REPLICATE HISTOGRAMS

by

Michael Sherman

and

Ed Carlstein

October 1992

Mimeo Series #2086

DEPARTMENT OF STATISTICS
Chapel Hill, North Carolina

2036 FAN, J., TRUONG, Y. & WANG, Y.: Nonparametric function estimation involving errors-in-variables, Sept. 1990. *Nonparametric Functional Estimation and Related Topics* (G. Roussas, ed.) *1991, 613–627.*

2037 JI, C.: Sieve estimators for pair-interaction potentials and local characteristics in Gibbs random fields, Oct. 1990. *Ann. Statist., to appear.*

2038 CHAKRAVARTI, I.M.: A three-class association scheme on the flags of a finite projective plane and a (PBIB) design defined by the incidence of the flags and the Baer subplanes in $PG(2,q^2)$, Oct. 1990. *Discrete Math., to appear.*

2039 CHAKRAVARTI, I.M.: Geometric construction of some families of two-class and three-class association schemes and codes from non-degenerate and deg… *Discrete Math., to appear.*

2040 CARLSTEIN, E.: Resampling techniques for stationary ti… Oct. 1990. *New Directions in Time Series Analysis, Part I* … *Murray Rosenblatt, eds.) 1992, Springer-Verlag, IMA Volu… to appear.*

2041 FAN, J. & MARRON, S.: Best possible constant for band… *Statist., to appear.*

2042 FAN, J., TRUONG, Y.K. & WANG, Y.: Measurement e… approach, Oct. 1990. *(30 pages)*

2043 REN, J.J.: On Hadamard differentiability and M-estima… *(dissertation)*

2044 SEN, P.K. & SENGUPTA, D.: On characterizations of … estimators, Dec. 1990. *Comm. Statist., Th. & Methods, 2…*

2045 TRUONG, Y.K. & FAN, J.: Deconvolution problems in…

2046 HU, T.-C. & FAN, J.: Bias correction and higher order … *Prob. Lttrs., 13, 1992, 235–248.*

2047 HU, T.-C. & WEBER, N.: On the invariance principle … Jan. 1991. *(6 pages)*

2048 HU, T.-C.: A study on the law of the iterated logarithm for arrays of r… *Comm. Statist., Th. & Meth., A20, 1991, 1989–1994.*

2049 FAN, J.: Design-adaptive nonparametric regression, Feb. 1991. *J. Amer. Statist. Ass…* *to appear.*

2050 SEN, P.K.: Some informative aspects of jackknifing and bootstrapping, May 1991. *Order Statistics and Nonparametrics: Theory and Applications* (eds. I.A. Salama & P.K. Sen), *1992, North Holland, to appear.*

2051 HALL, P., HU, T.-C. & MARRON, S.: Improved variable window kernel estimates of probability densities, Apr. 1991. *(19 pages)*

2052 FAN, J. & GIJBELS, I.: Variable Bandwidth and Local Linear Regression Smoothers, Apr. 1991. *Ann. Statist., to appear.*

2053 ALDERSHOF, B.: Estimation of integrated squared density derivatives, Apr. 1991. *(dissertation)*

2054 BROOKS, M.: Bandwidth selection methods for kernel estimators of the intensity function of a nonhomogeneous Poisson process, Apr. 1991. *(dissertation)*

2055 FAN, J. & GIJBELS, I.: Local linear smoothers in regression function estimation, May 1991. *(27 pages)*

2056 FAN, J. & GIJBELS, I.: Minimax estimation of a bounded squared mean, May 1991. *Statist. & Prob. Ltrs., 13, 1992, 383–390.*

2057 SIMONS, G. & YAO, Y.-C.: A three color problem, May 1991. *(17 pages)*

2058 SEN, P.K.: … statistical estimation, May 1991. *Festschrift in* … *65–82.*

… May 1991. *(thesis)*

… adaptive nonparametric function estimation: A …

… stimation with errors-in-variables: Asymptotic … *tivariate Anal., to appear.*

… ov random field texture models, July 1991.

… Use of moments in distribution theory: A …

… ency, Sept. 1991. *(19 pages)*

… etermining the best among three treatments …

… emal index, Nov. 1991. *(30 pages)*

… n statistics and chaos, Nov. 1991. *Statist. Sc.,*

… dmirovich Gnedenko, Nov. 1991. *Statist. Sc.*

… undaries: A nonparametric approach,

2070 SHERMAN, M. & CARLSTEIN, E.: Method of moments estimation: Simple and complicated settings, Jan. 1992. *(11 pages)*

2071 GHOSH, J.K., SEN, P.K. & MUKERJEE, R.: Second order Pitman closeness and Pitman admissibility, Feb. 1992. *(17 pages)*

2072 JOHNSON, N.L., KOTZ, S. & PEARN, W.L.: Flexible process capability indices, Feb. 1992. *(10 pages)*

2073 JUREČKOVÁ, J. & SEN, P.K.: Asymptotic equivalence of regression rank scores estimators and R-estimators in linear models, Feb. 1992. *(16 pages)*

REPLICATE HISTOGRAMS

Michael Sherman
Department of Biostatistics, University of Rochester

Ed Carlstein
Department of Statistics, University of North Carolina at Chapel Hill

ABSTRACT

The "replicate histogram" is introduced as a simple diagnostic tool for describing the sampling distribution of a general statistic. It can be applied to virtually any statistic that has an asymptotic distribution, and, the data on which the statistic is computed may be serially or spatially dependent. The method is completely sample-based, requiring no theoretical analysis by the user, no knowledge of the proper standardization for the statistic, and no specification of the underlying dependence mechanism generating the data. The replicate histogram warns the user of non-normal sampling distributions, and also indicates the type of departure from normality (e.g., skewness, peakedness). In the case of spatially dependent data, the statistic may be computed on observations from irregularly shaped index-sets. Large-sample validity of the replicate histogram is established via strong consistency results, which are proved under mild conditions for both the time-series and random field cases. Examples are presented illustrating the diagnostic power of the replicate histogram for time-series and spatial data-sets.

Research supported by National Science Foundation Grants DMS-8902973 and DMS-9205881.

1. INTRODUCTION

Here is a typical scenario for statistical inference: A series of n observations $\mathbf{X}_n := (X_1, X_2, ..., X_n)$ is obtained from a random process that is controlled by unknown parameters (θ,ν). The "target parameter" $\theta \in \mathbf{R}$ is estimated by a scalar statistic $s_n := s_n(\mathbf{X}_n)$, while ν is a "nuisance parameter" which may be a vector or even infinite-dimensional.

In order to draw statistical inferences from s_n to θ (e.g., confidence intervals, hypothesis tests), one must make assumptions about the sampling distribution of s_n, or, as a practical approximation, assumptions about the asymptotic distribution (F) of a standardized transform of s_n, say, $t_n := a_n(s_n - b_n)$, $a_n>0$. The validity of the inferences therefore relies upon the appropriateness of the chosen F (e.g., normal, χ^2). The "replicate histogram" is a simple diagnostic tool for assessing the appropriateness of F, using only the data $\mathbf{X}_n$ at hand.

To illustrate the need for using the replicate histogram, consider the following obstacles which confront the statistician in trying to correctly determine $F(y) := \lim_{n\to\infty} \mathbf{P}\{t_n \le y\}$.

(i) The statistic s_n may be complicated (e.g., a robustified or adaptively defined statistic, like Switzer's adaptive trimmed mean [see Efron (1982), Example 5.2]), so that a theoretical derivation of F is analytically intractable.

(ii) The observations might be serially dependent, so their joint probability structure must be accounted for in deriving F. This in turn may require knowledge or assumptions about the underlying serial dependence mechanism, and estimation of its concomitant parameters. For example, in an assumed AR(p) model, ν would include the noise distribution as well as the p autoregressive coefficients -- all of which could be relevant in deriving F.

