AMERICAN INSTITUTES FOR RESEARCH®

APRIL 2018

Weighted and Unweighted Correlation Methods for Large-Scale Educational Assessment: wCorr Formulas

AIR - NAEP Working Paper #2018-01

NCES Data *R* Project Series #02

Paul Bailey, *American Institutes for Research*
Ahmad Emad, *American Institutes for Research*
Ting Zhang, *American Institutes for Research*
Qingshu Xie, *MacroSys*
Emmanuel Sikali, *National Center for Education Statistics*

The research contained in this working paper was commissioned by the National Center for Education Statistics (NCES). It was conducted by the American Institutes for Research (AIR) in the framework of the Education Statistics Services Institute Network (ESSIN) Task Order 14: Assessment Division Support (Contract No. ED-IES-12-D-0002/0004) which supports NCES with expert advice and technical assistance on issues related to the National Assessment of Educational Progress (NAEP). AIR is responsible for any error that this report may contain. Mention of trade names, commercial products, or organizations does not imply endorsement by the U.S. Government.

This page intentionally left blank.

Weighted and Unweighted Correlation Methods for Large-Scale Educational Assessment: wCorr Formulas

AIR - NAEP Working Paper #2018-01

NCES Data *R* Project Series #02

April 2018

Paul Bailey
Ahmad Emad
Ting Zhang
Qingshu Xie
Emmanuel Sikali

AIR
AMERICAN INSTITUTES FOR RESEARCH®

1000 Thomas Jefferson Street NW
Washington, DC 20007-3835
202.403.5000

www.air.org

AIR

Established in 1946, with headquarters in Washington, D.C., the American Institutes for Research (AIR) is a nonpartisan, not-for-profit organization that conducts behavioral and social science research and delivers technical assistance both domestically and internationally in the areas of health, education, and workforce productivity. For more information, visit www.air.org.

NCES

The National Center for Education Statistics (NCES) is the primary federal entity for collecting and analyzing data related to education in the U.S. and other nations. NCES is located within the U.S. Department of Education and the Institute of Education Sciences. NCES fulfills a Congressional mandate to collect, collate, analyze, and report complete statistics on the condition of American education; conduct and publish reports; and review and report on education activities internationally.

ESSIN

The Education Statistics Services Institute Network (ESSIN) is a network of companies that provide the National Center for Education Statistics (NCES) with expert advice and technical assistance, for example in areas such as statistical methodology; research, analysis and reporting; and Survey development. This AIR-NAEP working paper is based on research conducted under the Research, Analysis and Psychometric Support sub-component of ESSIN Task Order 14 for which AIR is the prime contractor. The two other sub-components under Task 14 are Assessment Operations Support and Reporting and Dissemination.

The NCES Project officer for the Research, Analysis and Psychometric Support sub-component of ESSIN Task Order 14 is William Tirre (William.Tirre@ed.gov).

The NCES Project officer for the NCES Data R Project is Emmanuel Sikali (Emmanuel.Sikali@ed.gov).

Suggested citation:

Bailey, P., Emad, A., Zhang, T., & Xie, Q. (2018). *Weighted and Unweighted Correlation Methods for Large-Scale Educational Assessment: wCorr Formulas* [AIR-NAEP Working Paper #2018-01, NCES Data *R* Project Series #02]. Washington, DC: American Institutes for Research.

For inquiries, contact:

Paul Bailey, Senior Economist
Email: pbailey@air.org

Markus Broer, Project Director for Research under ESSIN Task 14
Email: mbroer@air.org

Mary Ann Fox, Project Director of ESSIN Task 14
Email: mafox@air.org

Contents

Page

Figures

Page

This page intentionally left blank.

Introduction

The wCorr package can be used to calculate Pearson, Spearman, polyserial, and polychoric correlations, in weighted or unweighted form.[1] The package implements the tetrachoric correlation as a specific case of the polychoric correlation and biserial correlation as a specific case of the polyserial correlation. When weights are used, the correlation coefficients are calculated with so called sample weights or inverse probability weights.[2]

This vignette introduces the methodology used in the wCorr package for computing the Pearson, Spearman, polyserial, and polychoric correlations, with and without weights applied. For the polyserial and polychoric correlations, the coefficient is estimated using a numerical likelihood maximization.

The weighted (and unweighted) likelihood functions are presented. Then simulation evidence is presented to show correctness of the methods, including an examination of the bias and consistency. This is done separately for unweighted and weighted correlations.

Numerical simulations are used to show:

- The bias of the methods as a function of the true correlation coefficient (ρ) and the number of observations (n) in the unweighted and weighted cases; and
- The accuracy [measured with root mean squared error (RMSE) and mean absolute deviation (MAD)] of the methods as a function of ρ and n in the unweighted and weighed cases.

Note that here “bias” is used for the mean difference between true correlation and estimated correlation.

The *wCorr Arguments* vignette[3] describes the effects the Maximum Likelihood(ML) and fast arguments have on computation and gives examples of calls to wCorr.

Specification of estimation formulas

Here we focus on specification of the correlation coefficients between two vectors of random variables that are jointly bivariate normal. We call the two vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$. The i^{th} members of the vectors are then called x_i and y_i.

[1] The estimation procedure used by the wCorr package for the polyserial is based on the likelihood function in Cox, N. R. (1974), "Estimation of the Correlation between a Continuous and a Discrete Variable." *Biometrics*, **30** (1), pp n171-178. The likelihood function for polychoric is from Olsson, U. (1979) "Maximum Likelihood Estimation of the Polychoric Correlation Coefficient." *Psyhometrika*, **44** (4), pp 443-460. The likelihood used for Pearson and Spearman is written down in many places. One is the "correlate" function in Stata Corp, Stata Statistical Software: Release 8. College Station, TX: Stata Corp LP, 2003.

[2] Sample weights are comparable to `pweight` in Stata.

