REMARKS ON SOME NONPARAMETRIC ESTIMATES OF A DENSITY FUNCTION[1]

BY MURRAY ROSENBLATT[2]

University of Chicago

1. Summary. This note discusses some aspects of the estimation of the density function of a univariate probability distribution. All estimates of the density function satisfying relatively mild conditions are shown to be biased. The asymptotic mean square error of a particular class of estimates is evaluated.

2. Estimates of the density function. Let $X_1, \cdots, X_n$ be independent and identically distributed random variables with continuous density function $f(y)$. Let $S(y; X_1, \cdots, X_n)$ be an estimate of $f(y)$. The function $S(y; x_1, \cdots, x_n)$ is assumed to be jointly Borel measurable in $(y, x_1, \cdots, x_n)$. It is also assumed that

$$S(y; x_1, \cdots, x_n) \geqq 0,$$

since $f(y) \geqq 0$.

It can easily be shown that

$$S(y; X_1, \cdots, X_n)$$

is not an unbiased estimate of $f(y)$. Suppose to the contrary that

$$ES(y; X_1, \cdots, X_n) \equiv f(y) \tag{1}$$

for all continuous f and all y. Condition (1) implies that for each y,

$$ES(y; X_1, \cdots, X_n) < \infty.$$

Assume that $S(y; x_1, \cdots, x_n)$ is a symmetric function of $x_1, \cdots, x_n$, since the symmetrized n-tuple $(X_1, \cdots, X_n)$ is a sufficient statistic for the problem. But then

$$\int_a^b S(y; X_1, \cdots, X_n)\, dy \tag{2}$$

is a symmetric estimate of

$$F(b) - F(a) = \int_a^b f(y)\, dy.$$

Moreover, (2) is an unbiased estimate of $F(b) - F(a)$, since

$$E \int_a^b S(y; X_1, \cdots, X_n)\, dy = \int_a^b ES(y; X_1, \cdots, X_n)\, dy$$
$$= \int_a^b f(y)\, dy = F(b) - F(a)$$

Received April 27, 1955.

[1] Research carried out at the Statistical Research Center, University of Chicago, under the sponsorship of the Statistics Branch, Office of Naval Research.

The comments of R. R. Bahadur have been very helpful.

[2] Now at Indiana University.

R.A. Davis et al. (eds.), *Selected Works of Murray Rosenblatt*, Selected Works in Probability and Statistics, DOI 10.1007/978-1-4419-8339-8_13,

by Fubini's theorem. However, the only unbiased estimate of $F(b) - F(a)$ symmetric in the observations $X_1, \cdots, X_n$ is $F_n(b) - F_n(a)$, where $F_n(y)$ is the sample distribution function. This follows immediately from the fact that the symmetrized n-tuple $(X_1, \cdots, X_n)$ is a complete statistic [2]. Thus,

$$F_n(b) - F_n(a) = \int_a^b S(y; X_1, \cdots, X_n)\, dy$$

for all a and b and almost all $X_1, \cdots, X_n$. But then $F_n(y)$ is absolutely continuous in y for almost all $X_1, \cdots, X_n$, which is impossible.

One need not require $S(y; x_1, \cdots, x_n)$ to be nonnegative. An assumption like

$$\int_a^b E|S(y; X_1, \cdots, X_n)|\, dy < \infty$$

for some two values, a, b, with $a < b$, would lead to the same conclusion, that is, that there are no unbiased estimates $S(y; X_1, \cdots, X_n)$ of $f(y)$ satisfying this condition.

3. The difference quotient of the sample distribution function. An obvious estimate of $f(y)$ is the difference quotient

$$S(y; X_1, \cdots, X_n) = f_n(y) = \frac{F_n(y + h) - F_n(y - h)}{2h}$$

of the sample distribution function $F_n(y)$, where $h = h_n$ is a function of the sample size n and approaches zero as $n \to \infty$. The asymptotic behavior of this estimate as $n \to \infty$ is examined in terms of its mean square error. Fix and Hodges have used an estimate of this form in their discussion of a nonparametric discrimination problem [1].

