2-25-2020

The Importance of Type I Error Rates When Studying Bias in Monte Carlo Studies in Statistics

Michael Harwell
University of Minnesota - Twin Cities, harwe001@umn.edu

Recommended Citation

INVITED ARTICLE

The Importance of Type I Error Rates When Studying Bias in Monte Carlo Studies in Statistics

Michael Harwell
University of Minnesota - Twin Cities
Minneapolis, MN

Two common outcomes of Monte Carlo studies in statistics are bias and Type I error rate. Several versions of bias statistics exist but all employ arbitrary cutoffs for deciding when bias is ignorable or non-ignorable. This article argues Type I error rates should be used when assessing bias.

Keywords: Monte Carlo studies, statistics, bias, type I errors

Introduction

Various outcomes are used to capture the behavior of estimators and statistical tests in Monte Carlo studies. Harwell et al. (2018) reviewed 677 articles in six journals appearing between 1985-2012 that reported Monte Carlo results and found 33.1%, 16%, and 44.1% of these studies presented results for bias, Type I error rate, and root mean square error (RMSE), respectively.

Bias

Bias of an estimator is defined as $\mathrm{E}(\hat{\theta} - \theta)$ which is the difference between an estimator's expected value and the true value of the parameter being estimated (Neter et al., 1996), where $\hat{\theta}$ is an estimate of θ. The central feature of bias is that it is non-random. The $(\hat{\theta}_i - \theta)$ generated across $i = 1, 2,…, J$ replications in a Monte Carlo study lead to several common measures of bias:

$$\text{Average bias} = \sum_{i=1}^{J} \frac{\left(\hat{\theta}_i - \theta\right)}{J} \tag{1}$$

$$\text{Absolute bias} = \sum_{i=1}^{J} \frac{\left|\hat{\theta}_i - \theta\right|}{J} \tag{2}$$

$$\text{Relative bias} = \frac{1}{J}\left[\sum_{i=1}^{J}\left(\frac{\hat{\theta}_i - \theta}{\theta}\right) \times 100\right] \left(\text{expressed as a percentage}\right) \tag{3}$$

$$\text{Absolute relative bias} = \frac{1}{J}\left[\sum_{i=1}^{J}\left|\frac{\hat{\theta}_i - \theta}{\theta}\right| \times 100\right] \left(\text{expressed as a percentage}\right) \tag{4}$$

A critical feature of the above bias measures in Monte Carlo studies are user-specified (arbitrary) cutoffs to distinguish important (non-ignorable) bias from less important (ignorable) bias: Cutoffs for average bias and absolute bias, equations (1) and (2), are unique to individual Monte Carlo studies (e.g., Brannick et al., 2019; Yuan et al., 2015), whereas 5% and 10% are typical cutoffs for relative bias and absolute relative bias, equations (3) and (4), although other values are sometimes used.

For example, Harring et al. (2012) used an absolute bias cutoff of .05 for structural equation modeling estimates; Jin et al. (2016) and Kim et al. (2016) used a relative bias cutoff of 5% for estimated factor loadings; Leite and Beretvas (2010) used 5% when examining bias after imputing missing Likert-type data; Li et al. (2011) used 5% when evaluating bias in estimated correlations; Wang et al. (2012) used 5% in their study of the impact of violating factor scaling assumptions, and Ye and Daniel (2017) used 5% for assessing bias in cross-classified random effect models as did Meyers and Beretvas (2006) and Chung et al. (2018).

Similarly, Enders et al. (2018) used a relative bias cutoff of 10% when evaluating the effect of a method for imputing missing data in multilevel models; Holtmann et al. (2016) used 10% for estimated coefficients for a structural equation model as did Wang and Kim (2017); McNeish (2016) used 10% for parameter estimates in a partially-nested multilevel model, and Chen and Leroux (2018) used 10% for evaluating estimates for a cross-classified random effects model. Other relative bias cutoffs appearing in the Monte Carlo literature include Bai and Poon's (2009) 2.5% for estimates in two-level structural equation modeling and Vallejo,

3

Fernandez, Cuesta, and Livacis-Rojas's (2015) 20% for evaluating the impact of heterogeneity in multilevel models.

