

Consider the usual scenario for analyzing data: A scientist observes some data,  $\mathbf{X}_n := (X_1, X_2, \dots, X_n)$  controlled by unknown parameters  $(\theta, \nu)$ . The “target parameter”  $\theta$  is estimated by a statistic  $\hat{\theta} := s_n(\mathbf{X}_n)$ , while  $\nu$  is a “nuisance parameter”.

Having selected the estimator  $\hat{\theta}$ , the scientist seeks to establish the accuracy of  $\hat{\theta}$  by computing a confidence interval for  $\theta$ . In order to do this, knowledge of the sampling distribution of  $s_n(\mathbf{X}_n)$  is necessary, or as a practical approximation, the asymptotic distribution ( $F$ ) of  $t_n := a_n(s_n - b_n)$ , a standardized transform of the statistic. The accuracy of the confidence interval thus relies crucially upon the chosen asymptotic distribution  $F$  (for example, normal,  $\chi^2$ ). The “replicate histogram” is a simple diagnostic tool for assessing the appropriateness of  $F$  using only the observed data.

To see the need for such a diagnostic, consider the following obstacles confronting the scientist:

(i) The statistic  $s_n$  may be complicated (e.g., a robust measure of location or an adaptively defined statistic), so that a theoretical derivation of  $F$  is difficult.

(ii) The observations may be serially, or spatially dependent, so their joint distribution must be accounted for in deriving  $F$ . This may require knowledge of the underlying dependence mechanism, and estimation of the nuisance parameter  $\nu$ .

(iii) The correct standardizing constants  $(a_n, b_n)$  may involve the unknown  $(\theta, \nu)$  and have a role in determining the basic characteristics of  $F$  (e.g., symmetric vs. skewed; normal vs.  $\chi^2$ ).

Let’s define the replicate histogram and discuss how it avoids these three obstacles. Let  $X_l^i := (X_{i+1}, X_{i+2}, \dots, X_{i+l})$  denote “subseries” of consecutive observations so that the observed data is  $X_n^0$ , and the collection of all subseries with length  $l$  is  $\{X_l^i : 0 \leq i \leq n - l\}$ . The associated “replicates” are simply  $s_l^i := s_l(X_l^i)$ , and the replicate histogram corresponds to the empirical distribution of the replicates.

We see that: i) the replicate histogram is directly computable from the data so no theoretical analysis is necessary, ii) by employing subseries replicates of the statistic  $s_n$  the correct dependence structure is automatically retained, without assumptions about the underlying dependence mechanism, iii) since the replicates do not depend on  $a_n$  and  $b_n$  there is no need to derive (nor to guess) these sequences.

As an example consider the usual  $s_n = \sum_{i=1}^n (x_i - \bar{x})^2 / (n - 1)$  computed from a time series to estimate the marginal variance of the series. Figure 1 shows a smooth of the replicates based on a time series,  $\{X_i\}$ , of length  $n = 200$  using subseries length  $l = 40$ . The graph clearly suggests a highly nonnormal skewed-left sampling distribution. The procedure was carried out on another time series  $\{\tilde{X}_i\}$ , using the same statistic  $s_n$ , sample size  $n$ , and subseries length  $l$ . The result is in Figure 2. This picture does seem compatible with a normal sampling distribution.

The underlying true dependence mechanism was the autoregression  $Z_i = \beta Z_{i-1} + \epsilon_i$  with  $\beta = .5$ ,  $\{\epsilon_i\}$  independent normal errors;  $\{X_i\}$  and  $\{\tilde{X}_i\}$  were then the threshold variables  $X_i = I\{Z_i > 0\}$  and  $\tilde{X}_i = I\{Z_i > 1\}$ . Asymptotic theory shows that, for the  $\{X_i\}$  data,  $n(s_n - \sigma^2) \xrightarrow{D} T$ , where  $T$  has the density shown in Figure 3; while for the  $\{\tilde{X}_i\}$  data,  $n^{1/2}(s_n - \tilde{\sigma}^2) \xrightarrow{D} Z$  (normal) shown in Figure 4. Thus, without any knowledge of proper standardization  $(n, n^{1/2})$ , centerings  $(\sigma^2, \tilde{\sigma}^2)$ , or underlying dependence structure, the (smoothed) replicate histogram provides useful diagnostic information about the sampling distributions. In particular, it gives a good indication of whether the normal approximation is reasonable or not. The method can be applied to spatial data, by computing the replicates from “subshapes” of the original data grid.

It turns out that a transformed (by a shift and scale) replicate histogram is consistent for the true  $F$ , whatever it is, for a broad choice of subseries lengths  $l$ , under a general assumption of weak dependence. Naturally, any method which achieves such seemingly far-reaching applicability must do so at some cost. The cost, for the replicate histogram, is that it returns only diagnostic information about the shape of  $F$ , but does not return numerical estimates of the percentiles of  $F$ . If the sequences  $a_n$  and  $b_n$  are known and  $F$  is the normal distribution, the replicates can actually be used to obtain second order correct inferences, i.e., more accurate than obtained from the limiting normal distribution. How generally (for which statistics, limiting distributions, dependence structures) this “higher order accuracy” holds is of great practical importance, and is an area of current research.