Research Article

David J. Torres* and Damain Rouson

Investigating the ecological fallacy through sampling distributions constructed from finite populations

https://doi.org/10.1515/mcma-2024-2013
Received February 8, 2024; revised July 12, 2024; accepted July 23, 2024

Abstract: Correlation coefficients and linear regression values computed from group averages can differ from correlation coefficients and linear regression values computed using individual scores. This observation known as the ecological fallacy often assumes that all the individual scores are available from a population. In many situations, one must use a sample from the larger population. In such cases, the computed correlation coefficient and linear regression values will depend on the sample that is chosen and the underlying sampling distribution. The sampling distribution of correlation coefficients and linear regression values for group averages will be identical to the sampling distribution for individuals for normally distributed variables for random samples drawn from infinitely large continuous distributions. However, data that is acquired in practice is often acquired when sampling without replacement from a finite population. Our objective is to demonstrate through Monte Carlo simulations that the sampling distributions for correlation and linear regression will also be similar for individuals and group averages when sampling without replacement from normally distributed variables. These simulations suggest that when a random sample from a population is selected, the correlation coefficients and linear regression values computed from individual scores will not be more accurate in estimating the entire population values compared to samples when group averages are used as long as the sample size is the same.

Keywords: Ecological fallacy, sampling distributions, Pearson R, Monte Carlo simulation, linear regression, multiple regression

MSC 2020: 62J05, 62H10, 62P25

1 Introduction

Linear regression coefficients, the Pearson R correlation, and the coefficient of determination R^2 have long been used to quantify the relationship between dependent and independent variables. The "ecological fallacy" has shown that linear regression and correlation coefficients based on group averages cannot be used to estimate linear regression and correlation coefficients based on individual scores [14]. Shih, Bradley and Yabroff acknowledge the ecological fallacy in health disparities research [16]. A census-based approach was evaluated by Geronimus and Bound [6]. Many have proposed methods to infer individual-level relationships from aggregate data [1, 7, 9]. Aggregate bias was removed by properly specifying the regression equations [8]. A combination of aggregate and individual data has been used [18] to predict associations between particulate matter and COVID-19 mortality.

A mathematical analysis of the ecological fallacy [13] assumes a fixed set of scores that generate different linear regression coefficients or correlation values depending on whether the individual scores are used or whether they are averaged first. However, the scores may not necessarily represent an entire population. In

***Corresponding author: David J. Torres,** Department of Mathematics and Physical Science, Northern New Mexico College, Española, NM 87532, USA, e-mail: davytorres@nnmc.edu
Damain Rouson, Computer Languages and Systems Software Group, Lawrence Berkeley National Laboratory, Berkeley, California, USA, e-mail: rouson@lbl.gov

many situations, the individual scores are themselves a sample. In such situations, the computed values of the correlation and regression coefficients of the sample of individuals are estimates of the population coefficients. The accuracy of the estimates depends on the underlying sampling distribution. Analytical sampling distributions have been derived for the Pearson R coefficient [2], coefficient of determination R^2 (see [4]), and simple regression slope b (see [15]) when individual scores are sampled from bivariate and multivariate normal distributions. The same sampling distributions also describe group averages for normally distributed variables [5, 17]. Thus when sampling is performed from infinitely large normally distributed populations, the population estimate of R, R^2, and b based on a random sample of individual scores will be no more accurate than a random sample of group averages if the sample size (n) is the same. The disadvantage of using group averages is that n can be greater if the individual scores had not been first averaged since the sampling distributions do become narrower and the variances smaller when n increases. The advantage of using group averages is that if the group size m is large $m \geq 30$, the central limit theorem does not require the variables to be normally distributed.

Less is known when sampling without replacement from finite populations which is the way data is acquired in many practical situations. This article employs Monte Carlo simulations to suggest that the R, R^2, and slope distributions are also similar for samples selected without replacement for both individual and group averaged data for equal sample sizes. Our observations afford another interpretation of the ecological fallacy and suggest that for samples drawn from finite populations, the correlation coefficients and linear regression values will be selected from approximately the same sampling distribution regardless of the group size that is used.

The paper is organized as follows. Section 2 introduces the notation used in the paper and describes the analytical sampling distributions of R, R^2, and slope b. Section 3 creates distributions by sampling without replacement from a population of size N using Monte Carlo simulations and compares them with analytical distributions for both simple regression (Section 3.1) and multiple regression (Section 3.2) using a small (0.5 %) and large (25 %) sample percent of the population ($S_n = 100\frac{n}{N}$). Section 4 explores the parameter space of ρ (the population correlation coefficient), N, m, and n further. Section 4.1 considers simple regression correlation, Section 4.2 discusses Fisher's approximation, and Section 4.3 considers the linear regression slope. Mixed groups are simulated in Section 4.4 and non-normal distributions are simulated in Section 4.5. Section 4.6 considers a limited set of multiple regression examples. We conclude and discuss our observations in Section 5.

2 Nomenclature and analytical sampling distributions

We refer the reader to Table 1 which lists the symbols and their descriptions that are used in the manuscript. Muirhead [12] notes that samples $\{(x_i, y_i) : 1 \leq i \leq n\}$ selected from a bivariate distribution,

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

with ρ the continuous population correlation, means μ_x, μ_y, and standard deviations σ_x and σ_y generate the Pearson R distribution

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

where Γ is the gamma function, ${}_2F_1$ is the generalized hypergeometric function. This result was originally derived by Fisher [2]. Fisher [3] also devised a transformation

$$R_z = \frac{1}{2}\ln\Big(\frac{1+R}{1-R}\Big) \tag{2.3}$$

whose distribution approaches a normal distribution

$$Z(R_z) = \frac{1}{\sigma_z\sqrt{2\pi}}\exp\Big(-\frac{1}{2}\Big(\frac{R_z-\mu_z}{\sigma_z}\Big)^2\Big)$$

N	Size of the population
n	Sample size
$S_n = 100\frac{n}{N}$	Sample percent of population
m	Group size
n_{total}	Total number of scores used in sample = nm
k	Number of independent variables
R	Sample correlation coefficient
R^2	Sample coefficient of determination
b	Sample regression slope
$\mathrm{E}(R)$, $\mathrm{E}(R^2)$, $\mathrm{E}(b)$	Expectation of analytical sampling distributions
$\mathrm{Var}(R)$, $\mathrm{Var}(R^2)$, $\mathrm{Var}(b)$	Variance of analytical sampling distributions
$\mathrm{E}(R_{\text{sim}})$, $\mathrm{E}(R^2_{\text{sim}})$, $\mathrm{E}(b_{\text{sim}})$	Expectation of simulation sampling distributions
$\mathrm{Var}(R_{\text{sim}})$, $\mathrm{Var}(R^2_{\text{sim}})$, $\mathrm{Var}(b_{\text{sim}})$	Variance of simulation sampling distributions
ρ	Correlation of continuous distribution
ρ^2	Coefficient of determination of continuous distribution
μ_x, μ_y	Means from bivariate distribution
σ_x, σ_y	Standard deviation from bivariate distribution
$\mathcal{B}$	Bivariate distribution
B	Beta function
${}_2F_1$	Generalized hypergeometric function
D_R	Percent relative difference in Pearson R variance, $100(\mathrm{Var}(R_{\text{sim}}) - \mathrm{Var}(R))/\mathrm{Var}(R)$
D_R^2	Percent relative difference in R^2 variance, $100(\mathrm{Var}(R^2_{\text{sim}}) - \mathrm{Var}(R^2))/\mathrm{Var}(R^2)$
D_b	Percent relative difference in slope b variance, $100(\mathrm{Var}(b_{\text{sim}}) - \mathrm{Var}(b))/\mathrm{Var}(b)$
D_z	Percent relative difference in Fisher variance, $100(\mathrm{Var}(R_z^{\text{sim}}) - \sigma_z^2)/\sigma_z^2$

Table 1: Nomenclature used in manuscript.

as $n \to \infty$, where

$$\mu_z = \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right), \quad \sigma_z^2 = \frac{1}{n-3}. \tag{2.4}$$

Gatignon [5] notes that averages $(\bar{x}_k, \bar{y}_k)$ of size m

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

follow the distribution

$$\mathcal{B}\left(\bar{x}_k, \bar{y}_k; \rho, \mu_x, \mu_y, \frac{\sigma_x}{\sqrt{m}}, \frac{\sigma_y}{\sqrt{m}}\right), \tag{2.5}$$

where $x_i^{(k)}$ and $y_i^{(k)}$ refer to the i'th member of the k'th group which are drawn from a bivariate distribution $\mathcal{B}$, see (2.1). While the standard deviations are different in the arguments of (2.1) and (2.5), the Pearson R distribution (2.2) does not depend on the standard deviations. Therefore the distribution (2.2) also applies to averages $(\bar{x}_k, \bar{y}_k)$ (see [17]).

The expectation $\mathrm{E}(R)$ and variation $\mathrm{Var}(R)$ of $f(R)$ are (see [12])

$$\mathrm{E}(R) = \frac{2\rho}{n-1}\frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{1}{2}(n-1))}{}_2F_1\left(\frac{1}{2}, \frac{1}{2}; \frac{1}{2}(n+1); \rho^2\right), \tag{2.6}$$

$$\mathrm{Var}(R) = 1 - \frac{n-2}{n-1}(1-\rho^2){}_2F_1\left(1, 1; \frac{1}{2}(n+1); \rho^2\right) - \mathrm{E}(R)^2. \tag{2.7}$$

The $\mathrm{Var}(R)$ decreases by approximately $\frac{1}{n-1}$ as n increases (see [12])

$$\mathrm{Var}(R) = \frac{(1-\rho^2)^2}{n-1} + O(n^{-2}). \tag{2.8}$$

In regards to the simple regression slope b, samples of size n drawn from a bivariate distribution generate the distribution $h(b)$ (see [15])

$$h(b) = \frac{(1-\rho^2)^{\frac{n-1}{2}}}{\sqrt{\pi}}\frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n-1}{2})}\frac{\sigma_x}{\sigma_y}\left[1 - \rho^2 + \left(\rho - \frac{\sigma_x}{\sigma_y}b\right)^2\right]^{-\frac{n}{2}}. \tag{2.9}$$

If group averages are used, both σ_x and σ_y will be reduced by the square root of the group size $\sqrt{m}$, but the ratio $\frac{\sigma_x}{\sigma_y}$ will remain the same. Thus the distribution $h(b)$ also applies to group averages for equal group sizes. The expectation $\mathrm{E}(b)$ and variance $\mathrm{Var}(b)$ are (see [15])

$$\mathrm{E}(b) = \rho\Big(\frac{\sigma_y}{\sigma_x}\Big), \tag{2.10}$$

$$\mathrm{Var}(b) = \frac{\sigma_y^2}{\sigma_x^2}\Big(\frac{1-\rho^2}{n-3}\Big). \tag{2.11}$$

If we let

$$t = \frac{(\frac{\sigma_x}{\sigma_y}b - \rho)\sqrt{\nu}}{\sqrt{1-\rho^2}}, \quad \nu = n-1,$$

the distribution (2.9) becomes a t-distribution with ν degrees of freedom which can be approximated by a normal distribution for large n.

When more than one independent variable is used, the distribution of samples of size n drawn from a multivariate normal distribution generate the coefficient of determination $g(R^2)$ distribution as derived by Fisher [4]

$$g(R^2) = \frac{(1-\rho^2)^{\frac{n-1}{2}}}{B(\frac{k}{2}, \frac{n-k-1}{2})}\,{}_2F_1\Big(\frac{n-1}{2}, \frac{n-1}{2}; \frac{k}{2}; \rho^2R^2\Big)(R^2)^{\frac{k-2}{2}}(1-R^2)^{\frac{n-k-3}{2}}, \tag{2.12}$$

where B is the beta function, ρ^2 is the coefficient of determination for the entire population, and k represents the number of independent variables. Again we note that the distribution does not depend on the standard deviations of the multivariate normal distribution. Thus the sampling distribution of group averages will be the same as individual scores. The expectation and variance of $g(R^2)$ are provided by Muirhead [12],

$$\mathrm{E}(R^2) = 1 - \Big(\frac{n-k-1}{n-1}\Big)(1-\rho^2)\,{}_2F_1\Big(1, 1; \frac{n+1}{2}; \rho^2\Big), \tag{2.13}$$

$$\begin{aligned}\mathrm{Var}(R^2) = {} & \Big[\frac{(n-k-1)(n-k+1)}{(n-1)(n+1)}\Big](1-\rho^2)^2\,{}_2F_1\Big(2, 2; \frac{n+3}{2}; \rho^2\Big) \\ & - \Big[\Big(\frac{n-k-1}{n-1}\Big)(1-\rho^2)\,{}_2F_1\Big(1, 1; \frac{n+1}{2}; \rho^2\Big)\Big]^2.\end{aligned} \tag{2.14}$$

3 Method: Monte Carlo simulations

3.1 Comparing simple linear regression distributions

We begin our investigation by generating a population of scores $\{(x_i, y_i) : i = 1, N\}$ of size N, where each pair (x_i, y_i) is generated by sampling from a bivariate distribution (2.1). We generate each pair (x_i, y_i) by sampling x_i^* and y_i^* independently from a standard normal distribution and correlating them using

$$x_i = \sigma_x x_i^* + \mu_x, \quad y_i = \sigma_y(\rho x_i^* + \sqrt{1-\rho^2}y_i^*) + \mu_y, \quad i = 1, N, \tag{3.1}$$

where (μ_x, μ_y) represent the means and (σ_x, σ_y) represent the standard deviations in the bivariate distribution (2.1). The use of (3.1) does not guarantee that $\{(x_i, y_i) : i = 1, N\}$ will be correlated at exactly the value ρ so multiple iterations are performed until they are correlated at the value of ρ to within 1×10^{-4}. We use the computed value of ρ in equations (2.6)–(2.14) when making comparisons.

Subsequently, we randomly select n groups of scores of size m. Since the sampling is done without replacement, we require that $n_{\text{total}} = nm \le N$ since the elements within each group will be unique and no element within the population will be used more than once in any of the groups. We also ensure that each sample is unique so that even if the same group of nm elements are chosen, the group arrangement will be different in each sample. Group averages are formed from the m scores and the Pearson R coefficient and linear regression slope are computed using the n averages. This type of sampling is used to model practical situations in which data is acquired.

Figure 1 shows the results of the Monte Carlo simulations of the Pearson R correlation coefficient and linear regression slope using a population of $N = 10{,}000$ scores, a sample size of $n = 50$, and three different values of $\rho = -0.5, -0.1, 0.3$ each with their respective group size $m = 4$, $m = 3$, and $m = 2$. The population of $N = 10{,}000$ scores was generated with $(\mu_x, \mu_y) = (0, 0)$ and $(\sigma_x, \sigma_y) = (1, 1.2)$ using equations (3.1). The analytical distributions for the Pearson R distribution (2.2) and the linear regression slope (2.9) are shown with solid lines and the simulations are shown using black dots. Each of the three simulations used a million randomly chosen samples to create the distribution. The Monte Carlo simulations visually match the analytical distributions well. However, we note that in these simulations, the sample size $n = 50$ is small compared to the population size $N = 10{,}000$. A sample size of 50 constitutes only an $S_n = 100\frac{n}{N} = 0.5\,\%$ sample percent of the population.

Figure 2 shows the results of the Monte Carlo simulations of the Pearson R correlation coefficient and linear regression slope conducted under the same conditions as Figure 1 except that a population of $N = 400$ scores and a sample size of $n = 100$ are used. The analytical distributions for the Pearson R distribution (2.2) and the linear regression slope (2.9) are shown with solid lines and the simulations are shown using black dots and lines. While the expectation values of the Monte Carlo simulations $\mathrm{E}(R_{\mathrm{sim}})$ and $\mathrm{E}(b_{\mathrm{sim}})$ seem to visibly match the expectation values $\mathrm{E}(R)$ and $\mathrm{E}(b)$ of the analytical distributions, the variances $\mathrm{Var}(R_{\mathrm{sim}})$ and $\mathrm{Var}(b_{\mathrm{sim}})$ of the simulations are visibly smaller than the analytical variances $\mathrm{Var}(R)$ and $\mathrm{Var}(b)$. Note that in these simulations, the sample size $n = 100$ represents $S_n = 100\frac{n}{N} = 25\,\%$ of the population of $N = 400$.

Table 2 shows the differences in the simulated and analytical expected value of R, $\mathrm{E}(R_{\mathrm{sim}}) - \mathrm{E}(R)$, the slope b, $\mathrm{E}(b_{\mathrm{sim}}) - \mathrm{E}(b)$, and the percent relative differences in the variances

$$D_R = 100\left(\frac{\mathrm{Var}(R_{\mathrm{sim}}) - \mathrm{Var}(R)}{\mathrm{Var}(R)}\right), \quad D_b = 100\left(\frac{\mathrm{Var}(b_{\mathrm{sim}}) - \mathrm{Var}(b)}{\mathrm{Var}(b)}\right)$$

for Figures 1 and 2. Table 2 shows that the differences in the expectation are small for both figures. However the percent relative differences in the variance of R and b are large in Figure 2 (ranging between −24.6 %

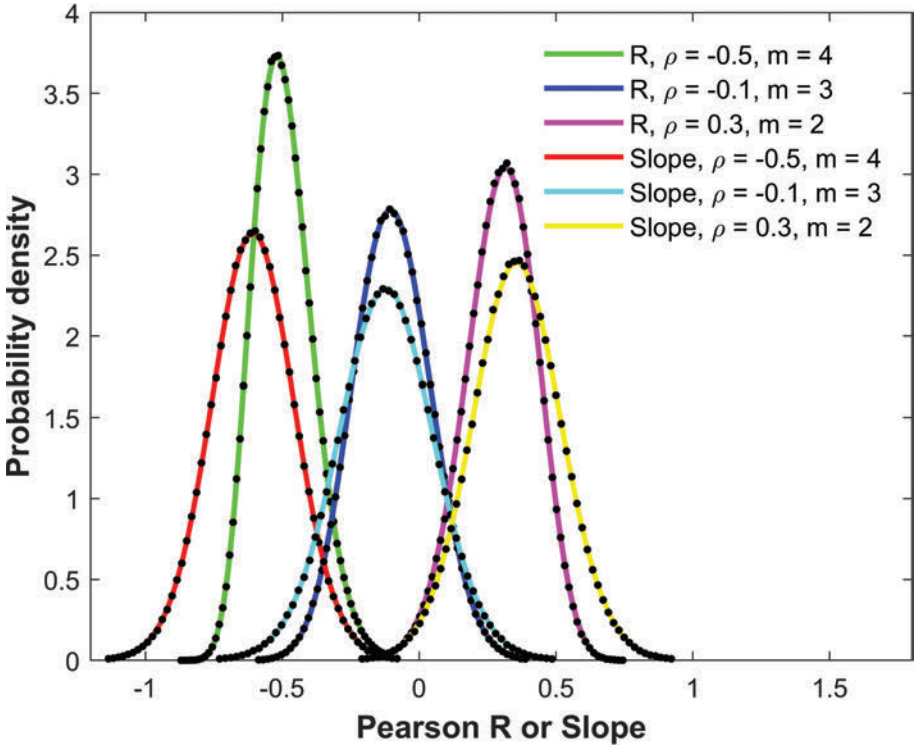

Figure 1: Comparison of the Monte Carlo simulation distributions (black dots) of Pearson R and linear regression slope using $n = 50$ and $N = 10{,}000$ with analytical distributions (2.2) in green, blue, and magenta and (2.9) in red, cyan, and yellow for $\rho = -0.5, -0.1$, and 0.3 with respective group sizes $m = 4$, $m = 3$, and $m = 2$.

