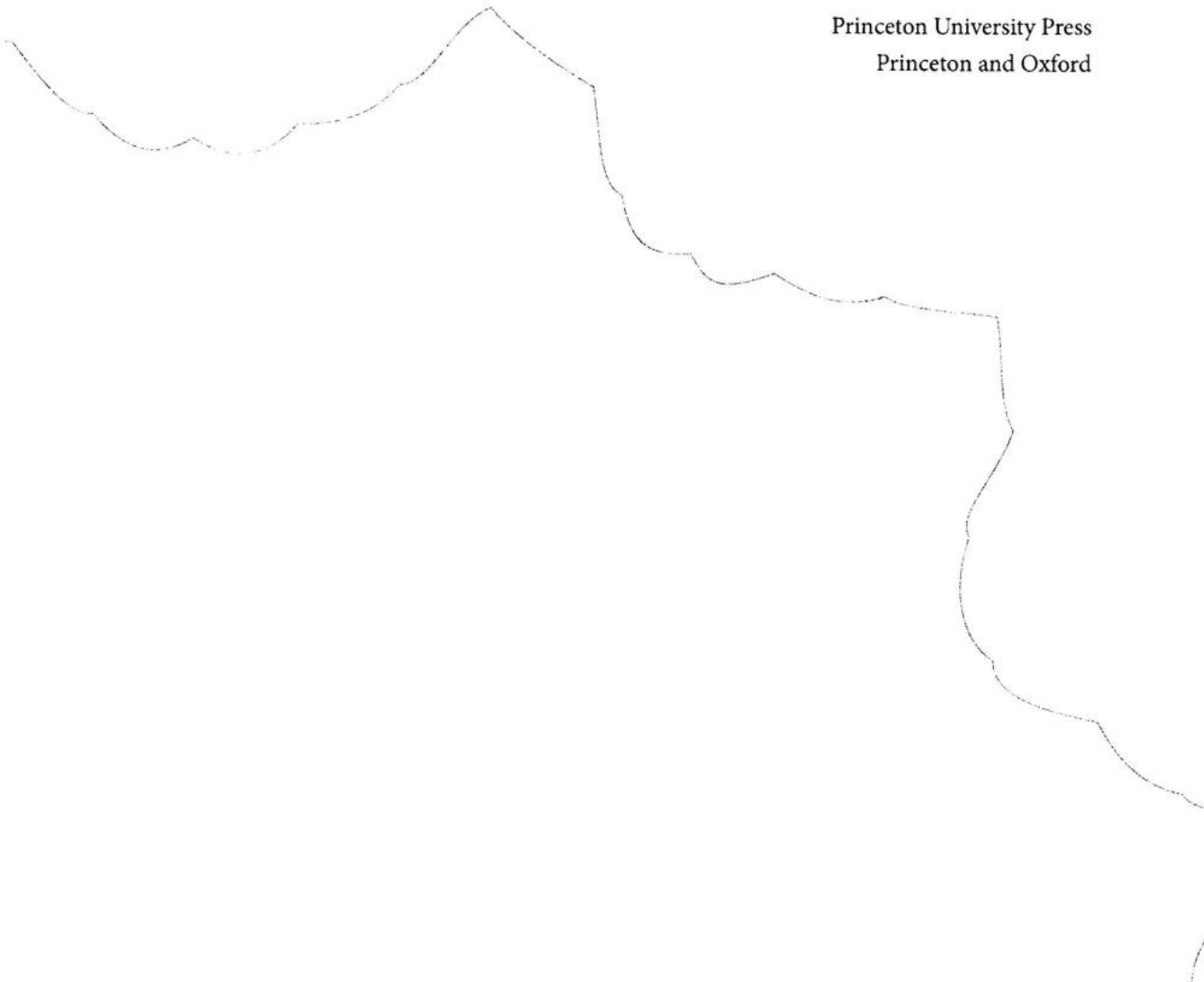
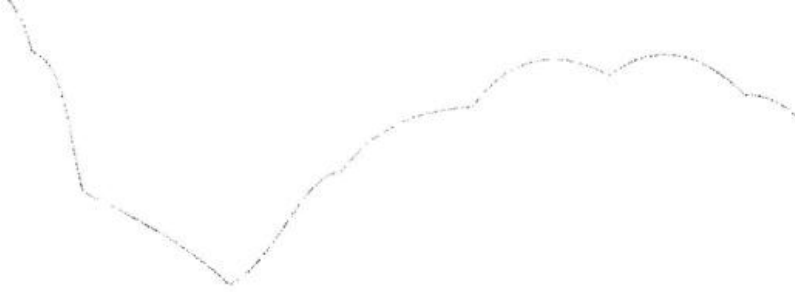


