Eric-Jan Wagenmakers and Dora Matzke

Bayesian Inference From The Ground Up

The Theory of Common Sense

Eric-Jan Wagenmakers and Dora Matzke

Bayesian Inference From The Ground Up

The Theory of Common Sense

JASP Publishing

TUFTE-LATEX.GOOGLECODE.COM

This printing, December 2023

Contents

Contents 3

Preface 9

Synopsis 13

JASP 19

PART I PROBABILITY

1 *Probability Belongs Wholly to the Mind?* 31

2 *Epistemic and Aleatory Uncertainty* 45

3 *The Rules of Probability* 57

4 *Interlude: Leibniz's Blunder* 81

5 *The Measurement of Probability* 89

6 *Coherence* 99

PART II COHERENT LEARNING, LAPLACE STYLE

7 *Learning from the Likelihood Ratio* 119

8 *An Infinite Number of Hypotheses* 135

9 *The Rule of Succession* 157

10 *The Problem of Points* 173

11 *Interlude: Buffon's Needle* 183

12 *The Pancake Puzzle* 205

13 *A Plethora of Pancakes* 227

PART III COHERENT LEARNING, JEFFREYS STYLE

14 *A Crack in the Laplacean Edifice* 245

15 *Wrinch and Jeffreys to the Rescue* 253

16 *Haldane's Rule of Succession* 273

17 *Jeffreys's Platitude* 285

18 *The Principle of Parsimony* 297

PART IV APPENDICES

33 *Jevons Explains Permutations* 511

34 *Pascal's Arithmetical Triangle* 517

35 *Statistical Analysis of the Binomial Distribution* 529

36 *Recommended Readings* 535

37 *Figure Listing* 541

Bibliography 555

In spite of its immense difficulties of application, and the aspersions which have been mistakenly cast upon it, the theory of probabilities, I repeat, is the noblest, as it will in course of time prove, perhaps the most fruitful branch of mathematical science. It is the very guide of life, and hardly can we take a step or make a decision of any kind without correctly or incorrectly making an estimation of probabilities.

W. Stanley Jevons

The Principles of Science, 1874

Preface

The purpose of this book is to present the key concepts of Bayesian inference in an intuitive and attractive fashion. The current treatment differs with respect to other 'introductions to Bayesian statistics' in five important ways. First and foremost, we have tried to present an introduction for undergraduate students in the social sciences, *not* an introduction geared toward associate professors of mathematics at MIT. This means that we focus on providing the right intuition, that we seek to solidify that intuition with concrete examples, and that we try to limit the number of equations (see also Lindley 1985; 2006). Of course, the one equation we cannot avoid is Bayes' theorem. Luckily, the theorem represents 'common sense expressed in numbers', and it is remarkable how much insight can be gained from just this single formula.

Contrary to popular belief, this is probably not Thomas Bayes (c. 1701-1761). For details, see the discussion by Prof. David R. Bellhouse at `http://www.york.ac.uk/depts/maths/histstat/bayespic.htm`.

The second way in which our book differs from other introductory treatments of Bayesian inference is that we approach the topic according to the philosophy of the geophysicist and polymath Sir Harold Jeffreys (1891-1989). Specifically, Jeffreys showed how the Bayesian paradigm can support both *hypothesis testing* ('is the effect present or absent?') and *parameter estimation* ('how big is the effect, assuming it is present?'). In contrast, many Bayesian textbooks fail to provide a coherent and compelling account of hypothesis testing – in our opinion, this is a serious omission that betrays a lack of familiarity with how scientists conduct experiments and interpret results.

The third way in which our treatment differs from most others is that we emphasize the central role of *prediction* in scientific learning. It may be intuitively clear that sound predictions ought to arise from our knowledge of the world; it is less clear that our knowledge of the world is adjusted as a function of predictive performance. Yet Bayes' theorem tells us that *accounts of the world that predicted observed data successfully enjoy a boost in plausibility, whereas accounts that predicted poorly suffer a decline*.[1]

[1] Repeated throughout this book, this specific mantra was first presented in Wagenmakers et al. (2016a) as suggested by our close colleague Michael Lee (`https://faculty.sites.uci.edu/mdlee/`). Note also that the emphasis on prediction is common in robotics and object tracking, where beliefs need to undergo constant revision according to changing inputs from the environment.

The fourth way in which this book stands out is in its emphasis on historical development. Among the heroes of this book are Pierre-Simon Laplace (1749–1827), Augustus De Morgan (1806–1871), William

Stanley Jevons (1835–1882), Henri Poincaré (1854–1912), J. B. S. Haldane (1892–1964), Dorothy Maud Wrinch (1894–1976), and of course Sir Harold Jeffreys (1891–1989). Many chapters provide abundant historical background and extensive quotations. Some students have told us that long quotations are boring. We heap scorn on this notion. Our heroes may no longer be around to give a Ted Talk or record a TikTok video, but their words have lost none of their eloquence, relevance, and vision. Poincaré advocated a similar approach to the teaching of mathematics:

> "In the edifices built up by our masters, of what use to admire the work of the mason if we can not comprehend the plan of the architect? (...)
>
> Zoologists maintain that the embryonic development of an animal recapitulates in brief the whole history of its ancestors throughout geologic time. It seems it is the same in the development of minds. The teacher should make the child go over the path his fathers trod; more rapidly, but without skipping stations. For this reason, the history of science should be our first guide." (Poincaré 1913, pp. 436-437)

The fifth way in which this book is unique is that we take full advantage of JASP, an open-source statistical software program with extensive support for Bayesian analyses. Available for free at `jasp-stats.org`, JASP makes it easy to perform comprehensive Bayesian analyses with just a few mouse clicks or keystrokes. The current volume, 'The Theory of Common Sense', primarily uses the JASP module *Learn Bayes*[2]; the second volume ('Common Sense in Practice' – in preparation) will take full advantage of the many standard Bayesian analyses implemented in JASP such as the comparison of two proportions, the comparison of means, hierarchical modeling, meta-analysis, and more.

[2] The development of this module was supported by the APS Fund for Teaching and Public Understanding of Psychological Science and the Erasmus+ 'QHELP' project.

To keep the concepts separate and the content digestible, we have chosen to present the material in a sequence of relatively short chapters. Most chapters include a summary, exercises, and suggested readings. Occasional interlude chapters provide material that is educational but not necessary to understand the remaining chapters. Note that this book is still a living document; the current version will be regularly updated as new chapters become available. We intend to continually update the book material, so we welcome any and all suggestions for improvement.

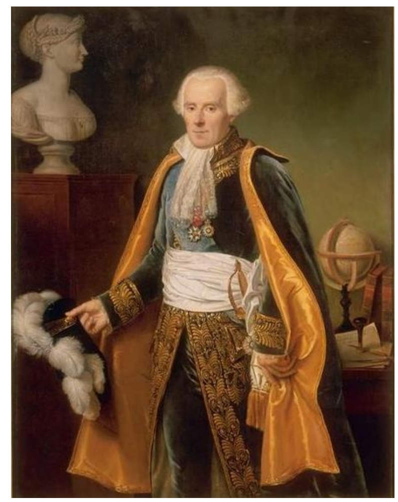
The goal of this volume is to outline philosophical ideas, sketch key historical developments, and generally to proceed systematically from scenarios that are simple to those that are more complex. Specifically, Part I introduces the Bayesian view on probability, Part II outlines the Laplacean estimation approach, and Part III provides an overview of the Jeffreyian hypothesis testing approach, which was explicitly developed to overcome the limitations of the Laplacean approach.[3] Part IV includes several technical appendices.

[3] It is ironic that some modern statisticians, unaware of century-old arguments, unwittingly regress and happily advocate the Laplacean approach over the Jeffreyian approach.

Pragmatic readers looking for a crash course in applied Bayesian statistics may skip the first volume altogether and proceed directly to the second volume. The first chapters of the second volume summarize the key points from the first volume.[4] We strongly feel that this is not just another course on just another topic. In the epigraph to this book, Jevons called the theory of probabilities "the very guide of life". To further underscore the importance of the topic, we cannot improve on the French genius Pierre-Simon Laplace, who ended his famous 1814 book *Essai Philosophique sur les Probabilités* in dramatic fashion:

[4] The second volume is still in preparation, so this advice is currently not very practical. Impatient readers may consult one of the many tutorials on applying Bayesian statistics (e.g., van Doorn et al. 2021).

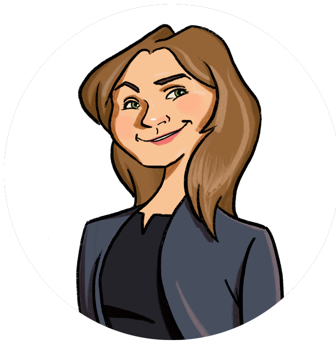
> "One sees in this essay that the theory of probabilities is basically only common sense reduced to a calculus. It makes one estimate accurately what right-minded people feel by a sort of instinct, often without being able to give a reason for it. It leaves nothing arbitrary in the choice of opinions and of making up one's mind, every time one is able, by this means, to determine the most advantageous choice. Thereby, it becomes the most happy supplement to ignorance and to the weakness of the human mind. If one considers the analytical methods to which this theory has given rise, the truth of the principles that serve as the groundwork, the subtle and delicate logic needed to use them in the solution of the problems, the public-benefit businesses that depend on it, and the extension that it has received and may still receive from its application to the most important questions of natural philosophy and the moral sciences; if one observes also that even in matters which cannot be handled by the calculus, it gives the best rough estimates to guide us in our judgements, and that it teaches us to guard ourselves from the illusions which often mislead us, one will see that there is no science at all more worthy of our consideration, and that it would be a most useful part of the system of public education." (Laplace 1814/1995, pp. 124)



Pierre-Simon Laplace (1749-1827). "On voit, par cet Essai, que la théorie des probabilités n'est, au fond, que le bon sens réduit au calcul; elle fait apprécier avec exactitude ce que les esprits justes sentent par une sorte d'instinct, sans qu'ils puissent souvent s'en rendre compte." Posthumous portrait by Jean-Baptiste Paulin Guérin, 1838.

About the Authors

Prof. dr. Eric-Jan ('EJ') Wagenmakers is a mathematical psychologist and a militant Bayesian. He works at the Psychological Methods Unit of the University of Amsterdam where he heads a lab that develops the JASP open-source software program for statistical analyses. Wagenmakers is also a strong advocate of Open Science and the preregistration of analysis plans. For more information, see `www.ejwagenmakers.com`.



Dr. Dora ('Dora') Matzke is also a mathematical psychologist and a dedicated Bayesian working at the Psychological Methods Unit of the University of Amsterdam. Matzke develops formal models for speeded decision making in psychology and cognitive neuroscience. Specifically, Matzke has proposed new models and Bayesian methods to measure response inhibition, that is, the time it takes to stop an action. For more information, see `https://www.ampl-psych.com/team/dora-matzke/`.

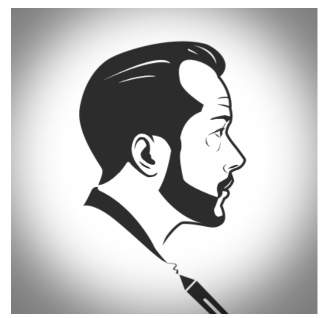
Acknowledgments

We gratefully acknowledge the assistance of Alexander Ly, Maarten Marsman, Quentin F. Gronau, Charlotte Tanis, Johnny van Doorn, Don van den Bergh, František Bartoš, Šimon Kucharský, and Jiashun Wang. We also thank the JASP team (`jasp-stats.org`) for creating free software that allows students to learn the basics of Bayesian inference without tears. In general, this book could not have been written without the continual interaction with our friends and colleagues at the Psychological Methods Unit at the University of Amsterdam.

A special thanks goes out to Viktor Beekman (`instagram.com/viktordepictor`) for his artwork which is on display throughout this book. Most of the graphs were created in `R` or in JASP (`jasp-stats.org`). We are grateful to those who kindly granted us permission to present copyrighted material. A figure listing is at the end of this book.

We are indebted to the creators of the Tufte LaTeX style files, the Overleaf editing system, and to Wikipedia. Special thanks go to LaTeX gurus Kevin Godby and Jonas Petter for upgrading the Tufte style file based on a series of complicated requests by EJ Wagenmakers and Michael Lee. We also thank our students and colleagues for their suggestions for improvement.

This work was supported by a Vici grant (016.Vici.170.083), an advanced ERC grant (743086 UNIFY), and an Erasmus+ grant (QHELP). The text is set in Lood, designed by Hans van Maanen for Canada Type.



Graphical artist Viktor Beekman (Viktor.Beekman@gmail.com).

Co-funded by the Erasmus+ Programme of the European Union

The writing of this book and the development of the associated JASP *Learn Bayes* module was supported by the Erasmus+ 'QHELP' project, whose aim is to develop software to facilitate quantitative learning. The project website can be found at `https://www.qhelp.eu/`.

Eric-Jan Wagenmakers and Dora Matzke, December 2023

Synopsis

The subject upon which we now enter must not be regarded as an isolated and curious branch of speculation. It is the necessary basis of nearly all the judgments and decisions we make in the prosecution of science, or the conduct of ordinary affairs.

Jevons, 1874

Chapter Goal

This chapter outlines the Bayesian learning cycle that forms the conceptual backbone of the entire paradigm.

The introduction to this chapter is a translation from Wagenmakers and Gronau (2018).

The Learning Cycle

There is a Dutch saying "not even a donkey bumps into the same stone twice".[5] Donkeys learn from experience, and they share this ability to adapt with all known animal species – cats, lizards, spiders...even single-cellular slime molds are capable of learning. It could hardly be any other way, of course, for evolution is a ruthless sculptor: organisms unable to adapt to their environment are doomed to extinction.

[5] In Dutch: "zelfs een ezel stoot zich in het gemeen niet tweemaal aan dezelfde steen". English versions: "once bitten twice shy", or "Fool me once, shame on you. Fool me twice, shame on me."

But how do organisms learn from their environment? In general, learning can only occur when there exist multiple rival hypotheses. If there is only a single hypothesis, this represents a religious belief, an unshakable conviction that is impervious to any empirical disconfirmation whatsoever. To learn, therefore, we must begin with multiple competing hypotheses, each with its own plausibility. In the Amazon, a young piranha detects movement in the water, far away; one hypothesis holds that the movement is triggered by wounded prey, the other holds that it is caused by a healthy fellow piranha. To find out more, our piranha swims closer. In this way, the piranha collects new observations, and these should lead to learning, that is, an adjustment of the relative plausibility of the competing hypotheses. It is intuitive that hypotheses increase and decrease in plausibility in proportion to their predictive success: the 'prey' hypothesis predicts a violent thrashing, whereas the 'fellow piranha' hypothesis predicts a more even motion pattern. When the new observations suggest a violent thrashing, this increases the plausibility of the 'prey' hypothesis and decreases the plausibility of the 'fellow piranha' hypothesis.

On the basis of such general considerations, we arrive at the following qualitative regularity:

$$\begin{array}{c}\text{Present knowledge}\\ \text{about the world}\end{array} = \begin{array}{c}\text{Past knowledge}\\ \text{about the world}\end{array} \times \begin{array}{c}\text{Predictive}\\ \text{updating factor.}\end{array}$$

This regularity states that the learning process –the adjustment of knowledge on the basis of observed data– is governed by the predictive adequacy of the rival hypotheses. This common-sense argument is formalized by what is known as Bayes' rule or Bayes' theorem, but for now we will discuss the rule without invoking the equation.

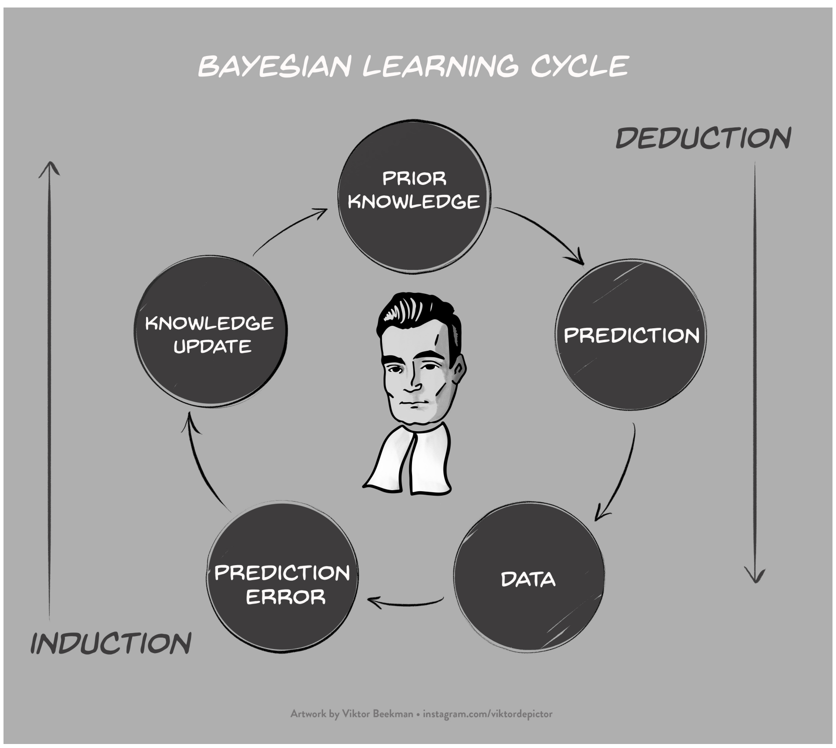


Figure 1: Bayesian learning can be conceptualized as a cyclical process of updating knowledge in response to prediction errors. The prediction step is deductive, and the updating step is inductive. For a detailed account, see Jevons (1874/1913, Chapters XI and XII). Figure available at `BayesianSpectacles.org` under a CC-BY license.

The learning process is depicted in Figure 1. It is important that the learning process can continue indefinitely, as long as new data keep flowing in; the updated (i.e., posterior) knowledge after one cycle of learning serves as the prior knowledge for the next cycle. This is not only theoretically elegant, but for a simple organism like our piranha, who is confronted by a life-long deluge of data, it is also practically relevant: after the knowledge has been updated, the old data have done their job and can safely be forgotten — the only thing the piranha needs to do is use incoming data to adjust the existing knowledge.