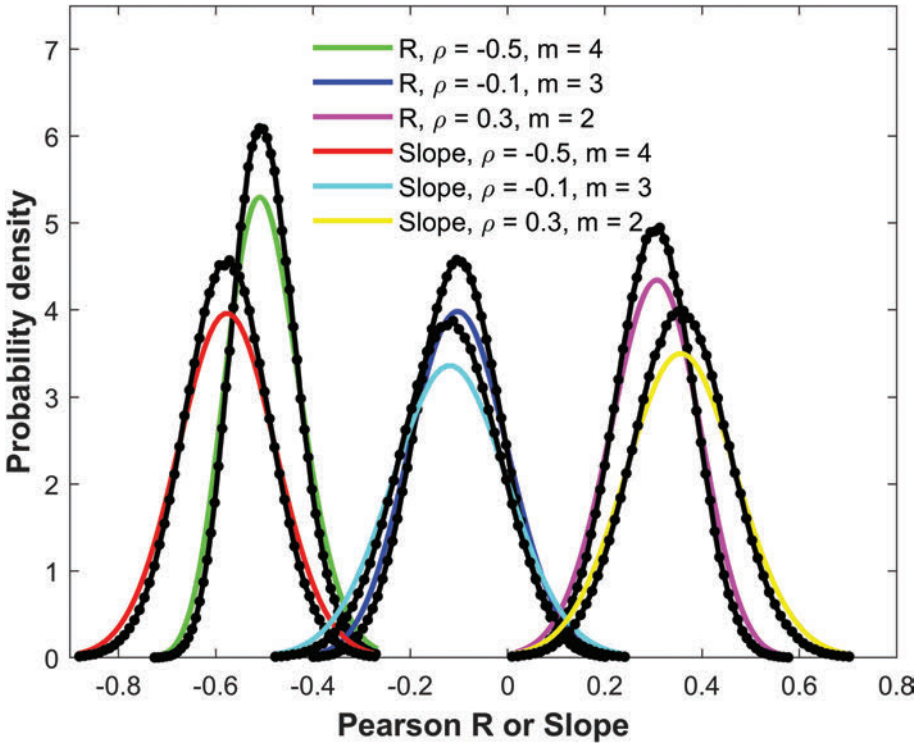

Figure 2: Comparison of Monte Carlo simulation distributions (black dots and lines) of Pearson R and linear regression slope using $n = 100$ and $N = 400$ with analytical distributions (2.2) in green, blue, and magenta and (2.9) in red, cyan, and yellow for $\rho = -0.5, -0.1$, and 0.3 with respective group sizes $m = 4$, $m = 3$, and $m = 2$.

Figure 1: Sample size $n = 50$, Population size $N = 10{,}000$				
ρ	$\mathrm{E}(R_{\mathrm{sim}}) - \mathrm{E}(R)$	Percent relative difference in $\mathrm{Var}(R)$	$\mathrm{E}(b_{\mathrm{sim}}) - \mathrm{E}(b)$	Percent relative difference in $\mathrm{Var}(b)$
−0.5	-5.9×10^{-5}	−0.26	-1.4×10^{-4}	−0.56
−0.1	-2.4×10^{-4}	−0.01	-3.2×10^{-4}	−0.04
0.3	1.7×10^{-4}	−1.45	2.8×10^{-4}	−1.11
Figure 2: Sample size $n = 100$, Population size $N = 400$				
ρ	$\mathrm{E}(R_{\mathrm{sim}}) - \mathrm{E}(R)$	Percent relative difference in $\mathrm{Var}(R)$	$\mathrm{E}(b_{\mathrm{sim}}) - \mathrm{E}(b)$	Percent relative difference in $\mathrm{Var}(b)$
−0.5	4.1×10^{-4}	−24.4	7.6×10^{-5}	−24.6
−0.1	5.8×10^{-5}	−23.4	-3.5×10^{-4}	−24.0
0.3	7.1×10^{-5}	−22.3	-3.9×10^{-4}	−23.2

Table 2: Differences when comparing the analytical and simulation distribution expectation and variance in Figures 1 and 2.

and −22.3 %). We also note that these percent relative differences are similar in magnitude to the sample percent of the population 25 % in Figure 2.

3.2 Comparing multiple regression distributions

For multiple regression simulations, we use the Cholesky decomposition [11] to correlate the variables. If $\mathbf{C}$ represents the $(k+1) \times (k+1)$ positive-definite symmetric correlation matrix, a lower triangular matrix $\mathbf{L}$ is found such that $\mathbf{C} = \mathbf{L}^{\mathrm{T}}\mathbf{L}$. A vector $\vec{z}$ of $k+1$ variables sampled from a standard normal distribution is multiplied by $\mathbf{L}\vec{z}$ to form the correlated variables. In our simulations, one dependent variable z and two independent variables (x and y) are used ($k = 2$).

Figure 3 shows the results of the Monte Carlo simulation that generates the distribution of the coefficient of determination R^2 and the linear regression slope between z and x for a population of $N = 10{,}000$ scores and a sample size of $n = 50$. In the first simulation, the group size is $m = 2$ and $\rho^2 = 0.26$. If R_{zx}, R_{zy} and R_{xy} refer to the Pearson R correlation coefficient between z and x, z and y, and x and y respectively, the correlation matrix **C** has the following off-diagonal elements: $R_{zx} = C_{1,2} = C_{2,1} = -0.5$, $R_{zy} = C_{1,3} = C_{3,1} = 0.1$, $R_{xy} = C_{2,3} = C_{3,2} = -0.055$. In the second simulation, the group size is $m = 10$, $\rho^2 = 0.73$, and the correlation matrix has the following off-diagonal elements: $R_{zx} = C_{1,2} = C_{2,1} = 0.7$, $R_{zy} = C_{1,3} = C_{3,1} = 0.8$, and $R_{xy} = C_{2,3} = C_{3,2} = 0.56$. Each simulation shown with black dots used a million randomly chosen samples. The analytical distributions for R^2 (2.12) are shown using the green and blue solid lines. The solid red and cyan line for the linear regression slope between z and x is generated by sampling **with** replacement and is used as a proxy for the analytical distribution. This was done since an analytical distribution for multiple regression slopes could not be found in our literature search. The simulations visibly match the analytical distributions. Note that the sample size of $n = 50$ is small compared to the population size $N = 10{,}000$ and constitutes only an $S_n = 100\frac{n}{N} = 0.5\,\%$ sample percent of the population.

Figure 4 shows the results of the Monte Carlo simulations of the coefficient of determination R^2 and the linear regression slope between z and x for a population of $N = 600$ scores and a sample size of $n = 150$. In the first simulation, the group size is $m = 2$ and $\rho^2 = 0.25$. The correlation matrix has the following off-diagonal elements: $R_{zx} = C_{1,2} = C_{2,1} = -0.5$, $R_{zy} = C_{1,3} = C_{3,1} = 0.1$, $R_{xy} = C_{2,3} = C_{3,2} = -0.12$. In the second simulation, the group size is $m = 10$, $\rho^2 = 0.71$, and the correlation matrix has the following off-diagonal elements: $R_{zx} = C_{1,2} = C_{2,1} = 0.7$, $R_{zy} = C_{1,3} = C_{3,1} = 0.8$, $R_{xy} = C_{2,3} = C_{3,2} = 0.60$. Each simulation shown with black lines and dots used a million randomly chosen samples. The analytical distributions for R^2 (2.12) are shown using the green and blue solid lines. The solid red and cyan line for slope between z and x is generated by sampling **with** replacement and is used as a proxy for the analytical distribution. While the expectation values of the Monte Carlo simulations $\mathrm{E}(R^2_{\mathrm{sim}})$ and $\mathrm{E}(b_{\mathrm{sim}})$ seem to visibly match the expectation values $\mathrm{E}(R^2)$ and $\mathrm{E}(b)$ of the analytical distributions, the variances $\mathrm{Var}(R^2_{\mathrm{sim}})$ and $\mathrm{Var}(b_{\mathrm{sim}})$ of the simulations are visibly smaller than

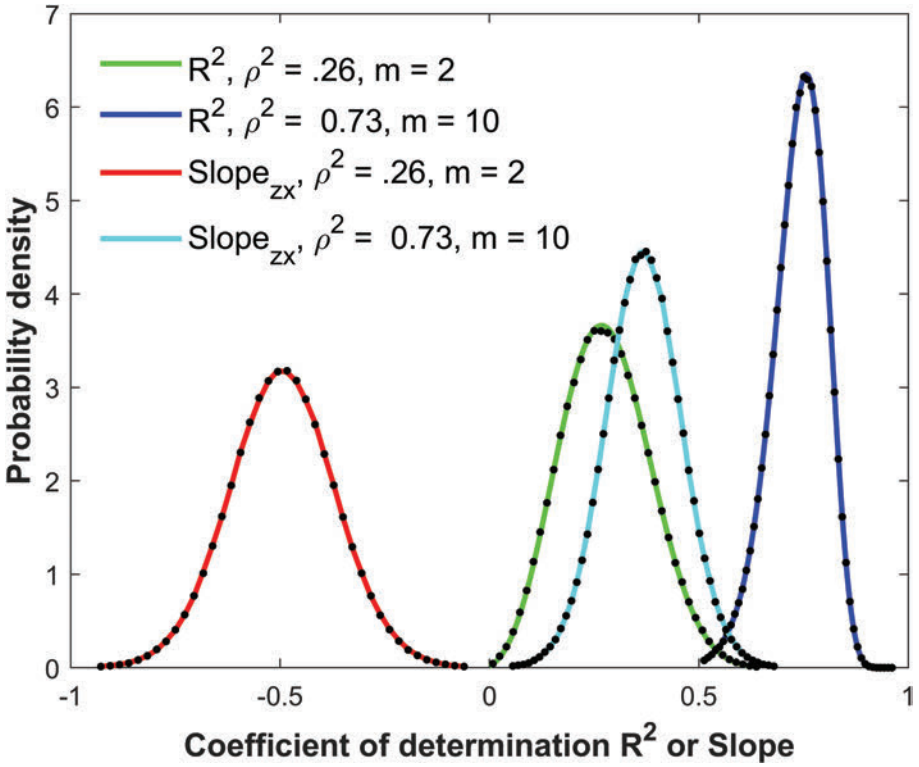

Figure 3: Comparison of Monte Carlo simulations (black dots) and analytical distributions of Pearson R^2 (solid green and blue) and linear regression slope between z and x (solid red and cyan) using a sample size of $n = 50$ and group sizes of $m = 2, 10$ in a population of $N = 10{,}000$.

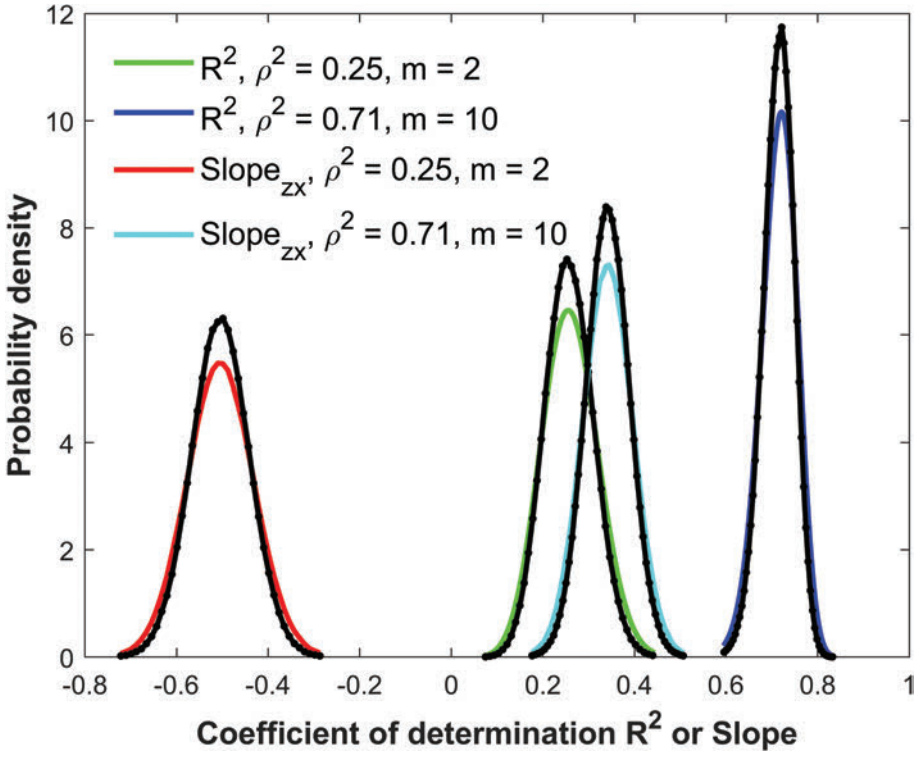

Figure 4: Comparison of Monte Carlo simulations (black lines and dots) and analytical distributions of Pearson R^2 (solid green and blue) and linear regression slope between z and x (solid red and cyan) using a sample size of $n = 150$ and group sizes of $m = 2, 10$ in a population of $N = 600$.

Figure 3: Sample size $n = 50$, Population size $N = 10000$				
ρ^2	$\mathrm{E}(R^2_{\mathrm{sim}}) - \mathrm{E}(R^2)$	**Percent relative differences in Var(R^2)**	$\mathrm{E}(b_{\mathrm{sim}}) - \mathrm{E}(b)$	**Percent relative differences in Var(b)**
0.26	-2.1×10^{-4}	1.1×10^{-2}	-5.6×10^{-5}	-5.3×10^{-3}
0.73	-7.1×10^{-5}	-7.6×10^{-3}	-2.8×10^{-5}	-5.2×10^{-3}
Figure 4: Sample size $n = 150$, Population size $N = 600$				
ρ^2	$\mathrm{E}(R^2_{\mathrm{sim}}) - \mathrm{E}(R^2)$	**Percent relative differences in Var(R^2)**	$\mathrm{E}(b_{\mathrm{sim}}) - \mathrm{E}(b)$	**Percent relative differences in Var(b)**
0.25	-1.7×10^{-3}	−23.0	-1.8×10^{-4}	−24.3
0.71	-3.5×10^{-4}	−24.7	4.2×10^{-7}	−23.5

Table 3: Differences when comparing the analytical and simulation distributions for Figures 3 and 4.

the analytical variances $\mathrm{Var}(R^2)$ and $\mathrm{Var}(b)$. Note that in these simulations, the sample size $n = 150$ represents $S_n = 100\frac{n}{N} = 25\,\%$ of the population of $N = 600$.

Table 3 shows the differences in the simulated and analytical expected value of R^2, $\mathrm{E}(R^2_{\mathrm{sim}}) - \mathrm{E}(R^2)$, the linear regression slope b between z and x, $\mathrm{E}(b_{\mathrm{sim}}) - \mathrm{E}(b)$, and the percent relative difference in the variances

$$D_R^2 = 100\left(\frac{\mathrm{Var}(R^2_{\mathrm{sim}}) - \mathrm{Var}(R^2)}{\mathrm{Var}(R^2)}\right), \quad D_b = 100\left(\frac{\mathrm{Var}(b_{\mathrm{sim}}) - \mathrm{Var}(b)}{\mathrm{Var}(b)}\right)$$

for Figures 3 and 4. Table 3 shows that the differences in the expectation are small for both figures. However the percent relative difference in the variance of R^2 and b are large in Figure 4 (ranging between −24.7 % and −23.0 %). We also note that these percent relative differences are similar in magnitude to the sample percent of the population $S_n = 25\,\%$ in Figure 4.

These first examples suggest that when sampling without replacement from normal distributions, the expectation differences between the analytical and simulated distributions are small regardless of the group

size m. Differences arise between the variance of the analytical and simulated distributions when the sample size n is a significant percent of the population size N regardless of the group size m. In the next section, we explore parameter space further to determine if the differences in the analytical and simulated expectations in R, R^2 and the linear regression slope remain small. We also investigate if the differences in variances are a function of the sample percent of the population $S_n = 100\frac{n}{N}$.

4 Results: Exploring parameter space

Our objective in Section 4 is to continue to explore the parameter space of ρ, m, n, and N further to determine if and where differences exist between the analytical distributions and the sampling distributions created without replacement. The parallel Fortran code is available at: https://github.com/davytorres/MonteCarloEcological.git. In our preliminary analysis in Section 3, only one population was chosen for the simulations. We increase the number of populations to 112 (chosen because it is divisible by 8 and 14 which are the number of processors available on our computers for parallel runs) to determine if the population selection affects the observed trends. In all the simulations in Section 4, a million randomly chosen samples are used to create each distribution. The most important parameter we identified in Section 3 was the sample percent of the total population $S_n = 100\frac{n}{N}$ which affects the variances of the distributions. Other parameters such as the group size and the Pearson ρ value did not seem to have a significant impact when comparing the analytical and simulated distributions. We also explore mixed groups sizes and non-normal distributions.

4.1 Simple regression – Pearson R

Figure 5 plots the difference $(\mathrm{E}(R_{\mathrm{sim}}) - \mathrm{E}(R))$ using a population size $N = 10{,}000$ sampled from a bivariate distribution with $\rho = 0.7$, $(\mu_x, \mu_y) = (0, 0)$, and $(\sigma_x, \sigma_y) = (1, 1)$; $\mathrm{E}(R)$ is computed using equation (2.6); $\mathrm{E}(R_{\mathrm{sim}})$ is generated with Monte Carlo simulations that use different sample sizes n (displayed along the horizontal axis using the sample percent of the population $S_n = 100\frac{n}{N}$) and different group sizes $m = 1, 2, 5, 10$. The symbol on the plot (square, circle, diamond, or triangle) shows the average difference and the error bars show the maximum and minimum differences from 112 different populations with size $N = 10{,}000$. Due to the number of computations required, simulations were run in parallel on fourteen processors on a laptop or eight processors on a desktop using MPI Fortran. Note that the sample percent of the population is limited according to the equation $nm \leq N$. For example, a group size of $m = 2$ can only use 50 % of the population. For this reason, the values we use for S_n along the horizontal axis for each group size vary according to $S_n = 10\frac{i}{m}$, $i = 1, 10$. The exception is $m = 1$. We modify the largest value of S_n when $m = 1$ and $i = 10$ to be 95 since only one sample can be chosen when $m = 1$ and all the elements of the population are used.

Figure 5 shows that the difference in $(\mathrm{E}(R_{\mathrm{sim}}) - \mathrm{E}(R))$ is small $(< 2 \times 10^{-4})$ confirming that the expectation values can be approximated using (2.6) regardless of the sample size n and group size m for normally distributed variables and $\rho = 0.7$. There is a small positive offset in the average difference. Simulations (figure not shown to save space) identical to Figure 5 were also conducted except that a small population $N = 400$ was used. In these simulations, all the absolute average differences were small $(< 5 \times 10^{-4})$ but larger than the average differences shown in Figure 5.

