Advancing Quantitative Science With Monte Carlo Simulation

PsyPag & MSCP-Section Simulation Summer School

Hok Chio (Mark) Lai, Winnie Wing-Yee Tse, & Yichi Zhang

2021/06/16

Overview

What is Monte Carlo (MC) simulation?

Simulating Data From a Normal Distribution

Properties of Statistical Methods

Monte Carlo Simulation Study/Experiment

Monte Carlo Methods

- 1930s-1940s: Nuclear physics (the Manhattan Project)
  - Key figures:
    - Stanislaw Ulam
    - John von Neumann
    - Nicholas Metropolis
  - Naming: Casino in Monaco

Image credit: sam garza from Los Angeles, USA, CC BY 2.0 https://creativecommons.org/licenses/by/2.0, via Wikimedia Commons

Why Do We Do Statistics?

- To study some target quantity in the population
  - Based on a limited sample
- How do we know that a statistics/statistical method gets us to a reasonable answer?
  - Analytic method
  - Simulation

> MC is one way to understand the properties of one or more statistical procedures

What is MC (in Statistics)?

A statistical technique that uses (psuedo-random) sampling to get numerical results

- Simulate the *process of repeated random sampling*
    - E.g., repeatedly drawing sample of IQ scores of size 10 from a population
- Approximate *sampling distributions*
    - Using **pseudorandom samples**
- Study properties of **statistical methods**
    - regression coefficients, fit index
    - compare multiple estimators or modeling approaches

Simulating Random Data From a Normal Distribution

Generating Random Data in R

With MC, one simulates the process of generating the data with an assumed **data generating model/mechanism**

```
rnorm(5, mean = 0, sd = 1)
```

```
## [1]  0.1185515 -1.0909555 -1.0258400  0.1501688  1.3313129
```

```
rnorm(5, mean = 0, sd = 1)  # numbers changed
```

```
## [1] -0.53826642  2.00587115 -0.73160714 -0.37485398 -0.04361177
```

Setting the Seed

- Most programs use algorithms to generate numbers that look like random, i.e., *pseudorandom*
    - Completely determined by the **state** of the random number generator, which can be set by the seed

> For replicability, set the seed explicitly

```r
state1 <- .Random.seed  # state of RNG
rnorm(5, mean = 0, sd = 1)
```

```
## [1]  0.049911283 -0.799108882 -0.791406078  1.481268818 -0.005218739
```

```r
set.seed(1)
state2 <- .Random.seed  # state of RNG changed
identical(state1, state2)
```

```
## [1] FALSE
```

```r
rnorm(5, mean = 0, sd = 1)
```

```
## [1] -0.6264538  0.1836433 -0.8356286  1.5952808  0.3295078
```

```r
set.seed(1)
state3 <- .Random.seed  # state of RNG unchanged with the same seed
identical(state2, state3)
```

```
## [1] TRUE
```

```r
rnorm(5, mean = 0, sd = 1)  # same seed, same numbers
```

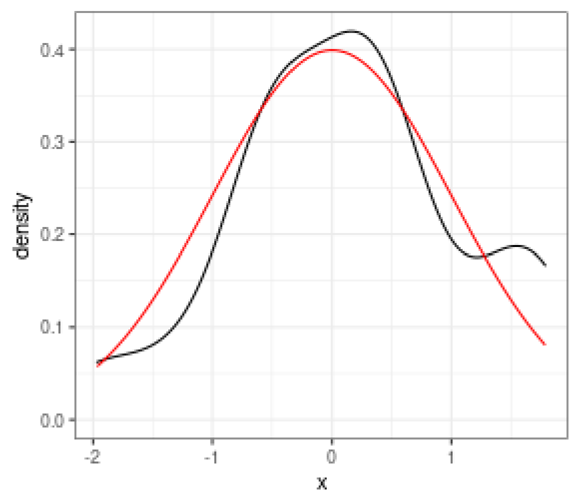
```
## [1] -0.6264538  0.1836433 -0.8356286  1.5952808  0.3295078
```

Generating Data From Univariate Distributions

```
rnorm(n, mean, sd)      # Normal distribution (mean and SD)
runif(n, min, max)      # Uniform distribution (minimum and maximum)
rchisq(n, df)           # Chi-squared distribution (degrees of freedom)
rbinom(n, size, prob)   # Binomial distribution
```