The Knowledge Pump

The Bayesian learning cycle, shown in Figure 1, can be viewed as a knowledge pump[6] with two fundamentally different processes working in alternation: *deduction* and *induction*. The deductive process specifies how rival hypotheses generate predictions for observed data (see the box 'The Data-Generating Process' below). Without such predictions, the learning process cannot get off the ground. Once the data are in, the relative adequacy of the predictions can be assessed, and this drives an inductive process: the adjustment of knowledge in light of experience. Once the inductive process has finished, the knowledge pump is ready for its next predict-update cycle.[7]

[6] Or an old-fashioned railroad handcar, now seen mostly in cartoons.

[7] Some incredibly smart researchers have argued that scientific reasoning should be based only on the deductive process. These researchers were probably mistaken (e.g., as argued in Jeffreys 1973, Chapter 1; Jeffreys 1961, pp. 1-8; Jevons 1874/1913).

The Data-Generating Process

One of the key goals of statistical inference is to use observed data to figure out ('infer') the unobserved processes that gave rise to those data. These unobserved (if you want to sound smart, call them 'latent') processes are generally known as a 'data-generating process' (DGP). In general, a DGP represents a statement about the world. Philosophers often prefer the term 'proposition', empirical researchers usually speak of 'hypotheses', whereas statisticians postulate 'models'. A statistical model can be considered a concrete implementation of a hypothesis; for instance, a hypothesis could be 'women play better chess than men', and a corresponding statistical model would stipulate that the average Elo-rating of women exceeds that of men (after correcting for baseline differences in participation rates).[8] A statistical model is often a composite of several DGPs. For example, in the model that postulates that women play better chess than men, the unknown true difference in mean Elo-rating can take on all kinds of values; it is therefore considered a *parameter* within the larger model: an instance of a larger class of DGPs. As we will see, the distinction between propositions, hypotheses, models, and parameters is mostly cosmetic: the Bayesian learning process governs the data-driven change in plausibility regardless of the label applied.

[7] See the article 'Why are (the best) women so good at chess? Participation rates and gender differences in intellectual domains' by Bilalić et al. (2009).

As noted above, and as demonstrated in later chapters, Bayes' rule formalizes the learning cycle shown in Figure 1. By doing so, it allows us to move beyond the data-generating perspective where we postulate only how *underlying causes lead to observed consequences*, that is, causes

→ `consequences`. Although this forms an essential ingredient of the learning process, in real life we are confronted with data and wish to gain knowledge about the underlying process. In other words, we want to move in the opposite direction and learn from *observed consequences about the underlying causes*, that is, `causes` ← `consequences`. By inverting the causal arrow, Bayes' rule allows us to reason about the world in a coherent fashion.[9]

[9] More on coherence in Chapter 6.

Adage of the *New Academy*, a group of influential Greek philosophers who believed that we cannot be absolutely certain of anything. To prevent this insight from resulting in behavioral paralysis, concrete action is based on whatever seems most plausible. For a riveting account, see Cicero (45BC/1956a) and Cicero (45BC/1956b). Figure available at `BayesianSpectacles.org` under a CC-BY license.

Exercises

1. Go online and read up on 'Cromwell's rule'. How does it connect to the foregoing argument?
2. The statement on the tile in the margin, "never assert absolutely", is attributed to Carneades, Russell, and Lindley. What did Russell say that warrants his inclusion on the tile?

Chapter Summary

The Bayesian learning cycle consists of a never-ending alternating sequence of deductive forecasting and inductive knowledge adjustment. At each point in time, rival accounts of the world make predictions, and the adequacy of these predictions in light of the observed data determines how the plausibility of the rival accounts gets updated: accounts that predicted the data relatively well enjoy a boost in plausibility, whereas those that predicted the data relatively poorly suffer a decline.

Want to Know More?

✓ Jevons, W. S. (1874/1913). *The Principles of Science: A Treatise on Logic and Scientific Method*. London: Macmillan. Timeless classic by a brilliant author, and freely available online.

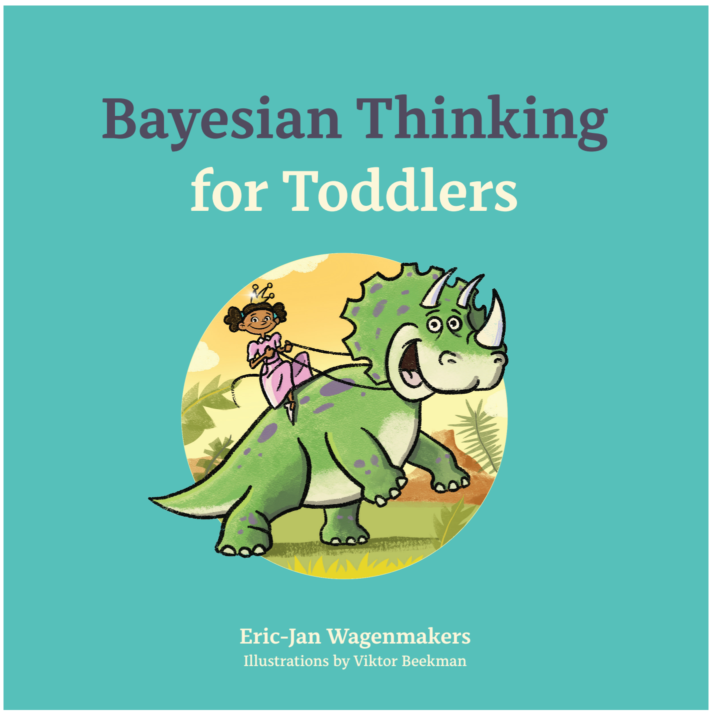
"In deduction we are engaged in developing the consequences of a law or identity. (...) Induction is the exactly inverse process. Given certain results or consequences, we are required to discover the general law from which they flow." (Jevons 1874/1913, p. 14)

✓ Wagenmakers, E.–J., Dutilh, G., & Sarafoglou, A. (2018). The creativity-verification cycle in psychological science: New methods to combat old idols. *Perspectives on Psychological Science*, 13, 418-427. A historical perspective on the interplay between deduction and induction.

✓ Wagenmakers, E.–J. (2020). *Bayesian Thinking for Toddlers*. Freely available at `psyarxiv.com/w5vbp/`. Dinosaurs courtesy of Viktor Beekman. Also available in Dutch, German, and Turkish.

✓ The predict-update description of the Bayesian learning cycle is common in the literature on *Bayesian filtering*, where the environment is dynamic (e.g., Thrun et al. 2005). For instance, as a robot moves across a room it needs to update its beliefs about its current position according to the information coming from its sensors. Another popular application is the tracking of moving objects such as cars or rockets. However, the same predict-update mechanism also underlies learning in static environments, although textbooks rarely emphasize this aspect. For a clear conceptual introduction to Bayesian filtering, we recommend the YouTube videos by Cyrill Stachniss.

"Doubt is not a pleasant condition, but certainty is an absurd one." – Voltaire.



Cover of *Bayesian Thinking for Toddlers*. "A must-have for toddlers with even a passing interest in Bayesian knowledge updating and the prequential principle."

JASP

> In order that a scientific method may be of any value, it must satisfy two conditions. In the first place, it must be possible to apply it in the actual cases to which it is meant to be relevant. In the second, its arguments must be sound. The main object of science is to increase knowledge of the world, and if a method is not applicable to anything in the world it obviously cannot lead to any knowledge. This principle is very elementary, and it is probably for that very reason that it is habitually overlooked in theories of scientific knowledge.
>
> Wrinch & Jeffreys, 1921

Chapter Goal

This chapter introduces JASP, an open-source statistical software program with an attractive graphical user interface. JASP makes it easy to conduct comprehensive Bayesian analyses with just a few mouse clicks or keystrokes. JASP will play an increasingly important role as you progress through the chapters of this book, and we recommend that you install JASP, free of charge, from `jasp-stats.org`.

JASP unlocks Bayesian advantages for practitioners unwilling to pursue a career in mathematical statistics.

A Bayesian Mousetrap

At its theoretical core, Bayesian inference is about *learning from experience*: accounts of the world that predict new data relatively well enjoy a boost in plausibility, whereas accounts that predict new data relatively poorly suffer a decline. This appears perfectly straightforward, and in the previous chapter we argued that even piranhas learn from experience and hence engage in some form of Bayesian inference. The idea that Bayesian inference is *easy* is reinforced by pithy statements such as "Bayesian inference is hard in the sense that thinking is hard" (Don Berry) and "Bayesian statistics is fundamentally boring" (Phil Dawid).

Unfortunately, between Bayesian theory and Bayesian practice, the gods have placed a healthy dose of *mathematical statistics* and *probabilistic programming*. This does not worry piranhas much because piranhas are content with a quick-and-dirty form of learning, good enough to help them survive. But when humans apply Bayesian inference to a data analysis problem, quick-and-dirty 'intuitive Bayes' will not suffice – common sense needs to be translated to numbers, and the reallocation of plausibility needs to happen with mathematical precision. Doing so is *hard*.

Figure available at `BayesianSpectacles.org` under a CC-BY license.

Consequently, practitioners with limited quantitative backgrounds –psychologists, physicians, ecologists, business analysts, neuroscientists– quickly discover the truth in the Russian proverb that "free cheese can only be found in a mousetrap". The 'cheese' represents the benefits that come with every Bayesian analysis: probability can be assigned to hypotheses and parameters, evidence for and against hypotheses can be quantified and monitored as the data accumulate, and prior knowledge can be seamlessly taken into account. The 'mousetrap' is that these Bayesian benefits are available only to those who are willing to pay for them with sweat and tears. This is off-putting. Most practitioners do not have the patience to take several courses in mathematical statistics and probabilistic programming before they can finally implement a Bayesian t-test to analyze their data. Who can blame them? Instead, the blame lies with Bayesian statisticians, who as a group have failed to develop user-friendly software that makes it easy for practitioners to reap the benefits of Bayesian techniques without first having to pursue a career in mathematical statistics.

Bayesian Inference Without Tears

To close the gap between Bayesian theory and Bayesian practice, our group (part of the Psychological Methods Unit at the University of Amsterdam) has developed JASP, a cross-platform, open-source statistical software program with an attractive graphical user interface (GUI).[10] Using JASP, practitioners can conduct Bayesian inference by dragging and dropping variables of interest into analysis panels, whereupon the associated statistical output becomes available for inspection. With JASP, the emphasis can shift from shallow problems of *implementation* and *computation* to deeper problems of *specification* and *interpretation*. Free cheese, and no mousetrap.

[10] In honor of the Bayesian pioneer Sir Harold Jeffreys (1891-1989), JASP stands for 'Jeffreys's Amazing Statistics Program'. Jeffreys is the hero of this book, and later chapters will discuss his statistical vision in detail.

JASP is a central component of this book. In 'Part II: Coherent Learning, Laplace Style' and 'Part III: Coherent Learning, Jeffreys Style', we encourage the reader to work with the *Learn Bayes* module in JASP.[11] Inspired by the Bayesian knowledge pump from Figure 1, the *Learn Bayes* module facilitates an interactive, step-by-step exploration of the cyclical process of Bayesian learning: specifying prior knowledge, making predictions, collecting data, assessing predictive success, and updating to posterior knowledge.

[11] The development of this module was supported by the APS Fund for Teaching and Public Understanding of Psychological Science and by the Erasmus+ project 'QHELP'.

Screenshot of the JASP website, September 2022.

In the second volume, 'Common Sense in Practice' (in preparation) we turn to a series of popular statistical tools such as the t-test, the A/B test, the correlation test, and others. With JASP, it is easy to conduct comprehensive Bayesian analyses for these tests with just a few mouse clicks. This allows students, teachers, and researchers to focus on the key concepts: setting up the models and interpreting the results. More advanced applications will make use of the *JAGS* module that presents a JASP GUI for probabilistic programming (Plummer 2003). Another relevant JASP module is *Distributions*, which offers students the opportunity to explore particular distributions and fit them to data.

The JASP Principles

JASP is based on the following collection of interrelated philosophies, convictions, and principles about science and software:

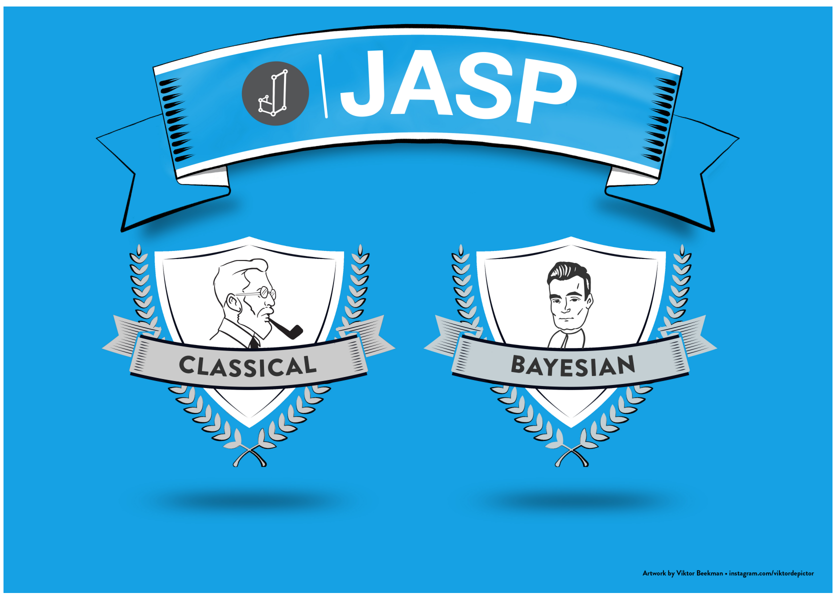
- ✓ JASP is free. The core functionality of JASP will *always* be available for free. We consider it a travesty that, every year, universities around the world pay hundreds of millions of dollars of public money for licensing fees so that their employees can execute analyses that –from a statistical perspective– are trivial.

- ✓ JASP is open-source. The source code for JASP is available on GitHub at `https://github.com/jasp-stats/jasp-desktop/`. Currently, the analysis code is based on R and supported by 475 different R packages[12]; for its Bayesian analyses, JASP uses close to 40 R packages including BayesFactor (Morey and Rouder 2018), BAS (Clyde et al. 2011, Clyde 2016), abtest (Gronau et al. 2021), bain (Gu et al. 2019), stanova (by Henrik Singmann), Bayesrel (Pfadt et al. 2022), conting (Overstall and King 2014), RoBMA (Maier et al. 2023), RStan (Stan Development Team 2020), jfa (Derks et al. 2021), and multibridge (Sarafoglou et al. in press). The graphical user interface is familiar to users of SPSS and has been programmed in C++, html, and javascript.

[12] A full listing is available at `https://jasp-stats.org/r-package-list/`.

✓ JASP is statistically inclusive. JASP implements both Bayesian and frequentist/classical procedures.[13] In addition, JASP allows for both parameter estimation and hypothesis testing. This gives the user the freedom to choose the method most appropriate for the question at hand. Moreover, users can check the robustness of their conclusions by conducting an alternative analysis.

[13] Throughout this book, the emphasis will be firmly on Bayesian methodology.



The JASP coat of arms. The left shield shows Sir Ronald Fisher (1890-1962), longtime proponent of classical statistics and vociferous opponent of Bayesian statistics.

✓ JASP has a graphical user interface (GUI). Part of the JASP interface is familiar to users of IBM's SPSS: data are available in spreadsheet format, variables can be dragged and dropped into input fields, and the results are generated in a separate output panel. An example of the input and output panels is shown in Figure 2.

✓ JASP is designed with the user in mind. The JASP GUI is dynamic and has *immediate feedback*, updating its output as the user alters the input. In addition, the JASP GUI is based on the principle of *progressive disclosure*: initial output is minimalist to avoid overwhelming

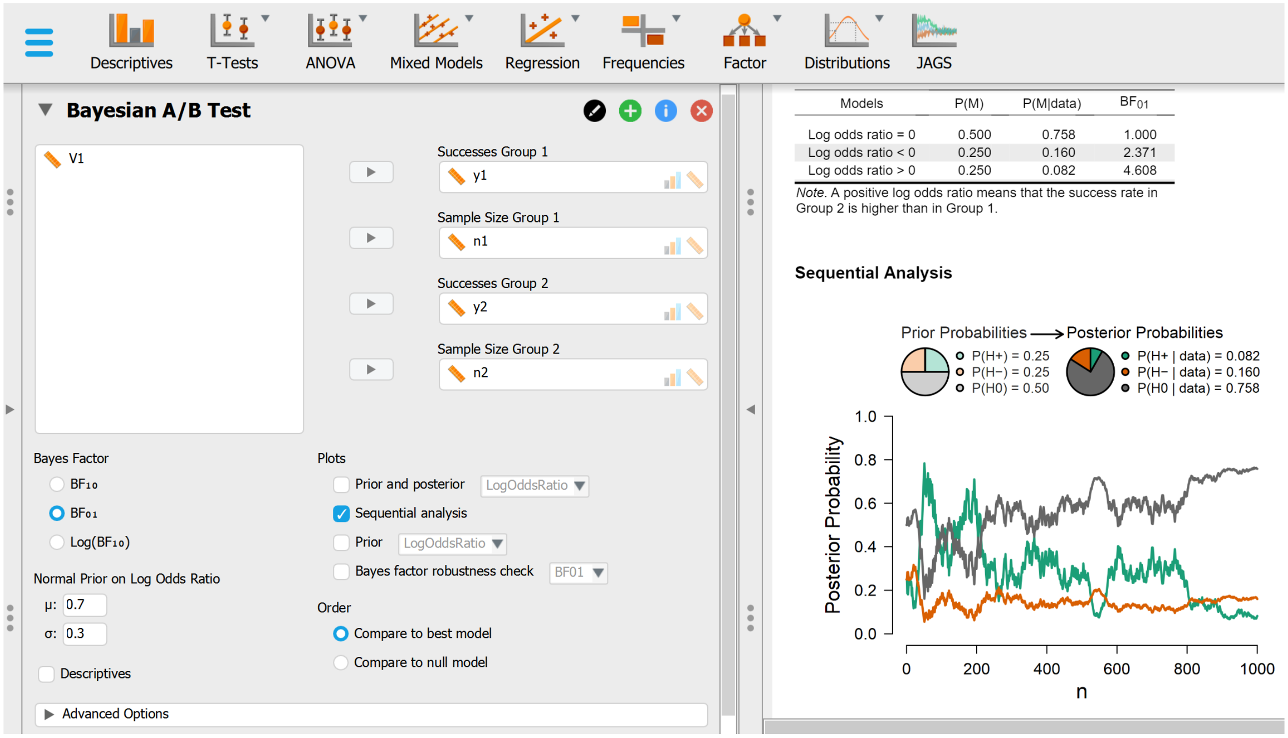


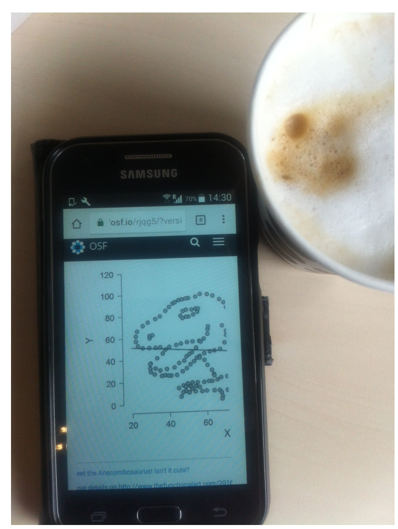
Figure 2: Screenshot of the JASP A/B test for the comparison of two proportions. Analysis options can be set in the left panel, and associated output is shown in the right panel.

the user; if desired, the user can request additional information by checking boxes. The JASP output was designed to be attractive and effective: figures are publication-ready and tables are in APA format, ready to be copy-pasted into a word processor.