We also note that the range of the error bars in Figure 5 increases as the sample percent of the population decreases. Define the range L_R of the error bars using (4.1),

$$L_R = \mathrm{Max}\{\mathrm{E}(R_{\mathrm{sim}}) - \mathrm{E}(R)\} - \mathrm{Min}\{\mathrm{E}(R_{\mathrm{sim}}) - \mathrm{E}(R)\}. \tag{4.1}$$

If the $\mathrm{Log}_{10}(L_R)$ is plotted against $\mathrm{Log}_{10}(S_n)$, a linear regression line can be fit through the data as shown in Figure 6 for $m = 1$ and $m = 10$. To avoid an overly cluttered plot, the plots for $m = 2$ and $m = 5$ are not shown. Additional points for small values of S_n are included in the plot $\{S_n = \frac{i}{m} : i = 1, 10\}$ in addition to the points $\{S_n = 10\frac{i}{m} : i = 1, 10\}$ included in Figure 5. Based on the slope of linear regression fit, the range of the error

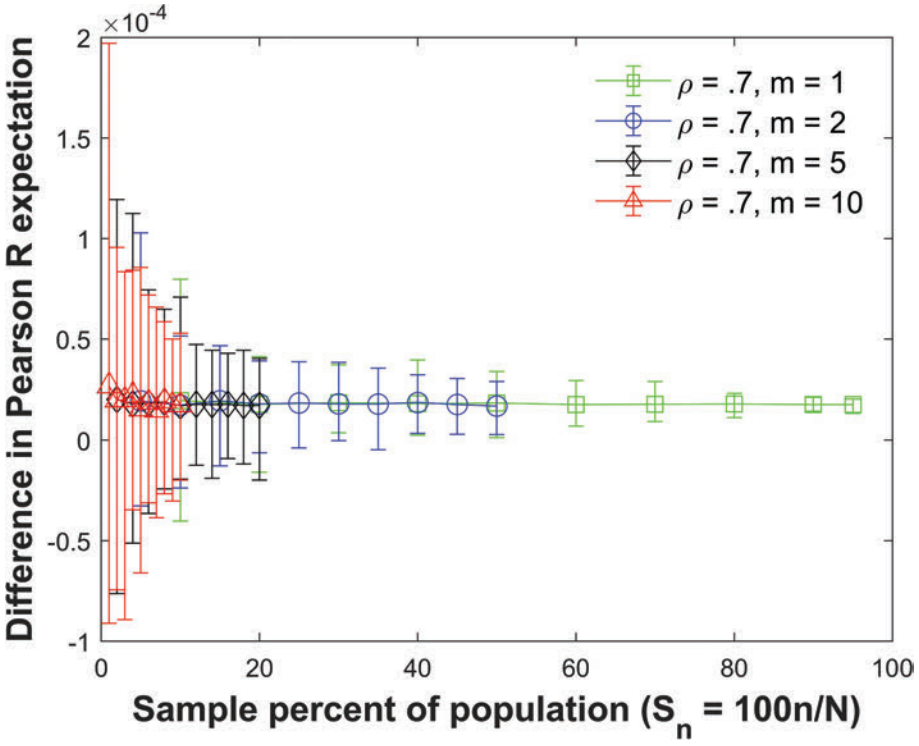

Figure 5: Plot of difference $(\mathrm{E}(R_{\mathrm{sim}}) - \mathrm{E}(R))$ using a population size $N = 10{,}000$ sampled from a bivariate distribution with $\rho = 0.7$ plotted against the sample percent of the population $S_n = 100\frac{n}{N}$. The plot symbol shows the average difference and the error bars show the maximum and minimum differences using 112 different populations of size $N = 10{,}000$.

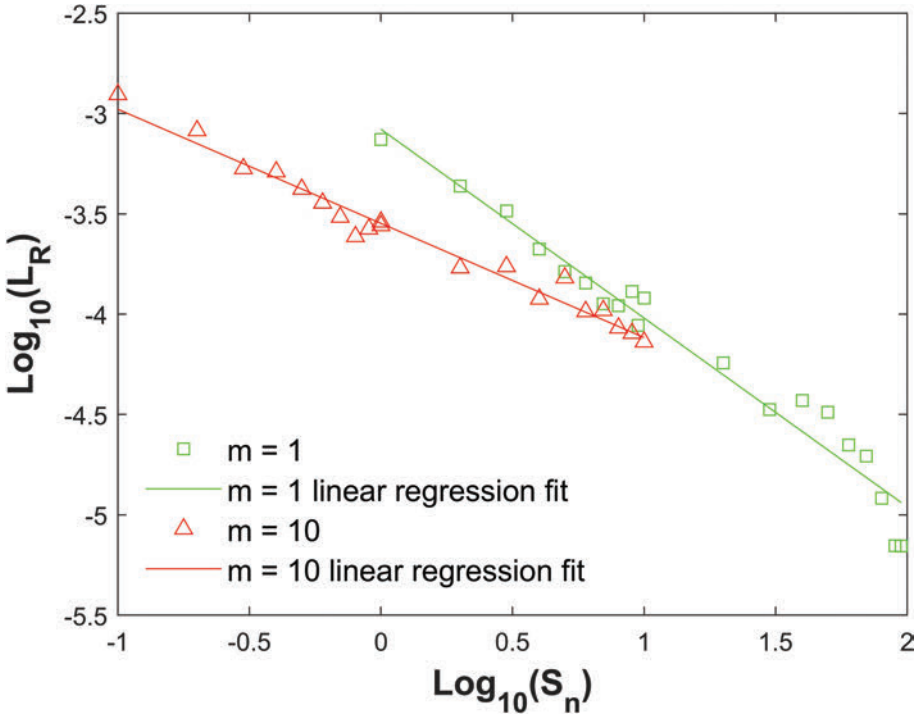

Figure 6: Plot of $\mathrm{Log}_{10}(L_R)$ vs $\mathrm{Log}_{10}(S_n)$ using a population size $N = 10{,}000$ sampled from a bivariate distribution with $\rho = 0.7$ for $m = 1$ and $m = 10$. The slopes of the linear regression lines are: -0.94 ($m = 1$), -0.71 ($m = 2$), -0.61 ($m = 5$), -0.57 ($m = 10$). The plots for $m = 2$ and $m = 5$ are not shown to avoid a cluttered plot.

bars L_R varies according to S_n^{α}, where $\alpha = -0.94$ ($m = 1$), $\alpha = -0.71$ ($m = 2$), $\alpha = -0.61$ ($m = 5$), and $\alpha = -0.57$ ($m = 10$) in their respective ranges of S_n.

Figure 7 plots the difference $(\mathrm{E}(R_{\mathrm{sim}}) - \mathrm{E}(R))$ using a population size $N = 10{,}000$ sampled from a bivariate distribution with different values of $\rho = \{-0.9, -0.6, -0.3, 0.0, 0.2, 0.5, 0.7\}$ and group size $m = 2$. Figure 8 shows the results for $m = 10$. Except for the value of ρ, the Monte Carlo simulation is conducted under the same conditions described in Figure 5. The plot symbol shows the average difference and the error bars show the maximum and minimum differences using 112 different populations. The differences remain small regardless of the value of ρ used. The expectation $\mathrm{E}(R_{\mathrm{sim}})$ is slightly larger than $\mathrm{E}(R)$ for $\rho > 0$, and $\mathrm{E}(R_{\mathrm{sim}})$ is slightly less than $\mathrm{E}(R)$ for $\rho < 0$.

Figure 9 plots the percent relative difference $D_R = 100(\frac{\mathrm{Var}(R_{\mathrm{sim}}) - \mathrm{Var}(R)}{\mathrm{Var}(R)})$ using a population size $N = 10{,}000$ sampled from a bivariate distribution with $\rho = 0.7$. The plot symbol (square, circle, diamond, or triangle) shows the average difference and the error bars show the maximum and minimum differences using 112 different populations. The variance $\mathrm{Var}(R)$ is computed using (2.7) and $\mathrm{Var}(R_{\mathrm{sim}})$ is generated using a Monte Carlo simulation in which a million samples were selected without replacement using different sample sizes n (displayed along the horizontal axis using the sample percent of the population $S_n = 100\frac{n}{N}$).

The average percent relative differences can be closely approximated by the linear plot

$$D_R = -S_n \tag{4.2}$$

which is **not a function of the group size** m. The size of the error bars increases as the sample percent of the population decreases which was also observed in the Pearson R expectation plot (Figure 5).

It is well known if all samples are considered, the variance of the sampling distribution σ_m^2 is reduced by the group size m when sampling with replacement

$$\sigma_m^2 = \frac{\sigma^2}{m}.$$

When computing the variance when sampling without replacement $(\sigma_m^2)_{wr}$, there is an additional finite population correction $\frac{N-m}{N-1}$,

$$(\sigma_m^2)_{wr} = \frac{\sigma^2}{m}\Big(\frac{N-m}{N-1}\Big).$$

The percent relative difference between the two variances is

$$100\Big(\frac{(\sigma_m^2)_{wr} - \sigma_m^2}{\sigma_m^2}\Big) = -100\Big(\frac{m-1}{N-1}\Big) \tag{4.3}$$

which is similar to the approximation (4.2) $D_R = -100\frac{n}{N} = -S_n$ (when n is substituted for $m-1$ and $N-1$ replaced with N). However, (4.3) does not apply to the Pearson R coefficient and linear regression slope since the variance is computed from the Pearson R coefficient and slope after they are calculated from each sample composed of n groups each of size m.

Figure 10 plots the error (from Figure 9) when approximating the plot of D_R versus S_n with the line $D_R = -S_n$. The error plotted is $D_R + S_n$. All absolute errors are less than 0.5 %. We note that the absolute error increases when the group size decreases and the sample percent of the population S_n decreases.

Simulations (not shown) identical to Figure 10 were conducted except that a small population $N = 400$ was used. In these simulations, all the absolute average errors were small (< 1.2 %) but larger than the average errors shown in Figure 10.

To further test the validity of (4.2), we fit a linear regression line through the plot of D_R versus S_n shown in Figure 9 for different group sizes $m = 1, 2, 5, 10$ and different values of $\rho = -0.9, -0.6, -0.3, 0.0, 0.2, 0.5$, and 0.7. The linear regression line fit through all values of m and all values of ρ has a Pearson R value in the range $-1 \le R < -1 + (2 \times 10^{-5})$, a slope b in the range $-1 - (1 \times 10^{-3}) < b < -1 + (7 \times 10^{-3})$, and a y-intercept y_{int} in the range $-6 \times 10^{-3} < y_{\mathrm{int}} < 1 \times 10^{-3}$ which shows that (4.2) is a good approximation to the plot of D_R vs S_n.

The ecological fallacy can be viewed from the following perspective for normally distributed variables and random sampling. Note that the total number of scores n_{total} used for each S_n is mn since n groups of m scores are averaged before the Pearson R coefficient is computed. For a fixed n_{total}, smaller group sizes (m) have the advantage of allowing larger sample sizes as $n = \frac{n_{\mathrm{total}}}{m}$. The variance of the sampling distribution decreases as the

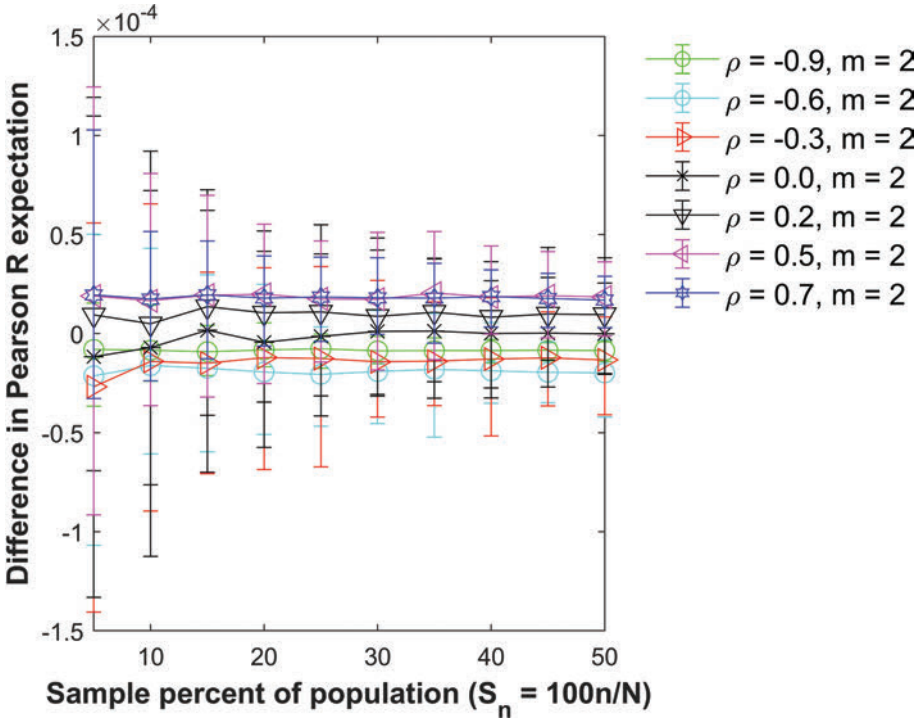

Figure 7: Plot of difference $(\mathrm{E}(R_{\mathrm{sim}}) - \mathrm{E}(R))$ using a population size $N = 10{,}000$ sampled from a bivariate distribution with different values of ρ with group size $m = 2$ plotted against the sample percent of the population $S_n = 100\frac{n}{N}$. The plot symbol shows the average difference and the error bars show the maximum and minimum differences using 112 different populations of size $N = 10{,}000$.

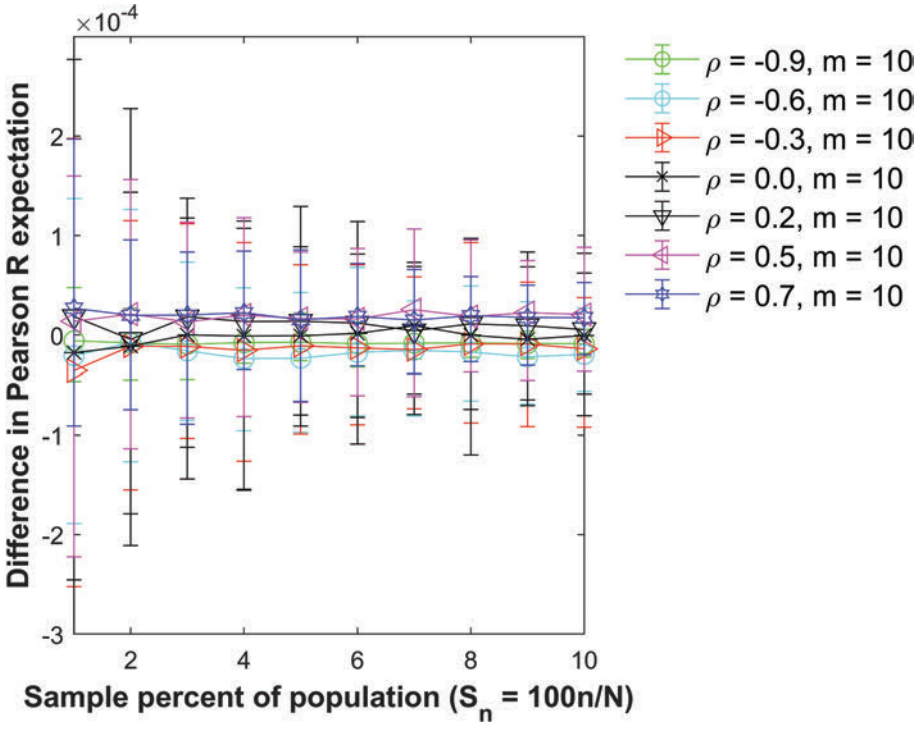

Figure 8: Plot of difference $(\mathrm{E}(R_{\mathrm{sim}}) - \mathrm{E}(R))$ using a population size $N = 10{,}000$ sampled from a bivariate distribution with different values of ρ with group size $m = 10$ plotted against the sample percent of the population $S_n = 100\frac{n}{N}$. The plot symbol shows the average difference and the error bars show the maximum and minimum differences using 112 different populations of size $N = 10{,}000$.

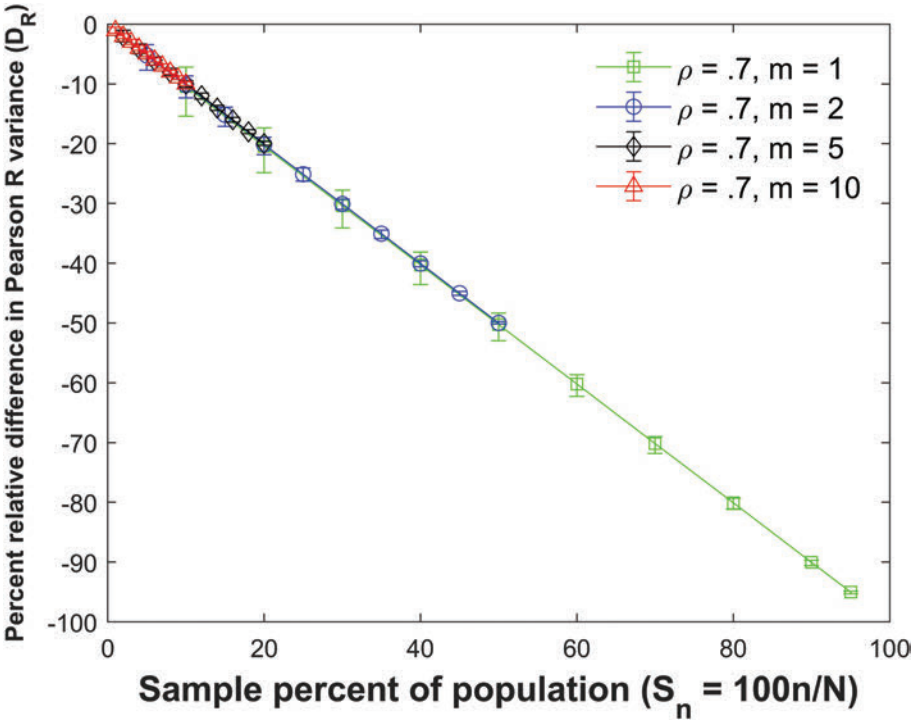

Figure 9: Plot of percent relative difference $D_R = 100(\frac{\text{Var}(R_{\text{sim}})-\text{Var}(R)}{\text{Var}(R)})$ using a population size $N = 10{,}000$ sampled from a bivariate distribution with $\rho = 0.7$ plotted against the sample percent of the population $S_n = 100\frac{n}{N}$. The plot symbol shows the average difference and the error bars show the maximum and minimum differences using 112 different populations. The percent relative difference can be described by the linear equation $D_R = -S_n$. The range of the error bars increases as the sample percent of the population decreases and the group size decreases.

