1 Subsampling

Suppose $X_i, i = 1, \ldots, n$ is an i.i.d. sequence of random variables with distribution P. Let $\theta(P)$ be some real-valued parameter of interest, and let $\hat{\theta}_n = \hat{\theta}_n(X_1, \ldots, X_n)$ be some estimate of $\theta(P)$. It is natural to assume in the i.i.d. setting that $\hat{\theta}_n$ is symmetric in its arguments, so we will make this assumption below, but it is not necessary. Consider the root

$$R_n = R_n(X_1, \ldots, X_n, \theta(P)) = \tau_n(\hat{\theta}_n - \theta(P)) \ ,$$

where τ_n is some normalizing sequence. Often, $\tau_n = \sqrt{n}$, but we do not wish to assume this. Let $J_n(x, P)$ denote the distribution of R_n; that is,

$$J_n(x, P) = \Pr\{ \ \tau_n(\hat{\theta}_n - \theta(P)) \leq x\} \ .$$

We wish to estimate $J_n(x, P)$ so we can make inferences about $\theta(P)$. For example, we would like to estimate quantiles of $J_n(x, P)$, so we can construct confidence sets for $\theta(P)$. Unfortunately, we do not know P, and, as a result, we do not know $J_n(x, P)$.

The bootstrap solved this problem simply by replacing the unknown P with an estimate $\hat{P}_n$. In the case of i.i.d. data, a typical choice of $\hat{P}_n$ is the empirical distribution of the $X_i, i = 1, \ldots, n$. For this approach to work, we essentially required that $J_n(x, P)$ when viewed as a function of P was continuous in a certain neighborhood of P. We will now explore an alternative to the bootstrap known as subsampling that does not impose this requirement. Subsampling is originally due to Politis and Romano (1994).

In order to motivate the idea behind subsampling, consider the following thought experiment. Suppose for the time being that $\theta(P)$ is known. Suppose that, instead of n i.i.d. observations from P, we had a very, very large number of i.i.d. observations from P. For concreteness, suppose $X_i, i = 1, \ldots, m$ is an i.i.d. sequence of random variables with distribution P with $m = nk$ for some very big k. We could then estimate $J_n(x, P)$ by looking at the empirical distribution of

$$\tau_n(\hat{\theta}_n(X_{n(j-1)+1}, \ldots, X_{nj}) - \theta(P)), j = 1, \ldots, k \ .$$

This is an i.i.d. sequence of random variables with distribution $J_n(x, P)$. Therefore, by the Glivenko-Cantelli theorem, we know that this empirical distribution is a good estimate of $J_n(x, P)$, at least for large k. In fact, with a simple trick, we could show that it is even possible to improve upon this estimate by using all possible sets of data of size n from the m observations, not just those that are disjoint; that is, estimate $J_n(x, P)$ with the empirical distribution of the

$$\tau_n(\hat{\theta}_{n,j} - \theta(P)), j = 1, \ldots, \binom{m}{n} ,$$

where $\hat{\theta}_{n,j}$ is the estimate of $\theta(P)$ computed using the jth set of data of size n from the original m observations.

In practice $m = n$, so, even if we knew $\theta(P)$, this idea won't work. The key idea behind subsampling is the following simple observation: replace n with some smaller number b that is much smaller than n. We would then expect

$$\tau_b(\hat{\theta}_{b,j} - \theta(P)), j = 1, \ldots, \binom{n}{b} ,$$

where $\hat{\theta}_{b,j}$ is the estimate of $\theta(P)$ computed using the jth set of data of size b from the original n observations, to be a good estimate of $J_b(x, P)$, at least if $\binom{n}{b}$ is large. Of course, we are interested in $J_n(x, P)$, not $J_b(x, P)$. We therefore need some way to force $J_n(x, P)$ and $J_b(x, P)$ to be close to one another. To ensure this, it suffices to assume that $J_n(x, P) \to J(x, P)$. Therefore, $J_b(x, P)$ and $J_n(x, P)$ are both close to $J(x, P)$, and thus close to one another as well, at least for large b and n. In order to ensure that both b and $\binom{n}{b}$ are large, at least asymptotically, it suffices to assume that $b \to \infty$, but $b/n \to 0$.

