ON THE EXACT VARIANCE OF PRODUCTS*

LEO A. GOODMAN
University of Chicago

A simple exact formula for the variance of the product of two random variables, say, x and y, is given as a function of the means and central product-moments of x and y. The usual approximate variance formula for xy is compared with this exact formula; e.g., we note, in the special case where x and y are independent, that the "variance" computed by the approximate formula is less than the exact variance, and that the accuracy of the approximation depends on the sum of the reciprocals of the squared coefficients of variation of x and y. The case where x and y need not be independent is also studied, and exact variance formulas are presented for several different "product estimates." (The usefulness of exact formulas becomes apparent when the variances of these estimates are compared.) When x and y are independent, simple unbiased estimates of these exact variances are suggested; in the more general case, consistent estimates are presented.

1. INTRODUCTION AND SUMMARY

THE usual formula, which has appeared in the statistical literature, for the variance of the product of two independent random variables is an approximation (see, for example, Yates [3, p. 198]). In this literature, it has also been suggested that this approximate formula for the variance is satisfactory only if the coefficients of variation of the two random variables are both relatively small. In the present note, we shall present a simple exact formula for this variance, which does not depend on any assumptions concerning the magnitudes of the coefficients of variation. The relative accuracy of the usual approximate formula for this variance will be computed. It will be seen that the "variance" obtained using the approximate formula is less than the exact variance, and that the approximation may be satisfactory in some cases where one or both of the coefficients of variation are relatively small. A simple unbiased estimate of the exact variance will also be presented. The exact formula for the variance will then be used to compute the relative efficiency of two different kinds of estimates of a parameter that is in fact equal to the product of two other parameters. (The reader will note that the usual approximate variance formula could not be used to compute correctly the relative efficiency of these two estimates. In fact, if the approximate variance formula had been applied, rather than the exact formula, an erroneous conclusion regarding the relative efficiency of these estimates might have been obtained.) This exact formula will also be generalized to obtain an exact formula for the variance of the product of three (or more) independent random variables. Finally, the situation where the random variables need not be independent will be investigated, and exact

* Research carried out at the Statistical Research Center, University of Chicago, under the sponsorship of the Statistics Branch, Office of Naval Research, and of the Social Science Research Committee, University of Chicago. Reproduction in whole or in part is permitted for any purpose of the United States Government.

I am indebted to R. L. Ashenhurst, H. L. Jones and R. Summers for some helpful comments. Formula (18) was obtained independently by Professor Jones using a somewhat different method from that presented in the present note.

variance formulas will be presented in this case along with consistent estimates of these variances.

The method of obtaining the exact variance formulas, which are presented here, is quite simple. This method can be generalized to obtain exact formulas for any of the central moments of the product of two or more random variables. These formulas could also be derived from general formulas relating product-moments about the origin to central product-moments (see, for example, Kendall and Stuart [1, p. 82] and Tschuprow [2]), but an additional set of calculations would then be required in order to modify the formulas obtained for the product-moments about the origin so that relatively simple formulas could be obtained for the central moments of the product. The present paper presents a more direct method of proof. To the best of my knowledge, the simple exact formulas presented in this paper are new. (It seems surprising that they should not have appeared in print before this.)

There are many situations where the variance of the product of two random variables is of interest (e.g., where an estimate is computed as a product of two other estimates), so that it will not be necessary to describe these situations in any detail in the present note.

2. THE CASE WHERE THE RANDOM VARIABLES ARE INDEPENDENT

Let x and y be two independent random variables. Let us denote the expected value of x by $E(x)=X$, the variance of x by $V(x)$, and the square of the coefficient of variation of x by $V(x)/X^2=G(x)$. A similar notation will be used for the random variable y. (For the sake of simplicity, we shall assume here that $E(x)=X$ and $E(y)=Y$ differ from zero, although some of the results presented here do not require this assumption.) Since

$$xy - XY = XY[(\delta x + 1)(\delta y + 1) - 1] = XY[\delta y + \delta x + \delta x \delta y] \tag{1}$$

where $\delta x=(x-X)/X$ and $\delta y=(y-Y)/Y$, we have that the variance $V(xy)$ of the product xy is equal to

$$\begin{aligned} V(xy) &= E\{(xy - XY)^2\} = (XY)^2[G(y) + G(x) + G(x)G(y)] \\ &= X^2V(y) + Y^2V(x) + V(x)V(y). \end{aligned} \tag{2}$$

