Research Article

David J. Torres*

Describing the Pearson R distribution of aggregate data

<https://doi.org/10.1515/mcma-2020-2054>

Received August 21, 2019; revised December 20, 2019; accepted January 7, 2020

Abstract: Ecological studies and epidemiology need to use group averaged data to make inferences about individual patterns. However, using correlations based on averages to estimate correlations of individual scores is subject to an “ecological fallacy”. The purpose of this article is to create distributions of Pearson R correlation values computed from grouped averaged or aggregate data using Monte Carlo simulations and random sampling. We show that, as the group size increases, the distributions can be approximated by a generalized hypergeometric distribution. The expectation of the constructed distribution slightly underestimates the individual Pearson R value, but the difference becomes smaller as the number of groups increases. The approximate normal distribution resulting from Fisher’s transformation can be used to build confidence intervals to approximate the Pearson R value based on individual scores from the Pearson R value based on the aggregated scores.

Keywords: Pearson R correlation, aggregate data, Monte Carlo simulations, confidence intervals

MSC 2010: 62H20, 91G60

1 Introduction

The relationship between the Pearson R and regression coefficients computed from individual scores and the Pearson R and regression coefficients computed from grouped averaged or aggregate scores has been the subject of many papers [4, 6, 8]. Ecological studies and epidemiology often need to use group averaged data to make inferences about individual patterns [14]. However, using correlations based on averages to estimate correlations of individual scores is subject to an “ecological fallacy”. Robinson [13] demonstrated that the correlation of averages generally does not agree with the correlation of the original scores.

Knapp [9] relates within-aggregate (i.e. within group) correlations R_w , the correlation coefficient $R_{\bar{x},\bar{y}}$ based on the group averages, and the correlation coefficient $R_{\text{individual}}$ based on the individual scores using the equation

$$R_{\text{individual}} = \eta_x \eta_y R_{\bar{x},\bar{y}} + \sqrt{1 - \eta_x^2} \sqrt{1 - \eta_y^2} R_w,$$

where η_x^2 and η_y^2 are correlation ratios (ratios of the between-aggregate variation to total variation). Hannan and Burstein [5] assert that the correlation coefficient based on groups is an unbiased estimate of the correlation coefficient based on individuals for random grouping. Ostroff [11] conducts an analysis of individual and aggregate correlations and plots ratios of aggregate correlations to individual correlations as a function of ratios of within and total variance. Piantadosi, Byar and Green [12] perform an analysis of individual and aggregate correlations and regression slopes and state that regression slopes based on aggregates are likely to be more accurate approximations of regression slopes based on individual values compared to correlation coefficients.

*Corresponding author: David J. Torres, Department of Mathematics and Physical Science, Northern New Mexico College, Española, NM, USA, e-mail: davytorres@nnmc.edu. <http://orcid.org/0000-0003-2469-5284>

Our objective is to describe the distribution of the Pearson $R_{\bar{x},\bar{y}}$ coefficients generated from the group averaged values and random sampling. In our literature search, we have not found work that describes the functional distribution of $R_{\bar{x},\bar{y}}$. We show through Monte Carlo simulations that the distribution of $R_{\bar{x},\bar{y}}$ can be approximated (for group sizes greater than 8) with a function based on the generalized hypergeometric function. Monte Carlo methods use repeated random sampling to solve a wide range of problems in the sciences and in applications involving measurements [16]. The expectation of the distribution of $R_{\bar{x},\bar{y}}$ slightly underestimates the individual Pearson R value, but the difference becomes smaller as the number of groups increase. The distribution of $R_{\bar{x},\bar{y}}$ values can be transformed into an approximately normal distribution under Fisher's transformation. The normal distribution can then be used to construct confidence intervals to approximate the individual $R_{\text{individual}}$ coefficient based on $R_{\bar{x},\bar{y}}$ assuming random sampling is used when constructing the group averages.

2 Distribution of Pearson R coefficient computed from a sample

2.1 Sampling from a bivariate distribution

If one samples $\{(x_i, y_i), 1 \leq i \leq n\}$ from a bivariate distribution

$$\mathcal{B}(x, y; \rho, \mu_x, \mu_y, \sigma_x, \sigma_y) = \frac{1}{2\pi\sigma_x\sigma_y(1-\rho^2)^{\frac{1}{2}}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 + \left(\frac{y-\mu_y}{\sigma_y}\right)^2 - 2\rho\frac{(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y}\right]\right\}, \quad (2.1)$$

where ρ is the level of correlation, μ_x and μ_y are the x -mean and y -mean, and σ_x and σ_y are the standard deviation for x and y , respectively, the sample Pearson R coefficient

$$R = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

will have the distribution

$$f(R) = \frac{(n-2)\Gamma(n-1)}{\sqrt{2\pi}\Gamma(n-\frac{1}{2})} (1-\rho^2)^{\frac{n-1}{2}} (1-\rho R)^{-n+\frac{3}{2}} (1-R^2)^{\frac{n}{2}-2} \times {}_2F_1\left(\frac{1}{2}, \frac{1}{2}; n-\frac{1}{2}; \frac{1}{2}(1+\rho R)\right) \quad (|R| < 1), \quad (2.2)$$

where Γ is the gamma function and ${}_2F_1$ is the generalized hypergeometric function [10]. The distribution (2.2) can be constructed using the series provided by [10]

$$f(R) = \frac{2^{n-3}(1-\rho^2)^{\frac{n-1}{2}}(1-R^2)^{\frac{n}{2}-2}}{\pi\Gamma(n-2)} \sum_{k=0}^{\infty} \left[\Gamma\left(\frac{n-1+k}{2}\right)\right]^2 \frac{(2R\rho)^k}{k!}. \quad (2.3)$$

