A CONSISTENT ESTIMATOR OF A COMPONENT OF A CONVOLUTION

By William R. Gaffey

University of California

1. Introduction and summary. Suppose the observed random variable X is the sum of two independent random variables Z and Y, where Z has a normal distribution with zero expectation and a known variance, and Y has a distribution function, say $G(y)$, which is completely unknown. Then the distribution function of X may be written as

$$F(x) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} G(y) \exp\left[-\frac{(x-y)^2}{2\sigma^2}\right] dy, \tag{1.1}$$

where $F(x)$ and $G(y)$ are unknown.

We consider here the problem of estimating $G(y)$ from a sample $x_1, x_2, \cdots x_n$. Such a problem may arise if, for example, each x_i represents a serum cholesterol determination on one human being randomly selected from some population. Then each x_i may be thought of as the true cholesterol value for that person, plus an "instrumental error" introduced by the complex chemical analysis. We may wish to "correct" for the instrumental error, so to speak, by estimating the distribution of true cholesterol levels in the population.

The maximum likelihood and minimum distance principles do not seem to yield estimators which may be expressed as explicit, more or less easily computable functions of the sample values. We present such an estimator, which is consistent at every continuity point of $G(y)$. (We consider only continuity points throughout the paper.) The estimator is constructed by first exhibiting an inversion formula for $G(y)$ in terms of the derivatives of $F(x)$, and then replacing the derivatives by the difference quotients of the empiric distribution function $F_n(x)$.

The asymptotic mean square error of the estimator is derived, and a rough rule is suggested for deciding when it is worthwhile to compute an estimate. The fact that the estimator is still consistent under certain kinds of dependence between Z and Y is indicated. Finally, some comments are made on the relationship between the present estimator and one derived by Eddington.

2. The inversion formula for $G(x)$. Denote by $F^{(2k)}(x)$ the $2k$th derivative of $F(x)$ evaluated at x. Pollard [3] has derived a formula which states, in effect, that

$$\lim_{n\to 0} \sum_{k=0}^{\infty} (-1)^k F^{(2k)}(x) (\sigma^2 t_n/2)^k / k! = G(x), \tag{2.1}$$

where $\{t_n\}$ is any increasing sequence of positive numbers with $\lim t_n = 1$.

We require a modification of this inversion formula. Let $\{t_n\}$ be an increasing

Received November 26, 1957; revised September 20, 1958.

sequence of positive numbers such that lim $t_n = 1$ and lim $n^k t_n^n = 0$ for any $k > 0$. Consider

$$G_n(x) = \sum_{k=0}^{n} (-1)^k F^{(2k)}(x)(\sigma^2 t_n/2)^k/k!. \tag{2.2}$$

Then lim $G_n(x) = G(x)$. To prove this it is sufficient to show that

$$\lim \sum_{k=n+1}^{\infty} (-1)^k F^{(2k)}(x)(\sigma^2 t_n/2)^k/k! = 0. \tag{2.3}$$

Now

$$F^{(2k)}(x) = \frac{1}{\sqrt{\pi}\,(\sqrt{2}\sigma)^{2k+1}} \int_{-\infty}^{\infty} H_{2k}\left(\frac{x - y}{\sqrt{2}\sigma}\right) \exp\left[-\frac{(x - y)^2}{2\sigma^2}\right] G(y)\,dy, \tag{2.4}$$

where $H_{2k}(x)$ is the $2k$th Hermite polynomial satisfying

$$H_{2k}(x) \leqq A2^k\sqrt{(2k)!}\exp[x^2/2], \tag{2.5}$$

A being independent of x and k. ([5], p. 236)

Therefore,

$$F^{(2k)}(x) \leqq A\sqrt{(2k!)}/\sigma^{2k}, \tag{2.6}$$

and consequently the absolute value of the sum in (2.3) is at most equal to

$$A \sum_{k=n+1}^{\infty} \sqrt{(2k)!}\,(t_n/2)^k/k!. \tag{2.7}$$

Since the coefficients of t_n^k are bounded above, the absolute value of (2.7) is no greater than

$$At_n^{n+1}(1 - t_n)^{-1}, \tag{2.8}$$

which approaches zero faster than any negative power of n.