The usual approximate formula, which has appeared in the literature, is

$$\tilde{V}(xy) = X^2V(y) + Y^2V(x) = (XY)^2[G(y) + G(x)]. \tag{3}$$

Thus, the relative inaccuracy of the approximation $\tilde{V}(xy)$ is

$$\begin{aligned} R &= [V(xy) - \tilde{V}(xy)]/V(xy) = G(x)G(y)/[G(x) + G(y) + G(x)G(y)] \\ &= 1/[A + 1], \end{aligned} \tag{4}$$

where $A=G^{-1}(x)+G^{-1}(y)$. From (4) we see that, if either $G(x)$ or $G(y)$ is quite small, then A will be relatively large, and the relative inaccuracy will be small. Thus, the approximate formula may be satisfactory even in some cases where only one of the two coefficients is small, contrary to what has usually been suggested in the literature.

We shall now present an unbiased estimate of the variance $V(xy)$. Since $E(x^2)-X^2=V(x)$, we have that

$$\begin{aligned} v(xy) &= [x^2 - v(x)]v(y) + [y^2 - v(y)]v(x) + v(x)v(y) \\ &= x^2v(y) + y^2v(x) - v(x)v(y) \end{aligned} \tag{5}$$

is an unbiased estimate of $V(xy)$, where $v(x)$ is an unbiased estimate of $V(x)$ and $v(y)$ is an unbiased estimate of $V(y)$. It is interesting to note that, while $V(x)V(y)$ is added to the usual approximate formula $\tilde{V}(xy)=X^2V(y)+Y^2V(x)$ to obtain the exact formula for $V(xy)$, the quantity $v(x)v(y)$ is subtracted from the estimate $\bar{v}(xy)=x^2v(y)+y^2v(x)$ to obtain the unbiased estimate $v(xy)$ of $V(xy)$.

Let us now consider the situation where a sample of x's and an independent sample of y's are obtained, and where the sample sizes are $n(x)$ and $n(y)$, respectively. Let $\bar{x}$ and $\bar{y}$ be the sample means of the x's and y's, respectively, and let $s^2(x)$ and $s^2(y)$ be the usual unbiased estimates of $V(x)$ and $V(y)$, respectively. Then $\bar{x}\bar{y}$ will be an unbiased estimate of XY whose variance is

$$\begin{aligned} V(\bar{x}\bar{y}) &= X^2V(\bar{y}) + Y^2V(\bar{x}) + V(\bar{x})V(\bar{y}) \\ &= X^2\frac{V(y)}{n(y)} + Y^2\frac{V(x)}{n(x)} + \frac{V(x)V(y)}{n(x)n(y)}\,. \end{aligned} \tag{6}$$

An unbiased estimate of $V(\bar{x}\bar{y})$ will be

$$\begin{aligned} v(\bar{x}\bar{y}) &= \bar{x}^2v(\bar{y}) + \bar{y}^2v(\bar{x}) - v(\bar{x})v(\bar{y}) \\ &= \bar{x}^2\frac{s^2(y)}{n(y)} + \bar{y}^2\frac{s^2(y)}{n(x)} - \frac{s^2(x)s^2(y)}{n(x)n(y)}\,. \end{aligned} \tag{7}$$

When $n(x)=n(y)=n$, the variance $V(\bar{x}\bar{y})$ becomes simply

$$V(\bar{x}\bar{y}) = [X^2V(y) + Y^2V(x) + V(x)V(y)/n]/n, \tag{8}$$

and the unbiased estimate of $V(\bar{x}\bar{y})$ becomes

$$v(\bar{x}\bar{y}) = [\bar{x}^2s^2(y) + \bar{y}^2s^2(x) - s^2(x)s^2(y)/n]/n. \tag{9}$$