[3] The *wCorr Arguments* vignette can be found at https://cran.r-project.org/web/packages/wCorr/vignettes/wCorrArguments.html

Formulas for Pearson correlations with and without weights

The weighted Pearson correlation is computed using the formula

$$r_{Pearson} = \frac{\sum_{i=1}^{n}[w_i(x_i - \overline{x})(y_i - \overline{y})]}{\sqrt{\sum_{i=1}^{n}(w_i(x_i - \overline{x})^2)\sum_{i=1}^{n}(w_i(y_i - \overline{y})^2)}}$$

where w_i is the weights, $\overline{x}$ and $\overline{y}$ are the weighted mean of the $\boldsymbol{X}$ and $\boldsymbol{Y}$ respectively, and n is the number pairs (x_i, y_i).[4]

The unweighted Pearson correlation is calculated by setting all of the weights to one.

Formulas for Spearman correlations with and without weights

For the Spearman correlation coefficient the unweighted coefficient is calculated by ranking the data and then using those ranks to calculate the Pearson correlation coefficient—so the ranks stand in for the $\boldsymbol{X}$ and $\boldsymbol{Y}$ data. Again, similar to the Pearson, for the unweighted case the weights are all set to one.

For the unweighted case the highest rank receives a value of 1 and the second highest 2, and so on down to the n^{th} value. In addition, when data are ranked, ties must be handled in some way. The chosen method is to use the average of all tied ranks. For example, if the second and third rank units are tied then both units would receive a rank of 2.5 (the average of 2 and 3).

For the weighted case there is no commonly accepted weighted Spearman correlation coefficient. Stata does not estimate a weighted Spearman and SAS neither documents nor cites their methodology in either of the corr or freq procedures.

The weighted case presents two issues. First, the ranks must be calculated. Second, the correlation coefficient must be calculated.

Calculating the weighted rank for an individual level is done via two terms. For the jth element the rank is

$$rank_j = a_j + b_j$$

The first term a_j is the sum of all weights $\boldsymbol{W}$ less than or equal to this value of the outcome being ranked (ξ_j)

$$a_j = \sum_{i=1}^{n} w_i \, \boldsymbol{I}(\xi_i < \xi_j)$$

[4] See the "correlate" function in Stata Corp, Stata Statistical Software: Release 8. College Station, TX: Stata Corp LP, 2003.

Where

$$\boldsymbol{I}(\xi_i < \xi_j) = \begin{cases} 1 \; if \; \xi_i < \xi_j \\ 0 \; if \; \xi_i \geq \xi_j \end{cases}$$

w_i is the ith weight and ξ_i and ξ_j are the ith and jth value of the vector being ranked, respectively.

The term b_j then serves two roles: making the ranks start with one (when there are no ties) and making the ranks the average of the ties (when there are ties).

When there are ties, each unit receives the mean rank for all of the tied units. In the simplest case the weights are all one and there are n tied units the vector of tied ranks would be $\mathbf{v} = (a_j + 1, a_j + 2, \ldots, a_j + n)$. The mean of this vector (here called $rank^1$ to indicate it is a specific case of $rank$ when the weights are all one) is then

$$rank_j^1 = \frac{1}{n}\sum_{i=1}^{n}(a_j + i)$$

$$= \frac{1}{n}\left(na_j + \frac{n(n+1)}{2}\right)$$

$$= a_j + \frac{n+1}{2}$$

$$= a_j + b_j^1$$

Where

$$b_j^1 = \frac{n+1}{2}$$

where the superscript one is again used to indicate that this is only for the unweighted case where all weights are set to one.

Performing an analogous computation as above, the total number of units tied for this rank is taken to be the sum of all of the weights. And the average of these ranks is then taken to be that value divided by the number of sample units tied for this rank. So

$$b_j^w = \frac{n+1}{2}\overline{w_j}$$

where w_j is the mean weight of all of the tied units. It is easy to see that if $w_j = 1$ for all j then $b_j^w = b_j^1$.

After the $\boldsymbol{X}$ and $\boldsymbol{Y}$ vectors are ranked, they are plugged into the weighted Pearson correlation coefficient formula shown earlier.

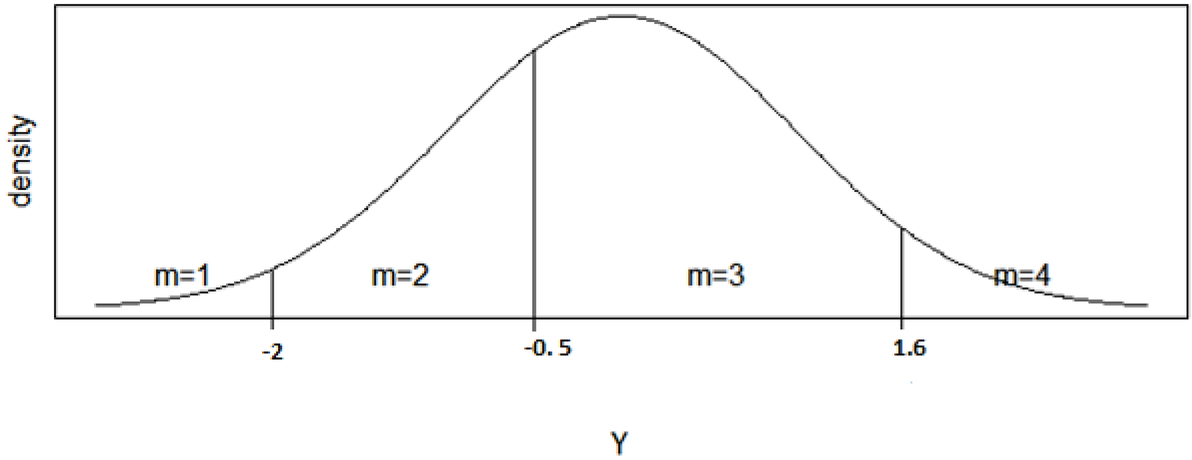
Polyserial correlation

For the polyserial correlation, it is again assumed that there are two continuous variables $\boldsymbol{X}$ and $\boldsymbol{Y}$ that have a bivariate normal distribution.[5]

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N[A, \boldsymbol{\Sigma}]$$

where $N(\mathbf{A}, \boldsymbol{\Sigma})$ is a bivariate normal distribution with mean vector $A = (\mu_x, \mu_y)$ and $\boldsymbol{\Sigma}$ is the covariance matrix of $\boldsymbol{X}$ and $\boldsymbol{Y}$. Where $\boldsymbol{X}$ is continuous but $\boldsymbol{Y}$ is discretized to $\boldsymbol{M}$. For example, if $\boldsymbol{Y}$ is partitioned as: $(-\infty, -2, -0.5, 1.6, \infty)$, then $m_i = 2$ when $-2 < y_i < -0.5$.

