Material for Psychological Methods website

Additional Procedural Details of the CI Methods

The BC and BC$_a$ CI methods. The bounds of a CI are defined as $\left[r^*_{\alpha_L},\ r^*_{\alpha_U}\right]$. For a percentile CI when α = 0.05, α_L and α_U are the resampled stats at the 0.025th and 0.975th quantile of the bootstrap distribution. The BC$_a$ can shift and widen the bounds, based on the approximated median bias ($\hat{z}_0$) and approximated acceleration ($\hat{a}$), whose formulas are given below. The quantiles for the BC$_a$ are: $\alpha_L = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z_{\alpha/2}}{1-\hat{a}\left(\hat{z}_0 + z_{\alpha/2}\right)}\right), \alpha_U = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z_{1-\alpha/2}}{1-\hat{a}\left(\hat{z}_0 + z_{1-\alpha/2}\right)}\right)$, where $\Phi()$ is the standard normal cumulative distribution function (e.g., $\alpha_U = \Phi(1.96) \doteq 0.975$). When both $\hat{z}_0$ and $\hat{a}$ are zero, the BC, BC$_a$, and percentile CI are equal; when $\hat{a}$ is zero, the BC and BC$_a$ are equal.

The approximation for median bias is $\hat{z}_0 = \Phi^{-1}\left(\frac{\#(r^* \leq r_{\text{obs}})}{\text{B}+1}\right)$, where $\Phi^{-1}()$ is the inverse function of a standard normal cumulative distribution function (e.g., $\hat{z}_0 = \Phi^{-1}(0.47) \doteq -0.07527$). There are two small differences in the formula between Efron and Tibshirani (1993, equation 14.14) and Davison and Hinkley (1997, p. 205). First, the denominator is different because the first authors define *B* so that $\alpha \times B$ is an integer, while the second authors define *B* so that $\alpha \times (\text{B} + 1)$ is an integer. Second, the first authors use '<', while the second authors use '≤'. We used Davison and Hinkley's version in both cases. A final complication is that the HI univariate sampling bootstrap replaces r_{obs} with ρ_{null}.

Several approaches for approximating acceleration have been proposed (see Efron, 1987; Efron & Tibshirani, 1993, equations 14.15 and 22.29). We chose the jackknife approach (1993, equation 14.15) because it was simpler and it permitted our unit tests (Hunt & Tomas, 2004; McConnell, 2004, p. 499) to verify the results from our routine against a BC$_a$ routine (written by Efron and Tibshirani in S-Plus and adapted for R), which also used the jackknife to approximate acceleration. We slightly modified their formula for acceleration to accommodate the condition when the sampling frame is not equal to

the observed sample. The size of the sampling frame is *M* rows by 2 columns. Recall that a bivariate sampling frame has *N* rows (i.e., the number of rows in the observed sample), whereas a univariate sampling frame has N^2 rows. The approximation for acceleration is

$$\hat{a} = \frac{\sum_{i=1}^{M}\left(r_{(\cdot)}^{*} - r_{(\neq i)}^{*}\right)^3}{6\left(\sum_{i=1}^{M}\left(r_{(\cdot)}^{*} - r_{(\neq i)}^{*}\right)^2\right)^{\frac{3}{2}}}$$

where $r_{(\cdot)}^{*} = \frac{1}{M}\sum_{i=1}^{M} r_{(i)}^{*}$ and $r_{(\neq i)}^{*}$ is the correlation of the i^{th} jackknifed sample (i.e., the i^{th} row is excluded).

A variation of the asymptotic adjusted CI method. We evaluated another CI method that is a variation of the asymptotic adjusted CI (Efron, 1982, p. 72, 84-86). It focuses on the size of univariate sampling frame (N^2) instead of the size of the observed sample (*N*). This affects univariate sampling only, because a bivariate sampling frame is the same size as the observed sample. The bounds of the percentile CI are widened by a factor of $\sqrt{(N^2+2)/(N^2-1)}$. However this alternative CI method did not perform well in the simulations. It was too conservative for the HI and too liberal for the OI.

Generation of Population Scores

Population scores were generated by an approach developed by Headrick (2002). The approach considers the first six moments of each univariate distribution to generate a correlated multivariate nonnormal distribution. In the special case of a bivariate distribution, three Gaussian deviates are used for each pair of scores: one deviate (E_X) is dedicated to generating the *X* value, one deviate (E_Y) is dedicated to generating the *Y* value, and one deviate (*W*) contributes to both the *X* and *Y* value.

For example, the following steps generate a pair of scores where *X* has a normal distribution and *Y* has a Chi-square distribution with 1 *df*, and $\rho_{\text{true}} = 0.4$.

