws. Let

$$\hat{G}_T(x) = \frac{1}{n} \sum_{k=1}^{n} I\left( \sqrt{T}(\hat{\theta}^*_T, k - \hat{\theta}_T) \leq x \right).$$

(iii) Let $\hat{\xi}_\alpha$ be such that $\hat{G}_T(\hat{\xi}_\alpha) = \alpha$. The 95% CI bootstrap CI of the mean $\theta_0$ is

$$[\hat{\theta}_T + \frac{1}{\sqrt{T}}\hat{\xi}_{\alpha/2}, \hat{\theta}_T + \frac{1}{\sqrt{T}}\hat{\xi}_{1-\alpha/2}].$$

• An alternative way to construct the CI is to use the likelihood ratio test.

We recall that if $f(\cdot; \theta_0)$ is the true distribution then $\mathcal{L}_T(\hat{\theta}_T) - \mathcal{L}_T(\theta_0) \xrightarrow{D} \chi^2_p$ and we can use this result to construct the $100(1 - \alpha)$% CI (see the section on confidence intervals).

But if the sample size is small, and we believe it that normality result is a poor approximation - hence the chi-squared result would also be a poor approximation. We can use a bootstrap method instead. In this case, every bootstrap estimator $\hat{\theta}^*_{T,k}$, we plug it into the log-likelihood

$$\mathcal{L}_T(\hat{\theta}^*_T) = \sum_{t=1}^{T} \log f(X_i; \hat{\theta}^*_T).$$

77
We can construct an estimator of the distribution function of $L_T(\hat{\theta}_T) - L_T(\theta_0)$, which we denote as $H_T$ as

$$\hat{H}_T(x) = \frac{1}{n} \sum_{k=1}^{n} I\left( \frac{L_T(\hat{\theta}_{k,T}) - L_T(\hat{\theta}_T)}{n} \leq x \right).$$

Let $\hat{\xi}_\alpha$ be such that $\hat{H}_T(\hat{\xi}_\alpha) = \alpha$. The 100$(1 - \alpha)$% CI for $\theta$ based on the log-likelihood ratio is

$$\{ \theta; L_T(\theta) > L_T(\hat{\theta}_T) - \hat{\xi}_{1-\alpha} \}.$$

In reality the parametric bootstrap is not used as much as the nonparametric bootstrap. The main reason is that in the misspecified case, the CIs produced have no meaning (they are incorrect) and will not even converge to the CIs produced by using the normal approximation (using the misspecified variance $I_{\theta_0}^{-1}J_{\theta_0}I_{\theta_0}$).

### 12.2 Using Edgeworth expansions to show why the nonparametric bootstrap works

On first reading the bootstrap may seem a little like magic. But really it is not. We recall that $G_T$ is the distribution of $\bar{X}$ based on sampling from $F$. Since in reality $F$ is unobserved, and can only be estimated using the empirical distribution function $F_T$, it does not seem unnatural that $\hat{G}_T$ can be used as an estimator of $G_T$.

We first state a consistency result, the proof can be found in various places, see for example Hall (1992) or van der Waart (2000). There are different ways this can be proven, in the more complex setting where we are not estimating the mean, using something like the Mallows distance (which is the measure which measures the distance between distributions) may be the most appropriate method for proving the result.

**Theorem 12.1 (Consistency)** Suppose that $\mathbb{E}(X_1^4) < \infty$. Then $\sqrt{T}(\bar{X} - \mu) \overset{D}{\to} \mathcal{N}(0,1)$ (noting that $\sqrt{T}(\bar{X} - \mu) \overset{D}{\to} \mathcal{N}(0,1)$).

The value of the above result is that is shows that the bootstrap distribution $\hat{G}_T$ converges to the standard normal, just like $G_T$ converges to the standard normal. Hence we do no lose by using the bootstrap approximation of the CIs.