The rationale for these cutoffs is not statistical but simply that they were used in previous Monte Carlo studies. For example, Myers and Beretvas (2006), Li et al. (2011), Leite and Beretvas (2010), Harring et al. (2012), Wang et al. (2012), Kim et al. (2016), Ye and Daniel (2017), and Chung et al. (2018) cited Hoogland and Boomsa (1998) as the basis of employing a .05 or 5% cutoff. Ironically, the rationale offered by Hoogland and Boomsa was arbitrary: "A boundary for acceptance of .05 is often used in robustness studies." (p. 364). Similarly, McNeish (2016) cited Flora and Curran (2004), Chen and Leroux (2018) cited Curran et al. (1996), Flora and Curran (2004), and Kaplan (1989) as the basis of their cutoff choice, Enders et al. (2018) cited Finch, West, and MacKinnon (1997) and Kaplan (1988), and Holtmann et al. (2017) cited Muthén and Muthén (2002) and Koch et al. (2014).

In some cases, cutoffs are ancillary to categorizing bias as ignorable or non-ignorable because bias values are far from zero. For example, the relative bias of 79% reported in Wang and Kim (2017) provided strong evidence of non-ignorable bias. But bias values close to zero or to a cutoff invite confusing and inconsistent interpretations. Consider the Harring et al. (2012) Monte Carlo study of five methods for estimating and testing a structural parameter representing a quadratic effect in nonlinear structural equation models. These authors employed an absolute bias cutoff of .05 and reported 450 bias values for varying estimation methods, sample sizes, distributions, and reliabilities, with 19.3% of the values exceeding .05. However, 13.3% of the 450 bias values were between .040 and .060 and 25.1% were between .030 and .070, raising the question of why, for example, reported values of .044 and .054, represented ignorable and non-ignorable bias. It's also possible that bias values near zero (e.g., 55.5% were $\leq$ .02) interpreted as ignorable bias simply reflect sampling error.

Similarly, Wang and Kim (2017) used a relative bias cutoff of 10% to identify non-ignorable bias in evaluating the effects of model misspecification on structural coefficients. These authors reported 144 relative bias values, 56.2% of which were described as representing "severe" bias because they exceeded 10%. An examination of these values shows that 24.3% were between 5% and 10% with 10 values (6.9%) equal to 9% and seven (4.9%) equal to 10%, leaving readers to wonder why an estimate that produced a relative bias of 10% represented "severe" (non-ignorable) bias but an estimate with 9% bias was ignorable. Moreover, bias values near zero (e.g., 14.4% were $\leq$ .03) may simply reflect sampling error. Similar patterns appear in many studies reporting Monte Carlo results (e.g., Chung

et al., 2018; Ye, 2015; Jin et al., 2016; Lachowicz et al., 2018; Li et al., 2011; McNeish, 2016).

Type I Error Rate

Another common outcome in Monte Carlo studies in statistics is Type I error rate. An empirical Type I error rate $(\hat{\alpha})$ is computed as the proportion of rejections of a true statistical null hypothesis across J replications; comparing $\hat{\alpha}$ to the user-specified Type I error rate (e.g., $\alpha = .05$) provides evidence of the ability of a statistical test to control its Type I error rate, which is crucial (Serlin, 2002). Relatedly, many Monte Carlo studies report confidence intervals about parameters of interest. For example, the coverage rate (CR) for confidence intervals about parameters of interest is frequently used as an indicator of standard error bias (Brannick et al., 2019; Chen & Leroux, 2018; Maas & Hox, 2004; Seco et al., 2013; Vallejo et al., 2015). A CR such as 88% for a confidence interval with a nominal coverage probability of .95 often reflects negatively-biased standard errors, whereas a CR of 98% reflects positively-biased standard errors. The empirical Type I error rate is simply 1- CR, for example, $\hat{\alpha} = 1 - .88 = .12$.