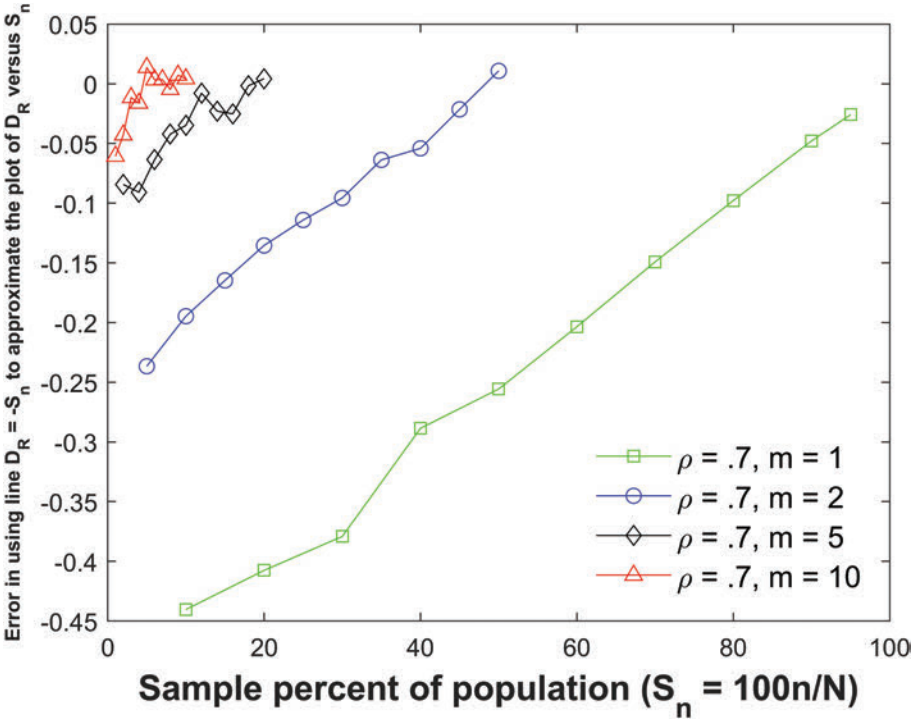

Figure 10: Plot of error $D_R + S_n$ (from Figure 9) when approximating the plot of D_R versus S_n with the line $D_R = -S_n$. The average value of the D_R from the 112 populations is used in the plot. The absolute error increases when the group size decreases and the sample percent of the population S_n decreases.

sample size (n) increases for two reasons. First $\mathrm{Var}(R)$ scales as $\frac{1}{n-1}$ according to (2.8). Secondly, when sampling without replacement, the variance is further reduced according to the $D_R = -S_n$ from Figure 9. Thus for a fixed set of available scores n_{total}, smaller group sizes translate into larger sample sizes and smaller variances in the sampling distributions. Smaller variances mean the sample correlation has a higher probability of being close to the population correlation.

4.2 Testing Fisher's approximation

Since the distribution of R is not symmetric about the mean, we investigated properties of the Fisher transformed (2.3) values of R as the transformed distribution is approximately normal and is useful in building confidence intervals. Recall from (2.4) that as $n \to \infty$, the Fisher transformed variables $R_z = \frac{1}{2}\ln(\frac{1+R}{1-R})$ approach a normal distribution with mean

$$\mu_z = \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right)$$

and variance

$$\sigma_z^2 = \frac{1}{n-3}.$$

Figures 11–13 use the same data and parameters from the simulations described in Figures 5–10.

Figure 11 plots the average difference $(\mathrm{E}(R_z^{\text{sim}}) - \mu_z)$ over the 112 populations where $\mathrm{E}(R_z^{\text{sim}})$ is the expectation of the distribution formed by the transformed

$$R_z = \frac{1}{2}\ln\left(\frac{1+R}{1-R}\right)$$

variables created by Monte Carlo simulations with $\rho = 0.7$. We note that the difference is small but increases as S_n decreases. A plot of $\mathrm{Log}_{10}(\mathrm{E}(R_z^{\text{sim}}) - \mu_z)$ vs $\mathrm{Log}_{10}(S_n)$ is linear and the least squares slope of the line is -1.1 with a Pearson R value of -0.997, from which one can conclude that $(\mathrm{E}(R_z^{\text{sim}}) - \mu_z)$ decreases as S_n^{α}, where $\alpha = -1.1$ for the ranges of S_n considered.

Figure 12 plots the average percent relative difference in the variance

$$\begin{aligned} D_z &= 100\left(\frac{\mathrm{Var}(R_z^{\text{sim}}) - \sigma_z^2}{\sigma_z^2}\right) \\ &= 100(n-3)\left(\mathrm{Var}(R_z^{\text{sim}}) - \frac{1}{n-3}\right), \end{aligned}$$

where $\mathrm{Var}(R_z^{\text{sim}})$ is the variance of the distribution formed by the Monte Carlo simulation after applying the Fisher transformation. Note that D_z can be approximated with the line $D_z = -S_n$. The absolute errors in approximating the plot of D_z vs S_n with the line $D_z = -S_n$ are all less than 0.5 percent. As with D_R, the absolute error increases when the group size decreases and the sample percent of the population S_n decreases.

We have compared the expectation and variances of the Pearson R distribution. To compare the simulated and analytical distribution functions, we use the Fisher transformed variables. Figure 13 plots the average error

$$E_{\text{area}} = \int_{-3.5}^{3.5} |f_{\text{sim}}(R_z) - Z(R_z)|\, dz,$$

where

$$Z(R_z) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{R_z - \mu}{\sigma}\right)^2\right)$$

and $f_{\text{sim}}(R_z)$ is the distribution formed by the Monte Carlo simulation. Since the variance decreases with S_n when sampling without replacement, the mean μ and standard deviation σ were selected to be the mean and standard deviation of the simulated distribution so Figure 13 is essentially a test of normality; E_{area} is small ($< 5.5 \times 10^{-3}$) for the tested range of S_n for all group sizes m.

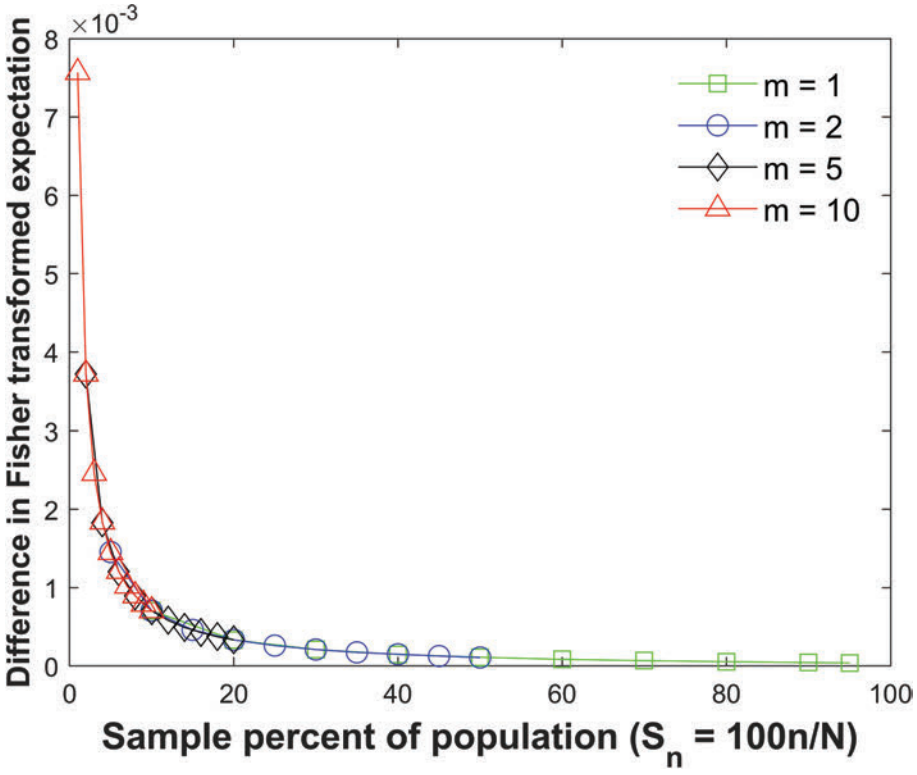

Figure 11: Plot of difference $(\mathrm{E}(R_z^{\mathrm{sim}}) - \mu_z)$, where $\rho = 0.7$, $\mu_z = \frac{1}{2}\ln(\frac{1+\rho}{1-\rho})$, and $\mathrm{E}(R_z^{\mathrm{sim}})$ is the expectation of the distribution formed by the Monte Carlo simulation after applying the transformation $R_z = \frac{1}{2}\ln(\frac{1+R}{1-R})$.

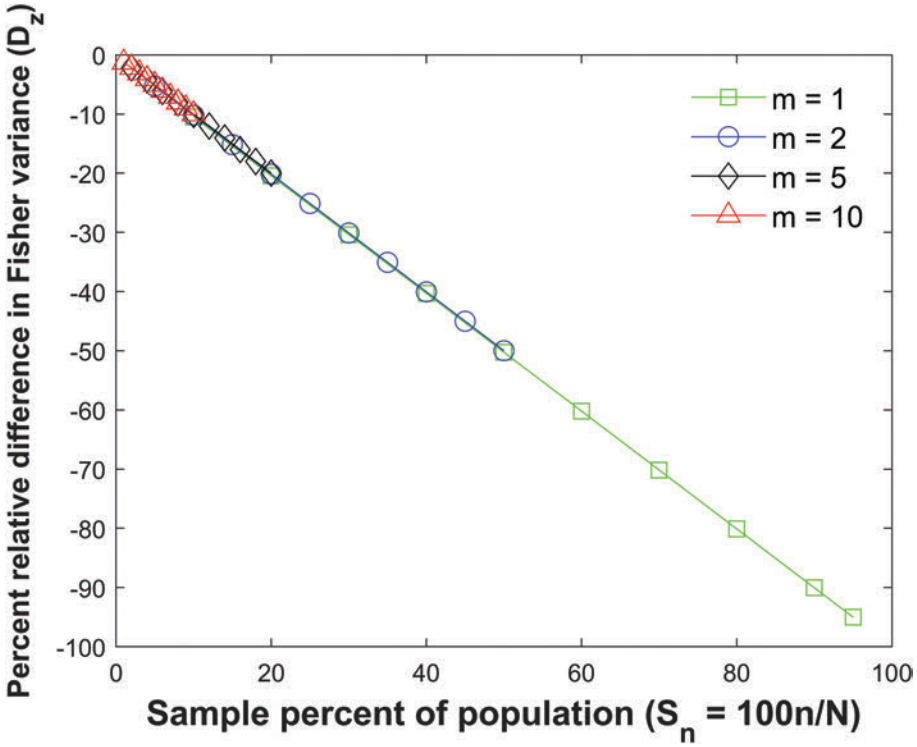

Figure 12: Plot of percent relative difference $100(n-3)(\mathrm{Var}(R_z) - \frac{1}{n-3})$, where $\mathrm{Var}(R_z^{\mathrm{sim}})$ is the variance of the distribution formed by the Monte Carlo simulation after applying the transformation $R_z = \frac{1}{2}\ln(\frac{1+R}{1-R})$ with $\rho = 0.7$.

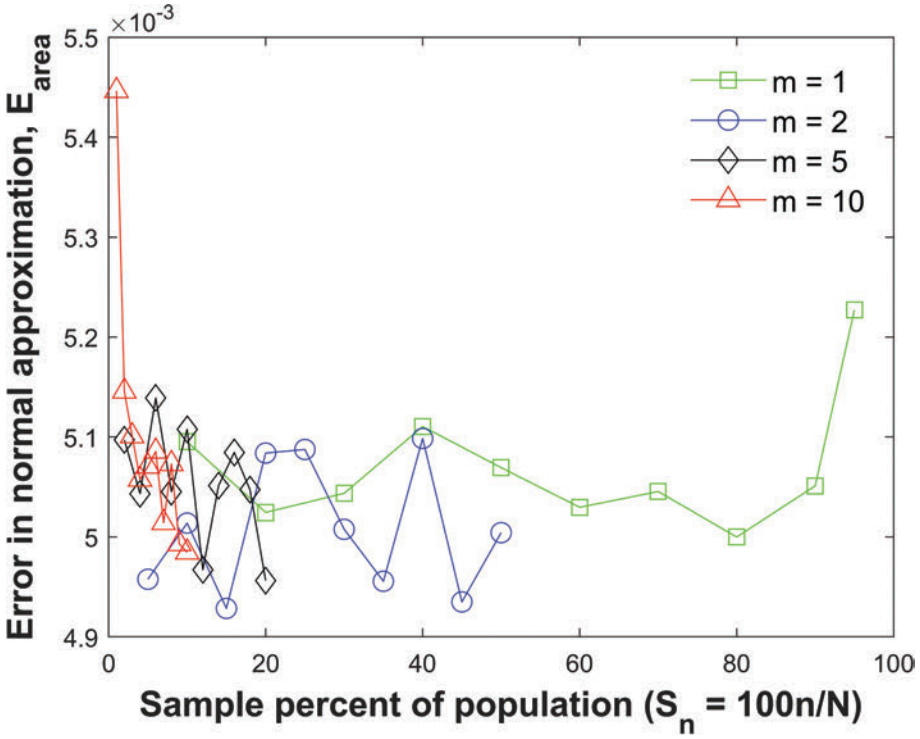

Figure 13: Plot of error $E_{\text{area}} = \int_{-3.5}^{3.5} |f_{\text{sim}}(R_z) - Z(R_z)|\, dz$, where $Z(R_z) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{1}{2}(\frac{R_z-\mu}{\sigma})^2)$ and $f_{\text{sim}}(R_z)$ is the distribution formed by the Monte Carlo simulation after applying the transformation $R_z = \frac{1}{2}\ln(\frac{1+R}{1-R})$.

4.3 Simple regression – Slope

Figures 14–18 use the same data from the simulations described in Figures 5–10. Figure 14 plots the difference $(\mathrm{E}(b_{\text{sim}}) - \mathrm{E}(b))$. The difference $(\mathrm{E}(b_{\text{sim}}) - \mathrm{E}(b))$ is small confirming that the expectation values can be approximated using (2.10) regardless of the sample size n and group size m for normally distributed variables and $\rho = 0.7$. Simulations identical to Figure 14 were also conducted except that a small population $N = 400$ was used. In these simulations, all the absolute average differences were small ($< 2 \times 10^{-4}$) but larger than the average differences shown in Figure 14.

Similar to Figure 5, the range of the error bars in Figure 14 increases as the sample percent of the population decreases. Define the range L_b of the error bars using (4.4),

$$L_b = \mathrm{Max}\{\mathrm{E}(b_{\text{sim}}) - \mathrm{E}(b)\} - \mathrm{Min}\{\mathrm{E}(b_{\text{sim}}) - \mathrm{E}(b)\}. \tag{4.4}$$

If the $\mathrm{Log}_{10}(L_b)$ is plotted against $\mathrm{Log}_{10}(S_n)$, a linear regression line can be fit through the data. Additional points for small values of S_n are included in the plot $\{S_n = \frac{i}{m} : i = 1, 10\}$ in addition to the points $\{S_n = 10\frac{i}{m} : i = 1, 10\}$. Based on the slope of linear regression fit, the range of the error bars L_b varies according to S_n^{α}, where $\alpha = -0.92$ $(m = 1)$, $\alpha = -0.72$ $(m = 2)$, $\alpha = -0.60$ $(m = 5)$, $\alpha = -0.58$ $(m = 10)$ for their respective ranges of S_n.

Figure 15 plots the difference $(\mathrm{E}(b_{\text{sim}}) - \mathrm{E}(b))$ using a population size $N = 10{,}000$ sampled from a bivariate distribution with different values of $\rho = \{-0.9, -0.6, -0.3, 0.0, 0.2, 0.5, 0.7\}$ and group size $m = 2$. Figure 16 shows the results for $m = 10$. The differences remain small regardless of the value of ρ used.

Figure 17 plots the percent relative difference

$$D_b = 100\left(\frac{\mathrm{Var}(b_{\text{sim}}) - \mathrm{Var}(b)}{\mathrm{Var}(b)}\right)$$

using a population size $N = 10{,}000$ sampled from a bivariate distribution with $\rho = 0.7$. $\mathrm{Var}(b)$ is computed using (2.11) and $\mathrm{Var}(b_{\text{sim}})$ is generated using a Monte Carlo simulation in which a million samples were selected without replacement. The average percent relative difference can be described by the linear plot

$$D_b = -S_n, \tag{4.5}$$

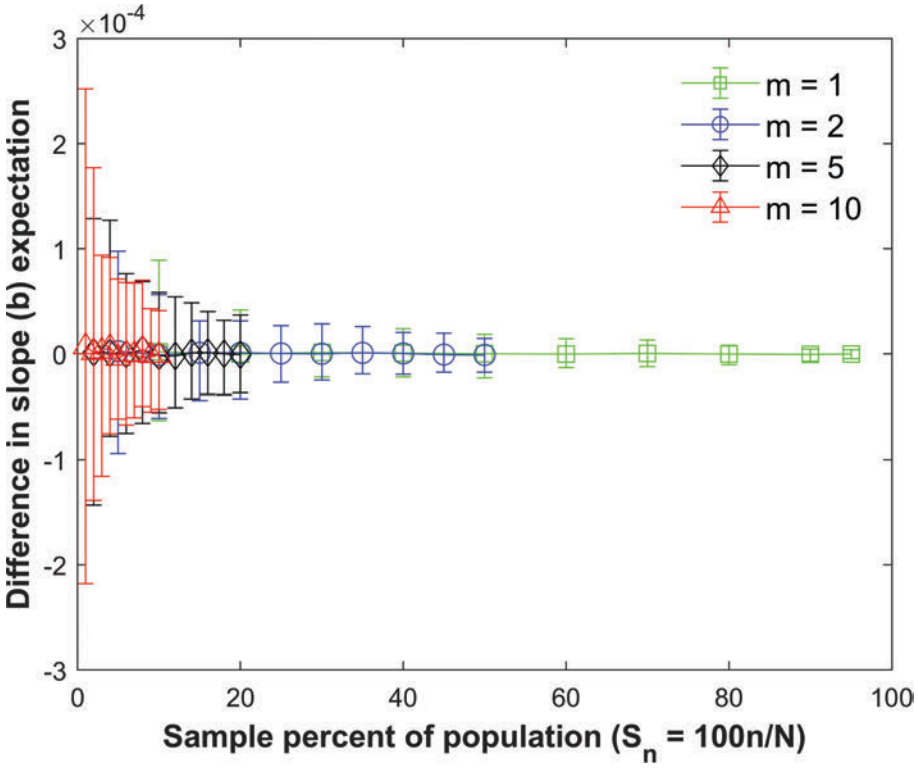

Figure 14: Plot of difference ($E(b_{sim}) - E(b)$) using a population size $N = 10{,}000$ sampled from a bivariate distribution with $\rho = 0.7$. The plot symbol shows the average difference and the error bars show the maximum and minimum differences using 112 different populations.

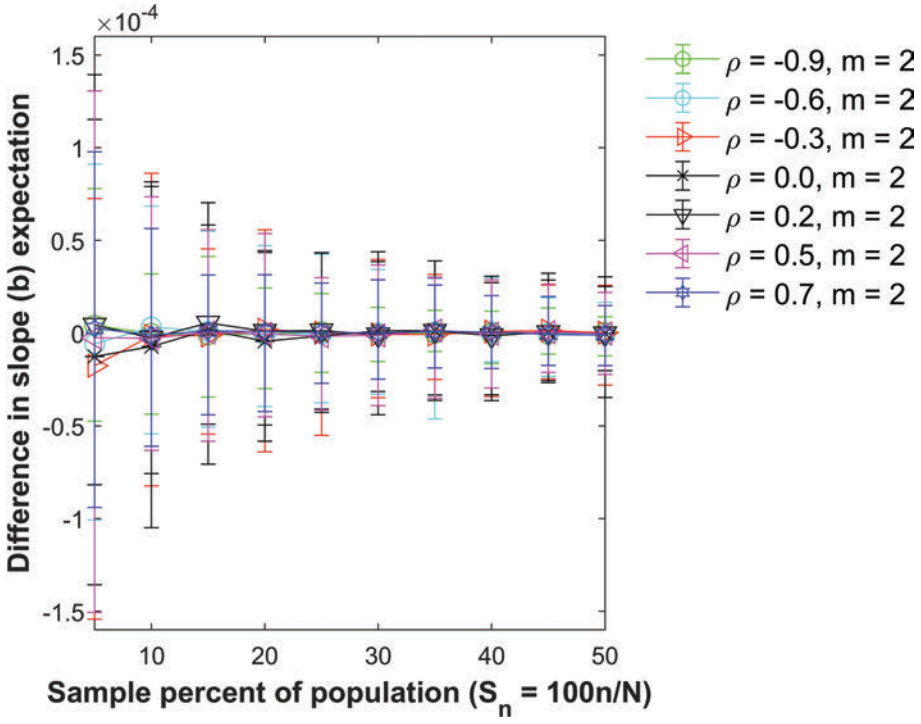

Figure 15: Plot of difference ($E(b_{sim}) - E(b)$) using a population size $N = 10{,}000$ sampled from a bivariate distribution with different values of ρ with group size $m = 2$ plotted against the sample percent of the population S_n.