If a sample of n paired observations (x_i, y_i) is obtained $(i=1, 2, \cdots, n)$, then

$$\sum_{i=1}^{n} x_iy_i/n = z$$

is also an unbiased estimate of XY in the special case where x and y are independent random variables. In this case, the variance of z is

$$V(z) = V(xy)/n = [X^2V(y) + Y^2V(x) + V(x)V(y)]/n, \tag{10}$$

and the relative efficiency of the estimate z as compared with the estimate $\bar{x}\bar{y}$ is

$$V(\bar{x}\bar{y})/V(z) = \frac{[X^2V(y) + Y^2V(x) + V(x)V(y)/n]}{[X^2V(y) + Y^2V(x) + V(x)V(y)]}, \tag{11}$$

which approaches

$$\tilde{V}(xy)/V(xy) = [G(y) + G(x)]/[G(y) + G(x) + G(y)G(x)] \tag{12}$$

as $n \to \infty$. Thus, the estimate z is less efficient than the estimate $\bar{y}\bar{x}$ in this case, and the relative decrease in the variance of $\bar{x}\bar{y}$ as compared with the variance of z approaches ($n \to \infty$)

$$1 - [\tilde{V}(xy)/V(xy)] = R = 1/[A + 1], \tag{13}$$

where $A = G^{-1}(x) + G^{-1}(y)$ as earlier herein. Thus, we see that, when x and y are independent, the effect of using the usual approximate formula $\tilde{V}(xy)$ rather than the exact formula $V(xy)$ is comparable to the effect of using the statistic z rather than $\bar{x}\bar{y}$ as an estimate of the product XY of the parameters X and Y (when $n \to \infty$).

The preceding results can be generalized to obtain exact formulas in the situation where the product of three (or more) independent random variables is of interest. For example, let the three random variables be x, y, and z, where X, Y, and Z are their respective means, $V(x)$, $V(y)$, $V(z)$ are their respective variances, and $G(x)$, $G(y)$, $G(z)$ are their respective squared coefficients of variation. Since

$$\begin{aligned} xyz - XYZ &= XYZ[(\delta x + 1)(\delta y + 1)(\delta z + 1) - 1] \\ &= XYZ[\delta x + \delta y + \delta z + \delta x\delta y + \delta x\delta z + \delta y\delta z + \delta x\delta y\delta z], \end{aligned} \tag{14}$$

where the δ's are defined as earlier herein, we have that the variance $V(xyz)$ of the product xyz is equal to

$$\begin{aligned} V(xyz) &= E\{(xyz - XYZ)^2\} \\ &= (XYZ)^2[G(x) + G(y) + G(z) + G(x)G(y) + G(x)G(z) + G(y)G(z) \\ &\qquad + G(x)G(y)G(z)]. \end{aligned} \tag{15}$$

An approximate formula for this variance (comparable to the usual approximate formula for the variance of the product of two independent random variables) would be

$$\tilde{V}(xyz) = (XYZ)^2[G(x) + G(y) + G(z)], \tag{16}$$

which we now see will be satisfactory only if the term

$$(XYZ)^2[G(x)G(y) + G(x)G(z) + G(y)G(z) + G(x)G(y)G(z)]$$

can be neglected. Formula (15) is a generalization of (2), while formula (16) is a generalization of (3).

3. THE CASE WHERE THE RANDOM VARIABLES NEED NOT BE INDEPENDENT

Let x and y be two random variables (not necessarily independent). Let us denote the expected value of xy by $E(xy) = M_{11}$ and the covariance between δx and δy by $E\{\delta x\delta y\} = D_{11}$. We also write $E\{(\delta x)^i(\delta y)^j\} = D_{ij}$ and $E\{(\Delta x)^i(\Delta y)^j\} = E_{ij}$, where $\Delta x = x - X$ and $\Delta y = y - Y$. Since

$$\begin{aligned} xy - M_{11} &= XY[(\delta x + 1)(\delta y + 1) - B_{11}], \\ &= XY[\delta x + \delta y + \delta x\delta y + F_{11}], \end{aligned} \tag{17}$$

where $B_{11}=M_{11}/(XY)$ and $F_{11}=1-B_{11}=-D_{11}$, we see that the variance $V(xy)$ of the product xy is equal to

$$\begin{aligned} V(xy) &= (XY)^2[G(y)+G(x)+2D_{11}+2D_{12}+2D_{21}+D_{22}-D_{11}^2] \\ &= X^2V(y)+Y^2V(x)+2XYE_{11}+2XE_{12}+2YE_{21}+E_{22}-E_{11}^2, \end{aligned} \tag{18}$$

where $E_{22}-E_{11}^2=V(\Delta x\Delta y)$ is the variance of $\Delta x\Delta y$. The usual approximate formula for $V(xy)$ is

$$\tilde{V}(xy) = X^2V(y)+Y^2V(x)+2XYE_{11}, \tag{19}$$

which we now see will be satisfactory only if the term $2X_{12}+2YE_{21}+V(\Delta x\Delta y)$ can be neglected.

