ESTIMATING THE DISTRIBUTION OF A STUDENTIZED STATISTIC BY SUBSAMPLING

Dimitris Politis, Department of Statistics
Purdue University, West Lafayette, IN 47907 USA

Joseph Romano, Department of Statistics
Stanford University, Stanford, CA 94305 USA

Nous considérons la construction des regions de confiance, par l'emploi de la distribution approximative d'une statistique quelconque. La vraie distribution est estimée par une normalisation convenable des valeurs de la statistique, qui est calculée en utilisant les échantillons partiels.

The construction of confidence regions by approximating the sampling distribution of some statistic is considered. The true sampling distribution is estimated by an appropriate normalization of the values of the statistic computed over subsamples of the data. The method yields asymptotically valid confidence regions under minimal conditions. Let $X_1, \ldots, X_n$ be a sample of n i.i.d. S-valued random variables with law P. The goal is to construct a confidence interval for $\theta(P)$. Let $T_n = T_n(X_1, \ldots, X_n)$ be an estimator of $\theta(P)$. Define $K_n(P)$ to be the sampling distribution of $\tau_n(T_n - \theta(P))/\hat{\sigma}_n$ based on a sample of size n from P, where $\hat{\sigma}_n$ is some estimate of scale, and the corresponding c.d.f. is denoted $K_n(\cdot, P)$. To describe the method, let $Y_1, \ldots, Y_{N_n}$ be equal to the $N_n = \binom{n}{b}$ subsets of $\{X_1, \ldots, X_n\}$, ordered in any fashion. Let $S_{n,i}$ be equal to the statistic T_b evaluated at the data set Y_i. Let $\hat{\sigma}_{n,i}$ be equal to the estimate of scale based on Y_i. Define

$$\hat{K}_n(x) = N_n^{-1} \sum_{i=1}^{N_n} 1\{\tau_b(S_{n,i} - T_n)/\hat{\sigma}_{n,i} \leq x\}.$$

The motivation behind the method is the following. For any i, Y_i is a random sample of size b from P. Hence, the *exact* distribution of $\tau_b(S_{n,i} - \theta(P))/\hat{\sigma}_{n,i}$ is $K_b(P)$. The empirical distribution of the N_n values of $\tau_b(S_{n,i} - \theta(P))/\hat{\sigma}_{n,i}$ should then serve as a good approximation to $K_n(P)$. Of course, $\theta(P)$ is unknown, so we replace $\theta(P)$ by T_n, which is asymptotically permissible under weak assumptions. The following theorem holds. It was considered in the special case of the mean by Wu (1990). The proof is similar to Theorem 2.1 of Politis and Romano (1992).

Theorem. *Assume $K_n(P)$ has a limit law $K(P)$, with corresponding c.d.f. $K(\cdot, P)$. Assume $\tau_b/\tau_n \to 0$, $b \to \infty$ and $b/n \to 0$ as $n \to \infty$. Suppose $\hat{\sigma}_n \to \sigma$*

in probability, where $\sigma = \sigma(P)$ *is a positive constant. Let* x *be a continuity point of* $K(\cdot, P)$*. Then,* $\hat{K}_n(x) \to K(x, P)$ *in probability. If* $K(\cdot, P)$ *is continuous, then* $\sup_x |\hat{K}_n(x) - K_n(x, P)| \to 0$ *in probability. Let* $d_n(1-\alpha) = \inf\{x : \hat{K}_n(x) \geq 1-\alpha\}$*. If* $K(\cdot, P)$ *is continuous at its* $1-\alpha$ *quantile,*

$$Prob_P\{\tau_n[T_n - \theta(P)]/\hat{\sigma}_n \leq d_n(1-\alpha)\} \to 1-\alpha$$

as $n \to \infty$*. Thus, the asymptotic coverage probability of* $[T_n - \hat{\sigma}_n \tau_n^{-1} d_n(1-\alpha), \infty)$ *is the nominal level* $1-\alpha$*. Assume, for every* $d > 0$*,* $\sum_n exp\{-d[n/b]\} < \infty$*,* $\tau_b(T_n - \theta(P)) \to 0$ *almost surely, and* $\hat{\sigma}_n \to \sigma(P)$ *almost surely. Then, the aforementioned convergence holds with probability one.*

In some special cases, it has been realized that a sample size trick can often remedy the inconsistency of the bootstrap. The usual bootstrap approximation to $K_n(P)$ is $K_n(\hat{P}_n)$, where $\hat{P}_n$ denotes the empirical measure. Rather than approximating $K_n(P)$ by $K_n(\hat{P}_n)$, the suggestion is to approximate $K_n(P)$ by $K_b(\hat{P}_n)$ for some b which usually satisfies $b/n \to 0$ and $b \to \infty$. The resulting estimator $K_b(x, \hat{P}_n)$ is obviously quite similar to $\hat{K}_n$. These approaches must be similar if b is so small that sampling with and without replacement are essentially the same. Indeed, if one resamples b numbers (or indices) from the set $\{1, \ldots, n\}$, then the chance that none of the indices is duplicated is $\Pi_{i=1}^{b-1}(1 - \frac{i}{n})$. This probability tends to 0 if $b^2/n \to 0$. (To see why, take logs and do a Taylor expansion analysis.) Hence, the following is true.

Corollary. *Under the further assumption that* $b^2/n \to 0$*, the convergence in probability results of the Theorem remain valid if* $\hat{K}_n(x)$ *is replaced by the bootstrap approximation* $K_b(x, \hat{P}_n)$*.*

In spite of the Corollary, we point out that $\hat{K}_n$ is more generally valid. Indeed, without the assumption $b^2/n \to 0$, $K_b(x, \hat{P}_n)$ can be inconsistent.

BIBLIOGRAPHY

Politis, D.N. and Romano, J.P. (1992), A general theory for large sample confidence regions based on subsamples under minimal conditions, Technical Report 399, Department of Statistics, Stanford University.

Wu, C,F.J. (1990), On the asymptotic properties of the jackknife histogram. *Ann. Statist.* **18**, 1438-1452.