(iii) The proper standardization constants (a_n, b_n) may involve the unknown parameters (θ,ν) or may not even be known in functional form; moreover they may actually be crucial in determining the fundamental characteristics of F (e.g., symmetric vs. skewed; normal vs. χ^2). For example, consider the simple case where s_n is the sample mean of i.i.d. observations, so that F is necessarily a stable distribution [see Ibragimov and Linnik (1971), Chapter 2]. The appropriate standardizing coefficient is $a_n=n^{1/2}h(n)$ (where $h(\cdot)$ is a slowly varying function) when F is the [symmetric] normal distribution. On the other hand, the appropriate standardizing coefficient is $a_n=n^{(\alpha-1)/\alpha}h(n)$ (where $\alpha \in (0,2)$) when F is an α-stable distribution [which may actually be quite skewed]. As another example, consider the situation where s_n is a U-statistic computed from i.i.d. observations [see Serfling (1980), Chapter 5]. When the U-statistic is "non-degenerate", the appropriate standardizing coefficient is $a_n=n^{1/2}$, and F is a normal distribution. But when the U-statistic is "degenerate", the appropriate standardizing coefficient is $a_n=n$, and F is then the distribution of a weighted sum of χ^2 random variables. The degenerate case is not an arcane theoretical construct: Several common goodness-of-fit statistics are degenerate U-statistics; and, moreover, a single U-statistic can behave as either degenerate or non-degenerate according to subtle changes in the underlying parameters (θ,ν) [see Example A in

Section 4].

Obstacles (i), (ii), (iii) could arise together in a particular statistical problem: When the statistic is analytically intractable (i), then the entire functional form of the standardizing coefficient a_n will typically remain unknown (iii). The presence of serial dependence (ii) would exacerbate any analytical difficulties encountered (i). The qualitative behavior of the two examples discussed in (iii) remains essentially unchanged when serial dependence (ii) is present [see Ibragimov and Linnik (1971), Chapter 18, and Carlstein (1988)]. The examples in Section 4 illustrate the use of replicate histograms for statistics whose behavior is fundamentally different from $\bar{X}$, in the presence of obstacles (ii) and (iii).

We now define the replicate histogram for time-series data and explain how it avoids these three obstacles. Section 3 will present theoretical asymptotic properties of the replicate histogram, as well as the generalization of replicate histograms to spatial data.

A general statistic is determined by a sequence of completely known functions $\{s_m(\cdot)\colon m \geq 1\}$, $s_m\colon \mathbf{R}^{md} \to \mathbf{R}^1$, where $d \geq 1$ is the dimension of X_i. Introduce the following notation for a "subseries" of consecutive observations:

$$X_l^i := (X_{i+1}, X_{i+2}, \ldots, X_{i+l}),$$

so that the observed data-set is X_n^0, and the collection of all available subseries with length l is $\{X_l^i\colon 0 \leq i \leq n-l\}$. The associated "replicates" are denoted by $s_l^i := s_l(X_l^i)$. The replicate histogram corresponds to the distribution:

$$G_n(y) := \frac{\sum_{i=0}^{n-l(n)} I\{s_{l(n)}^i \leq y\}}{n - l(n) + 1}, \qquad y \in \mathbf{R},$$

where $l(n) := \lfloor cn^\gamma \rfloor$, for any fixed $c > 0$ and $\gamma \in (0,1/2)$. This is simply the empirical distribution of the replicates. Note how the replicate histogram avoids obstacles (i), (ii), and (iii):

(i) The replicate histogram is directly computable from the available data, so no theoretical analysis whatsoever is necessary, regardless of how complicated the statistic may be. Moreover, the replicate histogram "works" [i.e., is a strongly consistent diagnostic tool, in a sense which will be made precise in Section 3] for virtually any statistic t_n that has an asymptotic distribution.

(ii) By employing subseries replicates of the statistic, the correct serial dependence structure is automatically retained, without any knowledge of or assumptions about the underlying dependence mechanism. Again, the replicate histogram "works" in the presence of serial dependence, provided only that the strength of the dependence satisfies a mild model-free "mixing" condition [see Section 3].

(iii) Since the replicate histogram is constructed from replicates of the unstandardized statistic s_n, there is no need to analytically derive (nor to guess) a_n and b_n.

Thus, the replicate histogram is an "omnibus" procedure: It applies to a general statistic in a general setting, so that each new scenario (e.g., a new $s_n(\cdot)$ or a new serial dependence mechanism) does not require the development of a new procedure. The replicate histogram shares "the charm of

the jackknife and the bootstrap", which is, quoting Efron (1982), "that they can be applied to complicated situations where parametric modeling and/or theoretical analysis is hopeless." Naturally, any method which achieves such seemingly far-reaching applicability must do so at some cost. This cost, for the replicate histogram, is that it returns only diagnostic information about the shape of F (e.g., symmetry vs. skewness), but does not return numerical estimates of the percentiles or probabilities induced by F.

2. COMPARISON WITH OTHER METHODS

Several methods have been proposed for estimating the distribution of a statistic in various specific settings, but none are diagnostic tools in as general a setting as that considered here. Wu (1990) says "a major purpose of resampling is to use the observed data to construct a distribution that mimics the unknown distribution of [the statistic]"; this is his motivation for developing the "jackknife histogram". However, he only considers asymptotically normal statistics computed from i.i.d. data.

The bootstrap has also been used for estimating the distribution of a statistic in many different settings: in the i.i.d. case, for statistics with normal limiting distributions (Bickel and Freedman (1981), Singh (1981)) and for statistics with non-normal limiting distributions (Athreya (1987), Bretagnolle (1983), Swanepoel (1986)); in the time-series case, for statistics with normal limiting distributions (Freedman (1984), Bose (1988), Rajarshi (1990), Künsch (1989)) and for statistics with non-normal limiting distributions (Basawa et al. (1989)); and in the spatially dependent case, for statistics with normal limiting distributions (Lele (1988)). But the bootstrap cannot be applied to these settings in an omnibus way: The bootstrap algorithm (and the theory justifying it) must be tailored to each specific situation, using knowledge or assumptions about (θ,ν). For example, in the Athreya (1987), Bretagnolle (1983), and Swanepoel (1986) results for asymptotically non-normal statistics, knowledge of the particular limiting distribution F (and the standardizing constants a_n and b_n) is needed to determine the correct standardization for the resampled statistic and the correct bootstrap resample size (which must differ from n). The Freedman (1984), Bose (1988), and Basawa et al. (1989) results rely crucially on the user's knowledge of the underlying serial dependence mechanism (autoregression) which generated the data [Freedman (1984) emphasizes this issue]. The Künsch (1989), Rajarshi (1990), and Lele (1988) results allow for a variety of dependence structures, but they all assume that the statistic has a normal limiting distribution. In some asymptotically normal cases (Singh (1981), Bose (1988)) the bootstrap does pick up an extra term in the Edgeworth expansion.

In summary, the above methods provide a piecemeal approach to describing the distribution of a general statistic. Each method is situation-specific and demands that the user have substantial prior knowledge: "Does my data come from an AR(p) process?" "Is the statistic asymptotically normal?" "If not, what is the correct limiting distribution?" "What are the corresponding appropriate standardizing constants?" "What is the correct resample size?" "Can the jackknife/bootstrap be

theoretically justified for my particular scenario?" If the user has this prior knowledge then the appropriate jackknife or bootstrap algorithm can and should be used (if it exists). In the absence of this knowledge, the replicate histogram still provides useful diagnostic information about the form of the limiting distribution, without any knowledge of the standardizations or the underlying dependence mechanism, and without any further situation-specific theoretical analysis. It will not, however, give percentiles of the limiting distribution of the (standardized) statistic. Thus, the replicate histogram is not a replacement for the bootstrap or jackknife, but rather is an omnibus diagnostic tool which should be used in a complementary or corroborative way.

In simultaneous and independent work, Politis and Romano (1992) study a variant of the replicate histogram for the non-diagnostic purpose of constructing confidence intervals in the case where the appropriate theoretical standardizations are known; for spatial data, their method requires observations on a rectangularly shaped index-set.

3. PROPERTIES OF THE REPLICATE HISTOGRAM

(A) Time-Series Data.