We now show what we can gain by using the bootstrap. Let us recall (57)

$$G_T(x) = P(S_T \leq x) = \Phi(x) + \frac{1}{T^{1/2}} p_1(x) \phi(x) + \frac{1}{T^{3/2}} p_2(x) \phi(x) + \frac{1}{T^{5/2}} p_3(x) \phi(x) \ldots, \quad (58)$$
where $\Phi$ is the distribution of the standard normal, $\phi(x)$ is the standard normal density and

$$p_1(x) = -\frac{1}{6} \tilde{\kappa}_3(x^2 - 1) \quad \text{and} \quad p_2(x) = -x\left\{ \frac{1}{24} \tilde{\kappa}_4(x^2 - 3) + \frac{1}{72} \tilde{\kappa}_5^2(x^4 - 10x^2 + 15) \right\}.$$  

We now rewrite the above results in terms of the underlying distribution of the random variables. Let us suppose the distribution of the iid random variables $\{X_t\}$ is $F$. Then rewriting the above we have

$$G_T(x) = P(S_T \leq x | F) = \Phi(x) + \frac{1}{T} \tilde{p}_1(x | F) \phi(x) + \frac{1}{T} \tilde{p}_2(x | F) \phi(x) + \ldots, \quad (59)$$

where $p_1(x | F), p_2(x | F)$ etc. are the polynomials, whose coefficients are determined by the cumulants

$$p_1(x | F) = -\frac{1}{6} \kappa_3 \left( \frac{X - \mu(F)}{\sigma(F)} | F \right) (x^2 - 1)$$
$$p_2(x | F) = -x \left\{ \frac{1}{24} \kappa_4 \left( \frac{X - \mu(F)}{\sigma(F)} | F \right) (x^2 - 3) + \frac{1}{72} \kappa_5 \left( \frac{X - \mu(F)}{\sigma(F)} | F \right)^2 (x^4 - 10x^2 + 15) \right\}.$$

$\mu(F) = E_F(X), \sigma(F)^2 = E_F(X^2) - (E_F(X))^2,$

$$\kappa_3(X | F) = E_F(X^3) - E_F(X)^3$$
$$\kappa_4(X | F) = E_F(X^4) - 3E_F(X^2)^2 - 3E(X)E(X^3) + E_F(X)^4$$

$$\kappa_3 \left( \frac{X - \mu(F)}{\sigma(F)} | F \right) = \sigma(F)^{-3/2} \kappa_3(X | F) \quad \kappa_4 \left( \frac{X - \mu(F)}{\sigma(F)} | F \right) = \sigma(F)^{-2} \kappa_4(X | F)$$

and $E_F(X) = \int x dF(x)$.

This leads us to something rather fascinating. We recall that the bootstrap distribution $\hat{G}_T(x)$ is an approximation of the finite sample distribution $G_T$. $G_T$ is determined by the measure $F$ and the bootstrap distribution is based entirely on the (random) measure $\hat{F}_T$. Hence conditioning on the distribution $\hat{F}_T$ by using (59), we have the Edgeworth expansion of the random measure $\hat{G}_T(x)$ conditioned on $\hat{F}_T$ which is

$$\hat{G}_T(x) = P(S_T \leq x | \hat{F}_T) = \Phi(x) + \frac{1}{T^{3/2}} \tilde{p}_1(x | \hat{F}_T) \hat{\phi}(x) + \frac{1}{T} \tilde{p}_2(x | \hat{F}_T) \hat{\phi}(x) + \ldots,$$

where $p_1(x | \hat{F}_T), p_2(x | \hat{F}_T)$, are random and given by

$$p_1(x | \hat{F}_T) = -\frac{1}{6} \hat{\sigma}(\hat{F}_T)^{-3/2} \kappa_3 \left( X | \hat{F}_T \right) (x^2 - 1) = -\frac{1}{6} \hat{\sigma}^{-3/2} \tilde{\kappa}_3(x^2 - 1)$$
$$p_2(x | \hat{F}_T) = -x \left\{ \frac{1}{24} \hat{\sigma}^{-2} \tilde{\kappa}_4(x^2 - 3) + \frac{1}{72} \hat{\sigma}^{-3} \tilde{\kappa}_5^2(x^4 - 10x^2 + 15) \right\},$$

since the mean of $\hat{F}_T$ is $X$ the variance of $\hat{F}_T$ is $\hat{\sigma}^2$, the rth order cumulant of $\hat{F}_T$ is the empirical cumulant $\hat{\kappa}_r$. 

79
Remark 12.1 Since $\hat{F}_T(x)$ is the distribution function of a discrete random variable, which gives the weight $1/T$ to the event $X_t$ and zero otherwise we see that

$$\mathbb{E}_{\hat{F}_T}(X^r) = \frac{1}{n} \sum_r X^r_t,$$

hence we obtain that the mean with respect to $\hat{F}_T$ contains the third order cumulant, whereas $\hat{\kappa}_3 = \frac{1}{n} \sum_r [X^r_t - \hat{E}_{\hat{F}_T}(X^r_t)]^3$.