MC Approximation of $N(0, 1)$

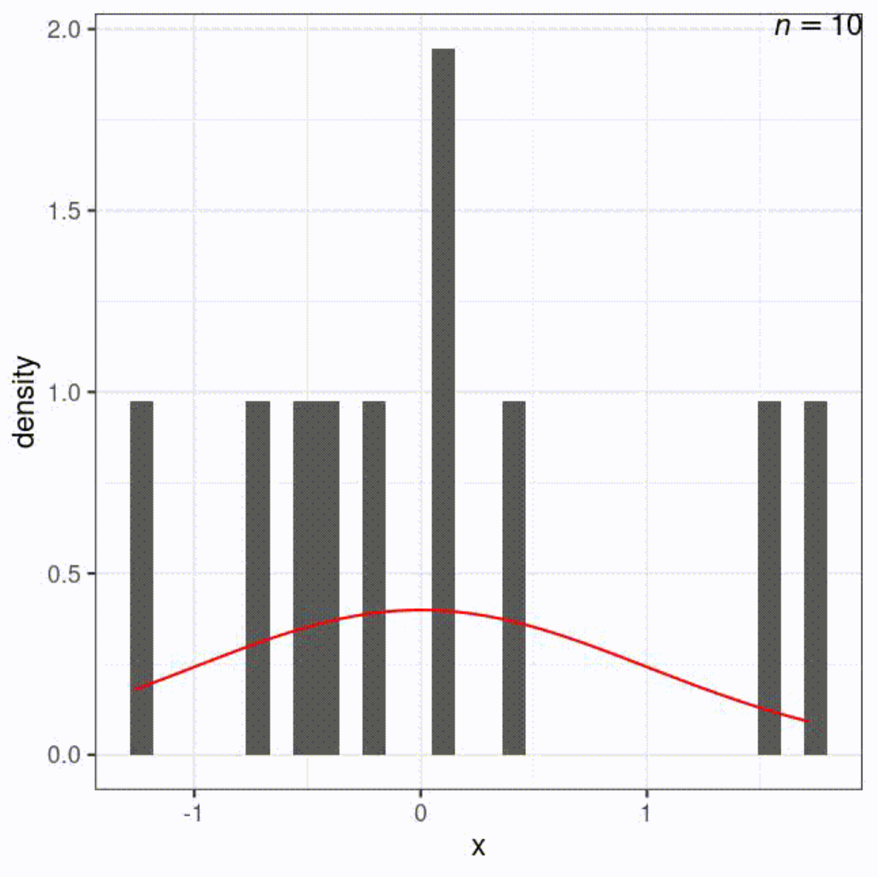
```r
library(tibble)
library(ggplot2)
set.seed(123)
nsim <- 20  # 20 samples
sam <- rnorm(nsim)  # default is mean = 0 and sd = 1
ggplot(tibble(x = sam), aes(x = x)) +
  geom_density(bw = "SJ") +
  stat_function(fun = dnorm, col = "red")  # overlay normal curve in red
```



Exercise

Try increasing `nsim` to 100, then 1,000

Exercise



Evaluating Properties of Statistical Methods

Some Types of Methods Studied by Simulations

Adapted from Table 3 of Morris, et al. (2019)