If a sample of n paired observations (x_i, y_i) are obtained $(i=1, 2, \cdots, n)$, then

$$\sum_{i=1}^{n} x_iy_i/n = z$$

is an unbiased estimate of M_{11} and $\bar{x}\bar{y}$ is a consistent estimate of XY. It is easy to see that the expected value of $\bar{x}\bar{y}$ is $E\{\bar{x}\bar{y}\}=XY(1-1/n)+M_{11}/n$ $=XY+E_{11}/n$, so that the statistic $(\bar{x}\bar{y}n-z)/(n-1)=w$ is an unbiased estimate of XY. The variance of z is

$$\begin{aligned} V(z) &= \frac{V(xy)}{n} \\ &= [X^2V(y)+Y^2V(x)+2XYE_{11}+2XE_{12}+2YE_{21}+V(\Delta x\Delta y)]/n, \end{aligned} \tag{20}$$

while the variance of $\bar{x}\bar{y}$ is

$$\begin{aligned} V(\bar{x}\bar{y}) = \Bigg[& X^2V(y)+Y^2V(x)+2XYE_{11}+2X\frac{E_{12}}{n}+2Y\frac{E_{21}}{n} \\ & +\frac{V(x)V(y)}{n}+\frac{\mathrm{Cov}[(\Delta x)^2,(\Delta y)^2]-E_{11}^2}{n^2}\Bigg]\Big/ n, \end{aligned} \tag{21}$$

where Cov $[(\Delta x)^2, (\Delta y)^2]=E_{22}-V(x)V(y)$ is the covariance between the random variables $(\Delta x)^2$ and $(\Delta y)^2$. The mean squared error of $\bar{x}\bar{y}$ as an estimate of XY is

$$MSE(\bar{x}\bar{y}) = V(\bar{x}\bar{y})+(E_{11}/n)^2. \tag{22}$$

Since the estimate w of XY is asymptotically equivalent to $\bar{x}\bar{y}$, the variance $V(w)$ of w can be simply approximated using the fact that the limiting value $(n\rightarrow\infty)$ of $nV(w)$ is equal to the limiting value of $nV(\bar{x}\bar{y})$ and $nMSE(\bar{x}\bar{y})$. Thus, we have that

$$V(w) \cong [X^2V(y)+Y^2V(x)+2XYE_{11}]/n = \tilde{V}(xy)/n. \tag{23}$$

Consistent estimates of $V(z)$, $V(\bar{x}\bar{y})$, $MSE(\bar{x}\bar{y})$, and $V(w)$ can be obtained by replacing the various population moments by the corresponding sample mo-

ments (which will be consistent estimates of the corresponding population moments) in equations (20), (21), (22), and (23), respectively.

The ratio of $V(\bar{x}\bar{y})$ to $V(z)$ approaches

$$\tilde{V}(xy)/V(xy) = \frac{[X^2V(y) + Y^2V(x) + 2XYE_{11}]}{[X^2V(y) + Y^2V(x) + 2XYE_{11} + 2XE_{12} + 2YE_{21} + V(\Delta x \Delta y)]} \tag{24}$$

as $n \to \infty$. We also see that $\tilde{V}(xy)/V(xy)$ is the limiting value for $MSE(\bar{x}\bar{y})/V(z)$ and $V(w)/V(z)$. Thus, the effect of using the usual approximate formula $\tilde{V}(xy)$ rather than the exact formula $V(xy)$ is comparable to the relative difference between the variance of the estimate z of M_{11} and the variance of the estimate w (or $\bar{x}\bar{y}$) of XY (when $n \to \infty$).

The preceding results can be generalized to deal with the situation where the product of three (or more) random variables is of interest (where these random variables need not necessarily be independent). The method of obtaining these results can also be generalized in order to obtain exact formulas for any of the central moments of the product of two or more random variables. These results can also be directly generalized to deal with situations where the ratio of random variables or the product of powers of random variables are of interest.

REFERENCES

[1] Kendall, Maurice G. and Stuart, Alan, *The Advanced Theory of Statistics*, Vol. 1, London: Charles Griffin and Co. Ltd., 1958.

[2] Tschuprow, Al. A., "On the mathematical expectation of the moments of frequency distributions in the case of correlated variables," *Metron*, 2 (1922), pp. 461–500.

[3] Yates, Frank, *Sampling Methods for Censuses and Surveys*, Second Edition, London: Charles Griffin and Co. Ltd., 1953.