The relationship between bias and Type I error rates in Monte Carlo studies suggest the latter can be important in evaluating the former. In data analysis retention of a statistical null hypothesis H$_0$: $\theta = 0$ implies ($\hat{\theta} - \theta$) represents sampling error whereas rejection implies ($\hat{\theta} - \theta$) represents sampling error plus an effect. In Monte Carlo studies in statistics retention of H$_0$: $\theta = 0$ (where it is known $\theta = 0$) implies ($\hat{\theta} - \theta$) represents sampling error not bias, whereas rejection of H$_0$: $\theta = 0$ (where it is known $\theta = 0$) implies ($\hat{\theta} - \theta$) represents sampling error plus bias. If relative bias is 4.4% when estimating θ and for the same simulated data $\hat{\alpha} = .048$ for a test of H$_0$: $\theta = 0$ ($\alpha = .05$ and it is known $\theta = 0$), bias equals zero and the 4.4% represents sampling error regardless of the chosen cutoff. On the other hand, if relative bias is 4.4% when estimating θ and for the same simulated data $\hat{\alpha} = .12$ for a test of H$_0$: $\theta = 0$ ($\alpha = .05$ and it is known $\theta = 0$), the difference between .12 and .05 should be treated as reflecting sampling error and bias. The decision of what constitutes $\hat{\alpha} \approx \alpha$ and $\hat{\alpha} \neq \alpha$, and for the latter whether bias is ignorable or non-ignorable, relies on the judgment of authors who can use Type I error rates to support claims an estimator is or is not biased.

It's possible to further quantify the magnitude of bias by computing RMSE and partitioning this quantity into (squared) bias and sampling variance:

$$\sum_{i=1}^{J} \frac{\left(\hat{\theta}_i - \theta\right)^2}{J} = \left(\bar{\hat{\theta}} - \theta\right)^2 + \sum_{i=1}^{J} \frac{\left(\hat{\theta}_i - \bar{\hat{\theta}}\right)^2}{J} , \tag{5}$$

where $\sum_{i=1}^{J} \left(\hat{\theta}_i - \theta\right)^2 \big/ J$ represents RMSE, $(\bar{\hat{\theta}} - \theta)^2$ represents squared bias, $\bar{\hat{\theta}}$ is the mean of the $\hat{\theta}_i$, and $\sum_{i=1}^{J} \left(\hat{\theta}_i - \bar{\hat{\theta}}\right)^2 \big/ J$ represents the sampling variance of an estimator (Gifford & Swaminathan, 1990). Equation (5) permits the contribution of squared bias to RMSE to be computed:

$$\text{Contribution of squared bias } = \frac{\left(\bar{\hat{\theta}} - \theta\right)^2}{\sum_{i=1}^{J} \frac{\left(\hat{\theta}_j - \theta\right)^2}{J}} \tag{6}$$

Values of equation (6) closer to one signal that (squared) bias is dominating differences between estimates and a parameter (sampling error is comparatively modest), and values closer to zero that (squared) bias is playing a modest or negligible role (sampling error is comparatively large).

For example, 42.2% of the empirical Type I error rates reported in Harring et al. (2012) exceeded the bounds of acceptability these authors employed via the Bradley liberal criterion (Bradley, 1978) $.025 \leq \hat{\alpha} \leq .075$, meaning 57.8% were within these bounds and the associated bias values should be treated as sampling error (Harring et al. reported the results of two Monte Carlo studies, one focused on bias when estimating a quadratic effect and the associated RMSE, and a second focused on empirical Type I error rates and power when testing a quadratic effect against zero. It's unclear whether the same simulated data were used in both studies but for illustrative purposes it is assumed the bias, Type I error rate, and RMSEs generated by the two Monte Carlo studies for the quadratic effect are comparable). Correspondingly, the bias values linked to error rates outside these bounds, such as the 11.1% of the error rates equal to .08, should be treated as reflecting both sampling error and bias. Similarly, Wang et al. (2012) reported that 66 of 144 (45.8%) bias values exceeded the selected 5% cutoff but an examination of empirical Type I error rates shows that only seven (4.8%) fell outside the cutoffs of .025 and .075 chosen by these authors using Bradley's liberal criterion. As a result, 57 (86.3%) of the 66 values described by these authors as showing bias

should be treated as representing sampling error, which changes the interpretation of the Monte Carlo results.