Moreover, the estimate provided by the sample is biased and underestimates ρ . The expectation $E(R)$ and the variance $\text{Var}(R)$ are provided by Muirhead [10],

$$E(R) = \rho - \frac{\rho(1-\rho^2)}{2(n-1)} + O(n^{-2}), \quad \text{Var}(R) = \frac{(1-\rho^2)^2}{n-1} + O(n^{-2}),$$

where $O(n^{-2})$ refers to omitted terms on the order of $\frac{1}{n^p}$, where $p \geq 2$. Thus the difference between ρ and $E(R)$ can be estimated to first order using

$$\rho - E(R) \approx \frac{\rho(1-\rho^2)}{2(n-1)}. \quad (2.4)$$

The distribution (2.2) is not symmetric. However, Fisher [2] proposed the transformation

$$z = \frac{1}{2} \ln\left(\frac{1+R}{1-R}\right) = \tanh^{-1} R, \quad (2.5)$$

and it can be shown [10] that the resulting distribution approaches a normal distribution with standard

n	$\rho = 0$	$\rho = 0.7$
4	1.1	1.15
8	1.006	1.03
16	1.001	1.009
32	1.0002	1.004
64	1.00004	1.002
128	1.00001	1.001
256	1.000003	1.0005

Table 1: Ratio of the standard deviation σ_z defined by (2.6) to the standard deviation of the distribution (2.2) after Fisher's transformation is applied for $\rho = 0$ and $\rho = 0.7$.

deviation

$$\sigma_z = \frac{1}{\sqrt{n-3}} \quad (2.6)$$

as $n \rightarrow \infty$. Table 1 reports the ratio of the standard deviation σ_z to the standard deviation of the distribution (2.2) after Fisher's transformation is applied for two values of $\rho = 0$ and $\rho = 0.7$. The ratio rapidly approaches 1.0 as n increases.

2.2 Bivariate distribution of averages

The distribution of averages sampled from a bivariate distribution (2.1)

$$\bar{x}_k = \frac{1}{m} \sum_{i=1}^m x_i^{(k)}, \quad \bar{y}_k = \frac{1}{m} \sum_{i=1}^m y_i^{(k)}, \quad 1 \leq k \leq n,$$

is described by [3] as

$$\mathcal{B}\left(\bar{x}_k, \bar{y}_k; \rho, \mu_x, \mu_y, \frac{\sigma_x}{\sqrt{m}}, \frac{\sigma_y}{\sqrt{m}}\right), \quad (2.7)$$

where $x_i^{(k)}$ and $y_i^{(k)}$ refer to the i -th member of the k -th group and $\mathcal{B}$ is the bivariate distribution defined in equation (2.1). While the standard deviations are different in the arguments of (2.1) and (2.7), the Pearson R distribution (2.2) does not depend on the standard deviations. Therefore, the distribution (2.2) also applies to averages $(\bar{x}_k, \bar{y}_k)$.

Figure 1 confirms this with a Monte Carlo simulation and random sampling for $\rho = 0.7$. If n groups of m scores are drawn from a bivariate distribution and the groups of m scores are averaged, the distribution of

Figure 1: The analytical distribution (2.2) accurately represents the distribution created by n groups, where each of the n groups is formed by averaging m scores sampled from a bivariate distribution with $\rho = 0.7$. In the figure, $n = \frac{512}{m}$. For each value of m , 100,000 different random samples of size $n \cdot m = 512$ were chosen, and therefore 100,000 R values based on the averages were used to form each of the distributions shown as “bivariate avg” in the figure.

the Pearson R based on the averaged scores matches the distribution (2.2) which is generated using (2.3). In Figure 1, $n = \frac{512}{m}$ groups are formed, where each group has m scores sampled from the bivariate distribution (2.1). The values of m used are $m = 2, 4, 8, 16, 32, 64$, and 128 . The number of samples of size $n \cdot m = 512$ used to create each distribution for each value of m is 100,000. Sampling is done by first finding each $x_i^{(k)}$ and $y_i^{(k)*}$ by randomly sampling from a normal distribution with a mean of zero and a standard deviation of one. The scores are then correlated using

$$y_i^{(k)} = \rho x_i^{(k)} + \sqrt{1 - \rho^2} y_i^{(k)*}. \quad (2.8)$$

3 Distribution of Pearson $R_{\bar{x}, \bar{y}}$ generated from partitions

3.1 Defining a partition

We now describe the distribution of Pearson R coefficients created from group averages formed from a fixed sample of scores of length n . This requires us to define a partition of a set of original scores, where the scores are divided into subsets of equal size m .

Given a set of original scores $X = \{x_i, 1 \leq i \leq n\}$, $Y = \{y_i, 1 \leq i \leq n\}$, define a set P_k to be a group of m indices, where $m < n$ and n is divisible by m ,

$$P_k = \{i_1^k, i_2^k, i_3^k, \dots, i_m^k\},$$

where each index $1 \leq i_j^k \leq n$ is unique (i.e. $i_{j_1}^k \neq i_{j_2}^k$ when $j_1 \neq j_2$). Construct all the $\frac{n}{m}$ sets P_k so that they are disjoint,

$$P_{k_1} \cap P_{k_2} = \emptyset,$$

and so the union of all the P_k sets forms the entire set of integers ranging from 1 to n . We refer to the collection of all P_k sets as a partition

$$P = \bigcup_k P_k = \{1, 2, 3, \dots, n\}.$$

There exist

$$n_m \stackrel{\text{def}}{=} \frac{n}{m} \quad (3.1)$$

groups P_k in the partition P . For each P_k in the partition P , construct the mean scores

$$\bar{x}_k = \frac{x_{i_1^k} + x_{i_2^k} + \dots + x_{i_m^k}}{m}, \quad \bar{y}_k = \frac{y_{i_1^k} + y_{i_2^k} + \dots + y_{i_m^k}}{m}.$$