This procedure is still not feasible because in practice we typically do not know $\theta(P)$. But we can replace $\theta(P)$ with $\hat{\theta}_n$. This would cause no problems if

$$\tau_b(\theta_n - \theta(P)) = \frac{\tau_b}{\tau_n}\tau_n(\theta_n - \theta(P))$$

were small. Since $\tau_n(\theta_n - \theta(P)) = O_P(1)$, it is enough to assume that

$\tau_b/\tau_n \to 0$. Typically, this assumption will be implied by the assumption that $b/n \to 0$, but it may not be, so we need to assume it separately.

The next theorem formalizes the above discussion.

Theorem 1.1 Let $X_i, i = 1, \ldots, n$ be an i.i.d. sequence of random variables with distribution P. Let $\theta(P)$ be a real-valued parameter of interest, and let $\hat{\theta}_n = \hat{\theta}_n(X_1, \ldots, X_n)$ be some estimate of $\theta(P)$. Assume that $\hat{\theta}_n$ is symmetric in its arguments. Let $J_n(x, P)$ be the distribution of the root $\tau_n(\hat{\theta}_n - \theta(P))$. Suppose $J_n(x, P)$ converges in distribution to $J(x, P)$. Let $b = b_n > 0$ be a sequence of integers such that $b \to \infty$, $b/n \to 0$, and $\tau_b/\tau_n \to 0$. Index by $i = 1, \ldots, N_n = \binom{n}{b}$ the different subsets of data of size b, and let $\hat{\theta}_{b,i}$ be $\hat{\theta}_b$ evaluated on the ith subset of data. Define

$$L_n(x) = \frac{1}{N_n} \sum_{1 \le i \le N_n} I\{\tau_b(\hat{\theta}_{b,i} - \hat{\theta}_n) \le x\} \ .$$

(i) $L_n(x) \xrightarrow{P} J(x, P)$ for all continuity points of $J(x, P)$.

(ii) If $J(x, P)$ is continuous at $c(1-\alpha) = J^{-1}(1-\alpha, P)$, then $\hat{c}_n(1-\alpha) = L_n^{-1}(1-\alpha) \xrightarrow{P} c(1-\alpha, P)$.

PROOF: (i) Let

$$U_n(x) = \frac{1}{N_n} \sum_{1 \le i \le N_n} I\{\tau_b(\hat{\theta}_{b,i} - \theta(P)) \le x\} \ .$$

Notice $U_n(x)$ only differs from $L_n(x)$ in that $\hat{\theta}_n$ is replaced with $\theta(P)$. Intuitively, we expect $U_n(x)$ and $L_n(x)$ to be close under our assumptions, so we will first show that $U_n(x) \xrightarrow{P} J(x, P)$ for all continuity points of $J(x, P)$.

Note that $U_n(x)$ is an average of the N_n terms $I\{\tau_b(\hat{\theta}_{b,i} - \theta(P)) \le x\}$ and

$$E[I\{\tau_b(\hat{\theta}_{b,i} - \theta(P)) \le x\}] = J_b(x, P) \ .$$

Therefore, we might expect $U_n(x) - J_b(x, P) \xrightarrow{P} 0$. If this were true, then $U_n(x) \xrightarrow{P} J(x, P)$ for all continuity points of $J(x, P)$, since $J_b(x, P) \to J(x, P)$ for all continuity points of $J(x, P)$ (this follows from the assumptions that $b \to \infty$ and $J_n(x, P)$ converges in distribution to $J(x, P)$).

The object $U_n(x)$ is an example of a U-statistic, and the fact that $U_n(x) - J_b(x,P) \xrightarrow{P} 0$ actually follows from the theory of U-statistics, but we will try to give a more direct proof. To this end, first consider the behavior of

$$\bar{U}_n(x) = \frac{1}{k} \sum_{1 \le i \le k} I\{\tau_b(\hat{\theta}_b(X_{b(i-1)+1}, \ldots, X_{bk}) - \theta(P)) \le x\} \ ,$$

where $k = k_n = \lfloor n/b \rfloor$. Since this is an average of i.i.d. random variables, it is easy to see that $\bar{U}_n(x) - J_b(x,P) \xrightarrow{P} 0$. Concretely, let $\epsilon > 0$ be given and use Chebychev's inequality to write

$$\Pr\{|\bar{U}_n(x) - J_b(x,P)| > \epsilon\} \le \frac{\mathrm{Var}(\bar{U}_n(x))}{\epsilon^2} \ .$$

But,

$$\mathrm{Var}(\bar{U}_n(x)) = \frac{\mathrm{Var}(I\{\tau_b(\hat{\theta}_b(X_{b(i-1)+1}, \ldots, X_{bk}) - \theta(P)) \le x\})}{k} \to 0 \ .$$

Thus, $\bar{U}_n(x) - J_b(x,P) \xrightarrow{P} 0$.

