

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
4	Prospective simulation of statistical power	73
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
 Part III		
	Post mensuram: Simulations in statistical analysis	103
5	Assumptions: Is that one important?	105
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
6	Folklore: Is that rule-of-thumb true or useful?	153
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
7	Workflows and pipelines can introduce and propagate artefacts	211
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
11	Putting it through its paces: Does this new method work?	375
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
12	Outroduction: How far should we push simulations?	403
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
 Appendix A		
	Useful R functions for data simulations	409
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	

1

General introduction

In 2009, a poster at the Annual Meeting of the Human Brain Mapping conference caused a stir. It showed significant brain activity in response to the subject being shown photographs of humans in emotional valence. The subject was an Arctic salmon—and it was dead (Bennett et al. 2009). In the same year, Kriegeskorte et al. (2009) highlighted a disturbing amount of circular reasoning and double-dipping in neuroscience, particularly in fMRI analyses, where the test statistics were not independent of the selection criteria and common analyses produced spurious results. Later research comparing how statistical software designed and used by different manufacturers of fMRI machines handled spatial autocorrelation in the voxels revealed an error rate of up to 70% (Eklund et al. 2016). Yet even when done correctly, effects can be tiny and to reliably detect them requires sample sizes much larger than usually published (Marek et al. 2022). Clearly, fMRI-based science needed improvement. But are the statistical approaches and analyses used in fMRI any different from those used in other scientific fields every day?

All statistical methods make some assumptions about the independence, distributions, representativeness, etc. of the samples and the data collected from them. Most classical statistical methods have been mathematically (analytically) proven to yield “asymptotically unbiased estimates.” This phrase means that a statistical estimate will converge to its “true” value (that’s the “unbiased” bit) given an infinite number of observations (that’s the “asymptotic” bit) of randomly-collected samples (observations or data points) that conform to all the assumptions made in the statistical model. However, no one collects an infinite amount of data and residual assumptions are never 100% met, so it is important to ask how reliable our estimates are for finite or even rather small datasets. As a corollary, we also want to know how small is too small (i.e., how many observations are too few for us to compute unbiased, or at least reliably informative, estimates), how large is large enough, and whether our conclusions are robust to minor violations of model assumptions.

Furthermore, statistical methods generally have been developed as stand-alone tools. For example, general linear models (GLMs, of which the familiar analysis of variance [ANOVA] is a special case) are used to identify the expected (or average) response of one variable as a function of one or more predictor variables. Principal component analysis (PCA) is used to reduce the dimensionality of a large and often unwieldy set of variables (either predictors or responses) to a more manageable, smaller number (usually two or three) of composite variables that are linear combinations of the original data. Used by itself, PCA works well as a descriptive, hypothesis-generating tool, whereas a GLM works well as a predictive, hypothesis-testing tool. But what happens if we combine a PCA and a GLM in workflow? Will the PCA-GLM approach still “work?” And what about more complicated methods and

workflows using machine learning or neural networks? Will the resulting estimates still be unbiased or will uncertainty and errors propagate throughout the workflow and create statistical artefacts?

Questions such as these rarely can be answered for even a small number of statistical methods, hardly ever by non-mathematicians, and may even defy analytical solutions. Instead, mathematicians and statisticians rely on simulations to answer them.

This book aims at providing an entry to understanding statistics using simulations. By going through many examples, both conceptually and with code, we hope to lower the bar so that scientists who use statistics can simulate data and associated analyses themselves. Although simulations are a standard tool in statistical research and the development of new statistical methods, we do not present or test general methods. Rather, we illustrate the value of simulations by working through key issues and exemplar datasets in specific contexts representing a broad range of research areas—from sociology and archaeology to ecology and physics, and many in between.

What you will learn from this chapter

- What simulated data are, why they are important, and important ways in which they are used.
- How the book is organized and what you need to know to use this book effectively.
- How to read, interpret, and use the Code Boxes.
- What notational conventions are used for mathematical expressions throughout the book.

1.1 What are simulated data?

Simulated data are data similar to what we plan to collect or have already collected, but are created using computational algorithms.¹ We think it is helpful to think of such data as emerging from a “data-generating model” (DGM). The central challenge addressed in this book is to define such a data-generating model and then to use it effectively. When simulating data, we usually take advantage of random-number generators to include random variation in the simulated dataset that is similar to the random variation that occurs in real data. Examples of such random variation include differences among sampled individuals (i.e., between-subject variance), random positions, and residual errors, among many others.

Computers can simulate data efficiently and rapidly. When we run many simulations (e.g., thousands of them, each starting with a different random “seed”), we end up with replicate datasets for which we know the true underlying parameter values, the sampling design, underlying correlations, etc. When we analyse these simulated data, we can quickly check for and identify biases in our estimates and determine whether our methods are reliable and robust. We also can use simulations to test model fits and generate p -values. Pedagogically, it can be argued that if we cannot simulate our data, we probably do not know what we are doing when we are analysing them. Even if we think we know what we are doing, simulating data opens our eyes to what is simple and what is difficult in statistical analysis.

¹ Other terms you may encounter for simulated data include “synthetic data,” “fake data,” or “invented data.” Fake data and invented data have negative connotations and suggest that we are “cooking the books” by doing something wrong or illegitimate. “Synthetic data” is fine, but in the statistical sciences, “simulated” is used more frequently than “synthetic.”

Simulated data can take many forms, from a small number of observations of a single variable to a very large number of observations generated by a sophisticated model of a complex system driven by different underlying stochastic processes and sampled with a different stochastic sampling process (e.g., Zurell et al. 2010). Think, for example, about modelling traffic jams or representing the associated emergency calls as a way of collecting data. Both processes can be simulated at the level of the data (which is the focus of this book) or at the deeper level of the process(es) that could plausibly generate the data. The latter approach is touched upon in Chapter 12.

One more thing before we start. Simulations and reality have a lot in common.² In both, strong effects will shine through but weak effects will be hard to discern through the fog of high variability (Silver 2012).

1.2 Simulated data are specific

Simulation sacrifices generality for ease of use. That is, we typically use simulations to ask questions about a specific dataset or a very specific type of statistical analysis. Unlike an analytical proof, the results of a simulation study rarely are applicable to a wide range of statistical methods, even those that share underlying mathematical structures. Despite its lack of generality, however, simulation is a great tool to use to explore how a particular kind of statistical method works and what it can actually tell us about the kinds of data we can collect or already have on hand.

Any real dataset is but a single instance (and, we hope, a random one) of all the possible datasets that could have been collected from the population of interest. Although we might have sampled more or fewer subjects or sites, used a different method, or collected the data at a different time, we did not. We are usually interested in analysing such specific datasets and using the analyses to make more general inferences about all the datasets we did not collect, i.e., the population.

When we simulate data, we have such a specific dataset in mind, but we do not restrict ourselves to the exact values in our real dataset. Rather, we are asking how much quantitative information can be gleaned from this specific instance of a more general kind or type of dataset.³ In other words, is the information in the data sufficient for us to extract a signal from the noise? To answer this question using data simulation, we repeatedly generate data with a similar underlying structure and apply our intended analysis to them. Since we defined and set the parameters in our simulations, we can now check: How often do we find the effect we programmed into the simulation? Are our parameter estimates biased? If we are interested in testing our model, we can also ask whether the p -value of the output correctly reflects the number of significant effects across the simulations.

If a statistical method applied to our simulated data yields unbiased estimates or the expected distribution of p -values, then we can be more confident that the application of the method to the analysis of our single and unique real dataset will be similarly unbiased and the p -value will accurately reflect the statistical “significance” of our results. But remember the dead salmon. If the application of our method to our simulated data yields, for example,

² In fact, many science-fiction writers and even some dyed-in-the-wool philosophers have asserted that reality may be nothing more than a simulation (e.g., Adams 1979; Bostrom 2003).

³ We distinguish here between types and instances of a variable or a dataset. A *type* is a more general category, whereas an *instance* is a specific example of the type. For example, the value 10 is an instance of a variable whose type is “integer.” An interesting feature of types is that they are hierarchical, so that objects that are usually thought of as types also can be used as instances of other objects.

biased estimates, numerical anomalies, or incorrect standard errors, then we should not trust a similar analysis of the real dataset.

1.3 Yes, scientists really simulate data

Although some mathematicians still frown upon simulations, there are many interesting problems for which closed-form analytical solutions have not yet been identified. Using computers to simulate data and explore a range of possible solutions of analytically insoluble problems now is recognised as a legitimate approach to solving them; one of the first and perhaps most familiar examples is the solution to the four-colour map problem (Appel & Haken 1989; see Meyer & Schmidt 2012 for others). For many statistical methods, simulations are the only way available to validate established and proposed methods.

Not only do many scientists simulate data and analyse them,⁴ statisticians and many others teaching both theoretical and applied statistics actively encourage this practice (e.g., Crawley 2007; Jones et al. 2014; Morris et al. 2019; McElreath 2020; Crump et al. 2021; Donovan et al. 2021). Really, there is no reason not to, and we illustrate the power of simulation with a few straightforward examples in the following subsections and in the next chapter. These motivate the more complex problems described in the rest of the book and should whet your appetite for doing simulations yourself.

1.3.1 Geographic biases in reported crime rates

Governments collect and analyse data on crimes to identify where they are committed and the kinds of individuals most likely to commit them. Systematic reviews of these data have suggested that most crimes reported to police occur in a small number of neighbourhoods (Lee et al. 2017). This result has been used to develop strategies that, their proponents assert, will most efficiently deploy people and resources to those areas where crimes are most likely to occur. Critics point to the racial, socioeconomic, and other biases on which these spatially-targeted strategies are based and that they perpetuate. Simulations have been used extensively to support both sides of the debate (Liu & Eck 2008). An important, but rarely asked question, however, is whether the data used to build the crime simulations are themselves biased.

Buil-Gil et al. (2022) used simulation to address this question. They examined potential sources of bias in reported crime statistics in the UK's Crime Survey for England and Wales (CSEW). These researchers simulated three things for the city of Manchester, England: a population of people, the number of crimes experienced by each individual in the population, and the number of those crimes that were actually reported to the police (Figure 1.1). Buil-Gil et al. (2022) then asked whether the bias in the number of reported crimes changed from the smallest spatial scale (a neighbourhood of ≈ 125 households) to larger ones (communities [collections of neighbourhoods], precincts [collections of communities], and wards [collections of precincts]).

The results were striking. First, the simulations accurately reflected the CSEW data that nearly two-thirds (62%) of all committed crimes are not reported to the police. But the variation in non-reporting was largest among neighbourhoods (inset plot in Figure 1.1). That is, different neighbourhoods, each of which tends to be demographically more homogeneous than the larger communities, precincts, or wards, report crimes at different rates. These

⁴ Any list of fewer than thousands of references is unfair to those doing simulations, but here are a few more recent examples: Royston & Sauerbrei (2014); Boulesteix et al. (2018); Jayasekera et al. (2018); Boulesteix et al. (2020); Adams & Collyer (2022); Martínez-Santalla et al. (2022); DiRenzo et al. (2023).

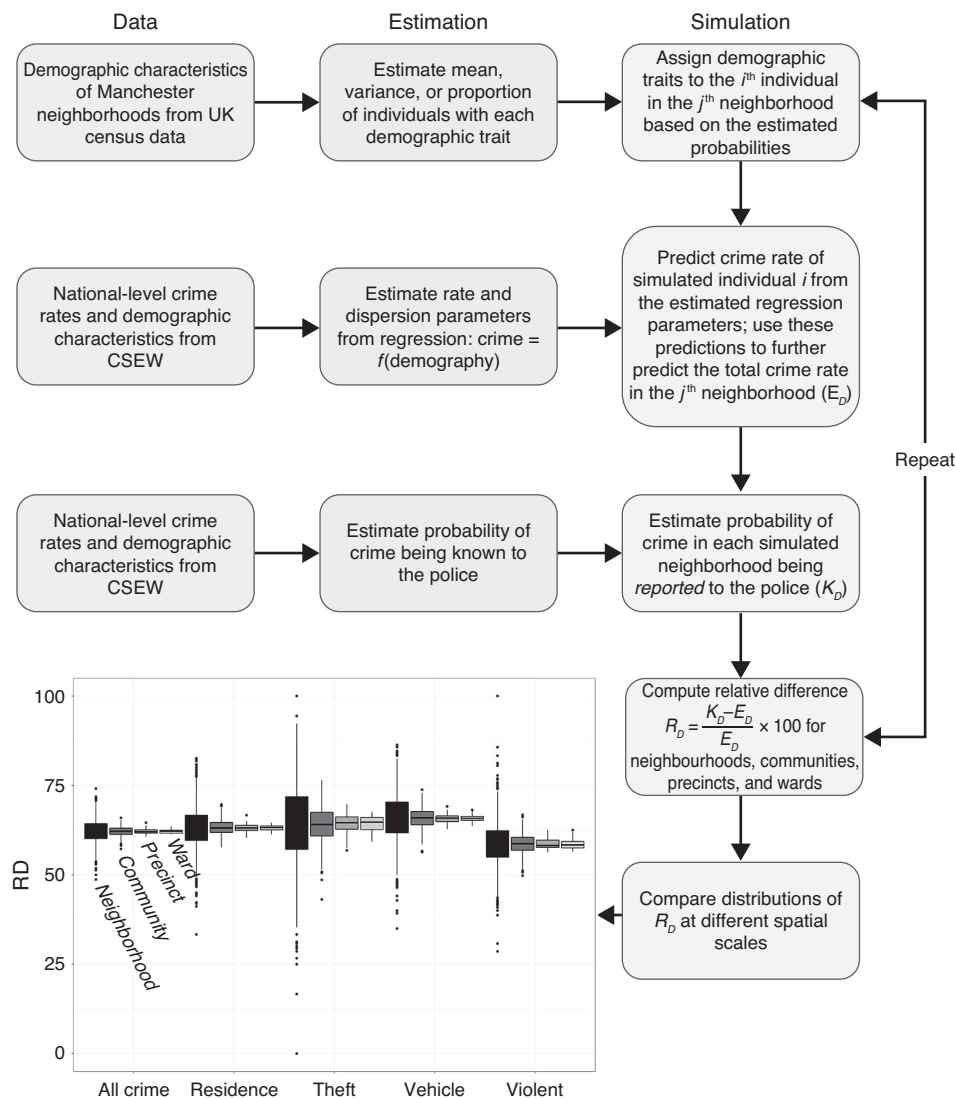


Figure 1.1 Workflow and results of the simulation exploring geographic bias in reported crime rates. The workflow, derived from the description of the simulation in Buil-Gil et al. (2022), includes obtaining data from government sources, estimation using negative binomial and logistic regressions, simulations, and visualisations. The inset figure is reproduced from Figure 1 of Buil-Gil et al. (2022) (CC-BY-4.0).

differences in reporting rates can lead to biases in the data that are used in criminological research and caution against blind acceptance of crime reduction strategies targeted at individual communities (“micro-geographies”).

1.3.2 Handling data gaps

FLUXNET⁵ is an international network of hundreds of stations that use the eddy-covariance technique to measure the flux of CO₂ and water above vegetation in a standardised way (Baldocchi 2003). The apparatus is very sensitive but requires air movement for measurements;

⁵ <https://fluxnet.org>.

