Journal of Statistical Research *ISSN 0256 - 422 X*
2009 Vol. 43, No. 1, pp. 109-115
Bangladesh

TWO NEW ESTIMATORS OF DISTRIBUTION FUNCTIONS

A. K. Md. Ehsanes Saleh

School of Mathematics and Statistics
Carleton University, Ottawa, Ontario, K1S 5B6, Canada
Email: esaleh@math.carleton.ca

Patrick J. Farrell

School of Mathematics and Statistics
Carleton University, Ottawa, Ontario, K1S 5B6, Canada
Email: pfarrell@math.carleton.ca

SUMMARY

This paper considers the estimation of a distribution function $F_X(x)$ based on a random sample $X_1, X_2, \ldots, X_n$ when the sample is suspected to come from a close-by distribution $F_0(x)$. Two new estimators, namely $F_n^{PT}(x)$ and $F_n^S(x)$ are defined and compared with the "empirical distribution function", $F_n(x)$, under local departure; that is $F_X(x) = F_0(x) + n^{-1/2}\delta$, where $\max_x |F_X(x) - F_0(x)| \leq n^{-1/2}\delta$. In this case, we show that $F_n^S(x)$ is superior to $F_n^{PT}(x)$ in the neighbourhood of $F_0(x)$.

Keywords and phrases: Empirical Distribution Function; Local Alternatives; Mean Square Error; Preliminary Test Estimator; Shrinkage Estimator.

1 Introduction

Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed random variables with distribution function $F_X(x)$. It is well known that the *empirical distribution function* of the sample $X_1, X_2, \ldots, X_n$ given by

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,x)}(X_i), \tag{1.1}$$

is a popular estimator of $F_X(x)$. Clearly, for every fixed x, $F_n(x)$ is the relative frequency of successes in a sequence of Bernoulli trials with

$$E[F_n(x)] = F_X(x) \text{ and } \mathrm{Var}[F_n(x)] = \frac{1}{n}F_X(x)[1 - F_X(x)] \leq \frac{1}{4n}. \tag{1.2}$$

That is, $F_n(x)$ is an unbiased estimator with variance given by (1.2). As a consequence, by the classical strong law of large numbers

$$F_n(x) \xrightarrow{a.s.} F_X(x) \text{ for } x \text{ fixed.} \tag{1.3}$$

Hence, $F_n(x)$ is an unbiased and strongly consistent estimator of $F_X(x)$ for each fixed x. Further, by the theorems given by Glivenko (1933) and Cantelli (1944)

$$\sup_{-\infty<x<\infty}|F_n(x) - F_X(x)| \xrightarrow{a.s.} 0, \tag{1.4}$$

which uniquely follows as a consequence of (1.3). This suggests that by sampling ad infinitum, $F_X(x)$ can be uniquely estimated by $F_n(x)$ with probability one.

Now, consider the process

$$V_n(x) = \sqrt{n}[F_n(x) - F_X(x)]. \tag{1.5}$$

Point-wise behaviour shows that

$$V_n(x) \overset{D}{\simeq} N\{0, F_X(x)[1 - F_X(x)]\}. \tag{1.6}$$

As for the rate of convergence of the Glivenko-Cantelli Theorem, one has

$$P\{\sup_{-\infty<x<\infty}|V_n(x)| \le y\} = \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2y^2}, y > 0, \tag{1.7}$$

and zero elsewhere, and

$$P\{\sup_{-\infty<x<\infty}V_n(x) \le y\} = \sum_{k=-\infty}^{\infty} (-1)^k e^{-2y^2}, y > 0, \tag{1.8}$$

and zero elsewhere.

Based on the point-wise estimator, $F_n(x)$, one may test the null hypothesis $H_0 : F_X(x) = F_0(x)$ versus $H_A : F_X(x) \neq F_0(x)$ using the statistic

$$L_n = V_{n0}^2(x)/\{F_0(x)[1 - F_0(x)]\}, \tag{1.9}$$

where

$$V_{n0}(x) = \sqrt{n}[F_n(x) - F_0(x)]. \tag{1.10}$$

As $n \to \infty, L_n$ converges to a χ^2 distribution with one degree of freedom under H_0, while under the local alternatives of the form

$$K_n : F_X(x) = F_0(x) + n^{-1/2}\delta, \tag{1.11}$$

where

$$\max_x|F_n(x) - F_0(x)| \le n^{-1/2}\delta, \tag{1.12}$$

for fixed positive δ, L_n follows a non-central chi-square distribution with one degree of freedom and non-centrality parameter Δ^2 given by

$$\Delta^2 = n[F_X(x) - F_0(x)]^2/\{F_0(x)[1 - F_0(x)]\} \ge 4\delta^2. \tag{1.13}$$

