The Annals of Mathematical Statistics
1968, Vol. 39, No. 3, 1083–1085

ON A SIMPLE ESTIMATE OF THE RECIPROCAL OF THE DENSITY FUNCTION

BY DANIEL A. BLOCH[1] AND JOSEPH L. GASTWIRTH[2]

The Johns Hopkins University

1. Introduction and summary. Let $x_1 < x_2 < \cdots < x_n$ be an ordered random sample of size n from the absolutely continuous cdf $F(x)$ with positive density $f(x)$ having a continuous first derivative in a neighborhood of the pth population quantile $\nu_p(= F^{-1}(p))$. In order to convert the median or any other "quick estimator" [1] into a test we must estimate its variance, or for large samples its asymptotic variance which depends on $1/f(\nu_p)$. Siddiqui [4] proposed the estimator $S_{mn} = n(2m)^{-1}(x_{[np]+m} - x_{[np]-m+1})$ for $1/f(\nu_p)$, showed it is asymptotically normally distributed and suggested that m be chosen to be of order $n^{\frac{1}{2}}$. In this note we show that the value of m minimizing the asymptotic mean square error (AMSE) is of order $n^{\frac{4}{5}}$ (yielding an AMSE of order $n^{-\frac{4}{5}}$). Our analysis is similar to Rosenblatt's [2] study of a simple estimate of the density function.

2. Large sample theory. In order to develop the large sample theory of S_{mn} it is convenient to consider the ordered sample $x_1 < x_2 < \cdots < x_n$ as a transform of an ordered sample $u_1 < u_2 < \cdots < u_n$ from a uniform distribution on $(0, 1)$ where $x_i = F^{-1}(u_i)$. For simplicity, let $G = F^{-1}$ (which exists as $f(x)$ is positive). We shall use the fact that the spacings from a uniform distribution have a beta distribution ([6], p. 236).

$$E(u_{[np]+m} - u_{[np]-m+1})^r = n!(2m + r - 2)!/[(n + r)!(2m - 2)!] \tag{2.1}$$

and

$$E(u^s_{[np]-m+1}u^r_{[np]+m}) = n!([np] + s - m)!([np] + m + r + s - 1)! \cdot [([np] - m)!([np] + m + s - 1)!(n + r + s)!]^{-1}. \tag{2.2}$$

As Siddiqui did not prove that the estimator is consistent we now do so.

THEOREM 1. *If $m = o(n)$ and $m \to \infty$ as $n \to \infty$, then the statistic S_{mn} is a consistent estimator of $g(p) = 1/f(\nu_p)$.*

PROOF. Since $u_{[np]+m}$ and $u_{[np]-m+1}$ converge to p in probability, expanding G about p yields the following representation of S_{mn} :

$$S_{mn} \sim n(2m)^{-1}g(p)(u_{[np]+m} - u_{[np]-m+1}) + o_p(u_{[np]+m} - u_{[np]-m+1}). \tag{2.3}$$

Received 3 April 1967; revised 15 January 1968.

[1] Research supported by the Office of Naval Research under contract NONR 4010(09) awarded to the Department of Statistics, The Johns Hopkins University. The author is now at the Biomathematics Division of Cornell University Graduate School of Medical Science and Sloan Kettering Institute, New York.

[3] Research supported by the National Science Foundation Research Grant No. GP 6225 awarded to the Department of Statistics, The Johns Hopkins University.

Using formula (2.1) and Chebyshev's inequality, the random variable $n(2m)^{-1}(u_{[np]+m} - u_{[np]-m+1})$ can be shown to converge in probability to one as $m \to \infty$. Thus, $S_{mn} = g(p) + o_p(1)$. (The consistency of S_{mn} can also be proved by using the methods of Sen [3].)

For completeness we include the asymptotic distribution of S_{mn} [4].

THEOREM 2 (Siddiqui). *Under the conditions of Theorem 1,*

$$(2.4) \qquad (2m)^{\frac{1}{2}}[S_{mn} - g(p)]/g(p) \to_{\mathcal{L}} N(0, 1).$$

For the remainder of the section we assume that the first three derivatives of f exist in a neighborhood of ν_p and proceed heuristically. From Theorem 2, the variance of S_{mn}

$$(2.5) \qquad \text{Var } (S_{mn}) \sim g^2(p)/2m$$

as $m \to \infty$ and $m/n \to 0$.

In order to obtain the asymptotically optimal choice of m, we require an expression for the asymptotic bias of S_{mn}. Expanding $G(u_{[np]+m}) - G(u_{[np]-m+1})$ in a Taylor series it can be shown, using (2.1) and (2.2) that

$$(2.6) \qquad E(S_{mn}) \sim g(p) + [g''(p)/6](m/n)^2$$

as $m \to \infty$ and $m/n \to 0$. Thus, the squared bias is given by

$$(2.7) \qquad (\text{bias } S_{mn})^2 = [E(S_{mn}) - g(p)]^2 \sim [g''(p)^2/36](m/n)^4$$

as $m \to \infty$ and $m/n \to 0$. Now m will be chosen to minimize the asymptotic mean square error

$$(2.8) \qquad E[S_{mn} - g(p)]^2 \sim g^2(p)/2m + [g''(p)^2/36](m/n)^4$$

as $m \to \infty$ and $m/n \to 0$. If m is set equal to cn^{γ}, $\gamma > 0$, it is easily seen from (2.8) that the optimal choice for γ is $\gamma = \frac{4}{5}$. Then (denoting S_{mn}, $m = cn^{\frac{4}{5}}$ by S_n)

$$(2.9) \qquad E[S_n - g(p)]^2 \sim g^2(p)/2cn^{\frac{4}{5}} + [g''(p)]^2c^4/36n^{\frac{4}{5}}$$

as $n \to \infty$. The value of c minimizing (2.9) is (assuming $g''(p) \neq 0$)

$$(2.10) \qquad \begin{aligned} c &= (9g^2(p)/2[g''(p)]^2)^{\frac{1}{5}} \\ &= (9f^8(\nu_p)/2[3(f'(\nu_p))^2 - f(\nu_p)f''(\nu_p)]^2)^{\frac{1}{5}}. \end{aligned}$$

