Statistica Sinica **11**(2001), 1105-1124

# ON THE ASYMPTOTIC THEORY OF SUBSAMPLING

Dimitris N. Politis, Joseph P. Romano and Michael Wolf

*University of California, Stanford University and Universidad Carlos III*

*Abstract:* A general approach to constructing confidence intervals by subsampling was presented in Politis and Romano (1994). The crux of the method is recomputing a statistic over subsamples of the data, and these recomputed values are used to build up an estimated sampling distribution. The method works under extremely weak conditions, it applies to independent, identically distributed (i.i.d.) observations as well as to dependent data situations, such as time series (possibly nonstationary), random fields, and marked point processes. In this article, we present some theorems showing: a new construction for confidence intervals that removes a previous condition, a general theorem showing the validity of subsampling for data-dependent choices of the block size, and a general theorem for the construction of hypothesis tests (not necessarily derived from a confidence interval construction). The arguments apply to both the i.i.d. setting and the dependent data case.

*Key words and phrases:* Confidence intervals, data-dependent block size choice, hypothesis tests, large sample theory, resampling.

## 1. Introduction

A general theory for the construction of confidence intervals or regions was presented in Politis and Romano (1992, 1994). The basic idea is to approximate the sampling distribution of a statistic based on the values of the statistic computed over smaller subsets of the data. For example, in the case where the data are $n$ observations which are independent and identically distributed, a statistic is computed based on the entire data set and is recomputed over all $\binom{n}{b}$ data sets of size $b$. Implicit is the notion of a statistic sequence, so that the statistic is defined for samples of size $n$ and $b$. These recomputed values of the statistic are suitably normalized to approximate the true sampling distribution.

This approach based on subsampling is perhaps the most general theory for the construction of first order asymptotically valid confidence regions. Other methods, such as the bootstrap, require that the distribution of the statistic is somehow locally smooth as a function of the unknown model. In fact, many papers have been devoted to showing the convergence of a suitably normalized statistic to its limiting distribution is appropriately uniform as a function of the unknown model in specific situations. In contrast, no such assumption or verification of such smoothness is required in the theory for subsampling. Indeed, the method here is applicable even in the several known situations which represent counterexamples to the bootstrap. To appreciate why subsampling behaves well under such weak assumptions, note that each subset of size $b$ (taken without replacement from the original data) is indeed a sample of size $b$ from the true model. Hence, it should be intuitively clear that one can at least approximate the sampling distribution of the (normalized) statistic based on a sample of size $b$. But, under the weak convergence hypothesis, the sampling distributions based on samples of size $b$ and $n$ should be close. The bootstrap, on the other hand, is based on recomputing a statistic over a sample of size $n$ from some estimated model which is hopefully close to the true model.

The method has a clear extension to the context of a stationary time series or, more generally, a homogeneous random field. The only difference is that the statistic is computed over a smaller number of subsets of the data that retain the dependence structure of the observations. For example, if $X_1, \ldots, X_n$ represent $n$ observations from some stationary time series, the statistic is recomputed only over the $n - b + 1$ subsets of size $b$ of the $\{X_i, X_{i+1}, \ldots, X_{i+b-1}\}$. The ideas extend to random fields and marked point processes as well.

The use of subsample values to approximate the variance of a statistic is well-known. The Quenouille-Tukey jackknife estimates of bias and variance based on computing a statistic over all subsamples of size $n - 1$ has been well-studied and is closely related to the mean and variance of our estimated sampling distribution with $b = n - 1$. Mahalanobis (1946) suggested the use of subsamples to estimate variability in studying crop yields, though he used the name interpenetrating samples. Half sampling methods have been well-studied in the context of sampling theory; see McCarthy (1969). Hartigan (1969) introduced what Efron (1982) calls a random subsampling method, based on the computation of a statistic over all $2^n - 1$ nonempty subsets of the data. His method is seen to produce exact confidence limits in the special context of the symmetric location problem. Hartigan (1975) adapted his finite sample results to a more general context of certain classes of estimators which have asymptotic normal distributions. But, even in this context, his asymptotic results assume the number of subsamples used to recompute the statistic remains fixed as $n \to \infty$, which results in a loss of efficiency.

Efron's (1979) bootstrap, while sharing some similar properties to the aforementioned methods, has corrected some deficiencies in the jackknife, and has tackled the more ambitious goal of approximating an entire sampling distribution. Shao and Wu (1989) have shown that, by basing a jackknife estimate of variance on the statistic computed over subsamples with $d$ observations deleted, many of the deficiencies of the usual $d = 1$ jackknife estimate of variance can be removed. Later, Wu (1990) used these subsample values to approximate an entire sampling distribution by what he calls a jackknife histogram, but only in regular i.i.d. situations where the statistic is appropriately linear so that asymptotic normality ensues. In more broad generality, Sherman and Carlstein (1996) considered the use of subsamples as a diagnostic tool to describe the shape of the sampling distribution of a general statistic, though formal inference procedures, such as the construction of confidence intervals, are not delivered. Here, we show how these subsample values can accurately estimate a sampling distribution without any assumptions of asymptotic normality, by only assuming the existence of a limiting distribution. Moreover, the asymptotic validity of confidence statements follows. In summary, while the method developed in this work is related to several well-studied techniques, the simplicity of our arguments leads to asymptotic justification under the most general conditions.

In Section 2, the method is described in the context of i.i.d. observations. The basic theory is quickly motivated. Theorem 2.1 presents a general theorem showing the validity of subsampling in the i.i.d. case, acknowledging a data-driven choice of block size, which inevitably is the situation used in practice. The basic argument is to show subsampling is consistent uniformly across a broad range of block sizes. A variation (Corollaries 2.1 and 5.1) of the basic confidence interval is presented which removes one of the original conditions. Although this condition is extremely weak, the new interval is more closely related to a construction presented in the next section on hypothesis testing. The use of subsampling in the context of hypothesis testing based on i.i.d. samples is described in Section 3. A basic result giving the behavior of the subsampling null distribution under the null hypothesis and contiguous alternatives is obtained. Sections 4 and 5 extend these ideas to the time series case. The same ideas apply, and the proofs only highlight the differences from the i.i.d. case. Section 6 presents an example and illustrates the idea of data-driven choice of the block size.

## 2. The i.i.d. Case with Data-Dependent Block Size

Throughout this section, $X_1, \ldots, X_n$ is a sample of $n$ independent and identically distributed random variables taking values in an arbitrary sample space $S$. The common probability measure generating the observations is denoted $P$. The goal is to construct a confidence region for some parameter $\theta(P)$. For now, $\theta$ is real-valued, but this can be considerably generalized to allow for the construction of confidence regions for multivariate parameters or confidence bands for functions.

Let $\hat{\theta}_n = \hat{\theta}_n(X_1, \ldots, X_n)$ be an estimator of $\theta(P)$. Nothing is assumed about $\hat{\theta}_n$, though it is natural to assume $\hat{\theta}_n$ is symmetric in its arguments in the i.i.d. case. It is desired to estimate or approximate the true sampling distribution of $\hat{\theta}_n$ in order to make inferences about $\theta(P)$.

Define $J_n(P)$ to be the sampling distribution of $\tau_n(\hat{\theta}_n - \theta(P))$ based on a sample of size $n$ from $P$, where $\tau_n$ is a normalizing constant. Also define the corresponding cumulative distribution function: $J_n(x, P) = Prob_P\{\tau_n[\hat{\theta}_n(X_1, \ldots, X_n) - \theta(P)] \leq x\}$.

For asymptotically valid confidence intervals for $\theta(P)$, we need the following.

**Assumption 2.1.** *There exists a limit law $J(P)$ such that $J_n(P)$ converges weakly to $J(P)$ as $n \to \infty$.*

We require this assumption for some sequence $\tau_n$; it is most useful when $\tau_n$ is such that the limit law $J(P)$ is nondegenerate. The assumption is clearly satisfied in numerous examples, and it is hard to conceive of a theory where it fails.