Figure 1. Density of Y for Cut Points $\boldsymbol{\theta} = (-\infty, \mathbf{-2}, \mathbf{-0.5}, \mathbf{1.6}, \infty)$.



Let μ_M and σ_M be the mean and standard deviation of the random variable $\boldsymbol{Y}$. Because a transformation of $\boldsymbol{Y}$ to $\boldsymbol{Y} + a$ where a is any real number can be offset by adjusting all cut points by a, the mean is irrelevant. A similar argument holds for σ_M. For convenience, we set, without any loss of generality, $\mu_Y = 0$ and $\sigma_Y = 1$.

Cox (1974) observed that the MLE mean and standard deviation of $\boldsymbol{X}$ are simply the average and (population) standard deviation of the data and do not depend on any other parameters.[6] This can be taken advantage of by substituting x by its standardized variable $\boldsymbol{Z}$.

Combining these simplifications, the probability of any given x_i, m_i pair can be written as:

[5] For a more complete treatment of the polyserial correlation, see Cox, N. R., "Estimation of the Correlation between a Continuous and a Discrete Variable" *Biometrics*, **50** (March), 171-187, 1974.

[6] The population standard deviation is used because it is the MLE for the standard deviation. Notice that, while the sample variance is an unbiased estimator of the variance and the population variance is not an unbaised estimator of the variance, they are very similar and the variance is also a nuisance parameter, not a parameter of interest when finding the correlation.

$$\Pr(\rho = r, \mathbf{\Theta} = \boldsymbol{\theta}; Z = z_i, M = m_i) = \phi(z_i)\int_{\theta_{m_i}}^{\theta_{m_i+1}} f(y|Z = z_i, \rho = r)dy$$

where $\Pr(\rho = r, \mathbf{\Theta} = \boldsymbol{\theta}; Z = z_i, M = m_i)$ is the probability of the event $\rho = r$ and the cut points are given by the vector $\boldsymbol{\theta}$, given the ith data point z_i and m_i; $\phi(\cdot)$ is the standard normal; and $f(Y|Z, \rho)$ is the distribution of ***Y*** conditional on ***Z*** and ρ. Because ***Y*** and ***Z*** are jointly normally distributed (by assumption)

$$f(Y|Z = z_i, \rho = r) \sim N\left(\mu_y + \frac{\sigma_y}{\sigma_z} r(z_i - \mu_z), (1 - r^2)\sigma_y{}^2\right) \sim N\big(r.z_i, (1 - r^2)\big)$$

because $\boldsymbol{Z} \sim N(0,1)$ and $\boldsymbol{Y} \sim N(0,1)$.

Because the random variable defined as $\frac{y - r \cdot z}{\sqrt{1 - r^2}}$ has a standard normal distribution, we can now write

$$\Pr(\rho = r, \mathbf{\Theta} = \boldsymbol{\theta}; Z = z_i, M = m_i) = \phi(z_i)\left[\Phi\left(\frac{\theta_{m_i+2} - r \cdot z_i}{\sqrt{1 - r^2}}\right) - \Phi\left(\frac{\theta_{m_i+1} - r \cdot z_i}{\sqrt{1 - r^2}}\right)\right]$$

where $\Phi(\cdot)$ is the standard normal cumulative density function.

Using the above probability function, the likelihood function is:

$$L(\rho = r, \mathbf{\Theta} = \boldsymbol{\theta}; \mathbf{Z} = \mathbf{z}, \mathbf{M} = \mathbf{m}) = \prod_{i=1}^{n} \Pr(\rho = r, \mathbf{\Theta} = \boldsymbol{\theta}; Z = z_i, M = m_i)^{w_i}$$

We then have to find the value of ρ that maximizes the likelihood function. Because the natural log function is monotonically increasing, we can maximize

$$\ln(L) = \sum_{i=1}^{n} w_i \ln[\Pr(\rho = r, \mathbf{\Theta} = \boldsymbol{\theta}; Z = z_i, M = m_i)]$$

instead, where w_i is the weight of the i^{th} member of the vectors ***Z*** and ***Y***. For the unweighted case, all of the weights are set to one.

In the unweighted case, the value of the nuisance parameter $\boldsymbol{\theta}$ is chosen to be

$$\hat{\theta}_{j+2} = \Phi^{-1}(n/N)$$

where n is the number of values to the left of the jth cut point (θ_{j+2} value) and N is the number of data points overall. Here two is added to j to make the indexing of θ agree with Cox (1974) as noted before. For the weighted cause n is replaced by the sum of the weights to the left of the jth cut point and N is replaced by the total weight of all units

$$\hat{\theta}_{j+2} = \Phi^{-1}\left(\frac{\sum_{i=1}^{N} w_i\, \boldsymbol{I}(m_i < j)}{\sum_{i=1}^{N} w_i}\right)$$

Because the numerical optimization is not perfect when the correlation is in a boundary condition ($\rho \in \{-1,1\}$), a check for perfect correlation is performed before the above optimization by simply examining if the values of ***X*** and ***M*** have weekly agreeing order (or opposite but agreeing order) and then the MLE correlation of 1 (or -1) is returned. Where weak agreement means that there is no disagreement—allowing for M to have different associated values of ***X***.

