An Effect of Violating Independent and Identically Distributed Random Variables in Bayesian Statistical Modeling

Author: Ani Gumruyan
BS in Data Science
Akian College of Science and Engineering
American University of Armenia

Supervisor: Dr. Tadamasa Sawada
Akian College of Science and Engineering,
American University of Armenia
Department of Psychology,
Russian-Armenian (Slavonic) University

***Abstract*—In statistical modeling, the precise choice of methods is closely related to the assumptions about the data, the violation of which can yield biased results. Bayesian statistics uses a descriptive approach to model formulation and, thus, heavily relies on data-related assumptions. One common assumption is the independent and identical distribution of variables (i.i.d.). The present research paper discusses the effect of violating i.i.d. in Bayesian modeling for two types of parameter estimation. The estimated parameters are the mean for one type and the correlation for the other. The effect of the violation was tested using synthetically generated data under two conditions: (i) the i.i.d. assumption was satisfied, and (ii) the i.i.d. assumption was not satisfied. The analysis of the obtained results showed that some models were more susceptible to the violation of i.i.d. than others, given the inconsistency between the assumptions used in the model formulation and the features of the data.**

***Index Terms*—Bayesian Statistical Modeling; Independent and Identically Distributed Random Variables; Parameter Estimation**

INTRODUCTION

Any statistical analysis method is based on assumptions about the data, and the outcomes of the analysis method become less reliable once the proposed assumptions are violated (Sawada 2023). It is, therefore, essential to understand the assumptions and how robust the analysis is to the violation of the assumptions beforehand to avoid misleading results. Various studies have tested the robustness of statistical analysis methods to the violation of assumptions, like the normality of distribution (Khan and Rayner 2003; Gottardo and Raftery 2009) and equal variance of distributions across conditions (Kasuya 2001; Ruscio and Roche 2012). While the extent of assumption violation is discussed differently with respect to approach, the common ground is that the violations and respective effects should be viewed as matters of relative degree, given that, in practice, it is hard to adhere to assumptions strictly, and yielded violations are of quantitative nature. (Bergh, Wagenmakers, and Aust 2023)

Rooted in the Bayes theorem, Bayesian Statistics discusses the probability of an event based on prior information about conditions given the observed data. (Eddy 2004) As a descriptive method of model formulation, Bayesian statistical modeling heavily relies on data-related assumptions. Given the specific features of the data, Bayesian models can be formulated accordingly to accommodate those features. Thus, assumptions for analysis emerge from the model formulation and give a relative freedom to reformulate models based on the data assumptions or introduce assumptions about the data aligning with the chosen model (Sawada 2023). This strength of Bayesian models requires careful consideration of assumption adjustments to avoid cases where data violates the assumptions in the proposed model, which can subsequently affect the analysis outcome and the reliability of results.

Sawada (2023) carried out the analysis of a single case using Bayesian modeling and discussed the effect of violating data assumptions on the results of an analysis. Aside from the already mentioned assumption of equal variance across conditions, the study reflected on yet another common assumption - the independent and identically distributed (i.i.d.) random variables. The idea behind i.i.d. implies that each random variable has the same probability distribution as the others and that all are mutually independent. Sawada (2023) showed that the observed violations of assumptions (unequal variance across participants and interdependence between means and variances) could have caused biased analysis results.

In the present research paper, the focus was on assessing the effect of violating the assumption of independent and identically distributed (i.i.d.) random variables in Bayesian statistical modeling. For estimating the mean of a distribution and for estimating the correlation coefficient between two sets of random variables, I implemented four Bayesian models to estimate the population mean and three models to estimate the correlation coefficient and tested how their outputs were affected by the violation of the assumption. These models were tested under two conditions in Monte Carlo simulations: (i) the assumption of i.i.d. was satisfied, and (ii) the assumption was not satisfied. The estimated parameters from the models were then compared between these conditions to assess the effect of the violation on the results of the implemented models. Based on the results of the simulation experiment, the robustness of Bayesian models to the violation of i.i.d. is discussed.

METHODOLOGY

The R programming language was chosen for analysis, and JAGS models (Just Another Gibbs Sampler) (Plummer 2003) were implemented and interfaced in RStudio by utilizing the package `rjags`.

