Simulation Studies

Christopher Schmid

Fall 2025

Simulation Studies

Definition Simulation studies are computer experiments that involve creating data by pseudo-random sampling from known probability distributions (Morris, White and Crowther, 2019)

- Evaluate and compare performance of alternative (new) statistical methods under different operating conditions
- Investigate a complex stochastic process (microsimulation)
- Evaluate type 1 and 2 error rates under different assumptions for given sample sizes
- Calculate sample size and power when designing a study

Other Uses of Simulation

- Estimate Bayesian posterior distributions (Markov chain Monte Carlo)
- Perform Monte Carlo integration
- Perform posterior predictive checking to determine how well model predicts actual data
- Get statistical properties of models in situations where standard techniques may not work well (e.g., standard errors via bootstrap)
- Examine properties of complex probabilistic models

ADEMP Framework for Simulation Study Protocols

- **A**ims and objectives
- **D**ata generating mechanisms
- **E**stimands/targets of analysis
- Statistical **M**ethods to evaluate
- **P**erformance measures

Careful design reduces need to redo simulations

Aims

- Purpose of simulation
- Properties of estimators or methods
- Show method works or does not work
- Show where method might fail

Example: Evaluate impacts of different models for survival

- misspecifying baseline hazard function in a survival model;
- fitting too complex a model;
- using semiparametric model

Data Generating Mechanisms

- Describes how random numbers used to generate data to produce different testing scenarios
- Can generate from parametric model or nonparametrically by resampling (e.g., by bootstrap)
- Often choose multiple data generating mechanisms
- Choosing mechanism that derives from particular method biases comparison

Data Generating Mechanisms: Choosing Factors to Vary

- Factors to vary
  - Sample size
  - True parameter values
  - Amount of missing data
- Ways to vary factors
  - One factor at a time holding others fixed
  - Factorial design
  - Partially factorial design

Data Generating Mechanisms Continued

- Scenarios should reflect most common circumstances and cover range of plausible parameter values if possible
- Number of scenarios and statistical methods to investigate depend on study objectives but may be constrained by
  - Available time (may want to proceed sequentially)
  - Programming efficiency
  - Availability of parallel computing
  - Complexity of presentation
  - Include interactions?
- Number of simulations depends on accuracy desired and resources available

Write protocol

Methods for generating datasets

- Carefully consider and fully describe methods by which data generated
- Requires assumed distribution for data and full specification of required parameters
- Need to resemble reality for results to be generalizable to real situations and have credibility
- Can use real data set as motivating example so data closely represent its structure
- Could use actual covariate data or covariate correlation structure and then generate outcome data
- Provide rationale for choices about data distributions, model parameters and correlation structure
- Verify generated data to ensure they resemble intended data structure

Generating multivariate data

- Specify correlations between covariates and outcomes
- Specification of means and covariance matrix is more straightforward if based on real data, especially with a large number of covariates, and generated data will reflect reality
- If chosen arbitrarily, need to make sure valid distributions (e.g., positive definite covariance matrices)
- Often assume multivariate normal distribution for simplicity
- If using real data, any continuous but non-normally distributed variables should be transformed to make assumption of normality more appropriate
- Binary variables can be generated as latent normal, i.e. generated as continuous variables and then dichotomized
    - Need to adjust correlation structure to be correct for binary variable

Generating time to event data

- Must specify model for multivariate covariate data and distribution for survival data, which may be censored
- To simulate censored survival data, need two survival distributions: one for uncensored survival times that would be observed if follow-up had been sufficiently long to reach event and another representing censoring mechanism
- Empirical survival distribution from a similar real data set would provide a reasonable choice for survival distribution
- Uncensored survival distribution could be generated to depend on a set of covariates with a specified relationship with survival, which represents the true prognostic importance of each covariate

Generating time to event data continued

- Time-dependent covariates could also be simulated and incorporated
- Random non-informative right censoring with a specified proportion of censored observations can be generated in a similar manner to uncensored survival times by assuming a particular distribution for censoring times
- Censoring mechanism can be extended to incorporate dependent, informative censoring
- If uncensored survival time for a case is less than or equal to censored time, then event is considered to be observed and survival time is uncensored survival time; otherwise, event is considered censored and survival time equals censored time