As an example, Harring et al. (2012) reported an average bias of .106 for the latent variable score (LVS) estimation method, a nonnormal distribution (NN-1), indicator reliability = .45, and $n = 50$ which implies bias is non-ignorable given the chosen cutoff of .05. These authors also reported Type I error rates assuming $\alpha = .05$. For the above conditions (LVS estimation, NN-1, indicator reliability = .45, $n = 50$) Harring et al. reported $\hat{\alpha} = .07$ which was within their bounds of acceptability using the Bradley criterion (Bradley, 1978), implying the reported bias of .106 should be treated as sampling error (bias = 0). Similarly, for the LVS method, a different nonnormal distribution (NN-2), indicator reliability = .45, and $n = 50$ Harring et al. reported a bias of .105 and $\hat{\alpha} = .08$. The latter falls outside their bounds of acceptability using the Bradley criterion and implies the reported bias value of .105 should be treated as reflecting sampling error and bias. Harring et al. also reported RMSEs which allows equation (6) to be used. For the LVS method, bias = .105, NN-2, indicator reliability = .45, $n = 50$, and $\hat{\alpha} = .08$ equation (6) produces

$$\frac{\left(\bar{\hat{\theta}} - \theta\right)^2}{\sum_{i=1}^{J} \frac{\left(\hat{\theta}_i - \theta\right)^2}{J}} = \frac{(.105)^2}{.103} = .107 \, ,$$

or almost 11%, suggesting bias plays a modest role in differences between estimates and the parameter (sampling error is comparatively large).

The same logic can be applied to other versions of bias reported in Monte Carlo studies. For example, McNeish and Harring (2017) cited Bradley (1978) in using cutoffs of 92% and 98% for CRs about parameters in 95% confidence intervals, with CRs outside these values treated as reflecting biased standard errors. Marrying Type I error rates with bias means the bias of standard errors associated with $92\% \leq CR \leq 98\%$ should be treated as zero, whereas the bias of standard errors associated with $CR < 92\%$ or $CR > 98\%$ is not zero. Similarly, Brannick et al. (2019) used Monte Carlo methods to study the impact of distribution, number of studies, and study sample size in a meta-analysis on the lower bound of confidence intervals for correlations and employed a cutoff of .02 for bias. Table 3 in this article shows that eight (11.1%) of the estimated correlations were biased using the .02 cutoff and of these six were between .02 and .03, whereas Table 4

shows that 13.2% of the conditions provided adequate coverage based on choice of CR $\geq$ 90% meaning $\hat{\alpha} < .10$ was deemed acceptable. If the Bradley liberal criterion was used bias values associated with 85% $\leq$ CR $\leq$ 95% would be treated as zero, whereas values associated with CR < 85% or CR > 95% would be treated as providing evidence of bias. Bias, Type I error rates or CRs, and RMSEs are frequently reported in Monte Carlo studies (e.g., Chen & Leroux, 2018; Chung et al., 2018; Lachowicz et al., 2018; Seco et al., 2013; Vallejo et al., 2015). Many that don't could, permitting the calculations illustrated above.