Let us now compare this result with the normal approximation in (58). This gives us

$$\hat{G}_T(x) = P(S_T \leq x|\hat{F}_T) = \Phi(x) + \frac{1}{T/2} \tau_1(x|\hat{F}_T) \phi(x) + \frac{1}{T} \tau_2(x|\hat{F}_T) \phi(x) + \ldots,$$

$$\bar{G}_T(x) = P(S_T \leq x|\bar{F}_T) = \Phi(x) + \frac{1}{T/2} \bar{\tau}_1(x|\bar{F}_T) \phi(x) + \frac{1}{T} \bar{\tau}_2(x|\bar{F}_T) \phi(x) + \ldots,$$

where

$$\bar{\tau}_1(x|\bar{F}_T) = -\frac{1}{6} \sigma^{-3/2} \kappa_3(x^2 - 1)$$

$$\bar{\tau}_2(x|\bar{F}_T) = -x \left\{ \frac{1}{24} \sigma^{-2} \kappa_4(x^2 - 3) + \frac{1}{72} \sigma^{-3} \kappa_3^2(x^4 - 10x^2 + 15) \right\}$$

$$\bar{\tau}_2(x|\bar{F}_T) = -x \left\{ \frac{1}{24} \sigma^{-2} \kappa_4(x^2 - 3) + \frac{1}{72} \sigma^{-3} \kappa_3^2(x^4 - 10x^2 + 15) \right\}.$$

Therefore taking differences gives

$$G_T(x) - \bar{G}_T(x) = \frac{1}{T^{1/2}} \left( \tau_1(x|F) - \bar{\tau}_1(x|\bar{F}_T) \right) \phi(x) + \frac{1}{T} \left( \tau_2(x|F) - \bar{\tau}_2(x|\bar{F}_T) \right) \phi(x) + \ldots.$$

Now this is whether the bootstrap distribution becomes very useful, we recall see that $\tau_1(x|F)$ contains the third order cumulant, whereas $\bar{\tau}_1(x|\bar{F}_T)$ contains it’s estimator. We recall that $|\hat{X} - \mu| = O_p(T^{-1/2})$, $|\hat{\sigma} - \sigma| = O_p(T^{-1/2})$, the same is true for $\hat{\kappa}_3$ and $\hat{\kappa}_4$, that is $|\hat{\kappa}_3 - \kappa_3| = O_p(T^{-1/2})$, $|\hat{\kappa}_4 - \kappa_4| = O_p(T^{-1/2})$. Substituting this into (60) leads to

$$|G_T(x) - \bar{G}_T(x)| = O_p\left( \frac{1}{T} \right).$$

Let us now compare this result with the normal approximation in (58). This gives us

$$|G_T(x) - \Phi(x)| = O_p\left( \frac{1}{T^{1/2}} \right)$$

Hence we observe that the bootstrap distribution $\hat{G}_T(x)$ leads to a better approximation of the finite sample distribution than the normal approximation.

Now by using the Cornish-Fisher expansions one can show that

$$|\hat{\xi}_x - \xi_x| = O_p(T^{-1})$$
comparing with

\[ |\xi_\alpha - z_\alpha| = O_p(T^{-1}). \]

Hence the confidence intervals constructed using the bootstrap are more accurate than the CIs using the normal approximation.

**Remark 12.2**  (i) In the case that \( \sigma \) is unknown, a similar result can be applied, but the calculations become more complicated.

However, it is always better to try and transform the estimator into a quantity which is asymptotically pivotal. We recall a distribution is asymptotically pivotal if its limiting distribution does not depend on the parameters. We recall that asymptotically the distribution of \( \sqrt{T \bar{X}} \) depends on the variance \( \sigma^2 \). If we bootstrap \( \sqrt{T \bar{X}} \) (instead of \( \sqrt{T \bar{X} - \mu} / \sigma \)) and the variance \( \sigma^2 \) is unknown, we may not gain by in terms of approximations.

(ii) Please observe that the calculations above are heuristic since we have not given conditions under which the expansion are true.

(iii) Similar arguments to those given above can also be applied to bootstrapping of other parameters, \( \theta \), besides the mean. They can also be generalised to dependent data and far more complicated situations.