Mean

Consider the following hypothetical scenario: the mean performance of people in a task is estimated by testing multiple participants. For instance, IQ or visual acuity measurements. Each person is tested in multiple trials, and the average of the trials of the person is computed. Subsequently, the population mean is estimated by computing the average across the individual averages of the participants. The data from such an experiment can be analyzed using the following four Bayesian models (Appendix), formulated by revising models from Lee and Wagenmakers (2013).

- Model-A considers each participant's individual mean and standard deviation. Trials of the individual participants are regarded as random samples from distributions with the common mean but with different standard deviations across the participants.
- Model-B is similar to Model-A, but it considers a common standard deviation across participants, and individual standard deviations do not affect the estimation.
- Model-C has a hierarchical structure; each participant's data is drawn from a common population mean with individual variation around that mean.
- Model-D is similar to Model-C, but it also considers participant-specific standard deviation, assuming each participant has a unique mean and standard deviation.

To generate the synthetic data for mean-estimating models, the `generate_data` custom function was implemented that iteratively generated participant means and standard deviations until the correlation between parameters fell below a threshold value (0.05 in the current case), ensuring relative independence. For each participant, trial data was generated using the custom `rnorm_revnorm` function (generating samples from a standard normal distribution, standardizing, and then reversing the normalization process), resulting in two sets of trial data: one with almost no correlation (data_nocorr) and another with almost perfect correlation (data_perfcorr, after sorting means and standard deviations in ascending order). Furthermore, two additional datasets that stored the separate mean and standard deviation values for almost no correlation (data_sep_nocorr) and almost perfect correlation (data_sep_perfcorr) conditions were generated.

The simulation stage for Mean estimation was automated with the custom function `simulate_mean`, carrying out 30 MC (Monte-Carlo) simulations for four models under two conditions for data with respect to the chosen parameter. The design of the simulation allows for three parameter choices:

1. Number of participants (by default 15): 5, 15, and 45.
2. Number of trials (by default 10): 5, 10, 20, and 40.
3. Standard deviation in normal distribution for standard deviations (by default 1.0): 0.2, 1.0, and 2.0.

Each simulation would consider a set of values for the chosen parameter, generate data utilizing the already mentioned custom function for data generation, run the four Bayesian JAGS models with respective sets of initial values and data input (data_sep_nocorr & data_sep_perfcorr for Models A & B, data_nocorr & data_perfcorr for Models C & D), and store the estimated values of population mean (mu_mu) from model output in tabular format. The obtained three data frames storing simulation results with respect to three parameters were then saved to the Rdata file for further analysis.

Correlation

Consider another hypothetical scenario: a set of values for two parameters is generated to estimate their correlation. For instance, the linear relationship between IQ and visual acuity measurements. This time, the data from such an experiment can be analyzed using the following three Bayesian models (Appendix), formulated by revising models from Lee and Wagenmakers (2013).

- In Model-1, observations of the two variables are assumed to follow a bivariate normal distribution, assuming a symmetric correlation structure between the variables.
- Model-2 extends Model-1 by adding a measurement error to the observed variables using lambda error, assuming that the observed variables (x) are the adjusted measurements of the true variables (y) with known and fixed measurement errors.
- Model-3 further extends Model-2 by allowing for different measurement errors for each variable. Nine scenarios of variable dependency for measurement errors are considered.

To generate the synthetic data for correlation-estimating models, the `generate_data` custom function was implemented to generate multivariate normal data (dataX & dataY) with a covariance matrix based on specified correlation coefficients and marginal variances. Additionally, a `generate_individual_lambdas` custom function was implemented to generate individual error measurement errors for the third model under nine scenarios of data dependency. Specifically:

- Dependency on dataX: Scenario 1-A, where the error parameter along the x-axis depends on dataX while the error parameter across the y-axis is constant, and vice versa for Scenario 1-B.
- Total dependency on dataX: Scenario 2, where the error parameters along the x-axis and y-axis depend on dataX.
- Dependency on dataY: Scenario 3-A, where the error parameter along the x-axis depends on dataY while the error parameter across the y-axis is constant, and vice versa for Scenario 3-B.
- Total dependency on dataY: Scenario 4, where the error parameters along the x-axis and y-axis depend on dataY.

- Dependency on dataXdataY: Scenario 5-A, where the error parameter along the x-axis depends on the product of dataX and dataY, while the error parameter across the y-axis is constant, and vice versa for Scenario 5-B.
- Total dependency on dataXdataY: Scenario 6, where the error parameters along the x-axis and y-axis depend on the product of dataX and dataY.