To describe the method studied in this section, let $Y_1, \ldots, Y_{N_n}$ be equal to the $N_n = \binom{n}{b}$ subsets of size $b$ of $\{X_1, \ldots, X_n\}$, ordered in any fashion. Of course, the $Y_i$ depend on $b$ and $n$, but this dependence is supressed. Only a very weak assumption on $b$ is required. In typical situations, it is assumed that $b/n \to 0$ and $b \to \infty$ as $n \to \infty$. Now, let $\hat{\theta}_{n,b,i}$ be equal to the statistic $\hat{\theta}_b$ evaluated at the data set $Y_i$. The approximation to $J_n(x, P)$ we study is defined by

$$L_{n,b}(x) = N_n^{-1} \sum_{i=1}^{N_n} 1\{\tau_b(\hat{\theta}_{n,b,i} - \hat{\theta}_n) \leq x\}. \tag{1}$$

Our motivation is the following. For any $i$, $Y_i$ is a random sample of size $b$ from $P$. Hence, the *exact* distribution of $\tau_b(\hat{\theta}_{n,b,i} - \theta(P))$ is $J_b(P)$. The empirical distribution of the $N_n$ values of $\tau_b(\hat{\theta}_{n,b,i} - \theta(P))$ should then serve as a good approximation to $J_n(P)$. Of course $\theta(P)$ is unknown, so we replace $\theta(P)$ by $\hat{\theta}_n$, asymptotically permissible because $\tau_b(\hat{\theta}_n - \theta(P))$ is of order $\tau_b/\tau_n \to 0$. These heuristics lead to the following theorem, first proved in the special case of fixed block size in Politis and Romano (1992). Here, we present a result which shows subsampling works very generally even with a data-driven choice of block size, as would happen in practice. The approach is to show subsampling works uniformly over a broad range of block sizes, and hence to a random choice as well.

**Theorem 2.1.** *Let $1 \leq j_n \leq k_n \leq n$ be integers satisfying $j_n \to \infty$, $k_n/n \to 0$, $\tau_{k_n}/\tau_n \to 0$, and, for every $d > 0$, $(k_n - j_n + 1)\exp(-d\lfloor \frac{n}{k_n} \rfloor) \to 0$ as $n \to \infty$. Assume $\{\tau_n\}$ is nondecreasing in $n$, and adopt Assumption* 2.1.

(i) *If $x$ is a continuity point of $J(\cdot, P)$, then $\sup_{j_n \leq b \leq k_n} |L_{n,b}(x) - J(x, P)| \to 0$ in probability.*

(ii) *If* $\{\hat{b}_n\}$ *is a data-dependent sequence* (*that is, a measurable function of* $X_1, \ldots, X_n$), *and* $Prob_P\{j_n \le \hat{b}_n \le k_n\} \to 1$, *then* $L_{n,\hat{b}_n}(x) \to J(x,P)$ *in probability.*

(iii) *If* $J(\cdot,P)$ *is continuous, then* $\sup_x |L_{n,\hat{b}_n}(x) - J(x,P)| \to 0$ *in probability. In fact,* $\sup_{j_n \le b \le k_n} \sup_x |L_{n,b}(x) - J(x,P)| \to 0$ *in probability.*

(iv) *Let* $c_{n,\hat{b}_n}(1-\alpha) = \inf\{x : L_{n,\hat{b}_n}(x) \ge 1-\alpha\}$. *Then, if* $J(\cdot,P)$ *is continuous,* $Prob_P\{\tau_n[\hat{\theta}_n - \theta(P)] \le c_{n,\hat{b}_n}(1-\alpha)\} \to 1-\alpha$ *as* $n \to \infty$. *Therefore, the asymptotic coverage probability under* $P$ *of the confidence interval* $[\hat{\theta}_n - \tau_n^{-1} c_{n,\hat{b}_n}(1-\alpha), \infty)$ *is the nominal level* $1-\alpha$.

**Proof.** Let $\hat{\theta}_{n,b,i}$ be the statistic $\hat{\theta}_b$ evaluated at the $i$th of the $\binom{n}{b}$ data sets of size $b$; any ordering of these $\binom{n}{b}$ values will do. Define

$$U_{n,b}(x,P) \equiv U_{n,b}(x) = \binom{n}{b}^{-1} \sum_{i=1}^{\binom{n}{b}} 1\{\tau_b[\hat{\theta}_{n,b,i} - \theta(P)] \le x\}. \tag{2}$$

First we claim that, for each continuity point $x$ of $J(\cdot,P)$,

$$\sup_{j_n \le b \le k_n} |U_{n,b}(x) - J(x,P)| \to 0 \text{ in probability.} \tag{3}$$

But $\sup_{j_n \le b \le k_n} |J_b(x,P) - J(x,P)| \to 0$, because, if this convergence failed, there would exist $\{b_n\}$ with $b_n \in [j_n, k_n]$ such that $J_{b_n}(x,P)$ does not converge to $J(x,P)$. This is a contradiction since $b_n \ge j_n \to \infty$. So, to show (3) it suffices to show

$$\sup_{j_n \le b \le k_n} |U_{n,b}(x) - J_b(x,P)| \to 0 \text{ in probability.} \tag{4}$$

But, for any $t > 0$,

$$\begin{aligned}
&Prob_P\{\sup_{j_n \le b \le k_n} |U_{n,b}(x) - J_b(x,P)| \ge t\} \\
&\le \sum_{b=j_n}^{k_n} Prob_P\{|U_{n,b}(x) - J_b(x,P)| \ge t\} \\
&\le (k_n - j_n + 1) \sup_{j_n \le b \le k_n} Prob_P\{|U_{n,b}(x) - J_b(x,P)| \ge t\} \\
&\le 2(k_n - j_n + 1) \sup_{j_n \le b \le k_n} \exp\{-2\lfloor \frac{n}{b} \rfloor t^2\}.
\end{aligned} \tag{5}$$

The last inequality makes use of Hoeffding's inequality for $U$-statistics, which applies since $U_{n,b}(x)$ is indeed a bounded $U$-statistic of degree $b$; see Serfling (1980), p.201. But this last expression is bounded above by $2(k_n - j_n + 1)\exp\{-2\lfloor \frac{n}{k_n} \rfloor t^2\}$, which tends to zero by assumption on $\{k_n\}$. Thus, (4) holds, as does (3). Now, note that

$$L_{n,b}(x) = \binom{n}{b}^{-1} \sum_{i=1}^{\binom{n}{b}} 1\{\tau_b[\hat{\theta}_{n,b,i} - \theta(P)] + \tau_b[\theta(P) - \hat{\theta}_n] \le x\}.$$

Fix any $\epsilon > 0$ so that $x \pm \epsilon$ are continuity points of $J(\cdot, P)$. Then

$$U_{n,b}(x-\epsilon) 1(E_{n,b}) \le L_{n,b} 1(E_{n,b}) \le U_{n,b}(x+\epsilon), \tag{6}$$

where $1(E_{n,b})$ is the indicator of the event $E_{n,b} \equiv \{\tau_b|\theta(P) - \hat{\theta}_n| \le \epsilon\}$. By the monotonicity of $\{\tau_n\}$, $1(E_{n,k_n}) \le 1(E_{n,b}) \le 1(E_{n,j_n})$ and $\tau_{k_n}/\tau_n \to 0$ implies $Prob_P(E_{n,k_n}) \to 1$. So $L_{n,b}(x)1(E_{n,k_n}) \le U_{n,b}(x+\epsilon)$. Thus, on the set $E_{n,k_n}$,

$$\begin{aligned} &\sup_{j_n \le b \le k_n} L_{n,b}(x) - J(x,P) \le \sup_{j_n \le b \le k_n} U_{n,b}(x+\epsilon) - J(x,P) \\ &\le \sup_{j_n \le b \le k_n} |U_{n,b}(x+\epsilon) - J(x+\epsilon,P)| + J(x+\epsilon,P) - J(x,P). \end{aligned}$$

But, by (3), it follows that, for every $\delta > 0$, $\sup_{j_n \le b \le k_n} L_{n,b}(x) - J(x,P) \le \delta + J(x+\epsilon,P) - J(x,P)$ with probability tending to one. Similarly, replacing $x+\epsilon$ by $x-\epsilon$ and using the first inequality in (6), we get, for every $\eta > 0$, $\sup_{j_n \le b \le k_n} |L_{n,b}(x) - J(x,P)| \le \eta$ with probability tending to one, which is equivalent to statement (i) of the theorem. Part (ii) is obvious. The rest of the theorem is proved as in the proof of Theorem 2.1 of Politis and Romano (1994).