✓ JASP facilitates transparent statistical reporting. JASP allows users to save data, input options, and annotated output in a single .jasp file.[14] This file can be opened and edited by colleagues and students who also have JASP installed; in addition, the Open Science Framework (`https://osf.io/`) has a JASP previewer that allows anyone to examine annotated JASP output from within a browser, even without having JASP installed. This means that students and colleagues can review JASP output on their tablet or cell phone. As of version 0.17, the underlying R syntax is visible by clicking the analysis-specific R icon. At the moment, the R syntax works only within JASP itself, where it can be used to reproduce analyses and control the GUI. In the near future, the R syntax produced by JASP will also work in R Studio.

✓ JASP keeps the interface simple. Many for-profit statistical software programs now include so many analyses that novice users find it hard to see the forest for the trees. JASP addresses this problem by using add-on modules, similar to how R users can add complexity by load-

[14] This file can be unzipped to explore the separate elements that together constitute a .jasp file.



The JASP previewer allows users to inspect the output of a .jasp file on the OSF. The graph shown on the cell phone displays the *Anscombosaurus*. Figure available at `https://osf.io/m6bi8/` under a CC-BY license.

ing R packages. Thus, 'base JASP' offers a clean and concise set of popular analyses. More advanced analyses are available through dedicated JASP modules, whose contents can be activated by checking boxes.

The JASP Community

There is a growing community of JASP users consisting of students, teachers, and researchers with widely different levels of statistical expertise. If you want to stay abreast of the latest JASP developments, or if you wish to learn more about JASP, we can recommend the following resources:

✓ JASP Website. The JASP website `jasp-stats.org` not only contains the latest version of the program but also offers background information, supporting materials, and teaching tools.

✓ JASP X and JASP Mastodon. The JASP X (formerly Twitter) account @JASPStats and the JASP Mastodon account @JASPStats@fosstodon.org bring all the latest news about JASP.

✓ JASP Facebook. The JASP Facebook group JASPStats keeps its members up to date about new releases and other important events.

✓ JASP Forum. The JASP/BayesFactor Forum at `http://forum.cogsci.nl/` is where you can discuss JASP input and output. You can also check the earlier topics to see whether your question has already been addressed.

✓ JASP Blog. The JASP blog (`https://jasp-stats.org/blog/`) features tutorial posts on particular statistical analyses, posts announcing new versions, and posts about new JASP materials.

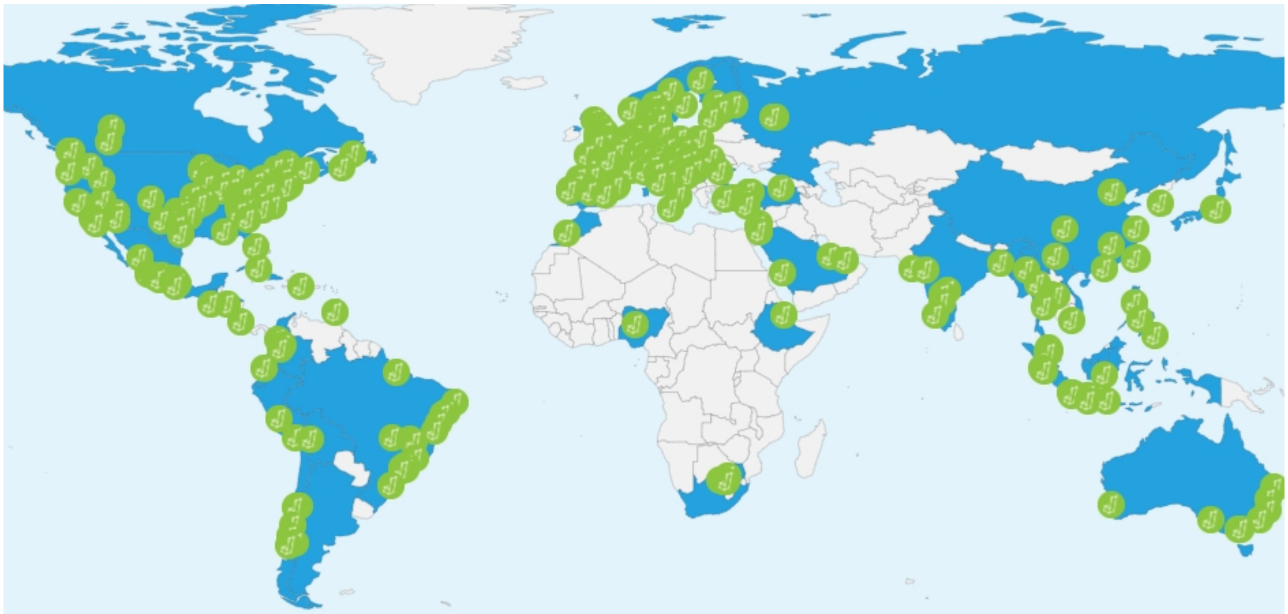
✓ JASP YouTube. The JASP YouTube channel (`https://www.youtube.com/channel/UCSulowI4mXFyBkw3bmp7pXg`) contains tutorial videos about JASP. If you search YouTube you will also find many other JASP tutorial videos.[15]

✓ JASP GitHub. The JASP GitHub page can be used for feature requests and for bug reports (both are considered 'issues', `https://github.com/jasp-stats/jasp-desktop/issues`). We pay keen attention to all suggestions for improvement. Advanced programmers can also use the GitHub page to contribute code.

✓ JASP Workshop. An excellent way to learn about Bayesian inference and JASP is to attend our annual two-day summer workshop in Amsterdam. You can register on the JASP website. We occasionally accept offers to organize the JASP workshop at other universities or institutes, either in a one-day or a two-day format.

[15] A good place to start is the list provided at `https://jasp-stats.org/2020/02/11/how-to-use-jasp-jasp-on-youtube/`.

✓ Bayesian Spectacles Blog. The blog at BayesianSpectacles.org covers all things Bayesian, and often features JASP-related content.



A world map showing 284 universities from 66 different countries where we know that teachers are using JASP. The map is not complete, so if your university is not listed, please let us know at communications@jasp-stats.org. Figure taken from https://jasp-stats.org/teaching-with-jasp/ on September 3rd, 2023. Not shown: University of Hawaii at Hilo.

Alternative Statistical Software Packages

There are other statistical software packages whose goals are similar to those of JASP. As far as inclusion of Bayesian procedures is concerned, JASP is closely aligned with the BayesFactor package in R (Morey and Rouder 2018). Another set of flexible Bayesian tools is offered by the popular programs BUGS (e.g., Lunn et al. 2012), JAGS (Plummer 2003), and Stan (Carpenter et al. 2017).[16] Other recently developed statistical packages for Bayesian analyses include blavaan (Merkle and Rosseel 2018), brms (Bürkner 2017), and Bayesian Regression (Karabatsos 2017). For classical analyses, we like to single out PSPP (https://en.wikipedia.org/wiki/PSPP) as a worthwhile alternative to for-profit statistical software such as IBM's SPSS.

[16] We are enthusiastic about these probabilistic programming languages (see, for instance, Lee and Wagenmakers 2013 and www.bayesmodels.com). If all students and researchers were comfortable programming in JAGS or Stan, the need for JASP would be much less acute.

Chapter Summary

Armed with JASP, a comprehensive Bayesian analysis is just a few mouse clicks away. Several add-on JASP modules (e.g., *Learn Bayes*, *JAGS*, and *Distributions*) have been developed to accompany this book and enhance your learning experience.

Want to Know More?

✓ Goss-Sampson, M. A. (2020). *Bayesian Inference in JASP: A Guide for Students*. Available from `https://jasp-stats.org/jasp-materials/`.

✓ Ly, A., van den Bergh, D. and Bartoš, F., & Wagenmakers, E.–J. (2021). Bayesian Inference With JASP. *The ISBA Bulletin*, 28, 7-15.

✓ Navarro, D. J., Foxcroft, D. R., & Faulkenberry, T. J. (2019). *Learning Statistics With JASP: A Tutorial for Psychology Students and Other Beginners*. Available from `https://learnstatswithjasp.com`.

✓ Wagenmakers, E.–J., Marsman, M., Jamil, T., Ly, A., Verhagen, A. J., Love, J., Selker, R., Gronau, Q. F., Šmíra, M., Epskamp, S., Matzke, D., Rouder, J. N., Morey, R. D. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review*, 25, 35-57.

✓ Wagenmakers, E.–J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, A. J., Selker, R., Gronau, Q. F., Dropmann, D., Boutin, B., Meerhoff, F., Knight, P., Raj, A., van Kesteren, E.–J., van Doorn, J., Šmíra, M., Epskamp, S., Etz, A., Matzke, D., de Jong, T., van den Bergh, D., Sarafoglou, A., Steingroever, H., Derks, K., Rouder, J. N., & Morey, R. D. (2018). Bayesian inference for psychology. Part II: Example applications with JASP. *Psychonomic Bulletin & Review*, 25, 58-76.

The contents of the last two articles may suggest that the presence of classical procedures in JASP is mostly an elaborate ruse to draw in as many unsuspecting users as possible, with the sole objective of turning them into Bayesians. We strongly deny this, of course.

Come for the p-value, stay for the posterior? Figure available at `BayesianSpectacles.org` under a CC-BY license.

Part I

Probability

1 *Probability Belongs Wholly to the Mind?*

This chapter is based almost entirely on a blog post for `BayesianSpectacles.org`: "The Merovingian, or why probability belongs wholly to the mind".

There is no doubt in lightning as to the point it shall strike; in the greatest storm there is nothing capricious; not a grain of sand lies upon the beach, but infinite knowledge would account for its lying there; and the course of every falling leaf is guided by the principles of mechanics which rule the motions of the heavenly bodies.

Jevons, 1874

Chapter Goal

This chapter makes the case that we are all victims of causality. Consequently, probability belongs wholly to the mind. The scientific verdict on this matter is still out –perhaps probability belongs only *mostly* to the mind– but the main purpose of this chapter is to have some philosophical fun and get accustomed to the fact that probability quantifies lack of knowledge.


Lambert Wilson (1958–), the French author who played the role of 'the Merovingian' in *The Matrix Reloaded* and *The Matrix Revolutions*. Photo taken by Georges Biard, available on Wikipedia under a CC BY-SA 3.0 license.

The Merovingian

The famous Matrix trilogy is set in a dystopian future where most of mankind has been enslaved by a computer network, and the remaining rebels find themselves on the brink of extinction. Just when the situation seems beyond salvation, a messiah –called Neo– is awakened and proceeds to free humanity from its silicon overlord. Rather than turn the other cheek, Neo's main purpose seems to be the physical demolition of his digital foes ('agents'), a task that he engages in with increasing gusto and efficiency. Aside from the jaw-dropping fight scenes, the Matrix movies also contain numerous references to religious themes and philosophical dilemmas. One particularly prominent theme is the concept of free will and the nature of probability.

Consider for instance the dialogue in the second movie, 'The Matrix Reloaded', where Neo and his friends Morpheus and Trinity visit an old computer program known as the Merovingian (played by Lambert Wilson) and his wife Persephone. Seated at a long table in an expensive restaurant, the Merovingian introduces himself as a "a trafficker of information". After a while, the following conversation ensues:

Merovingian: "It is, of course, the way of all things. You see, there is only one constant, one universal, it is the only real truth: causality. Action – reaction; cause – and effect."

Morpheus: "Everything begins with choice."

Merovingian: "No. Wrong. Choice is an illusion, created between those with power, and those without. (...) This is the nature of the universe. We struggle against it, we fight to deny it, but it is of course pretense, it is a lie. Beneath our poised appearance, the truth is we are completely out of control. Causality. There is no escape from it, we are forever slaves to it. Our only hope, our only peace is to understand it, to understand the 'why'." [The Merovingian stands up from the table]

Persephone: "Where are you going?"

Merovingian: "Please, ma cherie, I've told you, we are all victims of causality. I drink too much wine, I must take a piss. Cause and effect. Au revoir."[1]

[1] Dialogue taken from `http://www.scottmanning.com/content/merovingian-matrix-reloaded-transcript/`.

The philosophical position advocated by the Merovingian is known as *determinism*, the idea that nothing in the universe is capricious or random, but that everything is ultimately governed by cause-effect relations embodied in physical laws. In other words, everything that happens, happens for a reason, even though that reason (the Merovingian's 'why') may be unknown to an ignorant observer. In a deterministic universe, the past establishes the future without fail: for instance, the fact that you are reading these words right now was already in the stars millions of years ago, as no other world is possible other than the one that we currently inhabit.

One does not need to believe in a fully deterministic universe in order to embrace the Bayesian view on probability.[2] Yet, the Bayesian view is certainly consistent with the idea of a deterministic universe, because 'probability' in the Bayesian sense refers to a lack of information; complete certainty of knowledge is indicated by a probability of 0 or 1, with intermediate values specifying different degrees of belief. For Bayesians, 'probability' and 'plausibility' mean the same thing.

[2] Indeed, the Bayesian hero of this book, Sir Harold Jeffreys, rejected determinism.

Determinism was quite popular among Bayesian pioneers hundreds of years ago. For instance, Pierre-Simon Laplace proposed a particularly strong version of determinism – namely that a hypothetical being with a sufficiently high intelligence (a 'demon') could, from complete knowledge of the present, perfectly predict the future and perfectly reconstruct the past. The idea of determinism was also popular among philosophers in antiquity; for instance, the following fragment by Marcus Tullius Cicero anticipates Laplace's demon by almost 2,000 years:

> "Since, then, everything happens by fate (as will be shown elsewhere) if there could be any mortal who could observe with his mind the interconnection of all causes, nothing indeed would escape him. For he who knows the causes of things that are to be necessarily knows all the things that are going to be. But since no one but God could do this, what is left for man is that he should be aware of future things in advance by certain signs which make clear what will follow. For the things which are going to be do not come into existence suddenly, but the passage of time is like the unwinding of a rope, producing nothing new but unfolding what was there at first." (Cicero, de Divinatione I, lvi; part of Quintus Cicero's defense of divination)

Want of Art

William Stanley Jevons is mostly known for his groundbreaking work in the mathematical study of economics. In addition, Jevons was a prominent logician, and his 1874 book 'The Principles of Science: A Treatise on Logic and Scientific Method' stands as an enduring witness to his brilliance as a scientist and as a writer.

W. Stanley Jevons (1835-1882) at age 23. Copyright owned by the National Portrait Gallery, London, under a CC-BY-ND license.

Jevons' view on probability and statistical inference was influenced by Augustus De Morgan, who in turn was influenced by Laplace. Although many great scientists have enthusiastically advocated determinism, few have done so as eloquently as Jevons. Chapter 10 of the 'Principles' is devoted to the theory of probability. Jevons starts the chapter with a fragment that we are reprinting here in full:

> "The subject upon which we now enter must not be regarded as an isolated and curious branch of speculation. It is the necessary basis of the judgments we make in the prosecution of science, or the decisions we come to in the conduct of ordinary affairs. As Butler truly said, 'Probability is the very guide of life.' Had the science of numbers been studied for no other purpose, it must have been developed for the calculation of probabilities. All our inferences concerning the future are merely probable, and a due appreciation of the degree of probability depends upon a comprehension of the principles of the subject. I am convinced that it is impossible to expound the methods of induction in a sound manner, without resting them upon the theory of probability. Perfect knowledge alone can give certainty, and in nature perfect knowledge would be infinite knowledge, which is clearly beyond our capacities. We have, therefore, to content ourselves with partial knowledge knowledge mingled with ignorance, producing doubt.
>
> A great difficulty in this subject consists in acquiring a precise notion of the matter treated. What is it that we number, and measure, and calculate in the theory of probabilities? Is it belief, or opinion, or doubt, or knowledge, or chance, or necessity, or want of art? Does probability exist in the things which are probable, or in the mind which regards them as such? The etymology of the name lends us no assistance: for, curiously enough, *probable* is ultimately the same word as *provable*, a good instance of one word becoming differentiated to two opposite meanings.

The logic piano: a mechanical computer designed by Jevons in 1866 to solve problems in logic. Inv. 18230. ©History of Science Museum, University of Oxford. Usage granted until 2031.