Assemble all the pairs $\{(\bar{x}_k, \bar{y}_k), k = 1, 2, \dots, \frac{n}{m}\}$, and define $R_{\bar{x}, \bar{y}}$ to be the Pearson correlation coefficient computed from these pairs,

$$R_{\bar{x}, \bar{y}} = \frac{\sum_{k=1}^{n_m} (\bar{x}_k - \bar{x})(\bar{y}_k - \bar{y})}{\sqrt{\sum_{k=1}^{n_m} (\bar{x}_k - \bar{x})^2 \sum_{k=1}^{n_m} (\bar{y}_k - \bar{y})^2}}. \quad (3.2)$$

Let us illustrate the process with an actual example. Suppose we have the following $n = 8$ elements in sets X and Y : $X = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8\}$ and $Y = \{y_1, y_2, y_3, y_4, y_5, y_6, y_7, y_8\}$. Using a group size of $m = 2$, one can select the sets $P_1 = \{1, 5\}$, $P_2 = \{2, 4\}$, $P_3 = \{3, 7\}$ and $P_4 = \{6, 8\}$. This selection of P_k 's divides the 8 original elements in X and Y into $n_m = 4$ groups

$$\{(x_1, x_5), (x_2, x_4), (x_3, x_7), (x_6, x_8)\} \quad \text{and} \quad \{(y_1, y_5), (y_2, y_4), (y_3, y_7), (y_6, y_8)\}.$$

The average $\bar{x}_k$ values are $\bar{x}_1 = (x_1 + x_5)/2$, $\bar{x}_2 = (x_2 + x_4)/2$, $\bar{x}_3 = (x_3 + x_7)/2$ and $\bar{x}_4 = (x_6 + x_8)/2$. The corresponding average $\bar{y}_k$ values are also computed. The Pearson $R_{\bar{x}, \bar{y}}$ value is then computed from these average pairs $\{(\bar{x}_1, \bar{y}_1), (\bar{x}_2, \bar{y}_2), (\bar{x}_3, \bar{y}_3), (\bar{x}_4, \bar{y}_4)\}$ using (3.2).

The number of partitions that can be constructed using groups of size m from n individual scores is

$$\frac{n!}{(m!)^{n_m} (n_m!)}. \quad (3.3)$$

Figure 2: $\log_{10}$ of the number of partitions (3.3) for $n = 120$ original scores and different group sizes m .

The expression in (3.3) is derived by first choosing m elements from n elements, $\binom{n}{m}$. Given the remaining $n - m$ elements, there are $\binom{n-m}{m}$ ways of choosing the second group of m elements. Continuing, there are

$$\prod_{i=0}^{n_m-1} \binom{n - (i \cdot m)}{m} = \frac{n!}{(m!)^{n_m}} \quad (3.4)$$

number of ways of choosing all groups. However, since order does not matter when choosing a partition, equation (3.4) needs to be divided by the number of ways of reordering the groups (i.e. $n_m!$) which yields (3.3).

Figure 2 shows the number of partitions for different group sizes m for $n = 120$ original scores. The vertical axis uses a log scale because of the wide range of scores. For example, 80 on the vertical axis corresponds to 10^{80} partitions.

3.2 Constructing the Pearson $R_{\bar{x}, \bar{y}}$ distribution with simulations

Section 2.2 showed that the Pearson R coefficient generated from averaged data sampled from a bivariate distribution follows the distribution described by equation (2.2). In contrast, to create the Pearson R distribution from a partition, a set of n individual scores is chosen and remains fixed. Each partition rearranges the n original scores into n_m different groups each of size m . A Pearson $R_{\bar{x}, \bar{y}}$ value is generated based on the n_m average scores.

Constructing the exact distribution using partitions would require one to compute the fraction of all possible partitions that produce a Pearson $R_{\bar{x}, \bar{y}}$ coefficient that resides within a series of intervals that fill the range between -1 and 1 . We do not attempt to develop the theory that underlies the exact distribution but instead rely on a series of Monte Carlo simulations to describe the distribution of Pearson $R_{\bar{x}, \bar{y}}$ coefficients. Our Monte Carlo simulations show that the Pearson $R_{\bar{x}, \bar{y}}$ values generated from the randomly generated partitions are similar to the distribution described by (2.2). However, there are differences for small group sizes, which diminish as the group size m increases. The symbols we use in our simulations are summarized in Table 2. We refer to the distribution of $R_{\bar{x}, \bar{y}}$ values created from the randomly generated partitions as the partition distribution.

We begin by generating the scores

$$X = \{x_i, 1 \leq i \leq n\}, \quad Y^* = \{y_i^*, 1 \leq i \leq n\}, \quad (3.5)$$

where each x_i and y_i^* is randomly sampled from a normal distribution with a mean of zero and a standard deviation of one. The scores are then correlated using the equivalent of equation (2.8)

$$y_i = \rho x_i + \sqrt{1 - \rho^2} y_i^*. \quad (3.6)$$

$R_{\text{individual}}$	Pearson R based on individual scores
$R_{\bar{x},\bar{y}}$	Pearson R based on group averages computed from partition
$\bar{R}_{\bar{x},\bar{y}}$	Average of $R_{\bar{x},\bar{y}}$ from many partitions
ρ	Level of correlation from bivariate distribution
n	Number of original scores
m	Size of groups
$n_m = \frac{n}{m}$	Number of groups

Table 2: List of symbols.

