Supplement: A Simple, Interpretable Conversion from Pearson's Correlation to Cohen's d for Meta-Analysis

1. Derivation of point estimate d

Suppose that $Y = \beta_0 + \beta X + \epsilon$ with homoskedastic errors, $X \perp\!\!\!\perp \epsilon$, and $E[\epsilon] = 0$. Let s_X^2 and ${}^2_Y$ denote marginal sample variances, let $s_{Y|X}^2$ denote the conditional sample variance of Y, and let r denote the sample correlation between Y and X. Define the Cohen's d of interest as the increase in Y associated with a Δ-unit increase in X, taking c to be an arbitrary constant:

$$d = \frac{E[Y \mid X = c + \Delta] - E[Y \mid X = c]}{s_{Y|X}}$$

By homoskedasticity, $s_{Y|X}^2 = E[s_{Y|X}^2]$, yielding:

$$= \frac{E[Y \mid X = c + \Delta] - E[Y \mid X = c]}{\sqrt{s_Y^2 - \text{Var}\left(E[Y|X]\right)}}$$

$$= \frac{E[Y \mid X = c + \Delta] - E[Y \mid X = c]}{\sqrt{s_Y^2 - \widehat{\beta}^2 s_X^2}}$$

$$= \frac{\widehat{\beta}\Delta}{s_Y\sqrt{1 - r^2}}$$

$$= \frac{r\Delta}{(s_X^2)^{1/2}\sqrt{1 - r^2}}$$

2. Derivation of standard error for d

Let ρ denote the population correlation, and let σ_X and σ_Y respectively denote the population standard deviations of X and Y. We first develop a supporting lemma establishing the asymptotic independence of the sample estimates r and s_X.

Lemma 2.1. *Suppose that r and s_X are estimated in the same sample of size N. Assume the distribution of (X_i, Y_i) satisfies the following regularity condition. Namely, letting*

$$\kappa_{ab} = E\left[\left(\frac{X_i}{\sigma_X}\right)^a \left(\frac{Y_i}{\sigma_Y}\right)^b\right]$$

denote the ab^{th} mixed standardized moment, we assume that κ_{31}, κ_{22}, and κ_{40} are all finite (which we will show to hold for the bivariate normal distribution). Then $lim_{N\to\infty} Cov(r, s_X) = 0$.

Proof. Let $(V_1, \ldots, V_N)$ be an independently and identically distributed sample from a bivariate distribution $V_i = [X_i, Y_i]'$ fulfilling the above regularity condition, and let

$$\Sigma = \mathrm{Cov}(V_i) = \begin{bmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{bmatrix}$$

Without loss of generality, assume $E[V_i] = \mathbf{0}$. The sample covariance

$$S_N = \frac{1}{N-1}\sum_{i=1}^{N}(V_i - \bar{V}_N)(V_i - \bar{V}_N)' = \frac{1}{N-1}\sum_{i=1} V_i V_i' - \frac{N}{N-1}\bar{V}_N\bar{V}_N'$$

is asymptotically unbiased and consistent for Σ, and in fact

$$\sqrt{N}(S_N - \Sigma) \to_D N(\mathbf{0}, \Lambda)$$

for some asymptotic covariance matrix, Λ. We now focus on determining this matrix. First, for a symmetric matrix $\mathbf{A} = \begin{bmatrix} a & b \\ b & c \end{bmatrix}$, let $\widetilde{A} = [a, b, c]'$ denote the "vectorization" of its upper triangle. Now consider a single element of the average that enters into the first term (i.e., the scatter matrix) of S_N:

$$\widetilde{Z}_i = \widetilde{V_i V_i'} = \begin{bmatrix} X_i^2 \\ X_i Y_i \\ Y_i^2 \end{bmatrix}$$

The covariance matrix of these individual squared observations and their cross-product can be expressed in terms of the mixed standardized moments, κ_{ab}, as:

$$\Lambda = \mathrm{Cov}(\widetilde{Z}_i) = E[\widetilde{Z}_i \widetilde{Z}_i{}'] - E[\widetilde{Z}_i]E[\widetilde{Z}_i]' = \mathbf{K} - \widetilde{\Sigma}\widetilde{\Sigma}' \tag{2.1}$$

where

$$\mathbf{K} = \begin{bmatrix} \kappa_{40}\sigma_X^4 & \kappa_{31}\sigma_X^3\sigma_Y & \kappa_{22}\sigma_X^2\sigma_Y^2 \\ \kappa_{31}\sigma_X^3\sigma_Y & \kappa_{12}\sigma_X\sigma_Y^2 & \kappa_{13}\sigma_X\sigma_Y^3 \\ \kappa_{22}\sigma_X^2\sigma_Y^2 & \kappa_{13}\sigma_X\sigma_Y^3 & \kappa_{04}\sigma_Y^4 \end{bmatrix}$$

