Note on a Result by H. F. Smith

BU-80-M

D. S. Robson

October, 1956

In the paper entitled "An empirical law describing heterogeneity in the yields of agricultural crops", H. F. Smith has a result on page 8 which is not immediately obvious. The result referred to is:

$$\text{"}\left\{ \delta(\log_e s^2) \right\}^2 = \left\{ \frac{2s\delta s}{s^2} \right\}^2 = \frac{2}{n} ,$$

where $\delta s = s/\sqrt{2n}$ ...".

This result appears to be obtained by application of the following method for approximating the variance of a function f(X) of a chance variable X. Let μ_x be the mean value of the chance variable X and denote the deviation of X from its mean value by ϵ_X, so that $X = \mu_x + \epsilon_X$. Similarly, let μ_f be the mean value of the function f(X) and $\epsilon_{f(X)} = f(X) - \mu_f$. Now the Taylor series expansion of $f(X) = f(\mu_x + \epsilon_X)$ about the point μ_x is

$$f(\mu_x+\epsilon_X) = f(\mu_x) + f'(\mu_x)\,\epsilon_X + f''(\mu_x)\,\frac{\epsilon_X^2}{2!} + f'''(\mu_x)\,\frac{\epsilon_X^3}{3!} + \ldots.$$

where $f^{\upsilon}(\mu_x)$ is the υ'th derivative of f at the point μ_x. The error $\epsilon_{f(X)}$ therefore has the expansion

(1) $$\epsilon_{f(X)} = f(\mu_x+\epsilon_X) - \mu_f = [f(\mu_x)-\mu_f] + f'(\mu_x)\epsilon_X + f''(\mu_x)\,\frac{\epsilon_X^2}{2!} + \ldots$$

The approximation then consists of dropping all terms on the right side except $f'(\mu_x)\epsilon_X$, giving

(2) $$\epsilon_{f(X)} \sim f'(\mu_x)\epsilon_X,$$

which is exact only when f(X) is a linear function of X. This gives as an approximation for the variance

(3) $$\sigma_f^2 = E\epsilon_{f(X)}^2 \sim [f'(\mu_x)]^2 E\epsilon_X^2 = [f'(\mu_x)]^2\sigma_X^2 .$$

In the above example the chance variable X is a sample variance,

$$X = s^2 = \frac{\sum^{n}(Y_i-\bar{y})^2}{n-1} ,$$

and $f(s^2)$ is the natural logarithm,

$$f(X) = f(s^2) = \log_e s^2 .$$

Assuming $Y_1,\ldots,Y_n$ are normal, independent, and identdally distributed with variance σ_Y^2 we get

$$\mu_x = Es^2 = \sigma_Y^2$$

$$\epsilon_X = s^2 - \sigma_Y^2$$

$$\sigma_X{}^2 = E\epsilon_X{}^2 = \frac{2}{n-1}\,\sigma_Y{}^4$$

$$f'(\mu_X) = \frac{d(\log_e\ \sigma_Y{}^2)}{d(\sigma_Y{}^2)} = \frac{1}{\sigma_Y{}^2}$$

so that

$$\sigma_f{}^2 \sim \left[f'(\mu_X)\right]^2 \sigma_X{}^2 = \left[\frac{1}{\sigma_Y{}^2}\right]^2 \frac{2}{n-1}\,\sigma_Y{}^4 = \frac{2}{n-1}$$

This is essentially the answer arrived at by Smith, but we cannot be certain that this is the method that he used. His assertation

$$(4)\quad \delta\ \log_e\ \sigma_Y{}^2 = \frac{2\sigma_Y}{\sigma_Y{}^2}\ \delta\sigma_Y$$

suggests that he intends δ to be the operator $d/d\sigma_Y$, but if that's the case then $\delta\sigma_Y = 1$, contrary to his second assertion

$$\delta\sigma_Y = \frac{\sigma_Y}{\sqrt{2n}}\ .$$

Because of the resemblance in (4) to the operation of differentiation we conclude that the Taylor series approximation outlined above was originally employed to give the answer $2/n$ and that the derivation given in Smith's paper is erroneous.

As indicated earlier, this method for approximating the variance of $f(X)$ is exact only when f is a linear function of X. Thus, when

$$f(X) = aX + b$$

$$\mu_f = a\mu_X + b = f(\mu_X)$$

$$f'(\mu_X) = a$$

$$f^{\upsilon}(\mu_X) = 0 \quad \text{for } \upsilon > 1$$

so that, from (1)

$$\epsilon_{f(X)} = f(\mu_X+\epsilon_X) - f_\mu = f'(\mu_X)\epsilon_X$$

giving in place of (3) the equality

$$\sigma_f{}^2 = E\epsilon^2_{f(X)} = [f'(\mu_X)]^2\ E\epsilon_X{}^2 = a^2\sigma_X{}^2\ .$$

The errors of approximation committed when $f(X)$ is a second degree polynomial in X is easily computed. Let

$$f(X) = aX^2 + bX + c$$

then

$$\mu_f = a(\sigma_x^2+\mu_x^2) + b\mu_x + c$$

$$f(\mu_x) - \mu_f = a\sigma_x^2$$

$$f'(\mu_x) = 2a\mu_x + b$$

$$f''(\mu_x) = 2a$$

$$f^{(\nu)}(\mu_x) = 0 \quad \text{for } \mu > 2 .$$

Hence, by (1),

$$\epsilon_{f(X)} = a\sigma_x^2 + (2a\mu_x+b)\epsilon_X + (2a)\frac{\epsilon_X^2}{2!}.$$

so the variance σ_f^2 of $f(X)$ is

$$\sigma_f^2 = a^2\sigma_x^4 + (2a\mu_x+b)^2\sigma_x^2 + a^2E\epsilon_X^4 + 2a^2\sigma_x^4 + 2a(2a\mu_x+b)E\epsilon_X^3$$

while the approximation (3) gives

$$\sigma_f^2 \sim (2a\mu_x+b)^2 \sigma_x^2 .$$

Thus, if X is normally distributed, the approximation underestimates the true variance by an amount $6a^2\sigma_x^4$. Clearly, this method of approximation must be used with caution; it is always necessary to verify that the terms being ignored are truly negligible. When X is the maximum likelihood estimator of μ_x, based upon a sample of size n, then the above approximation may be expected to improve as n increases.