In Figure 3, we sample 1,000 different sets of $n = 64$ individual scores from a bivariate distribution (2.1) with $\rho = 0.7$ using (3.6). An $R_{\text{individual}}$ value is computed for each of the 1,000 different samples and plotted on the horizontal axis. For each of the 1,000 samples, 10,000 different partitions and their corresponding $R_{\bar{x},\bar{y}}$ values are generated with group size $m = 4$. The average Pearson value $\bar{R}_{\bar{x},\bar{y}}$ is computed by averaging the 10,000 $R_{\bar{x},\bar{y}}$ values from each partition distribution thus forming an approximation to the expected value for the distribution. The difference between $R_{\text{individual}}$ and $\bar{R}_{\bar{x},\bar{y}}$ is plotted on the vertical axis. Figure 3 shows that the expected value $\bar{R}_{\bar{x},\bar{y}}$ underpredicts $R_{\text{individual}}$ since the differences are all positive. Therefore, according to these simulations, the expected value from the partition distribution is a biased estimate of $R_{\text{individual}}$ at least for finite n . The difference $R_{\text{individual}} - \bar{R}_{\bar{x},\bar{y}}$ also appears to decrease as $R_{\text{individual}}$ increases.

To address the impact a specific individual sample has on the set of $R_{\bar{x},\bar{y}}$ coefficients and the average $\bar{R}_{\bar{x},\bar{y}}$ value, we generate 100 different samples from a bivariate distribution with $\rho = 0.7$ using (3.5) and (3.6). For consistency, equations (3.5) and (3.6) are now solved repeatedly until $R_{\text{individual}}$ and $\rho = 0.7$ agree to within 0.0001 before an individual sample is selected to be one of the 100 samples. The solving of (3.5) and (3.6) repeatedly until $R_{\text{individual}}$ and ρ agree to within 0.0001 is also used in the remaining figures with the exception of Figures 15 and 17. For each sample, 10,000 different partitions are generated, and $R_{\bar{x},\bar{y}}$ is generated for each partition. The 10,000 values of $R_{\bar{x},\bar{y}}$ are then averaged to form $\bar{R}_{\bar{x},\bar{y}}$. Figure 4 plots the average difference $R_{\text{individual}} - \bar{R}_{\bar{x},\bar{y}}$ of the 100 different samples as a function of the number of groups n_m . Different sample sizes n and group sizes m are used: $n = 210, m = 2, 3, 5, 7, 10, 15, 21$; $n = 512, m = 2, 4, 8, 16, 32, 64, 128$; $n = 1,024, m = 8, 16, 32, 64, 128, 256$; $n = 2,048, m = 32, 64, 128, 256$; $n = 4,096, m = 256$. The logarithm $\log_2(n_m)$ is used along the horizontal axis to compress the horizontal range. Figure 4 shows that the expected $\bar{R}_{\bar{x},\bar{y}}$ value is a biased estimate and underestimates $R_{\text{individual}}$. However, the graph also shows that, as the number of groups n_m in the partition increases, the average error and error bar range decrease. An error estimate $E(R_{\text{individual}}, n_m)$ shown in equation (3.7) is also plotted in yellow for comparison. The

Figure 3: Difference between $R_{\text{individual}}$ and the expected value $\bar{R}_{\bar{x},\bar{y}}$ generated by averaging 10,000 $R_{\bar{x},\bar{y}}$ values corresponding to 10,000 partitions with group size $m = 4$ generated from each of the 1,000 different samples of length $n = 64$ drawn from a bivariate distribution. Each partition contained $n_m = 16$ groups. The original scores were generated by sampling from a bivariate distribution with $\rho = 0.7$ using equations (3.5) and (3.6). $\bar{R}_{\bar{x},\bar{y}}$ underpredicts $R_{\text{individual}}$ in all 1,000 cases although all errors are less than 0.0017.

Figure 4: Difference between $R_{\text{individual}}$ and $\bar{R}_{\bar{x},\bar{y}}$ generated from 100 different samples as a function of the number of groups n_m . All samples are drawn from a bivariate normal distribution with $\rho = 0.7$. For each sample, 10,000 partitions are generated. The 10,000 $R_{\bar{x},\bar{y}}$ values generated for each partition are averaged to form the expected value $\bar{R}_{\bar{x},\bar{y}}$. The 100 differences $R_{\text{individual}} - \bar{R}_{\bar{x},\bar{y}}$ are averaged again and plotted. Equation (3.7) plotted in yellow estimates the difference $R_{\text{individual}} - \bar{R}_{\bar{x},\bar{y}}$ well for a large number of groups n_m .

Figure 5: Difference in maximum and minimum value $\bar{R}_{\bar{x},\bar{y}}$ generated from 100 different samples. All samples are drawn from a bivariate normal distribution with $\rho = 0.7$. For each sample, 10,000 partitions are generated. The 10,000 $R_{\bar{x},\bar{y}}$ values generated for each partition are averaged to form the expected value $\bar{R}_{\bar{x},\bar{y}}$. The maximum and minimum of the 100 values of $\bar{R}_{\bar{x},\bar{y}}$ are then found and the difference is plotted. The difference decreases as the number of groups n_m increases.

equation shows that $E(R_{\text{individual}}, n_m)$ estimates the first-order error $R_{\text{individual}} - \bar{R}_{\bar{x},\bar{y}}$ well for a number of groups $n_m > 8$,

$$E(R_{\text{individual}}, n_m) = \frac{R_{\text{individual}}(1 - R_{\text{individual}}^2)}{2(n_m - 1)}. \quad (3.7)$$

Equation (3.7) is generated by replacing ρ with $R_{\text{individual}}$ and n with n_m in equation (2.4). The difference between the maximum $\bar{R}_{\bar{x},\bar{y}}$ and minimum $\bar{R}_{\bar{x},\bar{y}}$ over the 100 samples is plotted in Figure 5. The difference decreases as the number of groups increases.