The random process $\{X_i\colon\ -\infty < i < +\infty\}$ is assumed to be stationary, and the strength of serial dependence is measured by the standard model-free "mixing coefficient"

$$\alpha(m) := \sup\Big\{\,|\,\mathbf{P}\{A\cap B\} - \mathbf{P}\{A\}\mathbf{P}\{B\}\,|\colon\ A\in\mathcal{F}(\ldots, X_{-1}, X_0),\ B\in\mathcal{F}(X_m, X_{m+1}, \ldots)\Big\},$$

as introduced by Rosenblatt (1956). Intuitively, the requirement that $\alpha(m)\to 0$ as $m\to\infty$ says that observations separated by a large time-lag behave approximately as if they were independent. This mixing property, together with stationarity, guarantees that the replicates defined in Section 1 are "valid".

In order to establish strong consistency for the replicate histogram, we need to consider the following theoretical transformation:

$$\hat{F}_n(y) \;:=\; G_n(b_{l(n)} + y/a_{l(n)}) \;=\; \frac{\sum_{i=0}^{n-l(n)} \mathbf{I}\{t^i_{l(n)} \le y\}}{n - l(n) + 1},$$

where $t^i_l := a_l(s^i_l - b_l)$. Notice that $\{G_n(y)\colon\ y\in\mathbf{R}\}$ (as defined in Section 1), which is directly computable from the observed data, contains essentially the same diagnostic information as the unknown $\{\hat{F}_n(y)\colon\ y\in\mathbf{R}\}$. For example, the standardization transformation preserves normality, symmetry, skewness (as measured by, e.g., the skewness coefficient $\beta(Y) := \mathbf{E}^2\{(Y-\mathbf{E}\{Y\})^3\}/\mathbf{V}^3\{Y\}$ [see Kendall and Stuart (1977), p. 87], or the skewness parameter β of the α-stable distributions [see Ibragimov and Linnik (1971), Theorem 2.2.2]). The following result justifies the replicate histogram as a diagnostic tool for describing F.

Theorem 1:

If $\alpha(m) = O(m^{-\epsilon})$ for some $\epsilon > 1/2\gamma$, then

$$\hat{F}_n(y) \overset{a.s.}{\underset{n\to\infty}{\Longrightarrow}} F(y), \quad \forall y \in \mathbf{R}.$$

Furthermore, if F is continuous then $\sup_y |\hat{F}_n(y) - F(y)| \overset{a.s.}{\underset{n\to\infty}{\Longrightarrow}} 0$.

The rate (ϵ) can be interpreted as follows: Recall that γ determines the length of the subseries. For large ϵ, the time-series is nearly i.i.d., and in this case γ can be small. For smaller ϵ, however, we see that γ needs to be larger; this is because small ϵ signifies strong serial dependence, and the subseries must be relatively long to capture the full extent of this dependence. The proof of Theorem 1 is in the Appendix.

(B) Spatial Data.

The replicate histogram can also be used to describe the distribution of a statistic computed on data from a possibly irregularly shaped set of indices in a stationary random field $\{X_i : i \in \mathbf{Z}^2\}$. One context where such a method is needed is in image analysis. As discussed in Section 1, there will typically be parameters to estimate; in order to make inferences from a statistic to the parameter it estimates, we will need information about the statistic's distribution.

The basic idea is the same as in the case of time-series data. Here we compute the statistic, s, on overlapping "subshapes" of data (analogous to subseries). For each n, let D_n be a finite set of lattice points in $\mathbf{Z}^2$, at which observations are taken; the cardinality of D_n is denoted $|D_n|$. In general, the "shape" D_n may be quite irregular. Formally, let $S_{D_n}(\cdot)$ be a function from $\mathbf{R}^{|D_n|}$ to $\mathbf{R}^1$, and let $s(D_n) := S_{D_n}(X_j : j \in D_n)$ be the corresponding statistic [$S_{D_n}(\cdot)$ is assumed to be invariant under translations of the set D_n]. As in the time-series case, we assume that the standardized statistic $t(D_n) := a_n(s(D_n) - b_n)$ has asymptotic distribution $F(y) := \lim_{n\to\infty} \mathbf{P}\{t(D_n) \le y\}$, and our goal is to obtain diagnostic information about F without any knowledge of the standardizations or the underlying spatial dependence mechanism, and without any theoretical analysis by the user. Let $D^i_{l(n)}$, $i = 1, 2, \ldots, k_n$ denote overlapping subshapes of D_n, where $l(n)$ determines the common size of each subshape [$l(n) = \lfloor cn^\gamma \rfloor$ as in Section 1] and k_n denotes the number of available subshapes. Analogously to Section 1 we compute the distribution:

$$G_n(y) := \frac{\sum_{i=1}^{k_n} \mathbf{I}\{s(D^i_{l(n)}) \le y\}}{k_n}, \quad y \in \mathbf{R},$$

and use the corresponding replicate histogram to obtain information about F.

We now describe the index sets D_n and $D^i_{l(n)}$ precisely. Let $A \subset (0,1] \times (0,1]$ be the interior of a simple closed curve which will serve as a template for D_n and $D^i_{l(n)}$. To avoid pathological cases we assume that the boundary of A has finite length. Now multiply the set A by n, to obtain the set

$A_n \subset (0,n] \times (0,n]$; i.e., A_n is the shape A inflated by a factor of n. The data are observed at the indices in $D_n := \{i \in A_n \cap \mathbb{Z}^2\}$. This formulation allows for a wide variety of shapes on which our data can be observed.

"Subshapes" are obtained by resolving $(0,n]^2$ into $l(n) \times l(n)$ overlapping subsquares [of which there are $(n - l(n)+1)^2$]. In $(0,l(n)] \times (0,l(n)]$ identify the set $D_{l(n)}$, and do the same in each subsquare by simply translating the origin. Since there is only data in D_n, we can only use the k_n subshapes $D^i_{l(n)}$, $i=1,2,...,k_n$ whose indices are all contained in A_n. Notice that each subshape replicate $s(D^i_{l(n)})$ mimics the spatial structure of $s(D_n)$.

In order to formally state the strong consistency result for $\hat{F}_n(y) := G_n(b_{l(n)} + y/a_{l(n)})$, the spatial dependence needs to be quantified. As in Section 3(A), we measure the strength of dependence by a model-free mixing coefficient,

$$\alpha_p(m) := \sup\Big\{ |P\{A \cap B\} - P\{A\}P\{B\}| : A \in \mathcal{F}(\Lambda_1),\ B \in \mathcal{F}(\Lambda_2),\ |\Lambda_1| \le p,\ |\Lambda_2| \le p,\ d(\Lambda_1,\Lambda_2) \ge m \Big\},$$

where $\mathcal{F}(\Lambda_i)$ contains the events depending on $\{X_j : j \in \Lambda_i\}$ and $d(\Lambda_1,\Lambda_2)$ is the minimal city-block distance between index-sets Λ_1 and Λ_2. Note that in the random field setup, the dependence between two sets of random variables (characterized by $\alpha_p(m)$) is a function not only of the distance (m) between the two sets, but also of each set's cardinality (p). In time-series, cardinality is generally not accounted for; this is considered acceptable because there are many standard examples which satisfy α-mixing, e.g., AR(1) processes with normal, double-exponential, or Cauchy errors (see Gastwirth and Rubin (1975)). In the random field case there is no consensus as to whether accounting for cardinality is necessary. Bradley (1991) has shown that, for some random fields, mixing conditions that account for cardinality [like (*) below] hold while mixing conditions that do not account for cardinality fail. For this reason we account for cardinality in our mixing coefficient. Assume that the following mixing condition holds:

$$\sup_p \frac{\alpha_p(m)}{p} = \mathcal{O}(m^{-\epsilon}) \text{ for some } \epsilon > 2+(1/\gamma). \qquad (*)$$

Condition (*) says that, at a fixed distance (m), as the cardinality increases we allow dependence to increase at a rate controlled by p. As the distance increases, the dependence must decrease at a polynomial rate in m. The relationship between this rate (ϵ) and the subshape size (γ) is similar to that in the discussion after Theorem 1.