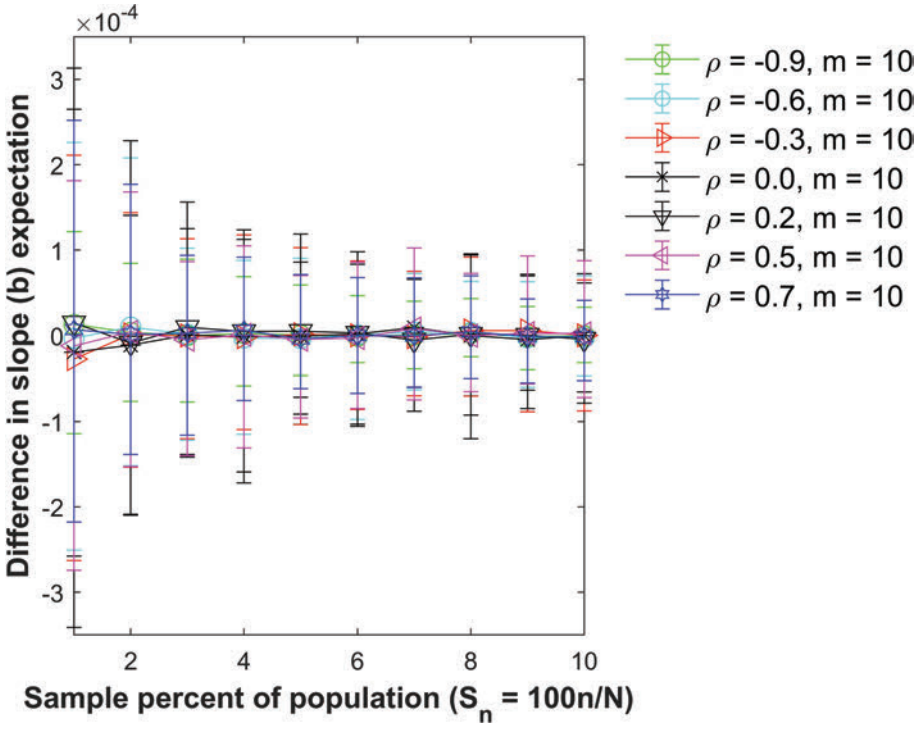

Figure 16: Plot of difference ($E(b_{sim}) - E(b)$) using a population size $N = 10{,}000$ sampled from a bivariate distribution with different values of ρ with group size $m = 10$ plotted against the sample percent of the population S_n.

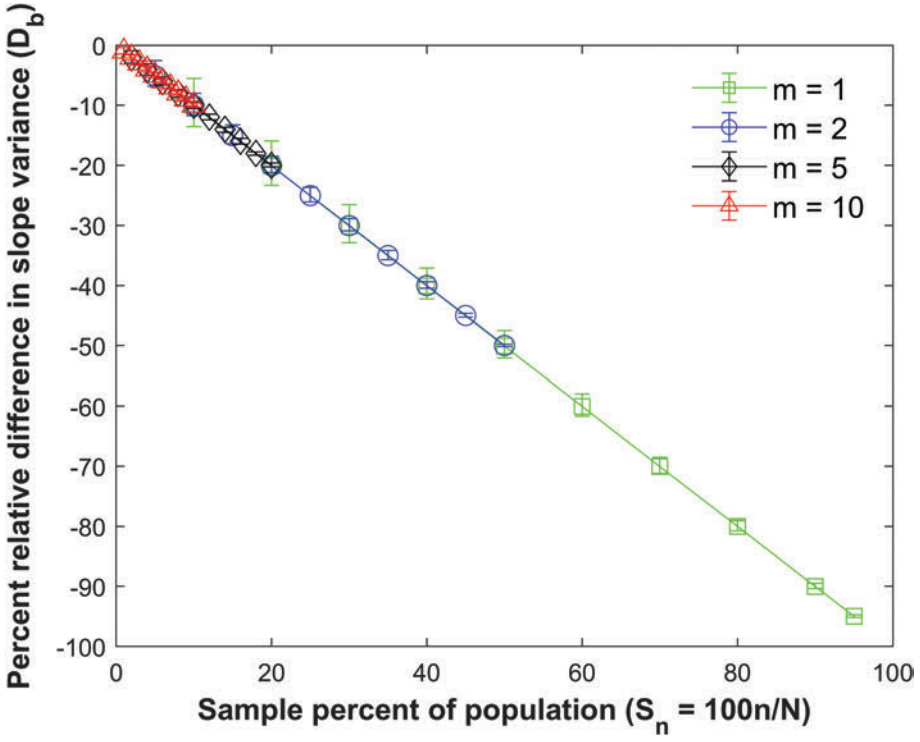

Figure 17: Plot of percent relative difference $D_b = 100(\frac{Var(b_{sim}) - Var(b)}{Var(b)})$ using a population size $N = 10{,}000$ sampled from a bivariate distribution with $\rho = 0.7$. The percent relative difference can be described by the linear plot $D_b = -S_n$. The size of the error bars increases as the sample percent of the population decreases and the group size decreases.

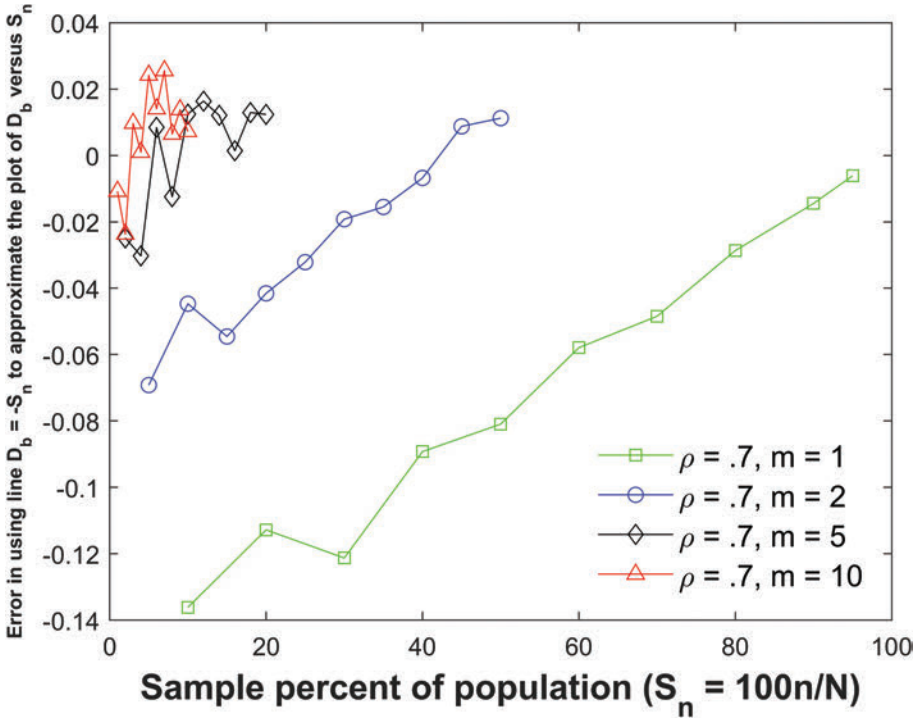

Figure 18: Plot of error $D_b + S_n$ (from Figure 17) when approximating the plot of D_b versus S_n with the line $D_b = -S_n$. The absolute value of the error increases when the group size decreases and the sample percent of the population S_n decreases (for $m = 1$ and $m = 2$).

where $S_n = 100\frac{n}{N}$ which is **not a function of the group size *m***. However, the size of the error bars increases as the sample percent of the population decreases and the group size decreases. Figure 18 plots the error (from Figure 17) when approximating the plot of D_b versus S_n with the line $D_b = -S_n$. The error plotted is $D_b + S_n$. All absolute average errors are less than 0.14 %. We note that the absolute value of the error increases when the group size decreases and the sample percent of the population S_n decreases (for $m = 1$ and $m = 2$).

Simulations (not shown) identical to Figure 18 were conducted except that a small population $N = 400$ was used. In these simulations, all the absolute average errors were small (0.6 %) but larger than the average errors shown in Figure 18.

To test the validity of (4.5), we fit a linear regression line through the plot of D_b versus S_n shown in Figure 17 for different group sizes $m = 1, 2, 5, 10$ and different values of $\rho = -0.9, -0.6, -0.3, 0.0, 0.2, 0.5$, and 0.7. The linear regression line fit through all values of m and all values of ρ has a Pearson R value in the range $-1 \le R < -1 + (1.8 \times 10^{-5})$, a slope b in the range $-1 + (-4 \times 10^{-3}) < b < -1 + (4 \times 10^{-3})$, and a y-intercept y_{int} in the range $-3 \times 10^{-3} < y_{\text{int}} < 3 \times 10^{-3}$ which shows that (4.5) is a good approximation to the plot of D_b vs S_n.

4.4 Groups of mixed size

In Figures 19–22, we study the behavior of groups of mixed size. In all the figures, the population size is $N =$ 10,000 and the samples are generated from a bivariate distribution with $\rho = 0.7$, $(\mu_x, \mu_y) = (0, 0)$, and $(\sigma_x, \sigma_y) = (1, 1)$. Figures 19 and 20 plot the difference $(\text{E}(R_{\text{sim}}) - \text{E}(R))$ and $(\text{E}(b_{\text{sim}}) - \text{E}(b))$ respectively for three types of mixed groups:

(1) $m = 2$ and $m = 8$,
(2) $m = 2$ and $m = 18$,
(3) $m = 3$, $m = 7$, and $m = 10$.

There are equal amounts of each group size in each sample. The plot symbol shows the average difference and the error bars show the maximum and minimum errors using 112 different populations. The differences in the

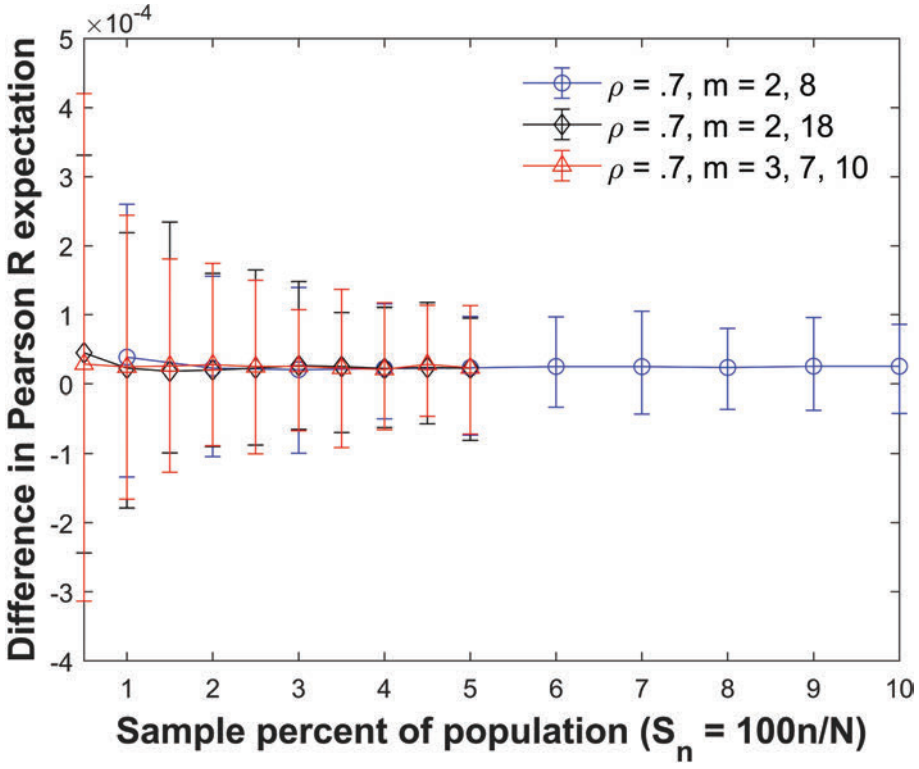

Figure 19: Plot of difference $(\mathrm{E}(R_{\mathrm{sim}}) - \mathrm{E}(R))$ using a population size $N = 10{,}000$ sampled from a bivariate distribution with $\rho = 0.7$ plotted against the sample percent of the population $S_n = 100\frac{n}{N}$ using groups of mixed size.

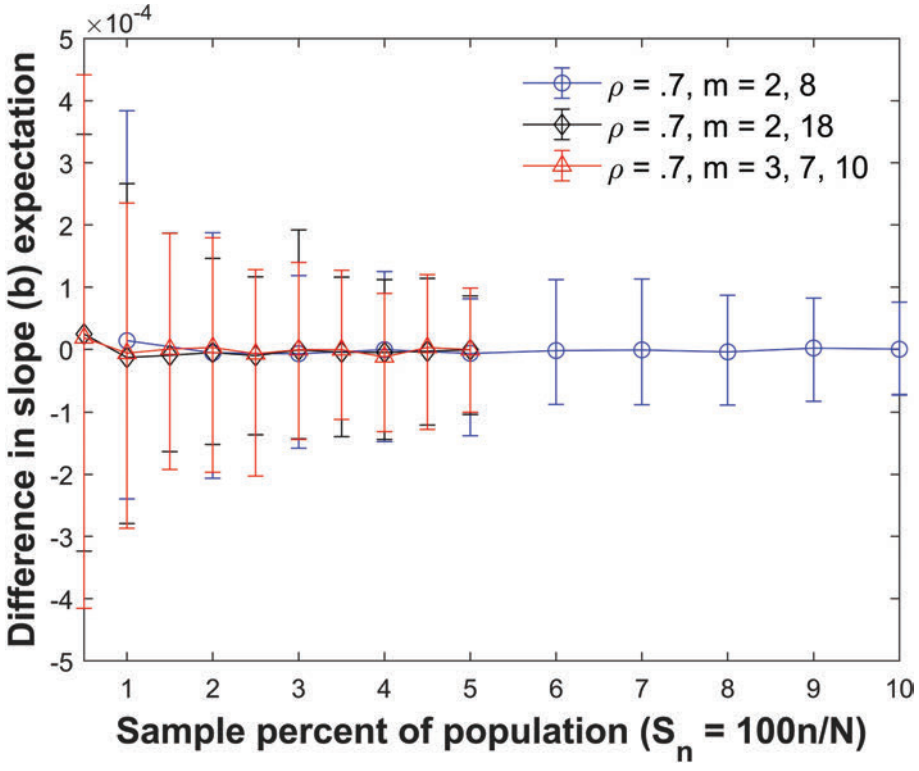

Figure 20: Plot of difference $(\mathrm{E}(b_{\mathrm{sim}}) - \mathrm{E}(b))$ using a population size $N = 10{,}000$ sampled from a bivariate distribution with $\rho = 0.7$ plotted against the sample percent of the population S_n using groups of mixed size.

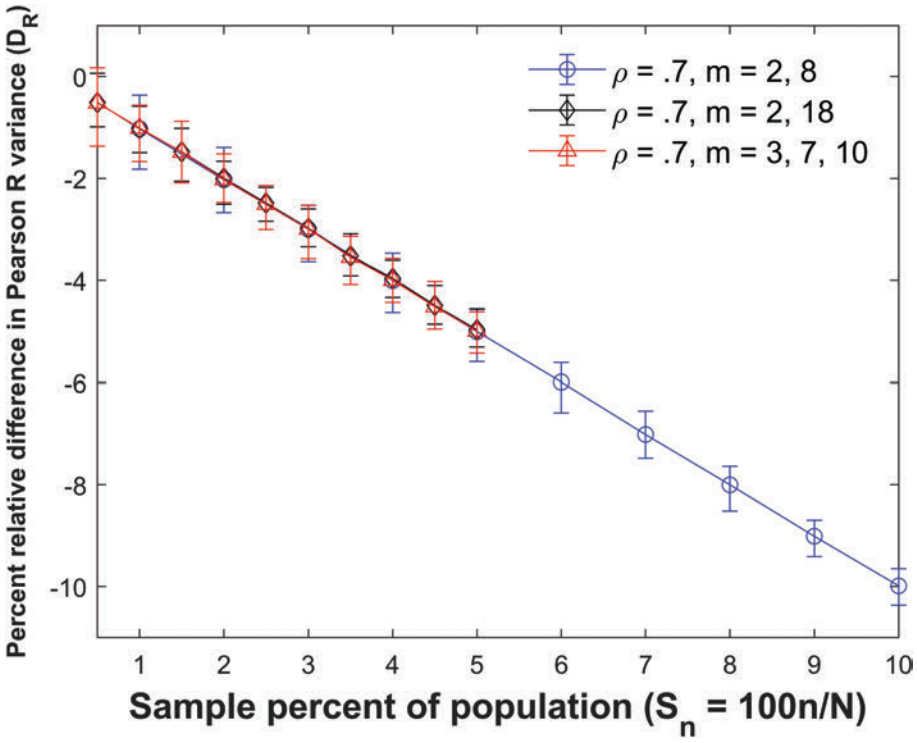

Figure 21: Plot of percent relative difference $D_R = 100(\frac{\text{Var}(R_{\text{sim}}) - \text{Var}(R)}{\text{Var}(R)})$ using a population size $N = 10{,}000$ sampled from a bivariate distribution with $\rho = 0.7$ plotted against the sample percent of the population S_n using groups of mixed size. The percent relative difference is closely approximated with the linear equation $D_R = -S_n$.

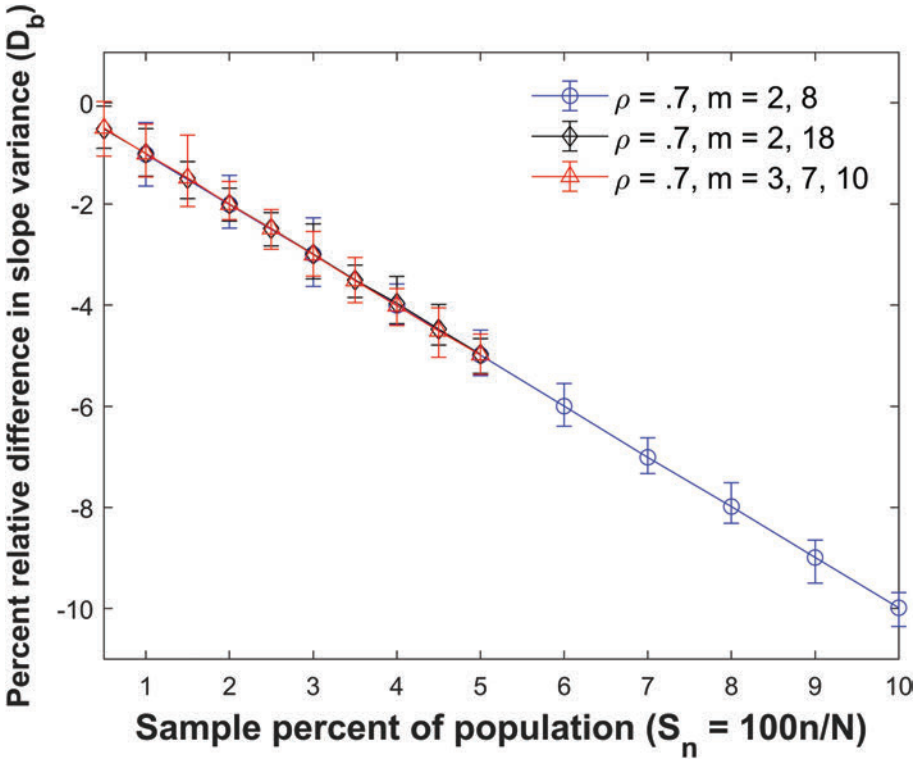

Figure 22: Plot of percent relative difference $D_b = 100(\frac{\text{Var}(b_{\text{sim}}) - \text{Var}(b)}{\text{Var}(b)})$ using a population size $N = 10{,}000$ sampled from a bivariate distribution with $\rho = 0.7$ plotted against the sample percent of the population S_n using groups of mixed size. The percent relative difference can be described by the linear equation $D_b = -S_n$.

expectation values remain small. However, the size of the error bars increases as the sample percent of the population S_n decreases.