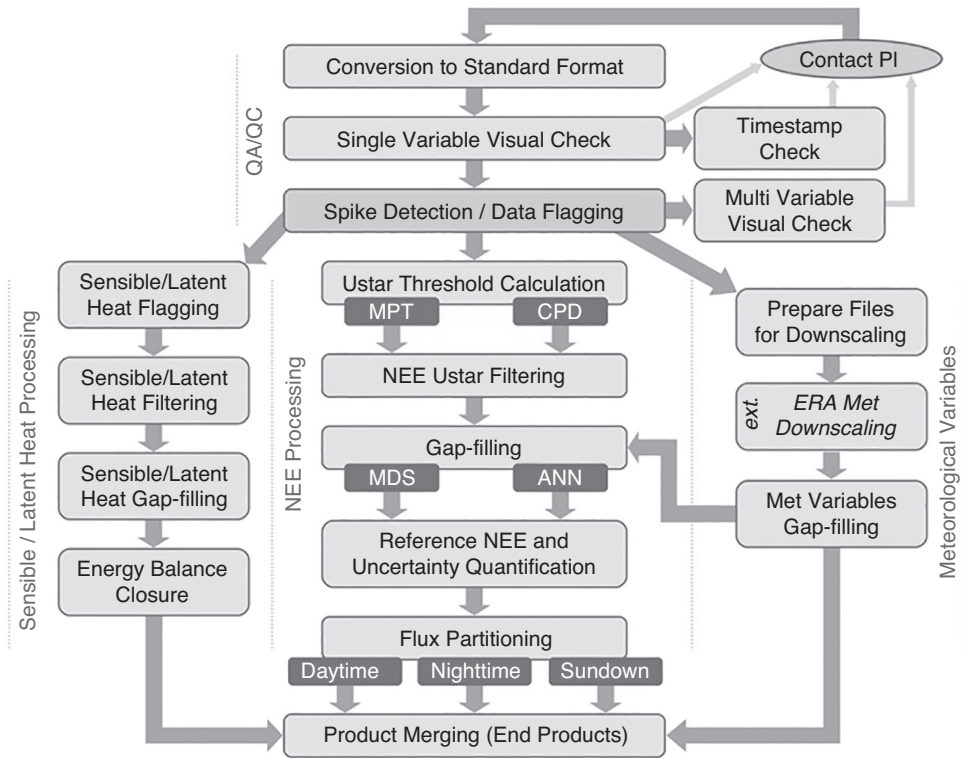


Figure 1.2 Workflow for gap filling of eddy covariance data in the FLUXNET network. Image in the public domain from FLUXNET (nd).

CO₂ flux cannot be measured at very low wind speeds, which happen regularly and especially at night and result in gaps in the data streams. Because these data gaps occur non-randomly, they need to be “filled” before the dataset can be analysed to avoid biased results.

Over the years, and after many, many simulations and comparisons of gap-filling methods, a sophisticated data pre-processing pipeline has been established (Figure 1.2). This is now the default workflow applied to all raw data entering the FLUXNET repository and data-processing chain. However, this workflow and the resulting estimates of the flux of carbon between forests and the atmosphere depends strongly on how the “U-star” (“Ustar” in Figure 1.2) threshold is determined. Simulations have shown that small changes in U-star could change the conclusion that a forest takes up more carbon from the atmosphere than it releases to its opposite—that the forest actually contributes planet-warming CO₂ into the atmosphere (Ellison et al. 2006).

1.3.3 Data augmentation in convolutional neural networks

Automatic classification of images is used widely—and often inappropriately or to ill effect—by many people, organisations, and government agencies. Classification, also called the “tyranny of the discontinuous mind” by Richard Dawkins (2011), can be improved by providing more and more data to the classifying algorithm. In lieu of genuine new data, computer scientists training convolutional neural networks (CNNs) on images use data augmentation: simulation of new data points based on old ones. Typical methods of data augmentation are flipping or rotating images, re-sizing or re-colouring them, adding noise to them, or increasing their grain. More recently, generative adversarial networks (GANs)

have introduced more sophisticated image mimicry but rely similarly on the variability in the training data. If each training image is altered in all combinations of such manipulations, sample size can be increased ten-fold with only marginal computational cost and at no cost to data collectors (e.g., Gopal et al. 2020; Nandhini Abirami et al. 2021; Li et al. 2023).

The second step of these simulations is to see whether data augmentation using GANs actually improves the classifier. Classifying the same images is uninformative; only classification of new, previously analysed, and validated images can tell us whether the method of data augmentation actually works. Indeed, it does improve classification accuracy for new images (e.g., Perez & Wang 2017; Mikołajczyk & Grochowski 2018), but not for the reasons initially suggested. Rather, the main reason seems to be that data augmentation using GANs adds noise, making the classifier avoid focusing on the specifics of an image, such as the stark contrast of the light reflection on the wet nose of a dog. As a result, data augmentation is now seen as a regularising or shrinkage approach rather than as a way to simulate new data (Steiner et al. 2022).

1.3.4 Null models in network analysis

When the value of an index depends in a complicated and non-linear way on its inputs, its distribution can be difficult to derive mathematically. One relevant case is the structure of bipartite (two-part) networks, which show the relationships between, for example, actors and the movies in which they appear, directors and the boards on which they serve, or pollinators and the flowers they visit. A description of what network structure we could expect based on how many movies and popular actors there are (i.e., in how many movies an actor has appeared) requires simulating the networks.

For example, when trying to identify modules (“cliques”) of actors (or pollinators), we use a “null model”: an algorithm to simulate a reference baseline. For either Boolean networks (e.g., an actor is in a movie or not) or networks with variable (but integer) number of interactions (e.g., a pollinator visits a flower N times), a modularity algorithm uses the sum of observations for each combination of actors and movies or pollinators and flowers, respectively, to simulate thousands of random networks against which the observed network is compared to determine if it is “structured” or random (e.g., Newman & Girvan 2004; Newman 2006; Barber 2007).

1.4 There are many good reasons to simulate data

Data simulation often is motivated by a specific problem with real data, such as violation of assumptions (e.g., normality, homoscedasticity), missing data, or unbalanced designs. However, there are many other benefits to simulating data besides demonstrating the validity or soundness of a particular method applied to a specific dataset. Perhaps most importantly, data simulation helps us to work through new problems that turn up while we are planning our research or analysing our data, and assess the validity of our proposed solutions when we are unsure of them.

While we were writing this book, ChatGPT, CoPilot, Llama, and many other large language models (LLMs) masquerading as artificial intelligence (AI) came online (cf. Searle 1982). Although many pundits (and not a few of our students and colleagues) have asserted that Big Data spelled the end of data analysis (e.g., Anderson 2008) and that AI and LLMs will make it unnecessary to learn how to write code (Hutson 2022), statisticians and programmers are still in high demand. Why? Even the best LLMs “hallucinate”: although they may confidently provide an answer to a question (a “prompt”), the answer may be inaccurate or completely wrong (e.g., Lehr et al. 2024). Hallucinations are most common when the

data used to train the LLM had errors or did not include enough information to address a particular prompt, when the prompt does not have sufficient context, or if the prompt concerns methods, packages, or programming languages that are not commonly used. Thus, it is incumbent on us to be sceptical of responses given by any AI or LLM. Answers should always be checked for accuracy and code should be tested carefully to determine whether it works the way we expect (e.g., Cooper et al., 2024). Independent data simulations are one of many useful tools to check the responses of an LLM and to use it effectively (see also Lubiana et al. 2023; van Dis et al. 2023).

In this book, our main focus is on using simulations to better understand what statistical models and analyses do, and in doing so, to improve our statistical competence. However, simulations increasingly are being used as part of scientific publications. In recent years, scientific societies and journal publishers have begun to develop guidelines for reporting the results of simulations so that readers can trust and maximally benefit from them (Morris et al. 2019; DiRenzo et al. 2023).

1.5 Useful background knowledge to use this book most effectively

This book is for anybody doing statistical analysis who already is familiar with basic concepts such as probability and likelihood and has skills with standard hypothesis-testing and model-fitting methods such as t -tests, χ^2 -statistics, and GLMs. Our previous books (Gotelli & Ellison 2012; Dormann 2020) can be used as refreshers. If you are starting from scratch, the OpenIntro project⁶ has excellent open-source resources for teaching and learning basic statistics. Its latest book, the second edition of *Introduction to Modern Statistics* (Çetinkaya-Rundel & Hardin 2024), covers basic concepts in modelling, hypothesis testing, and statistical inference; includes detailed exercises and scripts in the R programming language, and introduces simulation and randomisation tests with its discussion of the foundations of statistical inference. Fieberg (2024) focuses more tightly on alternatives to ordinary linear regression for analysing observational data collected by ecologists. His text includes many good examples of using data simulation and, like us, takes a pragmatic approach to the utility of both frequentist and Bayesian approaches.

We also assume familiarity with at least one statistical programming language, such as R, Python, Julia, Matlab, or Mathematica, and the ability to construct scripts in it from descriptive algorithms and pseudocode. We provide code for all our examples written in the R programming language (R Core Team 2023) and presented in a standard form for all the algorithms and examples in the text (Code Box 1.1).⁷

All the code may be downloaded from the online code repository associated with this book (Dormann & Ellison 2024). With only a few exceptions, we eschew the *tidyverse* and pipes (Wickham et al. 2019); all data wrangling and functions are written in “standard” R and will run in either the classic R GUI (Venables et al. 2023) or RStudio (RStudio Team 2020). Similarly, with the exception of the graphics in the Appendix, all plots are generated using *ggplot* (Wickham 2016). Unless the primary goal of a particular code snippet is generating a specific plot, we do not include the plotting commands in the Code Boxes. However, the code to generate all the plots is included in the online code repository. Finally, the online repository also includes additional examples and associated code that we have developed since writing the book, and we encourage our readers and others interested in data simulation to contribute their own examples and code to the repository.

⁶ <https://openintro.org>

⁷ There are packages in R or Python that simulate data for you (e.g., *conjurer* and *simstudy* in R; see Appendix A for others), but which hide what actually is going on. So, we use them sparingly.

Code Box 1.1: How to read the code in the Code Boxes

```
# Code is presented in the same way throughout the text

# Lines or phrases beginning with a "hash tag" or "pound sign" (#) are comments
  that describe the subsequent or associated executable commands. If these
  lines are pasted into an R script, they will not run. For example, the line

5 # Load required libraries

# tells you that the subsequent lines will include the contributed packages
  (R "libraries") necessary for the executable code to run successfully.

# Lines indented with one or more spaces and lacking a leading character (such
  as a # or a |) are executable code. If these are pasted into an R script,
  they will run. For example, the line
10 mean(x)

# will return the average of the variable "x".

15 # Lines beginning with a vertical bar (|) include the output (returned to the
  console) if executable code is executed. These lines should not be pasted
  into an R script, as they are neither comments nor executable code. For
  example,

| [1] 13.75

# would be the output from running the previous command (mean(x)).
```

1.6 Notational conventions

We have endeavoured to use consistent notation for mathematical expressions throughout the book, following Casella & Berger (2002). An upper-case letter (e.g., X) indicates a random variable, whereas its corresponding lower-case letter (e.g., x) is a specific value from it. The set of all N values of $\{X_1, X_2, \dots, X_N\}$ is denoted as X or $\{X\}$. We capitalise the sample size, N , throughout the main text, but in Code Boxes, it may appear in either lower case or upper case, depending on how it is defined in existing R functions.⁸ We do not use vector notation (e.g., $\mathbf{X}$) for a matrix X ; the difference between scalars and vectors should be clear from context.

Population-level (“true” and usually unknown) parameters are written with Greek letters (e.g., the population mean μ and its variance σ^2), whereas sample parameters normally are written with Roman letters (e.g., the sample mean, $\bar{x}$, and its variance, s^2). Modelled estimates are indicated with “hats” (e.g., the estimated population variance is written as $\hat{\sigma}^2$ and the estimated sample variance is written as $\hat{s}^2$).

⁸ Note that commands, object names, etc. in R are case-sensitive: “Bird” and “bird” would be two different kinds of flying animals.

The names of libraries, functions, and variables used in R (or other programming languages) are set in monospaced font; function names always include parentheses, as in `t.test()`, to serve as a reminder that they need to have objects, parameters, or options provided to them (e.g., `t.test(x, y, type="paired")`).

1.7 Structure, organisation, and flow

The book is constructed in such a way that the two chapters in this Part I set the tone. After that, we proceed through topics following the same path that we would normally use to develop a research project: ask a question and develop hypotheses; come up with a sampling protocol or experimental design; collect, visualise, summarise, and analyse the data; fit and test models and (re)check their assumptions; and state our conclusions together with our degree of certainty about them. We show that using data simulation helps us do all these steps better, improves our science, and give us more confidence in the conclusions we draw from our research.

In the two chapters in Part II of this book, we show how to use simulations to plan surveys, observations and experiments from the time we come up with an idea until we actually start the study. For example, we can investigate how a proposed sample size will affect the probability of detecting weak effects. Simulation might also suggest that we abandon a proposed study if we cannot muster the necessary resources and logistics to reliably test our hypotheses. Thinking longer about a study *before* we do it is always beneficial. Data simulation encourages us to think first and act later, and forces us to be explicit about our assumptions and expectations.

The three chapters in Part III get us ready to analyse our hard-earned data. Before we analyse the data we have collected, we should check to see if our data satisfy key distributional assumptions (Chapter 5). We also can use simulations to check whether the method or model we plan to use to analyse our data makes sense or is it simply folklore: a rule-of-thumb that “everybody knows” or has “always been done” (Chapter 6). In a similar vein, individual models may work fine by themselves but exhibit strange behaviour when linked together in a workflow or analytical pipeline. Data simulations can help us discover that before we get so far into a tunnel that we cannot turn around and extricate ourselves (Chapter 7).

The two chapters in Part IV showcase ways to use data simulations to evaluate the fit of our models and what to do if they exhibit strange behaviour or yield results that are not statistically significant. If our models work as expected, all is well; simulations let us evaluate the fit of models we know to be reliable (Chapter 8). If they do not and our model yields really strange or unexpected results when we analyse our real data, we also can use simulations to probe the method or model more deeply, identify the origin of the weird results, and have firmer grounds from which to judge whether the results are strange but true or only an odd artefact of the specific sample or method (Chapter 8). Finally, if—despite all our careful planning, scrupulous data collection, and robust analysis—our results are not statistically significant, simulation shows us that we should not seek absolution in retrospective power analysis (Chapter 9). At the same time, we can use retrospective design analysis (Gelman & Carlin 2014) to test whether our statistically significant results are as meaningful as we think they are (Chapter 9).

The three chapters in Part V close the circle. We can use all the approaches presented in Parts II–IV to synthesise existing work using meta-analysis or federated analyses (Chapter 10), or to design and test new statistical methods or create new indices (Chapter 11). Using simulations offers an efficient and effective way to demonstrate the appropriateness of a new procedure or detect its flaws. Finally, Chapter 12 returns to where we started this

Table 1.1: Where to find detailed discussion of key statistical topics that recur throughout this book.

