1 Introduction

Let $\theta = \phi(P)$ be a parameter of $P$. It could be finite-dimensional (e.g. mean or median), but not necessarily (e.g. characteristic function, density). Let $\hat{\theta}_n = \hat{\theta}(X_1, \ldots, X_n)$ be an estimator of $\theta$ given $n$ data points from $P$.

The frequentist methodology is about finding performance measures such as bias or error bars for estimators. For instance the bias of $\hat{\theta}_n$ is $E_P(\hat{\theta}_n - \theta)$. This expectation is an average over the possible data sets that the true distribution $P$ could have generated, but of course the difficulty is that we do not know $P$, and are given only one dataset. Then how can we hope to evaluate this quantity?

We have already seen several methods. One way to go is to compute the bias (or some other performance measure) of the estimator for any $P$. If we’re lucky and this error does not depend on $P$ then we do not need to know $P$. In that case asymptotics can be viewed as giving us a way to approximate the value of this error for large $n$ when the true answer is intractable. Yet even if it does depend on $P$, asymptotics may reveal that the bias of an estimator depends little on $P$ for large $n$, locally.

The bootstrap takes a different perspective. Even though we do not know $P$, we know an approximation of it: the empirical distribution $\hat{P}_n$. If we generate a new sample $X_1^*, \ldots, X_n^*$ of size $n$ from $\hat{P}_n$, we can compare $\hat{\theta}_n = \theta(X_1^*, \ldots, X_n^*)$ and the (known) true answer $\phi(\hat{P}_n)$. If $\hat{P}_n$ is close to $P$, the distribution of $\hat{\theta}_n^*$ relative to $\phi(\hat{P}_n)$ should be close to that of $\hat{\theta}_n$ relative to $\phi(P)$. Taking the bias as example, we can estimate the bias $E_P(\hat{\theta}_n - \phi(\hat{P}_n))$, where $\theta$ is unknown, using $E_{\hat{P}_n}(\hat{\theta}_n^* - \phi(\hat{P}_n))$, which can be computed.

We thus have a parallel between the “real world”:

$$P \quad \rightarrow \quad X_1, \ldots, X_n \quad \rightarrow \quad \hat{\theta}_n = \theta(X_1, \ldots, X_n)$$

(1)

and the “bootstrap world”:

$$\hat{P}_n \quad \rightarrow \quad X_1^*, \ldots, X_n^* \quad \rightarrow \quad \hat{\theta}_n^* = \theta(X_1^*, \ldots, X_n^*)$$

(2)

where every occurrence of $P$, including in the generation of the data, is replaced by $\hat{P}_n$.

More concretely, how do we compute performance measures in the bootstrap world, such as $E_{\hat{P}_n}(\hat{\theta}_n^* - \phi(\hat{P}_n))$?

Since $\hat{P}_n$ is discrete there only exists a finite number of $n$-samples that we can draw from it. Since we assume independence, $\theta$ in general does not depend on the order of the sample, so we only need to enumerate the vectors $M_n = (M_{n,1}, \ldots, M_{n,n})$, where $M_{n,i}$ is the number of $X_i^*$’s that are equal to $X_i$. The above expectation is thus a sum over each such vector, weighted by its (multinomial) probability:

$$E_{\hat{P}_n}(\hat{\theta}_n^* - \phi(\hat{P}_n)) = \sum_{M_n} \frac{n!}{\prod_i M_{n,i}!} \theta(M_n) - \phi(\hat{P}_n)$$

(3)

In practice, since there are (in general) exponentially many $M_n$, we must resort to an approximation, usually Monte Carlo integration. In other words we draw $B$ times a sample of $n$ data points i.i.d. from $\hat{P}_n$, or
equivalently we draw $B$ times $M_n$ from the multinomial distribution $\text{Mult}(n; \frac{1}{n}, \ldots, \frac{1}{n})$. Each sample gives us a value $\hat{\theta}_i^n$ for the estimator. By the law of large numbers we have

$$\frac{1}{B} \sum_{b=1}^{B} \theta(M_n^b) - \phi(P_n) \longrightarrow E_{\hat{P}_n}(\hat{\theta}_n^* - \phi(P_n)) \quad (4)$$

Sampling from $\hat{P}_n$ is natural since it parallels the way we obtained samples from $P$, and since Monte Carlo is the method most often used many people tend to see the bootstrap as a resampling technique, but the key characteristic of the bootstrap is the substitution of $\hat{P}_n$ for $P$ in the evaluation of the performance measure, not the sampling.

2 Performance measures

We have already mentioned the bias, and here we give other examples. When $P$ is a distribution over $\mathbb{R}$, it can equivalently be represented as $F$, its cumulative distribution function (cdf).

- bias: $\lambda_n(F) = E_{F}(\hat{\theta}_n) - \phi(F)$
- variance: $\lambda_n(F) = E_{F}(\hat{\theta}_n - E_{F}\hat{\theta}_n)^2$
- cdf of the error distribution: $f_n(t) = P_{F}(\sqrt{n}(\hat{\theta}_n - \phi(F)) \leq t)$
- scaled cdf of error dist.: $f_n(t) = P_{F}\left(\frac{\sqrt{n}(\hat{\theta}_n - \phi(F))}{\tau(F)} \leq t\right)$

In all these cases, the bootstrap substitutes $\hat{F}_n$ for $F$. The question is whether $\lambda_n(\hat{F}_n)$ converges to $\lambda_n(F)$. In particular the question of this convergence, a statistical question, should not be confused with that of the convergence to $\lambda_n(\hat{F}_n)$ of its approximation based on $B$ bootstrap re-samples, which is more a computational question and specific to the Monte Carlo method to compute $\lambda_n(\hat{F}_n)$.