Conclusion

Monte Carlo studies in statistics frequently report measures of bias that are judged to be ignorable or non-ignorable based on arbitrary cutoffs. Linking Type I error rates of statistical tests (or coverage rates of confidence intervals about parameters of interest) with the bias values of estimators used in those tests offers a quantitative framework based on simple statistical theory: If a statistical null hypothesis H_0: $\theta = 0$ (where it is known $\theta = 0$) is retained then $(\hat{\theta} - \theta)$ represents sampling error and if it is rejected in a Monte Carlo study $(\hat{\theta} - \theta)$ represents sampling error and bias. This strategy presumes a statistical test is available for an estimator either analytically or using a Monte Carlo-based sampling variance. If the root mean square error is reported the impact of squared bias can be estimated, helping to quantify the impact of bias on $(\hat{\theta} - \theta)$. Bias, empirical Type I error rates, and root mean square error values are usually straightforward to compute and report in a Monte Carlo study and marrying these quantities when interpreting bias should enhance interpretations of Monte Carlo results and move decisions about bias towards a quantitative framework.

References

Bai, Y., & Poon, W. Y. (2009). Using Mx to analyze cross-level effect in two-level structure. *Structural Equation Modeling, 16*(1), 163-178, doi: 10.1080/10705510802561527

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology, 31*(2), 44–152. doi: 10.1111/j.2044-8317.1978.tb00581.x

Brannick, M. T., Potter, S., & Teng, Y. (2019). Quantifying uncertainty in the meta-analytic lower bound estimate. *Psychological Methods, 24*(6), 754-773. doi: 10.1037/met0000217

Chung, H., Kim, J., Park, R. & Jean, H. (2018). The impact of sample size in cross-classified multiple membership multilevel models. *Journal of Modern Applied Statistical Methods, 17*(1), eP2491. doi: 10.22237/jmasm/1542209860

Curran, P. J., West, S. G., & Finch, J. F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods, 1*(1), 16-29. doi: 10.1037/1082-989X.1.1.16

Enders, C. K., Keller, B. T., & Levy, R. (2018). A full conditional specification approach to multilevel imputation of categorical and continuous variables. *Psychological Methods, 23*(2), 298-317. doi: 10.1037/met0000148

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods, 9*(4), 466-491. doi: 10.1037/1082-989X.9.4.466

Gifford, J. A., & Swaminathan, H. (1990). Bias and the effect of priors in Bayesian estimation of parameters of item response models. *Applied Psychological Measurement, 14*(1), 33-43. doi: 10.1177/014662169001400104

Harring, J. R., Weiss, B. A., & Hsu, J. (2012). A comparison of methods for estimating quadratic effects in nonlinear structural equation models. *Psychological Methods, 17*(2), 193-214. doi: 10.1037/a0027539

Harwell, M. R., Kohli, N., & Peralta-Torres, Y. (2018). A survey of reporting practices of computer simulation studies in statistical research. *The American Statistician, 72*(4), 321-327. doi: 10.1080/00031305.2017.1342692

Holtmann, J., Koch, T., Lochner, K., & Eid, M. (2016). A comparison of ML, WLSMV, and Bayesian methods for multilevel structural equation models in small samples: A simulation study. *Multivariate Behavioral Research, 51(5)*, 661-680. doi: 10.1080/00273171.2016.1208074

Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling: An overview and a meta-analysis. *Sociological Methods & Research, 26*(3), 329-367. doi: 10.1177/0049124198026003003

Jin, S., Luo, H., Yang-Wallentin, F. (2016). A simulation study of polychoric instrumental variable estimation in structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal, 23*(5), 680-694. doi: 10.1080/10705511.2016.1189334

Kaplan, D. (1988). The impact of specification error on the estimation, testing and improvement of structural equation models. *Multivariate Behavioral Research, 23*(1), 69-86. doi: 10.1207/s15327906mbr2301_4

Kaplan, D. (1989). A study of the sampling variability of the z-values of parameter estimates from misspecified structural equation models. *Multivariate Behavioral Research, 24*(1), 41-57. doi: 10.1207/s15327906mbr2401_3

Kim, E. S., Joo, S. H., Lee, P., Wang, Y., & Stark, S. (2016). Measurement invariance testing across between-level latent classes using multilevel factor mixture modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 23*(6), 870-877. doi: 10.1080/10705511.2016.1196108