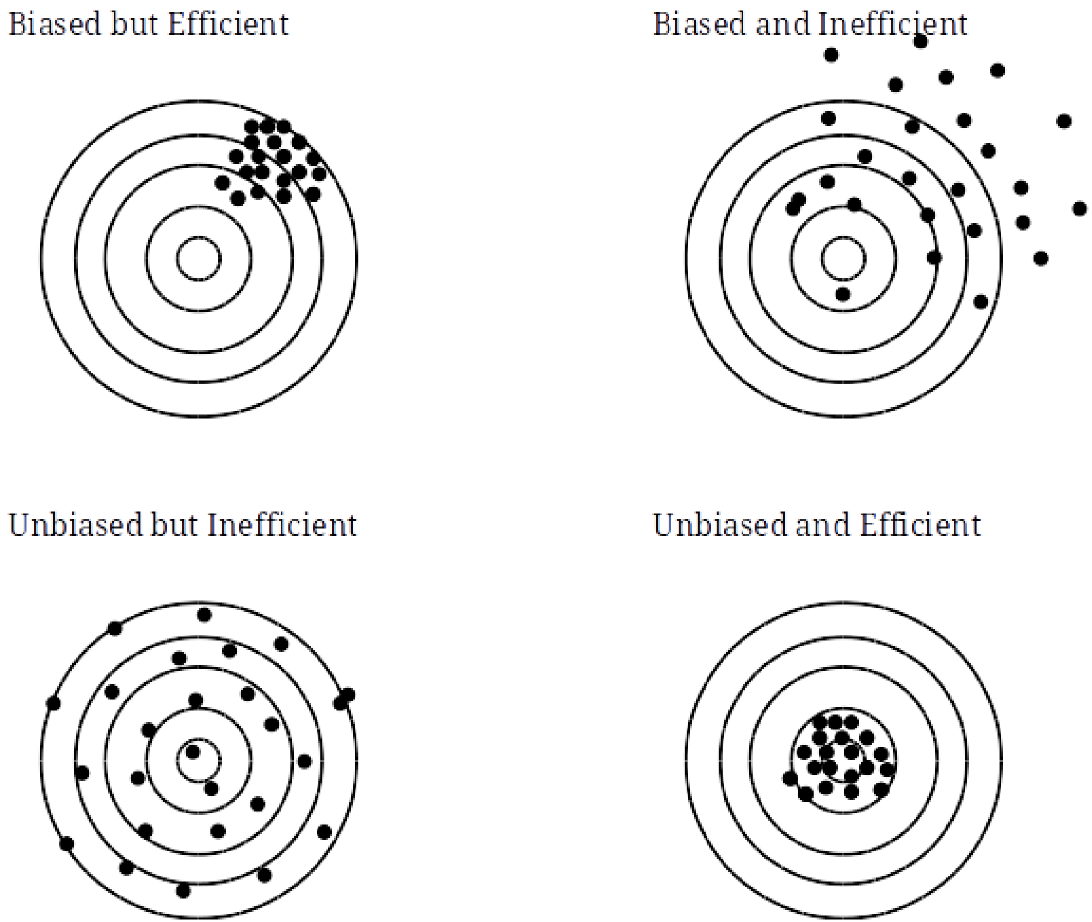
Task	Statistical Method	Properties
Estimation	Estimator	Bias, efficiency, consistency
Uncertainty	Standard error, confidence interval	SE bias, coverage
Inference	Hypothesis testing	Type I error rate, power
Model Selection	Model selection index	Correct model rate

One additional property: **Robustness**---resilience against outliers and assumption violations

Estimation: Parameter vs Estimator

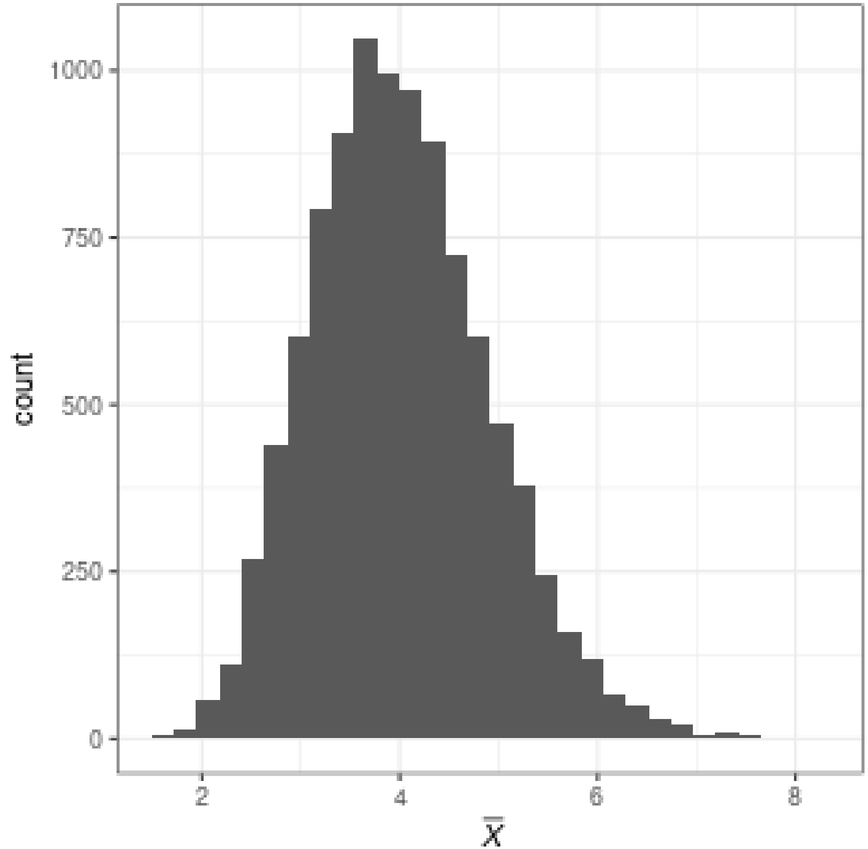
- **Estimator**/statistic: $T(\mathbf{X})$, or simply T
  - How good does it estimate the population parameter, θ?
- Examples:
  - $T = \bar{X}$ estimates $\theta = \mu$
  - $T = \dfrac{\sum_i (X_i - \bar{X})^2}{N-1}$ estimates $\theta = \sigma^2$