Figure 6 compares the analytical distribution

$$f(R_{\bar{x},\bar{y}}) = \frac{(n_m - 2)\Gamma(n_m - 1)}{\sqrt{2\pi}\Gamma(n_m - \frac{1}{2})} (1 - R_{\text{individual}}^2)^{\frac{n_m-1}{2}} (1 - R_{\text{individual}}R_{\bar{x},\bar{y}})^{-n_m+\frac{3}{2}} (1 - R_{\bar{x},\bar{y}}^2)^{\frac{n_m}{2}-2} \\ \times {}_2F_1\left(\frac{1}{2}, \frac{1}{2}; n_m - \frac{1}{2}; \frac{1}{2}(1 + R_{\text{individual}}R_{\bar{x},\bar{y}})\right) \quad (|R_{\bar{x},\bar{y}}| < 1) \quad (3.8)$$

to the partition distribution formed by randomly generating 50,000 partitions for $n = 512$ scores sampled from the bivariate distribution and different group sizes, $m = 2, 4, 8, 16, 32, 64$ and 128 corresponding to

Figure 6: Comparing the analytical distribution (3.8) to the partition distribution for $n = 512$, $\rho = 0.7$ and different group sizes m . The number of groups n_m is $\frac{n}{m}$. The 512 scores are sampled from a bivariate distribution. For each group size, 50,000 partitions were used to create the partition distribution.

Figure 7: Comparing the analytical distribution (3.8) to the partition distribution for $n = 512$, $\rho = 0$ and different group sizes m . The number of groups n_m is $\frac{n}{m}$. The 512 scores are sampled from a bivariate distribution. For each group size, 50,000 partitions were used to create the partition distribution.

the number of groups $n_m = 256, 128, 64, 32, 16, 8$ and 4 . The analytical distribution (3.8) is essentially (2.2) with R, ρ and n replaced with $R_{\bar{x}, \bar{y}}, R_{\text{individual}}$ and n_m , respectively. Higher peaks are clearly evident for group sizes $m = 2$ and $m = 4$ for the partition distribution labelled “partition” compared to the analytical distribution. The partition distributions for group sizes $m = 2$ and $m = 4$ also have smaller standard deviations compared to the analytical distribution (3.8). This trend is present for the same simulation performed with $\rho = 0$ shown in Figure 7. For larger group sizes ($m > 8$), the partition distribution agrees well with the analytical distribution. Differences between (3.8) and the partition distribution for $\rho = 0.7$ are shown in Figure 8.

We also create the partition distribution for scores randomly sampled from a uniform distribution, $0 \leq x_i \leq 1$, $0 \leq y_i \leq 1$, and correlated using (3.6) with $\rho = 0.7$. For each different group size, 50,000 partitions are generated. Differences between (3.8) and the partition distribution are shown in Figure 9. The high degree of similarity between Figures 8 and 9 shows that sampling from either a bivariate or uniform distribution does not affect the partition distributions.

Figure 10 uses different group sizes $m = 2, 4, 8$ and 64 and a fixed number of groups $n_m = 16$. For each group size, 50,000 partitions are generated. Samples are drawn from a bivariate distribution with $\rho = 0.7$. We see that, as the group size increases, the partition distributions converge to the analytical distribution (3.8). Figure 11 compares the analytical distribution described by (3.8) to the partition distribution formed by randomly generating 50,000 partitions for $n = 512$ scores sampled from a bivariate distribution with group size

Figure 8: Difference between analytical distribution (3.8) and partition distribution for $n = 512$, $\rho = 0.7$ and different group sizes m . The number of groups n_m is $\frac{n}{m}$. The 512 scores are sampled from a bivariate distribution. The difference decreases as the group size m increases. For each group size, 50,000 partitions were used to create the partition distribution.

Figure 9: Difference between analytical distribution (3.8) and partition distribution for $n = 512$, $\rho = 0.7$ and different group sizes m . The number of groups n_m is $\frac{n}{m}$. The 512 scores are sampled from a uniform distribution. The difference decreases as the group size m increases. For each group size, 50,000 partitions were used to create the partition distribution.

$m = 16$ and number of groups $n_m = 32$ for different values of $\rho = -0.7, -0.5, -0.2, 0.1, 0.4$ and 0.8 . The partition distribution agrees with the analytical distribution for all values of ρ .

Figure 12 compares the analytical distribution described by (3.8) to the partition distribution formed by randomly generating 50,000 partitions for $n = 512$ scores sampled from a bivariate distribution and a mixture of group sizes $m = 16$ and $m = 32$ for different values of $\rho = -0.7, -0.5, -0.2, 0.1, 0.4$ and 0.8 . Sixteen of the groups have size $m = 16$, and eight of the groups have size $m = 32$ in each partition. In the analytical distribution (3.8), we set $n_m = 24$ since there are 24 total groups in each partition. We see that the partition distribution agrees well with the analytical distribution for all values of ρ in spite of the fact that the partition is composed of groups of different size.

Figure 13 shows the ratio

$$\frac{s_{\text{analytical}}}{s_{\text{partition}}} \quad (3.9)$$

of the standard deviation of the analytical distribution $s_{\text{analytical}}$ (3.8) to standard deviation of the partition distribution $s_{\text{partition}}$ for $\rho = 0.7$ under the same conditions described in Figure 4. For each value of the sample size n and the group size m , 100 different samples were generated. The analytical distribution has a larger standard deviation than the partition distribution, which is quite evident for smaller group sizes. The logarithm $\log_2(m)$ is used along the horizontal axis to compress the horizontal range. We see that the standard

Figure 10: Comparing the analytical distribution (3.8) to the partition distribution for number of groups $n_m = 16$, $\rho = 0.7$ and different group sizes m . As the group size increases, the partition distributions converge to the analytical distribution. For each group size, 50,000 partitions were used to create the partition distribution. The original scores are sampled from a bivariate distribution.