Figure 21 and Figure 22 plot $D_R = 100(\frac{\mathrm{Var}(R_{\mathrm{sim}})-\mathrm{Var}(R)}{\mathrm{Var}(R)})$ and $D_b = 100(\frac{\mathrm{Var}(b_{\mathrm{sim}})-\mathrm{Var}(b)}{\mathrm{Var}(b)})$ respectively for the mixed groups. We observe again that $D_R = -S_n$ and $D_b = -S_n$ are good approximations to the plots of D_R versus S_n and D_b versus S_n. The absolute errors in approximating the plot of D_R vs S_n with the line $D_R = -S_n$ and D_b vs S_n with the line $D_b = -S_n$ are all less than 0.05 %.

4.5 Non-normal populations

In this subsection, we consider random samples drawn from non-normal distributions. In all Figures 23–35, the population size is $N = 10{,}000$ and $\rho = 0.7$. Figure 23, Figure 24, and Figure 25 plot the difference $(\mathrm{E}(R_{\mathrm{sim}}) - \mathrm{E}(R))$ sampled from a uniform ($f(x) = f(y) = 1, 0 \le x, y \le 1$), an exponential

$$f(x) = e^{-x}, \quad f(y) = e^{-y}, \quad 0 \le x, y < \infty,$$

and a bimodal distribution

$$f(x) = \frac{e^{-(x+1)^2} + e^{-(x-1)^2}}{2\sqrt{2\pi}}, \quad f(y) = \frac{e^{-\frac{y^2}{2}}}{\sqrt{2\pi}}$$

respectively. The plot symbol shows the average difference and the error bars show the maximum and minimum differences using 112 different populations. The average differences are small but their absolute values increase as the sample percent of the population S_n decreases.

Figure 26, Figure 27, and Figure 28 plot $D_R = 100(\frac{\mathrm{Var}(R_{\mathrm{sim}})-\mathrm{Var}(R)}{\mathrm{Var}(R)})$ sampled from a uniform, exponential, and bimodal distribution respectively. The magenta line $D_R = -S_n$ is included as a reference (as it served as a good approximation when sampling from normal distributions). The plots of D_R vs S_n appear linear. However, unlike

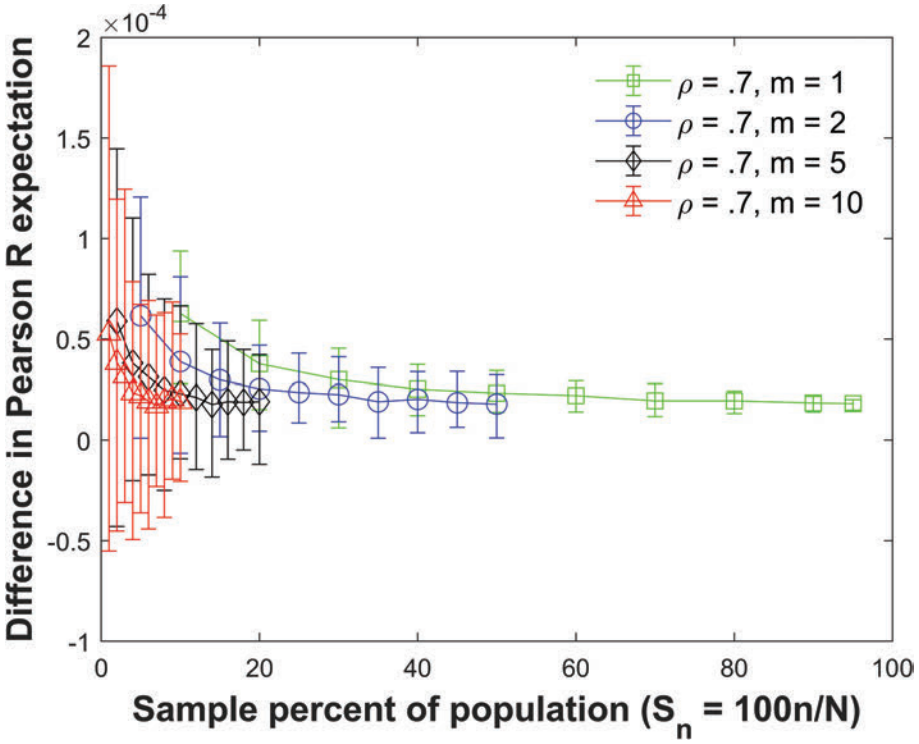

Figure 23: Plot of difference $(\mathrm{E}(R_{\mathrm{sim}}) - \mathrm{E}(R))$ using a population size $N = 10{,}000$ sampled from a **uniform** distribution with $\rho = 0.7$ plotted against the sample percent of the population $S_n = 100\frac{n}{N}$. The plot symbol shows the average difference and the error bars show the maximum and minimum differences using 112 different populations of size $N = 10{,}000$.

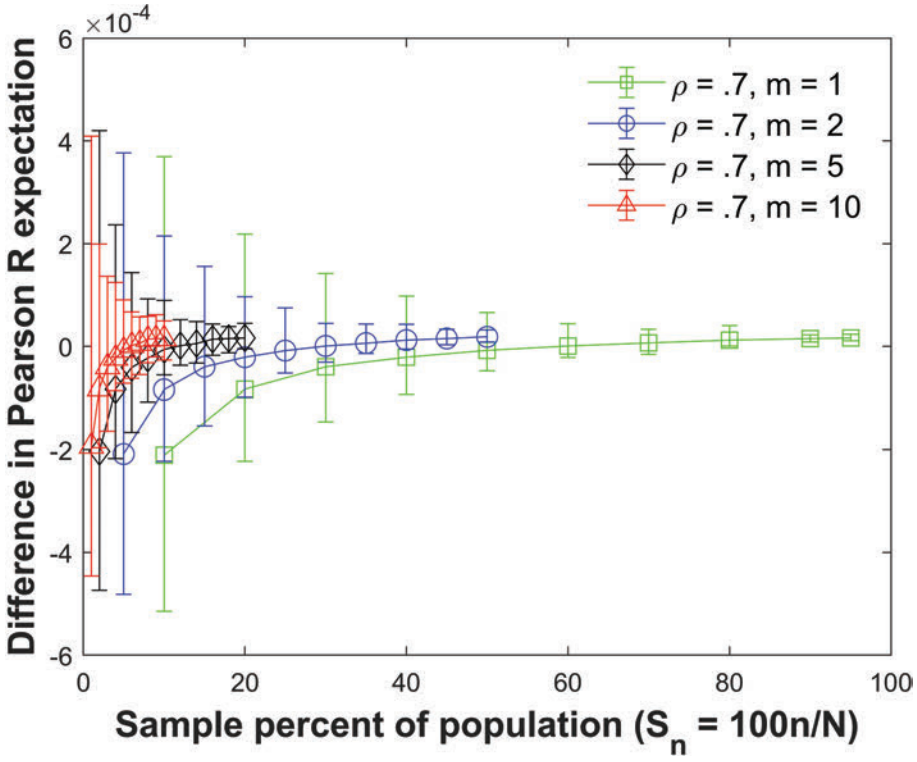

Figure 24: Plot of difference $(\mathrm{E}(R_{\mathrm{sim}}) - \mathrm{E}(R))$ using a population size $N = 10{,}000$ sampled from an **exponential** distribution with $\rho = 0.7$ plotted against the sample percent of the population S_n.

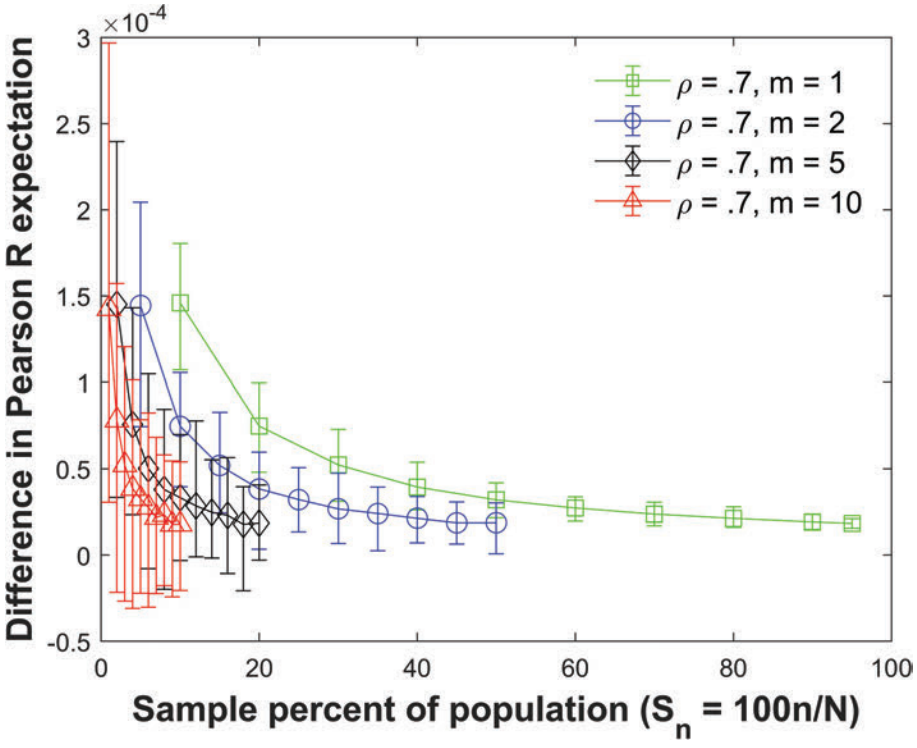

Figure 25: Plot of difference $(\mathrm{E}(R_{\mathrm{sim}}) - \mathrm{E}(R))$ using a population size $N = 10{,}000$ sampled from a **bimodal** distribution with $\rho = 0.7$ plotted against the sample percent of the population S_n.

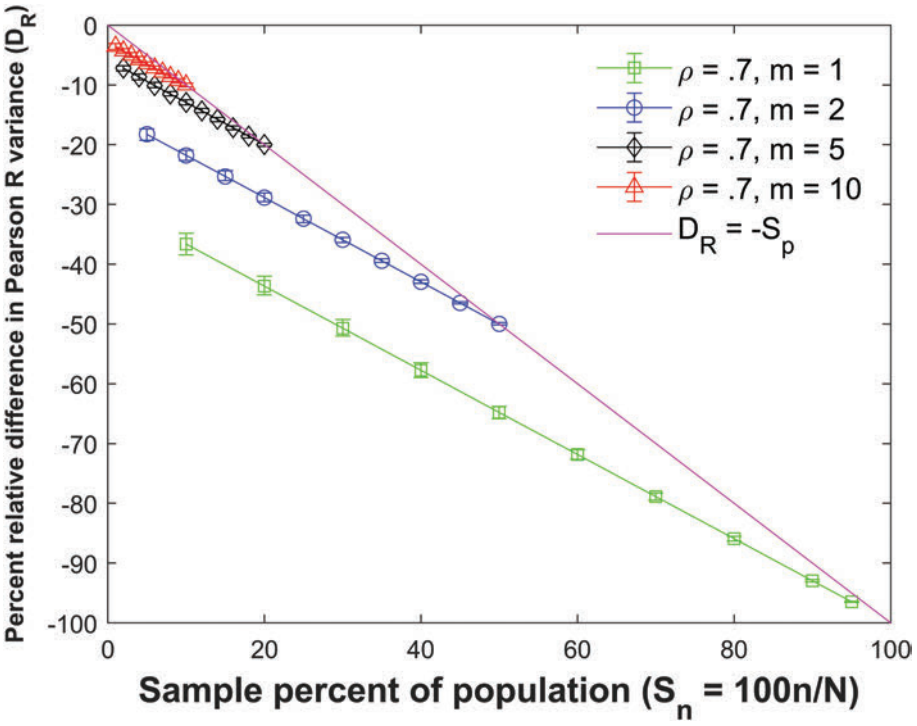

Figure 26: Plot of percent relative difference $D_R = 100(\frac{\text{Var}(R_{\text{sim}})-\text{Var}(R)}{\text{Var}(R)})$ using a population size $N = 10{,}000$ sampled from a **uniform** distribution with $\rho = 0.7$ plotted against the sample percent of the population S_n.

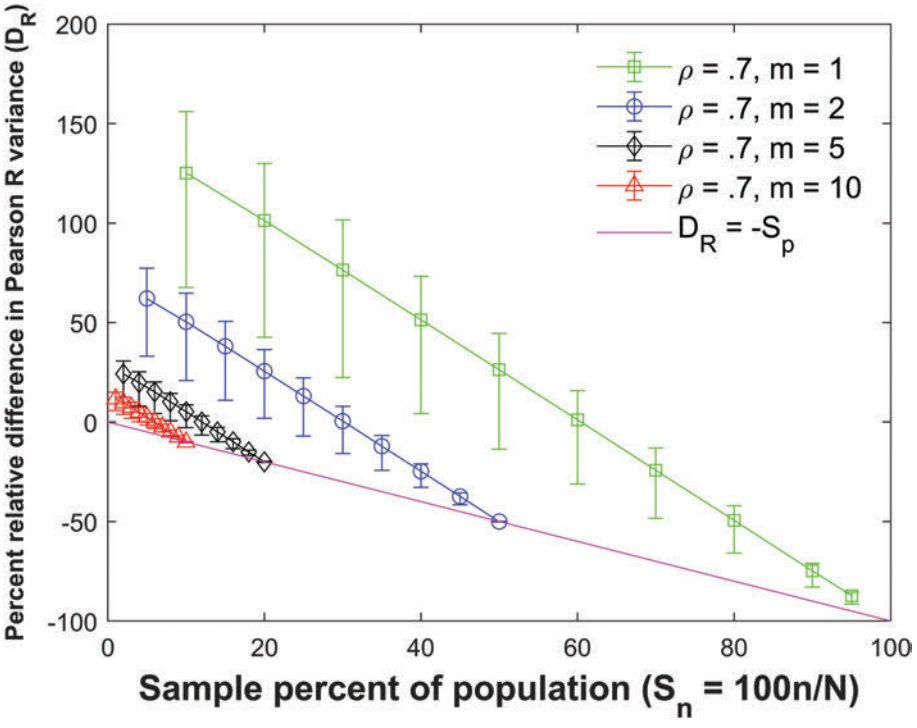

Figure 27: Plot of percent relative difference $D_R = 100(\frac{\text{Var}(R_{\text{sim}})-\text{Var}(R)}{\text{Var}(R)})$ using a population size $N = 10{,}000$ sampled from an **exponential** distribution with $\rho = 0.7$ plotted against the sample percent of the population S_n.

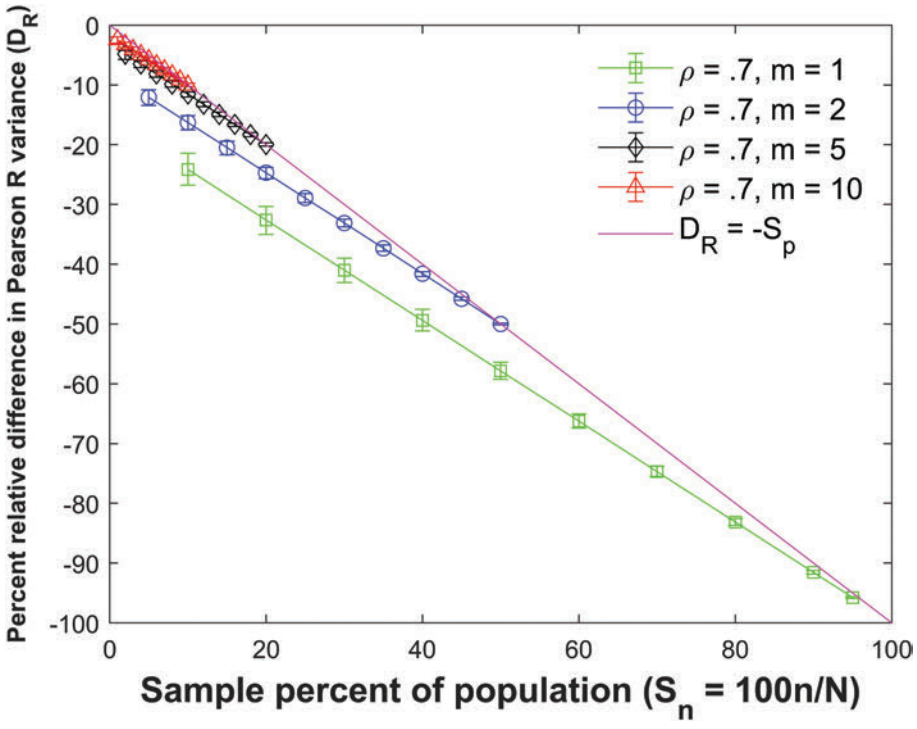

Figure 28: Plot of percent relative difference $D_R = 100(\frac{\text{Var}(R_{\text{sim}}) - \text{Var}(R)}{\text{Var}(R)})$ using a population size $N = 10{,}000$ sampled from a **bimodal** distribution with $\rho = 0.7$ plotted against the sample percent of the population S_n.

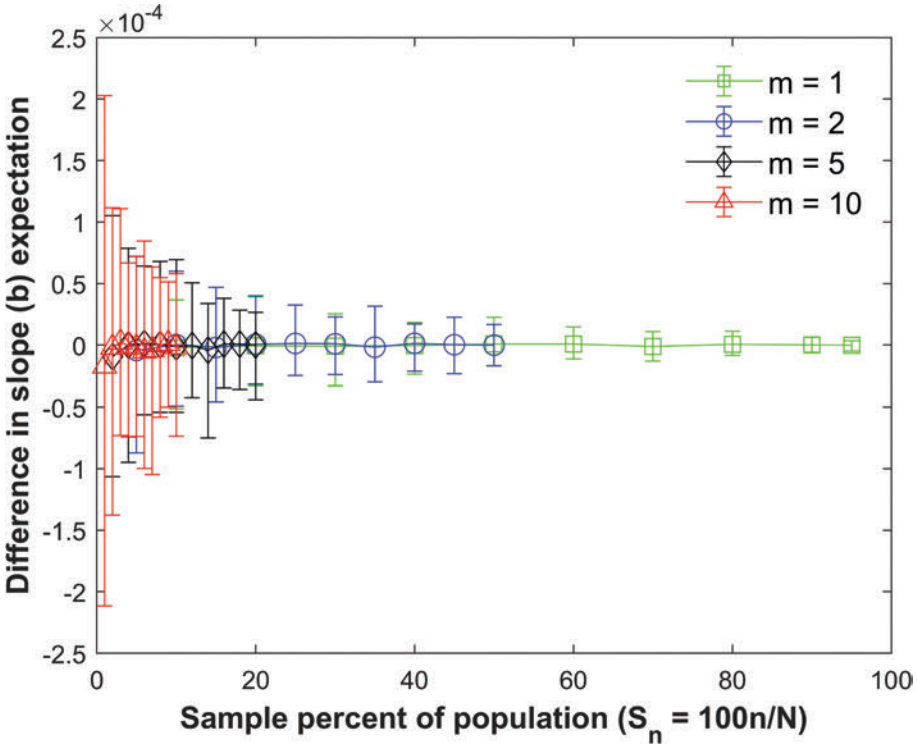

Figure 29: Plot of difference $(\text{E}(b_{\text{sim}}) - \text{E}(b))$ using a population size $N = 10{,}000$ sampled from a **uniform** distribution with $\rho = 0.7$ plotted against the sample percent of the population S_n.