Estimands and Targets

- Often a parameter of interest, e.g. mean, variance, regression coefficient
- Could be a method for testing a null hypothesis
- Model selection procedure
- Prediction
- Performance measure such as AUC

Statistical Methods

- Could be model for analysis, design or decision rule
- Evaluation of one method or comparison of several
- Comparators may include new methods, standard methods
- Need code to implement older methods
- Ensure methods address same estimands
- Method might involve a procedure which applies sequential methods

Performance Measures

- Measures to assess performance depend on aims and study targets
- Compare simulated results with true values used to simulate data
- Estimand targets will focus on frequentist properties of estimation such as bias and coverage
- If target is null hypothesis, power and type 1 error of primary interest
- Examine multiple measures as results may vary across criteria
- Trade-off between bias and variance
- Estimate uncertainty with which each is estimated

Simulation Failures

- Sometimes simulation fails (e.g., rare events or missing data may lead to lack of convergence)
- If allowance not made, then long program may break
- If failed simulations discarded, comparisons may be biased (e.g., one method may do better than another, but may sometimes fail) or may lead to inaccurate estimate of precision
- May repeat simulation to replace failure
- Should record failed simulations as performance measure
- May require post hoc change of protocol to omit scenarios that frequently fail
- Sometimes failure may lead to improved or at least revised algorithm that may mimic practice

Bias

- Bias is deviation of estimate from true quantity

$$
\begin{aligned}
&\bar{\hat{\beta}} - \beta && \text{bias} \\
&|\bar{\hat{\beta}} - \beta| && \text{absolute value of bias} \\
&100 * \left(\frac{\bar{\hat{\beta}} - \beta}{\beta} \right) && \text{percentage bias} \\
&100 * \left(\frac{\bar{\hat{\beta}} - \beta}{SE(\hat{\beta})} \right) && \text{standardized percentage bias}
\end{aligned} \tag{1}
$$

- Large bias can adversely impact efficiency, coverage and error rates
- Testing significance of amount of bias depends on number of simulations and so leads to statistical, but not practical significance

Mean Squared error

- Provides useful measure of overall accuracy incorporating both bias and variability
$$E[(\hat{\theta} - \theta)^2]$$
- Can be written as
$$(\bar{\hat{\theta}} - \theta)^2 + Var(\hat{\theta})$$
i.e., squared bias + variance
- Square root transforms it back onto same scale as parameter
- Relative contributions of bias and variance can depend on sample size so evaluation of MSE should be performed with different sample sizes

Precision

Can estimate either as

- Empirical standard error $\sqrt{Var(\hat{\theta})}$
- Average model standard error $\sqrt{E[\hat{Var}(\hat{\theta})]}$
- Empirical standard error measures precision of θ
- Does not require knowledge of θ
- Precision should be interpreted carefully in presence of bias because estimates biased toward null will be more precisely estimated
- Average model standard error estimates empirical SE
- Differences between two indicate bias in estimate of SE
- Can use relative % error in model SE

$$100(\frac{ModSE}{EmpSE} - 1)$$

Coverage

- Coverage is proportion of times that confidence interval contains true specified parameter value
- Should be approximately equal to nominal coverage rate, e.g., 95 percent of samples for 95 percent confidence intervals, to properly control type I error rate for testing null hypothesis of no effect

Undercoverage

Under-coverage, where coverage rates are lower than 95 per cent, indicates over-confidence in estimates since more simulations will incorrectly detect a significant result, which leads to higher than expected type I errors

Occurs if

- bias present
- Mod SE $<$ Emp SE
- distribution of $\hat{\theta}$ non-normal and normality assumed
- $Var(\hat{\theta}_i)$ too variable

Overcoverage

Over-coverage, where coverage rates are above 95 per cent, suggests that results are too conservative (i.e., fail to reject null hypothesis and too high type II error rate)

- Occurs if Mod SE $>$ Emp SE

Coverage

- Possible criterion for acceptability of coverage is that it should not fall outside of approximately two SEs of nominal coverage probability p

$$SE(p) = \sqrt{p(1-p)/n_{sim}}$$

- For example, if 95% CI with $n_{sim} = 1000$, $SE(\hat{p}) = 0.0069$ and hence empirical coverage between 93.6% and 96.4% support appropriate coverage
- Average length of CI for parameter estimates often used

$$\frac{\sum_{i=1}^{n_{sim}} 2Z_{1-\alpha/2}SE(\hat{\beta}_i)}{n_{sim}}$$

- If parameter estimates relatively unbiased, then narrower confidence intervals imply more precise estimates, suggesting gains in efficiency and power

