THE STANDARD ERROR OF GINI'S MEAN DIFFERENCE

By Z. A. Lomnicki

Polish University College, London

A general expression for the standard error of Gini's mean difference g was given in a paper under the same title by U. S. Nair [1]. See also [2], pp. 216–217.

The object of this note is to deduce in a more direct way a simpler formula for the variance of this statistic. The expression obtained is equivalent to that given by Nair except for an additional term overlooked in his final formula. The simplification is due to the fact that, for the evaluation of the expected values of g and g^2, it is not necessary to arrange the sample values in ascending order of magnitude as done by Nair.

Let n be the size of the sample, $f(x)$ the probability density function of the parent population, μ the mean and σ^2 the variance of x in the parent and let

$$F(x) = \int_{-\infty}^{x} f(t)\,dt, \qquad Z(x) = \int_{-\infty}^{x} tf(t)\,dt.$$

From the definition

$$g = \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j=i+1}^{n} |x_i - x_j| \tag{1}$$

(where the values x_i are not in order of magnitude but are numbered as they appear in the sample), we have

$$\begin{aligned} E(g) &= \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j=i+1}^{n} E(|x_i - x_j|) \\ &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} |x - y| f(x)f(y)\,dx\,dy = \Delta, \end{aligned} \tag{2}$$

where Δ is the mean difference (parameter) of the parent population. It is easy to check that Δ can also be written

$$\Delta = 2\int_{-\infty}^{+\infty} \{xF(x) - Z(x)\}f(x)\,dx = 2\int_{-\infty}^{\infty} xf(x)(2F(x) - 1)\,dx. \tag{3}$$

In order to find $E(g^2)$ let us write

$$g^2 = \frac{4}{n^2(n-1)^2} \{\sum (x_i - x_j)^2 + 2\sum |x_i - x_j||x_i - x_k| + 2\sum |x_i - x_j||x_k - x_l|\}. \tag{4}$$

The first sum should be read as the double sum extended to all pairs of different subscripts i, j, and has $n(n-1)/2$ terms; the second as a triple sum extended to all combinations of two pairs (i, j), (i, k) of different subscripts i, j, k and has $n(n-1)(n-2)/2$ terms; the third as a quadruple sum extended to all com-

binations of two pairs (i, j), (k, l) of different indices i, j, k, l and has $n(n-1)(n-2)(n-3)/8$ terms. Thus

$$E(g^2) = \frac{1}{n(n-1)} \{2E(x_i - x_j)^2 + 4(n-2)E(|x_i - x_j||x_i - x_k|) + (n-2)(n-3)E(|x_i - x_j||x_k - x_l|)\}. \tag{5}$$

The first expected value is equal to $2\sigma^2$; the third to Δ^2. Denoting the second by J we have

$$\operatorname{var}(g) = E(g^2) - \Delta^2 = \frac{1}{n(n-1)} (4\sigma^2 + 4(n-2)J - 2(2n-3)\Delta^2), \tag{6}$$

where

$$J = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} |x - y||x - z| f(x)f(y)f(z)\, dx\, dy\, dz. \tag{7}$$

This can be written as

$$\begin{aligned} J = \int_{-\infty}^{\infty} f(x) \Big\{ &\int_{-\infty}^{x}\int_{-\infty}^{x} (x-y)(x-z)f(y)f(z)\, dy\, dz \\ &+ \int_{-\infty}^{x}\int_{x}^{\infty} (x-y)(z-x)f(y)f(z)\, dy\, dz \\ &+ \int_{x}^{\infty}\int_{-\infty}^{x} (y-x)(x-z)f(y)f(z)\, dy\, dz \\ &+ \int_{x}^{\infty}\int_{x}^{\infty} (y-x)(z-x)f(y)f(z)\, dy\, dz \Big\}\, dx. \end{aligned} \tag{8}$$

Putting

$$G(x) = \int_{-\infty}^{x} (x-y)f(y)\, dy = xF(x) - Z(x), \tag{9}$$

$$H(x) = \int_{x}^{\infty} (y-x)f(y)\, dy = G(x) + \mu - x, \tag{10}$$

we obtain

$$\begin{aligned} J &= \int_{-\infty}^{\infty} [G^2(x) + 2G(x)H(x) + H^2(x)]f(x)\, dx \\ &= \int_{-\infty}^{\infty} [G(x) - H(x)]^2 f(x)\, dx + 4\int_{-\infty}^{\infty} G(x)H(x)f(x)\, dx, \end{aligned} \tag{11}$$

and finally

$$\operatorname{var}(g) = \frac{1}{n(n-1)} \{4(n-1)\sigma^2 + 16(n-2)I - 2(2n-3)\Delta^2\}, \tag{12}$$

where

$$I = \int_{-\infty}^{\infty} G(x)H(x)f(x)\,dx$$

$$= \int_{-\infty}^{\infty} \{[xF(x) - Z(x)]^2 + (\mu - x)[xF(x) - Z(x)]\}f(x)\,dx. \tag{13}$$

This integral can also be written as

$$I = \int_{-\infty}^{\infty}\int_{-\infty}^{x}\int_{x}^{\infty} (x - y)(z - x)f(x)f(y)f(z)\,dx\,dy\,dz, \tag{14}$$

and, according to the distribution involved, formula (13) or (14) may be more convenient in the evaluation of var (g).

Comparing (12) with the formulae given by Nair it is easy to show that an additional term $(n - 3)\mu^2$ has been omitted in his final formula for I_1. However, the values of var (g) for normal, exponential and rectangular distributions given in [1] are correct and agree with those obtained from formula (12) above.

REFERENCES

[1] U. S. NAIR, "The standard error of Gini's mean difference," *Biometrika*, Vol. 28 (1936), pp. 428–436.

[2] M. G. KENDALL, *Advanced Theory of Statistics*, Vol. I, Charles Griffin and Co., London 1943.

CORRECTION TO "A NOTE ON THE POWER OF A NONPARAMETRIC TEST"

BY F. J. MASSEY, JR.

University of Oregon

In the paper mentioned in the title (*Annals of Math. Stat.*, Vol. 21 (1950), pp. 440–443) the proof of the biasedness of a test based on the maximum deviation between sample and population cumulatives is incorrect. A proof is given below. Also, on page 442, line 2, "greater" should be replaced by "less". The notation refers to Fig. 1 of the original article.

Above point b (note $F_1(b) = F_0(b)$), there will be certain possible heights for $S_n(x)$ to attain and still remain in the band. Call these heights $b_1 = 1/n$, $b_2, b_3, \cdots, b_k = k/n$, where $k/n < 2d/\sqrt{n}$. Locate the point $x = c$ $(c < b)$ close enough to $x = b$ so that $F_0(c) + d/\sqrt{n} > b_k$. Then consider

(i) $$P_0 = P\{S_n(x) \text{ remain in band} \mid F(x) = F_0(x)\},$$

(ii) $$P_1 = P\{S_n(x) \text{ remain in band} \mid F(x) = F_1(x)\}.$$

Now $P_j = \sum_{i=1}^{k} P\{S_n(x)$ passes through b_i and remains in band $\mid F_j(x)\} = \sum_{i=1}^{k} P\{S_n(x)$ goes through $b_i \mid F_j(x)\} \cdot P\{S_n(x)$ stays in band for $x < b \mid F_j(x)$,