Polychoric correlation

As with the polyserial correlation, the polychoric correlation is a simple case of two continuous variables ***X*** and ***Y*** that have a bivariate normal distribution. But now both variables are discretized. The continuous (latent) variable ***Y*** was observed as a discretized variable ***M*** and the continuous (latent) variable ***X*** is discretized into ***P***.

The random variable ***P*** has the same properties as the ***M*** defined above.

Then the probability of any given pair (m_i, p_i) is

$$\Pr(\rho = r, \boldsymbol{\Theta} = \boldsymbol{\theta}, \boldsymbol{\Theta}' = \boldsymbol{\theta}'; P = p_i, M = m_i) = \int_{\theta'_{p_i}}^{\theta'_{p_i+1}} \int_{\theta_{m_i}}^{\theta_{m_i+1}} f(x, y|\rho = r) dy dx$$

where ρ is the correlation coefficient, θ is the cut points for ***M*** and θ' is the cut points for ***P***.

Using the probability, the log-likelihood is defined using the same logic as above. The natural logarithm of the likelihood function is then

$$\sum_{i=1}^{n} w_i \ln[\Pr(\rho = r, \boldsymbol{\Theta} = \boldsymbol{\theta}, \boldsymbol{\Theta}' = \boldsymbol{\theta}'; P = p_i, M = m_i)]$$

is then maximized. This is the weighted log-likelihood function. For the unweighted case set the weights to one.

Again, extreme values (correlations of 1 or -1) can be detected by testing for weakly ascending or descending associations between the discrete variables.

Simulation results

It is easy to prove the consistency of the $\boldsymbol{\theta}$ for the polyserial correlation and $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ for the polychoric correlation using the non-ML case. Similarly, for ρ, because it is a MLE that can be obtained by taking a derivative and setting it equal to zero, the results are asymptotically unbiased and obtain the Cramer-Rao lower bound.

This does not speak to the small sample properties of these correlation coefficients. Previous work has described their properties by simulation; and that tradition is continued below.[7]

Simulation study of unweighted correlations

In what follows, when the exact method of selecting a parameter (such as n) is not noted in the above descriptions, it is described as part of each simulation.

For each iteration (the exact number of times will be stated for each simulation), the following procedure is used:

- select a true Pearson correlation coefficient ρ;
- select the number of observations n;
- generate $\boldsymbol{X}$ and $\boldsymbol{Y}$ to be bivariate normally distributed using a pseudo-Random Number Generator (RNG);
- using a pseudo-RNG, select the number of bins for $\boldsymbol{M}$ and $\boldsymbol{P}$ (t and t') independantly from the set {2, 3, 4, 5};[8]
- select the bin boundaries for $\boldsymbol{M}$ and $\boldsymbol{P}$ ($\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$) by sorting the results of $(t-1)$ and $(t'-1)$ draws, respectively, from a normal distribution using a pseudo-RNG;
- confirm that at least 2 levels of each of $\boldsymbol{M}$ and $\boldsymbol{P}$ are occupied (if not, return to previous step); and
- calculate and record relevant statistics.

Bias, and RMSE of the unweighted correlations

This section shows the bias of the correlations as a function of the true correlation coefficient, ρ.

A simulation that consist of fifty computations, was done for is pair (ρ, n) where $\rho \in (-0.99, -0.95, -0.90, -0.85, \ldots, 0.95, 0.99)$, and $n \in \{10{,}100{,}1000\}$.

The bias and the RSME defined respectively as $\frac{1}{n}\sum_{i=1}^{n}(r_i - \rho_i)$ and $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(r_i - \rho_i)^2}$ where the true correlation coefficient is ρ_i and estimate correlation coefficient r_i were evaluated for the spearman, polyserial, polychoric and spearman correlations.

Figure 2 shows the bias as a function of the true correlation ρ. The simulation study shows that only the polyserial shows no bias at any level of n, shown by no clear deviation from 0 at any level of ρ. For the Pearson correlation there is bias when $n = 10$ that is not present when $n =$

[7] See, for example, the introduction to Rigdon, E. E. and Ferguson C. E., "The Performance of the Polychoric Correlation Coefficient and Selected Fitting Functions in Confirmatory Factor Analysis With Ordinal Data" *Journal of Marketing Research* **28** (4), pp. 491-497.

[8] This means that the simulation uses discrete ordinal variables (***M*** and ***P***) that have 2, 3, 4, or 5 discrete levels. Note that the number of levels in ***M*** and ***P*** are chosen independently so that one could be 2 while the other is 5 (or any other possible combination).

100 or 1,000. This is a well-known property of the estimator.[9] Similarly, the polychoric shows bias when $n = 10$.

The Spearman correlation shows bias at all of the tested levels of n. The bias is zero when the true correlation is 1, 0, or -1; is positive when ρ is below 0 (negative correlation); and is negative when ρ is above 0 (positive correlation). In this section, the Spearman correlation coefficient is compared with the true Pearson correlation coefficient. When this is done, the bias is expected because the Spearman correlation is not intended to recover a Pearson type correlation coefficient; it is designed to measure a separate quantity.