1) Calculate six coefficients for both *X* and *Y*, called c_{jx}s and c_{jy}s, where j = 1, 2, …, 6. The coefficients do not depend on the correlation between the two variables; their purpose is to reproduce the univariate moments so the marginal distribution is shaped correctly. The coefficients are solved with six simultaneous equations, which are explained further in Headrick (2002, equations 18, 22, & B.1-B.4). The coefficients for the normal distribution are {0, 1, 0, 0, 0, 0}; the Chi-square coefficients are {-0.40, 0.62, 0.42, 0.068, -0.0064, 0.000044}.

2) Calculate the intermediate correlation (ρ_Z) by substituting the desired population correlation (ρ_{true}) and the twelve coefficients from step 1 in the equation below.

$$\rho_{\text{true}} = 3c_{4,X}c_{0,Y} + 3c_{4,X}c_{2,Y} + 9c_{4,X}c_{4,Y} + c_{0,X}(c_{0,Y} + c_{2,Y} + 3c_{4,Y}) + c_{1,X}c_{1,Y}\rho_Z + 3c_{3,X}c_{1,X}\rho_Z + 15c_{5,X}c_{1,Y}\rho_Z + 3c_{1,X}c_{3,Y}\rho_Z + 9c_{3,X}c_{3,Y}\rho_Z + 45c_{5,X}c_{3,Y}\rho_Z + 15c_{1,X}c_{5,Y}\rho_Z + 45c_{3,X}c_{5,Y}\rho_Z + 225c_{5,X}c_{5,Y}\rho_Z + 12c_{4,X}c_{2,Y}\rho_Z^2 + 72c_{4,X}c_{4,Y}\rho_Z^2 + 6c_{3,X}c_{3,Y}\rho_Z^3 + 60c_{5,X}c_{3,Y}\rho_Z^3 + 60c_{3,X}c_{5,Y}\rho_Z^3 + 600c_{5,X}c_{5,Y}\rho_Z^3 + 24c_{4,X}c_{4,Y}\rho_Z^4 + 120c_{5,X}c_{5,Y}\rho_Z^5 + c_{2,X}(c_{0,Y} + c_{2,Y} + 3c_{4,Y} + 2c_{2,Y}\rho_Z^2 + 12c_{4,Y}\rho_Z^2)$$

For the specified population correlation and univariate distributions, ρ_Z = 0.695

3) Generate three Gaussian deviates: E_X, E_Y, and W. Calculate Z_X and Z_Y where

$$\begin{aligned} Z_X &= W\sqrt{\rho_Z} + E_X\sqrt{1-\rho_Z} = W\sqrt{0.695} + E_X\sqrt{1-0.695} \\ Z_Y &= W\sqrt{\rho_Z} + E_Y\sqrt{1-\rho_Z} = W\sqrt{0.695} + E_Y\sqrt{1-0.695} \end{aligned}$$

4) Calculate *X* and *Y*. In this example, these specific coefficients generate a normal (*X*) and a Chi-square (*Y*) distribution.

$$\begin{aligned} X &= c_{0,X} + c_{1,X}Z_X + c_{2,X}Z_X^2 + c_{3,X}Z_X^3 + c_{4,X}Z_X^4 + c_{5,X}Z_X^5 \\ &= 0 + Z_X + 0 + 0 + 0 + 0 \\ &= Z_X \\ Y &= c_{0,Y} + c_{1,Y}Z_Y + c_{2,Y}Z_Y^2 + c_{3,Y}Z_Y^3 + c_{4,Y}Z_Y^4 + c_{5,Y}Z_Y^5 \\ &= -.40 + 0.62Z_Y + 0.42Z_Y^2 + 0.068Z_Y^3 + -0.0064Z_Y^4 + 0.000044Z_Y^5 \end{aligned}$$

Steps 3 and 4 are repeated *N* times to generate a sample. The coefficients and intermediate correlation remain constant for the rest of the sample. Mathematica (version 5.2) was used to solve for

the coefficients (c_j) and the intermediate correlation corresponding to ρ_{true} values of 0.0, 0.1, … , 0.8. These constants were transferred into our C# program, which generated the data. Although only 2 digits are shown for each coefficient, our routine held 16 digits (a 64-bit floating point decimal). As a reviewer noted, distributions generated by this polynomial technique are only (close) approximations to the theoretical probability densities.

References

Hunt, A., & Thomas, D. (2004). *Pragmatic unit testing in C# with NUnit.* Dallas: Pragmatic Bookshelf.

McConnell, S. (2004). *Code complete* (2nd ed.). Redmond, WA: Microsoft Press.