Chance cannot be the subject of the theory, because there is really no such thing as chance[3], regarded as producing and governing events. The word chance signifies *falling*, and the notion of falling is continually used as a simile to express uncertainty, because we can seldom predict how a die, a coin, or a leaf will fall, or when a bullet will hit the mark. But everyone sees, after a little reflection, that it is in our knowledge the deficiency lies, not in the certainty of nature's laws. There is no doubt in lightning as to the point it shall strike; in the greatest storm there is nothing capricious; not a grain of sand lies upon the beach, but infinite knowledge would account for its lying there; and the course of every falling leaf is guided by the principles of mechanics which rule the motions of the heavenly bodies.

Chance then exists not in nature, and cannot coexist with knowledge; it is merely an expression, as Laplace remarked, for our ignorance of the causes in action, and our consequent inability to predict the result, or to bring it about infallibly. In nature the happening of an event has been pre-determined from the first fashioning of the universe. *Probability belongs wholly to the mind*." (Jevons 1874/1913, pp. 197-198)

[3] EWDM: The same sentiment was expressed by De Moivre (1718/1756, p. 253): "Chance (...) can neither be defined nor understood".

"There is no result in nature without a cause; understand the cause and you will have no need of the experiment." (Leonardo da Vinci)

An Interview with Einstein

In the 1920s, Nazi propagandist and Mussolini-admirer George Viereck managed to secure an interview with Albert Einstein. This interview was published in 1929 in *The Saturday Evening Post* under the title "What Life Means to Einstein". From the perspective of determinism, two of Einstein's statements stand out. First, when asked whom he felt was to blame for the downfall of Germany in World War I, Einstein concludes his answer as follows: "In a sense, we can hold no one responsible. I am a determinist. As such, I do not believe in free will." Second, later in the interview there is the following exchange:

Einstein: "I am happy because I want nothing from anyone. I do not care for money. Decorations, titles or distinctions mean nothing to me. I do not crave praise. The only thing that gives me pleasure, apart from my work, my violin and my sailboat, is the appreciation of my fellow workers."

Viereck: "Your modesty does you credit."

Einstein: "No. I claim credit for nothing. Everything is determined, the beginning as well as the end, by forces over which we have no control. It is determined for the insect as well as for the star. Human beings, vegetables or cosmic dust, we all dance to a mysterious tune, intoned in the distance by an invisible player."

A Deterministic View on Life

Many people believe that the future is partly in their own hands. We can usually choose freely whether to watch TV, or read a book, or go to the movies; we decide where to go on vacation, what to eat, whom to marry, and so on. There appears to be no external authority who commands us in such decisions, big and small; in this sense *we can do what we want*. This 'free will' perspective suggests that many possible futures remain open to us, and that we are in control of our own destiny, at least to some degree.[4]

The fact that we can do what we want, however, does not present a compelling argument against determinism. Yes, we may watch TV because we feel like it – but where did that feeling come from? A determinist believes that 'free will' is merely an illusion. You may experience the desire to do something and then do it, but that desire itself is the inevitable result of a myriad causal factors that were set in motion since the beginning of time. As summarized by Schopenhauer: "You can *do* what you *will*: but at each given moment of your life you can *will* only one determined thing and by no means anything other than this one."[5]

This deterministic perspective on life is visualized in Figure 1.1. The white lighting bolt running from top to bottom represents your life path, from which no deviation whatsoever is possible. The black lightning bolts in the top panel represent alternative life paths that you now know were always closed to you. It is not just that these alternative realities did not happen; they could never have happened. For instance, it would be tempting to think "had I not folded my hand but called her bluff instead then I would have won the poker tournament"; instead, the correct deterministic thought is "I now know that I did not call her bluff, and did not win the poker tournament". The purple lighting bolts in the bottom panel represent alternative life paths that you do not yet know will never materialize. It is tempting to think "If I participate in this lottery and I'm lucky, I may win the jackpot"; a determinist would correct this to "I do not yet know whether or not I will win the lottery. However, this is not an eventuality or a matter of luck – it is a certainty, but one of which I will only become aware after the fact."

An apt analogy is presented by Schopenhauer: "(...) we ought to regard events as they occur with the same eye as the print that we read, knowing full well that it stood there before we read it." When in the middle of a book, you know how the story started but you are still unsure about how it will end – but it can end in only one way, just as it started in only one way. For a determinist, the difference between what lies in the past and what lies in the future can therefore be attributed solely to a difference in knowledge.

[4] A figure that represents this perspective and formed the inspiration for this section is available at `https://twitter.com/waitbutwhy/status/1367871165319049221/photo/1`.

[5] See the section 'Want to Know More' for details on Schopenhauer's perspective.

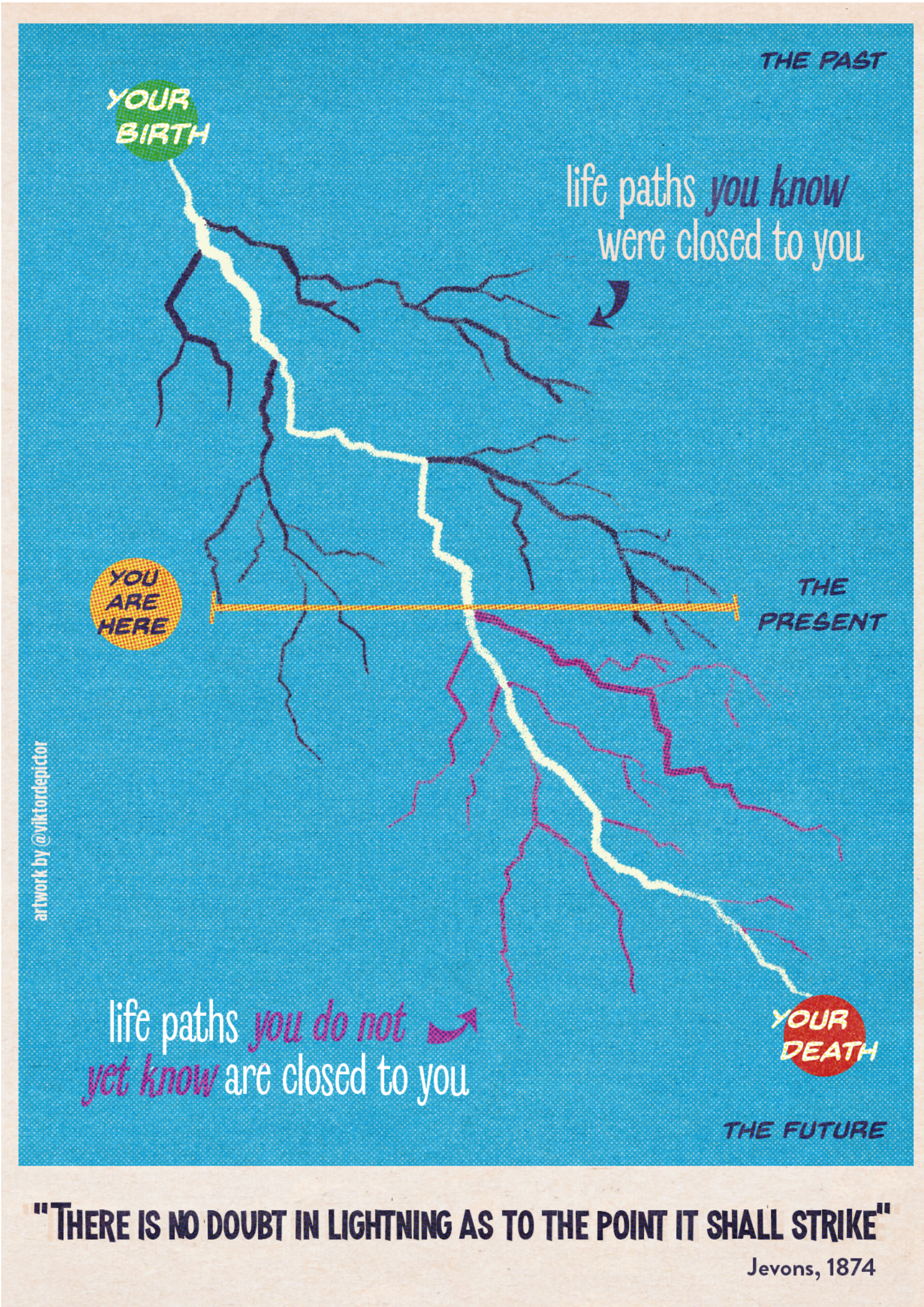


Figure 1.1: Figure available at `BayesianSpectacles.org` under a CC-BY license.

A Quantum Fly in the Deterministic Ointment

Readers with a background in physics may believe that hard-core determinists have gone the way of the dinosaur, with the theory of quantum mechanics providing the trigger for a mass extinction event. For instance, Hacking (1990, p. 1) remarks "The most decisive conceptual event of twentieth century physics has been the discovery that the world is not deterministic. Causality, long the bastion of metaphysics, was toppled, or at least tilted: the past does not determine exactly what happens next."

Specifically, the orthodox 'Copenhagen' interpretation of quantum mechanics holds that chance is inherent to nature, and that the behavior of the tiniest particles is fundamentally unpredictable. There exists no hidden deterministic structure that would allow us to calculate, say, the exact moment when a particular radioactive atom decays. The very fabric of our universe is capricious, and this is quite contrary to what most researchers believed in Jevons' time.[6]

Although the Copenhagen interpretation dominates the literature and the textbooks, there has always been opposition. The pragmatic attitude of many physicists towards discussions on the meaning of quantum mechanics is perhaps best summed up by the statement "shut up and calculate", that is, "stop philosophizing about the meaning of quantum uncertainty and make better use of your time by deriving the predictions for the next quantum experiment or application".

> "And yet...there are just *too many* loose ends in the conventional description of the quantum world. Phenomena that seem to make no sense. Assumptions that contradict themselves. Explanations that don't explain. And underneath it all is an uncomfortable truth, swept under the carpet with undue haste because it's deeply embarrassing: the 'shut up and calculate' brigade don't really understand it either." (Stewart 2019, p. 226)

In fact, there exist deterministic accounts of quantum phenomena (e.g., the de Broglie-Bohm *pilot wave theory*, or Hugh Everett III's *Many-Worlds Interpretation*) that provide an alternative to the Copenhagen interpretation. The relative popularity of the various interpretations has been assessed by polls at least six times, usually among physicists attending conferences on quantum mechanics. The Copenhagen interpretation was preferred by 13/48 (27%) respondents in Tegmark (1997); by 8/90 (9%) respondents in Tegmark and Wheeler (2001); by 14/33 (42%) respondents in Schlosshauer et al. (2013); by 2/18 (11%) respondents in Sommer (2013); by 3/76 (4%) respondents in Norsen and Nelson (2013); and by 59/149 (39%) respondents in Sivasundaram and Nielsen (2016). Overall, the Copenhagen interpretation was preferred by 99/414 (24%)

[6] But it is remarkably consistent with the physical universe postulated by the Greek philosopher Epicurus (341-270 BC), who believed that all matter was composed of atoms, and that these atoms sometimes behaved capriciously. Throughout history, Epicurus and his followers were widely ridiculed for propagating such absurdities.

The societal impact of quantum mechanics is *immense*. Tegmark and Wheeler (2001, p. 69) state that "today an estimated 30 percent of the U.S. gross national product is based on inventions made possible by quantum mechanics, from semiconductors in computer chips to lasers in compact-disc players, magnetic resonance imaging in hospitals, and much more."

respondents. The opinion on the matter does not appear settled, and poll-to-poll differences are substantial.

In conclusion, despite the onslaught from quantum mechanics, determinism is still alive. In the Netherlands, one of its most prominent advocates is the physics Nobel laureate Gerard 't Hooft (2016).[7] In the words of Cicero (45BC/1956b, I, vi), "Surely such wide diversity of opinion among men of the greatest learning on a matter of the highest moment must affect even those who think that they possess certain knowledge with a feeling of doubt."

[7] See also the YouTube videos by the physicist Sabine Hossenfelder, whose preferred account is known as *superdeterminism*. "I know it is somewhat boring coming from a German, but I think Einstein was right about quantum mechanics. Call me crazy if you want, but for me it is obvious that superdeterminism is the correct explanation for our observations. I just hope I'll live long enough to see that all those men who said otherwise will be really embarrassed." `https://youtu.be/ytyjgIyegDI`

Exercises

1. Based on the literature, what do you believe is the most compelling argument against determinism?

2. Why doesn't it matter for the Bayesian learning process whether or not the universe is deterministic?

3. In the section 'Want to know more?' below, read the summary of Schopenhauer's essay on free will. Suppose that the Copenhagen interpretation of quantum mechanics is correct. Does this salvage the concept of free will?

Chapter Summary

For a determinist, probability is nothing but a reflection of our knowledge, a number that quantifies our degree of reasonable belief, our certainty, or the intensity of our conviction.

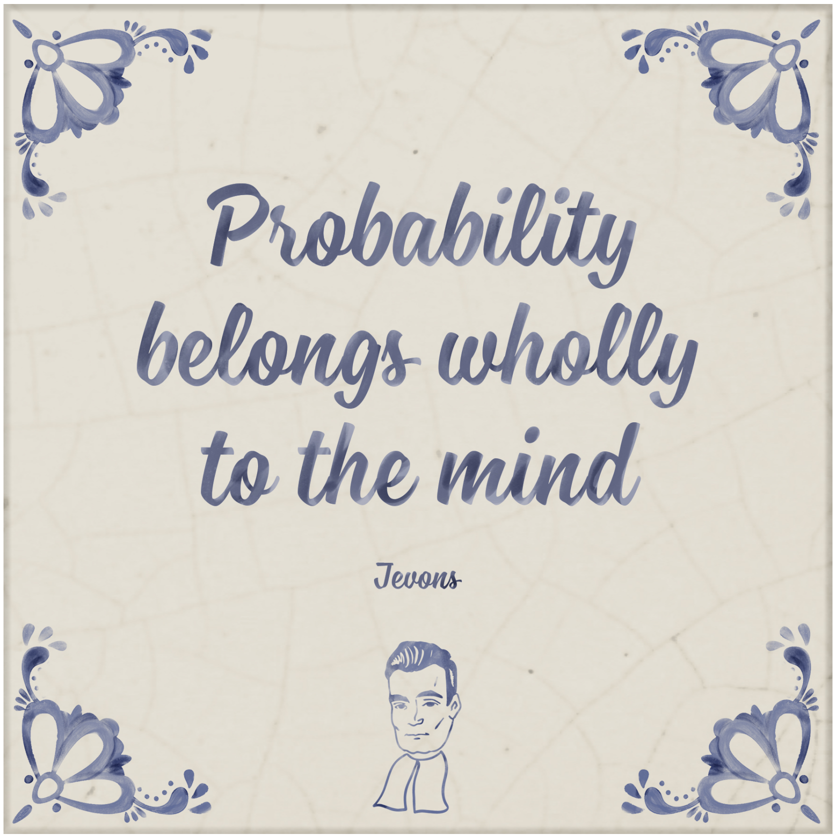
"Every phenomenon, however minute, has a cause; and a mind infinitely powerful, infinitely well-informed about the laws of nature, could have foreseen it from the beginning of the centuries. If such a mind existed, we could not play with it at any game of chance; we should always lose. In fact for it the word chance would not have any meaning, or rather there would be no chance. It is because of our weakness and our ignorance that the word has a meaning for us. And, even without going beyond our feeble humanity, what is chance for the ignorant is not chance for the scientist. Chance is only the measure of our ignorance. Fortuitous phenomena are, by definition, those whose laws we do not know." (Poincaré 1913, p. 395)

Want to Know More?

✓ Barrett, L., & Connell, M. (2005). Jevons and the Logic 'Piano'. *The Rutherford Journal*, 1, `http://rutherfordjournal.org/article010103.html`. Provides a brief account of Jevons' role in the development of logic. More details on the logic piano can be found in Jevons (1874/1913, pp. 123-131), Jevons (1870a), and Jevons (1870b).

✓ Cicero, M. T. (45 BC/1956). *Academica*. (H. Rackham, Trans.) London: William Heinemann LTD. All of Cicero's work is highly recommended.

✓ Cicero, M. T. (45 BC/1956). *de Natura Deorum*. (H. Rackham, Trans.) London: William Heinemann LTD. All of Cicero's work is highly recommended, but this is perhaps our favorite.



Statement by W. Stanley Jevons in *The Principles of Science*, 1874. Figure available at `BayesianSpectacles.org` under a CC-BY license.

✓ Cicero, M. T. (44 BC/1923). *de Devinatione*. (W. A. Falconer, Trans.) London: Harvard University Press. Did we mention that all of Cicero's work is highly recommended?

✓ Diaconis, P., & Skyrms, B. (2018). *Ten Great Ideas About Chance*. Princeton: Princeton University Press. "Consider tossing a coin just once. The thumb hits the coin; the coin spins upward and is caught in the hand. It is clear that if the thumb hits the coin in the same place with the same force, the coin will land with the same side up. Coin tossing is physics, not random! To demonstrate this, we had the physics department build us a coin-tossing machine. The coin starts out on a spring, the spring is released, the coin spins upward and lands in a cup (...) Because the forces are controlled, the coin always lands with the same side up. This is viscerally quite disturbing (even to the two of us). Magicians and crooked gamblers (including one of your authors) have the same ability." (pp. 10-11).

✓ Galavotti, M. C. (2005). *Philosophical Introduction to Probability*. Stanford: CSLI Publications. This highly recommended book provides a good overview of the main interpretations of probability.

✓ Earman, J. (1986). *A Primer on Determinism*. Dordrecht: Reidel. One of my colleagues, Louise, saw me read this book and asked 'so what is it about?' 'Well,' I answered, 'the author of this book investigates the claim that, millions of years ago, it was already 100% certain that you were going to ask me this very question at this particular time.' Louise immediately replied 'oh, so this book is just nonsense.' Despite Louise's negative first impression, the Earman book is the reference work on determinism, and will remain so for a long time to come. Unfortunately, the matter is complicated and a good understanding of the relevant concepts requires knowledge of classical physics, general relativity, and quantum theory.