Figure 2. Bias Versus ρ for Unweighted Correlations

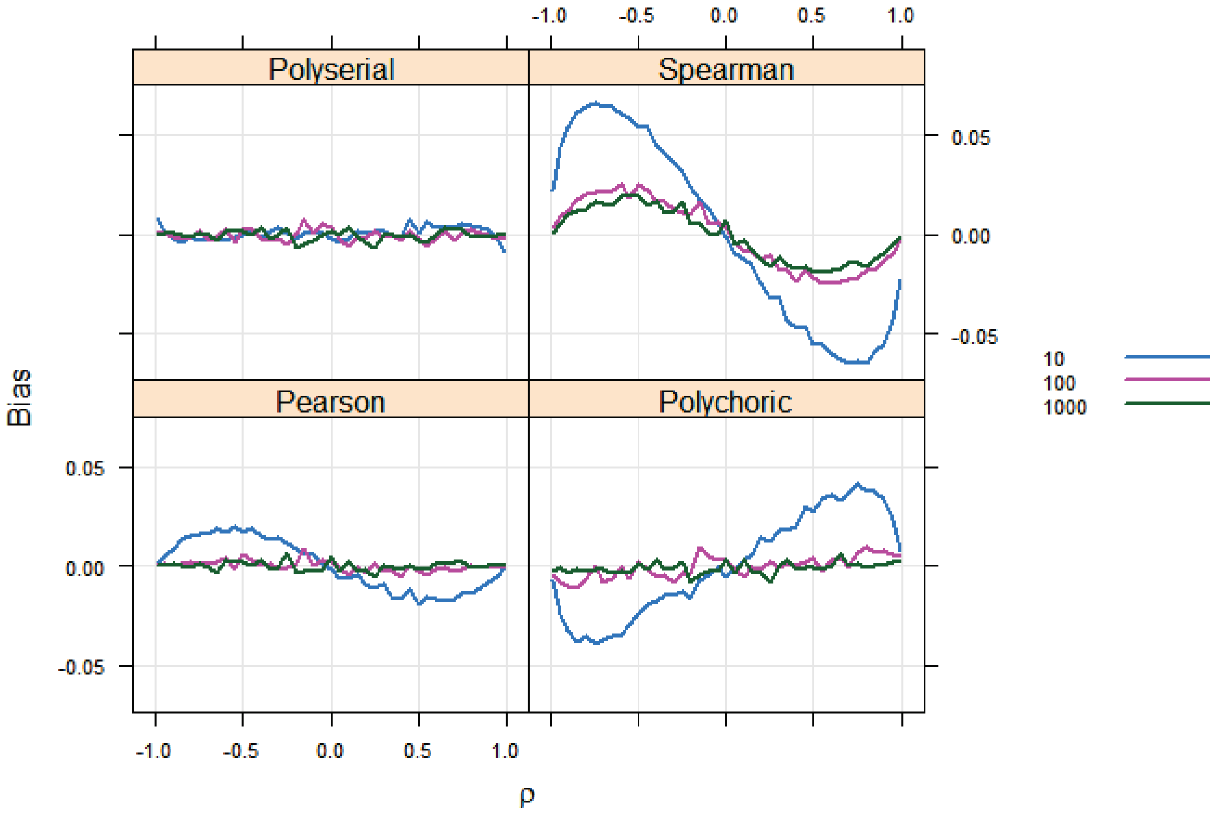
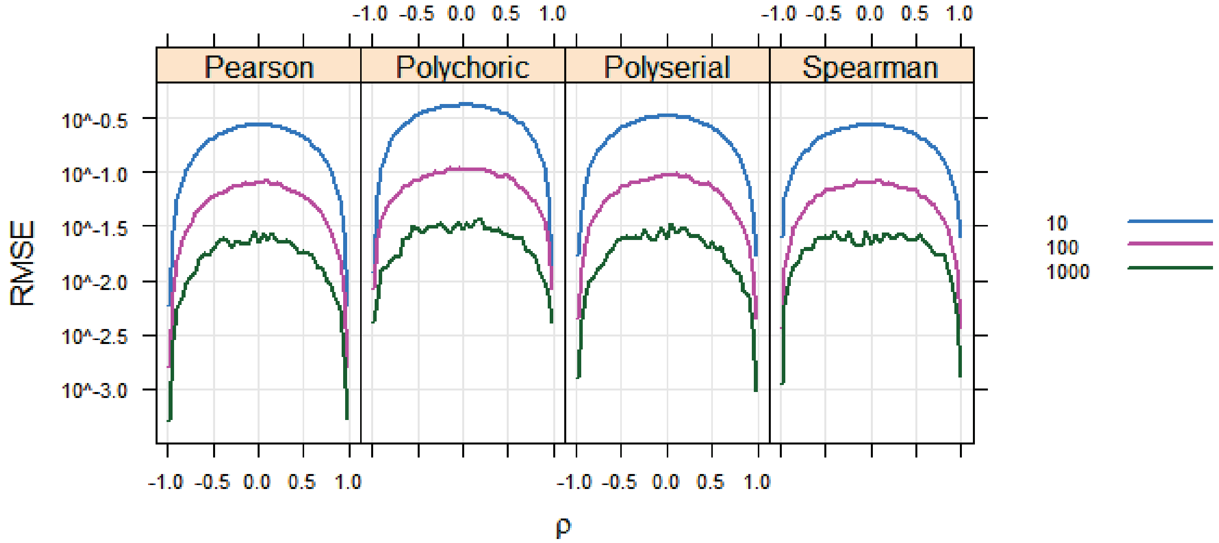


Figure 3 shows the RMSE as a function of ρ. All of the correlation coefficients have a uniform RMSE when as a function of $\rho -> 0$ that decreases when $0 < |\rho| < 1$. All plots also show a decrease in RMSE as n increases. This plot shows that there is no appreciable RMSE differences as a function of ρ. In addition, it shows that our attention to the MLE correlation of -1 or 1 at edge cases did not make the RMSE much worse in the neighborhood of the edges ($|\rho| \to 1$).

[9] See, for example, Olkin I. and Pratt, J. W. (1958). Unbiased estimation of certain correlation coefficients. *Annals of Mathematical Statistics, 29*(1), 201–211.