2 Two New Estimators of $F_X(x)$

Suppose now that one suspects that the sample $X_1, X_2, \ldots, X_n$ comes from the cdf $F_0(x)$. How do we incorporate this uncertain non-sample information into the estimation of $F_X(x)$ so that the estimation is somewhat improved in the sense of mean-square-error? A simple approach is to consider the preliminary test estimator (PTE) introduced by Han and Bancroft (1968), and expanded by Saleh (2006), among others. We may define the PTE as follows:

$$F_n^{PT}(x) = F_n(x) - [F_n(x) - F_0(x)]I[L_n < \chi_1^2(\alpha)], \tag{2.1}$$

where $I(A)$ is the indicator function of the set A. In this case, one estimates $F_X(x)$ by $F_0(x)$ if the test L_n accepts H_0 at level of significance α; otherwise $F_n(x)$ is used. Note that equation (2.1) is a discontinuous function, leading to extreme choices for the estimators. To make a smooth transition of equation (2.1), we define the following shrinkage estimator of $F_X(x)$:

$$F_n^S(x) = F_n(x) - c\sqrt{F_0(x)[1 - F_0(x)]}\frac{F_n(x) - F_0(x)}{\sqrt{n}|F_n(x) - F_0(x)|}, \tag{2.2}$$

or equivalently

$$F_n^S(x) = F_n(x) - c|L_n^{1/2}|^{-1}[F_n(x) - F_0(x)], \tag{2.3}$$

where c is a non-negative shrinkage constant. Note the similarity and dissimilarity between $F_n^{PT}(x)$ and $F_n^S(x)$. As $L_n \to \infty$, both $F_n^{PT}(x)$ and $F_n^S(x)$ become equal to $F_n(x)$. Thus, for large L_n, the two estimators behave in the same way. However, as $L_n \to 0$, $F_n^{PT}(x) \to F_0(x)$, while $F_n^S(x)$ tends to $F_0(x)$. On the other hand, if $|L_n^{1/2}| = c$, then $F_n^S(x) \to F_0(x)$. Further, $F_n^S(x)$ provides interpolated values between $F_0(x)$ and $F_n(x)$ for $|L_n| > c$. Later, we shall see that an appropriate choice for c is $\sqrt{2/\pi}$.

3 Asymptotic Bias and Mean Square Error of the Estimators of the Distribution Function

In this section, we consider the asymptotic distributional bias (ADB) and asymptotic distributional mean square error (ADMSE) of the three estimators $F_n(x)$, $F_n^{PT}(x)$, and $F_n^S(x)$. Due to a result of Heilers and Willers (1988), which states that point-wise stochastic convergence is equivalent to uniform stochastic convergence on a compact set, the asymptotic properties of the estimators hold uniformly in the x's.

First note that the test statistic L_n is a consistent test. Hence, for fixed alternatives; that is, where δ is equal to some fixed non-zero value, we have

$$E\{L_n I[L_n < \chi_1^2(\alpha)]\} \to 0, \tag{3.1}$$

as $n \to \infty$. This suggests that $F_n^{PT}(x)$ and $F_n(x)$ are asymptotically uniformly ADMSE equivalent, while $F_n^S(x)$ and $F_n(x)$ are not, since

$$E\{n[F_n^S(x) - F_n(x)]^2\} \leq c^2/4. \tag{3.2}$$

We now consider the asymptotic bias and mean square error of the estimators under the local alternatives $K_n : F_X(x) = F_0(x) + n^{-1/2}\delta$. The bias expression for $F_n(x)$ is given by

$$b_1[F_n(x)] = E[F_n(x) - F_X(x)] = 0 \text{ for all } n \text{ and } x. \tag{3.3}$$

Similarly, we define the ADB of the preliminary test estimator as

$$b_2[F_n^{PT}(x)] = \lim_{n \to \infty} E\{\sqrt{n}[F_n^{PT}(x) - F_X(x)]\}. \tag{3.4}$$