Topic	Chapter.Section
Assumptions:	
– Normally-distributed data	2.3; 5.1; 5.6; 8.4
– Homoscedasticity	2.2; 5.4; 8.4
– Independence	2.4; 4.2; 5.5–5.6; 6.5–6.6
– Stationarity	5.7; 12.2
Bayesian inference:	
– Bayes’ Theorem	3.3; 6.4
– Prior vs. posterior distributions	8.1; 8.6
– MCMC and its relatives	7.7; 8.1; 11.6
Categorical vs. continuous predictors	6.3; 8.5
Collinearity	6.2; 7.5; 7.8
Confidence intervals	2.3; 3.4; 8.5; 9.3
Effect size	3.4; 4.1; 7.6; 9.1–9.3; 10.3–10.7
Imputation of missing data	6.1; 7.3–7.4; 7.8
Machine learning vs. classical statistics	6.2; 6.6; 7.5; 8.4–8.5; 11.4–11.5
Meta-analysis	3.4; 10.3–10.4
Model selection	6.1; 7.8; 10.2; 11.3
Power analysis	
– a priori (prospective)	3.4; 4 [entire chapter]
– post hoc (retrospective)	9 [entire chapter]; 10.3
Time-series analysis	2.4; 4.4; 6.4–6.5; 11.5

book: using simulations for planning large and very complex research programs that are addressing some of the most difficult questions scientists have posed to date.

We encourage you to read and work through the chapters in order. However, each chapter is mostly self-contained and you should be able to jump to, and work with, any part or chapter of the book in which you are most interested. Some key statistical topics and concepts pertain to more than one method or model, or recur in different contexts. Table 1.1 provides an overview of the most important of these topics and points to where we present them in detail. For more fine-grained direction, please consult the table of contents and the index.

1.8 Summary

- Simulated data are created using computational algorithms.
- Simulated data are similar, but not identical, to data we have already collected or that we plan to collect.
- Simulating data requires a “data-generating model” that normally implies a specific question, dataset, or type of statistical analysis.
- One of the most important uses of data simulation is to determine how much information about a large population can be gleaned from the smaller sample of it that has been collected.

- Other common uses of data simulation include testing assumptions, working through unanticipated or unexpected problems that came up while planning a research project or analysing the data, assessing the validity of proposed solutions, and determining whether the results of using large language models are real or hallucinatory.
- To use this book most effectively, you should already be familiar with basic statistical concepts such as probability and likelihood, and have skills with standard hypothesis-testing and model-fitting methods such as t -tests, χ^2 -statistics, and GLMs.

References

- Adams, D. (1979). *The Hitchhiker's Guide to the Galaxy*. London, UK: Pan/Macmillan.
- Adams, D. C. & Collyer, M. L. (2022). Consilience of methods for phylogenetic analysis of variance. *Evolution*, 76(7), 1406–1419. doi: 10.1111/evo.14512.
- Anderson, C. (2008). The end of theory: The data deluge makes the scientific method obsolete. *WIRED*, 16(7). <https://www.wired.com/2008/06/pb-theory/>.
- Appel, K. & Haken, W. (1989). *Every Planar Map Is Four Colorable*. Providence, RI, USA: American Mathematical Society.
- Baldocchi, D. D. (2003). Assessing the eddy covariance technique for evaluating carbon dioxide exchange rates of ecosystems: Past, present and future. *Global Change Biology*, 9(4), 479–492. doi: 10.1046/j.1365-2486.2003.00629.x.
- Barber, M. (2007). Modularity and community detection in bipartite networks. *Physical Review E*, 76(6), 1–9. doi: 10.1103/PhysRevE.76.066102.
- Bennett, C. M., Miller, M. B., & Wolford, G. L. (2009). Neural correlates of interspecies perspective taking in the post-mortem Atlantic Salmon: An argument for multiple comparisons correction. *NeuroImage*, 47, S125. doi: 10.1016/S1053-8119(09)71202-9.
- Bostrom, N. (2003). Are you living in a computer simulation? *Philosophical Quarterly*, 53(211), 243–255.
- Boulesteix, A.-L., Binder, H., Abrahamowicz, M., et al. (2018). On the necessity and design of studies comparing statistical methods. *Biometrical Journal*, 60(1), 216–218. doi: 10.1002/bimj.201700129.
- Boulesteix, A.-L., Groenwold, R. H. H., Abrahamowicz, M., et al. (2020). Introduction to statistical simulations in health research. *British Medical Journal Open*, 10(12), e039921. doi: 10.1136/bmjopen-2020-039921.
- Buil-Gil, D., Moretti, A., & Langton, S. H. (2022). The accuracy of crime statistics: Assessing the impact of police data bias on geographic crime analysis. *Journal of Experimental Criminology*, 18, 515–541. doi: 10.1007/s11292-021-09457-y.
- Casella, G. & Berger, R. L. (2002). *Statistical Inference*. 2nd edition. Duxbury Press/Thomson Learning.
- Çetinkaya-Rundel, M. & Hardin, J. (2024). *Introduction to Modern Statistics*. <https://openintro.org>: OpenIntro, 2nd edition.
- Cooper, N., Clark, A. T., Lecompte, N., et al. (2024). Harnessing large language models for coding, teaching and inclusion to empower research in ecology and evolution. *Methods in Ecology and Evolution*, 15(10), 1757–1763. doi: 10.1111/2041-210X.14325.
- Crawley, M. J. (2007). *The R Book*. Chichester, West Sussex, UK: John Wiley & Sons.
- Crump, M. J. C., Navarro, D., & Suzuki, J. (2021). Answering questions with data: Introductory statistics for psychology students. DOI: 10.17605/OSF.IO/JZE52. Accessed 1 February 2024.
- Dawkins, R. (2011). The tyranny of the discontinuous mind. *New Statesman*, 100(5084/5085), 54–57.
- DiRenzo, G. V., Hanks, E., & Miller, D. A. W. (2023). A practical guide to understanding and validating complex models using data simulations. *Methods in Ecology and Evolution*, 14(1), 203–217. doi: 10.1111/2041-210X.14030.
- Donovan, T. M., Brown, M., & Katz, J. E. (2021). R for fledglings. <https://www.uvm.edu/~tdonovan/RforFledglings>. Accessed 1 February 2024.
- Dormann, C. & Ellison, A. M. (2024). Data ex machina. <https://osf.io/3skv4/>.
- Dormann, C. F. (2020). *Environmental Data Analysis: An Introduction with Examples in R*. Heidelberg, Germany: Springer.
- Eklund, A., Nichols, T. E., & Knutsson, H. (2016). Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proceedings of the National Academy of Sciences, USA*, 113, 7900–7905. doi: 10.1073/pnas.1612033113.
- Ellison, A. M., Osterweil, L. J., Hadley, J. L., et al. (2006). Analytic webs support the synthesis of ecological datasets. *Ecology*, 87(6), 1345–1358. doi: 10.1890/0012-9658(2006)87[1345:AWSTSO]2.0.CO;2.

- Fieberg, J. R. (2024). *Statistics for Ecologists: A Frequentist and Bayesian Treatment of Modern Regression Models*. Minneapolis, MN, USA: University of Minnesota Libraries Publishing.
- FLUXNET (n.d.). Data Processing 101: Pipeline and Procedures. <https://fluxnet.org/data/aboutdata/data-processing-101-pipeline-and-procedures/>. Accessed 16 January 2024.
- Gelman, A. & Carlin, J. (2014). Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science*, 9(6), 641–651. doi: 10.1177/1745691614551642.
- Gopal, A., Gandhimaruthian, L., & Ali, J. (2020). Role of general adversarial networks in mammogram analysis: A review. *Current Medical Imaging*, 16(7), 863–877. doi: 10.2174/1573405614666191115102318.
- Gotelli, N. J. & Ellison, A. M. (2012). *A Primer of Ecological Statistics*, 2nd edition. New York, NY, USA: Oxford University Press.
- Hutson, M. (2022). AI learns to write computer code in ‘stunning’ advance. DOI: 10.1126/science.adg2088. Accessed 26 May 2024.
- Jayasekera, J., Li, Y., Schechter, C. B., et al. (2018). Simulation modeling of cancer clinical trials: Application to omitting radiotherapy in low-risk breast cancer. *JNCI: Journal of the National Cancer Institute*, 110(12), 1360–1369. doi: 10.1093/jnci/djy059.
- Jones, O., Maillardet, R., & Robinson, A. (2014). *Introduction to Scientific Programming and Simulation Using R*, 2nd edition. Boca Raton, FL, USA: Chapman & Hall/CRC.
- Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S. F., & Baker, C. I. (2009). Circular analysis in systems neuroscience: The dangers of double dipping. *Nature Neuroscience*, 12(55), 535–540. doi: 10.1038/nn.2303.
- Lee, Y., Eck, J. E., O, S., & Martinez, N. N. (2017). How concentrated is crime at places? A systematic review from 1970 to 2015. *Crime Science*, 6, 6. doi: 10.1186/s40163-017-0069-x.
- Lehr, S. A., Calliskan, A., Liyanage, S., & Banaji, M. R. (2024). ChatGPT as research scientist: probing GPT’s capabilities as a research librarian, research ethicist, data generator, and data predictor. *Proceedings of the National Academy of Sciences USA*, 121(35), e2404328121.
- Li, J., Yu, Z., Yu, L., et al. (2023). A comprehensive survey on SAR ATR in deep-learning era. *Remote Sensing*, 15(5), 1454. doi: 10.3390/rs15051454.
- Liu, L. & Eck, J., Eds. (2008). *Artificial Crime Analysis Systems: Using Computer Simulations and Geographic Information Systems*. Hershey, PA, USA: Information Science Reference.
- Lubiana, T., Lopes, R., Medeiros, P., et al. (2023). Ten quick tips for harnessing the power of ChatGPT in computational biology. *PLoS Computational Biology*, 19(8), 1–9. doi: 10.1371/journal.pcbi.1011319.
- Marek, S., Tervo-Clemmens, B., Calabro, F. J., et al. (2022). Reproducible brain-wide association studies require thousands of individuals. *Nature*, 603(7902), 654–660. doi: 10.1038/s41586-022-04492-9.
- Martínez-Santalla, S., Martín-Devasa, R., Gómez-Rodríguez, C., et al. (2022). Assessing the nonlinear decay of community similarity: Permutation and site-block resampling significance tests. *Journal of Biogeography*, 49(5), 968–978. doi: 10.1111/jbi.14351.
- McElreath, R. (2020). *Statistical Rethinking: A Bayesian Course with Examples in R and STAN*, 2nd edition. Boca Raton, FL, USA: Chapman & Hall/CRC.
- Meyer, K. R. & Schmidt, D. S., Eds. (2012). *Computer Aided Proofs in Analysis*. New York, NY, USA: Springer.
- Mikołajczyk, A. & Grochowski, M. (2018). Data augmentation for improving deep learning in image classification problem. In *2018 International Interdisciplinary PhD Workshop (IIPhDW)* (pp. 117–122). Świnouście, Poland.
- Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, 38(11), 2074–2102. doi: 10.1002/sim.8086.
- Nandhini Abirami, R., Durai Raj Vincent, P. M., Srinivasan, K., et al. (2021). Deep CNN and deep GAN in computational visual perception-driven image analysis. *Complexity*, 2021, 5541134. doi: 10.1155/2021/5541134.
- Newman, M. E. J. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences, USA*, 103(23), 8577–8582. doi: 10.1073/pnas.0601602103.
- Newman, M. E. J. & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E*, 69(2), 1–15. doi: 10.1103/PhysRevE.69.026113.
- Perez, L. & Wang, J. (2017). The effectiveness of data augmentation in image classification using deep learning. *arXiv*, 1712.04621. doi: 10.48550/arXiv.1712.04621.
- R Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Royston, P. & Sauerbrei, W. (2014). Interaction of treatment with a continuous variable: Simulation study of power for several methods of analysis. *Statistics in Medicine*, 33(27), 4695–4708. doi: 10.1002/sim.6308.
- RStudio Team (2020). *RStudio: Integrated Development Environment for R*. RStudio, PBC., Boston, MA.
- Searle, J. (1982). The myth of the computer. *The New York Review of Books*, 29 (April 29), 3–7. <https://www.nybooks.com/articles/1982/04/29/the-myth-of-the-computer/>.
- Silver, N. (2012). *The Signal and the Noise: Why So Many Predictions Fail—But Some Don’t*. New York, NY, USA: Penguin Press.

- Steiner, A., Kolesnikov, A., Zhai, X., et al. (2022). How to train your ViT? Data, augmentation, and regularization in vision transformers. *arXiv*, 2106.10270. doi: 10.48550/arXiv.2106.10270.
- van Dis, E. A. M., Bollen, J., van Rooij, R., et al. (2023). ChatGPT: Five priorities for research. *Nature*, 614(7947), 224–226. doi: 10.1038/d41586-023-00288-7.
- Venables, W. N., Smith, D. M., & The R Core Team (2023). An Introduction to R. <https://cran.r-project.org/doc/manuals/r-release/R-intro.pdf>. Accessed 11 February 2024.
- Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. New York, NY, USA: Springer-Verlag, 2nd edition.
- Wickham, H., Averick, M., Bryan, J., et al. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43), 1686. doi: 10.21105/joss.01686.
- Zurell, D., Berger, U., Cabral, J. S., et al. (2010). The virtual ecologist approach: Simulating data and observers. *Oikos*, 119(4), 622–635. doi: 10.1111/j.1600-0706.2009.18284.x.

Index

Italic pagination indicates figures and tables, “n” indicates footnotes

- Adams, D., 137
- Addor, N., 252
- advanced simulations, 12–13; bias and, 403–6; covariance and, 409, 412; data-generating model (DGM) and, 407; errors and, 404; future of, 406–7; Gaussian distribution and, 405; hypothesis testing and, 406; LIDAR, 406; LIGO, 404–6; manipulation and, 404; NEON, 406; neural networks and, 406; noise and, 404–7; parameter values and, 404, 419; pipelines and, 403–5; sample size and, 407; signals and, 404–6; simulated data and, 404; stochastic weather forecasting, 403–4; uncertainty and, 403–4
- AIC: Akaike weight and, 155, 389; assumptions of statistical analysis and, 148; errors and, 254–55; evaluation and, 296, 389–97; GLS analysis and, 36; missing data and, 217–19; model selection and, 155, 157–59; predictor variables and, 238–39
- Akaike weight, 155, 389
- Alice’s restaurants, 271–273, 274
- American Statistical Association, xii
- analysis of variance (ANOVA): assumptions of statistical analysis and, 105, 116–23, 127, 145; bias and, 314; degrees of freedom and, 398; effect sizes and, 314; evaluation and, 398; frequency distribution and, 20; *F*-statistic and, 20–21, 116; *F*-values and, 20–21, 23, 117, 122; heteroscedasticity and, 20–21, 118–22, 154; homoscedasticity and, 13, 105–6, 116–23; normality and, 18; null hypothesis and, 118; predictor variables and, 179; residuals and, 18, 116–17; retrospective power analysis and, 314–15, 316; robustness and, 118–22; rules-of-thumb and, 154, 179; simulation basics and, 3, 17–22, 23; study designs and, 60
- Anderson-Darling test, 106
- apply, 413–16
- artificial intelligence (AI): Big Data and, 9–10, 201; evaluation and, 307; large language models (LLMs) and, 9, 14; machine learning and, 201, 205
- assumptions of statistical analysis, 12; AIC and, 148; analysis of variance (ANOVA) and, 105, 116–23, 127, 145; autocorrelated data and, 106, 149; Bayesian models and, 103, 131–32, 149; bias and, 105, 112, 114–15, 120, 134, 137, 145; confidence interval (CI) and, 103; correlation and, 103n1, 106, 114, 137–46, 149; degrees of freedom and, 116–17, 148; errors and, 103, 105–13, 116–17, 122, 123–28, 138, 147, 149; false discovery rate (FDR) and, 126–31; false nulls, 128–31; false positives, 105, 123–31, 145; family-wise rates and, 123–28; fixed effects and, 133–34, 138; frequentist analysis and, 103; *F*-statistic and, 116; Gaussian distribution and, 110–14, 131–32, 137n25; general linear models (GLMs) and, 103, 145, 147; hierarchical data and, 133–41; hyper-distributions and, 131–37; hypothesis testing and, 103; Kolmogorov-Smirnov (K-S) test and, 106, 135n24; machine learning and, 103, 105, 111; matrices and, 128–32; mixed-effect models and, 131–37; model structure and, 133, 149; multiple simulation analysis and, 115–16; multiple testing and, 123–31; noise and, 108–15; normal distribution and, 106–9, 116, 131–32, 135, 137, 140, 142; normality and, 105–9, 116, 132–37, 142, 145, 148; null hypothesis and, 118, 126–28; parameter values and, 103, 106, 107n2, 110, 114–16, 119, 131–35, 139–40, 142, 145, 148; post hoc analysis and, 118, 123, 145; predictor variables and, 137–45; principal component analysis (PCA) and, 142, 144–46; probability and, 123–24, 126–27, 145; *p*-values and, 107, 123, 126–27, 145, 148; replicates and, 108–9, 112, 113, 124–30, 134, 139, 141, 146; residuals and, 3, 106–9, 116–17, 121, 132–35, 138–41, 145, 147–48; robustness

