The Mean Absolute Difference between Correlated Variables

1. Professor DeFries's Derivation

The mean absolute difference (MAD) is the average of the absolute difference between two variables each from two populations. With the definition of probability density and cumulative distribution functions as $f(x)$ and $F(x)$, and $g(y)$ and $G(y)$ for two independent populations respectively, MAD can be expressed as:

$$\begin{aligned}
MAD(x,y) &= \int_a^b dx \int_a^b dy \left| x-y \right| f(x)g(y) \\
&= \int_a^b dx \int_a^x dy (x-y) f(x)g(y) + \int_a^b dx \int_x^b dy (y-x) f(x)g(y) \\
&= \int_a^b xG(x) f(x)dx - \int_a^b \left[\int_a^x yg(y)dy \right] f(x)dx \\
&\quad + \int_a^b \left[\int_x^b yg(y)dy \right] f(x)dx - \int_a^b x\left(1-G(x)\right) f(x)dx \\
&= \int_a^b xG(x) f(x)dx - \left[\int_a^b yg(y)dy - \int_a^b xg(x)F(x)dx \right] \\
&\quad + \int_a^b xg(x)F(x)dx - \left[\int_a^b xf(x)dx - \int_a^b xG(x) f(x)dx \right] \\
&= 2\left[\int_a^b xG(x) f(x)dx + \int_a^b xg(x)F(x)dx \right] - \left[\int_a^b yg(y)dy + \int_a^b xf(x)dx \right]
\end{aligned}$$

Nair (1936) showed:

$$MAD = \frac{2}{\sqrt{\pi}}\sigma \cong 1.13\sigma$$

if variables are chosen from two i.i.d. zero-mean normal populations (or the same population)[1].

Plomin and DeFries (1980) derived MAD when two zero-mean normal variables are correlated as:

$$MAD = 1.13\sigma\sqrt{1-r}\ .$$

Although the derivation process was omitted in their article, Professor DeFries has kindly sent me an unpublished note used then, on which the following is based.

Suppose we choose two variables, x and y, from the population with mean μ, variance σ^2,

---

[1] Table I (p. 433) has a typo. Normal distribution's first row should be $\frac{\sigma}{2\sqrt{\pi}}$ instead of $\frac{\sigma}{2\sqrt{2\pi}}$.

and correlation coefficient *r*. Then, the mean and variance of their (not absolute) difference

$$z = x - y$$

are:

$$\mu_z = 0, \quad \sigma_z^2 = 2(1-r)\sigma^2 .$$

If we further assume normality

$$f(z) = \frac{1}{\sqrt{2\pi \cdot 2(1-r)\sigma^2}} e^{-\frac{z^2}{2\cdot 2(1-r)\sigma^2}}$$

taking into the fact that its distribution is symmetric around zero and

$$f'(z) = -\frac{z}{2(1-r)\sigma^2} f(z)$$

we obtain the normal correlated version of MAD as:

$$\begin{aligned} MAD(x,y) == & \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dy \left|x-y\right| f(x)g(y) \\ & = 2\int_0^{\infty} z f(z) dz \\ & = -4(1-r)\sigma^2 \int_0^{\infty} f'(z) dz \\ & = -4(1-r)\sigma^2 \cdot \frac{1}{\sqrt{2\pi \cdot 2(1-r)\sigma^2}} e^{-\frac{z^2}{2\cdot 2(1-r)\sigma^2}} \Bigg|_0^{\infty} \\ & = \frac{2}{\pi}\sigma\sqrt{1-r} \cong 1.13\sigma\sqrt{1-r}. \end{aligned}$$

As expected, this formula coincides with the unrelated one if the correlation coefficient *r* equals zero, that is, uncorrelated.

2. My Derivation

I will derive the correlated case in a more "formal" fashion following Lomnicki's (1952) exposition of the independent case.

Suppose two zero-mean normal variables with the same variance, *x* and *y*, have a correlation coefficient *r*. Then,

$$y = rx + u$$

$$\mu_x = \mu_y = 0, \quad \sigma_x^2 = \sigma_y^2 = \sigma^2, \quad \sigma_{rx}^2 = r^2\sigma^2, \quad \sigma_u^2 = \sqrt{1-r^2} \cdot \sigma^2$$

If we define

$$z = (1-r)x$$

the absolute difference can be expressed as:

$$|x-y| = |x-rx-u| = |(1-r)x-u| = |z-u|$$

with

$$f(z) = \frac{1}{\sqrt{2\pi}\cdot(1-r)\sigma} e^{-\frac{z^2}{2(1-r)^2\sigma^2}}, \quad g(u) = \frac{1}{\sqrt{2\pi}\sqrt{1-r^2}\cdot\sigma} e^{-\frac{u^2}{2(1-r^2)\sigma^2}}.$$

Using the following relations

$$f'(z) = -\frac{z}{(1-r)^2\sigma^2} f(z), \quad g'(u) = -\frac{u}{(1-r^2)\sigma^2} g(u)$$

we obtain the normal correlated version of MAD as:

$$\begin{aligned}
MAD(x,y) = MAD(z,u) &= \int_{-\infty}^{\infty} dz \int_{-\infty}^{\infty} du\,|z-u|\, f(z)g(u) \\
&= 2\left[\int_{-\infty}^{\infty} zG(z)f(z)dz + \int_{-\infty}^{\infty} zg(z)F(z)dz\right] \\
&= -2\sigma^2\left[(1-r)^2\int_{-\infty}^{\infty} G(z)f'(z)dz + (1-r^2)\int_{-\infty}^{\infty} g'(z)F(z)dz\right] \\
&= 2\sigma^2\left[(1-r)^2\int_{-\infty}^{\infty} g(z)f(z)dz + (1-r^2)\int_{-\infty}^{\infty} g(z)f(z)dz\right] \\
&= 4(1-r)\sigma^2\int_{-\infty}^{\infty} g(z)f(z)dz \\
&= \frac{2\sqrt{2}}{\sqrt{\pi}\sqrt{1-r^2}}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}\left(\frac{2}{(1-r)(1-r^2)\sigma^2}\right)} dz \\
&= \frac{2\sqrt{2}}{\sqrt{\pi}\sqrt{1-r^2}}\cdot\sqrt{\frac{(1-r)(1-r^2)\sigma^2}{2}}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}\sqrt{\frac{(1-r)(1-r^2)\sigma^2}{2}}} e^{-\frac{z^2}{2}\left(\frac{2}{(1-r)(1-r^2)\sigma^2}\right)} dz \\
&= \frac{2}{\sqrt{\pi}}\sigma\sqrt{1-r} \cong 1.13\sigma\sqrt{1-r}
\end{aligned}$$

Although my derivation is a sledgehammer-to-crack-a-nut compared to Professor DeFries's much more elegant as well as simpler one, mine may have some merit to evaluate the degree of approximation when normality and other assumptions are relaxed.

Last but not least I thank Professor DeFries for his generosity to share his great idea with an inquisitive obscure person (myself).

References

Lomnicki, Z. A. 1952. The Standard Error of Gini's Mean Difference. *Annals of Mathematical Statistics* 23 (4): 635-637.

Nair, U. S. 1936. The Standard Error of Gini's Mean Difference. *Biometrika* 28 (3-4): 428-436

Plomin, R., and J. C. DeFries. 1980. Genetics and Intelligence: Recent Data. *Intelligence* 4 (1): 15-24.

Yoshitaka Fukui, ABS, Aoyama Gakuin University

10/18/2013, revised 10/21/2013