Theoretical physicist Dr. Sabine Hossenfelder (1976-), photographed in 2017. Hossenfelder is also a philosopher of science and author of several popular science books. In 2023, her YouTube channel has 728,000 subscribers.

✓ Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J., & Krüger, L. (1989). *The Empire of Chance*. Cambridge: Cambridge University Press.

✓ Hacking, I. (1990). *The Taming of Chance*. Cambridge: Cambridge University Press.

✓ Hossenfelder, S. (2022). *Existential Physics: A Scientist's Guide to Life's Biggest Questions*. Viking.

> "However, if you know one thing about quantum mechanics, it's that its physical interpretation has remained highly controversial. In 1964, more than half a century after the theory was established, Richard Feynman told his students, "I can safely say that nobody understands quantum mechanics." After another half century, in 2019, the physicist Sean Carroll wrote that "even physicists don't understand quantum mechanics." (...) if you don't believe the measurement update [the inherently probabilistic collapse of the wave function – EWDM] is fundamentally correct, that's currently a scientifically valid position to hold. I myself think it's likely the measurement update will one day be replaced by a physical process in an underlying theory, and it might come out to be both deterministic and time-reversable again." (pp. 16-17)

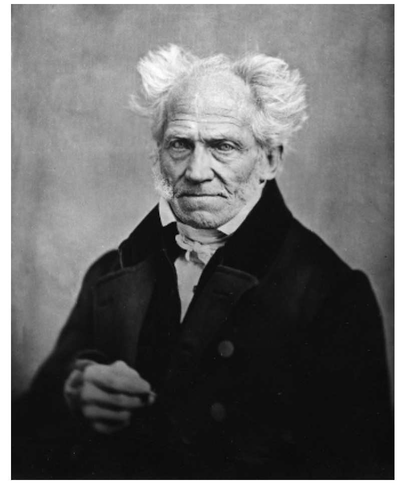
Portrait of W. Stanley Jevons (1835-1882) at age 42, by G. F. Stodart.

✓ Jevons, W. S. (1874/1913). *The Principles of Science: A Treatise on Logic and Scientific Method*. London: MacMillan. Timeless classic by a brilliant author, and freely available online.

✓ Laplace, P.–S. (1814/1902/1995). *[A] Philosophical Essay on Probabilities*. A surprisingly accessible essay by one of the most brilliant minds of all time. The French first edition, *Essai Philosophique sur les Probabilités*, was published in 1814; the 1902 English translation by Truscott and Emory leaves something to be desired; the 1995 English translation by Andrew I. Dale is superb, and also presents an appendix with useful notes that put the work in a modern perspective.

✓ Schabas, M. (1990). *A World Ruled by Number: William Stanley Jevons and the Rise of Mathematical Economics*. Princeton: Princeton University Press. A monograph on Jevons, reviewed by Zabell (1992). Other monographs include Peart (1996), Maas (2005), and Mosselmans (2007). Notable articles on Jevons include Jevons and Jevons (1934), Keynes (1936), and Robertson (1951).

✓ Schopenhauer, A. (2009). *The Two Fundamental Problems of Ethics*. Cambridge: Cambridge University Press. The original German edition dates from 1841 and is entitled *Die Beiden Grundprobleme der Ethik*. In the first treatise, Schopenhauer considers the idea of free will, and concludes that it is an illusion. Specifically, Schopenhauer argues that "You can *do* what you *will*: but at each given moment of your life you can *will* only one determined thing and by no means anything other than this one." (p. 48). This argument is based on determinism: "The law of causality stands firm *a priori* as the universal rule to which all real objects in the external world without exception are subordinated." (p. 50) Schopenhauer then explains that the exact nature of causality becomes more difficult to grasp when the systems under study become increasingly complex; however, this does not mean that causality is suddenly absent: "So, throughout this ever increasing heterogeneity, incommensurability and unintelligibility of the relation between cause and effect, has the *necessity* it presupposes also decreased at all? In no way, not in the slightest. As necessarily as the rolling ball sets the one at rest in motion, so too must the Leyden flask discharge itself when touched by the other hand, so must arsenic kill any living thing, so must the seed grain that was stored dry and showed no alteration through millennia germinate, grow and develop into a plant as soon as it is placed in the appropriate soil and exposed to the influences of air, light, heat and moisture. The cause is more complicated, the effect more heterogeneous, but the necessity with which it occurs is not one hair's breadth smaller." (p. 59) After some deeper reflections, Schopenhauer then concludes "It is definitely neither metaphor nor hyperbole, but a quite dry and literal truth, that just as a ball cannot start into motion on a billiard table until it receives an impact, no more can a human being stand up from his chair until a motive draws or drives him away: but then his standing up is as necessary and inevitable as the ball's rolling after the impact." (p. 65) Indeed, "Under presupposition of free will each



German philosopher Arthur Schopenhauer (1788-1860) photographed one year before his death, by J. Schäfer. "Under presupposition of free will each human action would be an inexplicable miracle – an effect without cause." (Schopenhauer 2009, p. 66)

The Schopenhauer paper also features some less compelling fragments. For instance, Schopenhauer claims that "we can stretch and considerably heighten our mental powers through wine or opium" (p. 53). Even more unsettling is that Schopenhauer tries to bolster the case for determinism by suggesting that people can foretell the future: "If we do not assume the strict necessity of all happening by way of a causal chain that links all events without distinction, and instead let it be interrupted in countless places by an absolute freedom, then all *foreseeing of the future*, in dreams, in clairvoyant somnambulism, and in second sight, becomes quite *objectively* and thus absolutely *impossible*, and so unthinkable – because then there is simply no objectively real future with the barest possibility of being foreseen, in contrast with the present situation where we doubt merely its *subjective* conditions and hence its *subjective* possibility. And even this doubt can no longer be accommodated among the well-informed these days, now that countless testimonies, from the most credible quarters, have confirmed such anticipations of the future." (pp. 79-80) human action would be an inexplicable miracle – an effect without cause. (p. 66)

It then follows that "*Everything that happens, from the greatest to the smallest, happens necessarily*. Whatever happens, necessarily happens. Whoever is alarmed at these propositions still has some things to learn and others to unlearn: but after that he will recognize that they are the most abundant source of comfort and relief. – Our deeds are truly no first beginning, and so in them nothing really new attains existence: rather *through what we do, we merely come to experience what we are*." (p. 79) And "Wishing that some incident had not happened is a foolish self-torment: for it means wishing something absolutely impossible, and is as irrational as the wish that the sun should rise in the West. Because every happening, great or small, occurs *strictly* necessarily, it is totally vain to reflect on how trivial and accidental were the causes that brought about that incident and how very easily they could have been different. For this is illusory, in that they all occurred with just as strict a necessity and had their effect with just as much power as those in consequence of which the sun rises in the East. Rather we ought to regard events as they occur with the same eye as the print that we read, knowing full well that it stood there before we read it."

✓ Stigler, S. M. (1999). *Statistics on the Table: The History of Statistical Concepts and Methods*. Cambridge, MA: Harvard University Press. Chapters 3 and 4 of this riveting book center on the contribution of Jevons to statistics.

✓ Tegmark, M., & Wheeler, J. A. (2001). 100 years of quantum mysteries. *Scientific American, 284*, 68-75. A historical overview of quantum mechanics and a positive evaluation of the Many-Worlds Interpretation (main problem: "The bizarreness of the idea"). For longer treatments critical of the Copenhagen dominance see Kumar (2009) and Becker (2018). A clear classical description is in Feynman (1965/1992).

"Probability, which necessarily implies uncertainty, is a consequence of our ignorance. To an omniscient Being there can be none. Why, for instance, if we throw up a shilling, are we uncertain whether it will turn up head or tail? Because the shilling passes, in the interval, through a series of states which our knowledge is unable to predict or to follow. If we knew the exact position and state of motion of the coin as it leaves our hand, the exact value of the final impulse it receives, the laws of its motion as affected by the resistance of the air and gravity, and finally the nature of the ground at the exact spot where it falls, and the laws regulating the collision between the two substances, we could predict as certainly the result of the toss as we can which letter of the alphabet will be drawn after twenty-five have been taken and examined. The probability, or amount of conviction accorded to any fact or statement, is thus essentially subjective, and varies with the degree of knowledge of the mind to which the fact is presented" (Crofton 1885, p. 768)

William Stanley Jevons and the Poor

Jevons's accomplishments in science are impressive. Robertson (1951, p. 247) states that "Within his theoretical framework, he moved incisively to the solution of problems in the real world in a way that no one before him had been able to do. If this does not constitute a claim to consideration as the founder of econometric method, I do not know what does." In this book, we will cite Jevons often and at length, as his writings on probability are clear, poetic, and compelling. However, the modern reader is likely to raise an eyebrow when it comes to Jevons's strong opposition to state support for the poor. As summarized by Keynes (1936, p. 544):

> "On the side of morals and sentiment Jevons was, and always remained, an impassioned individualist. There is a very odd early address of his, delivered to the Manchester Statistical Society in 1869, in which he deplores free hospitals and medical charities of all kinds, which he regarded as undermining the character of the poor (which he seems to have preferred to, and deemed independent of, their health). "I feel bound," he said, "to call in question the policy of the whole of our medical charities, including all free public infirmaries, dispensaries, hospitals, and a large part of the vast amount of private charity. What I mean is that the whole of these charities nourish in the poorest classes a contented sense of dependence on the richer classes for those ordinary requirements of life which they ought to be led to provide for themselves."."

William Stanley Jevons: A Burning Sense of Vocation

William Stanley Jevons (1835–1882) is primarily known for pioneering the mathematical treatment of economics. Based on carefully collected sets of observations, Jevons would model and predict the fluctuations of various economic indices including the price of gold and of wheat. In his first book, *The Coal Question*, Jevons suggested that an increasing demand on coal would exhaust the mines, resulting in dire economic consequences for the British Empire. This concern for a depletion of natural resources extended to Jevons's personal life: not only did he collect a vast number of books on economics, but he also hoarded thin brown packing paper, to such a degree that "even today, more than fifty years after his death, his children have not used up the stock he left behind him." (Keynes 1936, p. 523)

In the *The Theory of Political Economy*, Jevons put in mathematical form the idea of prospective utility based on the anticipation of pleasure and pain, that is, "If laborious action be regarded as having a positive value on account of its pecuniary reward and a negative value on account of the toilsome feelings which accompany it, the action will be carried on only so long as the individual contemplates a preponderating amount of satisfaction." (Robertson 1951, p. 237) In *The Principles of Science: A Treatise on Logic and Scientific Method*, "Jevons reduced logical inference to a simple but complete system, and defined the inductive or scientific method, showing its unity in all sciences, and the fundamental importance of the theory of probability." (Jevons and Jevons 1934, p. 232)

In *The Power of Numerical Discrimination*, Jevons describes the first experiment on what is now known as 'subitizing', the mind's ability to "comprehend and count" small numbers "by an instantaneous and apparently single act of mental attention." (Jevons 1871, p. 281) Jevons "had genius and divine intuition and a burning sense of vocation" (Keynes 1936, p. 545), but his frenzy of academic activity was unfortunately cut short at the age of 46:

> "Jevons was drowned while bathing on the south coast of England in August 1882, the shock of the cold water proving too much for his enfeebled health. He was a few weeks short of forty-seven years of age. He left a wife who had been a constant companion and help in his work, and three small children, too young to understand its nature." (Jevons and Jevons 1934, p. 231)

2 *Epistemic and Aleatory Uncertainty*

PROBABILITY DOES NOT EXIST

Bruno de Finetti, 1974, 'Theory of Probability'.

Chapter Goal

Probability is a notoriously ambiguous concept, and this chapter aims to clarify the difference between two of its Bayesian interpretations. According to the first interpretation, probability refers to a degree of reasonable belief, an intensity of conviction about the truth of some proposition (e.g., what is the probability that Julius Caesar, upon crossing the Rubicon, truly uttered the phrase "alea iacta est"? What is the probability that Italy will win the next Eurovision song contest? What is the probability that the 100$^{\text{th}}$ digit in the decimal expansion of π is even?). According to the second interpretation, probability (or better: *chance*) refers to the possible realization of a particular event given a data-generating process about which nothing more can be learned (e.g., what is the chance that a fair coin lands heads three times in a row?).[1]

[1] We regard the so-called frequentist definition of probability a historical accident; it is briefly mentioned at the end of this chapter, together with references to relevant background material.

Epistemic Uncertainty

In Bayesian inference, probability is generally understood to refer to a *degree of reasonable belief* (e.g., Jeffreys 1931). Complete confidence in the truth of a proposition is characterized by a probability of 1, a value that can be assigned to tautologies such as $3 = 3$; complete confidence in the falsity of a proposition is characterized by a probability of 0, a value that may be assigned to propositions that have been irrevocably disproved (e.g., 'all swans are white'; 'all Fermat numbers are prime'). Probabilities in between 0 and 1 represent a graded scale of *intensity of conviction*, or *degree of belief*. In this epistemic[2] interpretation, *probability* is synonymous with *plausibility*.

[2] From the Greek word for 'knowledge'.

Because probability refers to a state of uncertain knowledge, it is the property of an observer, not the property of an object. This is consistent with the deterministic idea, outlined in the previous chapter, that probability is a reflection of our ignorance, and hence that 'probability is wholly in the mind'. Consequently, early Bayesians had no trouble accepting that probability is defined by the person making the plausibility assessment, and that different people may have radically different probabilities for the same scenario:

> "(...) carry two men to a room in which are two boxes, one small and ribbed with steel, the other large and roughly put together. Let these men have come from the two most opposite points of the earth in manners and customs, yet they will immediately, when asked, point out which is the larger of the two boxes: if they are both sane, disagreement will be impossible. Now produce a piece of gold, and ask which of the two boxes is filled with that substance. One has seen gold, and knows its value, and also that it is rarely collected in large quantities, or placed in insecure receptacles. He would say that most likely the smaller box, if either, is full of gold. Or he may think that the question and circumstances are so extraordinary, that the former would not have been put unless this case had been a departure from ordinary rules, and may therefore pronounce for the larger box. In either case it is clear that the probability or improbability is the consequence of a state of his own mind, or of an impression existing in himself, in a sense which cannot be, in any view of the case, applied to the extension of the two boxes. If the other man knew nothing of gold, he would not be able to bring his mind to either of the preceding conclusions, in preference to the other. What we mean, then, by an event being probable or improbable, is this; that with regard to that event the mind of the spectator is in a state of disposition either to doubt or believe its happening; which evidently depends in no way upon the event itself, but upon the whole train of previous ideas and associations which the mind of the spectator possesses upon such circumstances as he thinks similar. *Therefore it is wrong to speak of any thing being probable or improbable in itself. The same thing may be really probable to one person and improbable to another. And thus men may be justified in drawing different conclusions upon the same subject.* [italics ours]" (De Morgan 1849, p. 394)

Augustus De Morgan (1806-1871), an early proponent of Bayesian inference and the work of Pierre-Simon Laplace.

Almost a century later, the mantra 'all probability is inherently subjective' resurfaced in the work by Bayesian statisticians such as Frank Ramsey, Jimmy Savage, Dennis Lindley, and Bruno de Finetti. For instance, in the preface to his famous monograph *Theory of Probability*, de Finetti argued explicitly that probability does not have an objective meaning:

> "The abandonment of superstitious beliefs about the existence of Phlogiston, the Cosmic Ether, Absolute Space and Time,..., or Fairies and Witches, was an essential step along the road to scientific thinking. Probability, too, if regarded as something endowed with some kind of objective existence, is no less a misleading misconception, an illusory attempt to exteriorize or materialize our true probabilistic beliefs." (de Finetti 1974, p. x)

Bruno de Finetti (1906–1985), the Bayesian statistician who promoted the idea that probability is always subjective. The 1979 photo is available at `http://www.brunodefinetti.it` and has been reproduced with permission from Fulvia de Finetti.

Instead, probability is a property of the observer:

"Probabilistic reasoning—always to be understood as subjective—merely stems from our being uncertain about something. It makes no difference whether the uncertainty relates to an unforseeable future, or to an unnoticed past, or to a past doubtfully reported or forgotten; it may even relate to something more or less knowable (by means of a computation, a logical deduction, etc.) but for which we are not willing or able to make the effort; and so on." (de Finetti 1974, pp. x-xi)

'Probabilis': Possessed of Verisimilitude

The word *probability* derives from the Latin *probare*, 'to try', which survives in the modern Italian 'provare', the English 'to probe' and the Germanic 'proberen/probieren'. The Latin 'probabilis' was Cicero's translation of the Greek 'pithanos' (persuasive). In Cicero's main works, 'probabilis' is synonymous with 'veri similia' (e.g., Cicero 45BC/1956a, Frag. 19; II, x, xxxi; see Glucker 1995 for a detailed treatment). The concept was proposed earlier by the skeptic philosopher Carneades, whose key ideas were as follows:
(I) The wise man withholds assent. "(…) what is so ill-considered or so unworthy of the dignity and seriousness proper to a philosopher as to hold an opinion that is not true, or to maintain with unhesitating certainty a proposition not based on adequate examination, comprehension and knowledge?" (Cicero 45BC/1956b, I,i)
(II) Even the perceptual information that enters our senses cannot be relied upon as veridical, as is demonstrated by visual illusions and the like. "What can be bigger than the sun, which the mathematicians declare to be nineteen times the size of the earth? How tiny it looks to us!" (Cicero 45BC/1956a, II, xxvi)
(III) In theory, the wise man never assents. In practice, when concrete decisions need to be taken, he is guided by probability, because some propositions are more truth-like than others. "Thus the wise man will make use of whatever apparently probable presentation he encounters, if nothing presents itself that is contrary to that probability, and his whole plan of life will be charted out in this manner." (Cicero 45BC/1956a, II, xxxi)

In Cicero's use, probability or verisimilitude has an epistemic interpretation, as it refers to the judgment of the wise man in deciding to go on a voyage, sow a crop, marry a wife, beget a family, and so on (Cicero 45BC/1956a, II, xxxiv; see also Popper 1972, p. 404). For the wise man, "Probability is the very guide of life" (a popular loose translation of Cicero 45BC/1956b, I, v, 12; see also Jevons' epigraph that starts this book).