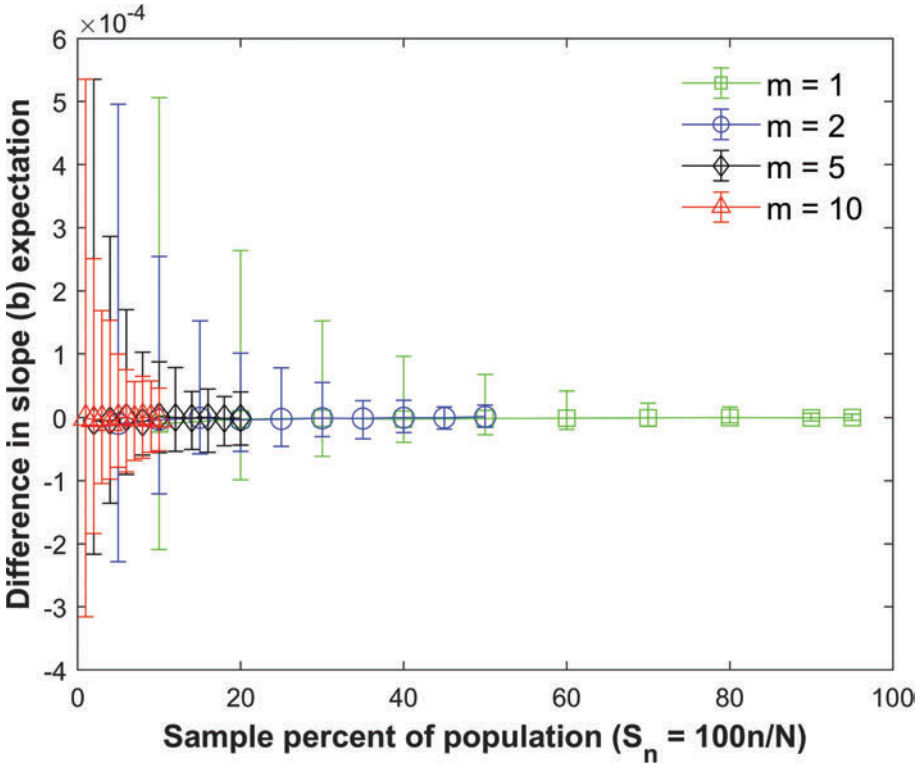

Figure 30: Plot of difference ($E(b_{sim}) - E(b)$) using a population size $N = 10{,}000$ sampled from an **exponential** distribution with $\rho = 0.7$ plotted against the sample percent of the population S_n.

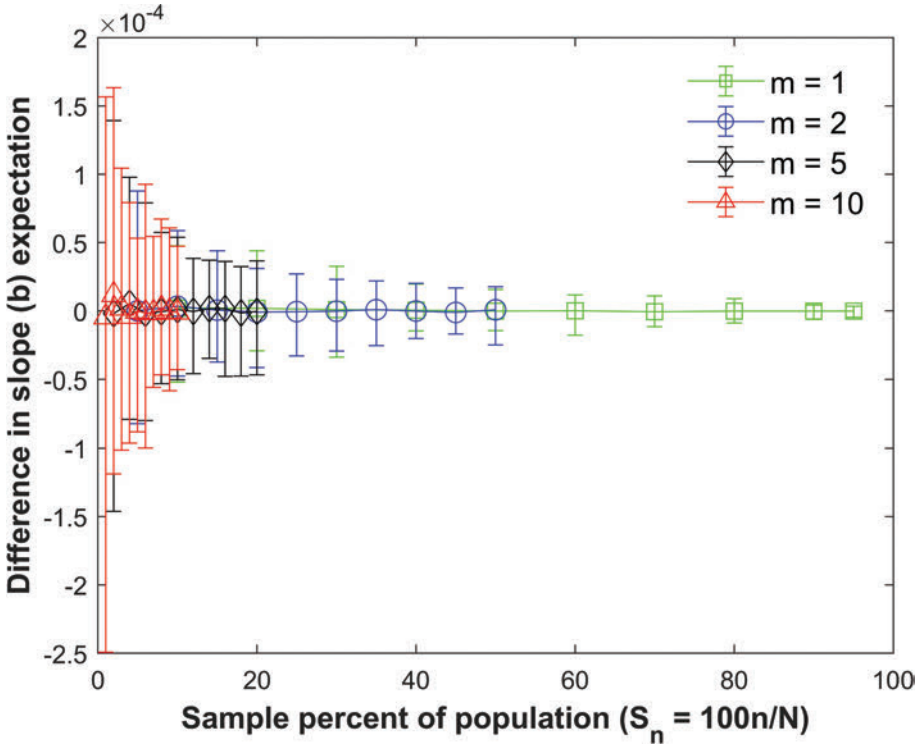

Figure 31: Plot of difference ($E(b_{sim}) - E(b)$) using a population size $N = 10{,}000$ sampled from a **bimodal** distribution with $\rho = 0.7$ plotted against the sample percent of the population S_n.

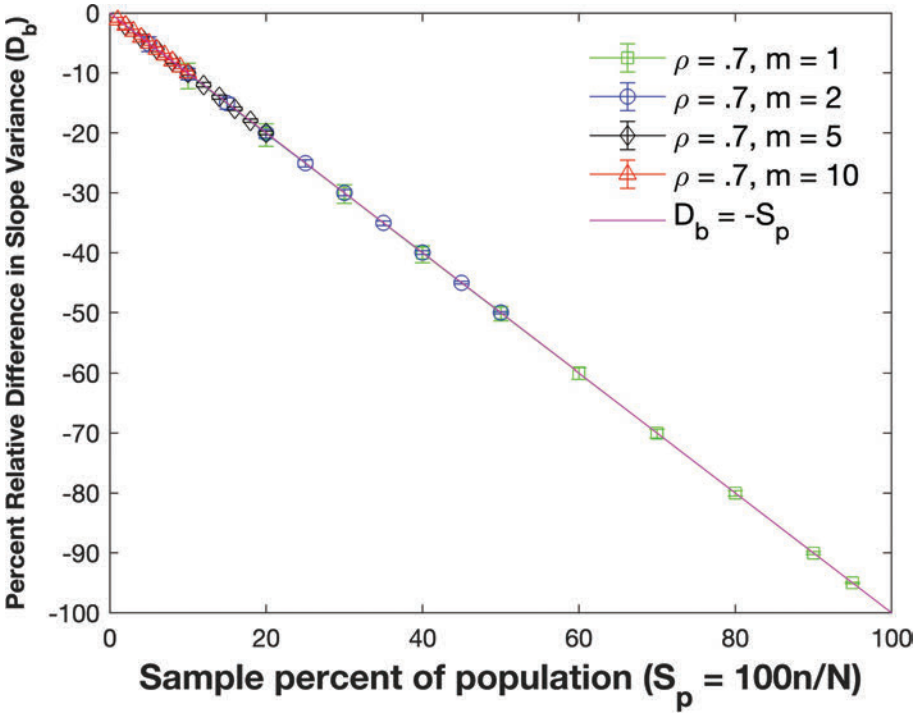

Figure 32: Plot of percent relative difference $D_b = 100(\frac{\text{Var}(b_{\text{sim}})-\text{Var}(b)}{\text{Var}(b)})$ using a population size $N = 10{,}000$ sampled from a **uniform** distribution with $\rho = 0.7$ plotted against the sample percent of the population S_n. The absolute errors in approximating the plot of D_b vs S_n with the line $D_b = -S_n$ are all less than 0.18 %.

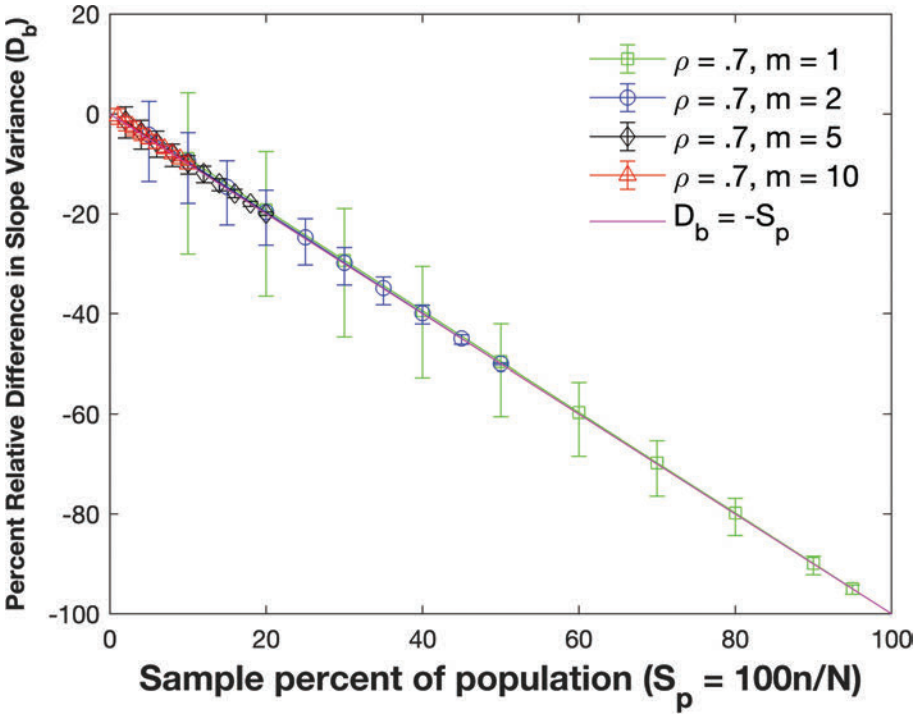

Figure 33: Plot of percent relative difference $D_b = 100(\frac{\text{Var}(b_{\text{sim}})-\text{Var}(b)}{\text{Var}(b)})$ using a population size $N = 10{,}000$ sampled from an **exponential** distribution with $\rho = 0.7$ plotted against the sample percent of the population S_n. The absolute errors in approximating the plot of D_b vs S_n with the line $D_b = -S_n$ are all less than 1.2 %.

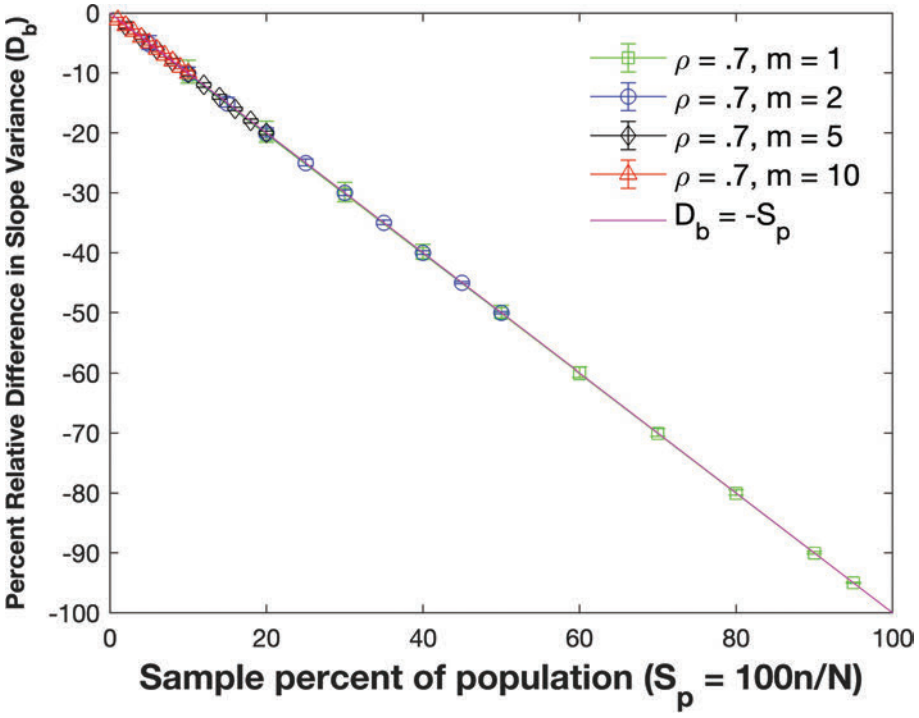

Figure 34: Plot of percent relative difference $D_b = 100(\frac{\text{Var}(b_{\text{sim}})-\text{Var}(b)}{\text{Var}(b)})$ using a population size $N = 10{,}000$ sampled from a **bimodal** distribution with $\rho = 0.7$ plotted against the sample percent of the population S_n. The absolute errors in approximating the plot of D_b vs S_n with the line $D_b = -S_n$ are all less than 0.35 %.

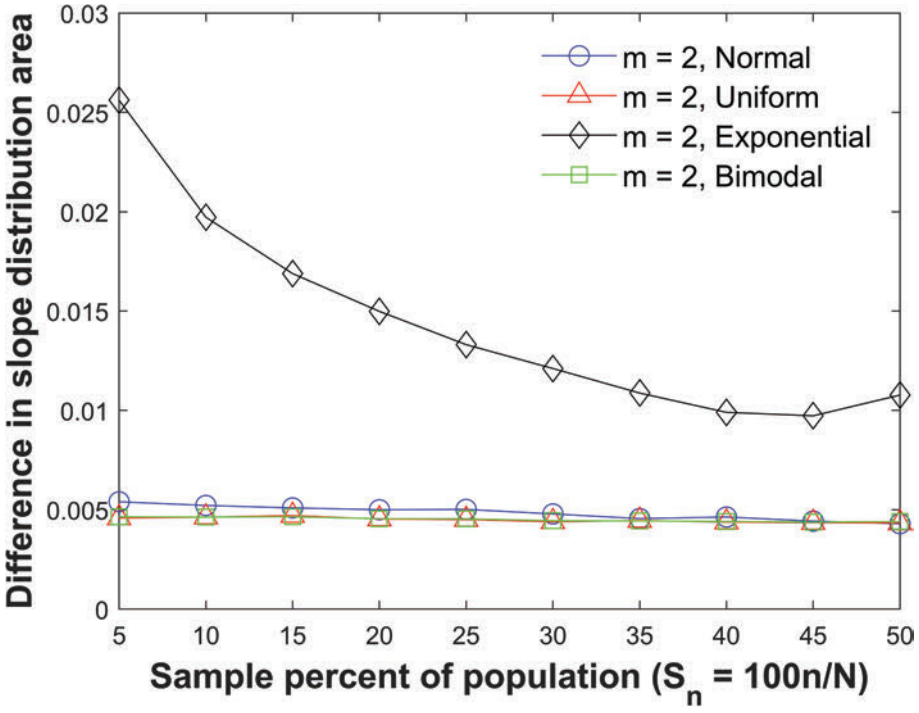

Figure 35: Plot of difference $\int |h_{\text{sim}}(b) - h(b)|\, db$, where $h_{\text{sim}}(b)$ is the distribution formed by the Monte Carlo simulation, $h(b)$ is the slope distribution (2.9), and $\{(x_i, y_i)\}$ are sampled from normal, uniform, exponential, and bimodal distributions.

samples from normal distributions, D_R is a function of **both** the sample percent of the population S_n and the group size m. The line for $m = 10$ does seem to match the $D_R = -S_n$ line better compared to smaller values of m due to the central limit theorem. Interestingly, we note that for $m > 1$ the plots intersect the line $D_R = -S_n$ when $mS_n = 100$. We also tried generating $\mathrm{Var}(R)$ by sampling **without** replacement for each non-normal population and found that D_R still could not be approximated with $D_R = -S_n$.

Figure 29, Figure 30, and Figure 31 plot the difference $(\mathrm{E}(b_{\mathrm{sim}}) - \mathrm{E}(b))$ sampled from a uniform, exponential, and bimodal distribution respectively. Since the difference is small, $\mathrm{E}(b)$ computed from (2.10) is a good approximation for the expectation of the sampling distribution.

Figure 32, Figure 33, and Figure 34 plot $D_b = 100(\frac{\mathrm{Var}(b_{\mathrm{sim}}) - \mathrm{Var}(b)}{\mathrm{Var}(b)})$ sampled from a uniform, exponential, and bimodal distribution respectively; $\mathrm{Var}(b)$ is computed using (2.11). The magenta line $D_b = -S_n$ is included as a reference. Remarkably, D_b is only a function of S_n and does not appear to depend on the group size m (unlike D_R). We note that Goodman's [7] method relies on the premise that regression coefficients are less easily affected by aggregation than correlation [10].

Figure 35 plots the area difference $\int |h_{\mathrm{sim}}(b) - h(b)|\, db$, where $h_{\mathrm{sim}}(b)$ is the distribution formed by the Monte Carlo simulation, $h(b)$ is the slope distribution (2.9), and $\{(x_i, y_i)\}$ are sampled without replacement from normal, uniform, exponential, and bimodal distributions. The variance of $h(b)$ is modified by adjusting the value of n in order to match the variance of the line $D_b = -S_n$. The largest area difference is generated with the exponential distribution, but all differences are less than 0.026. Figure 35 shows that the simulated distributions can be approximated with $h(b)$ for these populations types.

4.6 Multiple regression results

We move now to multiple regression simulations. Figure 36 plots the difference $(\mathrm{E}(R^2_{\mathrm{sim}}) - \mathrm{E}(R^2))$ using a population size $N = 10{,}000$ sampled from a normal distribution and correlated using Cholesky decomposition with $\rho^2 = 0.25$ $\rho^2 = 0.73$, and $\rho^2 = 0.72$. For $\rho^2 = 0.25$ and $m = 5$, the correlation matrix $\mathbf{C}$ has the following off-diagonal elements: $R_{zx} = C_{1,2} = C_{2,1} = -0.5$, $R_{zy} = C_{1,3} = C_{3,1} = 0.1$, $R_{xy} = C_{2,3} = C_{3,2} = -0.055$. We sample from both a normal and uniform distribution when $\rho^2 = 0.25$. For $\rho = 0.73$ and $m = 5$, the correlation matrix has the following off-diagonal elements: $R_{zx} = C_{1,2} = C_{2,1} = 0.7$, $R_{zy} = C_{1,3} = C_{3,1} = 0.8$, $R_{xy} = C_{2,3} = C_{3,2} = 0.56$. For $\rho = 0.72$ and $m = 3$, the correlation matrix has the following off-diagonal elements: $R_{zx} = C_{1,2} = C_{2,1} = -0.3$, $R_{zy} = C_{1,3} = C_{3,1} = 0.7$, $R_{xy} = C_{2,3} = C_{3,2} = -0.2$. The symbol shows the average difference and the error bars show the maximum and minimum differences using 112 different populations. The expectation $\mathrm{E}(R^2)$ is computed using (2.13) and $\mathrm{E}(R^2_{\mathrm{sim}})$ is generated using a Monte Carlo simulation in which a million samples are randomly selected without replacement. The differences $(\mathrm{E}(R^2_{\mathrm{sim}}) - \mathrm{E}(R^2))$ are small but the size of the error bars increase as S_n decreases.