Figure 11: Comparing the analytical distribution to the partition distribution for different values of ρ , $n = 512$ scores sampled from a bivariate distribution, and $m = 16$. For each value of ρ , 50,000 partitions were used to create the partition distribution.

Figure 12: Comparing the analytical distribution to the partition distribution for different values of ρ for $n = 512$ scores sampled from a bivariate distribution. Sixteen of the groups have size $m = 16$, and eight of the groups have size $m = 32$. The number of groups $n_m = 24$. For each value of ρ , 50,000 partitions were used to create the partition distribution.

Figure 13: Ratio of standard deviation of analytical distribution to the partition distribution (3.9) for $\rho = 0.7$, sample sizes $n = 210$, $n = 512$, $n = 1,024$, $n = 2,048$ and $n = 4,096$, and different group sizes m . For each value of n and m , 100 different samples are chosen, and 10,000 partitions are created for each sample.

Figure 14: In a follow-up to Figure 13, the difference between maximum and minimum standard deviation ratio (3.9) for all 100 samples for $\rho = 0.7$ and each value of n and m is plotted.

deviation of the partition distribution approaches the standard deviation of the analytical distribution as the group size increases.

The difference between the maximum ratio and the minimum ratio over all 100 samples for all values of n and m is plotted in Figure 14. For some sample sizes (e.g. $n = 512$, $n = 2,048$), there seems to be a slight increase in the difference as the group size m increases and the corresponding number of groups n_m decreases. However, the difference between the maximum ratio and the minimum ratio is less than 0.065 for all values of n and m .

These Monte Carlo simulations suggest that the distribution of Pearson $R_{\bar{x}, \bar{y}}$ values can be approximated (for group sizes $m > 8$) with the distribution described by (3.8).

4 Constructing confidence intervals

4.1 Confidence intervals for large group sizes m

Since the partition distribution will approach (3.8) as the group size m increases, Fisher's transformation (2.5) can be used to transform the set of $R_{\bar{x}, \bar{y}}$ values to z -values that become normally distributed as n_m grows

large. Confidence intervals can then be constructed from the normal distribution to predict the individual Pearson R value from the average Pearson $R_{\bar{x},\bar{y}}$ value according to equations (2.5) and (2.6). First the Fisher transformed (see (2.5)) z -value is computed from $R_{\bar{x},\bar{y}}$,

$$z_c = \frac{1}{2} \ln \left(\frac{1 + R_{\bar{x},\bar{y}}}{1 - R_{\bar{x},\bar{y}}} \right). \quad (4.1)$$

For a $C\%$ confidence interval, $z_{\alpha/2}$ is found, where $z_{\alpha/2}$ is the z -value above which $\frac{1}{2}(\frac{100-C}{100})$ of a standard normal distribution ($\mu = 0, \sigma = 1$) area lies,

$$\int_{z_{\alpha/2}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = \frac{1}{2} \left(\frac{100 - C}{100} \right). \quad (4.2)$$

For example, for a 95 % confidence interval, the value of $z_{\alpha/2}$ is the z -value above which 0.025 of the standard normal distribution area lies. The lower and upper z -values of the confidence interval are then computed using

$$z_l = z_c - \frac{1}{\sqrt{n_m - 3}} z_{\alpha/2}, \quad z_u = z_c + \frac{1}{\sqrt{n_m - 3}} z_{\alpha/2}, \quad (4.3)$$

where n_m is the number of groups used to compute the $R_{\bar{x},\bar{y}}$ value. The values of z_l and z_u are then transformed back to R values using

$$R_l = \tanh(z_l), \quad R_u = \tanh(z_u) \quad (4.4)$$

to generate the confidence interval (R_l, R_u) .

Equations (4.1)–(4.4) can be used to generate confidence intervals (R_l, R_u) for $R_{\text{individual}}$ given $R_{\bar{x},\bar{y}}$ as long as the group size is large ($m > 8$) and the number of groups n_m is large. A large value of n_m also reduces the bias according to (3.7). Table 3 shows the 95 % confidence intervals (R_l, R_u) for different values of $R_{\bar{x},\bar{y}}$ and different number of groups n_m . The confidence interval ranges are narrow for values of $R_{\bar{x},\bar{y}}$ close to 1 and large values of n_m . Figure 15 creates a filled contour plot of the 95 % confidence interval width $R_u - R_l$ as a function of $R_{\bar{x},\bar{y}}$ and the number of groups.

n_m	$R_{\bar{x},\bar{y}} = 0.1$	$R_{\bar{x},\bar{y}} = 0.3$	$R_{\bar{x},\bar{y}} = 0.5$	$R_{\bar{x},\bar{y}} = 0.7$	$R_{\bar{x},\bar{y}} = 0.9$
20	(−0.36, 0.52)	(−0.16, 0.66)	(0.07, 0.77)	(0.37, 0.87)	(0.76, 0.96)
40	(−0.22, 0.40)	(−0.01, 0.56)	(0.22, 0.70)	(0.50, 0.83)	(0.82, 0.95)
80	(−0.12, 0.31)	(0.09, 0.49)	(0.31, 0.65)	(0.57, 0.80)	(0.85, 0.93)
160	(−0.06, 0.25)	(0.15, 0.43)	(0.37, 0.61)	(0.61, 0.77)	(0.87, 0.93)
320	(−0.01, 0.21)	(0.20, 0.40)	(0.41, 0.58)	(0.64, 0.75)	(0.88, 0.92)

Table 3: 95 % confidence intervals (R_l, R_u) for $R_{\text{individual}}$ for different values of $R_{\bar{x},\bar{y}}$ and number of groups n_m . The confidence intervals are narrow for values of $R_{\bar{x},\bar{y}}$ close to 1 and large values of n_m .