By the same argument, we see that in order to prove that $U_n(x) - J_b(x,P) \xrightarrow{P} 0$, it is enough to prove that $\mathrm{Var}(U_n(x)) \to 0$. Intuitively, we might expect the $\mathrm{Var}(U_n(x)) \le \mathrm{Var}(\bar{U}_n(x))$. In order to formalize this idea, note that we may write

$$U_n(x) = E[\bar{U}_n(x) | X_{(1)}, \ldots, X_{(n)}] \ .$$

To see this, note that

$$\begin{aligned} k \sum_{\pi \in S_n} \bar{U}_n(x, X_{\pi(1)}, \ldots, X_{\pi(n)}) &= kb!(n-b)! \sum_{1 \le i \le N_n} I\{\tau_n(\hat{\theta}_{b,i} - \theta(P)) \le x\} \\ &= kb!(n-b)! \binom{n}{b} U_n(x) = kn! U_n(x) \ , \end{aligned}$$

where S_n is the set of all permutations π of $1, \ldots, n$. Therefore,

$$U_n(x) = \frac{1}{n!} \sum_{\pi \in S_n} \bar{U}_n(x, X_{\pi(1)}, \ldots, X_{\pi(n)}) = E[U_n(x) | X_{(1)}, \ldots, X_{(n)}] \ .$$

The last equality follows from the fact that since the data is i.i.d., conditional on $X_{(1)}, \ldots, X_{(n)}$, each of the $n!$ orderings are equally likely.

Now the desired result follows because

$$\begin{aligned}\mathrm{Var}(U_n(x)) &= E[(U_n(x) - J_b(x,P))^2] \\ &= E[E[\bar{U}_n(x) - J_b(x,P)|X_{(1)},\ldots,X_{(n)}]^2] \\ &\leq E[E[(\bar{U}_n(x) - J_b(x,P))^2|X_{(1)},\ldots,X_{(n)}]] = \mathrm{Var}(\bar{U}_n(x)) \ .\end{aligned}$$

We have therefore shown that $U_n(x) \xrightarrow{P} J(x,P)$ for all continuity points of $J(x,P)$. We now use this fact to establish the desired result, that is, $L_n(x) \xrightarrow{P} J(x,P)$ for all continuity points of $J(x,P)$

To this end, let x be a continuity point of $J(x,P)$ and note that

$$L_n(x) = \frac{1}{N_n} \sum_{1 \leq i \leq N_n} I\{\tau_b(\hat{\theta}_{b,i} - \theta(P)) - \tau_b(\hat{\theta}_n - \theta(P)) \leq x\} \ .$$

For $\epsilon > 0$, let

$$E_n = \{|\tau_b(\hat{\theta}_n - \theta(P))| < \epsilon\} \ .$$

Since

$$\begin{aligned}& I\{\tau_b(\hat{\theta}_{b,i} - \theta(P)) - \tau_b(\hat{\theta}_n - \theta(P)) \leq x\} \\ = \ & I\{\tau_b(\hat{\theta}_{b,i} - \theta(P)) \leq x + \tau_b(\hat{\theta}_n - \theta(P))\} \ ,\end{aligned}$$

it follows that when E_n is true,

$$\begin{aligned}I\{\tau_b(\hat{\theta}_{b,i} - \theta(P)) - \tau_b(\hat{\theta}_n - \theta(P)) \leq x\} &\leq I\{\tau_b(\hat{\theta}_{b,i} - \theta(P)) \leq x + \epsilon\} \\ I\{\tau_b(\hat{\theta}_{b,i} - \theta(P)) - \tau_b(\hat{\theta}_n - \theta(P)) \leq x\} &\geq I\{\tau_b(\hat{\theta}_{b,i} - \theta(P)) \leq x - \epsilon\} \ .\end{aligned}$$

Since $\tau_b/\tau_n \to 0$ and $J_n(x,P)$ converges in distribution to $J(x,P)E_n$ has probability tending to 1. Therefore, with probability tending to 1,

$$U_n(x-\epsilon) \leq L_n(x) \leq U_n(x+\epsilon) \ .$$

If $\epsilon > 0$ is such that $x - \epsilon$ and $x + \epsilon$ are continuity points of $J(x,P)$, then

$$\begin{aligned}U_n(x-\epsilon) &\xrightarrow{P} J(x-\epsilon,P) \\ U_n(x+\epsilon) &\xrightarrow{P} J(x+\epsilon,P)\end{aligned}$$