If $G^{(r)}(x)$, the rth derivative of $G(x)$, is integrable, and continuous at a particular value of x, a similar argument shows that lim $G_n^{(r)}(x) = G^{(r)}(x)$.

For later use, we note that by virtue of (2.6) and the continuity of any derivative of $F(x)$,

$$F^{(2k)}(a) = F^{(2k)}(b) + O(a - b)\sqrt{(2k)!}/\sigma^{2k}, \tag{2.9}$$

and that the upper bound of $O(a - b)$ is independent of k.

3. The estimator of $G(x)$ and its mean square error. Define the $2k$th difference quotient of the empiric distribution function $F_n(x)$ by

$$F_n^{(2k)}(x, h) = (2h)^{-2k} \sum_{j=0}^{2k} \binom{2k}{j} (-1)^j F_n(x + (k - j)2h), \tag{3.1}$$

for $h > 0$. The estimator of $G(x)$, for a sample of size n, will then be

$$\hat{G}_n(x) = \sum_{k=0}^{n} (-1)^k F_n^{(2k)}(x, h_n)(\sigma^2 t_n/2)^k/k!, \tag{3.2}$$

where lim $h_n = 0$.

Clearly, the asymptotic properties of the estimator will depend on the choice of sequences $\{h_n\}$ and $\{t_n\}$. We derive below an asymptotic expression for the mean square error on the assumption that $G^{(4)}(x)$ is integrable, and continuous at the particular value of x involved. Where there is no possibility of confusion, the subscripts are omitted from h_n and t_n.

Consider first the expectation of the $2k$th difference quotient.

$$(2h)^{2k}E[F_n^{(2k)}(x, h)] = \int_{x-h}^{x+h} \cdots \int_{y_{2k-1}-h}^{y_{2k-1}+h} F^{(2k)}(y_{2k})\, dy_{2k} \cdots dy_1 . \tag{3.3}$$

We may write

$$\begin{aligned} F^{(2k)}(y_{2k}) = F^{(2k)}(y_{2k-1}) &+ (y_{2k} - y_{2k-1})F^{(2k+1)}(y_{2k-1}) \\ &+ (1/2)(y_{2k} - y_{2k-1})^2 F^{(2k+2)}(\xi), \end{aligned} \tag{3.4}$$

where ξ is between y_{2k} and y_{2k-1}. When the integral with respect to y_{2k} is taken, the second term on the right vanishes. Complete integration of the third term on the right results in an expression which may be written as

$$(2h)^{2k}F^{(2k+2)}(\xi)h^2/6, \tag{3.5}$$

where $x - 2hk \leqq \xi \leqq x + 2hk$. Therefore,

$$\begin{aligned} (2h)^{2k}E[F_n^{(2k)}(x, h)] = 2h \int_{x-h}^{x+h} \cdots \int_{y_{2k-2}-h}^{y_{2k-2}+h} & F^{(2k)}(y_{2k-1})\, dy_{2k-1} \cdots dy_1 \\ &+ (2h)^{2k}F^{(2k+2)}(\xi)h^2/6. \end{aligned} \tag{3.6}$$

Repeating the process, we find finally that

$$(2h)^{2k}E[F_n^{(2k)}(x, h)] = (2h)^{2k}F^{(2k)}(x) + (2h)^{2k} \sum_{i=1}^{2k} F^{(2k+2)}(\xi_i)h^2/6, \tag{3.7}$$

where $x - 2hk \leqq \xi_i \leqq x + 2hk$ for all i, or more simply,

$$E[F_n^{(2k)}(x, h)] = F^{(2k)}(x) + 2kF^{(2k+2)}(\eta_k)h^2/6, \tag{3.8}$$

where $x - 2hk \leqq \eta_k \leqq x + 2kh$. Making this substitution in the expectation of $\hat{G}_n(x)$, we have

$$E[\hat{G}_n(x)] = G_n(x) + \frac{h^2}{6} \sum_{k=0}^{n} (-1)^k F^{(2k+2)}(\eta_k)(\sigma^2 t/2)^k 2k/k!. \tag{3.9}$$