The simulation stage for Correlation estimation was automated with the custom function `simulate_correlation`, carrying out 15 MC (Monte-Carlo) simulations for three models and the third one with 9 scenarios for measurement error with respect to the chosen parameter. The design of the simulation allows for four parameter choices:

1. Number of points (by default 500): 50, 100, and 500.
2. Values of lambdas for marginal variances (by default [1.0, 1.0]): [0.3, 2.0], [1.0, 1.0], and [2.0, 0.3].
3. The range of standard deviation in uniform distribution for measurement error dependency (by default [0.01, 0.2]): [0.01, 0.1] and [0.01, 0.2].
4. Correlation coefficient in data generation (by default 0.5): 0.2, 0.5, and 0.8.

Each simulation would consider a set of values for the chosen parameter, generate data utilizing the already mentioned custom function for data generation, and run the three Bayesian JAGS models (along with scenarios for model 3) with respective sets of initial values and data. The four data frames that were obtained storing the estimated values of the correlation coefficient (r) with respect to four parameters were then saved to the Rdata file for further analysis.

RESULTS

Mean

For the visual analysis of simulation results for each of the three parameters, two variants of plots were generated: (i) scatterplot of estimate values in each simulation, and (ii) scatterplot with error bars of the mean value of estimate across the simulations faceted by models and colored by correlation type.

Number of Participants: The results of simulations with respect to the number of participants are plotted in Figure 1. The results show that Models B and C are less affected by the violation of i.i.d since the blue and orange dots, representing estimates for correlated and uncorrelated data, respectively, are densely centered around close values of the population mean. On the other hand, Models A and D, especially A, are visibly affected by the violation of i.i.d, yielding estimates centered around more distant population mean values.

Notably, with more participants, the effect of violation for the affected models becomes more evident; we can see how the means for estimates for Models A and D are rationally closer when 5 participants are tested compared to the case with 45 participants.

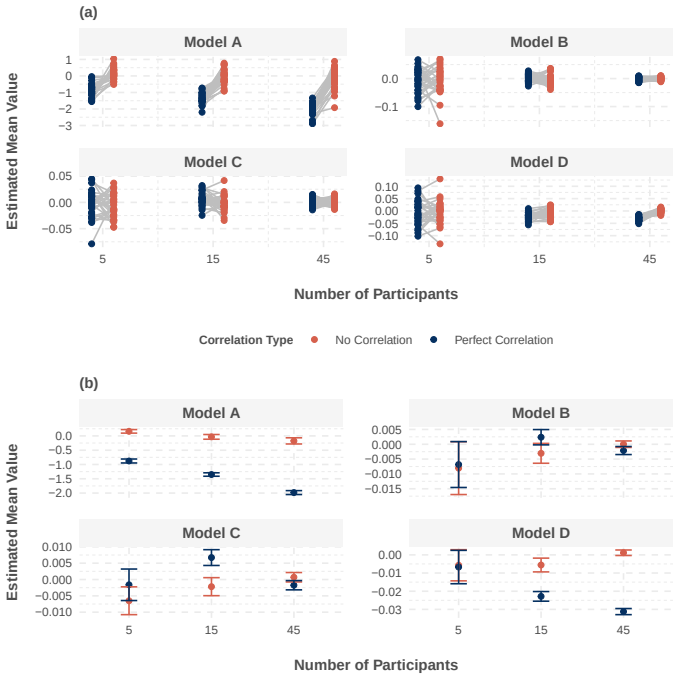


Fig. 1. (a) Scatterplot of estimated values for mean and (b) Scatterplot of the mean of estimated values for mean with error bars across simulations with respect to the number of participants, faceted by four models and colored by correlation type. In subfigure (a), the grey lines connect observations from the same simulation.