Figure 3. Root Mean Square Error Versus ρ for Unweighted Correlations

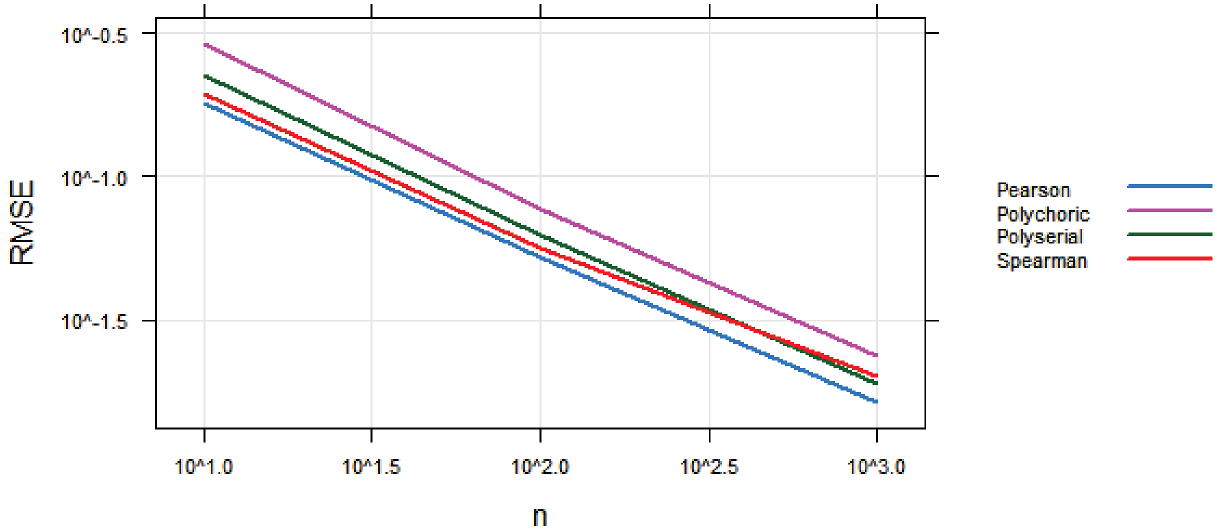


Consistency of the correlations

Figure 4 shows the RMSE as a function of n. The purpose of this plot is not to show an individual value but to show that the estimator is consistent. The plot shows a slope of about $-\frac{1}{2}$ for the Pearson, polychoric, and polyserial correlations. This is consistent with the expected first order convergence for each correlation coefficient under the assumptions of this simulation. Results for the Spearman also show approximate first order convergence but the slope increases slightly as n increases. Again, the Spearman is not estimating the same quantity as the Pearson and so is expected to diverge.

The plot also shows that the RMSE is less than 0.1 for all methods when $n > 100$.

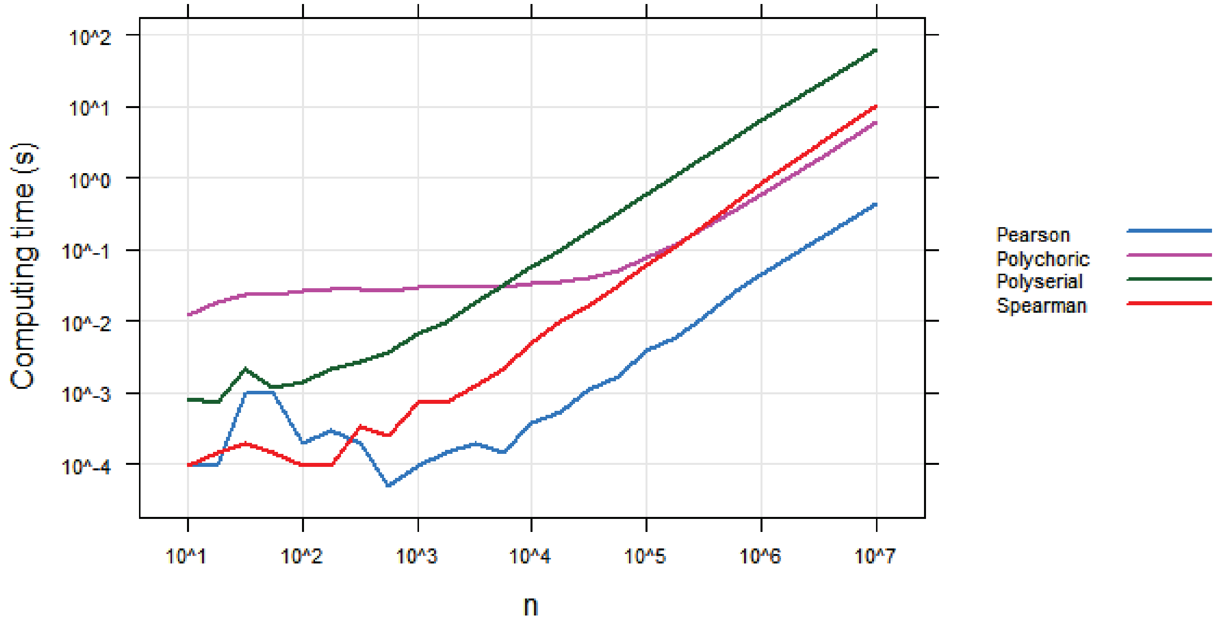
Figure 4. Root Mean Square Error Versus Sample Size for Unweighted Correlations



Computing Time

Figure 5 shows the mean time (in seconds) to compute a single correlation coefficient as a function of ρ and the value of n. The plot shows linearly rising computation times with slopes of about one. This is consistent with a linear computation cost. Using Big O notation to denote the computation cost, we see that the slopes are all about 1, so we can conclude that the overall complexity of these algorithms is O(n). The Spearman has a slightly higher slope, consistent with the use of a sort step that is O(n log(n)).

Figure 5. Computation time



Simulation study of weighted correlations

When complex sampling (other than simple random sampling with replacement) is used, unweighted correlations may or may not be consistent. In this section the consistency of the weighted coefficients is examined.

When generating simulated data, decisions about the generating functions have to be made. These decisions affect how the results are interpreted. For the weighted case, if these decisions lead to something about the higher weight cases being different from the lower weight cases then the test will be more informative about the role of weights. Thus, while it is not reasonable to always assume that there is a difference between the high and low weight cases, the assumption (used in the simulations below) that there is an association between weights and the correlation coefficients serves as a more robust test of the methods in this package.