Applying (2.9) we obtain, after some algebra,

$$\begin{aligned} E[\hat{G}_n(x)] = G_n(x) - \frac{h^2\sigma^2 t}{6} \sum_{k=0}^{n-1} & (-1)^k F^{(2k+4)}(x)(\sigma^2 t/2)^k/k! \\ &+ O\left[h^3 \sum_{k=1}^{n} k\sqrt{(2k+2)!}\, t^k/2^k(k-1)!\right]. \end{aligned} \tag{3.10}$$

Now by virtue of the properties of the sequence $\{t_n\}$,

$$\sum_{k=1}^{n} k\sqrt{(2k+2)!}\,t_n^k/2^k(k-1)! = O\left[\sum_{k=1}^{n} k^3 t_n^k\right] = O[(1-t_n)^{-4}]. \tag{3.11}$$

Therefore, if $\lim h_n(1-t_n)^{-4} = 0$, we have

$$[E[\hat{G}_n(x)] - G_n(x)]^2 \sim \frac{h_n^4\, t_n^2\, \sigma^4}{36}\,[G^{(4)}(x)]^2. \tag{3.12}$$

Now consider the variance of $\hat{G}_n(x)$. We note [4] that

$$E[F_n(x)F_n(y)] = \frac{n-1}{n}\,F(x)F(y) + \frac{1}{n}\,F\,(\min x, y). \tag{3.13}$$

Writing out the expectation of $\hat{G}_n^2(x)$ with this substitution, we obtain

$$E[\hat{G}_n^2(x)] = \frac{n-1}{n}\,E^2[\hat{G}_n(x)] + B_n\,, \tag{3.14}$$

or

$$\sigma^2(\hat{G}_n(x)) = B_n - \frac{1}{n}\,E^2[\hat{G}_n(x)], \tag{3.15}$$

where

$$\begin{aligned} B_n = \frac{1}{n}\sum_{k,r=0}^{n}\left(\frac{-\sigma^2 t}{8h^2}\right)^{k+r}\frac{1}{k!r!}\sum_{j=0}^{2k}\sum_{s=0}^{2r}\binom{2k}{j}\binom{2r}{s}(-1)^{j+s} \\ \cdot F[\min\,(x+(k-j)2h),\,(x+(r-s)2h)]. \end{aligned} \tag{3.16}$$

For convenience in writing, let $a_n = \sigma^2 t_n/8h_n^2$. After some manipulation it can be shown that

$$B_n = \frac{1}{n}\,E[\hat{G}_n(x)] + C_n\,; \tag{3.17}$$

where

$$\begin{aligned} C_n = \frac{1}{n}\sum_{k,r=0}^{n}(-a_n)^{k+r}\frac{1}{k!r!}\sum_{s=0}^{2r}\left[\sum_{j=0}^{k-r+s-1}\binom{2k}{j}(-1)^j\right]\binom{2r}{s} \\ \cdot(-1)^s[F(x+(r-s)2h_n) - F(x+(k-j)2h_n)]. \end{aligned} \tag{3.18}$$

Using the fact that, for $0 \leqq m < 2k$,

$$\sum_{j=0}^{m}(-1)^j\binom{2k}{j} = (-1)^m\binom{2k-1}{m}, \tag{3.19}$$

and, from (2.9), that

$$\begin{aligned} &F(x+(r-s)2h_n) - F(x+(k-j)2h_n) \\ &\quad = 2h_n[(r-s)-(k-j)] + O[h_n^2(r-s)^2 + h_n^2(k-j)^2], \end{aligned} \tag{3.20}$$

we find that

$$C_n = \frac{2h_n}{n}\sum_{k,r=0}^{n}\frac{a_n^{k+r}}{k!r!}\binom{2k+2r-2}{k+r-1} + O\left[\frac{h_n^2}{n}\sum_{k,r=0}^{n}\frac{a_n^{k+r}}{k!r!}\binom{2k+2r-2}{k+r-1}\frac{kr}{k+r}\right]. \tag{3.21}$$