- assumptions of statistical analysis (*continued*)
 - and, 105, 118, 123; sample size and, 106–10, 119, 124–25; seeds and, 107–11, 119–25, 128–30, 146–49; simulated data and, 107, 113, 127, 133–35, 143; single simulation analysis and, 115; standard deviation and, 107n3, 117, 131, 140; true positives and, 123–24; *t*-test and, 103, 120, 124, 127, 138; uncertainty and, 115, 147; violating, xi
- atrazine, 227–31
- autocorrelated data: assumptions of statistical analysis and, 106, 149; censored observations and, 226; change-point detection and, 196n27; errors and, 252; evaluation and, 281; machine learning and, 203–4; meta-analysis and, 343; Monte Carlo simulation and, 181, 183; noise and, 34, 86–87, 91; phylogenetic, 191n25; power of statistical analysis and, 86–92; residuals and, 33, 87–88, 91–92, 183, 193, 204, 281; sample size and, 3; simulation basics and, 3, 33–36; spatial, 3, 33, 191n25, 204; temporal, 34–36, 86–88, 91, 106, 181, 183, 203–4, 226; time series and, 191, 193, 196n27, 281
- Aydemir, A., 110
- base rate fallacy, 128–29
- Bayesian models: assumptions of statistical analysis and, 103, 131–32, 149; censored observations and, 226n9; didactic example of, 271–73; errors and, 246, 249–51; evaluation and, 269–72, 274, 283, 306n23, 387–89, 397; frequentist analysis and, 10, 103, 182, 306n23, 307, 312, 387; improving estimation in, 93, 98; learning from prior, 270–73; machine learning and, 203n33; Monte Carlo simulation and, 182, 187–89; power of statistical analysis and, 93, 98; predictor variables and, 232; recipes and, xi; retrospective power analysis and, 312, 319, 331, 334; simulation basics and, 10, 13, 22n6; study designs and, 45, 49
- before-after (BA) design, 99–100
- before-after control-impact (BACI), 99–100
- Benjamini-Hochberg procedure, 126–30
- Bennedson, M., 182, 184, 187–90
- Berger, R. L., 11
- Berlin, J. A., 314–16, 321–22
- Bernoulli data, 52n5, 55n8, 202, 236, 295, 387
- bias: advanced simulations and, 403–6; ANOVA and, 314; assumptions of statistical analysis and, 105, 112, 114–15, 120, 134, 137, 145; complete data analysis and, 222–32, 360–68, 370; crime rate reporting and, 6–7; effect size and, 314; errors and, 245–62; evaluation and, 273, 277, 375–77, 381–82; federated analysis and, 341–45, 357–58, 361, 362, 368–71; machine learning and, 201–3; meta-analysis and, 341–58, 369, 370; missing data and, 212–22; model selection and, 154n2, 159–60, 171; “network of chums” effect and, 344; parameter values and, 4–5, 69, 81, 115, 145, 159, 171–73, 211, 213, 222, 226, 232, 235, 242, 252–59, 362, 368–69, 371; power of statistical analysis and, 74, 78–85, 88, 89, 97, 101; predictor variables and, 171–73, 232–42; publication, 342–45, 353, 362, 365–67, 370; regression analysis and, 171, 242–45; relative, 238, 240–41, 242, 257, 262, 348–49, 351, 354–71; retrospective power analysis and, 314, 318, 320n18, 329, 331; simulation basics and, 3–8, 22n7, 34; study designs and, 45–48, 58n10, 65, 69; uncertainty and, 4, 47–48, 69, 115, 171, 211, 252, 254, 260, 273, 277, 404
- Big Data: artificial intelligence (AI) and, 9–10; era of, 341; machine learning and, 153, 200–4; meta-analysis and, 341; rules-of-thumb and, 153, 200–4; workflow and, 211
- binomial distribution: evaluation and, 286, 295, 297; Polya’s urn and, 18–20
- Bioclim, 146
- bipartite networks, 9
- BLOOM, xii
- bootstrapping: confidence interval (CI) and, 39–40, 43, 65–66, 69, 81, 83, 204, 288, 300–2, 304; data perturbation and, 39–40, 43, 416; errors and, 245–46, 251–59, 260; evaluation and, 273, 275–77, 288, 294n16, 299–303, 304, 306; general linear models (GLMs) and, 30; jackknife and, 17, 299–303, 304; likelihood and, 39–40, 43, 204, 273; machine learning and, 204; Monte Carlo simulation and, 181n14; power of statistical analysis and, 78–79, 81, 83; regression analysis and, 29, 31, 40, 64–66, 300–2, 304, 416; resampling and, 17, 40, 43, 416–18; R functions and, 416–19; study designs and, 48, 64–66, 69; uncertainty and, 29–30, 31, 39, 69, 245, 251–54, 260, 273, 275, 277, 288, 306, 418
- Borjas, G. J., 110
- Boulesteix, A.-L., 381
- breakpoint: Monte Carlo simulation and, 182n15, 189; randomness and, 37–42; residuals and, 40, 42; simulated data and, 37–42. *See also* change-point methods
- breast cancer, 235–38, 288
- Brezna, N., 331–32
- Broe, M., xii
- Broennimann, O., 142
- Buil-Gil, D., 6, 7
- carbon dioxide (CO₂): analyzing trends in, 87–88; ecological footprint and, 389; FLUXNET and, 7–8; monitoring atmospheric, 85–95, 96, 99n22; Monte Carlo simulation and, 181–90; power of statistical analysis and, 85–87, 89n15, 90–96, 99n22; predictor variables and, 177; simulation parameters and, 86–87; time series and, 85–87, 89n15, 90–91
- Carlin, J., 327–34
- Casella, G., 11
- categorisation: effect size and, 62n17; predictor variables and, 154, 172–80, 205
- C/C++, 166n8, 381, 384–85, 393

- censored observations: autocorrelated data and, 226; Bayesian models and, 226n9; correlation and, 222, 226; covariance and, 363–66; degrees of freedom and, 230–31; detection limit and, 222–33, 235; errors and, 228n11, 230; estimating values for, 222–32; false negatives and, 228n11; false positives and, 228n11; *F*-statistic and, 230; Gaussian distribution and, 231; imputation and, 224, 225, 232; left-censored data and, 222, 226–27, 230; likelihood and, 229; machine learning and, 226, 232; matrices and, 223; model structure and, 232; parameter values and, 226, 232; probability and, 225, 228n11; random effects and, 229; replicates and, 223, 225; residuals and, 230; right-censored data and, 224, 227; sample size and, 223, 224; seeds and, 223, 225; signals and, 228n11; truncated data and, 224n8
- Central Limit Theorem, 131
- Çetinkaya-Rundel, M. 10
- change-point methods, 192, 194–98. *See also* breakpoint
- ChatGPT, 9
- chemistry, xii, 1, 276, 318
- “Chinese restaurant” distribution, 20
- classification: automatic, 8; evaluation and, 288, 297, 302–3, 397; generative adversarial networks (GANs) and, 8–9; K-S test and, 28; machine learning and, 202; Monte Carlo simulation and, 29n17
- classification and regression trees (CARTs), 202
- cliques, 9
- code verification, 377. *See also* evaluation; functional verification; unit testing
- Cohen, J., 313, 314n7, 322, 329
- Cohen’s *d*, 58n10, 62n17, 314–15, 316
- collinearity: errors and, 252, 255, 258; evaluation and, 398; model selection and, 165–72; “pick one” approach and, 171–72; predictor variables and, 238; rules-of-thumb and, 154, 165–72; simulation basics and, 13
- complete data analysis: bias and, 222–32, 360–68, 370; as coordinated data analysis, 360; effect size and, 360–68; federated analysis and, 341–45, 361, 362–63, 368; matrices and, 362–66; meta-analysis and, 341–45; parameter values and, 360, 362, 368; predictor variables and, 360–68; randomness and, 222, 224, 225, 226, 229, 361, 363–64, 366; repeated simulation and, 362–64; replicates and, 363, 366; seeds and, 363–66; stepwise model selection and, 362; study identifiers and, 364–65; unbiased data and, 362
- compromise power analysis, 328
- conditional plots, 276–80
- confidence interval (CI): assumptions of statistical analysis and, 103; bootstrapping and, 39–40, 43, 65–66, 69, 81, 83, 204, 288, 300–2, 304; errors and, 249; evaluation and, 270, 286, 288, 290–307; hypothesis testing and, 334; machine learning and, 203–4, 295–305; misinterpretation of, 21–22; missing data and, 220; model variability and, 28; Monte Carlo simulation and, 182, 293; narrow, 212; non-Gaussian data and, 294–95; post hoc analysis and, 319–22; power of statistical analysis and, 78, 80, 81, 83; prediction interval and, 203–4, 270, 288–307; predictor variables and, 174, 320–22; regression analysis and, 243; retrospective power analysis and, 311n1, 313, 319–26, 333–34; R functions and, 419, 421; simulation basics and, 28, 35, 39–40, 43; study designs and, 63, 65–66, 69; true meaning of, 17, 21–26
- consistency, xii, 211, 358n15, 377
- control-impact (CI) design, 99–100
- convolutional neural networks (CNNs), 8
- Cook’s distance, 281, 282
- CoPilot, 9
- correlation: assumptions of statistical analysis and, 103n1, 106, 114, 137–46, 149; censored observations and, 222, 226; coefficient of, 34, 57–58, 74–78, 81, 83, 143, 166, 285, corridor of stability of, 78–82, 83; errors and, 248; 249, 252–53, 257, 260; evaluation and, 270–71, 285, 378; federated analysis and, 358; machine learning and, 203–4; meta-analysis and, 343, 346; missing data and, 214–15, 222; model selection and, 164–67; Monte Carlo simulation and, 181, 183; number of data points for, 76–80; Pearson’s correlation coefficient, 34, 74, 76; point of stability of, 78–79, 81, 83 power of statistical analysis and, 74–82, 83, 86–88, 91–95, 101; predictors and, 137–46; 149, 235–42; regression models and, 33, 58, 106, 153, 164–65, 181, 183, 215, 243; retro power analysis and, 312; simulation basics and, 3–4 (*see also* autocorrelated data); spatial, 3; Spearman’s coefficient of, 143, 285; study design and, 57–58; time series and, 191, 193, 196n27. *See also* collinearity; covariance
- corridor of stability: 78–82, 83. *See also* correlation
- covariance: advanced simulations and, 409, 412; censored observations and, 363–66; covariance matrix, 243, 272, 346, 355, 363–64, 366, 409; eddy, 7–8; evaluation and, 272, 293, 378, 390; meta-analysis and, 343, 346, 355; regression analysis and, 243–50; variance-covariance matrix, 33n22, 34, 243n17, 248–50, 293
- COVID-19 pandemic, 198
- Cramér-von Mises test, 106
- Crime Survey for England and Wales (CSEW), 6–7
- criterion analysis, 328
- Daszkiewicz, M., 106–8
- data gaps, 7–8, 215, 342n3. *See also* imputation; missing data
- data-generating model (DGM): advanced simulations and, 407; meta-analysis and, 345; power of statistical analysis and, 80; predictor variables and, 236; simulation

- data-generating model (DGM) (*continued*)
 - basics and, 4, 13, 17, 43; study designs and, 50, 69
- data mining, 68. *See also* questionable research practices (QRPs)
- data perturbation: bootstrapping and, 39–40, 43, 416; evaluation and, 389–90, 394, 397; randomness and, 40–43; R functions and, 416; sample size and, 40
- data snooping, 259. *See also* questionable research practices (QRPs)
- data wrangling, 10, 238, 421
- Dawkins, R., 8
- Dechêne, A., 48, 51, 55, 61–62
- DeepMind, 304
- degrees of freedom: ANOVA and, 398;
 - assumptions of statistical analysis and, 116–17, 148; censored observations and, 230–31; Efron's effective, 397n13; errors and, 254n29, 261; evaluation and, 272, 275, 278, 291–92, 296, 389–97; generalised (GDF), 389–97; missing data and, 217–19; model selection and, 155, 157–58, 160; Monte Carlo simulation and, 183, 185–86; parameter values and, 389–97; power of statistical analysis and, 90; predictor variables and, 176, 238–39; regression analysis and, 244–45; retrospective power analysis and, 314n7, 319–20, 330, 332; similarity to other approaches, 394–97; simulation basics and, 23, 32, 36–39; study designs and, 58n10, 67; testing, 390–94
- detrending: before modelling, 195–200;
 - change-point methods and, 192, 194–98; differencing and, 191–92, 193; errors and, 245; time series and, 154, 191–99, 205, 245
- diabetes, 213–22, 295
- diagnostics: assumption violations and, 105;
 - evaluation and, 267 (*see also* evaluation); generalised, 282–87; markers for, 235; model fit and, 275, 281–82; observed deviations and, 149 simulated random effect and, 135n24
- differencing, 191–92, 193
- dimensional analysis, 379–80, 398
- domain scientists, xi
- dormouse monitoring, 96–98
- dredging, 68, 154–55, 158, 235n13. *See also* questionable research practices (QRPs)
- Dunlap, W. P., 314–15
- echo chambers, 47–48
- eddy-covariance technique, 7–8
- Edwards, W., 318–19
- effect size: ANOVA and, 314; anticipating, 57–58;
 - bias in, 314; categorisation and, 62n17; complete data analysis and, 360–68; errors and, 58–59, 247, 248; estimating, 57–58; expressions of, 57; federated analysis and, 344, 360–61, 369, 371; hypothesis testing and, 58–59; meta-analysis and, 342, 345–58, 370; post hoc analysis and, 328–29; power of statistical analysis and, 73–76, 83, 95, 100–1; retrospective power analysis and, 311–34; simulation basics and, 13; standard deviation and, 24, 57–62, 76, 316, 330n27, 352; study designs and, 47, 49, 53, 57–62, 67, 69
- ELISA, 228n11
- Elman network, 383–85
- empirical cumulative density (ECD) plot, 284, 285
- entropy, 202, 386–87
- errors: adjusting tests and, 124–26; advanced simulations and, 404; AIC and, 254–55; assumptions of statistical analysis and, 103, 105–13, 116–17, 122, 123–28, 138, 147, 149; autocorrelated data and, 252; Bayesian models and, 246, 249–51; Benjamini-Hochberg procedure and, 126–30; bias and, 245–62; bootstrapping and, 245–46, 251–59, 260; censored observations and, 228n11, 230; collinearity and, 252, 255, 258; confidence interval (CI) and, 249; correlation and, 248, 249, 252–53, 257, 260; data entry, 260n34; degrees of freedom and, 254n29, 261; detrending and, 245; effect size and, 58–59, 247, 248; evaluation and, 272–78, 287–88, 291–92, 294–96, 299, 302, 304, 377–81, 392; false discovery rate (FDR), 126–31; false negatives, 21, 57–59, 66, 73, 176, 228n11, 311, 318–24; false nulls, 69, 128–31; false positives, 34–35, 57–60, 67, 105, 123–24, 129n18, 145, 154, 159–65, 228n11, 320, 322; family-wise, 123–28; federated analysis and, 252n23; Gaussian distribution and, 245n18, 249; hallucinations, 9–10, 14; HARKing and, 67–68; hypothesis testing and, 58–59; imputation and, 253, 255, 258; large language models (LLMs) and, 9–10; M, 159–63, 165, 331; machine learning and, 4, 33, 111, 202–4, 246, 249, 287, 294–95, 299n19, 302; matrices and, 246–50, 251n22, 255; mean squared, 116–17, 381; meta-analysis and, 350; missing data and, 217–20; model selection and, 154–71; model structure and, 256, 258; Monte Carlo simulation and, 180–86, 189, 249–51; multiple statistical steps and, 251–59; noise and, 245; normal distribution and, 246–47, 249n20; power of statistical analysis and, 73–74, 88, 90, 92–93; predictor variables and, 110–12, 174–80, 233, 238–39; probability and, 245, 249n20, 254; propagation of, 4, 103, 181n13, 211, 245–51, 378; questionable research practices (QRPs) and, 67–69; randomness and, 247–49, 253–55; recipes and, 252; regression analysis and, 110–12, 243–45; replicates and, 247, 256–57, 261; residual, 4, 33, 36, 38, 88, 90, 116–17, 176, 183–86, 204, 230, 244–48, 252, 261, 272, 275, 278, 287–88, 291–92, 392; retrospective power analysis and, 311–24, 327, 328–34; R functions and, 413, 416, 421; rules-of-thumb and, 153–71, 174–86, 189,