Power and Type 1 error

- Empirical power of a test is proportion of simulation samples in which null hypothesis is rejected at nominal significance level when null hypothesis is false
- Empirical type I error is proportion of simulation samples in which null hypothesis is rejected at nominal significance level when null hypothesis is true

Types of Datasets in Simulation Study

- Simulated data
- Random number generator states after each simulation
  - Simulation can be restarted if more runs needed
  - Simulation can be restarted where it breaks (e.g., non-convergence)
  - Can be reproduced by others
- Summaries of estimates across simulations
- Performance measures

Random number generation

- Fundamental requirement of any simulation are good random number generators
- Use random number generators with long sequence before repetition and for which subsets of the random number sequence are independent
- A random number generator must be able to reproduce identical set of random numbers when same starting value, known as a seed, is specified
  - Enables reproducibility for checking accuracy of program
  - Using same seed can make data sets partly dependent
- Random numbers for parallel independent simulations can be generated by setting different starting values for individual simulations that are greater than number of random numbers required for each simulation

Random number generation continued

- In general, seed determined by computer clock so if simulations are run in series, then seeds will give independent sequences
- Make sure default seed not being used by programs
- Can check by displaying current state of RNG and run program twice
- If initial state and state after first run are same, then program does not use random numbers
- If initial and first run are different, but first run and second run are same, then program is resetting seed

Dependence of Simulated Datasets

- For a given scenario and model/method, independent set of K simulations generated
- To compare methods in one scenario, usually want to use same set of K independent simulations in order to match on simulated data, thus eliminating within-sample variability
- For different scenarios, use independent data

Calculating Summary Measures

- Estimates include means, slopes, hazard ratios, odds ratios, etc.
- Also need within simulation standard error (SE) for estimates
- Summaries usually include the average estimate across simulations as well as the standard deviation
- Standard deviation of simulation estimates should be close to average within-simulation standard errors, if estimate is unbiased
- Can also summarize by quantiles of simulated values

Storing Estimates

- Need to store estimates to allow for
  - Error checking
  - Exploration of outliers, trends and patterns within individual simulations that may not be observed from summary measures alone
  - Revised summarization retrospectively without need to repeat all simulations
- Thorough consideration in design stage can ensure that all required estimates are included, analyzed and results stored

Software Issues

- Build code up in chunks
- Use built-in checks
- May want to simulate large dataset to make sure that method is working properly (i.e., avoid small sample variability)
- If methods to be compared implemented in different software packages (e.g. R and Stata) better to generate data in one package and port over to other so that variability from RNG is controlled

Choosing number of simulations

Two criteria as for any sample size calculation

- Accuracy desired (achieve certain Monte Carlo error)
  - Need variance of performance measure, σ^2 e.g., for measure which is proportion like coverage use $p(1-p)/n_{sim}$ where p is desired coverage level
  - Combined with desired accuracy δ can compute

$$N_{sim} = Z_{1-\alpha/2}\sigma^2/\delta^2$$

- Power to detect specific difference from true value as significant at level α with power $1-\beta$

$$N_{sim} = [Z_{1-\alpha/2} + Z_{1-\beta}]\sigma^2/\delta^2$$

Presentation of Results

- Tabular or graphical form
- Dimensions are methods, data generating mechanisms and performance measures
- Monte Carlo SEs should be reported either next to estimates or as maximum
- Can display Monte Carlo SE in graph as confidence interval

Simulation as an Experimental Design

- A simulation is a computerized experiment
- Outcome is performance of estimands in different models
- Parameter settings are factors
- Simulations provide data
- Objective is to evaluate and perhaps optimize performance under different parameter settings

Questions of Interest

- Which parameter settings affect which estimands on which performance measures?
- What is form of relationship? (nonlinearity)
- Does performance vary by combinations of parameters? (interaction)
- Where is performance optimized? (response surface)

Sequential design of experiments

- George Box described a sequential process for doing experiments
- Screening design to detect important factors out of many
  - Uses blocking along with factorial and fractional factorial designs
- Response surface designs to zero in on exact nature of effects
  - Central composite, Optimal designs
- In all designs, settings are chosen scientifically to optimize amount of information gained from experiment