**Remark 2.1.** In some cases, one finds that an optimal choice of $b = b_n$ should satisfy $b_n n^p \to \xi(P)$, for some $p \in (0,1)$, where $\xi(P)$ is a constant typically depending on the unknown probability mechanism $P$. In an ad hoc way, one can sometimes estimate $\xi(P)$ consistently by $\hat{\xi}_n$ (say by a plug-in approach), which leads to the choice of block size $\hat{b}_n = \lfloor \hat{\xi}_n n^p \rfloor$. Such a construction for $\hat{b}_n$ will easily satisfy the conditions of the theorem. Simply take $j_n = \lfloor \epsilon n^p \rfloor$ and $k_n = \lfloor n^p/\epsilon \rfloor$ for small enough $\epsilon$. Moreover, the condition $\tau_{k_n}/\tau_n \to 0$ will be satisfied in the typical case $\tau_n$ is proportional to $n^\beta$ for some $\beta \in (0,1)$. In practice, the parameter $\xi(P)$ may be difficult to estimate, and even if consistent estimation is possible, the resulting estimator may have poor finite-sample performance. The point of this section is to show subsampling has some asymptotic validity across a broad range of choices for the subsample size.

**Remark 2.2.** The monotonicity assumption on $\{\tau_n\}$ can be replaced by the condition $\sup_{j_n \le b \le k_n} [\tau_b/\tau_n] \to 0$, as the proof essentially shows. Actually, the assumption can be removed altogether if the interval is modified as in Corollary 2.1, where in fact the condition $\tau_{k_n}/\tau_n \to 0$ is removed altogether.

**Remark 2.3.** The convergence in probability statements in the theorem can be strengthened to be almost sure convergences, provided $\tau_{k_n}[\hat{\theta}_n - \theta(P)] \to 0$ almost surely, and for every $d > 0$, $\sum_{n=1}^{\infty} k_n \exp(-d\lfloor \frac{n}{k_n} \rfloor) < \infty$. The last condition holds whenever $k_n$ can be taken to be $O(n^p)$ with $p < 1$.

One can remove the assumption that $\tau_{k_n}/\tau_n \to 0$ if the goal is to construct an asymptotically valid confidence interval for $\theta(P)$, but at the small expense of bypassing consistent estimation of $J_n(\cdot, P)$. To see how, let $u_{n,b}(1-\alpha, P) = \inf\{x : U_{n,b}(x, P) \geq 1-\alpha\}$, where $U_{n,b}(\cdot, P)$ is defined in (2). Under continuity assumptions on $J(\cdot, P)$, the proof of Theorem 2.1 shows $U_{n,b}(x, P)$ converges in probability to $J(x, P)$ uniformly in $b \in [j_n, k_n]$. It follows that $u_{n,\hat{b}_n}(1 - \alpha, P)$ converges in probability (under $P$) to $c(1-\alpha, P)$; to argue why, one can easily conclude almost sure convergence along a subseqeuence. Moreover, the assumption $\tau_{b_n}/\tau_n \to 0$ is not used. Note, however, that $u_{n,\hat{b}_n}(1-\alpha, P)$ is not an estimator since it depends on $P$. Nevertheless, with $P$ fixed, the event

$$\{\tau_n(\hat{\theta}_n - \theta(P)) \leq u_{n,\hat{b}_n}(1-\alpha, P)\} \tag{7}$$

has an asymptotic probability of $1-\alpha$ under $P$ (assuming $J(\cdot, P)$ is continuous at $c(1-\alpha, P)$). But,

$$u_{n,\hat{b}_n}(1-\alpha, P) = c_{n,b}(1-\alpha) + \tau_{\hat{b}_n}(\hat{\theta}_n - \theta(P)). \tag{8}$$

Hence, the event (7) is exactly the same as

$$\{\tau_n(\hat{\theta}_n - \theta(P)) \leq c_{n,\hat{b}_n}(1-\alpha) + \tau_{\hat{b}_n}(\hat{\theta}_n - \theta(P))\}, \tag{9}$$

or equivalently,

$$\{(\tau_n - \tau_{\hat{b}_n})(\hat{\theta}_n - \theta(P) \leq c_{n,\hat{b}_n}(1-\alpha)\} \tag{10}$$

By solving for $\theta(P)$, the following nominal level $1-\alpha$ confidence interval is obtained:

$$[\hat{\theta}_n - (\tau_n - \tau_{\hat{b}_n})^{-1} c_{n,\hat{b}_n}(1-\alpha), \infty). \tag{11}$$

The interval (11) can be computed without knowledge of $P$, it has asymptotic coverage probability under $P$ of $1-\alpha$, and the assumption $\tau_{k_n}/\tau_n \to 0$ was not needed. Clearly, the only difference between (11) and the interval presented in Theorem 2.1 is that the factor $(\tau_n - \tau_{\hat{b}_n})^{-1}$ here replaces the factor $\tau_n^{-1}$ there.

**Corollary 2.1.** *Under Assumption* 2.1*, let* $1 \leq j_n \leq k_n \leq n$ *be integers such that* $j_n \to \infty$ *and* $k_n/n \to 0$ *as* $n \to \infty$*. Also assume, for every* $d > 0$*,* $k_n \exp(-d\lfloor \frac{n}{k_n} \rfloor) \to 0$ *as* $n \to \infty$*. If* $J(\cdot, P)$ *is continuous at* $c(1-\alpha, P)$*, then the interval (11) contains* $\theta(P)$ *with asymptotic probability* $1-\alpha$ *under* $P$*.*

**Remark 2.4.** The interval $I_1$ defined in (iii) of Theorem 2.1 corresponds to a one-sided hybrid percentile interval in the bootstrap literature (e.g., Hall (1992)).

A two-sided *equal-tailed* confidence interval can be obtained by forming the intersection of two one-sided intervals. The two-sided analogue of $I_1$ is $I_2 = [\hat{\theta}_n - \tau_n^{-1} c_{n,\hat{b}_n}(1-\alpha/2),\ \hat{\theta}_n - \tau_n^{-1} c_{n,\hat{b}_n}(\alpha/2)]$. $I_2$ is called equal-tailed because it has approximately equal probability in each tail: $Prob_P\{\theta(P) < \hat{\theta}_n - \tau_n^{-1} c_{n,\hat{b}_n}(1-\alpha/2)\} \simeq \alpha/2$ and $Prob_P\{\theta(P) > \hat{\theta}_n - \tau_n^{-1} c_{n,\hat{b}_n}(\alpha/2)\} \simeq \alpha/2$.

As an alternative approach, two-sided *symmetric* confidence intervals can be constructed. A two-sided symmetric confidence interval is given by $[\hat{\theta}_n - \hat{c},\ \hat{\theta}_n + \hat{c}]$, where $\hat{c}$ is chosen so that $Prob_P\{|\hat{\theta}_n - \theta(P)| > \hat{c}\} \simeq \alpha$. Hall (1988) showed that symmetric bootstrap confidence intervals may enjoy enhanced coverage and, even in asymmetric circumstances, can be shorter than equal-tailed confidence intervals. To construct two-sided symmetric subsampling intervals in practice, we follow the traditional approach and estimate the two-sided distribution function $J_{n,|\cdot|}(x,P) = Prob_P\{\tau_n|\hat{\theta}_n - \theta(P)| \le x\}$. The subsampling approximation to $J_{n,|\cdot|}(x,P)$ is defined by $L_{n,b,|\cdot|}(x) = N_n^{-1}\sum_{i=1}^{N_n} 1\{\tau_b \left|\hat{\theta}_{n,b,i} - \hat{\theta}_n\right|\} \le x\}$. An approximate $1-\alpha$ symmetric confidence interval is then given by $I_{SYM} = [\hat{\theta}_n - \tau_n^{-1} c_{n,\hat{b}_n,|\cdot|}(1-\alpha),\ \hat{\theta}_n + \tau_n^{-1} c_{n,\hat{b}_n,|\cdot|}(1-\alpha)]$, where $c_{n,\hat{b}_n,|\cdot|}(1-\alpha)$ is a $1-\alpha$ quantile of $L_{n,\hat{b}_n,|\cdot|}(\cdot)$. By Theorem 2.1 and the Continuous Mapping Theorem, the asymptotic validity of two-sided symmetric subsampling intervals easily follows.

## 3. Hypothesis Testing in the i.i.d. Case

In this section, we consider the use of subsampling for the construction of hypothesis tests. As before, $X_1, \ldots, X_n$ is a sample of $n$ independent and identically distributed observations taking values in a sample space $S$. The common unknown distribution generating the data is denoted by $P$. This unknown law $P$ is assumed to belong to a certain class of laws $\mathbf{P}$. The null hypothesis $H_0$ asserts $P \in \mathbf{P_0}$, and the alternative hypothesis $H_1$ is $P \in \mathbf{P_1}$, where $\mathbf{P_i} \subset \mathbf{P}$ and $\mathbf{P_0} \bigcup \mathbf{P_1} = \mathbf{P}$.