- 192, 202–7; S, 159–62, 165, 331; sample size and, 247, 248, 259–60; seeds and, 246–47, 250, 253–56, 261–62; simulated data and, 252, 254; simulation basics and, 3–6, 10, 21–38; stepwise model and, 254–55; study designs and, 49, 57n9, 58–62, 66–67; time series and, 192; unbiased data and, 252, 257; uncertainty and, 245–60; variance-covariance matrix and, 33n22, 34
- European Centre for Medium-range Weather Forecasting (ECMWF), 403–4
- evaluation, 12; AIC and, 296, 389–97; analysis of variance (ANOVA) and, 398; artificial intelligence (AI) and, 307; autocorrelated data and, 281; Bayesian models and, 269–72, 274, 283, 306n23, 387–89, 397; benchmark datasets and, 380–81; bias and, 273, 277, 375–77, 381–82; binomial distribution and, 286, 295, 297; bootstrapping and, 273, 275–77, 288, 294n16, 299–303, 304, 306; chemistry and, 276; classification and, 288, 297, 302–3, 397; code verification and, 377; collinearity and, 398; complex effects and, 276–81; conditional plots and, 276–81; confidence interval (CI) and, 270, 286, 288, 290–307; Cook's distance and, 281, 282; correlation and, 270–71, 285, 378; covariance and, 378, 390; coverage comparison and, 301–2; data perturbation and, 389–90, 394, 397; degrees of freedom and, 272, 275, 278, 291–92, 296, 389–97; DHARMA and, 387–88; didactic example of, 271–73; dimensional analysis and, 379–80, 398; empirical cumulative density (ECD) plot and, 284, 285; errors and, 272–78, 287–88, 291–92, 294–96, 299, 302, 304, 377–81, 392; frequentist analysis and, 306n23, 307, 387; *F*-statistic and, 272, 278, 292, 392; functional verification and, 377–78; Gaussian distribution and, 273, 281–83, 294–95, 297, 307, 378, 382n5, 387; general linear models (GLMs) and, 295–96, 299, 384, 387–92, 396–97; homoscedasticity and, 281; information gleaned from model, 273–77; intellectual advancement and, 382; intuitive understanding and, 339, 382–89; iterative learning and, 305–6; Kolmogorov-Smirnov (K-S) test and, 286–87, 289; learning from prior, 270–73; likelihood and, 269, 273, 288, 307, 386, 389, 392–97; limits of, 397–98; machine learning and, 270, 276, 280, 283–84, 287–88, 294–305, 307, 380, 387, 389–90, 397; manipulation and, 307; marginal plots and, 276–81; matrices and, 273, 291, 293, 299, 376–77, 389–90; method comparison and, 381–82; model diagnostics and, 281–88; model-fitting methods and, 267; model structure and, 273, 301; Monte Carlo simulation and, 270, 292–96; neural networks and, 295n18, 302–4, 383–84, 389–96; noise and, 284, 288–91, 295, 299–301, 375, 389–91, 394–95; non-Gaussian data and, 294–95; normal distribution and, 287n10; normality and, 281, 283; parameter values and, 267–76, 282n6, 294–97, 299n19, 301–2, 306, 384, 389–97; partial plots and, 276–81; peer review and, 68, 375; prediction interval and, 270, 288, 290–98, 302–7; predictor variables and, 288–305; probability and, 270–71, 274, 285n9, 286, 288n11, 289, 290n13, 295, 297, 301, 305, 377; quantile residuals and, 286–87; R&D, 375; randomness and, 272, 282n6, 288–90, 293, 295–96, 299–304, 377–78, 389–98; recurrent neural network (RNN) and, 383–85; replicates and, 271, 277, 284, 288–89, 300, 303–5, 306, 391, 394–95; residuals and, 267, 272, 275, 278–96, 283–88, 307, 383, 387–88, 392; robustness and, 376–77; sample size and, 302, 380, 398; seeds and, 272, 277–78, 284, 289, 296–99, 303, 305, 378, 391–92, 395; Shannon diversity and, 386; signals and, 375n2; simulated data and, 270, 282–84, 288, 299, 301, 380–82, 391; standard deviation and, 285, 301, 377–79, 384; stopping rules and, 378; testing data and, 275; training data and, 275; unbiased data and, 277, 381; uncertainty and, 269–75, 277, 288–91, 294–96, 297, 302, 305–7, 379; unit testing and, 376–79
- fake data, 4n1, 47–48, 404. *See also* simulated data
- false discovery rate (FDR), 126–31
- false negatives: censored observations and, 228n11; errors and, 21, 57–59, 66, 73, 176, 228n11, 311, 318–24; power of statistical analysis and, 73; predictor variables and, 176; retrospective power analysis and, 311, 318–24; study designs and, 57–59, 66. *See also* type-II errors
- false nulls, 69, 128–31
- false positives: assumptions of statistical analysis and, 105, 123–31, 145; avoiding, 35; censored observations and, 228n11; critical level of, 58; errors and, 34–35, 57–60, 67, 105, 123–24, 129n18, 145, 154, 159–65, 228n11, 320, 322; inflation of, 105, 123–31; model selection and, 154, 159–64, 165; multiple testing and, 123–31; retrospective power analysis and, 320, 322; study designs and, 57–60, 67. *See also* type-I errors
- Faul, F., 327–29, 330n27, 333
- Fazio, L. K.: bootstrapping and, 64; illusory truth effect and, 48–52, 60–65; plausibility and, 61; 68; study designs and, 48–55, 60–65, 66n22, 68
- federated analysis: bias and, 341–45, 357–58, 361, 362, 368–71; common model identification and, 354; complete data analysis and, 341–45, 361, 362–63, 368; correlation and, 358; data distribution and, 354; effect size and, 344, 360–61, 369, 371; embargoed datasets and, 370; errors and, 252n23; matrices and, 359; meta-analysis and, 12,

- federated analysis (*continued*)
 - 339, 341–45, 369; model application and, 354–57; multi-step, 358–60, 368, 370; one-step, 354–58, 361, 370; parameter estimates and, 343, 360, 368–69, 371; predictor variables and, 341–45, 354, 357–62, 368–69; replicates and, 359–60; seeds and, 359; two-step, 353, 363
- Fibonacci sequence, 377
- Fieberg, J. R., 10
- fire ants, 142–49
- fitted-model methods, 28–29
- Fitzpatrick, M. C., 142
- fixed effects: assumptions of statistical analysis and, 133–34, 138; randomness and, 132–33, 153, 351, 356; rules-of-thumb and, 153
- FLUXNET, 7–8
- fMRI imaging, 3, 375
- for-loops, 410–13
- Freckleton, R. P., 243
- frequency distribution, 20, 50
- frequentist analysis: assumptions of statistical analysis and, 103; Bayesian models and, 10, 103, 182, 306n23, 307, 312, 387; evaluation and, 306n23, 307, 387; Monte Carlo simulation and, 182, 188; retrospective power analysis and, 312; simulation basics and, xii, 10; study designs and, 49
- F*-statistic: ANOVA and, 20–21, 116; assumptions of statistical analysis and, 116; censored observations and, 230; evaluation and, 272, 278, 292, 392; homoscedasticity and, 116; Monte Carlo simulation and, 183, 185–86; regression analysis and, 244–45
- functional verification, 377–78. *See also* code verification; evaluation; unit testing
- F*-values, 20–21, 23, 117, 122
- Galileo, 341
- gamma-ray bursts, 37–42
- Gaussian distribution: advanced simulations and, 405; assumptions of statistical analysis and, 110–14, 131–32, 137n25; complete data analysis and, 231; errors and, 245n18, 249; evaluation and, 273, 281–83, 294–95, 297, 307, 378, 382n5, 387; machine learning and, 202; meta-analysis and, 346; Monte Carlo simulation and, 185, 249; noise and, 113, 114, 185, 346; power of statistical analysis and, 74n3, 92; random, 22–23, 51, 52n5, 74n3, 202, 378, 409; R functions and, 409; simulation basics and, 23, 29–33; standard deviation and, 23; study designs and, 51, 52n5
- Gelman, A., 327–34
- general additive mixed model (GAMM), 88
- generalized least squares (GLS), 34–36
- general linear models (GLMs): assumptions of statistical analysis and, 103, 145, 147; bootstrapping and, 30; evaluation and, 295–96, 299, 384, 387–92, 396–97; hypothesis testing and, 3, 10, 14; machine learning and, 203; meta-analysis and, 344; missing data and, 217–25; model selection and, 155–56, 158, 160, 163, 166; Monte Carlo simulation and, 29; predictor variables and, 3, 174, 176, 178, 232–41; principal component analysis (PCA) and, 3; R functions and, 418; simulation basics and, xii, 3, 29–30; threshold effect and, 103; uncertainty in, 30
- generative adversarial networks (GANs), 8–9
- Gijbels, I., 224n8
- González-Moreno, P., 142
- Goodman, S. N., 314–16, 321–22
- Google, 304
- Gould, E., 331–32
- G*Power, 327–29, 330n27
- GraphCast model, 304
- hallucinations, 9–10, 14
- happiness, 127
- Hardin, J., 10
- HARKing (Hypothesizing After the Results are Known), 67–68. *See also* questionable research practices (QRPs)
- Hartig, F., 387
- Heisey, D. M., 316, 323
- Hessian matrices, 28
- heteroscedasticity: ANOVA and, 20–21, 118–22, 154; assumptions of statistical analysis and, 116, 118, 121–23, 145; Monte Carlo simulation and, 180–90, 205; normality and, 116; post hoc analysis and, 123, 145
- hierarchical data: assumptions of statistical analysis and, 133–41; multiple analysis of, 134, 139–41
- Hoening, J. M., 316, 323
- homoscedasticity: ANOVA and, 13, 105–6, 116–23; assumptions of statistical analysis and, 13; evaluation and, 281; *F*-statistic and, 116; mixed-effect models and, 118; normality and, 9, 116, 281; testing effects of, 118–23
- hypothesis testing: advanced simulations and, 406; assumptions of statistical analysis and, 103; compromise power analysis and, 328; confident interval (CI) and, 334; errors in, 58–59; GLM and, 3, 10, 14; machine learning and, 203; PCA and, 3; reliable, 59; replicates for, 59; retrospective power analysis and, 328, 334; study designs and, 58–60. *See also* type-I errors; type-II errors
- if-then-else statements, 410, 421
- illusory truth effect: adequate replication and, 56–60; Fazio and, 48–52, 60–65; frequency distribution and, 50; hypotheses definitions and, 49–55; null expectation and, 52, 55; perceived accuracy and, 51–55, 56; prior expectations and, 331–32; processing fluency and, 54–55, 62; quantifying variables of interest, 50–55; reach of, 48; repeated statements and, 48; sample size and, 60–66;