Simulation Parameters in Jackson et al. (2018)

Setting	k	τ^2	Treatment	Control	Baseline probability	Correct models
1	10	0.024	$N \sim U(50, 500)$	N	$LO_c \sim N(\text{logit}(0.2), 0.3^2)$	2, 3, 6
2	10	**0**	$N \sim U(50, 500)$	N	$LO_c \sim N(\text{logit}(0.2), 0.3^2)$	2, 3, 6
3	10	**0.168**	$N \sim U(50, 500)$	N	$LO_c \sim N(\text{logit}(0.2), 0.3^2)$	2, 3, 6
4	**3**	0.024	$N \sim U(50, 500)$	N	$LO_c \sim N(\text{logit}(0.2), 0.3^2)$	2, 3, 6
5	**5**	0.024	$N \sim U(50, 500)$	N	$LO_c \sim N(\text{logit}(0.2), 0.3^2)$	2, 3, 6
6	**20**	0.024	$N \sim U(50, 500)$	N	$LO_c \sim N(\text{logit}(0.2), 0.3^2)$	2, 3, 6
7	10	0.024	$N \sim U(\mathbf{10}, \mathbf{100})$	N	$LO_c \sim N(\text{logit}(0.2), 0.3^2)$	2, 3, 6
8	10	0.024	$N \sim U(50, 500)$	N	$LO_c \sim N(\text{logit}(\mathbf{0.05}), 0.3^2)$	2, 3, 6
9	10	0.024	$N \sim U(50, 500)$	N	$LO_c \sim N(\text{logit}(\mathbf{0.01}), 0.3^2)$	2, 3, 6
10	10	0.024	$N \sim U(50, 500)$	**N/2**	$LO_c \sim N(\text{logit}(0.2), 0.3^2)$	2, 3, 6
11	10	0.024	$N \sim U(50, 500)$	**N/2 and N**	$LO_c \sim N(\text{logit}(0.2), 0.3^2)$	2, 3, 6
12	10	0.024	$N \sim U(50, 500)$	**N/2 and N (NR)**	$LO_c \sim N(\text{logit}(0.2), 0.3^2)$	None
13	10	0.024	$N \sim U(50, 500)$	N	$\mathbf{P}_c \sim \mathbf{U(0.1,0.3)}$	2
14	10	0.024	$N \sim U(50, 500)$	N	$\mathbf{LO}_a \sim N(\text{logit}(0.2), 0.3^2)$	4, 5, 6
15	10	**2**	$N \sim U(50, 500)$	N	$LO_c \sim N(\text{logit}(\mathbf{0.5}), 0.3^2)$	2, 3, 6

- Parameters changed one at a time
- Only examines linear effects
- Cannot detect interactions

Fixed vs Random Intercept IPD Models

- Long-running controversy about whether to use a fixed or random intercept when fitting IPD models (e.g., when fitting an exact model for a discrete outcome)

$$Y_{ij} = \mu_i + \delta_i Z_{ij} + \epsilon_{ij}$$
$$\delta_i \sim N(d, \sigma^2_\delta)$$
$$\epsilon_{ij} \sim N(0, \sigma^2_\epsilon)$$

- $Z_{ij} = 1$ if treated, 0 otherwise
- i indexes study, j indexes individuals
- Intercept μ_i can be treated as fixed constant or as random

$$\mu_i \sim N(m, \sigma^2_\mu)$$

- Slope and intercept may be correlated with correlation ρ_{md}

Potential Bias from Fixed vs Random Intercepts

- One fixed intercept for each study so number of parameters increases with number of studies
- Violates conditions for asymptotic properties of MLE
- But random intercepts borrow from each other which may bias treatment effect estimation
- Some previous work suggests that between-study variance may be affected
- What happens under Bayesian estimation?