If we let $k + r = j$, (3.21) may be rewritten as

$$C_n \sim \frac{2h_n}{n}\sum_{j=1}^{2n}\frac{(2a_n)^j}{j!}\binom{2j-2}{j-1} + O\left[\frac{1}{n}\sum_{j=0}^{2n-1}\frac{(2a_n)^j}{j!}\binom{2j}{j}\right]. \tag{3.22}$$

It is easy to verify the following equalities as Taylor series:

$$\begin{aligned}\sum_{j=1}^{\infty}\frac{b^j}{j!}\binom{2j-2}{j-1} &= \frac{1}{2\pi}\int_0^{\pi/2}\frac{\exp(4b\cos^2 x)-1}{\cos^2 x}\,dx \\ &= \frac{4b}{\pi}\int_0^{\pi/2}\sin^2 x\exp(4b\cos^2 x)\,dx = bO(e^{4b})\end{aligned} \tag{3.23}$$

and

$$\sum_{j=0}^{\infty}\frac{b^j}{j!}\binom{2j}{j} = \frac{2}{\pi}\int_0^{\pi/2}\exp(4b\cos^2 x)\,dx = O(e^{4b}). \tag{3.24}$$

Therefore, as b increases, (3.24) becomes negligible compared with (3.23).

By the use of the integral form of the remainder,

$$\begin{aligned}\sum_{j=1}^{2n}\frac{b^j}{j!}\binom{2j-2}{j-1} = \frac{1}{2\pi}\int_0^{\pi/2}&(\cos x)^{-2} \\ &\cdot\left[\frac{\exp(4b\cos^2 x)}{(2n)!}\int_{4b\cos^2 x}^{\infty}v^{2n}e^{-v}\,dv - 1\right]dx\end{aligned} \tag{3.25}$$

and

$$\sum_{j=0}^{2n-1}\frac{b^j}{j!}\binom{2j}{j} = \frac{2}{\pi(2n-1)!}\int_0^{\pi/2}\exp(4b\cos^2 x)\left[\int_{4b\cos^2 x}^{\infty}v^{2n-1}e^{-v}\,dv\right]dx. \tag{3.26}$$

Let $b = 2a_n$. Now if $a_n/n \rightarrow 0$, it is known ([7], Chap. 7) that, for any $y \geqq 0$,

$$\lim\frac{1}{(2n)!}\int_{8a_n y}^{\infty}v^{2n}e^{-v}\,dv = 1. \tag{3.27}$$

Therefore, under this assumption, the ratios of (3.25) to (3.23), and of (3.26) to (3.24), approach unity as n increases, so that (3.26) may be neglected. As a result we have

$$C_n \sim \frac{2h_n}{n}\sum_{j=1}^{2n}\frac{(2a_n)^j}{j!}\binom{2j-2}{j-1} \tag{3.28}$$

and

$$\sigma^2(\hat{G}_n(x)) \sim \frac{E(\hat{G}_n(x))[1-E(\hat{G}_n(x))]}{n} + \frac{2h_n}{n}\sum_{j=1}^{2n}\frac{(2a_n)^j}{j!}\binom{2j-2}{j-1}. \tag{3.29}$$

Putting together (3.29) and (3.12), and taking account of the rapidity of the convergence of $G_n(x)$ to $G(x)$, we have for the asymptotic mean square error,

$$E[(\hat{G}_n(x) - G(x))^2] \sim \frac{G(x)(1 - G(x))}{n} + \frac{h_n^4 t_n^2 \sigma^4}{36} [G^{(4)}(x)]^2 + \frac{2h_n}{n} \sum_{j=1}^{2n} \left(\frac{\sigma^2 t_n}{4h_n^2}\right)^j \frac{1}{j!} \binom{2j - 2}{j - 1}. \tag{3.30}$$

In order for this asymptotic expression to be valid and for the estimator to be consistent, C_n must approach zero, and h_n and t_n must obey the restrictions imposed during the derivation. These conditions on h_n and t_n are summarized here:

(A) $\lim h_n(1 - t_n)^{-4} = 0,$

and

(B) $\lim \frac{t_n}{nh_n} \exp\left[\frac{\sigma^2 t_n}{h_n^2}\right] = 0.$

Condition (B) is sufficient to ensure that C_n and a_n/n approach zero.