Figure 37 plots the percent relative difference $D^2_R = 100(\frac{\mathrm{Var}(R^2_{\mathrm{sim}}) - \mathrm{Var}(R^2)}{\mathrm{Var}(R^2)})$ using a population size $N = 10{,}000$ sampled from a bivariate distribution under the same conditions as Figure 36. The symbol shows the average difference and the error bars show the maximum and minimum differences using 112 different populations. The variance $\mathrm{Var}(R^2)$ is generated using (2.14) and $\mathrm{Var}(R^2_{\mathrm{sim}})$ is generated using a Monte Carlo simulation in which a million samples are randomly selected without replacement. The percent relative difference can be approximated with the line $D^2_R = -S_n$.

Figure 38 plots the difference $(\mathrm{E}(b_{\mathrm{sim}}) - \mathrm{E}(b))$ using a population size $N = 10{,}000$ sampled from a bivariate distribution with $\rho^2 = 0.25$, $\rho^2 = 0.73$, and $\rho = 0.72$ using the slope b between z and x. We also include a plot of $\rho^2 = 0.25$ which was created by sampling from a uniform distribution. The expectation $\mathrm{E}(b)$ is generated using a Monte Carlo simulation in which a million samples are selected **with** replacement and $\mathrm{E}(b_{\mathrm{sim}})$ is generated using a Monte Carlo simulation in which a million samples are selected without replacement. The differences $(\mathrm{E}(b_{\mathrm{sim}}) - \mathrm{E}(b))$ are small.

Figure 39 plots the percent relative difference $D_b = 100(\frac{\mathrm{Var}(b_{\mathrm{sim}}) - \mathrm{Var}(b)}{\mathrm{Var}(b)})$ under the same conditions as Figure 38 using the slope b between z and x. We also include a plot of $\rho^2 = 0.25$ which was created by sampling from a uniform distribution. The variance $\mathrm{Var}(b)$ is generated using a Monte Carlo simulation in which a million samples are selected **with** replacement and $\mathrm{Var}(b_{\mathrm{sim}})$ is generated using a Monte Carlo simulation in which

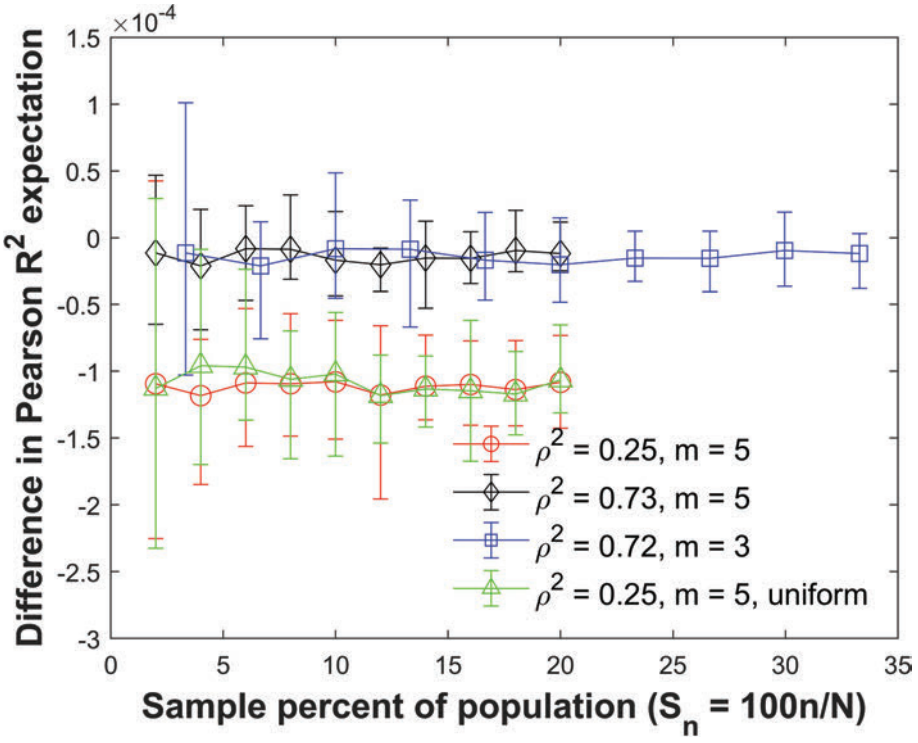

Figure 36: Plot of difference ($\mathrm{E}(R^2_{\mathrm{sim}}) - \mathrm{E}(R^2)$) using a population size $N = 10{,}000$ sampled from a bivariate distribution with $\rho^2 = .25$ and $m = 5$, $\rho^2 = 0.73$ and $m = 5$, and $\rho^2 = 0.72$ and $m = 3$. A plot of $\rho^2 = 0.25$ and $m = 5$ is also included which was sampled from a uniform distribution. The symbol shows the average difference and the error bars show the maximum and minimum differences using 112 different populations. The expectation $\mathrm{E}(R^2)$ is computed using (2.13) and $\mathrm{E}(R^2_{\mathrm{sim}})$ is generated using a Monte Carlo simulation in which a million samples are selected without replacement using different sample sizes n (displayed along the horizontal axis using the sample percent of the population $S_n = 100\frac{n}{N}$).

a million samples are selected without replacement. The percent relative difference can be approximated with the line $D_b = -S_n$ even for the case where samples were drawn from a uniform distribution.

5 Discussion and conclusion

In this article, we perform Monte Carlo simulations to select samples without replacement from finite populations to generate distributions of Pearson R, slope b, and the coefficient of determination R^2 in simple and multiple regression contexts. Our Monte Carlo simulations suggest that the expectations of the R, b, and R^2 distributions are similar to the expectations of the analytical sampling distributions for normally distributed data for both individual and group averaged data as long as the sample sizes n are the same. This observation is also true for groups of mixed sizes and for the three non-normal distributions we tested using simple regression simulations.

However the variances of the R, b, and R^2 distributions created without replacement are reduced compared to the variances of the analytical distributions. Our simulations show that the percent relative differences D_R, D_b, and D^2_R in the variances can be approximated with the linear equations $D_R = -S_n$, $D_b = -S_n$, and $D^2_R = -S_n$, where S_n is the sample percent of the population $S_n = 100\frac{n}{N}$. The group size that is used when selecting the samples does not affect the variances for normally distributed variables. We observe the same results when considering groups of mixed size.

The distribution of the Fisher transformed value of R denoted by R_z can also be approximated by a normal distribution as the sample size n increases when sampling without replacement for both individual and group averaged data. Similar to the other percent relative differences, D_z can be approximated with the line $D_z = -S_n$.

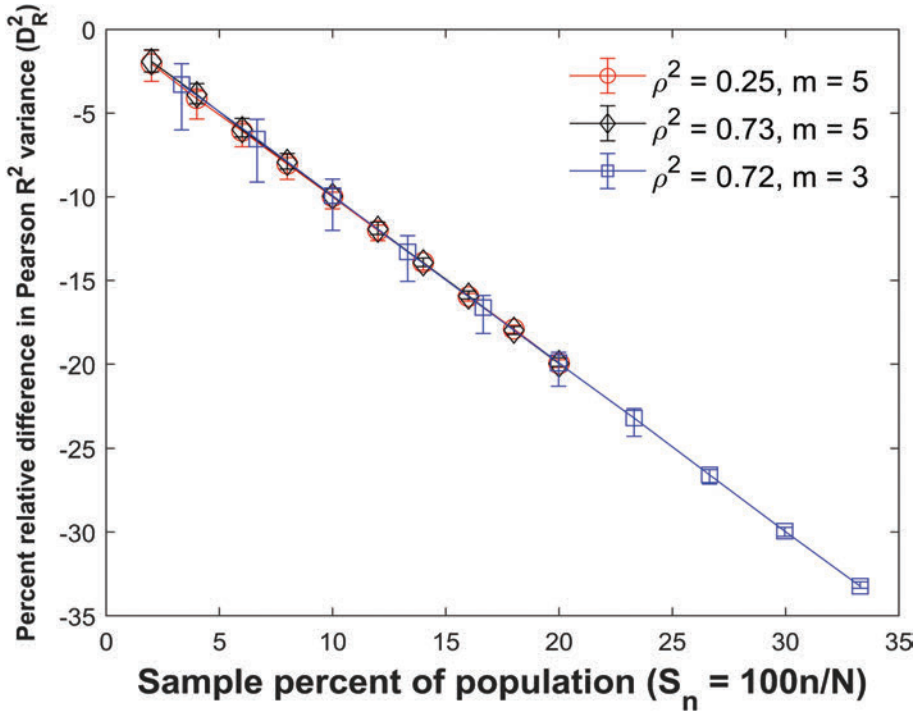

Figure 37: Plot of percent relative difference $D_R^2 = 100(\frac{\text{Var}(R^2_{\text{sim}})-\text{Var}(R^2)}{\text{Var}(R^2)})$ using a population size $N = 10{,}000$ sampled from a bivariate distribution with $\rho^2 = 0.25$ and $m = 5$, $\rho^2 = 0.73$ and $m = 5$, and $\rho^2 = 0.72$ and $m = 3$. The symbol shows the average difference and the error bars show the maximum and minimum differences using 112 different populations. The variance $\text{Var}(R^2)$ is computed using (2.14) and $\text{Var}(R^2_{\text{sim}})$ is generated using a Monte Carlo simulation in which a million samples are selected without replacement.

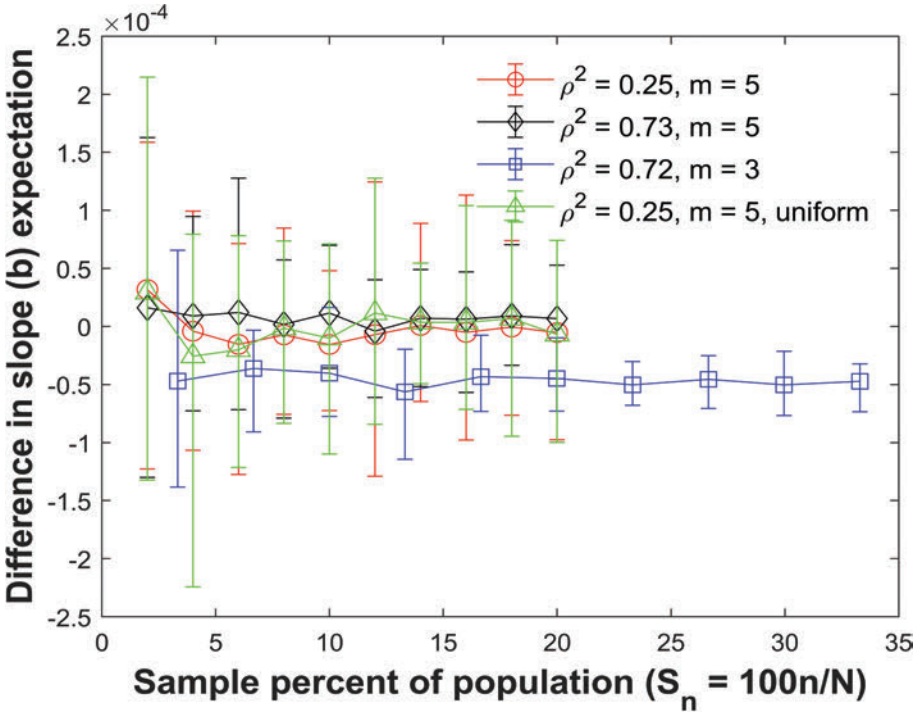

Figure 38: Plot of difference $(\text{E}(b_{\text{sim}}) - \text{E}(b))$ using a population size $N = 10{,}000$ sampled from a bivariate distribution with $\rho^2 = 0.25$ and $m = 5$, $\rho^2 = 0.73$ and $m = 5$, and $\rho^2 = 0.72$ and $m = 3$. A plot of $\rho^2 = 0.25$ and $m = 5$ is also included which was sampled from a uniform distribution. The symbol shows the average difference and the error bars show the maximum and minimum differences using 112 different populations. The expectation $\text{E}(b)$ is generated using a Monte Carlo simulation in which a million samples are selected **with** replacement and $\text{E}(b_{\text{sim}})$ is generated using a Monte Carlo simulation in which a million samples are selected without replacement.

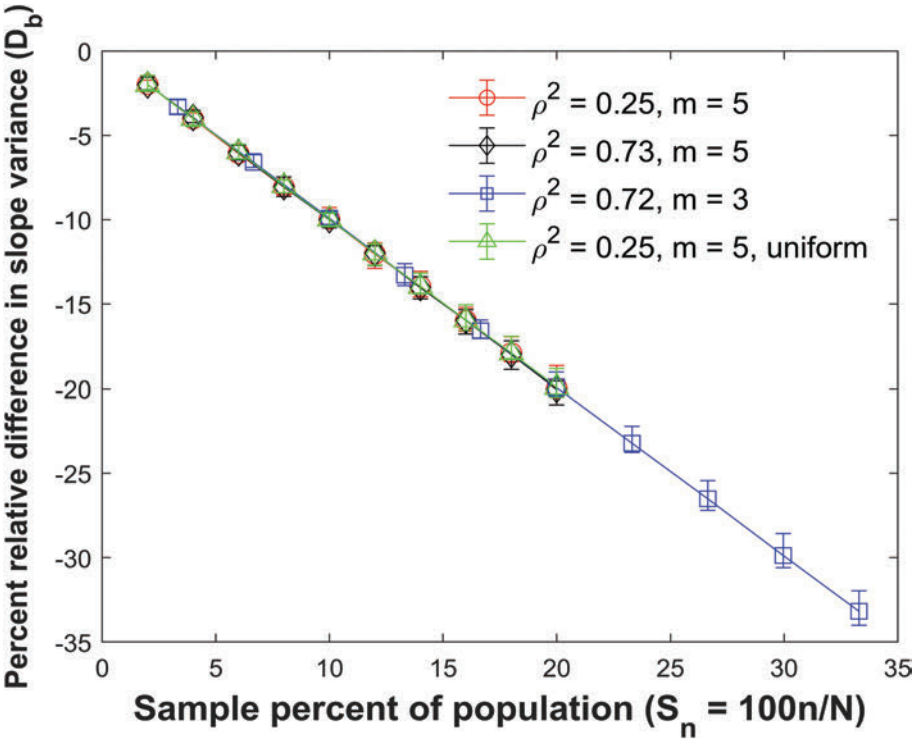

Figure 39: Plot of percent relative difference $D_b = 100(\frac{\text{Var}(b_{\text{sim}})-\text{Var}(b)}{\text{Var}(b)})$ using a population size $N = 10{,}000$ sampled from a bivariate distribution with $\rho^2 = 0.25$ and $m = 5$, $\rho^2 = 0.73$ and $m = 5$, and $\rho^2 = 0.72$ and $m = 3$. A plot of $\rho^2 = 0.25$ and $m = 5$ is also included which was sampled from a uniform distribution. The symbol shows the average difference and the error bars show the maximum and minimum differences using 112 different populations. The variance $\text{Var}(b)$ is generated using a Monte Carlo simulation in which a million samples were selected **with** replacement and $\text{Var}(b_{\text{sim}})$ is generated using a Monte Carlo simulation in which million samples are selected without replacement.

We also observed that for non-normal distributions, the percent relative differences in the correlation variances D_R depend on the sample size and the group size and cannot be approximated with a simple line. Interestingly, in contrast, the percent relative differences in **slope** variances D_b **can** be approximated by the line $D_b = -S_n$ for the non-normal distributions we considered.

Our observations afford another interpretation of the ecological fallacy and suggest that for random samples drawn without replacement from normally distributed finite populations, the correlation coefficients and linear regression slopes will be selected from approximately the same sampling distribution regardless of the group size m as long as the sample size n is the same.

A Appendix: Fortran code

The parallel Fortran code we use to run the Monte Carlo simulations evenly divides the 112 populations among the processors. Each processor then develops the Pearson R and slope distributions using its share of the populations for different values of group size m and different sample percents of the population S_n.

Each processor finds the maximum, minimum, and average difference of the expectation and variance of each distribution from the analytical value for its share of the populations. Then the maximum, minimum, and average difference of the expectation and variance are shared among all the processors, to find the maximum, minimum, and average difference for all populations which is stored on the first processor.

We have developed two versions of the code: an MPI version and a Coarray Fortran version. MPI uses the MPI_REDUCE calls for communication and the Coarray Fortran uses the co_sum, co_max, and co_min calls for communication.

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. We also received support from the Sustainable Research Pathways program and the Sustainable Horizons Institute.

References

[1] N. Cleave, P. J. Brown and C. D. Payne, Evaluation of methods for ecological inference, *J. Roy. Statist. Soc. Ser. A* **158** (1995), 55–72.
[2] R. A. Fisher, Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population, *Biometrika* **10** (1915), 507–521.
[3] R. A. Fisher, On the probable error of a coefficient of correlation deduced from a small sample, *Metron* **1** (1921), 3–32.
[4] R. A. Fisher, The general sampling distribution of the multiple correlation coefficient, *Proc. Roy. Soc. Lond. Ser. A* **121** (1928), 654–673.
[5] H. Gatignon, *Statistical Analysis of Management Data*, 2nd ed., Springer, New York, 2010.
[6] A. T. Geronimus and J. Bound, Use of census-based aggregate variables to proxy for socioeconomic group: Evidence from national samples, *Am. J. Epidemiol.* **148** (1988), 475–486.
[7] L. Goodman, Ecological regressions and behavior of individuals, *Amer. Sociological Rev.* **18** (1953), 663–664.
[8] L. Irwin and A. J. Lichtman, Across the great divide: Inferring individual level behavior from aggregate data, *Political Methodology* **3** (1976), 411–439.
[9] G. King, *A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data*, Princeton University, New Jersey, 1997.
[10] A. J. Lichtman, Correlation, regression, and the ecological fallacy: A critique, *J. Interdiscip. Hist.* **4** (1974), 417–433.
[11] S. Mahadevan, Monte Carlo simulation, in: *Reliability-Based Mechanical Design*, Marcel Dekker, New York (1997), 123–146.
[12] R. J. Muirhead, *Aspects of Multivariate Statistical Theory*, John Wiley & Sons, New Jersey, 2005.
[13] S. Piantadosi, D. P. Byar and S. B. Green, The ecological fallacy, *Am. J. Epidemiol.* **127** (1988), 893–904.
[14] W. S. Robinson, Ecological correlations and the behavior of individuals, *Amer. Sociological Rev.* **15** (1950), 351–357.
[15] V. Romanovskij, On the distribution of the regression coefficient in samples from normal population, *Bull. Acad. Sci. URSS* **20** (1926), no. 6, 643–648.
[16] Y. T. Shih, C. Bradley and K. R. Yabroff, Ecological and individualistic fallacies in health disparities research, *J. National Cancer Inst.* **115** (2023), 488–491.
[17] D. J. Torres, Describing the Pearson R distribution of aggregate data, *Monte Carlo Methods Appl.* **1** (2020), 17–32.
[18] S. M. Woodward, D. Mork, X. Wu, Z. Hou, D. Braun and F. Dominici, Combining aggregate and individual-level data to estimate individual-level associations between air pollution and COVID-19 mortality in the United States, *PLOS Global Public Health* **3** (2023), Article ID e0002178.