ADEMP for Our Problem

Aim Compare statistical properties of parameters in fixed vs random intercept Bayesian meta-analysis models with IPD

Data generating mechanism Continuous outcomes from studies with separate treatment effects and intercepts under varying numbers of studies and study sizes and varying true values of model parameters

Estimands Model parameters
(primary) d, σ_δ
(secondary) $m, \sigma_\mu, \sigma_\epsilon,\ \rho_{md}$

Models Fixed and random intercept models with and without centering, with and without correlation

Performance Measures Bias, coverage, width of 95% credible interval, root mean squared error

Models

Intercept	ρ_{md}	Treatment coding	Prior on $\sigma_\mu, \sigma_\delta, \sigma_\epsilon$
Fixed	–	0,1	Uniform; Half-Cauchy
Fixed	–	Study-specific centering	Uniform; Half-Cauchy
Random	No	0,1	Uniform; Half-Cauchy
Random	No	Study-specific centering	Uniform; Half-Cauchy
Random	Yes	0,1	Uniform; Half-Cauchy

Data Generating Mechanism: Model Parameter Settings

$$r_\mu = \frac{\sigma_\mu}{\sqrt{N}}$$

$$r_\delta = \frac{\sigma_\delta}{\sqrt{N}}$$

$$r_\epsilon = \frac{\sigma_\epsilon}{\sqrt{L}}$$

N number of studies L number of individuals in study

Data Generating Mechanism: Model Parameter Settings

$$r_\mu = \frac{\sigma_\mu}{\sqrt{N}}, r_\delta = \frac{\sigma_\delta}{\sqrt{N}}, r_\epsilon = \frac{\sigma_\epsilon}{\sqrt{L}}$$

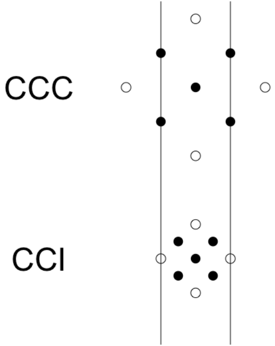
Parameters	Values
r_μ	0.1, 0.8, 1.5, 2.2, 3
r_δ	0.1, 0.5, 1, 1.5, 2
r_ϵ	0.01, 0.2, 0.5, 0.7, 1
ρ_{md}	-0.9, -0.5, 0, 0.5, 0.9
Number of Studies (N)	9, 25, 49, 81, 121
Study Sizes (L)	U(30,100)
	U(30,1000)
	U(900,1000)
	0.5*U(30,100) + 0.5*U(900,1000)
Allocation ratio	1:1 for all studies
d, m	0

Our project

- Parameters at many settings
- Many combinations of factors
- Many replicates of simulations at each of many settings running Bayesian model MCMC
- How to find appropriate response surface to evaluate model performance
- Computing nightmare!

Central Composite Design

- Three components: fractional factorial, star points and center points



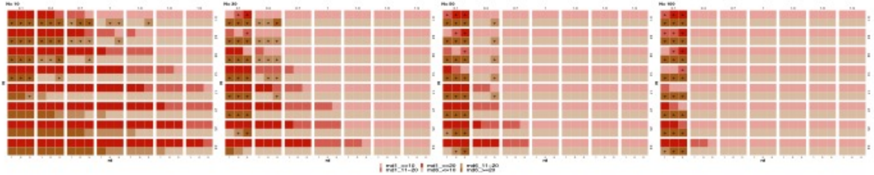
- Factorial points used to estimate main effects and 2-way interactions
- Star points used to estimate quadratic terms
- Center points provide replication and estimation of nonlinearity

Central Composite Design

- r_μ, r_δ, r_ϵ, N, ρ_{md} are quantitative factors of five different levels (1 center, 2 factorial, 2 star)
- Generate 2^5 runs using factorial levels of each quantitative factor
- L is a categorical factor with four levels used to generate 4 blocks
- Blocks confounded with combinations of quantitative factors so that blocks not confounded with second-order interactions
- Star block contains all star points of quantitative factors replicated 8 times each
- Simulations contains 122 runs (32 factorial points + 80 star points + 10 center points)
- Can fit quadratic model and use ANOVA to determine predictive model that identifies effect of different parameters

Bias for Between-Study Variance

- Red: fixed intercept model
- Brown: random intercept model
- N = 10, 30, 50, 100 are blocks
- Ratio of within-study variance to study size in rows (r_ϵ)
- Ratio of between-study variance to no. studies in columns (r_δ)
- Correlations within small blocks (ρ_{md})



Items Often Inadequately Presented

- Specifics of random number generator and choice of starting seeds
- Software package for generating data and for analysis
- Relationship between generated samples
- Number of simulations
  - Varies widely
  - Sometimes not reported
  - Rarely justified
- Source of generated data (real or typical)
- Incomplete reporting of results (use appendices)