There are several general approaches one can take for the construction of asymptotically valid tests, depending on the nature of the problem. In the special (but usual) case where the null hypothesis translates into a null hypothesis about a real- or vector-valued parameter $\theta(P)$, one can construct a confidence region for $\theta(P)$—by subsampling, bootstrapping, asymptotic approximations, or other methods—and then exploit the usual duality between the construction of confidence regions for parameters and the construction of hypothesis tests about those parameters. This is the approach taken in Politis and Romano (1996). However, not all hypothesis testing problems fit nicely into the aforementioned framework. An alternative bootstrap approach can be based on bootstrapping from a distribution obeying the constraints of the null hypothesis; see Beran (1986) and Romano (1988, 1989). None of the above approaches easily handles the following example, taken from Bickel and Ren (1997), but we will see that an appropriate simple subsampling scheme applies here as well. Bickel and Ren (1997) consider the related bootstrap with smaller resample size.

**Example 3.1. (Goodness of Fit for Censored Data)** Suppose that $U_1, \ldots,$ $U_n$ are i.i.d. random variables with cumulative distribution function $F$. The null hypothesis $H_0$ asserts $F = F_0$, where $F_0$ is some specified distribution. In this problem, however, we do not necessarily observe the full data $U_1, \ldots, U_n$ because the observations $U_i$ are left and right censored. Specifically, assume $(Y_i, Z_i)$ are independent and identically distribution pairs with $Z_i < Y_i$ (with probability one), and the $(Y_i, Z_i)$ pairs are independent of $U_1, \ldots U_n$. Define

$$V_i = \begin{cases} U_i, & \text{if } Z_i < U_i \leq Y_i; \\ Y_i, & \text{if } U_i > Y_i; \\ Z_i, & \text{if } X_i \leq Z_i \end{cases}$$

$$\delta_i = \begin{cases} 1, & \text{if } Z_i < U_i \leq Y_i; \\ 2, & \text{if } U_i > Y_i; \\ 3, & \text{if } X_i \leq Z_i. \end{cases}$$

The actual observations available are $X_i = (V_i, \delta_i)$. Let $\hat{F}_n$ be the nonparametric maximum likelihood estimator of $F$ based on $X_1, \ldots X_n$; this can be computed numerically by the algorithms described in Mykland and Ren (1996). Now, consider the Cramér-von Mises test statistic given by $T_n = n \int_{-\infty}^{\infty} [\hat{F}_n(x) - F_0(x)]^2 dF_0(x)$.

Under suitable conditions and when $F$ is the true distribution for $U_i$, $n^{1/2}$ $[\hat{F}_n(\cdot) - F(\cdot)]$, viewed as a process on $D[-\infty, \infty]$, converges weakly to a mean zero Gaussian process with covariance depending on the joint distribution of $(Z_i, Y_i)$; see Giné and Zinn (1990) and Bickel and Ren (1996). Hence, $T_n$ possesses a limiting distribution as well, both under the null hypothesis and against a sequence of contiguous alternatives; the notion of contiguity is presented in Bickel Klassgn, Rnov and Wellnar ((1993), Section A.9). The difficulty the bootstrap has in approximating this limiting distribution is that $Y_i$ and $Z_i$ are never observed together for any $i$, so that any information on the joint distribution is not available. Note, however, in the right censoring case (with $Z_i = -\infty$), $\hat{F}_n$ is the Kaplan-Meier estimator, and the distribution of the censoring variables can be estimated and the bootstrap offers a viable approach.

We now return to the general setup of testing the null hypothesis $H_0$ that $P \in$ $\mathbf{P_0}$ versus the alternative hypothesis $H_1$ that $P \in \mathbf{P_1}$. The goal is to construct an asymptotically valid test based on a given test statistic $T_n = \tau_n t_n(X_1, \ldots, X_n)$, where, as before, $\tau_n$ is a fixed nonrandom normalizing sequence (though even this assumption can be weakened; see Bertail, Politis and Romano (1999)). Let $G_n(x, P) = Prob_P\{\tau_n t_n(X_1, \ldots, X_n) \leq x\}$.

At this point, not too much is assumed about $T_n$, though it is certainly natural in the i.i.d. case that $t_n(X_1,\ldots,X_n)$ be symmetric in its arguments. As before, we assume $G_n(\cdot,P)$ converges in distribution, at least for $P\in\mathbf{P_0}$. Of course, this would imply (as long as $\tau_n\to\infty$) that $t_n(X_1,\ldots,X_n)\to 0$ in probability for $P\in\mathbf{P_0}$. Naturally, $t_n$ should somehow be designed to distinguish between the competing hypotheses. Our next theorem assumes $t_n$ is constructed to satisfy the following: $t_n(X_1,\ldots,X_n)\to t(P)$ in probability, where $t(P)$ is a constant which satisfies $t(P)=0$ if $P\in\mathbf{P_0}$ and $t(P)>0$ if $P\in\mathbf{P_1}$. This assumption can be made to hold in every conceivable example.

To describe the test construction, 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 $t_{n,b,i}$ be equal to the statistic $t_b$ evaluated at the data set $Y_i$. The sampling distribution of $T_n$ is then approximated by

$$\hat{G}_{n,b}(x)=N_n^{-1}\sum_{i=1}^{N_n}1\{\tau_b t_{n,b,i}\le x\}. \tag{12}$$

Using this estimated sampling distribution, the critical value for the test is obtained as the $1-\alpha$ quantile of $\hat{G}_{n,b}(\cdot)$; specifically, define

$$g_{n,b}(1-\alpha)=\inf\{x:\hat{G}_{n,b}(x)\ge 1-\alpha\}. \tag{13}$$

Finally, the nominal level $\alpha$ test rejects $H_0$ if and only if $T_n>g_{n,b}(1-\alpha)$.

The following theorem gives the consistency of this procedure, under the null hypothesis, the alternative hypothesis, and a sequence of contiguous alternatives. For reasons of notational simplicity and fewer assumptions, we just consider a nonrandom block size $b$, but the proof can easily generalize by using the ideas of Theorem 2.1.

**Theorem 3.1.**

(i) *Assume, for $P\in\mathbf{P_0}$, $G_n(P)$ converges weakly to a continuous limit law $G(P)$, whose corresponding cumulative distribution function is $G(\cdot,P)$ and whose $1-\alpha$ quantile is $g(1-\alpha,P)$. Assume $b/n\to 0$ and $b\to\infty$ as $n\to\infty$. If $G(\cdot,P)$ is continuous at $g(1-\alpha,P)$ and $P\in\mathbf{P_0}$, then $g_{n,b}(1-\alpha)\to g(1-\alpha,P)$ in probability and $Prob_P\{T_n>g_{n,b}(1-\alpha)\}\to\alpha$ as $n\to\infty$.*

(ii) *Assume the test statistic is constructed so that $t_n(X_1,\ldots,X_n)\to t(P)$ in probability, where $t(P)$ is a constant which satisfies $t(P)=0$ if $P\in\mathbf{P_0}$ and $t(P)>0$ if $P\in\mathbf{P_1}$. Assume $b/n\to 0$, $b\to\infty$, and $\liminf_n(\tau_n/\tau_b)>1$. Then, if $P\in\mathbf{P_1}$, the rejection probability satisfies $Prob_P\{T_n>g_{n,b}(1-\alpha)\}\to 1$ as $n\to\infty$.*

(iii) *Suppose $P_n$ is a sequence of alternatives such that, for some $P_0\in\mathbf{P_0}$, $\{P_n^n\}$ is contiguous to $\{P_0^n\}$. Assume $b/n\to 0$ and $b\to\infty$ as $n\to\infty$. Then $g_{n,b}(1-\alpha)\to g(1-\alpha,P_0)$ in $P_n^n$-probability. Hence, if $T_n$ converges in distribution* *to $T$ under $P_n$ and $G(\cdot, P_0)$ is continuous at $g(1-\alpha, P_0)$, $P_n^n\{T_n > g_{n,b}(1-\alpha)\} \to Prob\{T > g(1-\alpha, P_0)\}$.*

**Proof.** To prove (i), note again that $\hat{G}_{n,b}(x)$ is a U-statistic of degree $b$, with expectation under $P$ equal to $G_b(x,P)$. An argument analogous to the one used in the proof of Theorem 2.1 (but easier because there is no centering) shows that $\hat{G}_{n,b}(x) \to G(x,P)$ in probability. Indeed, the variance of the U-statistic tends to zero by the same exponential inequality. It follows that $g_{n,b}(1-\alpha) \to g(1-\alpha, P)$ in probability. Thus, by Slutsky's theorem, the asymptotic rejection probability of the event $T_n > g_{n,b}(1-\alpha)$ is exactly $\alpha$.