Now,

$$EF_n(y) = F(y),$$

$$E[F_n(y)F_n(y')] = \frac{1}{n} F(\min(y, y')) + \frac{n-1}{n} F(y)F(y'),$$

so that

$$\operatorname{cov}(F_n(y), F_n(y')) = \frac{1}{n}[F(\min(y, y')) - F(y)F(y')].$$

But then,

$$\operatorname{cov}(f_n(y), f_n(y')) = \frac{1}{4h^2n}[F(\min(y + h, y' + h)) - F(y + h)F(y' + h)$$
$$- F(\min(y + h, y' - h)) + F(y + h)F(y' - h) - F(\min(y - h, y' + h))$$
$$+ F(y - h)F(y' + h) + F(\min(y - h, y' - h)) - F(y - h)F(y' - h)].$$

On setting $y = y'$,

$$\sigma^2(f_n(y)) = \frac{1}{4h^2n}[F(y + h) - F(y - h) + (F(y + h) - F(y - h))^2].$$

Now consider the behavior of $f_n(y)$ where y is fixed as $n \to \infty$ and $h \to 0$. The mean square error

$$E|f_n(y) - f(y)|^2 = \sigma^2(f_n(y)) + (Ef_n(y) - f(y))^2$$
$$= \frac{1}{4h^2n}[F(y+h) - F(y-h) + (F(y+h) - F(y-h))^2]$$
$$+ \left[\frac{1}{2h}(F(y+h) - F(y-h)) - f(y)\right]^2$$

is a reasonable measure of how good an estimate $f_n(y)$ is of $f(y)$ locally at y. The density function f is assumed to be sufficiently regular for the following evaluation of the mean square error to be carried through. It will be enough to assume that the first three derivatives of f exist at y. Then

$$F(y+h) - F(y-h) = \int_{y-h}^{y+h} f(u)\,du$$
$$= \int_{y-h}^{y+h} [f(y) + (u-y)f'(y)$$
$$+ \tfrac{1}{2}(u-y)^2 f''(y) + 0\,|u-y|^3]\,du$$
$$= 2hf(y) + \tfrac{1}{3}f''(y)h^3 + 0(|h|^4).$$

Assume that $f''(y) \neq 0$. Then

$$(Ef_n(y) - f(y))^2 \sim \left(\frac{h^2}{6}f''(y)\right)^2 = \frac{h^4}{36}|f''(y)|^2$$

as $h \to 0$. The variance of the estimate

$$\sigma^2(f_n(y)) \sim \frac{f(y)}{2hn}$$

as $h \to 0$. The asymptotic mean square error

$$E|f_n(y) - f(y)|^2 \sim \frac{f(y)}{2hn} + \frac{h^4}{36}|f''(y)|^2 + o\left(\frac{1}{hn} + h^4\right) \tag{3}$$

as $h \to 0$ and $n \to \infty$. The question of an optimal choice of $h = h_n$ as a function of n now arises. If h is set equal to $kn^{-\alpha}$, $\alpha > 0$, it is easily seen from (3) that the optimal choice of α is $\alpha = \frac{1}{5}$. The optimal value of k is then the one minimizing

$$\frac{f(y)}{2k} + \frac{k^4}{36}|f''(y)|^2.$$

This value of k is

$$k = \left[\frac{9}{2}\frac{f(y)}{|f''(y)|^2}\right]^{1/5}.$$

With this choice of k and α, we find that

$$E|f_n(y) - f(y)|^2 \sim \tfrac{5}{4}9^{-1/5}2^{-4/5}f(y)^{4/5}|f''(y)|^{2/5}n^{-4/5}.$$

The choice of k would be based on guesses as to the magnitude of $f(y), f''(y)$.

One is led to a choice of h as a function of n that is independent of y by considering a global measure of how good f_n is as an estimate of f. The integrated mean square error

$$\int_{-\infty}^{\infty} E\,|f_n(y) - f(y)|^2\,dy$$

is a simple measure of this type. Let $f(y), f''(y)$ be bounded continuous functions that are square integrable. One is then led to the following asymptotic expression

$$\int_{-\infty}^{\infty} E\,|f_n(y) - f(y)|^2\,dy \sim \frac{1}{2hn} + \frac{h^4}{36}\int_{-\infty}^{\infty} |f''(y)|^2\,dy + o\left(\frac{1}{hn} + h^4\right)$$

as $h \to 0$ and $n \to \infty$. The optimal choice of h as a function of n is

$$h = kn^{-1/5},$$

where k is now

$$k = \left[\frac{9}{2\int_{-\infty}^{\infty} |f''(y)|^2\,dy}\right]^{1/5},$$

and

$$\int_{-\infty}^{\infty} E\,|f_n(y) - f(y)|^2\,dy \sim 2^{-4/5}9^{-1/5}\frac{5}{4}\left[\int_{-\infty}^{\infty} |f''(y)|^2\,dy\right]^{1/5} n^{-4/5}$$

as $n \to \infty$.

