APPROXIMATE EXPRESSIONS FOR THE CONDITIONAL MEAN AND VARIANCE OVER SMALL INTERVALS OF A CONTINUOUS DISTRIBUTION

By Gunnar Ekman

Institute of Mathematical Statistics, University of Stockholm

1. Summary. Approximate expressions of more or less simple analytical form are derived for the conditional mean and variance over small intervals of a distribution having a probability density of somewhat restricted nature. An alternative formula for the mean is derived. This result is applied to an extremal problem in stratified sampling.

2. Derivation of results. Consider a positive function $f(t)$, defined on some finite interval containing points x and y, and having continuous derivatives to the fourth order. Let us define functions

$$I_i(y, x) = \int_y^x (x - t)^i f(t)\, dt, \qquad \begin{cases} i = 0, 1, 2 \\ a \leqq x, y \leqq b. \end{cases} \tag{1}$$

These functions exist and may be partially integrated, whereby, using the mean value theorem for integrals and writing $(x - y) = k$ for ease of notation, the following identities are obtained:

$$\begin{aligned} I_0(y, x) &= kf + \frac{k^2}{2!}f' + \frac{k^3}{3!}f'' + \frac{k^4}{4!}f''' + \frac{k^5}{5!}f^{(4)}, \\ I_1(y, x) &= \frac{k^2}{2!}f + \frac{k^3}{3!}f' + \frac{k^4}{4!}f'' + \frac{k^5}{5!}f''' + \frac{k^6}{6!}f^{(4)}, \\ I_2(y, x) &= 2\left\{\frac{k^3}{3!}f + \frac{k^4}{4!}f' + \frac{k^5}{5!}f'' + \frac{k^6}{6!}f''' + \frac{k^7}{7!}f^{(4)}\right\}, \end{aligned} \tag{2}$$

where all $f^{(\nu)}$ but $f^{(4)}$ are taken at the point y, $f^{(4)}$ being taken at points θ_i within (x, y). Let us now define three new functions by

$$\begin{aligned} H_1(y, x) = \frac{I_1(y, x)}{I_0(y, x)} = \frac{k}{2}\Bigg\{1 - \frac{1}{6}\Bigg[k \cdot \frac{f'}{f} + \frac{k^2}{2!}\left(\frac{ff'' - (f')^2}{f^2}\right) \\ + \frac{k^3}{3!}\left(\frac{9f'''}{10f} + \frac{3(f')^3}{2f^3} - \frac{5f'f''}{2f^2}\right)\Bigg] + O(k^4)\Bigg\}, \end{aligned} \tag{3}$$

$$\begin{aligned} H_2(y, x) = \frac{I_2(y, x)}{I_0(y, x)} = \frac{k^2}{3}\Bigg\{1 - \frac{k}{4} \cdot \frac{f'}{f} - k^2\left(\frac{7f''}{60f} - \frac{(f')^2}{8f^2}\right) \\ - k^3\left(\frac{f'''}{30f} + \frac{(f')^3}{16f^3} - \frac{f'f''}{10f^2}\right) + O(k^4)\Bigg\}, \end{aligned} \tag{4}$$

Received December 1, 1958; revised April 13, 1959.

$$H_3(y,x) = H_2(y,x) - [H_1(y,x)]^2 = \frac{k^2}{12}\left\{1 + k^2\left(\frac{f''}{30f} - \frac{(f')^2}{12f^2}\right) + k^3\left(\frac{f'''}{60f} + \frac{(f')^3}{12f^3} - \frac{f'f''}{10f^2}\right) + O(k^4)\right\}, \tag{5}$$

where we have imposed yet another condition on $f(t)$, namely $f(y) \neq 0$. Furthermore we have the Taylor expansion

$$\log f(x) - \log f(y) = k \cdot \frac{f'}{f} + \frac{k^2}{2!}\left(\frac{ff'' - (f')^2}{f^2}\right) + \frac{k^3}{3!}\left(\frac{f'''}{f} + \frac{2(f')^3}{f^3} - \frac{3f'f''}{f^2}\right) + O(k^4). \tag{6}$$