To prove (ii), rather than considering $\hat{G}_{n,b}(x)$, just look at the empirical distribution of the values of $t_{n,b,i}$ (not scaled by $\tau_b$). So, define $\hat{G}^0_{n,b}(x) = N_n^{-1}\sum_{i=1}^{N_n} 1\{t_{n,b,i} \le x\} = \hat{G}_{n,b}(\tau_b x)$. But, by a now familiar argument, $\hat{G}^0_{n,b}$ is a U-statistic with expectation $E_P[\hat{G}^0_{n,b}(x)] = Prob_P\{t_b(X_1,\ldots,X_b) \le x\}$, and so $\hat{G}^0_{n,b}(\cdot)$ converges in distribution to a point mass at $t(P)$. It also follows that a $1-\alpha$ quantile, say $g^0_{n,b}(1-\alpha)$, of $\hat{G}^0_{n,b}(\cdot)$ converges in probability to $t(P)$. But our test rejects when $(\tau_n/\tau_b)\cdot t_n(X_1,\ldots,X_n)$ exceeds $g^0_{n,b}(1-\alpha)$. Since $\liminf_n(\tau_n/\tau_b) > 1$ and $t_n(X_1,\ldots,X_n) \to t(P)$ in probability (with $t(P) > 0$), it follows by Slutsky's theorem that the asymptotic rejection probability is one.

Finally, to prove (iii), we know that $g_{n,b}(1-\alpha) \to g(1-\alpha, P_0)$ in probability under $P_0$; contiguity forces $g_{n,b}(1-\alpha) \to g(1-\alpha, P_0)$ in probability under $P_n$.

**Remark 3.1.** Consider the special case of testing a real-valued parameter. Specifically, suppose $\theta(\cdot)$ is a real-valued function from $\mathbf{P}$ to the real line. The null hypothesis is specified by $\mathbf{P_0} = \{P : \theta(P) = \theta_0\}$. Assume the alternative hypothesis is one-sided and specified by $\{P : \theta(P) > \theta_0\}$. Suppose we simply take $t_n(X_1,\ldots,X_n) = \hat{\theta}_n(X_1,\ldots,X_n) - \theta_0$. Then, it can be checked that the test construction accepts the null hypothesis if and only if the confidence interval (11) (with $\hat{b}_n \equiv b$) contains the value $\theta_0$. Thus in this special case, the test construction presented in this section has an exact duality with the interval presented in (11). This is not surprising, because the argument leading up to (11) was based on the relationship (8) and the asymptotic coverage probability of the event (7). Moreover, in the testing context, $\theta(P) = \theta_0$ is fixed and known under the null hypothesis, in which case $u_{n,b}(\alpha, P)$ in (8) can be computed, at least under the null hypothesis.

In addition, if $\hat{\theta}_n$ is a consistent estimator of $\theta(P)$, then the hypothesis on $t_n$ in part (ii) of the theorem is satisfied (just take the absolute value of $t_n$ for a two-sided alternative). Thus the hypothesis on $t_n$ in part (ii) of the theorem boils down to verifying a consistency property and is rather weak, though this assumption can in fact be weakened further. The convergence hypothesis of part (i) is satisfied by typical test statistics; in regular situations, $\tau_n = n^{1/2}$.

**Remark 3.2.** In Example 3.1, simply take $t_n = \int_{-\infty}^{\infty}[\hat{F}_n(x) - F_0(x)]^2 dF_0(x)$. Then, $t_n$ (under reasonable conditions) will converge to $t(F) = \int_{-\infty}^{\infty}[F(x) - F_0(x)]^2 dF_0(x)$, if $F$ is the distribution of $U_i$. Clearly, $t(F) = 0$ if and only if the null hypothesis is true.

**Remark 3.3.** The interpretation of part (iii) of the theorem is the following. Suppose, instead of using the subsampling construction, one could use the test that rejects when $T_n > g_n(1-\alpha, P)$, where $g_n(1-\alpha, P)$ is the exact $1-\alpha$ quantile of the true sampling distribution $G_n(\cdot, P)$. Of course, this test is not available in general because $P$ is unknown and so is $g_n(1-\alpha, P)$. Then, the asymptotic power of the subsampling test against a sequence of contiguous alternatives $\{P_n\}$ to $P$ with $P$ in $\mathbf{P_0}$ is the same as the asymptotic power of this fictitious test against the same sequence of alternatives. Hence, to the order considered, there is no loss in efficiency in terms of power.

## 4. The Time Series Case

Suppose $\{\ldots, X_{-1}, X_0, X_1, \ldots\}$ is a sequence of random variables taking values in an arbitrary sample space $S$, and defined on a common probability space. Denote the joint probability law governing the infinite sequence by $P$, assumed stationary. The goal is to construct a confidence interval for some real-valued parameter $\theta(P)$, on the basis of observing $X_1, \ldots, X_n$. The sequence $\{X_t\}$ is assumed to satisfy a certain weak dependence condition. To make this condition precise, we introduce the concept of strong mixing coefficients following Rosenblatt (1956).

**Definition 4.1.** Given a random sequence $\{X_t\}$, let $\mathcal{F}_n^m$ be the $\sigma$-algebra generated by $\{X_t, n \leq t \leq m\}$, and define the corresponding $\alpha$-mixing sequence by

$$\alpha_X(k) = \sup_n \sup_{A,B} |P(A \cap B) - P(A)P(B)|, \tag{14}$$

where $A$ and $B$ vary over the $\sigma$-fields $\mathcal{F}_{-\infty}^n$ and $\mathcal{F}_{n+k}^{\infty}$, respectively. (Note that in case the sequence $\{X_t\}$ is strictly stationary, the $\sup_n$ in this definition becomes redundant.) The sequence $\{X_t\}$ is called *$\alpha$-mixing* or *strong mixing* if $\alpha_X(k) \to 0$ as $k \to \infty$.

Throughout this section, we assume the sequence is strictly stationary, but this condition can be relaxed somewhat, as in Politis, Romano, and Wolf (1997).

Let $\hat{\theta}_n = \hat{\theta}_n(X_1, \ldots, X_n)$ be an estimator of $\theta(P) \in \mathbb{R}$, the parameter of interest.

In the context of independent data, subsamples of size $b < n$ are generated by sampling $b$ observations without replacement from the original data sequence of size $n$. Since this approach does not take the order of the original sequence into account, it generally fails for time series data. The key, therefore, is to only use *blocks* of size $b$ of consecutive observations as legitimate subsamples, the first one being $\{X_1, X_2, \ldots, X_b\}$, the last $\{X_{n-b+1}, X_{n-b+2}, \ldots, X_n\}$. There are $q = n - b + 1$ such blocks, obviously, many fewer available subsamples than in the independent case.

Define $\hat{\theta}_{n,b,t} = \hat{\theta}_b(X_t, \ldots, X_{t+b-1})$, the estimator of $\theta(P)$ based on the subsample $\{X_t, \ldots, X_{t+b-1}\}$. Let $J_b(P)$ be the sampling distribution of $\tau_b(\hat{\theta}_{n,b,1} - \theta(P))$, where $\tau_b$ is an appropriate normalizing constant. Also define the corresponding cumulative distribution function:

$$J_b(x, P) = Prob_P\{\tau_b(\hat{\theta}_{n,b,1} - \theta(P)) \leq x\}. \tag{15}$$

Essentially, to consistently estimate $J_n(P)$, we only need Assumption 2.1.