Aleatory Uncertainty

Although the Bayesian position is strongly associated with the epistemic interpretation of probability, Bayesians also use an aleatory interpretation.[3] The aleatory interpretation comes into play when we consider a series of similar events in which there is a generally accepted limit on our knowledge. Standard examples include tosses of a coin, throws of a die, and drawings from a deck of cards or from an urn filled with marbles. Concretely, suppose we are about to toss a fair coin. The probability that it lands heads is not a random event – it is governed by the laws of physics and determined by factors such as the rate of spin, the initial velocity, and air resistance (Diaconis et al. 2007, Diaconis and Skyrms 2018). Nevertheless, when asked "what is the probability that a fair coin will land heads on the next toss?" it is assumed that these determining factors are beyond reach, and that, given this lack of knowledge, the degree of belief that the coin will come up heads corresponds to a probability of $.50$, irrespective of the outcomes of previous tosses.[4] Note that, in the Bayesian interpretation, the aleatory probability still refers to a degree of belief; it is not, for instance, defined as a hypothetical limit on a frequency of occurrence.

[3] From the Latin word 'alea', which means 'die'. For instance, "alea iacta est" means "the die is cast".

[4] This is under the crucial assumption that the coin is fair.

Geophysicist and Bayesian statistician Sir Harold Jeffreys gave a pithy definition of chance. If, given a particular state of the world, "(...) the probability of an event is the same at every trial, no matter what may have happened at previous trials, we say that the probability is a *chance*"[5] (Jeffreys 1973, p. 46; see also Jeffreys 1936a, p. 356; Jeffreys 1961, pp. 51-52).

[5] Jeffreys adds: "the term was used in this sense by N. R. Campbell and revived by M. S. Bartlett."

In statistical jargon, the irreducible unpredictability associated with aleatory processes is called *sampling variability*. In terms of the Bayesian learning cycle shown in Figure 1 (p. 14), it refers to the deductive prediction of data from a specific state of the world. To illustrate, Figure 2.1 shows the predicted number of heads when a fair coin is tossed ten times. The chance is small that a fair coin would land heads ten times in a row (i.e., ${}^1/_2{}^{10} = {}^1/_{1024}$); the chance is almost 0.25 that a fair coin shows 5 heads out of 10 tosses.

Epistemic and Aleatory Uncertainty in Practice

In practical application, both epistemic and aleatory uncertainty play a role: there are both unknowns and unknowables. Consider for instance the following scenario:

> "In October 2009, the Dutch newspaper Trouw reported on research conducted by H. Trompetter, a student from the Radboud University in the city of Nijmegen. For her undergraduate thesis, Trompetter had interviewed 121 older adults living in nursing homes. Out of these 121

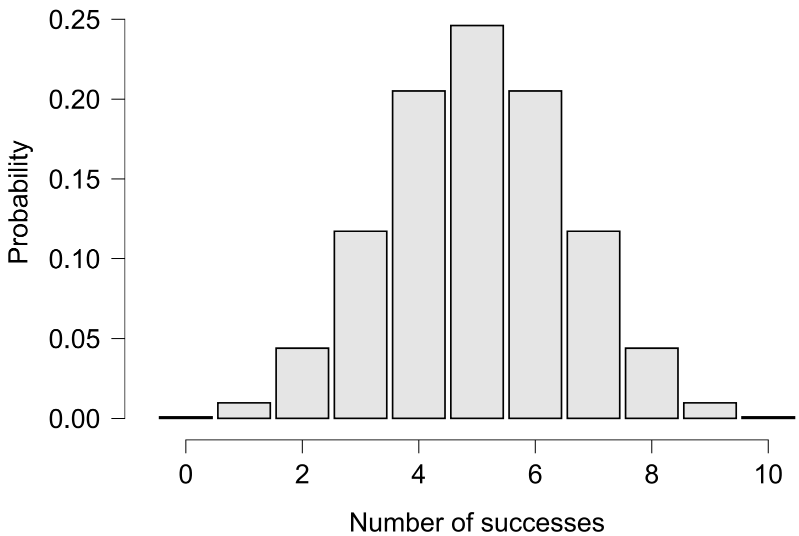


Figure 2.1: Aleatory uncertainty demonstrated for the scenario where a fair coin will be tossed ten times. The *Number of successes* on the x-axis refers to the number of times the coin is predicted to land heads. Figure from the JASP module *Learn Bayes*.

The Unknown and the Unknowables

"There are things that I am uncertain about simply because I lack knowledge, and in principle my uncertainty might be reduced by gathering more information. Others are subject to random variability, which is unpredictable no matter how much information I might get; these are the unknowables. The two kinds of uncertainty have been debated by philosophers, who have given them the names epistemic uncertainty (due to lack of knowledge) and aleatory uncertainty (due to randomness)." (O'Hagan 2004, p. 132)

> older adults, 24 (about 20%) indicated that they had at some point been bullied by their fellow residents. Trompetter rejected the suggestion that her study may have been too small to draw reliable conclusions: "If I had talked to more people, the result would have changed by one or two percent at the most." (Lee and Wagenmakers 2013, p. 47)

Let's keep things simple and assume that the nursing homes in the Netherlands are comparable with respect to the occurrence of bullying – that is, we assume that, as far as bullying is concerned, the nursing homes are statistically *exchangeable*. Next, based on Trompetter's data, let's predict the number of older adults who report being bullied if we were to survey a different nursing home with, say, 100 older adults. Given that we know the true underlying chance to be .20 (the proportion in the Trompetter data), the prediction is determined solely by sampling variability, that is, all that matters for the prediction is aleatory uncertainty. The aleatory predictions are shown in Figure 2.2 as the peaked histogram. For these purely aleatory predictions, there is a summed probability of 95% that the number of bullied older adults will fall in the range from 13 to 28; also, the probability that the number of bullied older adults will fall between 18 and 22 ('two percent at the most' difference from Trompetter's 20%) equals .47. Clearly, if we know that the true chance is .20 and we survey 100 older adults, the result cannot be predicted with much accuracy.

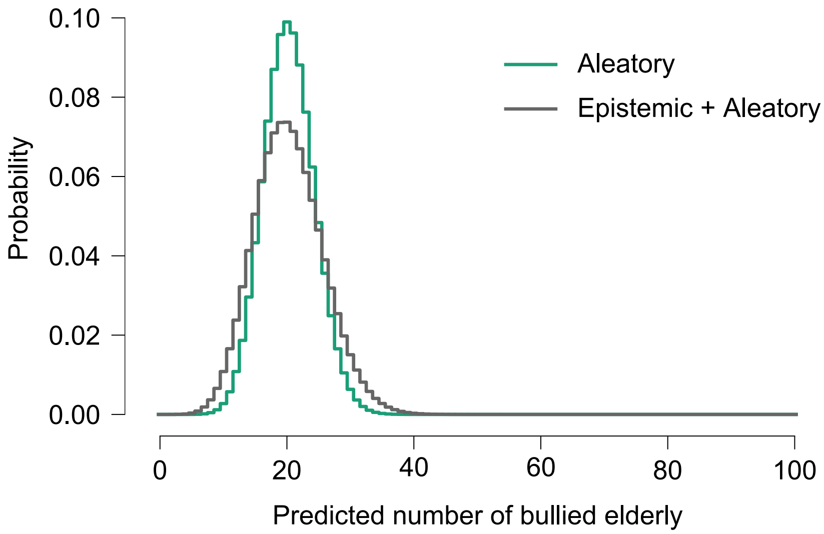


Figure 2.2: Predictions from the Trompetter scenario described in the main text. The 'aleatory' curve is based on the assumption that older adults from nursing homes have a .20 chance of reporting being bullied. The 'epistemic + aleatory' curve respects the fact that the true chance is not known exactly, and therefore allows other chances than .20 to play a role; consequently, the predictions become more spread out (i.e., more uncertain). The *Predicted number of bullied elderly* on the x-axis refers to the predicted number of bullied elderly from a nursing home of 100 inhabitants. Figure from the JASP module *Learn Bayes*.

The preceding analysis is seriously incomplete, however, as it assumes that an 'unknown' factor (i.e., the proportion of older adults in the Netherlands who report being bullied) was actually known exactly, and equals .20, the proportion of bullied older adults in Trompetter's relatively small sample. But based on Trompetter's observations (i.e., 24 bullied older adults out of 121) we are still uncertain about the true proportion of bullied elderly in the population – in particular, values such as .18 and .23 cannot be ruled out based on the initial sample. In other words, after learning about Trompetter's findings there remains considerable epistemic uncertainty about the true proportion in the population, and by ignoring this uncertainty (as was done in the above analysis) the predictions are overconfident. Realistic predictions need to consider not only sampling variability given a true state of the world, but also epistemic uncertainty, the fact that we do not exactly know the true state of the world (e.g., Aitchison and Dunsmore 1975). The broader histogram in Figure 2.2 shows the predictions based on the combination of epistemic and aleatory uncertainty.[6]

[6] The epistemic uncertainty was quantified in a standard Bayesian manner. to be discussed later in more detail. To appease the impatient reader: the epistemic posterior uncertainty was obtained by updating a flat prior distribution with Trompetter's observations (i.e., 24 bullied older adults out of 121).

The predictions that include epistemic uncertainty are now more spread out than they were before. For the predictions that make up the 'Epistemic + Aleatory' histogram, there is a summed probability of 95% that the number of bullied older adults will fall in the range from 11 to 32 (for aleatory-only this was 13 to 28); the probability that the number of bullied older adults will fall between 18 and 22 now equals .36 (for aleatory-only this was .47).

In sum, predictions about to-be-observed data should respect epistemic uncertainty; predictions that only involve aleatory uncertainty ('sampling variability') will falsely suggest that the future is more predictable than it really is.

Exercises

1. Borel wrote: "Indeed in all rigor, a judgment enunciated by Peter at a given time has a determinate probability, but the same judgment enunciated by him at a different time doesn't necessarily have the same probability, even if during the interval between these two times, he has received no external information." (Borel 1964, p. 51). How can this be?

2. In the Trompetter example, we assumed that the nursing homes were *exchangeable* in terms of bullying. (1) Is this a plausible assumption? How may it be violated? (2) What would happen to our predictions if we drop the assumption of exchangeability?

3. What would have had to happen in the Trompetter example to reduce the *epistemic* uncertainty that was involved in the prediction?

4. What would have had to happen in the Trompetter example to reduce the *aleatory* uncertainty involved in the prediction concerning the proportion of bullied elderly?

5. In antiquity, Carneades' idea that probability is the practical guide to life did not go unchallenged. As mentioned in Franklin (2015, p. 200), "Carneades has given no adequate reason why those appearances that *are* like the truth are in fact reliable guides for action.". Provide a response to this critique.

6. Does it make sense to speak of "the probability that the 10,000th figure in the digital expansion of Euler's number e is a 5"?

7. On Monday, September 7th 2020, one of us (EJ) was tested for COVID-19. Among those who are tested, about 3% receive a 'positive' outcome (i.e., the test detects the presence of COVID-19). Among those who receive a positive test outcome, about 75% really have COVID-19. It took 48 hours before EJ learned about the test outcome. On Tuesday, September 8th 2020, what would have been a reasonable estimate of the probability that EJ has COVID-19 (a) according to the doctor who administered the test (b) according to EJ (who has knowledge –albeit incomplete– of his own behavior and the people he interacted with in the past week) (c) according to an epidemiologists with knowledge about the prevalence of COVID-19 in Hilversum, where EJ lives?

Chapter Summary

In the Bayesian framework, probability is defined as a degree of reasonable belief.[7] When the belief concerns an 'unknown', that is, a proposition about which more can be learned, then the probability is called *epistemic*. Epistemic probabilities can be attached to unique events. For instance, one may assign a probability to the proposition that 100 years from now, The Netherlands will be largely underwater. Epistemic probabilities can also be attached to historic events. For instance, one may assign a probability to the proposition that the biologist Haldane spied for Stalin. But beliefs can also concern 'unknowables'; in repeated trials of observations (e.g., coin tosses, dice throws), the relevant knowledge to differentiate individual trial outcomes is often unavailable. When, given a particular state of the world, the probability of an outcome is the same for all trials, irrespective of what outcomes materialized previously, then that probability is called a *chance* (Jeffreys 1961, pp. 51-52). For instance, given that the state of the world is 'the coin is fair', the probability (chance) that the coin will land heads on the next toss is .50, irrespective of how many times it has landed heads in previous tosses. Chances reflect *aleatory* uncertainty, which gives rise to sampling variability. When the goal is to predict future events, both epistemic uncertainty and aleatory uncertainty have to be taken into account simultaneously. Ignoring epistemic uncertainty leads to predictions that are overconfident.

[7] Some Bayesian statisticians disagree, but immediately struggle to explain what would then be an acceptable alternative definition.

Want to Know More?

✓ Clayton, A. (2021). *Bernoulli's Fallacy: Statistical Illogic and the Crisis of Modern Science*. New York: Columbia University Press. "Consider this, instead, a piece of wartime propaganda, designed to be printed on leaflets and dropped from planes over enemy territory to win the

Probability and the Feeling of the Mind

"Probability is the feeling of the mind, not the inherent property of a set of circumstances. (...) Say that the question is, whether a red or a green ball shall be drawn, and suppose that A feels certain that all the balls are red, B, that all are green, while C knows nothing whatever about the matter. We have here, then, in reference to the drawing of a red ball, absolute certainty for or against, with absolute indifference, in three different persons, coming under different previous impressions. And thus we see that the real probabilities may be different to different persons. The abomination called intolerance, in most cases in which it is accompanied by sincerity, arises from inability to see this distinction. (...) In the mean time, we bring it forward as not the least of the advantages of this study, that it has a tendency constantly to keep before the mind considerations necessarily corrective of one of the most fearful taints of our intellect." (De Morgan 1838, pp. 7-8)

hearts and minds of those who may as yet be uncommitted to one side or the other. My goal with this book is not to broker a peace treaty; my goal is to win the war." (p. xv)

✓ de Finetti, B. (1974). *Theory of Probability*. New York: John Wiley & Sons. "More recently the subjectivist view has been seen as the best that is currently available and de Finetti appreciated as the great genius of probability." (Lindley 2000, p. 336)

✓ Eagle, A. (Ed.) (2011). *Philosophy of Probability: Contemporary Readings*. New York: Routledge. Includes a series of famous essays on probability, including Frank Ramsey's 1926 "Truth and Probability".

✓ Jeffreys, H. (1961). *Theory of Probability* (3rd edn.). Oxford, UK: Oxford University Press. The best book on statistical inference of all time, and by a landslide.

✓ Kyburg Jr., H. E., & Smokler, H. E. (Eds; 1964). *Studies in Subjective Probability*. New York: Wiley. A great collection of foundational papers on epistemic/subjective probability, including translated contributions from Borel and from de Finetti.

✓ Lindley, D. V. (1985). *Making Decisions* (2nd edn.). London: Wiley. Simple, straightforward, and compelling. A must-read.

✓ Lindley, D. V. (2006). *Understanding Uncertainty*. Hoboken: Wiley. If every student had to read this book, the world would be a better place.

✓ O'Hagan, A. (2004). Dicing with the unknown. *Significance, 1*, 132-133. A wonderful paper.

✓ Świątkowski, W., & Carrier, A. (2020). There is nothing magical about Bayesian statistics: An introduction to epistemic probabilities in data analysis for psychology starters. *Basic and Applied Social Psychology, 42*, 387-412. An accessible introduction to epistemic probabilities and Bayesian inference.

Probability of Effects and Probability of Causes

"It often happens that instead of trying to guess an event, by means of a more or less imperfect knowledge of the law, the events may be known and we want to find the law; or that instead of deducing effects from causes, we wish to deduce the causes from the effects. These are the problems called *probability of causes*, the most interesting from the point of view of their scientific applications.

I play écarté with a gentleman I know to be perfectly honest. He is about to deal. What is the probability of his turning up the king? It is $1/8$. This is a problem of the probability of effects.

I play with a gentleman whom I do not know. He has dealt ten times, and he has turned up the king six times. What is the probability that he is a sharper? This is a problem in the probability of causes.

It may be said that this is the essential problem of the experimental method. I have observed n values of x and the corresponding values of y. I have found that the ratio of the latter to the former is practically constant. There is the event, what is the cause?

Is it probable that there is a general law according to which y would be proportional to x, and that the small divergencies are due to errors of observation? This is a type of question that one is ever asking, and which we unconsciously solve whenever we are engaged in scientific work." (Poincaré 1913, p. 160; italics in original)

Afterthought: The Frequentist Definition of Probability

Instead of defining probability as degree of reasonable belief, some philosophers have proposed to define it as the limiting proportion of occurrence. For instance, the probability that a fair coin lands heads on the next throw is $.50$ because, in the limit of tossing the coin very often, the coin will land heads in 50% of the cases.