Using (6), we have from (3) and (5)

$$\left[H_1(y,x) - \frac{k}{2} + \frac{k}{12} \cdot \log\frac{f(x)}{f(y)}\right] = \frac{-k^4}{720} \cdot R_1(y) + O(k^5), \tag{7}$$

$$\left[H_3(y,x) - \frac{k^2}{12} + \left(\frac{k}{12} \cdot \log\frac{f(x)}{f(y)}\right)^2\right] = \frac{k^4}{360} \cdot \frac{f''(y)}{f(y)} + \frac{k^5}{720} \cdot R_2(y) + O(k^6), \tag{8}$$

where

$$R_1(t) = -\left[\frac{f'''(t)}{f(t)} + \frac{5(f'(t))^3}{(f(t))^3} - \frac{5f'(t)f''(t)}{(f(t))^2}\right],$$
$$R_2(t) = \left[\frac{f'''(t)}{f(t)} - \frac{f'(t)f''(t)}{(f(t))^2}\right]. \tag{9}$$

From the definition of $H_1(y, x)$ and $H_3(y, x)$, by (1) and (3)–(5), we see that the functions $[H_1(y, x) - (k/2)]$ and $H_3(y, x)$ are symmetrical in x and y, so that the same may be said for the left hand members of (7) and (8). This implies that x and y may be interchanged in the right hand members of these identities without changing the order of magnitude of the terms.

Now we see from (1) and (3)–(5), if $\mu(y, x)$ and $\sigma^2(y, x)$ denote the conditional mean and conditional variance respectively over (y, x) of the function $f(t)$ considered as a probability density, that we have

$$H_1(y,x) = x - \mu(y,x), \qquad H_3(y,x) = \sigma^2(y,x),$$

so that we are led from (7) and (8) to the following approximations:

$$\mu(y,x) \sim \frac{(x+y)}{2} + c(y,x) + \frac{(x-y)^4}{720} \cdot R_1(y), \tag{10}$$

$$\sigma^2(y,x) \sim \frac{(x-y)^2}{12} - [c(y,x)]^2 + \frac{(x-y)^4}{360} \cdot \frac{f''(y)}{f(y)} + \frac{(x-y)^5}{720} \cdot R_2(y), \tag{11}$$

where

$$c(y, x) = \frac{(x - y)}{12} \cdot \log \frac{f(x)}{f(y)}, \tag{12}$$

and where $R_1(t)$, $R_2(t)$ are given by (9); x and y may be interchanged in (10) and (11). We note that by taking only values of $f(t)$ at the end points (x, y) into consideration, that is, by neglecting all but the first two terms in (10) and (11), we may approximate μ and σ^2 correctly to $O(x - y)^4$. We call attention to the fact that the logarithms are to the base e, and that the conditions imposed on $f(t)$ are the following: $0 < f(x), f(y) < \infty$, and the first four derivatives of f exist and are continuous; these conditions may be weakened and were imposed for ease of derivation. Finally, it may be remarked that approximations containing $I_0(y, x)$ explicitly may be derived in the case $f(x)$ or $f(y) = 0$.

3. Further results. We shall derive another approximation to $\mu(y, x)$, assuming the existence and continuity of f' and f'' and with $f(x), f(y) \neq 0$. By cubing both sides of (3) we find

$$[x - \mu(y, x)]^3 = \frac{k^3}{8}\left[1 - \frac{k}{2} \cdot \frac{f'}{f} + O(k^2)\right]. \tag{13}$$

On the other hand, using the Taylor expansion

$$f(x) = f + k \cdot f' + O(k^2)$$

together with (2), we obtain from (13)

$$\frac{k^2}{8} \cdot \frac{I_0(y, x)}{f(x)} = \frac{k^3}{8}\left[1 - \frac{k}{2} \cdot \frac{f'}{f} + O(k^2)\right] = [x - \mu(y, x)]^3 + O(k^5). \tag{14}$$