Koch, T., Schultze, M., Eid, M., & Geiser, C. (2014). A longitudinal multilevel CFA-MTMM model for interchangeable and structurally different methods. *Frontiers in Psychology, 5*, 311. doi: 10.3389/fpsyg.2014.00311

Lachowicz, M. J., Preacher, K. J., & Kelley, K. (2018). A novel measure of effect size for mediation analysis. *Psychological Methods, 23*(2), 244-261. doi: 10.1037/met0000165

Leite, W., & Beretvas, S. N. (2010). The performance of multiple imputation for Likert-type Items with missing data. *Journal of Modern Applied Statistical Methods, 9*(1), 64-74. doi: 10.22237/jmasm/1272686820

Li, J. C., Chan, W., & Cui, Y. (2011). Bootstrap standard error and confidence intervals for the correlations corrected for indirect range restriction. *British Journal of Mathematical and Statistical Psychology, 64*(3), 367-387. doi: 10.1348/2044-8317.002007

Maas, C. J. M., & Hox, J. J. (2004). Robustness issues in multilevel regression analysis. *Statistica Neerlandica, 58*(2), 127-137. doi: 10.1046/j.0039-0402.2003.00252.x

McNeish, D. M. (2016). Estimation methods for mixed logistic models with few clusters. *Multivariate Behavioral Research, 51*(6), 790-804. doi: 10.1080/00273171.2016.1236237

McNeish, D. M., & Harring, J. R. (2017). Clustered data with small sample sizes: Comparing the performance of model-based and design-based approaches. *Communications in Statistics – Simulation and Computation, 46*(2), 855-869. doi: 10.1080/03610918.2014.983648

Meyers, J. L., & Beretvas, S. N. (2006). The impact of inappropriate modeling of cross-classified data structures. *Multivariate Behavioral Research, 41*(4), 473-497. doi: 10.1207/s15327906mbr4104_3

Muthén, L. K., & Muthén, B. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling: A Multidisciplinary Journal, 9*(4), 599-620. doi: 10.1207/S15328007SEM0904_8

Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman, W. (1996). *Applied linear statistical models* (4th ed.). Chicago, IL: Irwin.

Seco, G. V., Garcia, M. A., Garcia, M. P. F., & Rojas, P. E. L. (2013). Multilevel bootstrap analysis with assumptions violated. *Piscothema, 25*(4), 520-528. doi: 10.7334/psicothema2013.58

Serlin, R. C. (2002). Constructive criticism. *Journal of Modern Applied Statistical Methods, 1*(2), 202-227. doi: 10.22237/jmasm/1036109700

Wang, Y., & Kim, E. S. (2017). Evaluating model fit and structural coefficient bias: A Bayesian approach to multilevel bifactor model misspecification. *Structural Equation Modeling: A Multidisciplinary Journal, 24(5)*, 699-713. doi: 10.1080/10705511.2017.1333910

Wang, D., Whittaker, T. A., & Beretvas, S. N. (2012). The impact of violating factor scaling method assumptions on latent mean difference testing in structured means models. *Journal of Modern Applied Statistical Methods, 11*(1), 24-41. doi: 10.22237/jmasm/1335844920

Ye, F. (2015). Latent growth curve analysis with dichotomous items: Comparing four approaches. *British Journal of Mathematical and Statistical Psychology, 69*(1), 43-61. doi: 10.1111/bmsp.12058

Ye, F., & Daniel, L. (2017). The impact of inappropriate modeling of cross-classified data structures on random-slope models. *Journal of Modern Applied Statistical Methods, 16*(2), 458-484. doi: 10.22237/jmasm/1509495900

Yuan, K. H., Tong, X., & Zhang, Z. (2015). Bias and efficiency for SEM with missing data and auxiliary variables: Two-stage robust method versus two-stage ML. *Structural Equation Modeling: A Multidisciplinary Journal, 22*(2), 178-192. doi: 10.1080/10705511.2014.935750