This is a Bayesian book and so we will not discuss the frequentist definition in detail. Wrinch and Jeffreys (1919, p. 731) summarized their early examination as follows: "It is shown that the attempt to give a definition of probability in terms of frequency is unsuccessful." Indeed, Harold Jeffreys considered the non-frequentist definition a cornerstone of his Bayesian theory of scientific learning: "*The essence of the present theory is that no probability, direct, prior, or posterior, is simply a frequency. The fundamental idea is that of a reasonable degree of belief (...)*" (Jeffreys 1961, p. 401)

Jeffreys's concrete objections to the frequency definition can be found in *Theory of Probability*, Chapter VII, "Frequency definitions and direct methods". Jeffreys appears exasperated that his critique of the frequency definitions were generally ignored (see also Jeffreys 1936a):

> "Adherents of frequency definitions of probability have naturally objected to the whole system. But they carefully avoided mentioning my criticisms of frequency definitions, which any competent mathematician can see to be unanswerable. In this way they contrive to present me as an intruder into a field where everything was already satisfactory. I speak from experience in saying that students have no difficulty in following my system if they have not already spent several years in trying to convince themselves that they understand frequency theories." (Jeffreys 1961, viii)

One common objection to the frequentist definition, also mentioned by Jeffreys, is that it is unable to assign probabilities to unique events, and essentially deals only with aleatory uncertainty, severely restricting the application domain:

> "Probability is a purely epistemological notion. For something over one hundred years, however, people have tried to define probability in terms of some notion of limiting frequency in an infinite series. There are two objections to this. First, even if such a definition could be given, the epistemological problem would be completely untouched. Secondly, even if the limiting frequency in an infinite series was known, we could draw no conclusions whatever about any finite set without some further principle, which cannot be contained in either pure logic or experience; and all applications in practice are to finite sets." (Jeffreys 1955, p. 283; see also Jeffreys 1973, pp. 193-197)

In the frequentist interpretation, then, probability cannot be "the very guide of life". We suggest that a serious study of the frequentist definition of probability ought to begin with a serious study of Jeffreys's critique of the concept (see also Clayton 2021, Jaynes 2003).

3 The Rules of Probability [with Quentin F. Gronau]

There may seem to be an intricacy in this subject which may prove distasteful to some readers; but this intricacy is essential to the subject at hand.

Jevons, 1874

Chapter Goal

The Bayesian reasoning process is governed by the laws of probability theory. Here we provide a brief and intuitive account of the most important concepts.

"Without [the calculus of probabilities] science would be impossible, without it we could neither discover a law nor apply it. Have we the right, for instance, to enunciate Newton's law? Without doubt, numerous observations are in accord with it; but is not this a simple effect of chance? Besides how do we know whether this law, true for so many centuries, will still be true next year? To this objection, you will find nothing to reply, except: 'That is very improbable.' " (Poincaré 1913, p. 157)

Terminology and Axioms

We have a sample space Ω ('omega') of possible outcomes. Outcomes and their combinations form 'events'. Toss a die once: the sample space consists of the possible number of pips that may be observed (i.e., 1, 2, 3, 4, 5, or 6). An example event is "the pips are even in number". Although the interpretation of probability remains the topic of considerable debate (e.g., Galavotti 2005), the warring parties[1] agree that for something to be a probability, it needs to adhere to three basic rules –the Kolmogorov axioms– from which all others can be derived:

[1] Mostly Bayesians, who view probability as a degree of belief, and frequentists, who view probability as the limit of a proportion.

- Probabilities are not negative.
- Some outcome always happens.
- For mutually exclusive ('disjoint') events, probability adds.

Thus, for the probability that either event A *or* event B will occur (and only one may occur), we have $p(A \cup B) = p(A) + p(B)$. A Venn diagram[2] is shown in Figure 3.1.

[2] Venn had a distinctly negative opinion on Bayesian inference. Nonetheless, his diagrams are useful for obtaining an intuitive idea of the structure of a probabilistic problem.

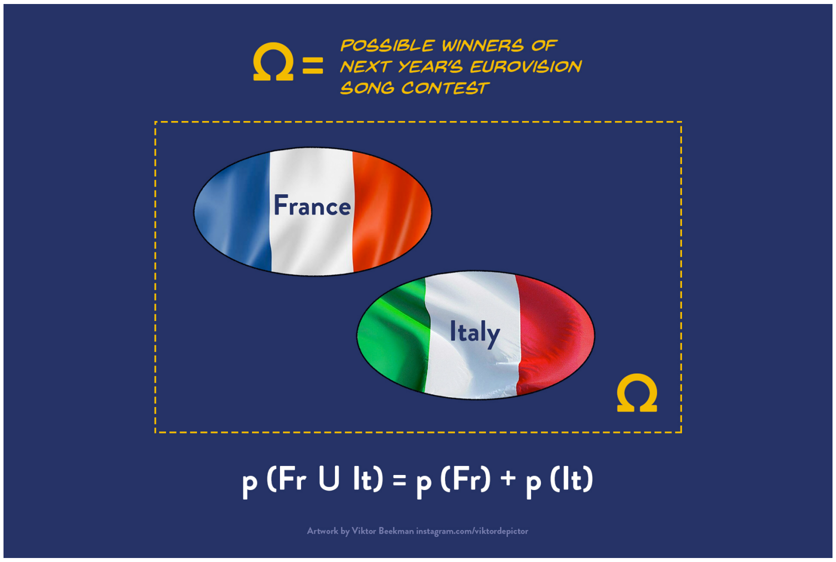


Figure 3.1: The probability for disjoint (i.e., mutually exclusive) events is their sum. The symbol '$\cup$' stands for 'union', the probability of A *or* B. The symbol 'Ω' represents the sample space of all possible winners. Figure available at `BayesianSpectacles.org` under a CC-BY license.

The Sum Rule

According to the sum rule, the probability of events A *or* B is given by the sum of their individual probabilities minus the probability of A *and* B (i.e, $p(A \cap B)$):

$$p(A \cup B) = p(A) + p(B) - p(A \cap B).$$

The Venn diagram in Figure 3.2 clarifies that the intersection (i.e., A *and* B) is subtracted because it would otherwise be counted twice. Note that when the events do not have any overlap, we obtain the third Kolmogorov axiom as a special case. Also note that the 'or' in $A \cup B$ is inclusive: at stake is the summed probability for event A happening and B not happening, for event B happening and A not happening, and for *both* A and B happening.

The Multiplication Rule

According to the multiplication rule, the probability of both *independent* events A *and* B arising is given by multiplying their individual probabilities:

$$p(A \cap B) = p(A) \times p(B).$$

When two fair dice are thrown, the probability that the first die will show six pips is $1/6$, and the probability that the second die will show six

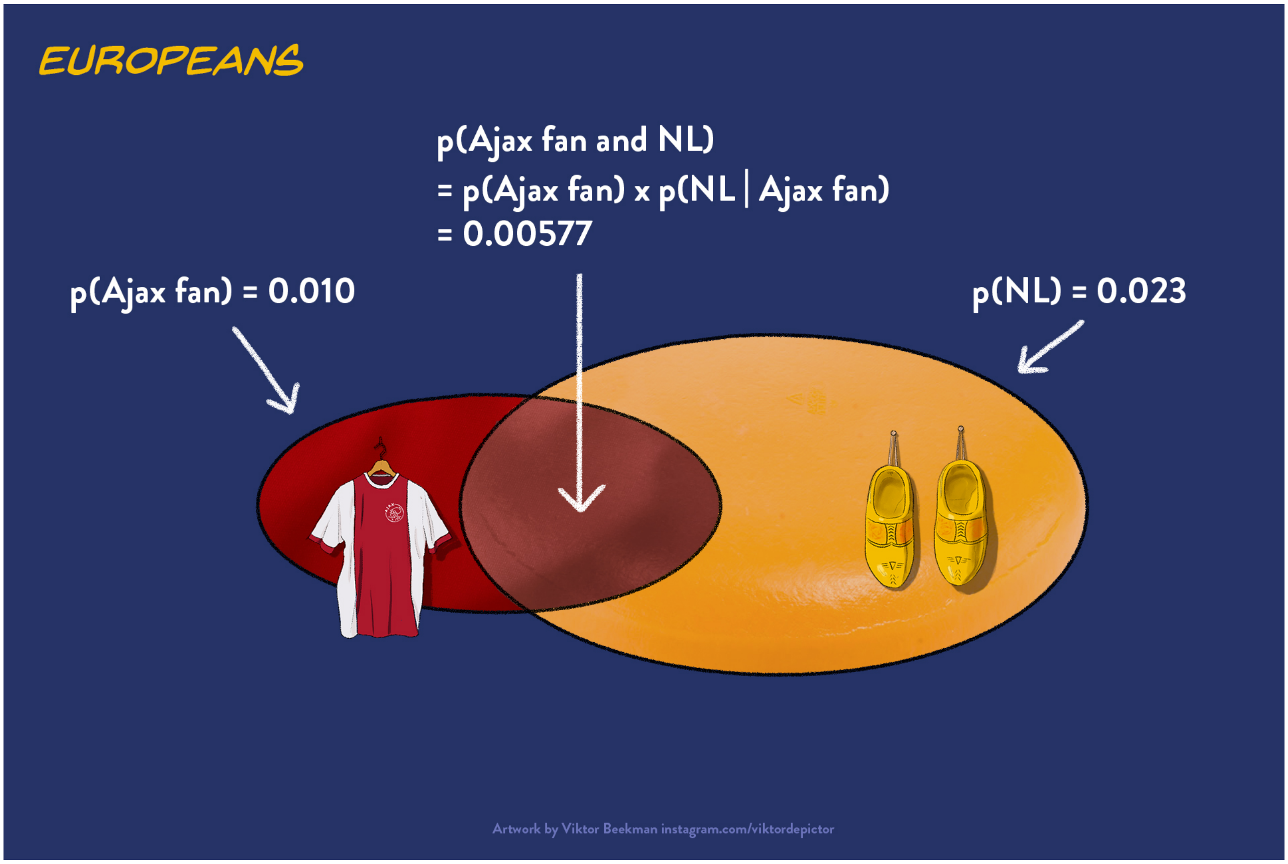


Figure 3.2: Venn diagrams provide an intuition for the sum rule, which states that $p(A \cup B) = p(A) + p(B) - p(A \cap B)$. Subtracting the intersecting area $p(A \cap B)$ (i.e., A *and* B) is needed to prevent that area from being counted twice. In this specific example, the probability that a randomly chosen European is either Dutch (i.e., NL) *or* a fan of Ajax is the sum of the individual probabilities minus the probability of a person being both Dutch and an Ajax fan. Figure available at BayesianSpectacles.org under a CC-BY license.

pips is $1/6$; according to the multiplication rule, the probability that *both* will show six pips is $1/6 \times 1/6 = 1/36$. Often, however, the constituent events are not independent, and this brings us to the next section.

Conditional Probability

The rule of conditional probability states that the probability of A *conditional on B holding true* equals the intersection (i.e., A *and* B) normalized to the probability of B:

$$p(A \mid B) = \frac{p(A \cap B)}{p(B)}.$$

The vertical stroke symbol '|' is usually read as 'given that'.[3]

The intuition for this rule can be obtained by considering another Venn diagram, shown in Figure 3.2. Suppose we wish to learn, from the information provided, the probability that a randomly selected European is Dutch, given that we are told they are an Ajax fan. The

[3] This notation was first proposed by our Bayesian hero Sir Harold Jeffreys (Jeffreys 1931, p.15; Jeffreys 1939, p.25). For details see the post "The man who rewrote conditional probability" on BayesianSpectacles.org.

probability of interest involves $p(\text{Ajax fan} \cap \text{NL})$. But this intersection of events has probability 0.00577, and clearly that is much too low. This probability would be correct if we were sampling randomly from the population of Europeans – in other words, if we blindly threw a dart onto the entire Venn diagram. But we know that our person is an Ajax fan; hence, our relevant universe Ω has reduced to the red oval in Figure 3.2. In other words, we are interested in the probability that a randomly thrown dart lands in the intersection area, given that we already know it landed in the red oval area — what we need, therefore, is the proportion of the red oval that is brownish. To obtain the desired result, we apply the definition of conditional probability and obtain[4]:

$$p(\text{NL} \mid \text{Ajax fan}) = \frac{p(\text{Ajax fan} \cap \text{NL})}{p(\text{Ajax fan})} = \frac{0.00577}{0.010} = 0.577.$$

The rule of conditional probability can also be written like this:

$$p(A \cap B) = p(A \mid B) \times p(B).$$

This way of writing the rule is consistent with another intuition, one that is provided by a tree diagram. Figure 3.3 shown an example. The tree progresses from left to right; the first branching factor is according to whether a randomly selected person is an Ajax fan or not, and the second branching factor is according to nationality.[5] Importantly, the second branch is conditional on what happened in the first branch, and this is why tree diagrams automatically encode conditional probabilities. For instance, the top path first leads to the selection of an Ajax fan; then, given that an Ajax fan was selected, there is a particular probability that this person is also Dutch. As indicated in the tree diagram, this probability is 0.577 – the same conditional probability that we already calculated above. A little reflection reveals that the top path of the tree diagram tells us everything we need to know to arrive at the rule for conditional probability: the probability of being an Ajax fan *and* Dutch is the probability of going up in both branches: first, with probability 0.010, we go up to select our Ajax fan; then, with probability 0.577, we go up once more to select a Dutch person, given that we find ourselves among the branches that only contain Ajax fans. In other words, we have:

$$p(\text{Ajax fan} \cap \text{NL}) = p(\text{Ajax fan}) \times p(\text{NL} \mid \text{Ajax fan}),$$

which is the definition of conditional probability.

Note that if $p(\text{NL} \mid \text{Ajax fan})$ were equal to $p(\text{NL})$ (i.e., whether or not one has selected an Ajax fan leaves unaltered the probability of having selected a Dutch person), we recover the multiplication law for independent events.

Bayesians such as Harold Jeffreys, Ed Jaynes, and Dennis Lindley have argued that *all* probability assignments are conditional, in the

[4] As a mnemonic, note that the vertical stroke symbol '|' for 'given that' was originally written as the slanted stroke symbol '/' that is now exclusively used to represent 'divided by'. Thus, when you see $p(\text{NL} \mid \text{Ajax fan})$ you immediately know that the definition involves a division by $p(\text{Ajax fan})$.

[5] The tree invites a temporal interpretation, but that is not necessary and the tree may just as will be constructed the other way around, with nationality as the first branching factor.

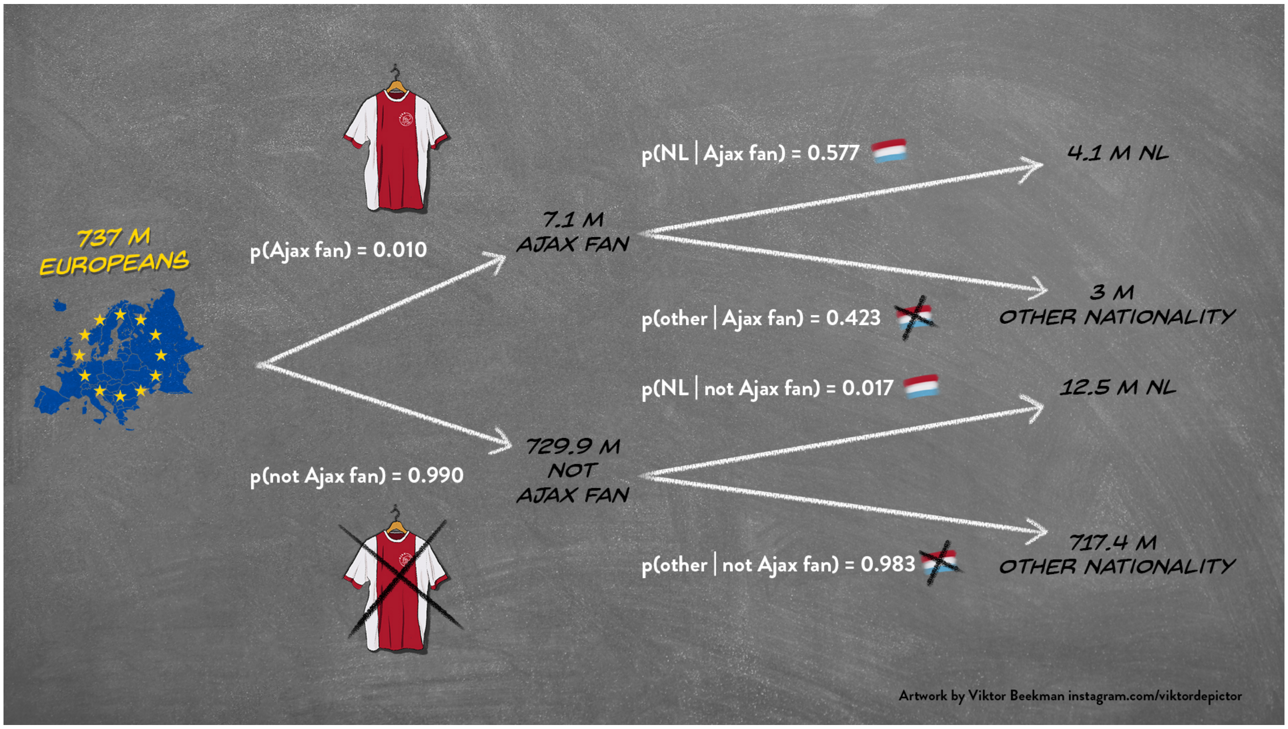


Figure 3.3: Tree diagrams help provide an intuition for the law of conditional probability and the law of total probability. See text for details. Figure available at `BayesianSpectacles.org` under a CC-BY license.

sense that they are conditional on background knowledge K. For instance, Wrinch and Jeffreys (1921, p. 381) wrote: "Now it appears certain that no probability is ever determined from experience alone. It is always influenced to some extent by the knowledge we had before the experience." To make this explicit, we should really write $p(A \mid K)$ instead of $p(A)$; however, it is unusual and cumbersome to pay tribute to K in every equation, and we will not do so here.[6] Nevertheless, it is important to realize that all probability assignments occur against the backdrop of an existing knowledge base.

[6] Harold Jeffreys often conditioned his probability statements on background knowledge or *history* 'H'; for current-day readers this can be confusing, as nowadays 'H' stands for 'hypothesis'.

Marginal Probability

The law of total probability establishes how the overall ('marginal') probability for an event can be computed from conditional probabilities involving an exhaustive partition of the sample space. Before we show the equation, consider again the tree diagram in Figure 3.3. Suppose we wish to derive, from the information given in the tree, the probability of selecting a Dutch person, $p(\text{NL})$. This number is not shown in the tree directly, because the first branch involves the probability of selecting an Ajax fan, which is not something we are interested in. For the question at hand, whether or not someone is an Ajax fan is a *nuisance*

Defining the Probable by the Probable

"Has probability been defined? Can it even be defined? And if it can not, how dare we reason about it? The definition, it will be said, is very simple: the probability of an event is the ratio of the number of cases favorable to this event to the total number of possible cases.