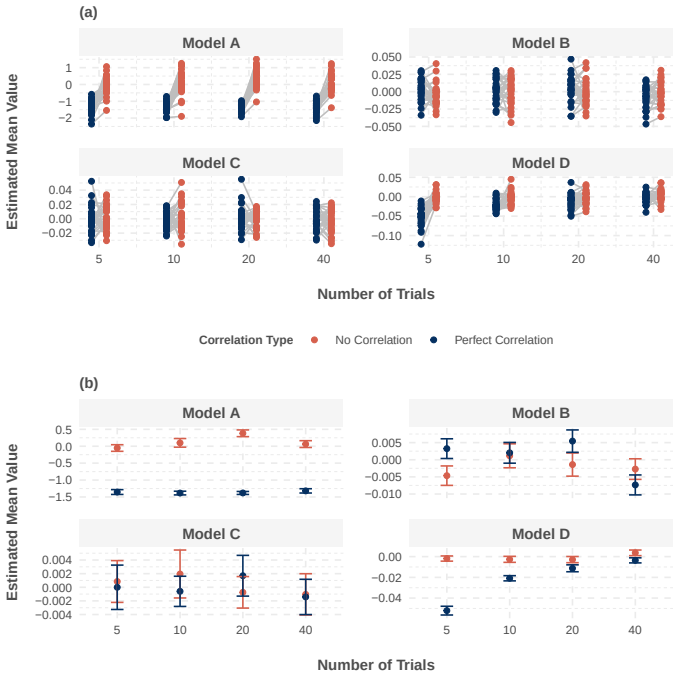


Fig. 2. (a) Scatterplot of estimated values for mean and (b) Scatterplot of the mean of estimated values for mean with error bars across simulations with respect to the number of trials, faceted by four models and colored by correlation type. In subfigure (a), the grey lines connect observations from the same simulation.

Number of Trials: The results of simulations with respect to the number of trials (Figure 2) support the previous observation that Models B and C are less affected by the violation of i.i.d than Models A and D, especially Model A. Interestingly, unlike the previous case with number of participants, with more trials, the effect of violation for Model D becomes less apparent, and the general pattern shows how the increased amount of trials yields more stable value across three out of four models, with estimates across simulations getting more densely centered.

Standard Deviation Value: Finally, the results of simulations with respect to the value of standard deviation in normal distribution for standard deviations (Figure 3) strengthen the insight that Models A and D are more affected by the violation of i.i.d than Models B and C. Notably, similar to the case with the number of participants, with a wider range of uniform distribution of standard deviation, the effect of violation for Model D becomes more apparent.

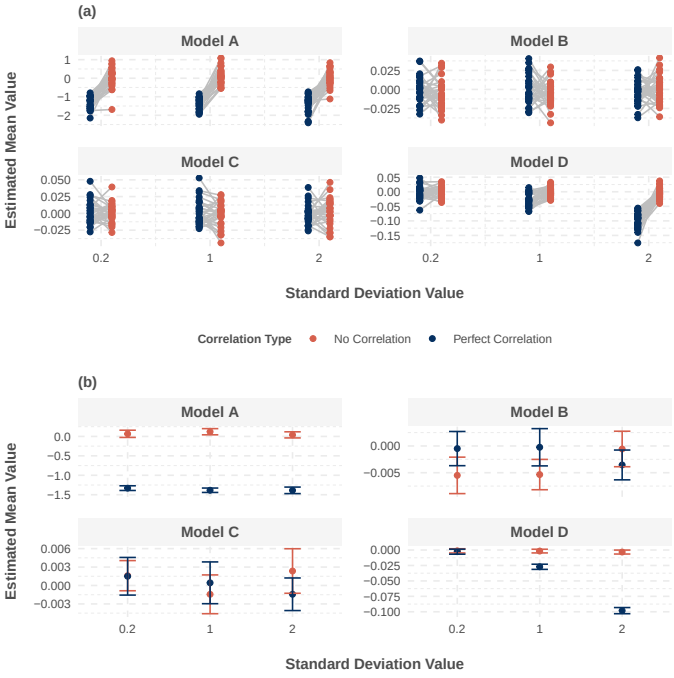


Fig. 3. (a) Scatterplot of estimated values for mean and (b) Scatterplot of the mean of estimated values for mean with error bars across simulations with respect to the value of standard deviation, faceted by four models and colored by correlation type. In subfigure (a), the grey lines connect observations from the same simulation.

The results of simulations estimating the population mean with respect to three parameters show that the implemented Bayesian models are not always robust to violations of i.i.d in data. Despite the similarity of Models A and B, the mean and the standard deviation are estimated individually in Model B, not affecting each other, and Model A introduces individual standard error directly to mean estimation, which makes Model A a reliable example of a non-robust model to assumption violation.

Correlation

For the visual analysis of simulation results for each of the four parameters, two variants of plots were generated: (i) scatterplot of estimate values in each simulation, and (ii) scatterplot with error bars of the mean value of estimate across the simulations faceted by parameter values and colored by models.