- study designs and, 48–52, 55–56, 60–66;
testing hypotheses and, 55–67
- immigration, 110–11, 113, 332, 406
- imputation: censored observations and, 224, 225,
232; errors and, 253, 255, 258; issues with,
214–15; missing data and, 13, 103, 155n5,
156, 160, 175, 211–22, 224, 225, 232–34,
253–55, 258; model selection and, 155n5,
156, 160; predictor variables and, 175,
213–22, 232–34; qualitative comparison
and, 215–19; retrospective power analysis
and, 311; simulation basics and, 13. *See also*
data gaps; missing data
- individual participant-level (IPL) meta-analysis,
352–55, 356, 360, 368. *See also* meta-analysis
- interocular traumatic test, 318–19intuitive
understanding: evaluation and, 339, 382–89;
Monte Carlo simulation and, 187n18;
multiple descriptions and, 383–86; recurrent
neural network (RNN) and, 383–85; result
presentation and, 386–87
- invented data, 4n1. *See also* fake data; simulated
data
- iterative learning, 305–6
- jackknife, 17, 299–303, 304. *See also* bootstrapping
- James Webb Space Telescope, 73n1
- Jena Experiment, 113–14, 119, 120
- Julia, xii, 10
- Keeling, C. D., 85
- Kolmogorov-Smirnov (K-S) test: assumptions of
statistical analysis and, 106, 135n24;
evaluation and, 286–87, 289; model
variability and, 28; normality and, 106;
quantile residuals and, 286–87; robustness
of, 26–28
- Kreyling, J., 179
- Kriegeskorte, N., 3
- Kruskal-Wallis test, 120
- Lane, D. M., 314–15
- large language models (LLMs), 9, 14
- Laser Interferometer Gravitational-Wave
Observatory (LIGO), 404–6
- Law of Small Numbers, 81
- left-censored data, 222, 226–27, 230
- Lev, J., 313
- LIDAR, 406
- likelihood: bootstrapping and, 39–40, 43, 204, 273;
censored observations and, 229; curvature
of, 28; evaluation and, 269, 273, 288, 307,
386, 389, 392–97; log, 38–41, 202, 204, 273,
288, 389, 392–97; machine learning and,
202, 204; maximum, 28, 190, 314; Monte
Carlo simulation and, 187, 189; null
hypothesis and, 39, 41; retrospective power
analysis and, 314, 331; simulation basics
and, 10, 14; study designs and, 45, 50; time
series and, 190
- likelihood-ratio test (LRT), 38–40, 43
- Lilliefors test, 26–28, 106
- Llama, xii, 9
- lmodel2, 111
- Locally Interpretable Model-agnostic Explanations
(LIME), 307
- logical operators, 421
- log-normal distribution, 31–32, 139n27
- Lyell, C., 138n26
- machine learning: artificial intelligence (AI) and,
201, 205; assumptions of statistical analysis
and, 103, 105, 111; autocorrelated data and,
203–4; Bayesian models and, 203n33; bias
and, 201–3; Big Data and, 153, 200–4;
bootstrapping and, 204; censored
observations and, 226, 232; chemistry and,
276; classical statistics and, 13; classification
and, 202; confidence interval (CI) and,
203–4, 295–305; correlation and, 203–4;
distributional assumptions and, 201–3;
errors and, 4, 33, 111, 202–4, 246, 249, 287,
294–95, 299n19, 302; evaluation and, 270,
276, 280, 283–84, 287–88, 294–305, 307,
380, 387, 389–90, 397; extrapolation and,
203–4; Gaussian distribution and, 202;
general linear models (GLMs) and, 203;
hypothesis testing and, 203; likelihood and,
202, 204; missing data and, 212, 214; model
selection and, 165–72; non-independence
and, 201–3; non-linear functions and, 203;
predictor variables and, 165–72, 203–4, 232,
238; probability and, 202; *p*-values and, 203;
randomness and, 201–4; regression analysis
and, 165–72; residuals and, 204; R functions
and, 419; robustness and, 201–2;
rules-of-thumb and, 153–54, 165–720,
200–5; sample size and, 202; simulation
basics and, 17, 33; unbiased data and, 201;
uncertainty and, 203; workflow and, 4
- Mair, M. M., 324–25
- manipulation: advanced simulations and, 404;
assumptions of statistical analysis and,
112–16; categorisation and, 177–80; combi-
nations of, 9; evaluation and, 307; predictor
variables and, 112–16, 177; simulation
basics and, 9; study designs and, 45, 49
- “Many Analysts, One Data Set” 331–32
- marginal plots, 276–80
- Markov chain Monte Carlo (MCMC) simulation,
50n4, 249–50, 389
- Mathematica, 10
- mathematicians, 4, 6, 18, 29n17
- Matlab, 10, 377
- matrices: assumptions of statistical analysis and,
128–32; binary, 377; censored observations
and, 223; complete data analysis and,
362–66; covariance, 243, 272, 346, 355,
363–64, 366, 409; errors and, 246–50,
251n22, 255; evaluation and, 273, 291, 293,
299, 376–77, 389–90; federated analysis and,
359; hat, 389–90; Hessian, 28; meta-analysis
and, 346–50, 355, 357; missing data and,
220–21; model selection and, 161–63, 166;

- matrices (*continued*)
 - power of statistical analysis and, 79–88, 95;
 - predictor variables and, 175, 236, 240–41;
 - random effects and, 132; regression analysis and, 243–44; R functions and, 409, 411–14;
 - study designs and, 61, 66; time series and, 198; variance-covariance, 33n22, 34, 243n17, 248–50, 293; workflow and, 212n1, 220–23, 236, 240–50, 251n22, 255
- mean squared errors, 116–17
- mean squared prediction error (MSPE), 381
- median retrospective power (MRP), 350–52
- M-error: retrospective power analysis and, 329–31; rules-of-thumb and, 159–60, 162–63, 165
- meta-analysis: autocorrelated data and, 343; bias and, 341–58, 369–370; Big Data and, 341; classical, 341–45, 352–57, 360, 368, 370; complete data analysis and, 341–45; correlation and, 343, 346; covariance and, 343, 346, 355; data distribution and, 347; data-generating model (DGM) and, 345; data sources for, 341–42; distributed studies and, 345–46; effect size and, 342, 345–58, 370; errors and, 350; federated analysis and, 12, 339, 341–45, 369; Gaussian distribution and, 346; general linear models (GLMs) and, 344; improved, 343, 345, 370; individual participant-level (IPL), 352–55, 356, 360, 368; individual studies and, 347–48; large dataset and, 346–47; matrices and, 346–50, 355, 357; median retrospective power (MRP) and, 350–52; mixed-effect models and, 344, 353; model structure and, 344, 368; noise and, 346; normal distribution and, 346; parameter values and, 343, 352; post hoc analysis and, 350, 352; predictor variables and, 343–60; privacy and, 368; probability and, 350, 352; *p*-values and, 345; randomness and, 347–50, 351, 353–56; repeated simulations and, 348–50; replicates and, 344, 355; retrospective power analysis and, 350–52; robustness and, 341, 344, 369; sample mean and, 342; sample size and, 342–49, 352, 357, 370; seeds and, 346, 349, 355–56; signals and, 342n3; simulated data and, 346, 354, 369; standard deviation and, 352; stepwise model selection and, 345, 351, 354; unbiased data and, 342, 343, 347n11, 352, 354, 370
- Michaelis-Menten equation, 114–15, 179, 380
- minimum detectable differences (MDDs), 322–26, 333
- minimum detectable effects (MDEs), 322–26, 333
- missing at random (MAR) data, 213, 215, 222, 225, 226. *See also* data gaps; imputation; missing data
- missing completely at random (MCAR) data, 213, 215, 220–23, 224, 253. *See also* data gaps; imputation; missing data
- missing data: AIC and, 217–19; bias and, 212–22; confidence interval (CI) and, 220; correlation and, 214–15, 222; degrees of freedom and, 217–19; errors and, 217–20; general linear models (GLMs) and, 217–25; imputation of, 13, 103, 155n5, 156, 160, 175, 211–22, 224, 225, 232–34, 253–55, 258; machine learning and, 212, 214; matrices and, 220–21; missing at random (MAR), 213, 215, 222, 225, 226; missing completely at random (MCAR), 213, 215, 220–23, 224, 253; missing not at random (MNAR), 213, 222; parameter values and, 213, 217–22; predictor variables and, 213–19; probability and, 212–13, 222; *p*-values and, 215; qualitative comparison and, 215–19; random values and, 215; residuals and, 217–19; sample size and, 212, 219–20; seeds and, 220; types of, 212–13; unbiased data and, 213n2, 219
- missing not at random (MNAR) data, 213, 222. *See also* data gaps; imputation; missing data
- mixed-effect models: assumptions of statistical analysis and, 131–37; homoscedasticity and, 118; hyper-distributions and, 131–37; linear (LMMs), 131–34; meta-analysis and, 344, 353; randomness and, 153; rules-of-thumb and, 153
- model-fitting methods: conditional plots and, 276–81; Cook's distance and, 281, 282; diagnostics and, 281–88; evaluation and, 267–77 (*see also* evaluation); GLMs and, 10 (*see also* general linear models (GLMs)); marginal plots and, 276–81; obscuring of, 277–78; partial plots and, 276–81; robustness and, 39; subset regression and, 154; *t*-test and, 10, 14, 24, 29–32, 57n9, 59–62, 74–77, 103, 120, 124, 127, 138, 324, 418–21; uncertainty and, 28–30; workflow and, 211; χ^2 statistics, 10, 14
- model selection: AIC and, 155, 157–59; bias and, 154n2, 159–60, 171; collinearity and, 165–72; correlation and, 164–67; degrees of freedom and, 155, 157–58, 160; dredging and, 154–55, 158; effect size and, 154, 159–64, 165; errors and, 154–71; false positives and, 154, 159–64, 165; general linear models (GLMs) and, 155–56, 158, 160, 163, 166; improved interpretation and, 154–64; imputation and, 155n5, 156, 160; interpretability and, 164; machine learning and, 165–72; matrices and, 161–63, 166; model structure and, 155–60; Monte Carlo simulation and, 154; parameter values and, 157, 159–60, 165–66, 171–72, 371; predictor variables and, 165–72; principal component analysis (PCA) and, 165n7; probability and, 155, 164, 171; *p*-values and, 160, 163; randomness and, 166–72; replicates and, 160, 163, 170; residuals and, 155–58, 168–69; sample size and, 164; seeds and, 160, 167, 170; simulated data and, 165–67, 172; stepwise, 154–55, 157, 159–60, 164–65, 236, 254–55, 345, 351, 354, 362; structural correctness and, 155–60; true positives and,

- 159–64, 165; unbiased data and, 159, 171; uncertainty and, 171
- model structure: assumptions of statistical analysis and, 133, 149; censored observations and, 232; definition of, 159; errors and, 256, 258; evaluation and, 273, 301; meta-analysis and, 344, 368; model selection and, 155–60; predictor variables and, 240
- Monte Carlo simulation: autocorrelated data and, 181, 183; Bayesian models and, 182, 187–89; bootstrapping and, 181n14; breakpoint and, 182n15, 189; carbon dioxide (CO₂) and, 181–90; classification and, 29n17; confidence interval (CI) and, 182, 293; correlation and, 181, 183; degrees of freedom and, 183, 185–86; errors and, 180–86, 189, 249–51; evaluation and, 270, 292–96; fitted-model methods and, 29; frequentist analysis and, 182, 188; *F*-statistic and, 183, 185–86; Gaussian distribution and, 185, 249; general linear models (GLMs) and, 29; heteroscedasticity and, 180–90, 205; intuitive understanding and, 187n18; likelihood and, 187, 189; Markov chain (MCMC), 50n4, 249–50, 389; model selection and, 154; noise and, 181–86; normal distribution and, 184, 187; normality and, 182–83; null hypothesis and, 184; observed trends and, 184–87; parameter values and, 181n13, 190; plausibility and, 188–89; probability and, 181, 184, 187; randomness and, 181, 212–22; reference distribution and, 184–87; regression analysis and, 182–84; replicates and, 181, 185–86; residuals and, 183–86; sample size and, 180; seeds and, 185–86; simulated data and, 184–86; standard deviation and, 184, 187; time series and, 182–84; uncertainty and, 181–84, 188
- Mueller, D. K., 228n11
- multiple testing: HARKing and, 67; inflation of false positives, 123–31
- Nakagawa, S., 377
- NASA (National Aeronautics and Space Administration), 342n3
- NEON (National Ecological Observatory Network), 406
- Neptune, 98
- “network of chums” effect, 344
- neural networks: advanced simulations and, 406; convolutional (CNN), 8; data augmentation and, 8–9; evaluation and, 295n18, 302–4, 383–84, 389–96; generative adversarial (GAN), 8–9; recurrent (RNN), 383, 384
- neurobiology, xii, 1
- Neyman, J., 22n6, 313
- niche shifts: assumptions of statistical analysis and, 142–49; demonstrating spurious, 143–44; species invasion and, 144–45
- Nightingale, F., 341
- noise: advanced simulations and, 404–7; assumptions of statistical analysis and, 108–15; autocorrelated data and, 34, 86–87, 91; convolutional neural networks (CNNs) and, 8; errors and, 245; evaluation and, 284, 288–91, 295, 299–301, 375, 389–91, 394–95; generative adversarial networks (GANs) and, 8–9; meta-analysis and, 346; Monte Carlo simulation and, 181–86; power of statistical analysis and, 73, 86–87, 91–95, 100–1; random, 40–41, 100, 301; signals and, 5, 87, 405; study designs and, 49; time series and, 197, 198
- normal distribution: assumptions of statistical analysis and, 106–9, 116, 131–32, 135, 137, 140, 142; Central Limit Theorem and, 131; errors and, 246–47, 249n20; evaluation and, 287n10; Gaussian, 22 (*see also* Gaussian distribution); log, 31–32, 139n27; meta-analysis and, 346; Monte Carlo simulation and, 184, 187; power of statistical analysis and, 74–78; random values and, 22–23; regression analysis and, 243–44; retrospective power analysis and, 315, 317, 330; R functions and, 412; simulation basics and, 22–33; study designs and, 53
- normality: ANOVA and, 18; assumptions of statistical analysis and, 105–9, 116, 132–37, 142, 145, 148; Cook’s distance and, 281, 282; evaluation and, 281, 283; heteroscedasticity and, 116; homoscedasticity and, 9, 116, 281; K-S test and, 26–28; Monte Carlo simulation and, 182–83; OLS and, 182; testing, 135–37; testing for, 106–9; *t*-test and, 29
- notational conventions, 4, 11–12, 143, 382n5
- NSF (National Science Foundation), 406
- null expectation, 52, 55
- null hypothesis: ANOVA and, 118; assumptions of statistical analysis and, 118, 126–28; false discovery rate (FDR) and, 126–31; likelihood and, 39, 41; Monte Carlo simulation and, 184; randomness and, 39; retrospective power analysis and, 311, 313, 323, 327, 329, 332; study designs and, 56–61, 69
- null models, 9, 56, 64
- Nutrient Network, 346n9
- OpenIntro project, 10
- ordinary least squares (OLS): normality and, 182; regression analysis and, 34–36, 111, 115n12, 254–55, 285; residuals and, 36, 285
- parameter values: advanced simulations and, 404, 419; assumptions of statistical analysis and, 103, 106, 107n2, 110, 114–16, 119, 131–35, 139–40, 142, 145, 148; bias and, 4–5, 69, 81, 115, 145, 159, 171–73, 211, 213, 222, 226, 232, 235, 242, 252–59, 362, 368–69, 371; censored observations and, 226, 232; complete data analysis and, 360, 362, 368; degrees of freedom and, 389–97; dispersion, 7, 148, 157, 217–19, 238–39, 296; evaluation and, 267–76, 282n6, 294–97, 299n19, 301–2, 306, 384, 389–97; meta-analysis and, 343,