Let $Y_t$ be the block of size $b$ of the consecutive data $\{X_t, \ldots, X_{t+b-1}\}$. Only a very weak assumption on $b$ is required; typically, $b/n \to 0$ and $b \to \infty$ as $n \to \infty$. The approximation to $J_n(x, P)$ we study is the analogue of (1) for the i.i.d. case, defined by

$$L_{n,b}(x) = \frac{1}{n-b+1} \sum_{t=1}^{n-b+1} 1\{\tau_b(\hat{\theta}_{n,b,t} - \hat{\theta}_n) \leq x\}. \tag{16}$$

The motivation behind the method is the following. For any $t$, $Y_t$ is a true subsample of size $b$ from the true model $P$. Hence, the *exact* distribution of $\tau_b(\hat{\theta}_{n,b,t} - \theta(P))$ is $J_b(P)$. By stationarity, the empirical distribution of the $n-b+1$ values of $\tau_b(\hat{\theta}_{n,b,t} - \theta(P))$ should serve as good approximation to $J_n(P)$, at least for large $n$. Replacing $\theta(P)$ by $\hat{\theta}_n$ is permissible because $\tau_b(\hat{\theta}_n - \theta(P))$ is of order $\tau_b/\tau_n$ in probability and we assume $\tau_b/\tau_n \to 0$. Just as in Theorem 2.1, we show that subsampling with a general data-driven choice of block size is consistent. In order to support this claim, one must show the convergence of $L_{n,b}(\cdot)$ to $J(\cdot, P)$ is uniform in a broad range of $b$ values, say $j_n \leq b \leq k_n$ (as expressed in Theorem 2.1). Analogous to the proof of Theorem 2.1, define $U_{n,b}(x) = q_{n,b}^{-1} \sum_{t=1}^{q_{n,b}} 1\{\tau_b[\hat{\theta}_{n,b,t} - \theta(P)] \leq x\}$, where $q = q_{n,b} = n - b + 1$. Then, the proof in the i.i.d. case goes through if we can bound

$$(k_n - j_n + 1) \sup_{j_n \leq b \leq k_n} Prob_P\{|U_{n,b}(x) - J_b(x, P)| \geq t\} \tag{17}$$

by something tending to zero; see equation (5). To do this, we appeal to an exponential type inequality for mixing sequences, as provided in Theorem 1.3 of Bosq (1996). Then one can obtain uniform consistency over $b$ in $\{b : j_n \leq b \leq k_n\}$ under the assumption $k_n = o(n)$ if one is willing to slightly strengthen the mixing assumption. The result is the following.

**Theorem 4.1.** *Let $X_1, \ldots, X_n$ be observations from a stationary model with mixing coefficients $\alpha_X(\cdot)$. Let $1 \leq j_n \leq k_n \leq n$ be integers satisfying $j_n \to \infty$,*

*$k_n/n \to 0$, and $\tau_{k_n}/\tau_n \to 0$, as $n \to \infty$, $\{\tau_n\}$ nondecreasing in $n$, assume, for some $\beta > 1$,*

$$\text{limsup}_{m\to\infty} m^\beta \alpha_X(m) < \infty. \tag{18}$$

*Under Assumption 2.1, we have the following.*

(i) *If $x$ is a continuity point of $J(\cdot, P)$, then $\sup_{j_n \le b \le k_n} |L_{n,b}(x) - J(x,P)| \to 0$ in probability.*

(ii) *If $\{\hat{b}_n\}$ is a data-dependent sequence (that is, a measurable function of $X_1, \ldots, X_n$), and $Prob_P\{j_n \le \hat{b}_n \le k_n\} \to 1$, then $L_{n,\hat{b}_n}(x) \to J(x,P)$ in probability.*

(iii) *If $J(\cdot, P)$ is continuous, then $\sup_x |L_{n,\hat{b}_n}(x) - J(x,P)| \to 0$ in probability In fact, $\sup_{j_n \le b \le k_n} \sup_x |L_{n,b}(x) - J(x,P)| \to 0$ in probability.*

(iv) *Let $c_{n,\hat{b}_n}(1-\alpha) = \inf\{x : L_{n,\hat{b}_n}(x) \ge 1-\alpha\}$. Then, if $J(\cdot, P)$ is continuous, $Prob_P\{\tau_n[\hat{\theta}_n - \theta(P)] \le c_{n,\hat{b}_n}(1-\alpha)\} \to 1-\alpha$ as $n \to \infty$. Therefore, the asymptotic coverage probability under $P$ of the confidence interval $[\hat{\theta}_n - \tau_n^{-1} c_{n,\hat{b}_n}(1-\alpha), \infty)$ is the nominal level $1-\alpha$.*

**Proof.** For the proof we just need to bound (17), because the rest of the argument from Theorem 2.1 carries over. Note that $U_{n,b}(x)$ is an average of the variables $1\{\tau_b[\hat{\theta}_{n,b,t} - \theta(P)]\}$ as $t$ ranges between 1 and $n-b+1$. Moreover, as $t$ varies between 1 and $n-b+1$, these variables form a stationary sequence, each between 0 and 1, and with mixing coefficients $\alpha_{n,b}(\cdot)$ satisfying $\alpha_{n,b}(j) \le \alpha_X[\max(0, j-b+1)]$. Also, $E[U_{n,b}(x)] = J_b(x,P)$. Then, according to Theorem 1.3 in Bosq (1996), for any $q$ in $[1, \frac{n}{2}]$ and any $t > 0$, (17) is bounded above by

$$k_n 4\exp(-t^2 q/8) + k_n 22(1+\frac{4}{t})^{1/2} q\, \alpha_X([\frac{n-b+1}{2q}]).$$

Let $p = (\beta-1)/(\beta+1)$ and choose $q = n^p$. The first term in the last expression is bounded above by $4\,n\exp(-t^2 n^p/2) \to 0$. Letting $C_t = 22(1+\frac{4}{t})^{1/2}$, the second term is bounded above by

$$C_t k_n n^p \alpha_X([\frac{n-b+1}{2n^p}]) \le C_t \frac{k_n}{n} n^{p+1} \alpha_X([\frac{n-k_n+1}{2n^p}]),$$

which is bounded above by $C_t \frac{k_n}{n} n^{p+1} \alpha_X([n^{1-p}/4])$ as soon as $k_n/n \le 1/2$. Letting $m_n = n^{1-p}$ and noting that $\beta = (1+p)/(1-p)$, this bound becomes

$$C_t \frac{k_n}{n} m_n^{\frac{1+p}{1-p}} \alpha_X([m_n/4]) = C_t \frac{k_n}{n} m_n^\beta \alpha_X([m_n/4]),$$

which tends to zero by assumptions on the mixing coefficients and the fact that $k_n/n \to 0$.

Just as in the i.i.d. case, the assumption that $\tau_b/\tau_n$ can be removed if the interval is modified appropriately.

**Corollary 4.1.** *Adopt the assumptions of Theorem* 4.1*, but do not assume* $\tau_{k_n}/\tau_n \to 0$ *or that* $\tau_n$ *is monotonic. If* $J(\cdot, P)$ *is continuous at* $c(1-\alpha, P)$*, then the interval* $[\hat{\theta}_n - (\tau_n - \tau_{\hat{b}_n})^{-1} c_{n,\hat{b}_n}(1-\alpha), \infty)$ *contains* $\theta(P)$ *with asymptotic probability* $1-\alpha$ *under* $P$.

**Remark 4.1.** One can weaken the mixing condition (18) at the expense of a stronger condition on $j_n$ and $k_n$, which allows us to recover the nonrandom block size case under just mixing (as pointed out by an anonymous referee). To see how, define $I_{b,t} = 1\{\tau_b[\hat{\theta}_{n,b,t} - \theta(P)] \le x\}, t = 1, \ldots, q$ and $s_{q,h} = q^{-1} \sum_{t=1}^{q-h} Cov(I_{b,t}, I_{b,t+h})$. Then, $Var(U_n(x)) = A_n^* + A_n$, where $A_n^* = q_{-1}(s_{q,0} + 2\sum_{h=1}^{b-1} s_{q,h})$ and $A_n = 2q^{-1} \sum_{h=b}^{q-1} s_{q,h}$. By Chebychev, (17) is bounded above by $(k_n - j_n + 1)\sup_{j_n \le b \le k_n} |A_n^*| + (k_n - j_n + 1)\sup_{j_n \le b \le k_n} |A_n|$. But,

$$(k_n - j_n + 1) \sup_{j_n \le b \le k_n} |A_n^*| \le (k_n - j_n + 1) \sup_{j_n \le b \le k_n} \frac{2(b-1)+1}{n-b+1} \le \frac{2k_n(k_n - j_n + 1)}{(n - j_n + 1)}.$$

By Davydov's (1970) inequality for bounded mixing variables, $|Cov(I_{b,t}, I_{b,t+h})| \le 4\alpha_X(h-b+1)$ and so $|A_n| \le 8q^{-1} \sum_{h=1}^{q-h} \alpha_X(h)$. Therefore,

$$(k_n - j_n + 1) \sup_{j_n \le b \le k_n} |A_n| \le (k_n - j_n + 1) \cdot \frac{1}{(n - j_n + 1)} \sum_{h=1}^{n-j_n+1} \alpha_X(h).$$

Thus, if we replace the conditions on $j_n$, $k_n$, and $\alpha_X(\cdot)$ in the theorem, with $(k_n - j_n + 1)/(n - j_n + 1) \to 0$ and $(k_n - j_n + 1)\sum_{h=1}^{n-j_n+1} \alpha_X(h)/(n - j_n + 1)$, the theorem goes through (as long as we still retain $j_n \to \infty$ for the rest of the argument). In particular, in the case $j_n = k_n = b$, the conditions are satisfied if $\alpha_X(h) \to 0$ as $h \to \infty$, $b \to \infty$, and $b/n \to 0$, thereby recovering the nonrandom $b$ case. We also find the theorem goes through under just $\alpha$-mixing as long as $(k_n - j_n)$ is uniformly bounded in $n$.