Therefore, with probability tending to 1, for any $\delta > 0$, we have that

$$\begin{aligned} U_n(x-\epsilon) &> J(x-\epsilon, P) - \frac{\delta}{2} \\ U_n(x+\epsilon) &< J(x+\epsilon, P) + \frac{\delta}{2} \ . \end{aligned}$$

Choose $\epsilon > 0$ so that $x - \epsilon$ and $x + \epsilon$ are continuity points of $J(x, P)$ and

$$\begin{aligned} J(x-\epsilon, P) - \frac{\delta}{2} &> J(x, P) - \delta \\ J(x+\epsilon, P) + \frac{\delta}{2} &< J(x, P) + \delta \ . \end{aligned}$$

This is possible because $J(x, P)$ is continuous at x. Putting this all together, we have that with probability tending to 1,

$$J(x, P) - \delta < L_n(x) < J(x, P) + \delta \ .$$

Since the choice of $\delta > 0$ was arbitrary, the desired result follows.

(ii) This follows from an argument almost identical to the proof of Lemma 2.1 in the lecture notes on the bootstrap. ■

In practice, N_n is too large to actually compute $L_n(x)$, so what one would do is randomly sample B of the N_n possible data sets of size b and just use B in place of N_n when computing $L_n(x)$. Provided $B = B_n \to \infty$, all the conclusions of the theorem remain valid. This approximation step is similar in spirit to approximating the bootstrap distribution $J_n(x, \hat{P}_n)$ using simulations from $\hat{P}_n$.

It is in fact possible to show that

$$\sup_{x \in \mathbf{R}} |U_n(x) - J_b(x, P)| \xrightarrow{P} 0 \ .$$

Remarkably, this convergence is in fact even uniform in P over *any* set of distributions for the observed data! This fact can be used to analyze when the coverage probability of confidence sets constructed using subsampling converges to the nominal level not just for a fixed distribution of the observed data, but uniformly over a set of distributions for the observed data.

See Romano and Shaikh (2006) and Andrews and Guggenberger (2007) for details.

Essentially, all we required was that $J_n(x, P)$ converged in distribution to a limit distribution $J(x, P)$, whereas for the bootstrap we required this and additionally that $J_n(x, P)$ was continuous in a certain sense. Showing continuity of $J_n(x, P)$ was very problem specific. We also saw an example where $J_n(x, P) \to J(x, P)$, but this continuity failed (e.g., the extreme order statistic). Subsampling would have no problems handling the extreme order statistic.

Typically, when both the bootstrap and subsampling are valid, the bootstrap works better in the sense of higher-order asymptotics (see the lecture notes on the bootstrap), but subsampling is more generally valid.

There is a variant of the bootstrap known as the m-out-of-n bootstrap. Instead of using $J_n(x, \hat{P}_n)$ to approximate $J_n(x, P)$, one uses $J_m(x, \hat{P}_n)$ where m is much smaller than n. If one assumes that $m^2/n \to 0$, then all the conclusions of the theorem remain valid with $J_m(x, \hat{P}_n)$ in place of $L_n(x)$. This follows because if $m^2/n \to 0$, then (i) $m/n \to 0$ and (ii) with probability tending to 1, the approximation to $J_m(x, \hat{P}_n)$ is the same as the approximation to $L_n(x)$ because the probability of drawing all distinct observations tends to 1. To see this, note that this probability is simply equal to

$$\frac{n(n-1)(n-2)\cdots(n-b+1)}{n^b} = \prod_{1 \leq i \leq b-1} (1 - \frac{i}{n}) \ .$$

Since $1 - \frac{i}{n} \geq 1 - \frac{b}{n}$, we have that

$$\prod_{1 \leq i \leq b-1} (1 - \frac{i}{n}) \geq (1 - \frac{b}{n})^b = (1 - \frac{\frac{b^2}{n}}{b})^b \ .$$

If $b^2/n \to 0$, then for every $\epsilon > 0$ we have that $b^2/n < \epsilon$ for all n sufficiently large. Therefore,

$$(1 - \frac{\frac{b^2}{n}}{b})^b > (1 - \frac{\epsilon}{b})^b \to \exp(-\epsilon) \ .$$

By choosing $\epsilon > 0$ sufficiently small, we see that the desired probability converges to 1.