Number of Points: The results of simulations with respect to the number of points (Figure 4) show that with the increase in the number of points, results across models and simulations stabilize and are closely centered around the initial correlation coefficient input for data generation. Furthermore, with the increase in the number of points, the mean estimate values across simulations with respect to models more strictly show which models tend to overestimate the correlation coefficient. Specifically, Model 3 with Scenario 6 (total dependency on the product of both dataX and dataY) gives the most different estimate of correlation.

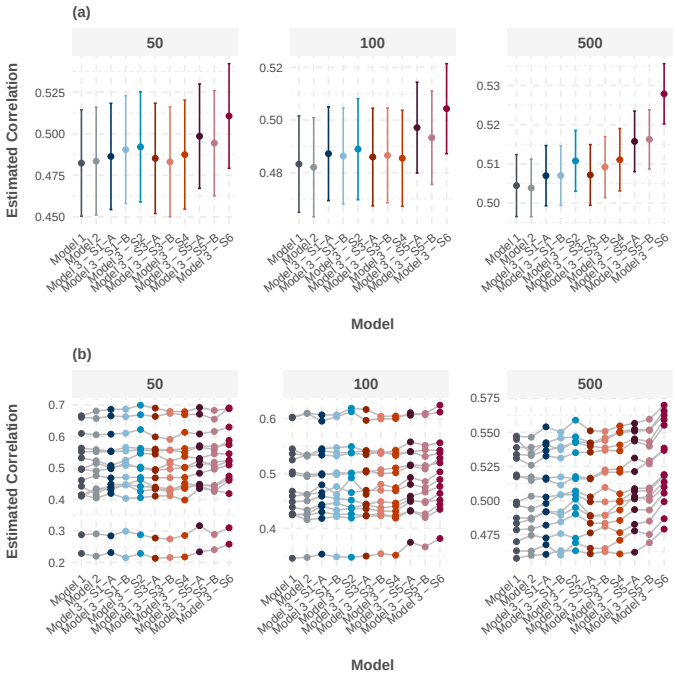


Fig. 4. (a) Scatterplot of the mean of estimated values for correlation with error bars and (b) Scatterplot of estimated values for correlation across simulations with respect to the number of points, faceted by parameter value and colored by model. In subfigure (b), the grey lines connect observations from the same simulation.

Values of lambdas for marginal variances: The results of simulations with respect to the values of lambdas for marginal variances (Figure 5) show that when marginal variances are equal, the values of estimates across simulations in Model 3 are relatively close for dependency-wise paired scenarios. In contrast, in the cases when either of the marginal variances is greater, the estimates by paired scenarios vary in favor of the greater variance.

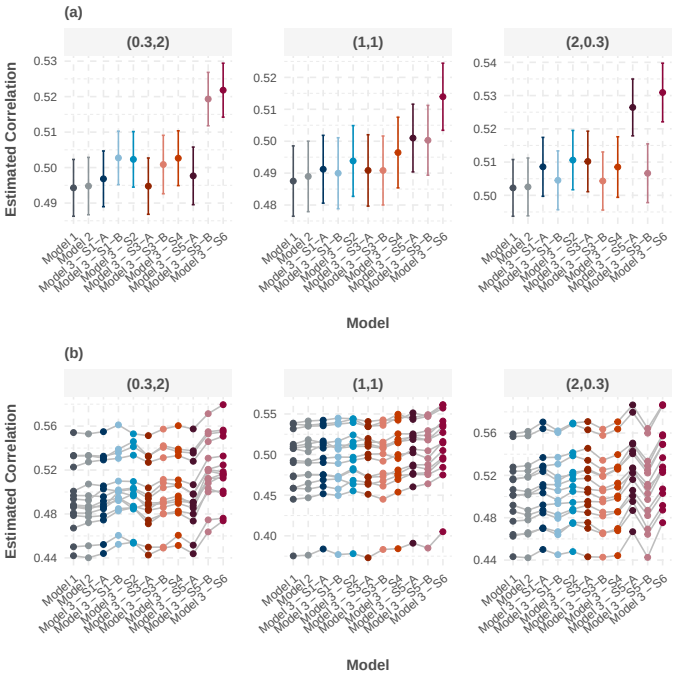


Fig. 5. (a) Scatterplot of the mean of estimated values for correlation with error bars and (b) Scatterplot of estimated values for correlation across simulations with respect to the values of lambdas for marginal variances, faceted by parameter value and colored by model. In subfigure (b), the grey lines connect observations from the same simulation.