Figure 15: Width of confidence interval $R_u - R_l$ for $R_{\text{individual}}$ as a function of $R_{\bar{x},\bar{y}}$ and the number of groups n_m . Narrow widths (represented by darker shades of blue) occur for absolute values of $R_{\bar{x},\bar{y}}$ close to one and for a large number of groups.

Figure 16: Comparing the analytical distribution (3.8) to the partitioning distribution for different group sizes m for $\rho = 0.7$ after Fisher's transformation (2.5) is applied. Small group sizes $m = 2$ and $m = 4$ have a larger peak and a smaller standard deviation compared to the analytical distribution. For each group size, 50,000 partitions were used to create the partition distribution.

4.2 Confidence intervals for small group sizes $m \leq 8$

If the group sizes m are not large, the process of generating the confidence intervals shown in equations (4.1)–(4.4) will need to be revised to account for the fact that the analytical distribution has a larger standard deviation compared to the partition distribution. Figure 16 compares the analytical distribution (3.8) to the partition distribution created through Monte Carlo simulation and random sampling after Fisher's transformation (2.5) is applied for $\rho = 0.7$. Similar to Figure 6, small group sizes (e.g. $m = 2$ and $m = 4$) have a smaller standard deviation compared to the analytical distribution.

We assess whether the partition distributions for small group sizes $m = 2$, $m = 4$ and $m = 8$ can be approximated by a normal distribution by plotting the Fisher z -values (4.1) arranged in ascending order against the z -values (extracted from a normal distribution) corresponding to a measure f_i of the distribution area below each Fisher z -value,

$$f_i = \frac{i - 0.375}{I + 0.25}. \quad (4.5)$$

In (4.5), i refers to the location of the Fisher z -value when all the values are arranged in ascending order, and I refers to the total number of values in the distribution. If the plot of the Fisher z -values versus the z -values extracted from the normal distribution is linear, the partition distribution is normal [15]. This plot is shown in Figure 17. Except for extreme z -values, $|z| > 3.2$, which constitute only 0.13 % of all the z -values, the distributions can be approximated by a normal distribution.

Figure 18 plots the ratio

$$R_{sd} = \frac{s_{\text{analytical}}^{\text{Fisher}}}{s_{\text{partition}}^{\text{Fisher}}} \quad (4.6)$$

of the standard deviation of Fisher's transformed analytical distribution

$$s_{\text{analytical}}^{\text{Fisher}} = \frac{1}{\sqrt{n_m - 3}} \quad (4.7)$$

to the standard deviation of Fisher's transformed partition distribution $s_{\text{partition}}^{\text{Fisher}}$ for $\rho = 0.7$. Different group sizes m are used with a variety of different sampling lengths $n = 210$, $n = 512$, $n = 1,024$, $n = 2,048$ and $n = 4,096$. For each value of n and m , 100 different samples were generated, and 10,000 randomly created partitions were used to create the partition distribution for each of the 100 different samples. Similar to Figure 13, Figure 18 shows that the Fisher transformed analytical distribution has a larger standard deviation than the Fisher transformed partition distribution for small group sizes. The standard deviation of the partition distribution approaches the standard deviation of the analytical distribution as the group size increases.

Figure 17: The z-values extracted from a normal distribution are plotted against the Fisher z-values for group sizes $m = 2$, $m = 4$ and $m = 8$. The figure shows that, except for extreme z-values $|z| > 3.2$ which constitute only 0.13 % of all the z-values, the distributions can be approximated by a normal distribution. The 30,000 partitions and 30,000 $R_{\bar{x}, \bar{y}}$ values are computed for each of the partition distributions which are generated from of sample of length $n = 1,024$ formed using a bivariate distribution with $\rho = 0.7$. The black lines are generated by constructing the least squares line for values of $|z| \leq 3.2$.

Figure 18: Ratio of standard deviation of analytical normal distribution (4.6) to the partition distribution for different group sizes m and sample sizes n for $\rho = 0.7$ after Fisher's transformation is applied. The curve fit function $\beta(m)$ from equation (4.8) is also plotted.

Figure 19: Ratio of standard deviation of analytical normal distribution (4.6) to the partition distribution for different group sizes m for $\rho = 0$ after Fisher's transformation is applied. The curve fit function $\beta(m)$ from equation (4.8) is also plotted.

m	$\rho = -0.9$	$\rho = -0.5$	$\rho = 0$	$\rho = 0.3$	$\rho = 0.7$
2	(94.4, 95.6)	(94.3, 95.5)	(94.5, 95.6)	(94.5, 95.7)	(94.5, 95.5)
4	(94.4, 95.6)	(94.5, 95.5)	(94.5, 95.7)	(94.5, 95.5)	(94.5, 95.5)
8	(95.0, 96.0)	(95.0, 95.9)	(94.9, 96.0)	(95.0, 96.0)	(95.1, 95.9)
16	(95.0, 96.2)	(95.0, 95.9)	(95.1, 95.9)	(94.9, 96.0)	(95.1, 96.1)

Table 4: Simulation to assess the validity of the confidence intervals (4.9) for different group sizes m and different values of ρ . Each value of m and ρ shows the minimum and maximum percentage of $R_{\text{individual}}$ values that reside within the confidence interval for 100 different samples.

However, to account for the difference in the standard deviations for smaller group sizes, we also fit a curve $\beta(m)$ to the data to approximate R_{sd} which is shown in yellow,

$$\beta(m) = 1.01 \sqrt{\frac{2^{1.5} + m^{1.5}}{m^{1.5}}}. \quad (4.8)$$

Figure 18 only uses combinations of n and m , so the number of groups $n_m = \frac{n}{m}$ is at least 16 since the number of groups need to be fairly large in order for (4.7) to be valid. Figure 19 performs the same simulation as Figure 18 except $\rho = 0$. (Also, to reduce the computational cost, $n = 4,096$ is not used, and only $m = 32$ is used for $n = 1,024$, and only $m = 64$ is used for $n = 2,048$.) Figure 19 shows that (4.8) is still a good approximation to the ratio of the standard deviations when $\rho = 0$.