What is a Good Estimator?



Sampling Distribution

- What is it?



Example I

Simulating Means and Medians

Monte Carlo Simulation Study

Examples in the Literature

- Curran, West, & Finch (1996, Psych Methods) studied the performance of the χ^2 test for nonnormal data in CFA
- Kim & Millsap (2014, MBR) studied the performance of the Bollen-Stine Bootstrapping method for evaluating SEM fit indices
- MacCallum, Widaman, Zhang, & Hong (1999, Psych Methods) studied sample size requirement for getting stable EFA results
- Maas & Hox (2005, Methodology) studied the sample size requirement for multilevel models

A Simulation Study is an Experiment

Experiment	Simulation
Independent variables	Design factors
Experimental conditions	Simulation conditions
Controlled variables	Other parameters
Procedure/Manipulation	Data generating model
Dependent variables	Evaluation measures
Substantive theory	Statistical theory
Participants	Replications

Framework

(Sigal, et al., 2016; Chalmers, et al., 2020; Morris, et al., 2019)

- Research questions
  - *What is the effect of ignoring random slopes in a growth model?*

- Design
  - *3 (N = 50, 100, 200)* $\times$ *2 (slope variance = 0.1, 0.5) design*
  - *Constant: 4 time points, maximum likelihood estimation, etc*
  - *500 replications*

- Date-generating model (fixed and random components)
  - *linear growth model with normally distributed errors*

Framework (cont'd)

- Statistical methods
    1. *slope estimate and standard error under correctly specified latent growth model with lavaan*
    2. *slope estimate and standard error under misspecified model*

- Evaluative measures
    - *convergence, bias, SE bias, relative efficiency*

- Summary and reporting
    - *Table, plot*

Design

Like experimental designs, conditions should be carefully chosen

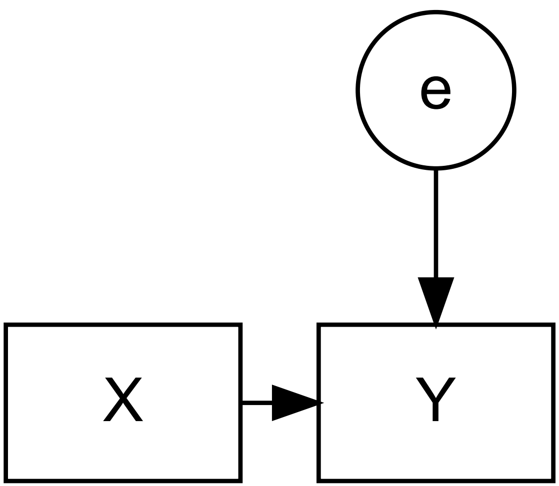
- What to manipulate? Sample size? Effect size? Why?
  - Based on statistical theory and reasoning
  - E.g., Gauss-Markov theorem: regression coefficients are unbiased with violations of distributional assumptions
- What levels? Why?
  - Needs to be realistic for empirical research
  - Maybe based on previous systematic reviews,
  - Or a small review of your own