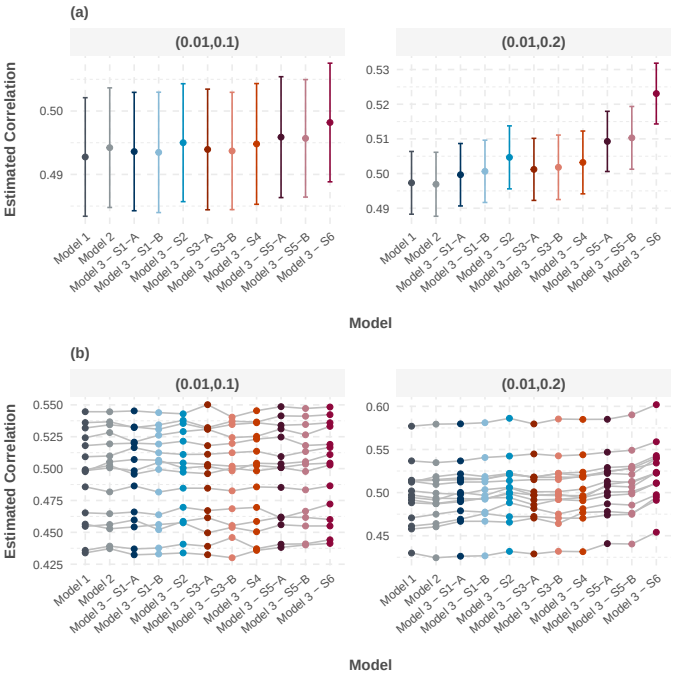


Fig. 6. (a) Scatterplot of the mean of estimated values for correlation with error bars and (b) Scatterplot of estimated values for correlation across simulations with respect to the range of standard deviation in uniform distribution for error measurement, faceted by parameter value and colored by model. In subfigure (b), the grey lines connect observations from the same simulation.

The distribution range of standard deviation for measurement error dependency: The results of simulations with respect to the range of standard deviation in uniform distribution for error measurement in Figure 6 show that with a broader range of uniform distribution for the measurement error, the average estimate of correlation across the models leads to overestimation compared to the narrower one, where average values not only underestimate the correlation between parameters but are densely distributed.

Correlation coefficient in data generation: Finally, the results of simulations with respect to the correlation coefficient in data generation in Figure 7 show that with a greater initial value of correlation coefficient in data generation, the average estimates of correlation across dependency-wise paired scenarios for Model 3 tend to get closer to each other. Moreover, with a greater correlation coefficient, estimated values across simulations become relatively more stable.

The results of simulations estimating the correlation between two variables with respect to the four parameters show that the Bayesian models implemented with and without the consideration of marginal variances of variables yield varying results and are affected by the parameter changes in data formulation. This part of the analysis shows that the parameter used for assessing the robustness of models to the violation of i.i.d - correlation, on its own, is subject to a mismatch in the data vs. model formulation, causing an underlying violation of i.i.d for individual variables.

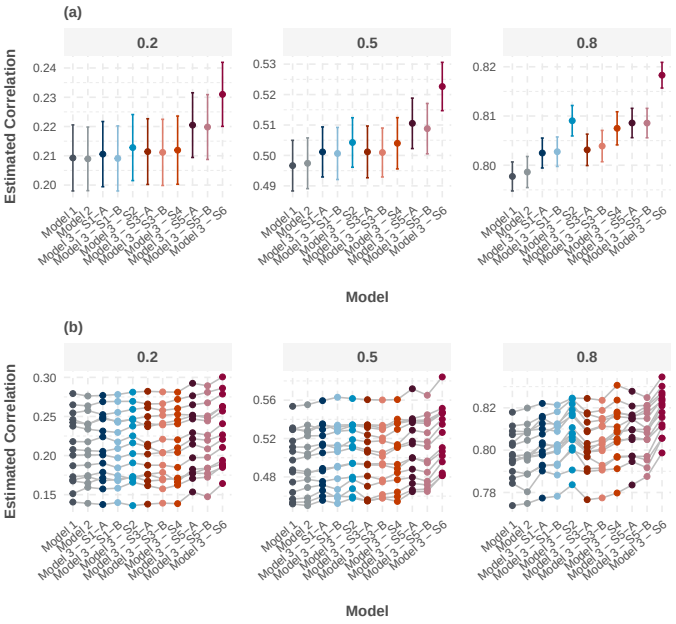


Fig. 7. (a) Scatterplot of the mean of estimated values for correlation with error bars and (b) Scatterplot of estimated values for correlation across simulations with respect to the value of correlation coefficient, faceted by parameter value and colored by model. In subfigure (b), the grey lines connect observations from the same simulation.