# STATISTICS BY SIMULATION

**Carsten F. Dormann and  
Aaron M. Ellison**

Princeton University Press  
Princeton and Oxford





# Table of contents

Preface xi  
 Acknowledgments xv

## Part I

**Propositi: Why and how to simulate . . . . . 1**

**1 General Introduction . . . . . 3**

1.1 What are simulated data? . . . . . 4

1.2 Simulated data are specific . . . . . 5

1.3 Yes, scientists really simulate data . . . . . 6

1.4 There are many good reasons to simulate data . . . . . 9

1.5 Useful background knowledge to use this book most effectively . . . . 10

1.6 Notational conventions . . . . . 11

1.7 Structure, organisation, and flow . . . . . 12

1.8 Summary . . . . . 13

**2 The basics of simulating data and the need for computational competence . . . . . 17**

2.1 A road map for simulation in statistics . . . . . 17

2.2 Two simple examples . . . . . 18

2.3 More complex examples . . . . . 21

2.4 Simulating autocorrelated data . . . . . 33

2.5 Simulation versus randomisation techniques . . . . . 35

2.6 Summary . . . . . 43

## Part II

**Ante mensuram: Prospective simulations of study designs and their power . . . . . 45**

**3 Think before you act . . . . . 47**

3.1 The illusion of truth: A case study . . . . . 47

3.2	The question comes first . . . . .	48
3.3	Setting expectations, defining hypotheses . . . . .	49
3.4	Testing hypotheses and assessing their support . . . . .	55
3.5	Pre-registration . . . . .	67
3.6	Summary . . . . .	69
<b>4</b>	<b>Prospective simulation of statistical power . . . . .</b>	<b>73</b>
4.1	Simple group comparisons . . . . .	74
4.2	How many data points do we need for a simple correlation? . . . . .	76
4.3	Is “recruit until significant” problematic? . . . . .	80
4.4	How long does a time series have to be? . . . . .	85
4.5	Improving estimates: Is the experiment powerful enough? . . . . .	91
4.6	Summary . . . . .	101
<b>Part III</b>		
<b>Post mensuram: Simulations in statistical analysis . . . . .</b>		<b>103</b>
<b>5</b>	<b>Assumptions: Is that one important? . . . . .</b>	<b>105</b>
5.1	Linear regression requires the data to be normally distributed . . . . .	106
5.2	Regression models also assume that errors in predictor variables are negligible or unimportant . . . . .	110
5.3	The intended, rather than the realised, manipulation is an admissible predictor variable . . . . .	112
5.4	ANOVA requires homoscedasticity . . . . .	116
5.5	Multiple testing and the inflation of false positives . . . . .	123
5.6	Hyper-distributions in mixed-effect models are normal . . . . .	131
5.7	Correlations among predictors are the same outside the range of the observed data . . . . .	137
5.8	Summary . . . . .	145
<b>6</b>	<b>Folklore: Is that rule-of-thumb true or useful? . . . . .</b>	<b>153</b>
6.1	Model selection does not always improve interpretation . . . . .	154
6.2	Selecting one of two correlated predictors does not mitigate collinearity in regression and machine learning . . . . .	165
6.3	It is not OK to categorise continuous predictor variables . . . . .	172
6.4	Use Monte Carlo simulation when data are heteroscedastic . . . . .	180
6.5	Time series should not be detrended by default . . . . .	190
6.6	Machine learning and Big Data do not obviate rules-of-thumb . . . . .	200
6.7	Summary . . . . .	204
<b>7</b>	<b>Workflows and pipelines can introduce and propagate artefacts . . . . .</b>	<b>211</b>
7.1	What can we do about missing data? . . . . .	212
7.2	Types of missing data . . . . .	212