Equations (4.1)–(4.4) can still be used to compute the confidence intervals, but equation (4.3) is replaced with

$$\begin{aligned} z_l &= z_c - \frac{1.03}{\beta(m)\sqrt{n_m - 3}} z_{\alpha/2}, \\ z_u &= z_c + \frac{1.03}{\beta(m)\sqrt{n_m - 3}} z_{\alpha/2}. \end{aligned} \quad (4.9)$$

We also increase the standard deviation by 3 % (reflected by the 1.03 in equation (4.9)) to account for the bias (3.7) and the range in standard deviations seen in Figure 14.

4.3 Assessing the accuracy of the confidence intervals

A simulation is performed to assess the accuracy of the confidence intervals computed using (4.9). A value of ρ is selected from $\rho = -0.9, -0.5, 0, 0.3$ and 0.7 , and a group size is selected from $m = 2, 4, 8, 16$ for $n_m = 32$ number of groups. The number of original scores n is $n = 32m$ according to (3.1). One hundred samples are generated from a bivariate distribution. Equations (3.5) and (3.6) are repeatedly solved until the difference between $R_{\text{individual}}$ and ρ is less than 0.0001 for each sample. For each sample, 10,000 different partitions and therefore 10,000 $R_{\bar{x},\bar{y}}$ values are generated. For each $R_{\bar{x},\bar{y}}$ value, a confidence interval is created and the percent P_r of the 10,000 partitions that contain $R_{\text{individual}}$ is computed. Table 4 shows the minimum percent ($P_r^{\min}$) and maximum percent ($P_r^{\max}$) as a range ($P_r^{\min}, P_r^{\max}$) for all 100 samples. Ideally, the ranges should be tightly centered around 95 %. These ranges do confirm that the confidence intervals can be used to approximate the value of $R_{\text{individual}}$.

5 Conclusion

We have constructed the partition distribution of the Pearson $R_{\bar{x},\bar{y}}$ coefficients generated from group averaged values using random sampling and Monte Carlo simulations. The simulations suggest that the distributions of $R_{\bar{x},\bar{y}}$ can be approximated (for group sizes greater than 8) with a function based on the generalized hypergeometric function (3.8). While the expectation of the $R_{\bar{x},\bar{y}}$ distribution slightly underestimates the individual

Pearson R value, the difference becomes smaller as the number of groups increase. The distribution of $R_{\bar{x},\bar{y}}$ can be transformed into an approximate normal distribution using Fisher's transformation. Confidence intervals for the individual correlation value can then be constructed from the normal distribution. The confidence intervals become more narrow as the number of groups increase and the Pearson $R_{\bar{x},\bar{y}}$ value approaches one.

Acknowledgment: Thank you to Ana Vasilic and Jose Pacheco for their valuable insights and feedback.

Funding: This research is supported by an Institutional Development Award (IDeA) from the National Institute of General Medical Sciences of the National Institutes of Health under grant number P20GM103451. The content is solely the responsibility of the author and does not necessarily represent the official views of the National Institutes of Health.

References

- [1] G. Firebaugh, A rule for inferring individual-level relationships from aggregate data, *Amer. Sociological Rev.* **43** (1978), no. 4, 522–557.
- [2] R. A. Fisher, On the probable error of a coefficient of correlation deduced from a small sample, *Metron* **1** (1921), 3–32.
- [3] H. Gatignon, *Statistical Analysis of Management Data*, 2nd ed., Springer, New York, 2010.
- [4] L. A. Goodman, Ecological regressions and the behavior of individuals, *Amer. Sociological Rev.* **18** (1953), 663–664.
- [5] M. T. Hannan and L. Burstein, Estimation from grouped observations, *Amer. Sociological Rev.* **39** (1974), no. 3, 374–392.
- [6] J. W. Halliwell, Dangers inherent in correlating averages, *J. Educational Res.* **55** (1962), no. 7, 327–329.
- [7] J. L. Hammond, Two sources of error in ecological correlations, *Amer. Sociological Rev.* **38** (1973), no. 6, 764–777.
- [8] L. Irwin and A. J. Lichtman, Across the great divide: Inferring individual level behavior from aggregate data, *Polit. Methodology* **3** (1976), 411–439.
- [9] T. R. Knapp, The unit-of-analysis problem in applications of simple correlation analysis to educational research, *J. Educational Stat.* **2** (1977), no. 3, 171–186.
- [10] R. J. Muirhead, *Aspects of Multivariate Statistical Theory*, John Wiley & Sons, Hoboken, 2005.
- [11] C. Ostroff, Comparing correlations based on individual-level and aggregated data, *J. Appl. Psychol.* **78** (1993), no. 4, 569–582.
- [12] S. Piantasosi, D. P. Byar and S. B. Green, The ecological fallacy, *Amer. J. Epidemiology* **127** (1998), no. 6, 893–904.
- [13] W. S. Robinson, Ecological correlations and the behavior of individuals, *Amer. Sociological Rev.* **15** (1950), 351–357.
- [14] S. Swartz, The fallacy of the ecological fallacy: The potential misuse of a concept and the consequences, *Amer. J. Public Health* **84** (1994), no. 5, 819–824.
- [15] M. Sullivan, *Fundamentals of Statistics. Informed Decision Using Data*, 5th ed., Pearson, Boston, 2018.
- [16] T. Vesala, U. Rannik, M. Leclerc, T. Foken and K. Sabelfeld, Flux and concentration footprints, *Agricultural Forest Meteorol.* **127** (2004), no. 3–4, 111–116.