## 5. Hypothesis Testing in the Stationary Case

In Section 3, it was discussed how to use subsampling for hypothesis testing when the null hypothesis does not translate into a null hypothesis on a parameter, and thus the duality between hypothesis tests and confidence regions cannot be exploited. The discussion was limited to i.i.d. observations but the problem, of course, also exists for dependent observations. Goodness of fit tests are one of many examples. The approach presented here will be analogous to the one of Section 3. To provide a general framework, assume $X_1, \ldots, X_n$ is a sample of stationary observations taking values in a sample space $S$. Denote the probability law governing the infinite, stationary sequence $\ldots, X_{-1}, X_0, X_1, \ldots$ by $P$. This unknown law $P$ is assumed to belong to a certain class of laws $\mathbf{P}$. The null hypothesis $H_0$ asserts $P \in \mathbf{P_0}$, and the alternative hypothesis $H_1$ is $P \in \mathbf{P_1}$, where $\mathbf{P_i} \subset \mathbf{P}$ and $\mathbf{P_0} \bigcup \mathbf{P_1} = \mathbf{P}$. The goal is to construct an asymptotically valid test based on a given test statistic, $T_n = \tau_n t_n(X_1, \ldots, X_n)$, where, as usual, $\tau_n$ is a fixed nonrandom normalizing sequence (but this assumption could be relaxed).

Let $G_n(x, P) = Prob_P\{\tau_n t_n(X_1, \ldots, X_n) \leq x\}$. As before, we assume that $G_n(\cdot, P)$ converges in distribution, at least for $P \in \mathbf{P_0}$. The theorem we present assumes $t_n$ is constructed to satisfy the following: $t_n(X_1, \ldots, X_n) \to t(P)$ in probability, where $t(P)$ is a constant which satisfies $t(P) = 0$ if $P \in \mathbf{P_0}$ and $t(P) > 0$ if $P \in \mathbf{P_1}$.

Let $t_{n,b,j}$ be equal to the statistic $t_b$ evaluated at the block of data $\{X_j, \ldots, X_{j+b-1}\}$. The sampling distribution of $T_n$ is then approximated by

$$\hat{G}_{n,b}(x) = \frac{1}{n-b+1} \sum_{j=1}^{n-b+1} 1\{\tau_b t_{n,b,j} \leq x\}. \tag{19}$$

Given the estimated sampling distribution, the critical value for the test is obtained as the $1 - \alpha$ quantile of $\hat{G}_{n,b}(\cdot)$; specifically, define

$$g_{n,b}(1-\alpha) = \inf\{x : \hat{G}_{n,b}(x) \geq 1 - \alpha\}. \tag{20}$$

Finally, the nominal level $\alpha$ test rejects $H_0$ if and only if $T_n > g_{n,b}(1 - \alpha)$.

**Theorem 5.1.**

(i) *Assume, for $P \in \mathbf{P_0}$, $G_n(P)$ converges weakly to a continuous limit law $G(P)$, whose corresponding cumulative distribution function is $G(\cdot, P)$ and whose $1-\alpha$ quantile is $g(1-\alpha, P)$. Assume $b/n \to 0$ and $b \to \infty$ as $n \to \infty$. Also, assume that $\alpha_X(m) \to 0$ as $m \to \infty$, where $\alpha_X(\cdot)$ is the mixing sequence corresponding to $\{X_t\}$. If $G(\cdot, P)$ is continuous at $g(1 - \alpha, P)$ and $P \in \mathbf{P_0}$, then $g_{n,b}(1-\alpha) \to g(1-\alpha, P)$ in probability and $Prob_P\{T_n > g_{n,b}(1-\alpha)\} \to \alpha$ as $n \to \infty$.*

(ii) *Assume the test statistic is constructed so that $t_n(X_1, \ldots, X_n) \to t(P)$ in probability, where $t(P)$ is a constant which satisfies $t(P) = 0$ if $P \in \mathbf{P_0}$ and $t(P) > 0$ if $P \in \mathbf{P_1}$. Assume $b/n \to 0$, $b \to \infty$, and $\tau_b/\tau_n \to 0$ as $n \to \infty$. Also, assume that $\alpha_X(m) \to 0$ as $m \to \infty$, where $\alpha_X(\cdot)$ is the mixing sequence corresponding to $\{X_t\}$. Then if $P \in \mathbf{P_1}$, the rejection probability satisfies $Prob_P\{T_n > g_{n,b}(1 - \alpha)\} \to 1$ as $n \to \infty$.*

(iii) *Suppose $P_n$ is a sequence of alternatives such that, for some $P_0 \in \mathbf{P_0}$, $\{P_n^{[n]}\}$ is contiguous to $\{P_0^{[n]}\}$ ($P_n^{[n]}$ denotes the law of the finite segment $X_1, \ldots, X_n$ when the law of the infinite sequence $\ldots, X_{-1}, X_0, X_1, \ldots$ is given by $P_n$; the meaning of $\{P_0^{[n]}\}$ is analogous). Assume $b/n \to 0$ and $b \to \infty$ as $n \to \infty$.*

*Then* $g_{n,b}(1-\alpha) \to g(1-\alpha, P_0)$ *in* $P_n^{[n]}$*-probability. Hence, if* $T_n$ *converges in distribution to* $T$ *under* $P_n$ *and* $G(\cdot, P_0)$ *is continuous at* $g(1-\alpha, P_0)$, $P_n^{[n]}\{T_n > g_{n,b}(1-\alpha)\} \to Prob\{T > g(1-\alpha, P_0\}$.

**Proof.** The proof mimicks the proof of Theorem 3.1, with differences being analogous to the differences between the proofs of Theorems 2.1 and 4.1.

**Remark 5.1.** Remarks 3.1 and 3.3 also apply here.

## 6. An Example

The goal of this section is to illustrate the idea of data-dependent choice of block size by presenting a heuristic algorithm and a small simulation study. More extensive simulations can be found in Politis, Romano and Wolf (1997). Our algorithm is based on the fact that for the subsampling method to be consistent, the block size $b$ needs to tend to infinity with the sample size $n$ but at a smaller rate, satisfying $b/n \to 0$. Indeed, for $b$ too close to $n$ all subsample statistics ($\hat{\theta}_{n,b,i}$ or $\hat{\theta}_{n,b,t}$) will be almost equal to $\hat{\theta}_n$, resulting in the subsampling distribution being too tight and in undercoverage of subsampling confidence intervals. Lahiri (1998) makes this intuition precise by proving, in the context of mean-like statistics, that for $b/n \to 1$, the subsampling approximation collapses to a point mass at zero. On the other hand, if $b$ is too small, the intervals can undercover or overcover depending on the state of nature; e.g., see Table 1. This leaves a number of $b$ values in the 'right range' where we would expect almost correct results, at least for large sample sizes. Hence, in this range, the confidence intervals should be 'stable' when considered as a function of the block size. This idea is exploited by computing subsampling intervals for a large number of block sizes $b$, and then looking for a region where the intervals do not change very much. Within this region, an interval is picked according to some reasonable criterion.

While this method can be carried out by 'visual inspection', it is desirable to also have some automatic selection procedure, at the very least when simulation studies are to be carried out. The procedure we propose is based on minimizing a running standard deviation. Assume one computes subsampling intervals for block sizes $b$ in the range of $b_{small}$ to $b_{big}$. The endpoints of the confidence intervals should change in a smooth fashion, as $b$ changes. A running standard deviation applied to the endpoints determines the volatility around a specific $b$ value, and the value of $b$ associated with the smallest volatility is chosen. Here is a more formal description of the algorithm.

Table 1. Univariate mean, AR(1) model, $n = 250$. Estimated coverage probabilities of nominal 95% symmetric confidence intervals for the

univariate mean. The estimates are based on 1000 replications for each scenario.