7.3	Imputation of missing predictors . . . . .	213
7.4	Estimating values for censored observations . . . . .	222
7.5	Pre-selecting predictors . . . . .	232
7.6	Regression on residuals . . . . .	242
7.7	Error propagation . . . . .	245
7.8	Workflow: Stringing multiple statistical steps into an analytical pipeline . . . . .	251
7.9	Summary . . . . .	259

## Part IV

### Post exemplum: Diagnostic simulations . . . . . 267

#### 8 Evaluating models: How well do they really fit? . . . . . 269

8.1	Learning from the prior . . . . .	270
8.2	What does a model tell us, and what does it not tell us? . . . . .	273
8.3	Visualising more complex effects: conditional, marginal, and partial plots . . . . .	276
8.4	Model diagnostics . . . . .	281
8.5	Predicting with confidence is not the same as confidence in prediction . . . . .	288
8.6	Iterative learning: New priors from old posteriors . . . . .	305
8.7	Outlook . . . . .	306
8.8	Summary . . . . .	307

#### 9 Post hoc alternatives to retrospective power analysis . . . . . 311

9.1	Reprise: Prospective power analysis . . . . .	312
9.2	What is retrospective power analysis? . . . . .	313
9.3	Post hoc alternatives to retrospective power analysis . . . . .	318
9.4	Summary: Most retrospective analyses should be avoided . . . . .	332
9.5	Coda: What would a Bayesian do instead? . . . . .	334

## Part V

### In posterum: Simulations for new methods . . . . . 339

#### 10 Combining studies: Meta-analysis and federated analysis . . . . . 341

10.1	Whence the data? . . . . .	341
10.2	From meta-analysis through federated analysis to complete analysis . . . . .	342
10.3	Meta-analysis . . . . .	345
10.4	Individual participant-level meta-analysis . . . . .	352
10.5	One-step federated analysis . . . . .	354
10.6	Multi-step federated analysis . . . . .	358
10.7	Complete data analysis . . . . .	360

10.8	Conclusions and outlook .....	367
10.9	Summary .....	370
<b>11</b>	<b>Putting it through its paces: Does this new method work? .....</b>	<b>375</b>
11.1	Unit testing .....	376
11.2	Dimensional analysis .....	379
11.3	Comparisons .....	380
11.4	Intellectual advancement .....	382
11.5	Intuitive understanding .....	382
11.6	Model-agnostic number of parameters: Generalised degrees of freedom .....	389
11.7	Know your limits .....	397
11.8	Summary .....	398
<b>12</b>	<b>Outroduction: How far should we push simulations? .....</b>	<b>403</b>
12.1	Stochastic weather forecasting .....	403
12.2	Infusing fake signals to test the workflow at LIGO .....	404
12.3	Virtual LIDAR scanning .....	406
12.4	Advanced simulation may be neither possible nor desirable .....	406
<b>Appendix A</b>		
<b>Useful R functions for data simulations .....</b>		<b>409</b>
A.1	Drawing random values from a distribution .....	409
A.2	Doing things repeatedly: for-loops and replicate .....	410
A.3	Shuffling, resampling, and bootstrapping: sample() .....	417
A.4	Little helpers .....	418
A.5	Dedicated simulation packages .....	421
Index	423	