4. Specific sequences $\{h_n\}$ and $\{t_n\}$. The logical step, after deriving the mean square error, is to determine sequences $\{h_n\}$ and $\{t_n\}$ which minimize it. In the present case the complexity of (3.30) makes this extremely difficult. Alternatively, we may search for easily computed sequences satisfying conditions (A) and (B). Suppose we let

$$h_n = a(\ln n)^{-\alpha}, \qquad \alpha > 0 \tag{4.1}$$

and

$$t_n = 1 - (\ln n)^{-\beta}, \qquad \beta > 0. \tag{4.2}$$

Then it may be verified that if $\beta < \alpha/4$, condition (A) is satisfied. If, in addition, $\alpha = 1/2$ and $a \geqq \sigma$, then condition (B) is satisfied, and in fact C_n becomes negligible.

In order to minimize the bias, whose square is the second term of (3.30), it is reasonable to let a assume its minimum value. Finally, the convergence of G_n to G depends on the fact that $\beta > 0$. A value convenient for computation, say $\beta = 0.1$, is suggested. These "convenient" sequences are then

$$h_n = \sigma(\ln n)^{-.5} \tag{4.3}$$

and

$$t_n = 1 - (\ln n)^{-.01}. \tag{4.4}$$

5. Remarks on the bias of $\hat{G}_n(x)$. It is possible that, with smaller sample sizes, the bias introduced by the estimating procedure may be greater than the bias involved in ignoring the whole problem and simply using the original sample distribution function to estimate $G(x)$. A reasonable rule of thumb for deciding if it is worthwhile to compute $\hat{G}_n(x)$ is to do so if the maximum bias of $\hat{G}_n(x)$ does not exceed the maximum bias of $F_n(x)$ for the given sample size. To get some idea of the order of magnitude of the sample sizes required, we will assume that $G(x)$ is a normal distribution function with variance τ^2.

The maximum bias which can result from using $F_n(x)$ is

$$\max_x |E[F_n(x)] - G(x)| = \max_{x>0} [\Phi(x/\tau) - \Phi(x/\sqrt{\sigma^2 + \tau^2})], \tag{5.1}$$

where $\Phi(x)$ is the standard normal distribution function. The maximum is attained when

$$x = \frac{\tau}{\sigma}\left[(\sigma^2 + \tau^2) \ln\left(1 + \frac{\sigma^2}{\tau^2}\right)\right]^{\frac{1}{2}}.$$

The maximum asymptotic bias of $\hat{G}_n(x)$ is

$$\max_x \frac{h_n^2 \sigma^2 t_n}{6} | G^{(4)}(x) | = \frac{h_n^2 \sigma^2 t_n}{6} \max_x | G^{(4)}(x) |, \tag{5.2}$$

where

$$G^{(4)}(x) = \frac{1}{\sqrt{2\pi}}\left(\frac{3x}{\tau^5} - \frac{x^3}{\tau^7}\right) e^{-x^2/2\tau^2}. \tag{5.3}$$

$| G^{(4)}(x) |$ attains its maximum at $x = .742\ \tau$, and

$$\max_x | G^{(4)}(x) | = \frac{1.38}{\sqrt{2\pi}\tau^4}. \tag{5.4}$$

Therefore, the maximum asymptotic bias of $\hat{G}_n(x)$ is no greater than the maximum bias of $F_n(x)$ if (using the forms (4.3) and (4.4) for h_n and t_n)

$$\begin{aligned}(\ln n)^{-1}[1 - (\ln n)^{-0.1}] &\leqq 4.35\sqrt{2\pi}\left(\frac{\tau}{\sigma}\right)^4 \\ &\cdot \left[\Phi(\{(1 + \tau^2/\sigma^2) \ln (1 + \sigma^2/\tau^2)\}^{\frac{1}{2}}) - \Phi\left(\frac{\tau}{\sigma}\{\ln (1 + \sigma^2/\tau^2)\}^{\frac{1}{2}}\right)\right].\end{aligned} \tag{5.5}$$