**Gaussian innovations**

| Parameter | $b = 4$ | $b = 8$ | $b = 16$ | $b = 32$ | Data-driven |
|---|---|---|---|---|---|
| $\rho = 0.2$ | 0.93 | 0.92 | 0.91 | 0.89 | 0.93 |
| $\rho = 0.5$ | 0.87 | 0.90 | 0.89 | 0.88 | 0.92 |
| $\rho = 0.8$ | 0.74 | 0.84 | 0.87 | 0.87 | 0.87 |
| $\rho = 0.95$ | 0.41 | 0.53 | 0.64 | 0.73 | 0.74 |
| $\rho = -0.5$ | 0.97 | 0.95 | 0.94 | 0.92 | 0.93 |

**Exponential innovations**

| Parameter | $b = 4$ | $b = 8$ | $b = 16$ | $b = 32$ | Data-driven |
|---|---|---|---|---|---|
| $\rho = 0.2$ | 0.92 | 0.92 | 0.92 | 0.89 | 0.92 |
| $\rho = 0.5$ | 0.88 | 0.90 | 0.90 | 0.88 | 0.91 |
| $\rho = 0.8$ | 0.70 | 0.81 | 0.86 | 0.86 | 0.90 |
| $\rho = 0.95$ | 0.44 | 0.57 | 0.67 | 0.74 | 0.72 |
| $\rho = -0.5$ | 0.97 | 0.95 | 0.93 | 0.91 | 0.94 |

**Algorithm 6.1.** (Minimizing confidence interval volatility)

(i) For $b = b_{small}$ to $b = b_{big}$, compute a subsampling interval for $\theta(P)$ at the desired confidence level, resulting in endpoints $I_{b,low}$ and $I_{b,up}$.

(ii) For a small integer $k$, let $VI_b$ be the standard deviation of the endpoints $\{I_{b-k,low}, \ldots, I_{b+k,low}\}$ plus the standard deviation of the endpoints $\{I_{b-k,up}, \ldots, I_{b+k,up}\}$.

(iii) Pick $b^*$ corresponding to the smallest volatility index $VI_b$ and report $[I_{b^*,low}, I_{b^*,up}]$ as the final confidence interval.

**Remark 6.1.** The range of $b$ values, determined by $b_{small}$ and $b_{big}$, which is included in the minimization algorithm is not very important, as long as it is not too narrow. In the terminology of Sections 2 and 4, we can think of $b_{small}$ as corresponding to $j_n$ and $b_{big}$ as corresponding to $k_n$. Dependence on $n$ has been suppressed.

**Remark 6.2.** The algorithm contains a model parameter $k$. Simulation studies have shown that the algorithm is not very sensitive to the choice of this parameter. We typically employ $k = 2$ or $k = 3$.

Using a simulation study, we can compare the performance of this data-driven choice of block size with that of the best *fixed* block size, which in practice is unknown. Performance will be measured by empirical coverage probability of nominal 95% symmetric confidence interval for the univariate mean. As the data generating process, a simple AR(1) model is used, given by $X_t = \rho X_{t-1} + \epsilon_t$, where the $\epsilon_t$ are i.i.d. standard normal or (centered) exponential with mean 1.

The closer the AR(1) parameter $\rho$ is to one in absolute value, the stronger is the dependence of the $\{X_t\}$ sequence. The values of $\rho$ included in the study are $\rho = 0.2$, 0.5, 0.8, 0.95, and -0.5 and the sample size considered is $n = 250$. We compare the fixed block sizes $b = 4$, 8, 16, and 32 with the above data-dependent choice of block size using $b_{small} = 4$ and $b_{big} = 40$. The results are presented in Table 1.

One can see that the best fixed block size changes significantly with the AR(1) parameter $\rho$ and the larger is $\rho$ in absolute value, the larger is in general the optimal block size. This is not surprising, since bigger block sizes should be needed to capture stronger dependence structures. For positive $\rho$, the intervals tend to undercover and, again not surprising, the performance decreases for larger $\rho$. For the negative value $\rho = -0.5$, the intervals overcover for small block sizes, but undercover eventually (which is a consequence of the formerly stated theoretical results). The data-driven method of choosing the block size does about as well as the best fixed block size. This is encouraging, since the data-driven method is feasible while the optimal block size is unknown in practice.

## Acknowledgements

We are grateful to an anonymous referee for improving the conditions in Theorems 2.1 and 4.1 (especially Remark 4.1), thereby showing how the fixed block size case follows from these theorems.

## References

Beran, R. (1984). Bootstrap methods in statistics. *Jahresberichte des Deutschen Mathematischen Vereins* **86**, 14-30.

Beran, R. (1986). Simulated power function. *Ann. Statist.* **14**, 151-173.

Bertail, P., Politis, D. N. and Romano, J. P. (1999). On subsampling estimators with unknown rate of convergence. *J. Amer. Statist. Assoc.* **94**, 569-579.

Bickel, P. and Ren, J. (1996). The $m$ out of $n$ bootstrap and goodness of fit tests with doubly censored data. *Robust Statistics, Data Analysis, and Computer Intensive Methods.* Lecture Notes in Statistics, Springer, 35-47.

Bickel, P. and Ren, J. (1997). On choice of $m$ for the $m$ out of $n$ bootstrap in hypothesis testing. Preprint, Department of Statistics, University of California, Berkeley.

Bosq, D. (1996). Nonparametric statistics for stochastic processes: estimation and prediction. *Lecture Notes in Statist.* **110**. Springer, New York.

Davydov, Y. A. (1970). The invariance principle for stationary processes. *Theory Probab. Appl.* **14**, 487-498.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Ann. Statist.* **7**, 1-26.

Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans.* CBMS-NSF Regional Conference Series in Applied Mathematics, Society for Industrial and Applied Mathematics, Philadelphia.

Giné, E. and Zinn, J. (1990). Bootstrapping general empirical measures. *Ann. Probab.* **18**, 851-869.

Hall, P. (1988). On symmetric bootstrap confidence intervals. *J. Roy. Statist. Soc. Ser. B* **50**, 35-45.

Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*. Springer, New York.

Hartigan, J. (1969). Using subsample values as typical values. *J. Amer. Statist. Assoc.* **64**, 1303-1317.

Hartigan, J. (1975). Necessary and sufficient conditions for asymptotic joint normality of a statistic and its subsample values. *Ann. Statist.* **3**, 573-580.

Lahiri, S. N. (1998). Effects of block lengths on the validity of block resampling methods. Preprint, Department of Statistics, Iowa State University.

Mahalanobis, P. (1946). Sample surveys of crop yields in India. *Sankhyā Ser. A* **7**, 269-280.

McCarthy, P. (1969). Pseudo-replication: half-samples. *Rev. Internat. Statist. Inst.* **37**, 239-263.

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

Politis, D. N. and Romano, J. P. (1994). Large sample confidence regions based on subsamples under minimal assumptions. *Ann. Statist.* **22**, 2031-2050.

Politis, D. N. and Romano, J. P. (1996). Subsampling for econometric models—Comments on bootstrapping time series models. *Econom. Rev.* **15**, 169-176.

Politis, D. N., Romano, J. P. and Wolf, M. (1997). Subsampling for heteroskedastic time series. *J. Econom.* **81**, 281-317.

Romano, J. P. (1988). A bootstrap revival of some nonparametric distance tests. *J. Amer. Statist. Assoc.* **83**, 698-708.

Romano, J. P. (1989). Bootstrap and randomization tests of some nonparametric hypotheses. *Ann. Statist.* **17**, 141-159.

Rosenblatt, M. (1956). A central limit theorem and a strong mixing condition. *Proc. Nat. Acad. Sci.* **42**, 43-47.

Serfling, R. (1980). *Approximation Theorems of Mathematical Statistics*. John Wiley, New York.

Shao, J. and Wu, C. F. (1989). A general theory for jackknife variance estimation. *Ann. Statist.* **17**, 1176-1197.

Sherman, M. and Carlstein, E. (1996). Replicate histograms. *J. Amer. Statist. Assoc.* **91**, 566-576.

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

Department of Mathematics, University of California, San Diego, U.S.A.

Department of Statistics, Seguoia Hall Stanford University, Stanford, CA 94305, U.S.A.

E-mail: romano@stat.stanford.edu

Departamento de Estadística y Econometría, Universidad Carlos III, Madrid, Spain.

(Received May 1999; accepted April 2001)