Results of weighted correlation simulations

Simulations are carried out in the same fashion as previously described but include a few extra steps to accommodate weights. The following changes were made:

- Weights are assigned according to $w_i = (x - y)^2 + 1$, and the probability of inclusion in the sample was then $Pr_i = \frac{1}{w_i}$.
- For each unit, a uniformly distributed random number was drawn. When that value was less than the probability of inclusion (Pr_i), the unit was included.

Units were generated until n units were in the sample.

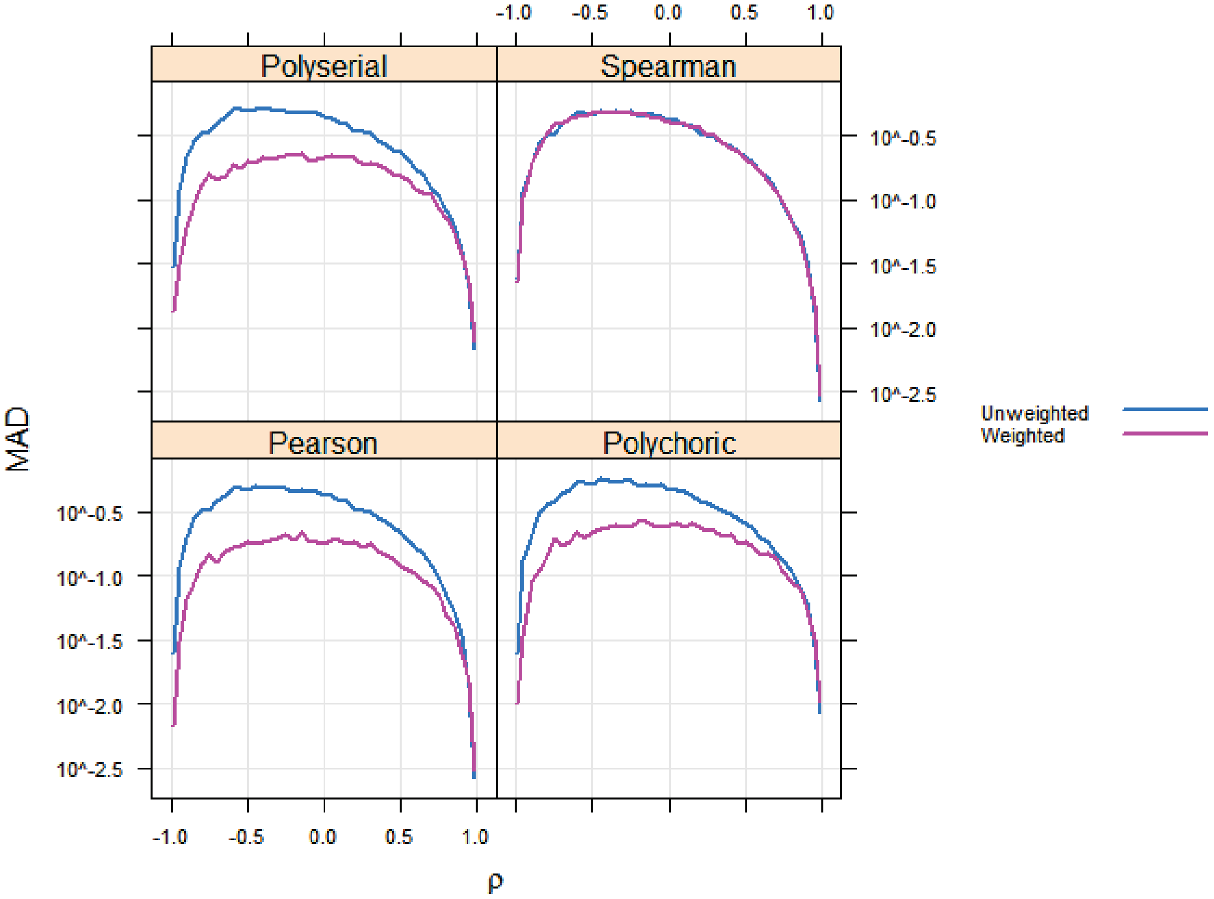
Two simulations were run. The first shows the mean absolute deviation (MAD)

$$MAD = \frac{1}{n}\sum_{i=1}^{n} | r_i - \rho_i |$$

as a function of ρ and was run for $n = 100$ and $\rho \in (-0.99, -0.95, -0.90, -0.85, \ldots, 0.95, 0.99)$, with 100 iterations run for each value of ρ.

The following plot shows the MAD for the weighted and unweighted results as a function of ρ when $n = 100$. This shows that for values of ρ near zero, under our simulation assumptions (for all but the Spearman correlation) the weighted correlation performs better than (has lower MAD than) the unweighted correlation for all correlation coefficients. Over the entire range, the difference between the two is never such that the unweighted has a lower MAD. Thus, under the simulated conditions at least, the weighted correlation has lower or approximately equal MAD for every value of the true correlation coefficient (ρ).

Figure 6. Mean Absolute Deviation Versus ρ (Weighted)



The second simulation (shown in Figure 7) used the same values of ρ and used $n \in \{10,100,1000,10000\}$ and shows how RMSE and sample size are related. In particular, it shows first order convergence of the weighted Pearson, polyserial, and polychoric correlation coefficient.