Since this inequality involves the asymptotic bias of $\hat{G}_n(x)$, it is presumbly not too trustworthy for small n. Substituting $n = 30$ and solving for σ/τ, we find that (5.5) holds if $\sigma/\tau \leqq 3$. In other words, a sample of size 30 justifies computing $\hat{G}_n(x)$ if the standard deviation of the known component is no more than three times the standard deviation of the unknown component. Therefore, even though (5.5) is an asymptotic expression, it seems reasonable to say that with samples of size 30 or larger, it is worthwhile to compute $\hat{G}_n(x)$ for most situations of practical interest.

6. Dependence between Z and Y. Suppose now that the normal random variable Z, instead of having expectation zero, is dependent on Y in the sense that it has an expectation μ_y when $Y = y$. Then if $y + \mu_y$ is a continuous, strictly monotone function of y, the analyses leading to (2.1) and (3.2) are still valid, provided the derivatives and difference quotients are now evaluated at the point $x + \mu_x$. Therefore, under this kind of dependence, $\hat{G}_n(x + \mu_x)$ estimates $G(x)$.

7. Other estimators of $G(x)$. Although the present paper arose from consideration of some public health problems, the problem of instrumental error has had a long history in astronomy. In particular, several solutions to the integral equation (1.1), under varying restrictions on $G(x)$, have been given by astronomers. (See [2] and [6] for bibliographies.) With the exception of [2], however, the solutions themselves are offered as estimators, without taking into account the error involved in using an estimate of $F(x)$. Consequently, their consistency is open to question. In [2], an estimator for $G^{(1)}(x)$ when $F^{(1)}(x)$ is observed is given, and its maximum bias computed when $G^{(1)}(x)$ is a normal probability density. The approximate variance of the estimator when $F^{(1)}(x)$ is subject to error is also given, but the form of the bias shows that the estimator is not consistent.

It is of some interest to examine one of the first solutions to (1.1), given by Eddington [1]. It is

$$\sum_{k=0}^{\infty} (-1)^k F^{(2k)}(x)(\sigma^2/2)^k/k! = G(x), \tag{7.1}$$

which may be thought of as Pollard's formula (2.1) with the limit and summation operations interchanged. In practice, only the first two terms of (7.1) are used as an estimator, and difference quotients are apparently used to approximate the derivatives. However, even if we consider the whole series, and assume the derivatives known, it is clear that the convergence of the solution depends on the form of $G(x)$, so that (7.1) is not consistent for arbitrary $G(x)$.

REFERENCES

[1] A. S. Eddington, "On a formula for correcting statistics for the effects of a known probable error of observation," *Mon. Not. R. Astrom. Soc.*, Vol. 73 (1913), pp. 359–360.

[2] F. D. Kahn, "The correction of observational data for instrumental band width," *Proc. Camb. Philos. Soc.*, Vol. 51 (1955), pp. 519–525.

[3] H. Pollard, "Distribution functions containing a Gaussian factor," *Proc. Amer. Math. Soc.*, Vol. 4 (1953), pp. 578–582.

[4] M. Rosenblatt, "Remarks on some nonparametric estimates of a density function," *Ann. Math. Stat.*, Vol. 27 (1956), pp. 832–837.

[5] G. Szego, *Orthogonal Polynomials*, American Mathematical Society, New York, 1939.

[6] R. J. Trumpler and H. F. Weaver, *Statistical Astronomy*, University of California Press, Berkeley, 1953.

[7] D. V. Widder, *The Laplace Transform*, Princeton University Press, Princeton, 1946.