Assuming $x > y$ for definitiveness and writing $I_0(y, x) = P(y, x) =$ the area under $f(t)$ in (y, x), we have from (14)

$$\mu(y, x) = x - \left\{\frac{P(y, x) \cdot (x - y)^2}{8f(x)}\right\}^{\frac{1}{3}} + O(k^3), \tag{15}$$

from which, by permuting x and y, we obtain also

$$\mu(y, x) = y + \left\{\frac{P(y, x) \cdot (x - y)^2}{8f(y)}\right\}^{\frac{1}{3}} + O(k^3). \tag{16}$$

These approximations, (15) and (16), are less accurate than (10), even when the last term of the latter is neglected, but may be used to obtain an approximate solution to the following problem arising in the theory of stratified sampling with proportionate allocation, see [1]. Given a density $f(x)$ over a range (x_0, x_n), $(n - 1)$ variable points x_i, $i = 1, \cdots, (n - 1)$, $x_{i-1} < x_i$, and denoting by P_h, μ_h and σ_h^2 the area, conditional mean and conditional variance in the interval (x_{h-1}, x_h), we are to minimize

$$\sum_{h=1}^{n} P_h \sigma_h^2. \tag{17}$$

In [1] it is shown that the points minimizing (17) satisfy

$$x_h - \mu_h = \mu_{h+1} - x_h , \qquad h = 1, \cdots, (n-1), \tag{18}$$

and it is seen that the determination of such points may give rise to some computational difficulties. Let us now assume $(x_n - x_0)$ finite, $0 < f(x) < \infty$, and that $f'(x)$ and $f''(x)$ exist and are continuous over the whole range. Taking $y = x_{h-1}$ and $x = x_h$ in (15) and $y = x_h$ and $x = x_{h+1}$ in (16), we see that if we neglect terms of order $O(k^3)$, the equations

$$(x_h - x_{h-1})^2 P_h = (x_{h+1} - x_h)^2 P_{h+1} , \qquad h = 1, \cdots, (n-1),$$

that is,

$$(x_h - x_{h-1})^2 P_h = K_n , \qquad h = 1, \cdots, n, \tag{19}$$

where K_n is a constant dependent on $f(x)$ and n, may be substituted for (18). This result may be compared with the approximate solution $(x_h - x_{h-1})P_h = C_n$ given in [2] to the similar problem of minimizing $\sum_1^n P_h\sigma_h$; we see by (11) that (19), and (11) of [2], may be replaced by $P_h\sigma_h^2 = K_n'$ and $P_h\sigma_h = C_n'$, respectively, without affecting the degree of approximation, whereby a certain analogy between the two results is discerned. Proceeding as in [2] we come to the same results as to the respective degree of approximation to the true minimizing values of the points satisfying (19) and the thereof resulting sum (17). We see that when $f(x_0) = 0$ or $x_0 = -\infty$ and/or $f(x_n) = 0$ or $x_n = +\infty$ we may substitute

$$8 f(x_1)\cdot(x_1 - \mu_1)^3 = K_n \quad \text{and/or} \quad = 8 f(x_{n-1})\cdot(\mu_n - x_{n-1})^3 \tag{20}$$

for those equations of (19) with $h = 1$ and/or $h = n$, also that K_n varies with n about as n^{-3}, and that an iterative method of finding K_n may be employed. Finally we note that the methods of the last section of [2] may be used even in the present case, if we put $(x_h - \mu_h) = A_h$ and $(\mu_{h+1} - x_h) = B_{h+1}$, which results in

$$1 + \frac{\partial B_h}{\partial x_{h-1}} = -\frac{\partial A_h}{\partial x_{h-1}} = \frac{f(x_{h-1}) \cdot (\mu_h - x_{h-1})}{P_h} ,$$

$$1 - \frac{\partial A_h}{\partial x_h} = \frac{\partial B_h}{\partial x_h} = \frac{f(x_h) \cdot (x_h - \mu_h)}{P_h} .$$

REFERENCES

[1] Tore Dalenius, "The problem of optimum stratification," *Skand. Aktuarietids.*, Vol. 33 (1950), pp. 202-213.

[2] Gunnar Ekman, "An approximation useful in univariate stratification," *Ann. Math. Stat.*, Vol. 30 (1959), pp. 219-229.