- parameter values (*continued*)
 - 352; missing data and, 213, 217–22; model selection and, 157, 159–60, 165–66, 171–72, 371; Monte Carlo simulation and, 181n13, 190; notational conventions for, 11; population-level, 11; power of statistical analysis and, 74n3, 79–87, 90–93; predictor variables and, 173, 176, 234–42; regression, 7, 110, 142, 301; regression analysis and, 243–46, 249–61; retrospective power analysis and, 315, 319–21, 325, 329, 332, 334; simulation basics and, 4–5, 7, 11–12, 18–21, 24–31, 33n22, 35–36, 40–42; study designs and, 49–50, 55n7, 56, 60–61, 68–69, 369; time series and, 191; workflow and, 211, 213, 217–19, 220, 222, 226, 232–46, 249–61
- partial plots, 276–80
- Pearson, E. S., 313
- Pearson's correlation coefficient, 34, 74, 76
- peer review, 68, 375
- perceived accuracy, 51–55, 56
- Perugini, M., 77–79, 81, 83
- Pima Indian Diabetic Disease dataset, 214
- pipelines: advanced simulations and, 403–5; artefacts and, 211, 228n11, 245, 251–58; simulation basics and, 8, 12; workflow and, 8, 12, 211, 228n11, 245, 251–58, 403–5
- Planet 9, 98–99
- plausibility: Fazio and, 61; Monte Carlo simulation and, 188–89; study designs and, 50–56, 61–62, 66
- Pluto, 98n20
- pollinators, 9
- Pólya, G., 18–20
- post hoc analysis: assumptions of statistical analysis and, 118, 123, 145; compromise power analysis, 328; confidence intervals (CI) and, 319–22; criterion analysis and, 328; design analysis and, 329–32; effect size and, 328–29; federated analysis and, 341–45, 354, 357–62, 368–69; G*Power and, 327–29, 330n27; heteroscedasticity and, 123, 145; interocular traumatic test and, 318–19; meta-analysis and, 350, 352; minimum detectable differences (MDDs) and, 322–26, 333; minimum detectable effects (MDEs) and, 322–26, 333; need for, 318–19; predictor variables and, 179, 320–22; retrospective power analysis and, 312–15, 318–34; sensitivity analysis, 328; simulation basics and, 13, 13; study designs and, 67, 69
- posterior probability distribution, 271, 305, 312, 334. *See also* Bayesian models
- power of statistical analysis: autocorrelated data and, 86–92; Bayesian models and, 93, 98; bias and, 74, 78–85, 88, 89, 97, 101; bootstrapping and, 78–79, 81, 83; carbon dioxide (CO₂) and, 85–87, 89n15, 90–96, 99n22; confidence interval (CI) and, 78, 80, 81, 83; correlation and, 74–82, 83, 86–88, 91–95, 101; data-generating model (DGM) and, 80; data points for correlation, 76–80; degrees of freedom and, 90; effect size and, 73–76, 83, 95, 100–1; errors and, 73–74, 88, 90, 92–93; false negatives and, 73; Gaussian distribution and, 74n3, 92; impact assessment and, 99–100; improving estimates and, 91–100; matrices and, 79–88, 95; monitoring atmospheric carbon dioxide (CO₂) and, 85–95, 96, 99n22; noise and, 73, 86–87, 91–95, 100–1; normal distribution and, 74–78; parameter values and, 74n3, 79–87, 90–93; probability and, 73–74, 84, 91, 97; puerperal fever deaths study and, 81–84; *p*-values and, 81, 84, 95, 96; randomness and, 74–76, 100; “recruit until significant” and, 80–85; reliability and, 78; replicates and, 77, 80, 82, 86–89, 94–95; residuals and, 87–92; R functions and, 75; sample size and, 73–88, 98, 101; seeds and, 76–77, 80–95; signals and, 87; simple group comparisons and, 74–76; simulated data and, 76, 83n8, 87; stability and, 78; standard deviation and, 74–76, 96; stopping rules and, 74, 81–85, 87, 89, 95, 101; time series and, 85–91; *t*-test and, 74–77; type-II errors and, 73–74; unbiased data and, 74, 79, 83–85, 97; uncertainty and, 78, 91–93. *See also* type-II errors
- prediction interval: confidence interval (CI) and, 203–4, 270, 288–307; evaluation and, 270, 288, 290–98, 302–7; retrospective power analysis and, 321n19
- predictor variables: AIC and, 238–39; analysis of variance (ANOVA) and, 179; assumptions of statistical analysis and, 137–45; Bayesian models and, 232; bias and, 171–73, 232–42; carbon dioxide (CO₂) and, 177; categorisation of, 154, 172–80, 205; collinearity and, 238; complete data analysis and, 360–68; confidence interval (CI) and, 174, 288–305, 320–22; correlations among, 137–46, 149, 235–42; data-generating model (DGM) and, 236; degrees of freedom and, 176, 238–39; errors and, 110–12, 174–80, 233, 238–39; evaluation and, 288–305; false negatives and, 176; general linear models (GLMs) and, 3, 174, 176, 178, 232–41; imputation and, 175, 213–22, 232–34; machine learning and, 165–72, 203–4, 232, 238; manipulation and, 112–16, 177; matrices and, 175, 236, 240–41; mean squared prediction error (MSPE), 381; meta-analysis and, 343–60; missing data and, 213–19; model selection and, 165–72; model structure and, 240; non-Gaussian data and, 294–95; parameter values and, 173, 176, 234–42; post hoc analysis and, 179; pre-selecting, 232–42; randomness and, 236, 238–41, 242; random values and, 112; regression analysis and, 110–12; replicates and, 178–79, 234, 241; residuals and, 176, 238–39; retrospective power analysis and, 320–22; robustness and, 174, 238; sample size and, 236; seeds and, 178,

- 233–37, 241; simulated data and, 232, 239; stepwise model selection and, 236; true positives and, 180; unbiased data and, 172, 235, 238, 242
- pre-registration, 66–69, 267, 311, 314n8, 342n4, 358, 369
- principal component analysis (PCA): assumptions of statistical analysis and, 142, 144–46; general linear models (GLMs) and, 3; model selection and, 165n7; niche shifts and, 142, 144–46
- prior probability distribution, 22n6, 45, 50, 270, 313n5. *See also* Bayesian models
- probability: assumptions of statistical analysis and, 123–24, 126–27, 145; censored observations and, 225, 228n11; detection, 97; errors and, 245, 249n20, 254; evaluation and, 270–71, 274, 285n9, 286, 288n11, 289, 290n13, 295, 297, 301, 305, 377; improving estimation and, 91–100; likelihood and, 14 (*see also* likelihood); machine learning and, 202; meta-analysis and, 350, 352; missing data and, 212–13, 222; model selection and, 155, 164, 171; Monte Carlo simulation and, 181, 184, 187; Pólya urn and, 18–20; power of statistical analysis and, 73–74, 84, 91, 97; retrospective power analysis and, 311–22, 329–34; R functions and, 409; selection, 97; simulation basics and, 7, 10, 12, 14, 17, 18n1, 20, 22n6, 26, 29n17, 33n20, 37; study designs and, 45, 50, 56, 58, 66–67, 69; time series and, 190, 197
- processing fluency, 54–55, 62. *See also* illusory truth effect
- Programme for International Student Assessment (PISA), 132
- prospective power analysis, 67, 73, 312–13, 327
- puerperal fever deaths, 81–84, Semmelweis and, 81
- p*-values: assumptions of statistical analysis and, 107, 123, 126–27, 145, 148; machine learning and, 203; meta-analysis and, 345; missing data and, 215; model selection and, 160, 163; power of statistical analysis and, 81, 84, 95, 96; retrospective power analysis and, 34n23, 312–16, 318n14, 320–21, 333–34; simulation basics and, xii, 5, 20, 27–28; study designs and, 58n12, 67; workflow and, 211, 215. *See also* hypothesis testing; type-I errors;
- Python, xii, 10, 377, 380
- QA/QC (quality assurance and quality control) procedures, 8, 342
- Qucit, 295, 299–300, 302
- questionable research practices (QRPs), 67–69
- R&D (Research & Development), 375
- random effects: assumptions of statistical analysis and, 132–33, 135, 137–38, 141, 142; censored observations and, 229; matrices and, 132; rules-of-thumb and, 153
- random-forest models: assumptions of statistical analysis and, 145; errors and, 255; evaluation and, 288–90, 295n18, 299–304, 390–94; model selection and, 166–72; predictor variables and, 238–41
- randomness: assumptions of statistical analysis and, 105, 107, 112, 118, 131–38, 141–47; breakpoint and, 37–42; complete data analysis and, 222, 224, 225, 226, 229, 361, 363–64, 366; data perturbation and, 40–43; errors and, 247–49, 253–55; evaluation and, 272, 282n6, 288–90, 293, 295–96, 299–304, 377–78, 389–98; fixed effects and, 132–33, 153, 351, 356; Gaussian, 22–23, 51, 52n5, 74n3, 202, 378, 409; machine learning and, 201–4; meta-analysis and, 347–50, 351, 353–56; missing data and, 212–22; mixed-effect models and, 153; model selection and, 166–72; Monte Carlo simulation and, 181; noise and, 40–41, 100, 301; null hypothesis and, 39; power of statistical analysis and, 74–76, 100; predictor variables and, 236, 238–41, 242; resampling and, 35, 37, 40n28, 41, 43, 416–18; retrospective power analysis and, 311n1, 319, 321n20; R functions and, 409–11, 412, 416–20; seeds and, 4 (*see also* seeds); simulated data vs., 35–43; simulation basics and, 3–5, 8–11, 17–19, 22–24, 27–43; study designs and, 51–52, 55n8, 57, 61; uncertainty and, 29, 31, 39, 249, 254
- random values: Gaussian distribution and, 22–23; missing data and, 215; Monte Carlo simulation and, 181; multivariate, 409; predictor variables and, 112; R functions and, 409–10, 412, 418
- reality, 5, 58
- recipes, xi, 1, 153, 252
- recurrent neural network (RNN), 383–85
- registered reports, 68
- regression analysis: bias and, 171, 242–45; bootstrapping and, 29, 31, 40, 64–66, 300–2, 304, 416; confidence interval (CI) and, 243; correlation and, 33, 58, 106, 153, 164–65, 181, 183, 215, 243; covariance and, 243–50; degrees of freedom and, 244–45; detrending and, 190–200; dilution of, 110–15; errors and, 110–12, 243–45; *F*-statistic and, 244–45; generalized least squares (GLS), 34–36; machine learning and, 165–72; matrices and, 243–44; Monte Carlo simulation and, 182–84; multiple regression and, xii, 153, 212, 243–47, 248, 291, 347; normal distribution and, 243–44; ordinary least squares (OLS), 34–36, 111, 115n12, 254–55, 285; parameter values and, 243–46, 249–61; predictor variables and, 110–12; residuals and, 242–45; sample size and, 243; seeds and, 244; segmented, 37–41; subset, 154–55, 397n13; time series and, 182–84, 190–200; unbiased data and, 243
- replicate function, 416–17
- replicates: analysis of, 89–91; assumptions of statistical analysis and, 108–9, 112, 113, 124–30, 134, 139, 141, 146; censored

- replicates (*continued*)
 - observations and, 223, 225; complete data analysis and, 363, 366; errors and, 247, 256–57, 261; evaluation and, 271, 277, 284, 288–89, 300, 303–5, 306, 391, 394–95; federated analysis and, 359–60; meta-analysis and, 344, 355; model selection and, 160, 163, 170; Monte Carlo simulation and, 181, 185–86; power of statistical analysis and, 77, 80, 82, 86–89, 94–95; predictor variables and, 178–79, 234, 241; retrospective power analysis and, 319, 320n18; R functions and, 410, 416–21; simulation data and, 4, 19n2, 22–32, 40–43, 89–91; study designs and, 56–59, 65, 73; time series and, 194, 197–98
- resampling: bootstrapping and, 17, 40, 43, 416–18; jackknife and, 17, 299–303, 304; random, 35, 37, 40n28, 41, 43, 416–18; R functions and, 416–21; signals and, 40
- residuals: ANOVA and, 18, 116–17; assumptions of statistical analysis and, 3, 106–9, 116–17, 121, 132–35, 138–41, 145, 147–48; autocorrelated data and, 33, 87–88, 91–92, 183, 193, 204, 281; breakpoint and, 40, 42; censored observations and, 230; data perturbation and, 41; empirical cumulative density (ECD) plot and, 284, 285; errors and, 4, 33, 36, 38, 88, 90, 116–17, 176, 183–86, 204, 230, 244–48, 252, 261, 272, 275, 278, 287–88, 291–92, 392; evaluation and, 267, 272, 275, 278–96, 307, 383, 387–88, 392; machine learning and, 204; missing data and, 217–19; model selection and, 155–58, 168–69; Monte Carlo simulation and, 183–86; OLS, 36, 285; partial plots and, 279; power of statistical analysis and, 87–92; predictor variables and, 176, 238–39; quantile, 283–88; regression analysis and, 242–45; rules-of-thumb and, 155–58, 168–69, 176, 183–86, 191, 193, 198–99, 204; simulated data and, 4; skewed, 29–32; standard deviation and, 40, 117, 283, 285; time series and, 191, 193, 198–99; workflow and, 212, 217–19
- retrospective power analysis, 12; analysis of variance (ANOVA) and, 314–15, 316; Bayesian models and, 312, 319, 331, 334; bias and, 314, 318, 320n18, 329, 331; chemistry and, 318; compromise power analysis, 328; concept of, 313–18; confidence interval (CI) and, 311n1, 313, 319–26, 333–34; correlation and, 312; criterion analysis, 328; degrees of freedom and, 314n7, 319–20, 330, 332; design analysis and, 329–32; discredited standard approach and, 313; effect size and, 311–34; errors and, 311–24, 327, 328–34; false negatives and, 311, 318–24; false positives and, 320, 322; frequentist analysis and, 312; G*Power and, 327–29, 330n27; hypothesis testing and, 328, 334; imputation and, 311; interocular traumatic test and, 318–19; lack of recommendation of, 312; likelihood and, 314, 331; median retrospective power (MRP) and, 350–52; meta-analysis and, 350–52; minimum detectable differences (MDDs) and, 322–26, 333; minimum detectable effects (MDEs) and, 322–26, 333; normal distribution and, 315, 317, 330; null hypothesis and, 311, 313, 323, 327, 329, 332; parameter values and, 315, 319–21, 325, 329, 332, 334; post hoc analysis and, 312–15, 318–34; prediction interval and, 321n19; predictor variables and, 320–22; probability and, 311–22, 329–34; problems with, 312–18, 332–34; prospective power analysis and, 67, 312–13, 327; *p*-values and, 34n23, 312–16, 318n14, 320–21, 333–34; randomness and, 311n1, 319, 321n20; replicates and, 319, 320n18; sample mean and, 319; sample size and, 311–13, 316, 319–30; seeds and, 315, 332; sensitivity analysis, 328; signals and, 331n30; standard deviation and, 314n7, 315, 316, 319, 330n27, 332; true positives and, 324, 328n25; *t*-test and, 324; type-II errors and, 311–15, 328, 333; unbiased data and, 320n18, 331; uncertainty and, 334
- R functions: apply, 413–16; bootstrapping and, 416–19; confidence interval (CI) and, 419, 421; convenient, 418–21; data perturbation and, 416; data wrangling, 10, 238, 421; dedicated simulation packages, 421–22; drawing random values from distribution, 409–10; errors and, 413, 416, 421; for-loops and, 410–13; Gaussian distribution and, 409; general linear models (GLMs) and, 418; if-then-else statements, 410, 421; lapply, 413; logical operators, 421; machine learning and, 419; mapply, 415; matrices and, 409, 411–14; normal distribution and, 412; power of statistical analysis and, 75; probability and, 409; randomness and, 409–11, 412, 416–20; random values and, 409–10, 412, 418; repetition, 410–17; replicate, 416; replicates and, 410, 416–21; resampling and, 416–21; sample, 417; sample mean and, 419, 421; sapply, 413; seeds and, 409–11, 416–21; shuffling, 417–18; simulate, 418; simulation basics and, 38n26; study designs and, 62n17, 65; system.time, 418–21; Sys.time, 418–21; tapply, 414; *t*-test and, 418–21; uncertainty and, 418; update, 419
- right-censored data, 224, 227
- RMS *Titanic*, 173–75; Birkenhead drill and, 173n9
- robustness: ANOVA and, 118–22; assumptions of statistical analysis and, 105, 118, 123; evaluation and, 376–77; hypothesis tests and, 47; Kolmogorov-Smirnov test and, 26–28; machine learning and, 201–2; meta-analysis and, 341, 344, 369;