The following result justifies the replicate histogram as a diagnostic tool for describing the sampling distribution of a statistic computed from spatially dependent data (in a possibly irregularly shaped index-set). The proof of this result is in the Appendix.

Theorem 2:

If mixing condition (*) is satisfied, then the conclusion of Theorem 1 again holds.

4. EXAMPLES

In this Section, we illustrate how the replicate histogram can be used for diagnostics on a sampling distribution, based on a finite data-set, without any knowledge of the proper standardizations or the underlying dependence mechanism, and without any theoretical analysis by the user. Eight situations are considered.

(A) Time-Series Data.

The usual $s_n(x_1, x_2, \ldots, x_n) := \sum_{i=1}^{n}(x_i - \sum_{j=1}^{n} x_j/n)^2/(n-1)$ will be computed on a time-series of length n, in order to estimate marginal variance. [The statistic s_n^0 is a reasonable estimator because, e.g., $E\{s_n^0\} \to V\{X_1\}$ for any stationary time-series such that $\sum_{i=1}^{\infty} C\{X_1, X_i\}$ converges.] To describe the sampling distribution of s_n^0, we then construct the replicate histogram from the same available data (i.e., the time-series of length n). Figure A.1 shows the resulting [smoothed] replicate histogram based on a realization of n=200 observations $\{X_i\}$, using $l(n)=40$. The replicate histogram clearly suggests a highly non-normal skewed-left sampling distribution. This procedure was carried out on another time-series $\{\tilde{X}_i\}$, using the same function $s_n(\cdot)$, sample size n, and subseries length $l(n)$ as above; the resulting replicate histogram is shown in Figure A.$\tilde{1}$. This replicate histogram does seem compatible with a normal sampling distribution. For larger sample size (n=1000, $l(n)=88$), there is an analogous (but even more dramatic) contrast between the replicate histogram of $s_n(\cdot)$ computed from $\{X_i\}$ (Figure A.2) vs. the replicate histogram of $s_n(\cdot)$ computed from $\{\tilde{X}_i\}$ (Figure A.$\tilde{2}$).

The diagnostic messages being sent by these four replicate histograms are correct: The underlying mechanism generating the data was actually an AR(1) process $Z_i = \beta Z_{i-1} + \xi_i$ with $\beta=.5$ and $\xi_i \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$; $\{X_i\}$ and $\{\tilde{X}_i\}$ were then threshold variables with $X_i = I\{Z_i>0\}$ and $\tilde{X}_i = I\{Z_i>1\}$. Asymptotic theory confirms that, for the $\{X_i\}$ data, $n(s_n^0 - \sigma^2) \xrightarrow{\mathcal{D}} T$ where T has density

$$f_T(y) := \frac{8^{1/2}\exp(\frac{4y-1}{2\tau^2})}{\pi^{1/2}(1-4y)^{1/2}\tau}, \quad -\infty < y < \tfrac{1}{4},$$

[see Carlstein (1988), Example 4] shown in Figure A.3; while for the $\{\tilde{X}_i\}$ data, $n^{1/2}(s_n^0 - \tilde{\sigma}^2) \xrightarrow{\mathcal{D}} \mathcal{N}(0,\omega^2)$ [see Ibragimov and Linnik (1971), Theorem 18.5.4] shown in Figure A.$\tilde{3}$.

As expected, the replicate histograms give a more accurate approximation to the theoretical asymptotic distributions when the sample size is larger (n=1000: Figures A.2 and A.$\tilde{2}$). However, even for sample time-series of moderate length (n=200: Figures A.1 and A.$\tilde{1}$), the user still gets a clear warning that normality is doubtful in the $\{X_i\}$ case, while the $\{\tilde{X}_i\}$ picture is compatible with normality. Thus, without any knowledge of the proper standardizations $(n, n^{1/2})$, centerings $(\sigma^2, \tilde{\sigma}^2)$, and underlying dependence mechanism, the replicate histograms provide useful diagnostic information about the sampling distributions. In particular, the replicate histogram gives a good indication of

whether the sampling distribution is normal or not, and if not, how it departs from normality.

(B) Spatial Data.

Here we will compute the magnitude of the sample mean, $S_{D_n}(x_1, x_2, \ldots, x_{/D_n/}) := |\sum_{i=1}^{/D_n/} x_i / |D_n||$, on data in a triangular random field [specifically, A is the triangle with vertices (0,0),(1,0),(0,1)]. To describe the sampling distribution of $s(D_n)$, we then construct the replicate histogram from the same available data (i.e., the triangular random field D_n). Figure B.1 shows the resulting replicate histogram based on a realization of random field $\{X_i\}$ with n=50, l(n)=20. The replicate histogram clearly suggests a non-normal skewed-right sampling distribution. This procedure was carried out on another random field $\{\tilde{X}_i\}$, using the same function $S_{D_n}(\cdot)$, the same shape A, the same sample-size factors n and l(n) as above; the resulting replicate histogram is shown in Figure B.$\tilde{1}$. This replicate histogram does seem compatible with a normal sampling distribution. For larger sample-size factors (n=80, l(n)=25), there is an analogous contrast between the replicate histogram of $S_{D_n}(\cdot)$ computed from $\{X_i\}$ (Figure B.2) vs. the replicate histogram of $S_{D_n}(\cdot)$ computed from $\{\tilde{X}_i\}$ (Figure B.$\tilde{2}$). Note that the replicate histograms computed from $\{X_i\}$ data (Figures B.1 and B.2) are skewed but are not dramatically peaked near the lower end of their support [compare this to the sharp peak at the upper end of Figures A.1, A.2, A.3].

The diagnostic messages being sent by these four replicate histograms are again correct: The underlying mechanism generating the data was actually an isotropic symmetric nearest-neighbor binary response model with bonding strength β=.1; the binary response values are $\{+1, -1\}$ for the $\{X_i\}$ random field, and $\{+3, -1\}$ for the $\{\tilde{X}_i\}$ random field. Asymptotic theory [e.g., Ellis (1985), Theorems V.7.2 and V.7.7] confirms that, for the $\{X_i\}$ data, $|D_n|^{1/2} s(D_n)$ converges in distribution to a half-normal (shown in Figure B.3); while for the $\{\tilde{X}_i\}$ data, $|D_n|^{1/2}(s(D_n) - 1)$ is asymptotically normal. For moderately sized spatial data-sets, the replicate histogram correctly warns the user of departures from normality (Figures B.1 and B.2); in particular, note that Figure B.3 is skewed right but is not sharply peaked. Thus, without any knowledge of the centerings (0,1), scaling ($|D_n|^{1/2}$), and underlying dependence mechanism, the replicate histogram provides useful diagnostic information about the sampling distributions.

FIGURE A.1: Replicate Histogram

$\{X_i\}$ data (time-series)

n=200, ℓ(n)=40

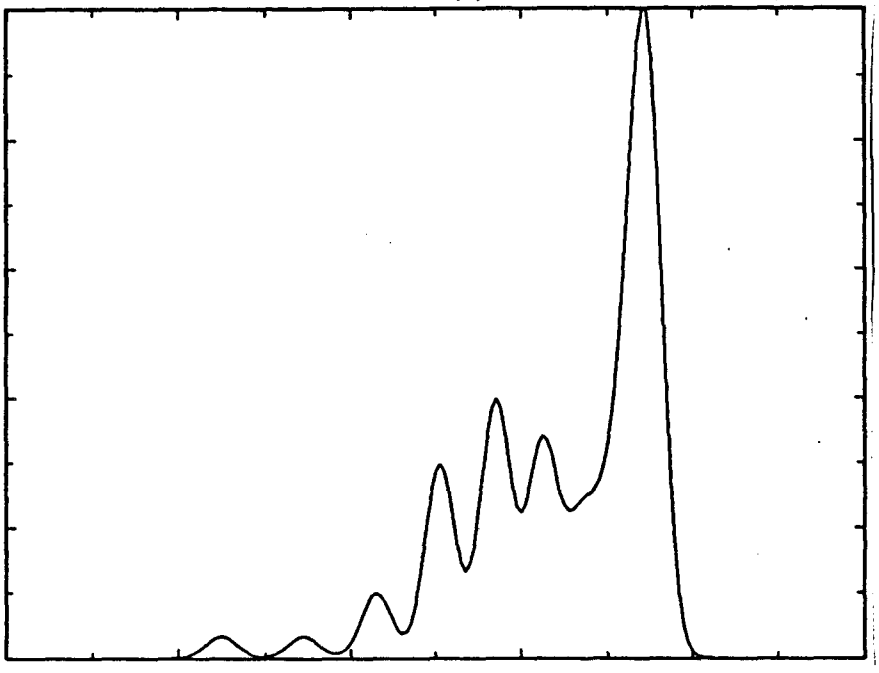


FIGURE A.1̃: Replicate Histogram

$\{\tilde{X}_i\}$ data (time-series)