Therefore, we have

$$b_2[F_n^{PT}(x)] = -\lim_{n \to \infty} E\{\sqrt{n}[F_n(x) - F_0(x)]I[L_n < \chi_1^2(\alpha)]\} = -\delta H_3[\chi_1^2(\alpha); \Delta^2], \tag{3.5}$$

where $H_\nu(\cdot; \Delta^2)$ is the cdf of a non-central chi square distribution with ν degrees of freedom and non- centrality parameter Δ^2 for all x. Finally,

$$b_3[F_n^S(x)] = -c\sqrt{F_0(x)[1 - F_0(x)]} \lim_{n \to \infty} E(Z_n/|Z_n|) \leq -(c/2) \lim_{n \to \infty} E(Z_n/|Z_n|), \tag{3.6}$$

where

$$Z_n = \frac{\sqrt{n}[F_n(x) - F_0(x)]}{\sqrt{F_0(x)[1 - F_0(x)]}}, \tag{3.7}$$

is asymptotically a standard normal random variable. Hence

$$b_3[F_n^S(x)] \leq -(c/2)[1 - 2\Phi(-\Delta)], \tag{3.8}$$

by Theorem 1 of Chapter 3 of Saleh (2006).

The expressions for mean square error may be obtained similarly. Specifically, for $F_n(x)$, the asymptotic distributional mean square error is given by

$$M_1[F_n(x)] = \lim_{n \to \infty} E\{\sqrt{n}[F_n(x) - F_X(x)]\}^2 = F_X(x)[1 - F_X(x)] \leq 1/4. \tag{3.9}$$

The ADMSE of $F_n^{PT}(x)$ is given by

$$M_2[F_n^{PT}(x)] = \lim_{n \to \infty} E\{n[F_n^{PT}(x) - F_X(x)]^2\}. \tag{3.10}$$

Thus

$$M_2[F_n^{PT}(x)] = F_0(x)[1 - F_0(x)]\{1 - H_3[\chi_1^2(\alpha); \Delta^2] + \Delta^2(2H_3[\chi_1^2(\alpha); \Delta^2] - H_5[\chi_1^2(\alpha); \Delta^2])\}, \tag{3.11}$$

which implies that

$$M_2[F_n^{PT}(x)] \leq (1/4)\{1 - H_3[\chi_1^2(\alpha); \Delta^2] + \Delta^2(2H_3[\chi_1^2(\alpha); \Delta^2] - H_5[\chi_1^2(\alpha); \Delta^2])\}. \tag{3.12}$$

Finally, the ADMSE of $F_n^S(x)$ is given by

$$M_3[F_n^S(x)] = \lim_{n \to \infty} E\{n[F_n^S(x) - F_X(x)]^2\}. \tag{3.13}$$

As a result, we have

$$M_3[F_n^S(x)] = F_0(x)[1 - F_0(x)](1 + c^2 - 2c\sqrt{2/\pi}e^{-\Delta^2/2}), \tag{3.14}$$

so that

$$M_3[F_n^S(x)] \leq (1/4)(1 + c^2 - 2c\sqrt{2/\pi}e^{-\Delta^2/2}). \tag{3.15}$$

Note that the minimum of $M_3[F_n^S(x)]$ occurs at $c^* = \sqrt{2/\pi}e^{-\Delta^2/2}$. Therefore, in order to make c^* independent of Δ^2, we choose c as $c_0 = \sqrt{2/\pi}$, which allows us to express $M_3[F_n^S(x)]$ as

$$M_3[F_n^S(x)] \leq (1/4)[1 - (2/\pi)(2e^{-\Delta^2/2} - 1)]. \tag{3.16}$$

In the next section, we present an analysis of the mean square errors of the estimators.

4 Analysis of the MSE of the Estimators

First we note that the MSE of $F_n(x)$ is constant, that $F_0(x)[1 - F_0(x)] \leq 1/4$, and that

$$\max_x\{M_2[F_n^{PT}(x)]\} = (1/4)\{1 - H_3[\chi_1^2(\alpha); \Delta^2] + \Delta^2(2H_3[\chi_1^2(\alpha); \Delta^2] - H_5[\chi_1^2(\alpha); \Delta^2])\}. \tag{4.1}$$

Hence, $F_n^{PT}(x)$ is better than $F_n(x)$ if

$$\Delta^2 \leq H_3[\chi_1^2(\alpha); \Delta^2]/(2H_3[\chi_1^2(\alpha); \Delta^2] - H_5[\chi_1^2(\alpha); \Delta^2]), \tag{4.2}$$

otherwise, $F_n(x)$ is better than $F_n^{PT}(x)$. The asymptotic relative efficiency of $F_n^{PT}(x)$ relative to $F_n(x)$ is

$$ARE[F_n^{PT}(x) : F_n(x)] = \{1 - H_3[\chi_1^2(\alpha); \Delta^2] + \Delta^2(2H_3[\chi_1^2(\alpha); \Delta^2] - H_5[\chi_1^2(\alpha); \Delta^2])\}^{-1}. \tag{4.3}$$

If we set $\text{ARE}[F_n^{PT}(x) : F_n(x)] = Ef(\alpha; \Delta^2)$ say, the optimum level of significance of the test may be obtained by solving the maximin problem:

$$\max_\alpha\{\min_{\Delta^2}[ARE(\alpha; \Delta^2)]\} = Ef(\alpha^*; \Delta^2_{\min}) = Ef_0, \tag{4.4}$$

where Ef_0 is prespecified. For some results, see Saleh (2006, Chapter 3). Thus, α^* is the optimum level of significance to be chosen in order to achieve a prespecified efficiency, Ef_0.