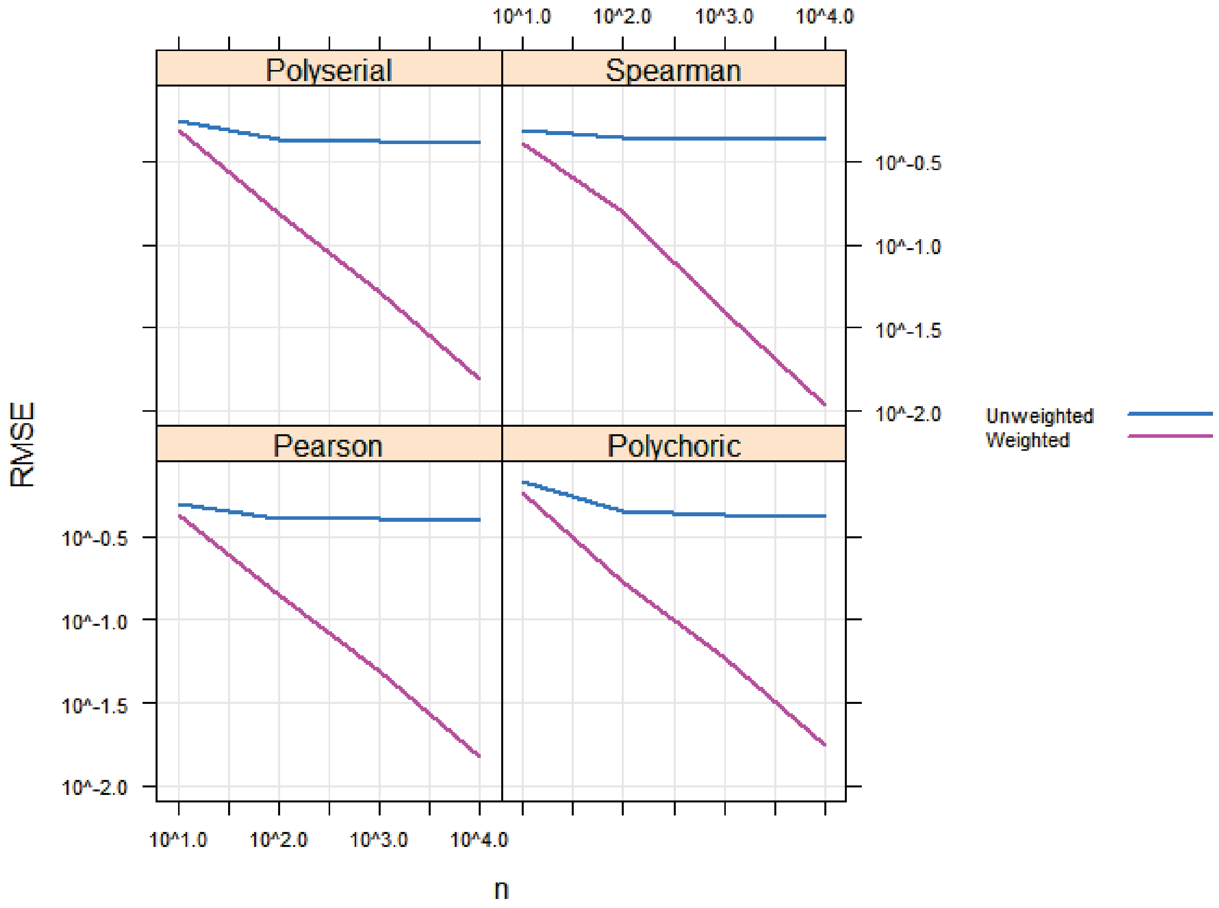
For the previous plots the calculated Spearman correlation coefficient was compared to the generating Pearson correlation coefficient. For this plot only, the Spearman correlation coefficient to the true Spearman correlation coefficient. This is because the Spearman coefficient is not attempting to estimate the Pearson correlation. To do this, the simulation is modified slightly. A population of data is generated and the true Spearman correlation coefficient then is calculated as the population coefficient.[10] Then, a sample from the population with varying probability as described in the weighted simulation section is used to calculate sample Spearman correlation coefficient. Then, the root mean squared difference between the sample and population coefficients are calculated as with the Pearson—except that the population Spearman correlation coefficient is used in place of the Pearson correlation coefficient (ρ).

[10] The R `stats` package `cor` function is used to calculate the population Spearman correlation coefficient; this results in an unweighted coefficient, which is appropriate for the population parameter.

Thus, the results in Figure 7 show that, when compared to the true Spearman correlation coefficient, the weighted Spearman correlation coefficient is consistent.

In all cases the RMSE is lower for the weighted than the unweighted. Again, the fact that the simulations show that the unweighted correlation coefficient is not consistent does not imply that it will always be that way—only that this is possible for these coefficients to not be consistent.

Figure 7. Root Mean Square Error vs *ρ* (Polyserial, Pearson, Polychoric panels) or Population Spearman Correlation Coefficient (Spearman Panel) for Weighted and Unweighted Correlations



Conclusion

Overall, the simulations show first-order convergence for each unweighted correlation coefficient with an approximately linear computation cost. Further, under our simulation assumptions, the weighted correlation performs better than (that is, has lower MAD or RMSE than) the unweighted correlation for all correlation coefficients.

We show the first-order convergence of the weighted Pearson, polyserial, and polychoric correlation coefficient. The Spearman is shown to not consistently estimate the population Pearson correlation coefficient but is shown to consistently estimate the population Spearman correlation coefficient—under the assumptions of our simulation

This page intentionally left blank.

This page intentionally left blank.

ABOUT AMERICAN INSTITUTES FOR RESEARCH

Established in 1946, with headquarters in Washington, D.C., American Institutes for Research (AIR) is an independent, nonpartisan, not-for-profit organization that conducts behavioral and social science research and delivers technical assistance both domestically and internationally. As one of the largest behavioral and social science research organizations in the world, AIR is committed to empowering communities and institutions with innovative solutions to the most critical challenges in education, health, workforce, and international development.

AIR®
AMERICAN INSTITUTES FOR RESEARCH®

1000 Thomas Jefferson Street NW
Washington, DC 20007-3835
202.403.5000

www.air.org

Making Research Relevant

LOCATIONS

Domestic

Washington, D.C.

Atlanta, GA

Austin, TX

Baltimore, MD

Cayce, SC

Chapel Hill, NC

Chicago, IL

Columbus, OH

Frederick, MD

Honolulu, HI

Indianapolis, IN

Metairie, LA

Naperville, IL

New York, NY

Rockville, MD

Sacramento, CA

San Mateo, CA

Waltham, MA

International

Egypt

Honduras

Ivory Coast

Liberia

Zambia