n=200, ℓ(n)=40

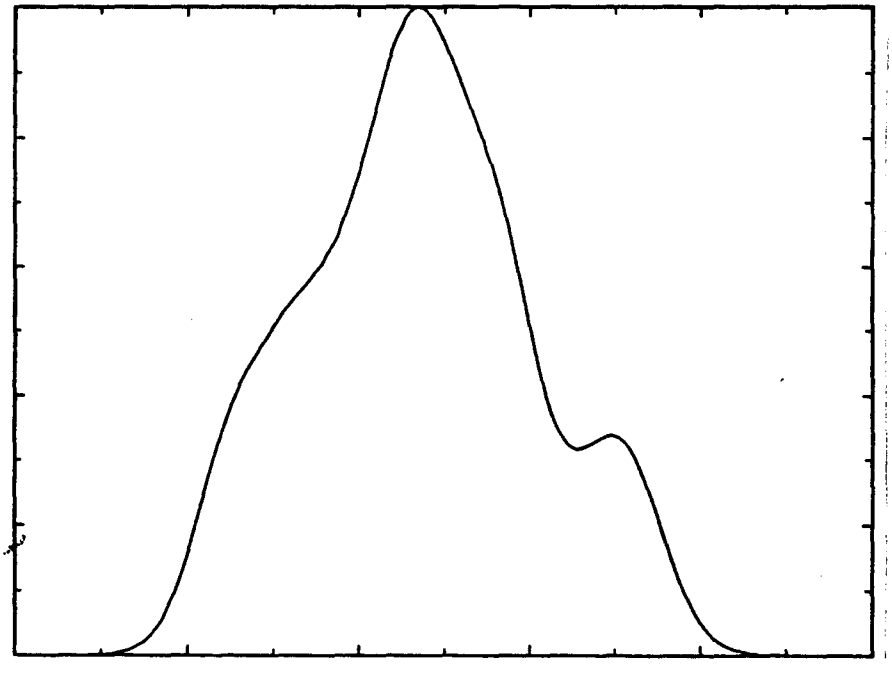


FIGURE A.2: Replicate Histogram

$\{X_i\}$ data (time-series)

n=1000, ℓ(n)=88

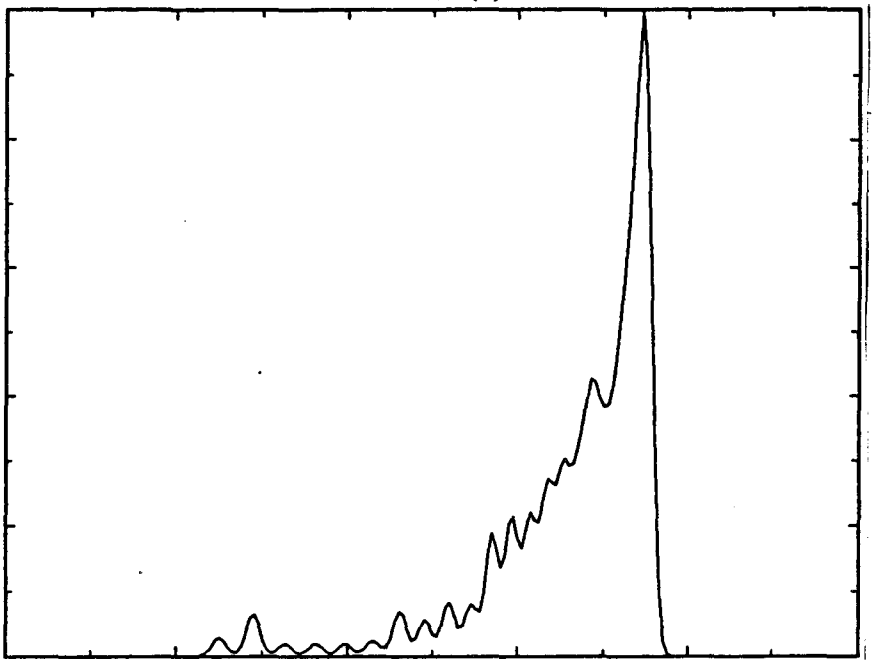


FIGURE A.$\tilde{2}$: Replicate Histogram

$\{\tilde{X}_i\}$ data (time-series)