Conclusion

Statistical modeling goes hand in hand with formulating and adjusting data-related assumptions. Working with Bayesian models provides the flexibility of assumption adjustment, the misuse of which can result in assumption violation, affecting the reliability of obtained results.

The present research paper discussed the effects of violating i.i.d in Bayesian statistical modeling for the two types of parameter estimation. It showed that the results of the simulation could be affected by the violation of i.i.d. depending on how Bayesian statistical models were formulated. The models were formulated to estimate the mean and the correlation under different assumptions regarding the data. Based on the obtained results, I found that some models are more susceptible to the violation of i.i.d. While the effects across models might have been limited, the contrast of results was enough to doubt the robustness of models. The cause of such behavior was the mismatch of assumptions used in model formulations, which were also subsequently violated.

The verdict is simple: The quest for data analysts is to be able to identify the underlying assumptions to mitigate the risks of inaccurate model selection in a thorough process of data scrutinization before performing the analysis.

Appendix

Mean

Figures A1-A4 show the graphical representations of four Bayesian statistical models for mean estimation, formulated by revising the models from Lee & Wagenmakers (2013, p. 54 - 59). Note that, to avoid introducing extra bias to models through individually adjusted priors for the simulations, "uninformative" priors with uniform distribution were used.

Figure A1: Graphical model for the structure of Model-A. The parameters mu_i and sigma_i, shaded in orange, refer to the data input to the model. The single-bordered circles represent variables sampled from respective priors.

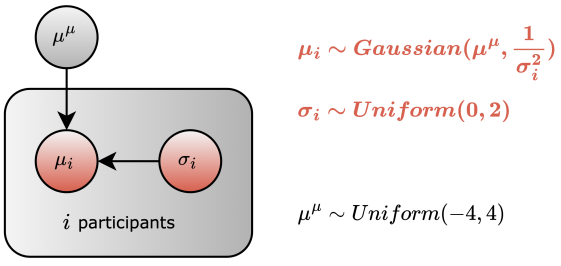


Figure A2: Graphical model for the structure of Model-B. The parameters mu_i and sigma_i, shaded in orange, refer to the data input to the model. The single-bordered circles represent variables sampled from respective priors.

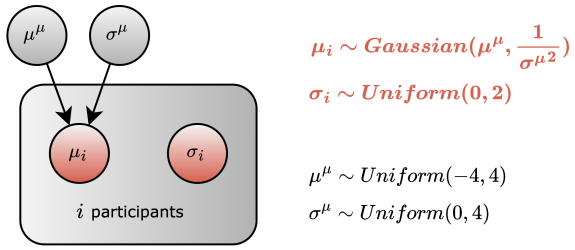


Figure A3: Graphical model for the structure of Model-C. The parameter x_ij, shaded in orange, refers to the data input to the model. The single-bordered circles represent variables sampled from respective priors.

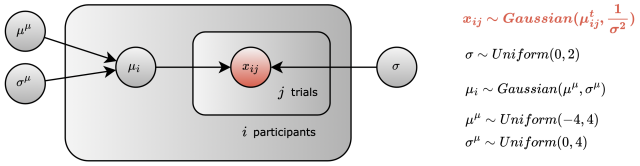
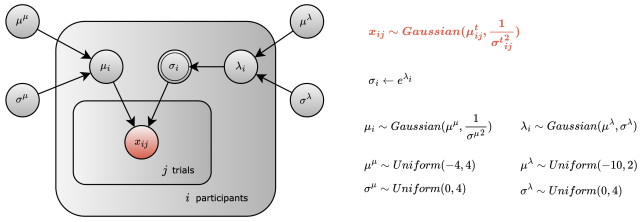


Figure A4: Graphical model for the structure of Model-D. The parameter x_ij, shaded in orange, refers to the data input to the model. The single-bordered circles represent variables sampled from respective priors, and the double-bordered circles represent variables computed from other variables.



Correlation

Figures A5-A7 show the graphical representations of three Bayesian statistical models for correlation estimation, formulated by revising the models from Lee & Wagenmakers (2013, p. 60 - 63).