4. A class of estimates of the density function. The discussion of the previous section suggests that the following class of estimates will be of interest. Let $w_n(u)$ be a nonnegative function such that

$$\int_{-\infty}^{\infty} w_n(u)\,du = 1.$$

The sequence of functions $\{w_n(u)\}$ is chosen so that the total mass concentrates in the neighborhood of zero as $n \to \infty$; that is, given any $\epsilon > 0$,

$$\int_{|u|<\epsilon} w_n(u)\,du \to 1$$

as $n \to \infty$. Corresponding to each sequence of weight functions $\{w_n(u)\}$ of this type, there is an estimate

$$f_n(y) = \int_{-\infty}^{\infty} w_n(y - u)\,dF_n(u) = \frac{1}{n}\sum_{j=1}^{\infty} w_n(y - X_j).$$

Now

$$Ef_n(y) = \int_{-\infty}^{\infty} w_n(y - x)\, dF(x) = \int_{-\infty}^{\infty} w_n(y - x) f(x)\, dx,$$

and

$$\operatorname{cov}\,(f_n(y), f_n(y')) = \frac{1}{n}\left[\int_{-\infty}^{\infty} w_n(y - x) w_n(y' - x) f(x)\, dx - \left(\int_{-\infty}^{\infty} w_n(y - x) f(x)\, dx\right)^2\right].$$

On setting $y = y'$, we have

$$\sigma^2(f_n(y)) = \frac{1}{n}\left[\int_{-\infty}^{\infty} w_n^2(y - x) f(x)\, dx - \left(\int_{-\infty}^{\infty} w_n(y - x) f(x)\, dx\right)^2\right].$$

Note that all estimates of this form are themselves density functions; that is,

$$f_n(y) \geqq 0,$$

and

$$\int_{-\infty}^{\infty} f_n(y)\, dy = 1.$$

An estimate $f_n(y)$ with any desired regularity properties can be obtained by choosing a weight function $w_n(u)$ with these same regularity properties. Thus, $f_n(y)$ will be analytic if $w_n(u)$ is.

As an example, consider

$$w_n(u) = \frac{1}{h} w\left(\frac{u}{h}\right),$$

where $h = h_n \to 0$ as $n \to \infty$, and

$$\int_{-\infty}^{\infty} w(u)\, du = 1.$$

The estimate discussed in the previous section is obtained on setting

$$w(u) = \begin{cases} \frac{1}{2} & \text{when } |u| < 1, \\ 0 & \text{otherwise.} \end{cases}$$

The function $w(u)$ is assumed to be square integrable. Then

$$\sigma^2(f_n(y)) = \frac{1}{n}\left[\frac{1}{h}\int_{-\infty}^{\infty} w^2(u) f(y + hu)\, du - \left(\int_{-\infty}^{\infty} w(u) f(y + hu)\, du\right)^2\right],$$

and

$$Ef_n(y) - f(y) = \int w(u) f(y + hu)\, du - f(y).$$

It is clear that

$$\int w(u)f(y + hu)\,du \to f(y),$$

and that

$$\int w^2(u)f(y + hu)\,du \to f(y)\int w^2(u)\,du$$

as $h \to 0$ for every continuous density function f. Hence,

$$\sigma^2(f_n(y)) \sim \frac{1}{nh} f(y)\int w^2(u)\,du.$$

The integral

$$\int w(u)\,|u|^3\,du$$

is assumed to be finite, and $f(y)$ is assumed to have continuous derivatives of the first three orders in the following computation of the bias. The bias of the estimate is then

$$Ef_n(y) - f(y) = \int w(u)[f(y + hu) - f(y)]\,du$$

$$= hf'(y)\int w(u)u\,du + \tfrac{1}{2}h^2f''(y)\int w(u)u^2\,du + 0(|h|^3).$$

It is now clear that it would be advantageous to have

$$\int w(u)u\,du = 0.$$

This condition will be satisfied if $w(u)$ is symmetric about zero. Using the same sort of argument as was used in the last section, it is easily seen that the mean square error of these estimates can be made no smaller than $0(n^{-4/5})$ for all admissible f. It would be very interesting to find out whether there are other estimates $f_n(y)$ with an asymptotic behavior of the order $1/n$ as $n \to \infty$.

REFERENCES

[1] E. Fix and J. L. Hodges, Jr., *Discriminatory Analysis, Nonparametric Discrimination: Consistency Properties*, USAF School of Aviation Medicine, Project No. 21-49-004, Report No. 4.

[2] E. Lehmann, "Notes on the theory of estimation," University of California, Berkeley (1950).