n=1000, ℓ(n)=88

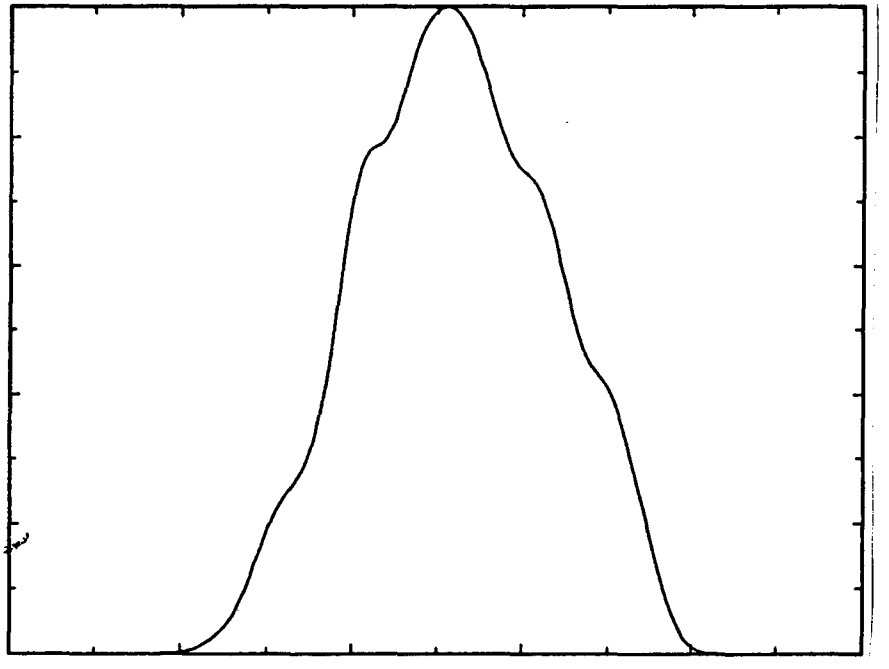


FIGURE A.3: $f_T(\cdot)$

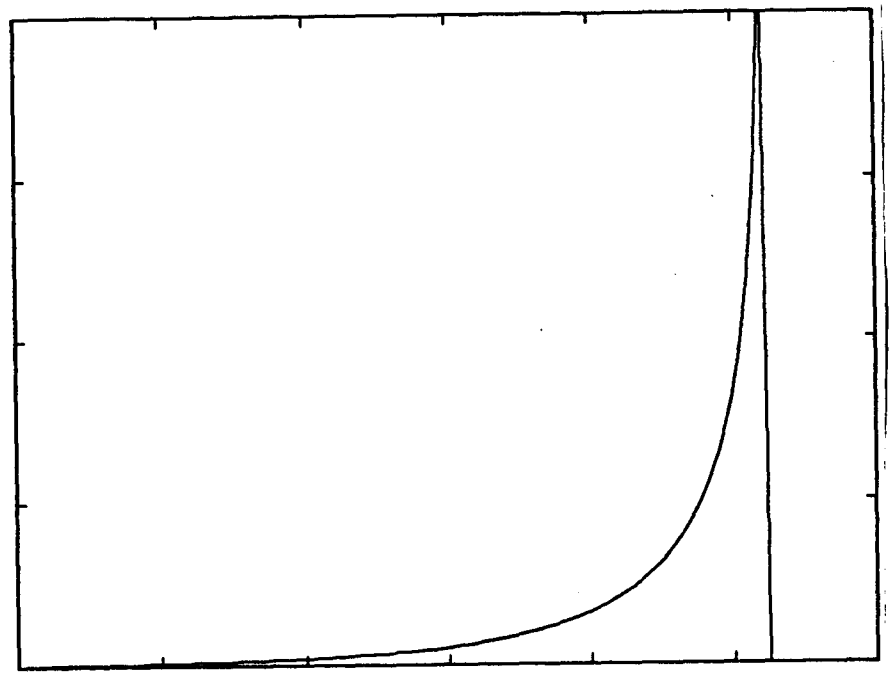


FIGURE A.3̃: Normal

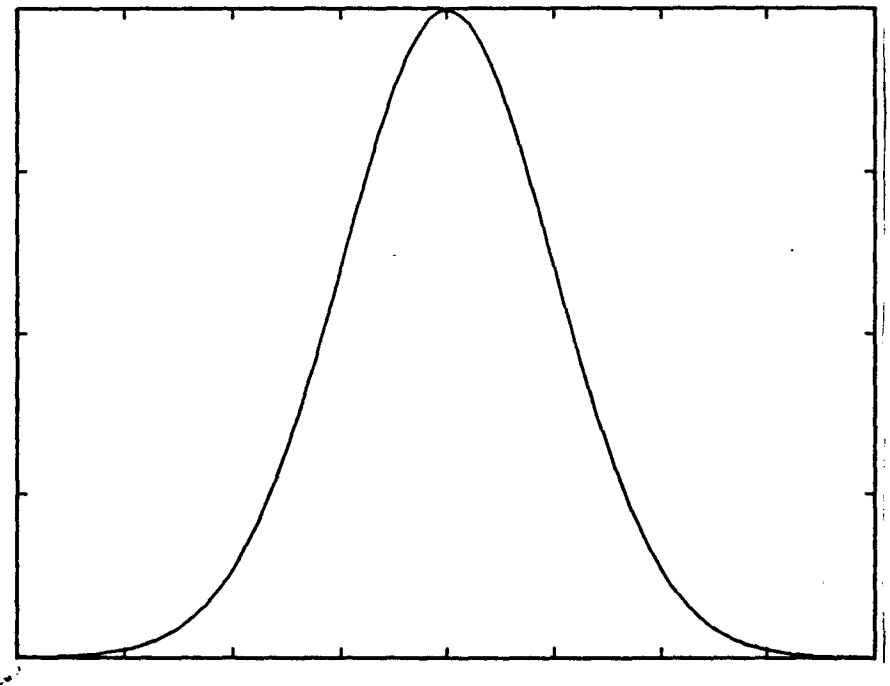


FIGURE B.1: Replicate Histogram

$\{X_i\}$ data (random field)

n=50, ℓ(n)=20, k_n=496

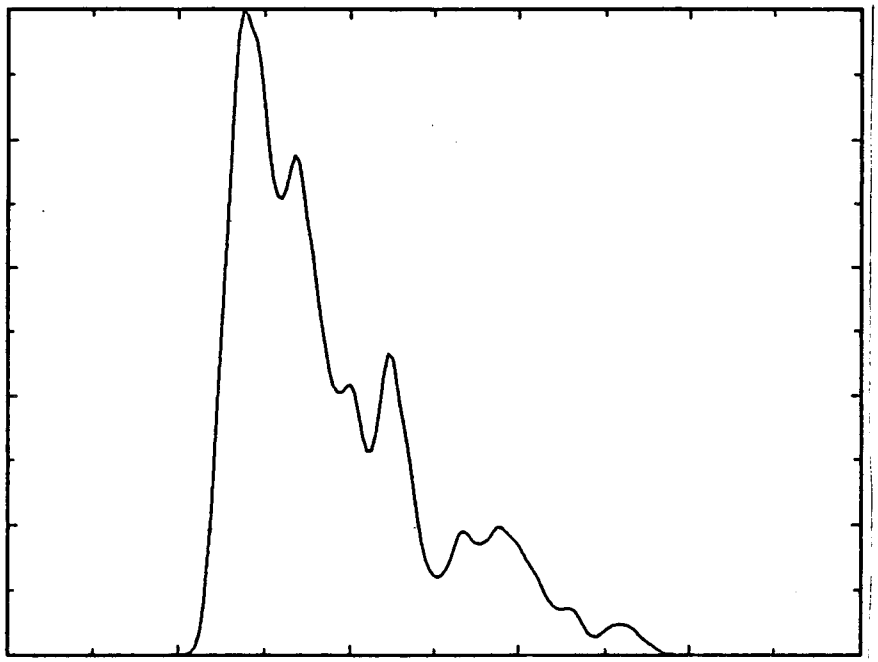


FIGURE B.$\tilde{1}$: Replicate Histogram

$\{\tilde{X}_i\}$ data (random field)

n=50, ℓ(n)=20, k_n=496

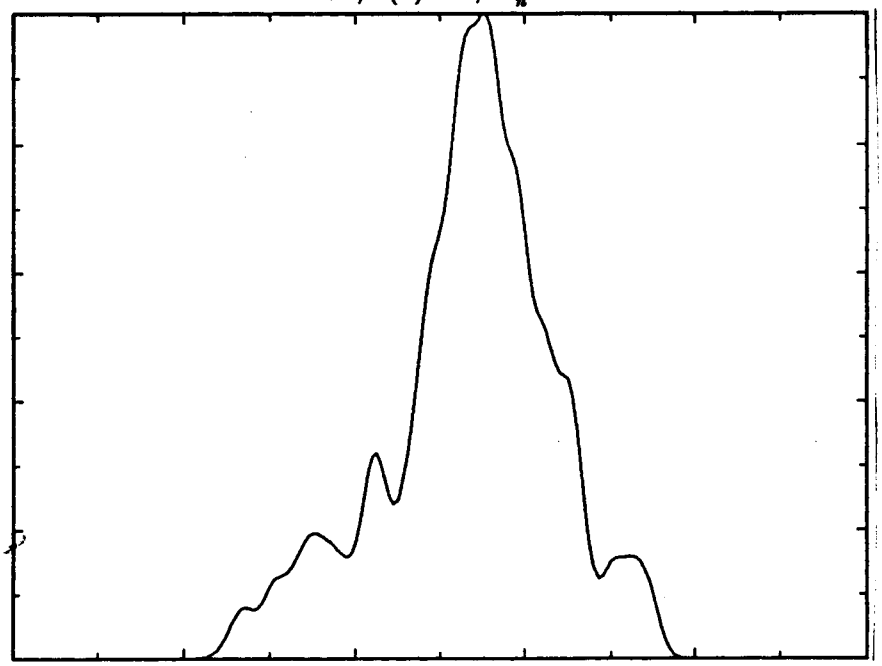


FIGURE B.2: Replicate Histogram

$\{X_i\}$ data (random field)