Figure A5: Graphical model for the structure of Model-1. The parameter x_i, shaded in orange, refers to the data input to the model. The single-bordered circles represent variables sampled from respective priors. Bold variables indicate a pair of variables.

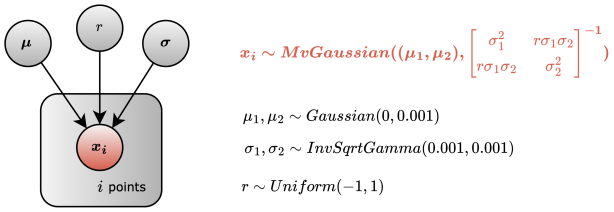


Figure A6: Graphical model for the structure of Model-2. The parameters x_i and lambda_error, shaded in orange, refer to the data input to the model. The single-bordered circles represent variables sampled from respective priors. Bold variables indicate a pair of variables.

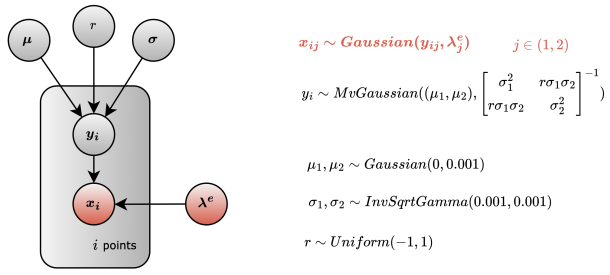
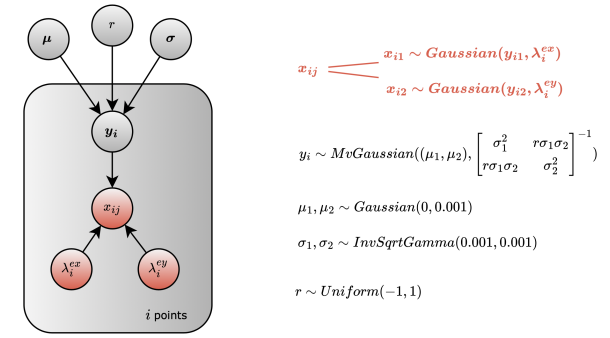


Figure A7: Graphical model for the structure of Model-3. The parameters x_i, lambdaerror_x, and lambdaerror_y, shaded in orange, refer to the data input to the model. The single-bordered circles represent variables sampled from respective priors. Bold variables indicate a pair of variables.



References

Bergh, D. van den, E.-J. Wagenmakers, and F. Aust. 2023. "Bayesian Repeated-Measures Analysis of Variance: An Updated Methodology Implemented in Jasp." *Advances in Methods and Practices in Psychological Science* 6. https://doi.org/10.1177/25152459231168024.

Eddy, S. 2004. "What Is Bayesian Statistics?" *Nat Biotechnol* 22: 1177–78. https://doi.org/10.1038/nbt0904-1177.

Gottardo, R., and A. Raftery. 2009. "Bayesian Robust Transformation and Variable Selection: A Unified Approach." *Canadian Journal of Statistics* 37(3): 361–80. https://doi.org/10.1002/cjs.10021.

Kasuya, E. 2001. "Mann-Whitney u Test When Variances Are Unequal." *Animal Behaviour* 61(6): 1247–49. https://doi.org/10.1006/anbe.2001.1691.

Khan, A., and G. D. Rayner. 2003. "Robustness to Nonnormality of Common Tests for the Many-Sample Location Problem." *Journal of Applied Mathematics and Decision Sciences* 7(4).

Lee, M. D., and E.-J. Wagenmakers. 2013. "Bayesian Cognitive Modeling: A Practical Course." *Cambridge University Press*, 54–63. https://doi.org/10.1017/CBO9781139087759.

Plummer, M. 2003. "Jags: A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling." *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003), Vienna*.

Ruscio, J., and B. Roche. 2012. "Variance Heterogeneity in Published Psychological Research: A Review and a New Index." *Methodology* 8(1): 1–11. https://doi.org/10.1027/1614-2241/ a000034.

Sawada, T. 2023. "Effects of Violating the Assumptions of Equal Variance and Independent and Identically Distributed Random Variables: A Case Using Bayesian Statistical Modeling." *The Quantitative Methods for Psychology* 19(3): 281–95. https://doi.org/10.20982/tqmp.19.3.p281.