Similarly, we consider the ARE of $F_n^S(x)$ relative to $F_n(x)$, which is

$$ARE[F_n^S(x) : F_n(x)] = [1 - (2/\pi)(2e^{-\Delta^2/2} - 1)]^{-1}. \tag{4.5}$$

Under $H_0 : F_X(x) = F_0(x)$, we have that $\Delta^2 = 0$. Hence

$$ARE[F_n^S(x) : F_n(x)] = [1 - (2/\pi)]^{-1} = \pi/(\pi - 2) = 2.75. \tag{4.6}$$

On the other hand

$$ARE[F_n^{PT}(x) : F_n(x)] = \{1 - H_3[\chi_1^2(\alpha); 0]\}^{-1} \geq 1, \tag{4.7}$$

which depends on the size of α. As $\Delta^2 \to \infty$,

$$ARE[F_n^S(x) : F_n(x)] = [1 + (2/\pi)]^{-1} = \pi/(2 + \pi) = 0.61, \tag{4.8}$$

while ARE$[F_n^{PT}(x) : F_n(x)] \to 1$. These facts imply that $F_n^S(x)$ is superior to $F_n^{PT}(x)$ when $F_X(x)$ is close to $F_0(x)$. On the other hand, the minimum guaranteed efficiency of $F_n^S(x)$ relative to $F_n(x)$ is 0.61, and that of $F_n^{PT}(x)$ is relative, depending upon α. In general, ARE$[F_n^S(x) : F_n(x)]$ decreases from $\pi/(\pi - 2)$ at $\Delta^2 = 0$ to a value of one at $\Delta^2 = \ln(4) = 1.38$, and then drops to the minimum value $\pi/(2 + \pi) = 0.61$ as $\Delta^2 \to \infty$. By contrast, ARE$[F_n^{PT}(x) : F_n(x)]$ has maximum value $\{1 - H_3[\chi_1^2(\alpha); 0]\}^{-1}$ at $\Delta^2 = 0$, dropping to a value of one at $\Delta^2 = 1$. It continues to drop, reaching the minimum value of ADMSE, and then increases towards a value of one as $\Delta^2 \to \infty$. From this, one may note that the range of Δ^2 for which $F_n^S(x)$ is better than $F_n(x)$ is wider than the range produced by $F_n^{PT}(x)$. Further, $F_n^S(x)$ is independent of α, while the minimum of the ARE$[F_n^{PT}(x) : F_n(x)]$ depends on the value of α. In general, $F_n^S(x)$ does not dominate $F_n^{PT}(x)$ uniformly except in the range $(0, \ln(4))$. Thus, considering the high efficiency of $F_n^S(x)$, and also the fact that it is independent of the size α, of the test, the estimate $F_n^S(x)$ is preferable over $F_n^{PT}(x)$ if $F_0(x)$ is close to $F_X(x)$.

Acknowledgements

This research was supported by the Natural Sciences and Engineering Research Council of Canada.

References

Bancroft, T.A. (1944). On biases in estimation due to the use of preliminary tests of significance. *Ann. Math. Statist.* **15**, 190-204.

Cantelli, F.P. (1944). Su due applicazioni di un teorema G. Boole, R.C. Lincei. **26**, p. 39.

Csorgo, M. and Revesz, P. (1981). *Strong Approximations in Probability and Statistics.* Academic Press, New York.

Glivenko, V.I. (1933). Sulla determinazione empirica di una legge di probabilita. *Gion. del l'Inst. Itali. deqli Attuari.* **4**, 92-99.

Han, C.P. and Bancroft, T.A. (1968). On pooling means when variance is unknown. *Jour. Amer. Statist. Assoc.* **63**, 1333-1342.

Heilers, S. and Willers, R. (1988). Asymptotic normality of R-estimators in linear models. *Statistics.* **19**, 173-184.

Saleh, A.K.Md.E. (2006). *Theory of Preliminary Test and Stein-Type Estimation with Applications.* Wiley, New York.