n=80, ℓ(n)=25, k_n=1596

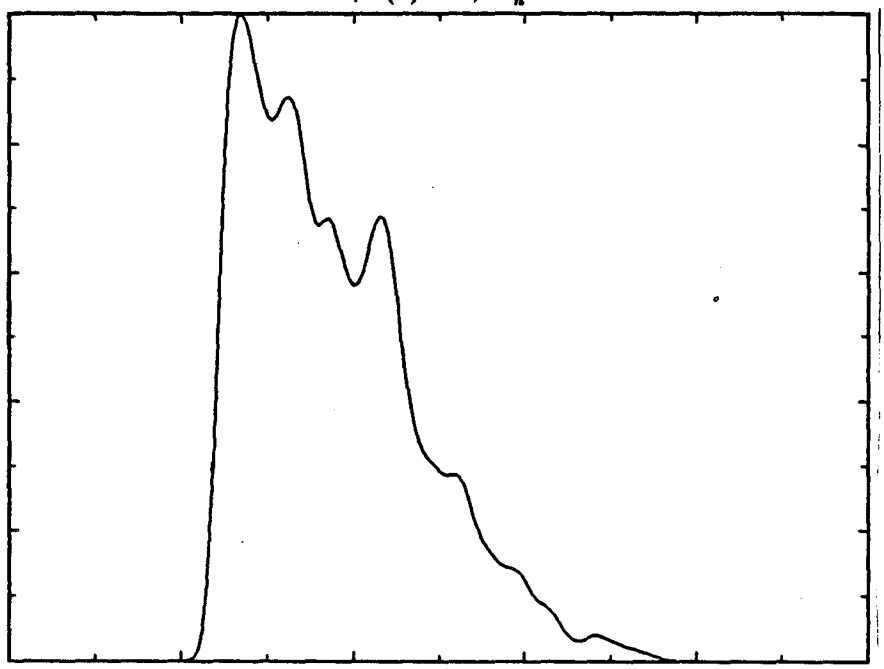


FIGURE B.$\tilde{2}$: Replicate Histogram

$\{\tilde{X}_i\}$ data (random field)

n=80, ℓ(n)=25, k_n=1596

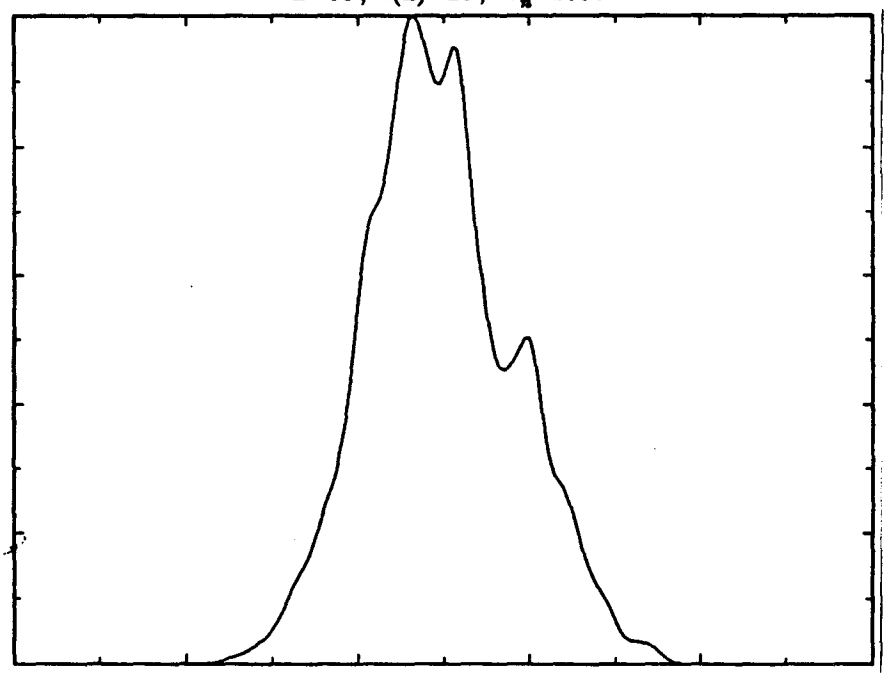
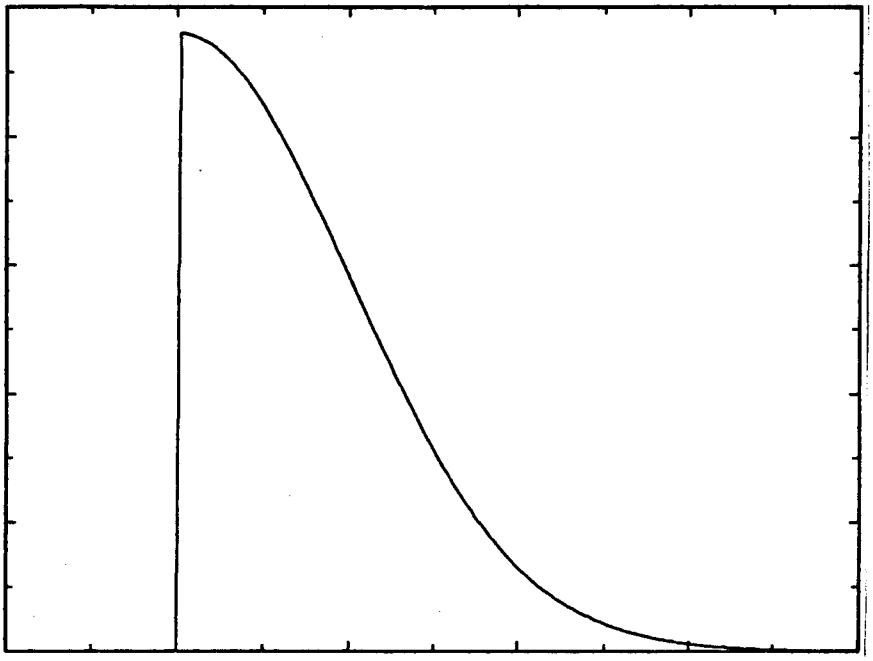


FIGURE B.3: Half-Normal



APPENDIX

Proof of Theorem 1:

For each $k \in \mathbf{N}$ and $l \in \mathbf{N}$ define $F_l(y) := \mathbf{P}\{t_l^0 \le y\}$, $\tilde{F}_{k,l}(y) := \sum_{i=0}^{k^2-1} \frac{\mathbf{I}\{t_l^i \le y\}}{k^2}$, and $\Delta_k := \max_{\{l:\, l(k^2) \le l \le l((k+1)^2)\}} |\tilde{F}_{k,l}(y) - F_l(y)|$. Note that for each $n \in \mathbf{N}$, $\exists\ k(n) \in \mathbf{N}$ such that $k^2(n) \le n < (k(n)+1)^2$. Now,

$$|\hat{F}_n(y) - F(y)| \le |\hat{F}_n(y) - \tilde{F}_{k(n),l(n)}(y)| + |\tilde{F}_{k(n),l(n)}(y) - F_{l(n)}(y)| + |F_{l(n)}(y) - F(y)|. \qquad (\dagger)$$

We have $|F_{l(n)}(y) - F(y)| \to 0$ as $n \to \infty$ by assumption. We will use the following Lemma to show that the second term on the r.h.s. of $(\dagger)$ tends to 0.

Lemma: $\Delta_k \overset{a.s.}{\to} 0$ as $k \to \infty$.

Proof:

$\mathbf{P}\{\Delta_k > \tau\} \le \sum_{\{l:\, l(k^2) \le l \le l((k+1)^2)\}} \mathbf{P}\{|\tilde{F}_{k,l}(y) - F_l(y)| > \tau\}$, and $\mathbf{P}\{|\tilde{F}_{k,l}(y) - F_l(y)| > \tau\} \le \mathbf{V}\{\tilde{F}_{k,l}(y)\}/\tau^2$. Writing $d(i,j) := |i-j|$, we obtain:

$$\mathbf{V}\{\tilde{F}_{k,l}(y)\} =$$

$$\frac{1}{k^4}\Big[\sum\sum_{\{(i,j):\, d(i,j) < 2l\}} \mathbf{C}\{\mathbf{I}\{t_l^i \le y\}, \mathbf{I}\{t_l^j \le y\}\} + \sum\sum_{\{(i,j):\, d(i,j) \ge 2l\}} \mathbf{C}\{\mathbf{I}\{t_l^i \le y\}, \mathbf{I}\{t_l^j \le y\}\}\Big]$$

$$\le [k^2 4l + k^4 \alpha(l)]/k^4$$

by letting each summand in the first sum be bounded in absolute value by 1 and by noting that there are at most $k^2 4l$ terms in the first sum ($0 \le i, j \le k^2 - 1$), and by observing that each of the at most k^4 summands in the second sum is bounded by $\alpha(l)$ in absolute value. For k sufficiently large, $(c/2)k^{2\gamma} < l(k^2) \le l \le l((k+1)^2) \le l(4k^2) \le 2ck^{2\gamma}$ by definition, and $\alpha(l) \le \tilde{C} l^{-\epsilon}$ by assumption; hence

$$[k^2 4l + k^4 \alpha(l)]/k^4 \le C[k^{2\gamma-2} + k^{-2\gamma\epsilon}].$$

Combining this bound with the inequalities from the beginning of the proof, we find

$$\sum_{k=1}^{\infty} \mathbf{P}\{\Delta_k > \tau\} \le \sum_{k=1}^{\infty} \sum_{\{l:\, l(k^2) \le l \le l((k+1)^2)\}} (C/\tau^2)[k^{2\gamma-2} + k^{-2\gamma\epsilon}] \le (C/\tau^2)(c+2) \sum_{k=1}^{\infty} [k^{2\gamma-2} + k^{-2\gamma\epsilon}] < \infty.$$