We now turn to estimating $\mathrm{Cov}(r, s_X)$ using Equation (2.1) and the delta method. Define the transformation $g(x, y, z) = \left(x, \frac{y}{\sqrt{x}\sqrt{z}}\right)$ on $\widetilde{S}_N$. This provides the bivariate distribution of (s_x, r). By the delta method and the asymptotic normality of S_N:

$$\sqrt{N}\left(g(\widetilde{S_N}) - (\rho, \sigma_X^2)'\right) \to_D N\left(\mathbf{0}, \mathbf{J}(\widetilde{\Sigma})'\tilde{\Lambda}\mathbf{J}(\widetilde{\Sigma})\right) \tag{2.2}$$

where $\mathbf{J}(\widetilde{\Sigma})$ is the Jacobian of g evaluated at $(\sigma_X^2, \rho\sigma_X\sigma_Y, \sigma_Y^2)$, which is equal to:

$$\mathbf{J}(\widetilde{\Sigma}) = \begin{bmatrix} 1 & -\frac{\rho}{2\sigma_X^2} \\ 0 & \frac{1}{\sigma_X\sigma_Y} \\ 0 & -\frac{\rho}{2\sigma_Y^2} \end{bmatrix}$$

From Equation (2.2), we thus have:

$$\widetilde{\Lambda} \approx \mathbf{J}(\widetilde{\Sigma})'\left[\mathbf{K} - \widetilde{\Sigma}\widetilde{\Sigma}'\right]\mathbf{J}(\widetilde{\Sigma})$$

In terms of the mixed standardized moments of the bivariate distribution of (X_i, Y_i), the entry of interest simplifies to:

$$N \times \mathrm{Cov}(r, s_X) \approx \widetilde{\Lambda}_{22} \approx \sigma_X^2\left(\kappa_{31} - \frac{\rho\,(\kappa_{22} + \kappa_{40})}{2}\right)$$

Thus, $\lim_{N\to\infty} \mathrm{Cov}(r, s_X) = 0$ for any bivariate distribution satisfying the regularity condition. For example, for the bivariate normal distribution, applying Isserlis' Theorem regarding mixed standardized moments[1] yields $k_{31} = 3\rho$, $k_{22} = 1 + 2\rho^2$, and $k_{40} = 3$, so the regularity condition holds and $\lim_{N\to\infty} \mathrm{Cov}(r, s_X) = 0$. □

We now derive an approximate standard error of d using the delta method. Let $z_f = \mathrm{arctanh}(\rho)$ denote the Fisher-transformed population correlation and $\widehat{z}_f$ its sample estimate, which is asymptotically normal with variance $\frac{1}{N-3}$. Suppose s_X^2 is estimated in a sample of size N_s, such that $N_s = N$ if s_X^2 is estimated in the same sample used to estimate r. Assuming that

X is approximately normal, then asymptotically in N_s, s_X^2 is approximately normal with variance $\frac{2(s_x^2)^2}{N_s-1}$. Let δ denote the population standardized mean difference and correlation. Define g as the function mapping z_f and σ_X^2 to δ:

$$g(x_1, x_2) = \frac{\Delta \tanh(x_1)}{(x_2)^{1/2}\sqrt{1 - \tanh^2(x_1)}}$$

Thus, $g(z_f, \sigma_X^2) = \delta$. The gradient evaluated at the population parameters is:

$$\nabla\Big|_{(z_f, \sigma_X^2)} = \begin{bmatrix} \frac{\partial g}{\partial z_f} \\ \frac{\partial g}{\partial \sigma_X^2} \end{bmatrix} = \begin{bmatrix} \frac{\Delta}{\sigma_X \sqrt{\text{sech}^2(z_f)}} \\ -\frac{1}{2}(\sigma_X^2)^{-3/2} \frac{\rho\Delta}{\sqrt{1-\rho^2}} \end{bmatrix} = \begin{bmatrix} \frac{\Delta}{\sigma_X \sqrt{1-\rho^2}} \\ -\frac{1}{2}(\sigma_X^2)^{-1}\delta \end{bmatrix} = \begin{bmatrix} \delta/\rho \\ -\frac{1}{2}(\sigma_X^2)^{-1}\delta \end{bmatrix}$$

Note that Lemma (2.1) regarding the asymptotic independence of z_f and s_x also implies that r and s_x are asymptotically independent because r is a function only of z_f. Thus, applying the delta method yields:

$$\widehat{\text{SE}}(d) \approx \sqrt{\widehat{\nabla}_1^2 \, \widehat{\text{SE}}^2(z_f) + \widehat{\nabla}_2^2 \, \widehat{\text{SE}}^2(s_X^2)}$$

$$= \sqrt{(d/r)^2 \frac{1}{N-3} + \frac{1}{4}(s_X^2)^{-2} d^2 \frac{2(s_X^2)^2}{N_s - 1}}$$

$$= |d| \sqrt{\frac{1}{r^2(N-3)} + \frac{1}{2(N_s - 1)}}$$

If $r = 0$ exactly, the standard error estimate is undefined, so could be replaced by:

$$\lim_{r \to 0} \widehat{\text{SE}}(d) = \frac{|d|}{s_X} \sqrt{\frac{1}{N-3}}$$

References

[1] Leon Isserlis. On a formula for the product-moment coefficient of any order of a normal frequency distribution in any number of variables. *Biometrika*, 12(1/2):134–139, 1918.