Full Factorial designs are most commonly used

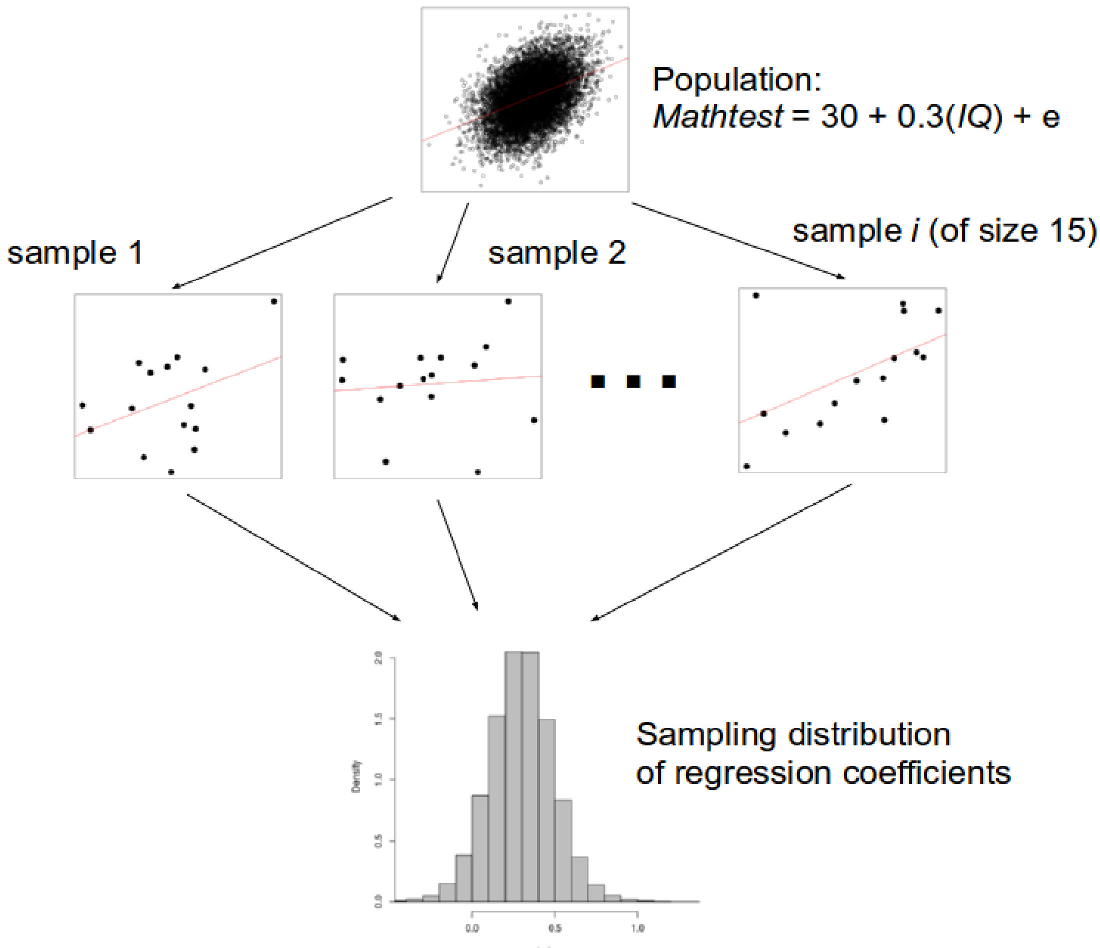
- Fractional factorial may sometimes be beneficial (Skrondal, 2000)

Data Generation

- Starts with a statistical data generating model
  - E.g., $Y_i = \beta_0 + \beta_1 X_i + e_i, \quad e_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$
    - Systematic (deterministic) component: X_i
    - Random (stochastic) component: e_i
    - Constants (parameters): β_0, β_1
  - Y_i completely determined by $X_i, e_i, \beta_0, \beta_1$



Model-Based Simulation



Statistical Methods

- Analyze each simulated data set with one or more approaches/models
- Obtain statistics of interest (e.g., estimate, SE, CI, p value)

Evaluative Measures

Some definitions:

Mean estimate	$\bar{\hat{\theta}} = \sum_{i=1}^{R} \hat{\theta}_i / R$
Average estimated SE	$\bar{\hat{\mathrm{SE}}}(\hat{\theta}) = \sum_{i=1}^{R} \hat{\mathrm{SE}}(\hat{\theta}_i) / R$
Empirical SE	$\hat{SD}(\hat{\theta})$ = $\sqrt{\frac{\sum_{i=1}^{R}(\theta_i - \bar{\hat{\theta}})^2}{R}}$