3 Example: bias of the median

Let $\theta = \phi(F)$ be the median of $F$, and take the sample median as estimator: $\hat{\theta}_n = \theta(X_1, \ldots, X_n) = \phi(\hat{F}_n)$, i.e. we use a plugin estimator. The performance measure we look at is the bias: $\lambda_n(F) = E_{F}(\hat{\theta}_n) - \theta$.

Now let’s go through the calculation for $n = 3$.

We are given three data points $X_1, X_2, X_3$ from $F$, with order statistics $X_{(1)} = b, X_{(2)} = c, X_{(3)} = d$ (i.e. $b < c < d$ if we assume the data points are distinct). The order statistics of the bootstrap sample, $(X_{(1)}^*, X_{(2)}^*, X_{(3)}^*)$ have the following distribution:

<table>
<thead>
<tr>
<th>$(X_{(1)}^<em>, X_{(2)}^</em>, X_{(3)}^*)$</th>
<th>bbb</th>
<th>bbc</th>
<th>bbd</th>
<th>bbc</th>
<th>bcd</th>
<th>bdd</th>
<th>ccc</th>
<th>ccd</th>
<th>cdd</th>
<th>ddd</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>b</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>$\hat{\theta}_n^*$</td>
<td>b</td>
<td>b</td>
<td>b</td>
<td>c</td>
<td>c</td>
<td>d</td>
<td>c</td>
<td>c</td>
<td>d</td>
<td>d</td>
</tr>
</tbody>
</table>

In particular we obtain that

$$P(\hat{\theta}_n^* = b) = \frac{7}{27} \quad P(\hat{\theta}_n^* = c) = \frac{13}{27} \quad P(\hat{\theta}_n^* = d) = \frac{7}{27} \quad (5)$$
\[ \lambda_n(\hat{F}_n) = \frac{1}{27} \left( 7X(1) + 13X(2) + 7X(3) \right) \quad (7) \]

\[ = 14 \left( \frac{X(1)}{2} + \frac{X(3)}{2} - X(2) \right) \quad (8) \]

### 4 Consistency

The following example comes from Lehmann [1999], Example 6.5.7, pp. 429-430. Say we look at this performance measure: \( \lambda_n(F) = E(\sqrt{n}\theta_n)^2 \). Assume that \( \hat{\theta}_n \) is a U-statistic: 

\[ \hat{\theta}_n = \frac{1}{n(n-1)} \sum_{i \neq j} \psi(X_i, X_j) \]

A previous result gives us, in the notation of Lehmann [1999]:

\[ \lambda_n(F) = 4 \left( \frac{n-2}{n-1} \gamma_1^2 + \frac{2}{n-1} \gamma_2^2 \right) \quad \] where 
\[ \gamma_1^2 = E_F \psi(X_1, X_2) \psi(X_1, X_3) \quad \text{and} \quad \gamma_2^2 = E_F \psi(X_1, X_2)^2 \quad \text{(Note that these are the raw moments of \( \psi \) while van der Vaart [1998] p. 163 uses the centered moments).} \]

Therefore \( \lambda_n(F) \to 4\gamma_1^2 \).

All the above quantities are expectations with respect to the true density \( F \). The bootstrap estimate of \( \lambda_n(F) \) is \( \lambda_n(\hat{F}_n) = 4 \left( \frac{n-2}{n-1} \gamma_1^2 + \frac{2}{n-1} \gamma_2^2 \right) \) where

\[ \gamma_1^2 = \frac{1}{n^3} \sum_{X_i, X_j, X_k \in \{X_1, X_2, X_3\}} \psi(X_i, X_j) \psi(X_i, X_k) \quad (9) \]

\[ \gamma_2^2 = \frac{1}{n^2} \sum_{i,j} \psi(X_i, X_j)^2 \quad (10) \]

If \( \gamma_1^2, \gamma_2^2 \) and \( \gamma_3^2 \) are finite, then by the law of large numbers\(^1\) \( \gamma_1^2 \to \gamma_1^2 \) and \( \gamma_2^2 \to \gamma_2^2 \), which implies consistency.

Now consider the case where \( \gamma_3^2 = \infty \), so that in

\[ \gamma_2^2 = \frac{1}{n^2} \sum_{i \neq j} \psi(X_i, X_j)^2 + \frac{1}{n^2} \sum_i \psi(X_i, X_i)^2 \quad (12) \]

the second term does not converge. For example take:

\[ X_i \overset{i.i.d.}{\sim} U(0, 1) \quad (13) \]

\[ |\psi(X, Y)| \leq M \quad (14) \]

\[ \psi(X, X) = e^X \quad (15) \]

The last term we can define any way we want to construct our example. It does not affect the value of the non-bootstrap estimator since data points have probability 0 of being equal under the true distribution. Therefore \( \lambda_n(F) \) is finite but divergence of the bootstrap estimator happens in the second term of (12) if 

\[ P\left( \frac{1}{n^2} \sum_i e^{X_i} > A \right) \to 1. \]

It suffices to show that 

\[ P(\max_i e^{X_i} > An^2) \to 1 \]

i.e. 

\[ P(e^{\frac{X_i}{n}} < An^2)^n \to 0 \]

or 

\[ (1 - \frac{2 \log(An^2)}{2n})^n \to 0. \]

This is true since \( 2 \log(An^2) < \sqrt{n} \) for large enough \( n \).

\(^1\)or is it by the consistency of V-statistics?
References