- model-fitting methods and, 39; predictor variables and, 174, 238; rules-of-thumb and, 153–54, 174, 199–202; simulation basics and, 3–4, 12, 18, 20, 25–29, 39; time series and, 199
- Rotenberry, J. T., 322
- R packages: *brms*, 272, 387–88; *DHARMA*, 135n24, 283–84, 387–88; *dplyr*, 30, 421; *faraway*, 38, 121, 154n3, 156, 160, 214n3, 214n6, 215, 276, 296, 298–99, 415, 417, 419; *ggplot2*, 30, 63, 189, 332; *lmerTest*, 138, 353, 355; *mgcv*, 92, 276, 280n4, 298, 393, 418; *mice*, 156, 160, 175, 215, 233, 255; *mvtnorm*, 30, 79–80, 82, 244, 246–47, 250, 272, 293, 296, 346, 349, 355, 359, 363–64, 366, 409, 411; *pwr*, 61, 62n17, 75n4, 313n4; *ranger*, 149, 166n8, 167, 170, 239, 241, 255, 289, 295n18, 300, 393, 394
- R programming language, xii, 1, 10, 12, 377, 380
- rules-of-thumb, xii, 12; abstract concepts and, 1; analysis of variance (ANOVA) and, 154, 179; avoiding mistakes and, xi; Big Data and, 153, 200–4; collinearity and, 154, 165–72; concept of, 153n1; errors and, 153–71, 174–86, 189, 192, 202–7; fixed effects and, 153; machine learning and, 153–54, 165–72, 200–5; mixed-effect models and, 153; model selection and, 154–72; Monte Carlo simulation and, 180–90; predictor variables and, 172–80; random effects and, 153; recipes and, 153; residuals and, 155–58, 168–69, 176, 183–86, 191, 193, 198–99, 204; robustness and, 153–54, 174, 199–202; standard deviation and, 24; time series and, 190–200
- sample, 417–18
- sample mean: meta-analysis and, 342; notation for, 11; retrospective power analysis and, 319; R functions and, 419, 421; standard deviation and, 23, 26, 28, 31, 60, 74, 316, 319; study designs and, 54, 57, 60
- sample size: advanced simulations and, 407; assumptions of statistical analysis and, 106–10, 119, 124–25; autocorrelated data and, 3; censored observations and, 223, 224; computational cost and, 9; data perturbation and, 40; errors and, 247, 248, 259–60; evaluation and, 302, 380, 398; illusory truth effect and, 60–66; Kolmogorov-Smirnov test and, 26–28; machine learning and, 202; meta-analysis and, 342–49, 352, 357, 370; missing data and, 212, 219–20; model selection and, 164; Monte Carlo simulation and, 180; normal distributions and, 23n8; notational conventions and, 11; Pólya urn and, 19–20; power of statistical analysis and, 73–88, 98, 101; predictor variables and, 236; regression analysis and, 243; retrospective power analysis and, 311–13, 316, 319–30; R functions and, 418; standard deviation and, 24, 28, 40, 58, 60, 62, 75, 316, 330n27; stopping rules and, 74, 81; study designs and, 47, 49, 57n9, 58–62, 66, 69; weak effects and, 12
- Schapiro, R. E., 382–83
- Schönbrodt, F. D., 77–79, 81, 83
- seeds: ANOVA and, 22; assumptions of statistical analysis and, 107–11, 119–25, 128–30, 146–49; censored observations and, 223, 225; complete data analysis and, 363–66; errors and, 246–47, 250, 253–56, 261–62; evaluation and, 272, 277–78, 284, 289, 296–99, 303, 305, 378, 391–92, 395; federated analysis and, 359; meta-analysis and, 346, 349, 355–56; missing data and, 220; model selection and, 160, 167, 170; Monte Carlo simulation and, 185–86; Pólya's urn and, 18–20; power of statistical analysis and, 76–77, 80–95; predictor variables and, 178, 233–37, 241; regression analysis and, 244; reproducibility and, 18, 54, 76; retrospective power analysis and, 315, 332; R functions and, 409–11, 416–21; simulation basics and, 4, 18–31, 34, 40–43; study designs and, 54, 65; time series and, 192, 195, 197–98
- Semmelweis, I., 81
- sensitivity analysis, 328
- S-error, 159–62, 165, 329–30, 331
- Shannon, C., 386
- Shannon-Weiner diversity index, 386
- Shapiro-Wilk statistic, 106–9
- signals: advanced simulations and, 404–6; censored observations and, 228n11; evaluation and, 375n2; meta-analysis and, 342n3; noise and, 5, 87, 405; power of statistical analysis and, 87; resampling, 40; retrospective power analysis and, 331n30; study designs and, 55n6; time series and, 195; truth and, 55n6
- Silberzahn, R., 331
- simstudy, 133
- simulated data: advanced simulations and, 404; assumptions of statistical analysis and, 107, 113, 127, 133–35, 143; comparing, 21, 23, 381–82; computational competence and, 17–43; concept of, 4–5; errors and, 252, 254; evaluation and, 270, 282–84, 288, 299, 301, 380–82, 391; generalized least squares (GLS) and, 34–35; hallucinations and, 9–10, 14; improving estimation in, 91–100; K-S test and, 28n12; likelihood-ratio test (LRT) and, 41; meta-analysis and, 346, 354, 369; model selection and, 165–67, 172; Monte Carlo simulation and, 184–86; power of statistical analysis and, 76, 83n8, 87; predictor variables and, 232, 239; randomisation techniques and, 35–43; reasons for, 9–10; replicates and, 4, 19n2, 22–32, 40–43; residuals and, 4; specific, 5–6; time series and, 196; workflow and, 211, 232, 239, 252, 254. *See also* fake data; simulated data
- simulation basics: abstract concepts and, 1, 148, 377, 413–14; analysis of variance (ANOVA) and, 3, 17–18, 20–22, 23; assumptions of statistical analysis, 3 (*see also* assumptions of

simulation basics (*continued*)

statistical analysis); autocorrelated data and, 3, 33–36; Bayesian models and, 10, 13, 22n6; bias and, 3–8, 22n7, 34; chemistry and, xii, 1; collinearity and, 13; computational competence and, 17–43; confidence interval (CI) and, 17, 21–26, 28, 35, 39–40, 43; data-generating model (DGM) and, 4, 13, 17, 43; degrees of freedom and, 23, 32, 36–39; effect size, 13; errors, 3–6, 10, 21–38; frequentist analysis and, xii, 10; Gaussian distribution and, 23, 29–33; general linear models (GLMs), xii, 3, 29–30; imputation and, 13; Kolmogorov-Smirnov (K-S) test and, 26–28; machine learning and, 17, 33; manipulation and, 9; model variability and, 28–29; motivations for, 1; nine steps for, 17–18; normal distribution and, 22–33; notational conventions, 4, 11–12, 143, 382n5; parameter values and, 4–5, 7, 11–12, 18–21, 24–31, 33n22, 35–36, 40–42; pipelines and, 8, 12; Pólya's urn and, 18–20; post hoc analysis and, 13; probability and, 7, 10, 12, 14, 17, 18n1, 20, 22n6, 26, 29n17, 33n20, 37; *p*-values and, xii, 5, 20, 27–28; randomness and, 3–5, 8–11, 17–19, 22–24, 27–43; reality and, 5, 58; recipes and, 1; retrospective power analysis and, xii; R functions and, 38n26; robustness and, 3–4, 12, 18, 20, 25–29, 39; seeds and, 4, 18–31, 34, 40–43; skewed residuals and, 29–32; *t*-test and, 10, 14, 24, 29–32; type-II errors and, 21, 28n12; unbiased data and, 3–5, 22n7

standard deviation: effect size and, 24, 57–62, 76, 316, 330n27, 352; evaluation and, 285, 301, 377–79, 384; Gaussian distribution and, 23; log-normal distribution and, 31–32; meta-analysis and, 352; Monte Carlo simulation and, 184, 187; power of statistical analysis and, 74–76, 96; residuals and, 40, 117, 283, 285; retrospective power analysis and, 314n7, 315, 316, 319, 330n27, 332; rules-of-thumb and, 24; sample mean and, 23, 26, 28, 31, 60, 74, 316, 319; sample size and, 24, 28, 40, 58, 60, 62, 75, 316, 330n27; study designs and, 52–53, 57–58, 60, 62

Standardised Potential Evapotranspiration Index (SPEI), 380

Standard Model, 95–96, 311

Stan Development Team, 388

Stanley, T. D., 350, 352

StarCoder, xii

stepwise model selection: complete data analysis and, 362; errors and, 254–55; meta-analysis and, 345, 351, 354; predictor variables and, 236; rules-of-thumb and, 154–55, 157, 159–60, 164–65

stochastic weather forecasting, 403–4

stopping rules: evaluation and, 378; power of statistical analysis and, 74, 81–85, 87, 89, 95, 101; principle of, 81; puerperal fever deaths

study and, 81–84; sample size and, 74, 81; study designs and, 66. *See also* censored observations

study designs: adequate replication and, 56–60; analysis of variance (ANOVA) and, 60; Bayesian models and, 45, 49; bias and, 45–48, 58n10, 65, 69; bootstrapping and, 48, 64–66, 69; confidence interval (CI) and, 63, 65–66, 69; correlation and, 57–58; data-generating model (DGM) and, 50, 69; degrees of freedom and, 58n10, 67; effect size and, 47, 49, 53, 57–62, 67, 69; errors and, 49, 57n9, 58–62, 66–67; false negatives and, 57–59, 66; false positives and, 57–60, 67; frequency distribution and, 50; frequentist analysis and, 49; Gaussian distribution and, 51, 52n5; HARKing and, 67–68; hypotheses definitions and, 49–55; hypothesis testing and, 58–60; illusory truth effect and, 47–52, 55–56, 60–66; likelihood and, 45, 50; manipulation and, 45, 49; matrices and, 61, 66; noise and, 49; normal distribution and, 53; null expectation and, 52, 55; null hypothesis and, 56–61, 69; parameter values and, 49–50, 55n7, 56, 60–61, 68–69, 369; perceived accuracy and, 51–55, 56; plausibility and, 50–56, 61–62, 66; post hoc analysis and, 67, 69; power of, 45, 47, 49, 57–62, 66–69; prior expectations and, 331–32; probability and, 45, 50, 56, 58, 66–67, 69; processing fluency and, 54–55, 62; *p*-values and, 58n12, 67; quantifying variables of interest, 50–55; questionable research practices (QRPs) and, 67–69; randomness and, 51–52, 55n8, 57, 61; replicates and, 56–59, 65, 73; R functions and, 62n17, 65; sample mean and, 54, 57, 60; sample size and, 47, 49, 57n9, 58–62, 66, 69; seeds and, 54, 65; signals and, 55n6; standard deviation and, 52–53, 57–58, 60, 62; stopping rules and, 66; testing hypotheses and, 55–67; true positives and, 59, 66; *t*-test and, 57n9, 59–62; type-II errors and, 58, 59n13, 66; uncertainty and, 47–48, 69

subset regression, 154–55, 397n13

synthetic data, 4n1

system.time, 418–21

Sys.time, 418–21

Thurman, E. M., 228n11

tidyverse, 10

time series: autocorrelated data and, 191, 193, 196n27, 281; carbon dioxide (CO₂) and, 85–87, 89n15, 90–91; change-point methods and, 192, 194–98; correlation and, 191, 193, 196n27; detrending and, 154, 190–200, 205, 245; differencing and, 191–92, 193; errors and, 192; length of, 85–91; likelihood and, 190; matrices and, 198; Monte Carlo simulation and, 182–84; noise and, 197, 198; parameter values and, 191; probability and,

- 190, 197; regression analysis and, 182–84, 190–200; replicates and, 194, 197–98; residuals and, 191, 193, 198–99; robustness and, 199; seeds and, 192, 195, 197–98; signals and, 195; simulated data and, 196
- true negatives, 35, 67. *See also* type-II errors
- true positives: assumptions of statistical analysis and, 123–24; model selection and, 159–64, 165; predictor variables and, 180; retrospective power analysis and, 324, 328n25; study designs and, 59, 66. *See also* type-I errors
- truncated data, 224n8
- t*-test: assumptions of statistical analysis and, 103, 120, 124, 127, 138; normality and, 29; power of statistical analysis and, 74–77; retrospective power analysis and, 324; R functions and, 418–21; simulation basics and, 10, 14, 24, 29–32; skewed residuals and, 29–32; study designs and, 57n9, 59–62; Welch's generalisation of, 120
- type-I errors: assumptions of statistical analysis and, 105–6, 117, 122, 123–27; inflated, 252; model selection and, 154, 164; Neyman and, 313; Pearson and, 313; power of statistical analysis and, 74; retrospective power analysis and, 311–22, 327–28; simulation basics and, 21, 26, 28n12, 29, 31–34; study designs and, 58, 59n13, 60, 66–67. *See also* hypothesis testing
- type-II errors: assumptions of statistical analysis and, 105, 126; inflated, 252; missing data and, 219; Neyman and, 313; Pearson and, 313; predictor variables and, 174, 176, 178; retrospective power analysis and, 311–15, 328, 333; simulation basics and, 21, 28n12; study designs and, 58, 59n13, 66. *See also* power of statistical analysis
- unbiased data: assumptions of statistical analysis and, 115, 134; complete data analysis and, 362; errors and, 252, 257; evaluation and, 277, 381; machine learning and, 201; meta-analysis and, 342, 343, 347n11, 352, 354, 370; missing data and, 213n2, 219; model selection and, 159, 171; power of statistical analysis and, 74, 79, 83–85, 97; predictor variables and, 172, 235, 238, 242; regression analysis and, 243; retrospective power analysis and, 320n18, 331; simulation basics and, 3–5, 22n7; workflow and, 211, 213n2, 219, 235, 238, 242, 243, 252, 257
- uncertainty: advanced simulations and, 403–4; assumptions of statistical analysis and, 115, 147; bias and, 4, 47–48, 69, 115, 171, 211, 252, 254, 260, 273, 277, 404; bootstrapping and, 29–30, 31, 39, 69, 245, 251–54, 260, 273, 275, 277, 288, 306, 418; communication of, 47–48, 69, 275, 302; errors and, 245–60; evaluation and, 269–75, 277, 288–91, 294–96, 297, 302, 305–7, 379; improving estimation and, 91–100; machine learning and, 203; model-fitting methods and, 28–30; model selection and, 171; Monte Carlo simulation and, 181–84, 188; power of statistical analysis and, 78, 91–93; propagation of, 247–49; quantification of, 8, 245, 250, 254, 269, 302; randomness and, 29, 31, 39, 249, 254; retrospective power analysis and, 334; R functions and, 418; study designs and, 47–48, 69; workflow and, 211–12, 245–60
- unit testing, 376–79. *See also* code verification; evaluation; functional verification; unit testing
- van Marle, M. J. E., 180–90
- variance-covariance matrices, 33n22, 34, 243n17, 248–50, 293. *See also* correlation
- Vasisht, S., xii
- Walker, H., 313
- W boson, 95–96
- Weisser, W. W., 113–15
- Wiens, J. A., 322
- Wilcox, R. R., 29n18, 117
- workflow: bias and, 211 (*see also* bias); Big Data and, 211; machine learning and, 4; matrices and, 212n1, 220–23, 236, 240–50, 251n22, 255; model-fitting methods and, 211; parameter values and, 211, 213, 217–19, 220, 222, 226, 232–46, 249–61; pipelines and, 8, 12, 211, 228n11, 245, 251–58, 403–5; *p*-values and, 211, 215; random-forest models and, 166–72; residuals and, 212, 217–19; simulated data and, 211, 232, 239, 252, 254; unbiased data and, 211, 213n2, 219, 235, 238, 242, 243, 252, 257; uncertainty and, 211–12, 245–60, 254–56
- χ^2 -statistics, 10, 14
- Yates, K. L., 145
- Zayed, A. A., 211
- Z-score, 58n11, 75