For estimators

Raw bias	$\bar{\hat{\theta}} - \theta$
Relative bias	Bias / θ
Standardized bias	Bias / $\hat{SD}(\hat{\theta})$
Relative efficiency (RE; for unbiased estimators)	$\mathrm{RE}(\hat{\theta}, \tilde{\theta})$ = $\frac{\hat{SD}^2(\tilde{\theta})}{\hat{SD}^2(\hat{\theta})}$
Mean squared error (MSE)	$\mathrm{Bias}^2 + \hat{\mathrm{Var}}(\hat{\theta})$
Root Mean squared error (RMSE)	$\sqrt{\mathrm{MSE}}$

For uncertainty

SE bias	$\bar{SE}(\hat{\theta}) - \hat{SD}(\hat{\theta})$
Relative SE bias	SE bias / $\hat{SD}(\hat{\theta})$
Coverage	proportion of sample CIs containing θ

For statistical inferences:

Power/Empirical Type I error rates	proportion with $p < \alpha$ (usually α = .05)

Summary and Reporting

Same as analyzing real data

- Plots, figures
- ANOVA, regression
    - E.g., 3 (sample size) × 4 (parameter values) 2 (models) design: 2 between factors and 1 within factor

Example II

Simulation Example on Structural Equation Modeling

Number of Replications

MC requires large number of replications. But how large?

- Monte Carlo (MC) Error
  - Like standard error (SE) for a point estimate
- For expectations (e.g., bias)
  - MC Error = $\hat{SD}(\hat{\theta})/\sqrt{R}$

E.g., if one wants the MC error to be ≤2.5% of the sampling variability, *R* needs to be 1 / $.025^2$ = 1,600

For power/Type I error/CI coverage,

- MC Error = $\sqrt{\frac{p(1-p)}{R}}$

E.g., with *R* = 250, and empirical Type I error = 5%, MC Error = 1.38%

Further Readings

Carsey, et al. (2014); Morris, et al. (2019) for a gentle introduction

Chalmers, et al. (2020) and Sigal, et al. (2016) for using the R package `SimDesign`

Harwell, et al. (2018) for a review of design and reporting practices

Skrondal (2000), Serlin (2000), and Bandalos, et al. (2013) for additional topics

Thanks!

Slides created via the R package xaringan.

Contact:

Mark Lai (hokchiol@usc.edu)

Winnie Wing-Yee Tse (wingyeet@usc.edu)

Yichi Zhang (yzhang97@usc.edu)

References

Bandalos, D. L. et al. (2013). "Use of Monte Carlo studies in structural equation modeling research". In: *Structural equation modeling. A second course*. Ed. by G. R. Hancock and R. O. Mueller. 2nd ed. Charlotte, NC: Information Age, pp. 625-666.

Boomsma, A. (2013). "Reporting Monte Carlo studies in structural equation modeling". In: *Structural Equation Modeling. A Multidisciplinary Journal* 20, pp. 518-540. DOI: 10.1080/10705511.2013.797839.

Bradley, J. V. (1978). "Robustness?" In: *British Journal of Mathematical and Statistical Psychology* 31, pp. 144-152. DOI: 10.1111/j.2044-8317.1978.tb00581.x.

Carsey, T. M. et al. (2014). *Monte Carlo simulation and resampling methods for social science*. Thousand Oaks, CA: Sage.

Chalmers, R. P. et al. (2020). "Writing Effective and Reliable Monte Carlo Simulations with the SimDesign Package". In: *The Quantitative Methods for Psychology* 16.4, pp. 248-280. DOI: 10.20982/tqmp.16.4.p248.

References (cont'd)

Collins, L. M. et al. (2001). "A comparison of inclusive and restrictive strategies in modern missing data procedures". In: *Psychological Methods* 6, pp. 330-351. DOI: 10.1037//1082-989X.6.4.330.

Harwell, M. et al. (2018). "A survey of reporting practices of computer simulation studies in statistical research". In: *The American Statistician* 72, pp. 321-327. DOI: 10.1080/00031305.2017.1342692.

Hoogland, J. J. et al. (1998). "Robustness studies in covariance structure modeling". In: *Sociological Methods & Research* 26, pp. 329-367. DOI: 10.1177/0049124198026003003.