The second inequality follows from the fact that $l((k+1)^2) - l(k^2) \le c[(k+1)^{2\gamma} - k^{2\gamma}] + 1 \le c+1$ (because $\gamma < 1/2$). Finiteness of the last sum follows from $2 - 2\gamma > 1$ and $2\gamma\epsilon > 1$. Since $\tau > 0$ is arbitrary, we have $\Delta_k \overset{a.s.}{\to} 0$. □

Now, $|\tilde{F}_{k(n),l(n)}(y) - F_{l(n)}(y)| \le \max_{\{l:\, l(k^2(n)) \le l \le l((k(n)+1)^2)\}} |\tilde{F}_{k(n),l}(y) - F_l(y)| = \Delta_{k(n)} \overset{a.s.}{\to} 0$, using the Lemma and the fact that $k(n) \to \infty$. From $(\dagger)$ it only remains to show that $|\hat{F}_n(y) - \tilde{F}_{k(n),l(n)}(y)| \overset{a.s.}{\to} 0$. Write

$$|\hat{F}_n(y)-\tilde{F}_{k(n),l(n)}(y)| =$$

$$\Big|\sum_{i=0}^{k^2(n)-1}\frac{I\{t^i_{l(n)}\le y\}}{n-l(n)+1}-\sum_{i=0}^{k^2(n)-1}\frac{I\{t^i_{l(n)}\le y\}}{k^2(n)}+\sum_{i=k^2(n)}^{n-1}\frac{I\{t^i_{l(n)}\le y\}}{n-l(n)+1}-\sum_{i=n-l(n)+1}^{n-1}\frac{I\{t^i_{l(n)}\le y\}}{n-l(n)+1}\Big|$$

$$\le \sum_{i=0}^{k^2(n)-1}\Big|\frac{1}{n-l(n)+1}-\frac{1}{k^2(n)}\Big|+\sum_{i=k^2(n)}^{n-1}\frac{1}{n-l(n)+1}+\sum_{i=n-l(n)+1}^{n-1}\frac{1}{n-l(n)+1}$$

$$\le (|k^2(n)-n|+|l(n)-1|+n-k^2(n)+l(n)-1)/(n-l(n)+1) \le 2(n-k^2(n)+l(n)-1)/(n-l(n))$$

$$\le 2(2k(n)+l(n))/(n-l(n)) \to 0. \quad .$$

The fourth inequality holds because $n-k^2(n) \le 2k(n)$. Convergence to zero is a consequence of $k^2(n) \le n$ and $l(n)/n \to 0$. Uniform convergence follows in the usual way. □

Proof of Theorem 2:

Let $\lambda(A)$ denote the area of A and let $\|\partial A\|$ denote the length of the boundary of A. $\lambda(A)>0$ and $\|\partial A\|<\infty$ implies that $\exists\ \delta>0$ such that a $\delta\times\delta$ square is completely contained in A and hence the corresponding $\delta n\times\delta n$ square is completely contained in A_n. This $\delta n\times\delta n$ square contains a square set of $\lfloor\delta n\rfloor^2$ lattice points, from which we form the $(\lfloor\delta n\rfloor-l(n)+1)^2$ overlapping subsquares, each $l(n)\times l(n)$, and the corresponding $D^i_{l(n)}$, $i=1,2,\ldots,(\lfloor\delta n\rfloor-l(n)+1)^2$. Note that $(\lfloor\delta n\rfloor-l(n)+1)^2 \le k_n$ and hence that $k_n \ge Cn^2$, $C>0$.

Using the same logic as in the proof of the Lemma, we see that $V\{\hat{F}_n(y)\} \le \frac{1}{k_n^2}[k_n(4l(n))^2+k_n^2\alpha_{l^2(n)}(l(n))]$. The r.h.s. is $O(n^{-\omega})$ for some $\omega>1$, by assumption. This implies that $\sum_{n=1}^{\infty}P\{|\hat{F}_n(y)-E\{\hat{F}_n(y)\}|>\tau\}<\infty$ and hence that $\hat{F}_n(y)\overset{a.s.}{\to}F(y)$. □

REFERENCES

Athreya, K. (1987). Bootstrap of the mean in the infinite variance case. In *Proceedings of the First World Congress of the Bernoulli Society* (Y. Prohorov and V. Sazonov, ed.s), 2, 95-98. VNU Science Press, The Netherlands.

Basawa, I., Mallik, A., McCormick, W., and Taylor, R., (1989). Bootstrapping explosive autoregressive processes. *Annals of Statistics*, 17, 1479-1486.

Bickel, P., and Freedman, D. (1981). Some asymptotic theory for the bootstrap. *Annals of Statistics*, 9, 1196-1217.

Bose, A. (1988). Edgeworth correction by bootstrap in autoregressions. *Annals of Statistics*, 16, 1709-1722.

Bradley, R. (1991). Some examples of mixing random fields. *Technical Report* #342, Center for Stochastic Processes, Department of Statistics, University of North Carolina, Chapel Hill.

Bretagnolle, J. (1983). Lois limites du bootstrap de certaines fonctionnelles. *Annales de l'Institut Henri Poincaré*, 19, 281-296.

Carlstein, E. (1988). Degenerate U-statistics based on non-independent observations. *Calcutta Statistical Association Bulletin*, 37, 55-65.

Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans.* SIAM, Philadelphia.

Ellis, R. (1985). *Entropy, Large Deviations, and Statistical Mechanics.* Springer-Verlag, N.Y.

Freedman, D. (1984). On bootstrapping two-stage least-squares estimates in stationary linear models. *Annals of Statistics*, 12, 827-842.

Gastwirth, J., and Rubin, H. (1975). The asymptotic distribution theory of the empiric c.d.f. for mixing stochastic processes. *Annals of Statistics*, 3, 809-824.

Ibragimov, I., and Linnik, Y. (1971). *Independent and Stationary Sequences of Random Variables.* Wolters-Noordhoff, The Netherlands.

Kendall, M., and Stuart, A. (1977). *The Advanced Theory of Statistics, Vol. I.* Griffin, London.

Künsch, H. (1989). The jackknife and the bootstrap for general stationary observations. *Annals of Statistics*, 17, 1217-1241.

Lele, S. (1988). Non-parametric bootstrap for spatial processes. *Technical Report* #671, Department of Biostatistics, Johns Hopkins University, Baltimore.

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

Rajarshi, M. (1990). Bootstrap in Markov sequences based on estimates of transition density. *Annals of the Institute of Statistical Mathematics*, 42, 253-268.

Rosenblatt, M. (1956). A central limit theorem and a strong mixing condition. *Proceedings of the National Academy of Sciences*, 42, 43-47.

Serfling, R. (1980). *Approximation Theorems of Mathematical Statistics.* Wiley, N.Y.

Singh, K. (1981). On the asymptotic accuracy of Efron's bootstrap. *Annals of Statistics*, 9, 1187-1195.

Swanepoel, J. (1986). A note on proving that the (modified) bootstrap works. *Communications in Statistics, Theory and Methods*, 15, 3193-3203.

Wu, C. (1990). On the asymptotic properties of the jackknife histogram. *Annals of Statistics*, 18, 1438-1452.