With this choice of c and γ we find that

$$(2.11) \qquad E[S_n - g(p)]^2 \sim \tfrac{5}{4}9^{-\frac{1}{5}}2^{-\frac{4}{5}}[g(p)]^{\frac{8}{5}}[g''(p)]^{\frac{2}{5}}n^{-\frac{2}{5}}$$

as $n \to \infty$.

It should be noted that the above formulas are very similar to those obtained by Rosenblatt [2], pp. 835–836, for his estimate of the density function. While the problems considered by Rosenblatt and in this note are different, the solutions are isomorphic.

Recently [5] Weiss and Wolfowitz proposed another estimator of the density function which, in effect, estimates c. Presumably, their approach can be extended to the problem considered here.

The choice of c should be based on prior knowledge of the values of $f(\nu_p)$,

$f'(\nu_p)$ and $f''(\nu_p)$. The values of c when $p = \frac{1}{2}$ for some common densities are as follows: Normal (.5), Cauchy (.4), and Logistic (.58).

3. Extensions. It is also of interest to estimate $p(1 - q)/nf(\nu_p)f(\nu_q)$, $p < q$, the asymptotic covariance between the pth and qth sample quantiles. The discussion of the previous sections suggests that $1/f(\nu_p)f(\nu_p)$ be estimated by

$$(3.1) \qquad S_{m_1m_2n} = n^2(x_{[np]+m_1} - x_{[np]-m_1+1})(x_{[nq]+m_2} - x_{[nq]-m_2+1})/4_{m_1m_2}.$$

The consistency of $S_{m_1m_2n}$ (if $m_1 = o(n)$, $m_2 = o(n)$, $m_1 \to \infty$ as $n \to \infty$ and $m_2 \to \infty$ as $n \to \infty$) follows from the fact that the product of two consistent Siddiqui estimators is consistent. It is also easy to show that

$$(3.2) \qquad ((S_{m_1m_2n} - g(p)g(q))/g(p)g(q)[(2m_1)^{-1} + (2m_2)^{-1}]^{\frac{1}{2}}) \to_{\mathfrak{L}} N(0, 1).$$

Thus, the variance of the estimate

$$(3.3) \qquad \text{Var}(S_{m_1m_2n}) \sim g^2(p)g^2(q)[(2m_1)^{-1} + (2m_2)^{-1}]$$

as $m_1 \to \infty$, $m_2 \to \infty$, $m_1/n \to 0$ as $n \to \infty$ and $m_2/n \to 0$ as $n \to$. Since $S_{m_1m_2n}$ is asymptotically the product of two independent Siddiqui-estimators, the mean of the estimate

$$(3.4) \qquad E(S_{m_1m_2n}) \sim g(p)g(q) + g(p)g''(q)(m_2/n)^2/6 + g''(p)g(q)(m_1/n)^2/6$$

as $m_1 \to \infty$ $m_2 \to \infty$, $m_1/n \to 0$ as $n \to \infty$ and $m_2/n \to 0$ as $n \to \infty$. We now assume that $g''(p) \neq 0$, $g''(q) \neq 0$ and $g''(p)/g''(q) > 0$. As before we choose m_1 and m_2 to minimize the asymptotic mean square error.

$$(3.5) \qquad \begin{aligned} &E[S_{m_1m_2n} - g(p)g(q)]^2 \\ &\quad \sim [g(p)g''(q)(m_2/n)^2/6 + g(q)g''(p)(m_1/n)^2/6] \\ &\qquad\qquad + g^2(p)g^2(q)[(2m_1)^{-1} + (2m_2)^{-1}]. \end{aligned}$$

Letting $m_1 = c_1n^{\gamma_1}$, $\gamma_1 > 0$ and $m_2 = c_2n^{\gamma_2}$, $\gamma_2 > 0$ we find that the optimum choices of γ_1 and γ_2 are $\gamma_1 = \gamma_2 = \frac{4}{5}$. The optimal value of c_1 is

$$(3.6) \qquad c_1 = (9g^2(p)/2[g''(p)]^2[1 + (g(p)g''(q)/g''(p)g(q))^{\frac{1}{3}}])^{\frac{1}{5}}.$$

The optimal value of c_2 is given by the same formula with p and q interchanged.

Acknowledgment. The authors would like to thank the referee for his helpful suggestions.

REFERENCES

[1] GASTWIRTH, J. L. (1966). On robust procedures. *J. Amer. Statist. Assoc.* **61** 929–948.

[2] ROSENBLATT, M. (1956). Remarks on some non-parametric estimates of a density function. *Ann. Math. Statist.* **27** 832–837.

[3] SEN, P. K. (1966). On a distribution-free method of estimating asymptotic efficiency of a class of non-parametric tests. *Ann. Math. Statist.* **37** 1759–1770.

[4] SIDDIQUI, M. M. (1960). Distribution of quantiles in samples from a bivariate population. *J. Res. N.B.S.* **64B** 145–150.

[5] WEISS, L. and WOLFOWITZ, J. (1967). Estimation of a density function at a point. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete* **7** 327–335.

[6] WILKS, S. S. (1962). *Mathematical Statistics.* Wiley, New York.