A simple example will show how incomplete this definition is. I throw two dice. What is the probability that one of the two at least turns up a six? Each die can turn up in six different ways; the number of possible cases is $6 \times 6 = 36$; the number of favorable cases is 11; the probability is $11/36$.

That is the correct solution. But could I not just as well say: The points which turn up on the two dice can form $6 \times 7/2 = 21$ different combinations? Among these combinations 6 are favorable; the probability is $6/21$.

Now why is the first method of enumerating the possible cases more legitimate than the second? In any case it is not our definition that tells us. We are therefore obliged to complete this definition by saying '...to the total number of possible cases provided these cases are equally probable.' So, therefore, we are reduced to defining the probable by the probable." (Poincaré 1913, pp. 155-156)

factor. How do we get rid of it? Well, we observe that there are two paths in the tree diagram that result in the selection of a Dutch person. The first path involves 'Ajax fan' and then 'NL', for a probability of $0.010 \times 0.577 = 0.00577$; the second path involves 'not Ajax fan' and then 'NL', for a probability of $0.990 \times 0.017 = 0.01683$. Adding these two probabilities provides the marginal or overall probability of selecting a Dutch person: $0.00577 + 0.01683 = 0.0226$. What we have done, in fact, is to compute a weighted average between the result within the group of Ajax fans (with a corresponding probability of 0.577) and within the group of not Ajax fans (with a corresponding probability of 0.017); the averaging weights are provided by the probability of being an Ajax fan. That this is required can be intuited from the tree diagram, and also from imagining that the probability of finding an Ajax fan is zero; in that case, only the lower of the two 'NL' paths is relevant, and $p(\text{NL} \mid \text{no Ajax fan})$ is equal to $p(\text{NL})$.

When the nuisance factor B can take on two values (e.g., Ajax fan vs. no Ajax fan; winning vs losing; left or right, etc.) the law of total probability can be written as follows:

$$p(A) = p(A \mid B_1) \times p(B_1) + p(A \mid B_2) \times p(B_2).$$

When the nuisance factor can take on many values (e.g., day of the year), say n of them, we simplify notation by using Euler's summation sign Σ:

$$p(A) = \sum_{i=1}^{n} p(A \mid B_i) \times p(B_i),$$

indicating that the partition runs from B_1 to B_n. Although n may be large, the principle remains the same: the marginal probability is obtained by simply summing over the weighted conditional distributions.[7]

However, we have to face one more complication: sometimes, the number of partitions n is infinite. For instance, imagine you are playing an online game and the software spawns a synthetic opponent i with strength S_i. This strength S_i is determined by drawing a value from a continuous distribution – say a bell-shaped distribution with mean 100 and standard deviation 15, like the population distribution of IQ. So sometimes your opponent will be very weak, sometimes very strong, but most of the time your opponent will be average. Suppose that when we know S_i, we know your chances of beating the opponent – in other words, we know the conditional probabilities $p(\text{win} \mid S_i)$. But now the question is, without yet knowing what specific opponent you are going to face, what are your chances of winning the next game? The law of total probability appears to tell us that we should compute

$$p(\text{win}) = \sum_{i=1}^{n} p(\text{win} \mid S_i) \times p(S_i),$$

but we cannot do this, because under a continuous distribution, the probability $p(S_i)$ of drawing any specific value S_i is...zero.[8] As shown in the right panel of Figure 3.4, 'probability' in a continuous distribution is defined as the area under the curve, that is, the probability that a value falls between a and b is the area of the continuous distribution in the interval from a to b. As the interval narrows, the probability decreases, until, when the interval is zero, it vanishes entirely.

The standard solution to this dilemma is to switch from summing (which is defined for discrete quantities) to *integration* (which is defined for continuous quantities). The equation then becomes:

$$p(\text{win}) = \int_S p(\text{win} \mid S) \times p(S)\,\mathrm{d}S,$$

where $p(S)$ indicates the continuous distribution from which particular S_i are drawn. The symbols of integration are explained in Figure 3.5 (Thompson 1910). Whenever the integral cannot be solved analytically, one may resort to numerical approximations. One of these approximations is particularly straightforward: we draw a large number of S_i from distribution $p(S)$, and for each we compute $p(\text{win} \mid S_i)$, which we then average to obtain the desired result.[9]

The concept of marginal probability is of fundamental importance for Bayesian inference. Whenever an analysis is complicated by the pres-

[7] The events B_i that are conditioned on must be exhaustive and exclusive.

[8] For more details see the YouTube channel '3Blue1Brown', episode "Why "probability of 0" does not mean "impossible" | Probabilities of probabilities, part 2".

[9] By increasing the number of draws the analytical result can be approximated to any desired degree of accuracy.

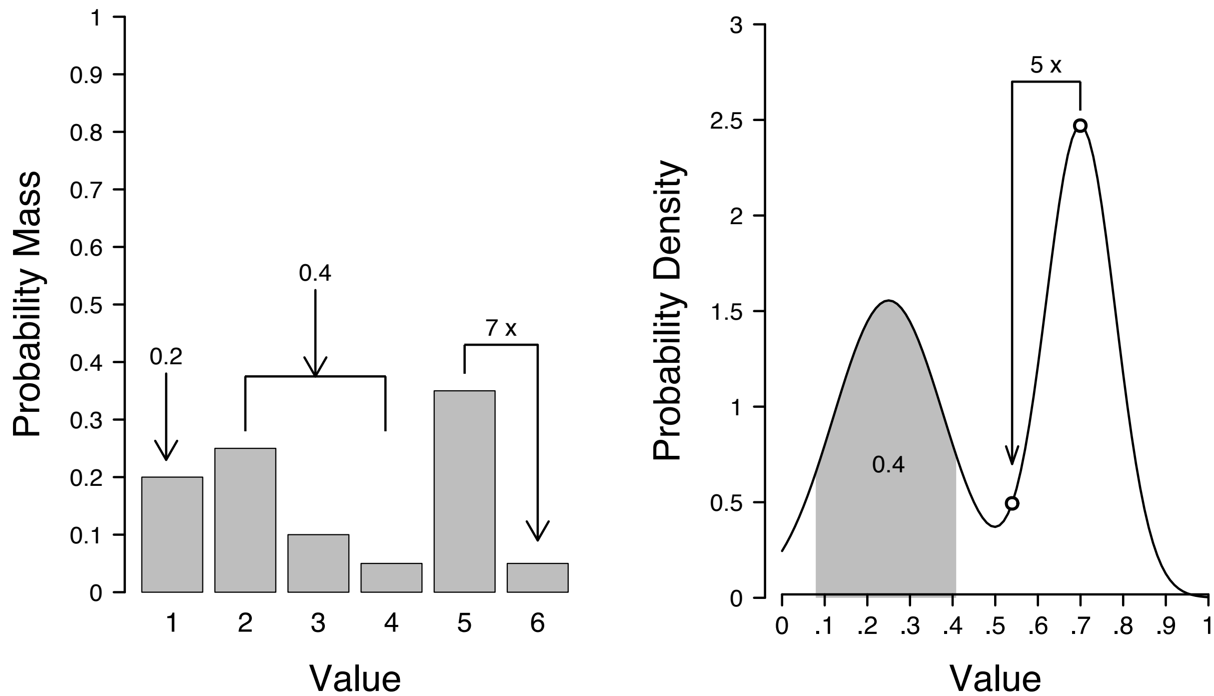


Figure 3.4: Discrete and continuous probability distributions. Left panel: In a discrete distribution, probability is the mass assigned to each point, as indicated by its height. Right panel: In a continuous distribution, probability is the area under the curve. The height of the curve does have meaning, but only relative to another height. Code from `http://shinyapps.org/apps/RGraphCompendium`.

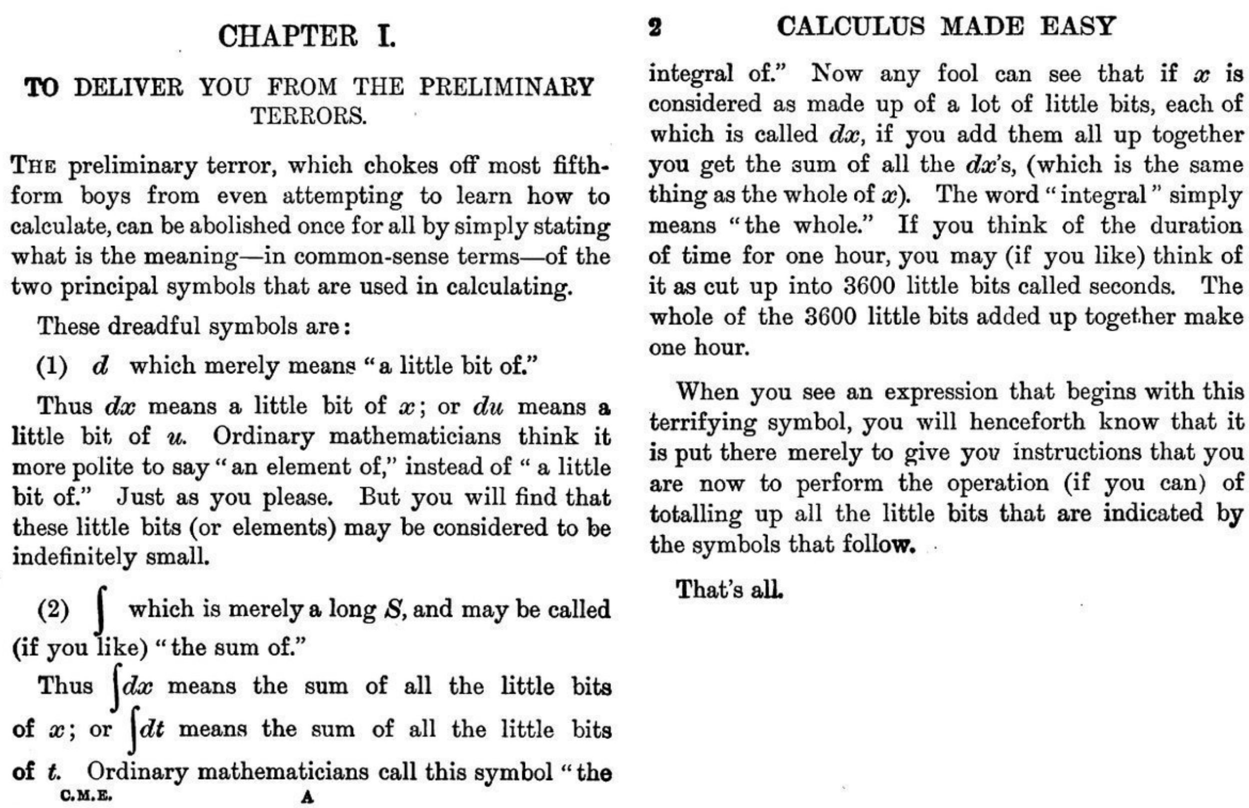
ence of a nuisance factor that exerts an influence but is not of immediate interest, the law of total probability dictates how this nuisance factor may be 'averaged out'. To drive this intuition home we now consider a geometric interpretation.

Excursion: A Geometric Interpretation of Marginal Probability

Roger and Zita are going to play a tennis match. Without wind, they are equally matched; but Roger is a relatively good wind player, so when it is windy the probability of Roger winning increases to 0.70. The probability that it will be windy is 0.60. A tree diagram is shown in Figure 3.6.

Given the information from the tree diagram, what is the probability that Roger wins the match? To answer this question, we need to remove the wind factor and compute a weighted average – in statistics lingo, we need to *marginalize* across the wind factor.[10] From the tree diagram, we can see that two paths lead to Roger winning. The first path is 'wind' $\rightarrow$ 'Roger wins' that has probability $.60 \times .70 = .42$; the second path

[10] The term 'marginalize' originates from the analysis of contingency tables, where the row sums are presented in the table margin.



CHAPTER I.

TO DELIVER YOU FROM THE PRELIMINARY TERRORS.

THE preliminary terror, which chokes off most fifth-form boys from even attempting to learn how to calculate, can be abolished once for all by simply stating what is the meaning—in common-sense terms—of the two principal symbols that are used in calculating.

These dreadful symbols are:

(1) d which merely means "a little bit of."

Thus dx means a little bit of x; or du means a little bit of u. Ordinary mathematicians think it more polite to say "an element of," instead of "a little bit of." Just as you please. But you will find that these little bits (or elements) may be considered to be indefinitely small.

(2) $\int$ which is merely a long S, and may be called (if you like) "the sum of."

Thus $\int dx$ means the sum of all the little bits of x; or $\int dt$ means the sum of all the little bits of t. Ordinary mathematicians call this symbol "the

C.M.E. A

2 CALCULUS MADE EASY

integral of." Now any fool can see that if x is considered as made up of a lot of little bits, each of which is called dx, if you add them all up together you get the sum of all the dx's, (which is the same thing as the whole of x). The word "integral" simply means "the whole." If you think of the duration of time for one hour, you may (if you like) think of it as cut up into 3600 little bits called seconds. The whole of the 3600 little bits added up together make one hour.

When you see an expression that begins with this terrifying symbol, you will henceforth know that it is put there merely to give you instructions that you are now to perform the operation (if you can) of totalling up all the little bits that are indicated by the symbols that follow.

That's all.

Figure 3.5: The first chapter of S. P. Thompson's 1910 classic work 'Calculus Made Easy' explains how to interpret the symbols of integration.

is 'no wind' → 'Roger wins' that has probability $.40 \times .50 = .20$. The total probability that Roger wins is the sum across these two paths, so $.42 + .20 = .62$. As in the Ajax example, we have effectively applied the law of total probability to remove the wind factor, as follows:

$$\begin{aligned} p(\text{Roger wins}) = {} & p(\text{Roger wins} \mid \text{wind}) \times p(\text{wind}) \\ & + p(\text{Roger wins} \mid \text{no wind}) \times p(\text{no wind}). \end{aligned}$$

Instead of using a tree diagram, we can also display the information by plotting the probability that Roger wins against the probability that it is windy, creating a Venn diagram with four non-overlapping areas, as shown in Figure 3.7 In this figure, the area of the left square (i.e., 'wind' and 'Roger wins') is $.60 \times .70 = .42$, equalling the probability for the first path in the tree diagram. The area of the right square (i.e., 'no wind' and 'Roger wins') is $.40 \times .50 = .20$, the same as the second path in the tree diagram. The marginal probability of Roger winning therefore equals the summed area of the two squares, that is, the area for Roger winning under the curve across the wind factor. Because the x-axis ranges from 0 to 1, this total area equals the average height of the curve.

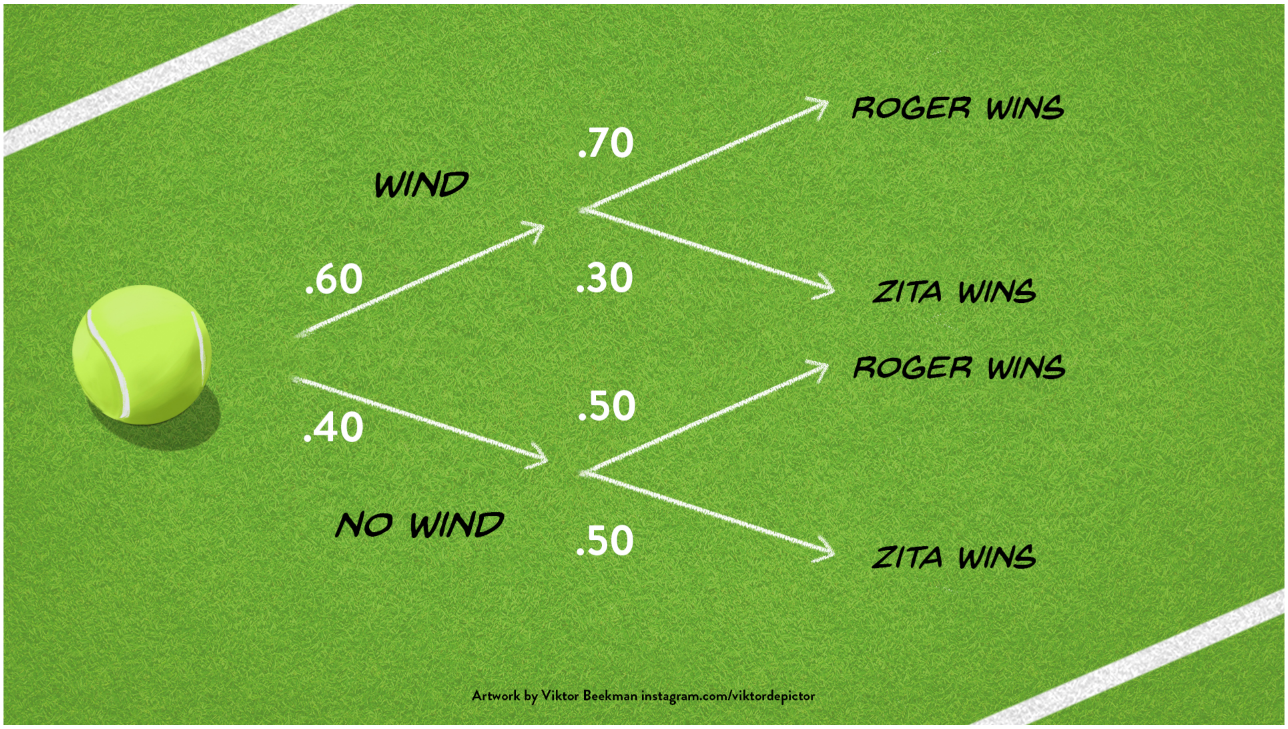


Figure 3.6: Tree diagram for a tennis match. When there is no wind, Roger and Zita are equally matched; when it is windy, however, Roger's chances of winning increase. What is the marginal probability of Roger winning the match? Figure available at `BayesianSpectacles.org` under a CC-BY license.

This geometric interpretation of marginal probability makes it clear that it is a weighted average across the nuisance variable (in this case, the wind factor). For instance, if the probability of it being windy increases from .60 to .80, the area under the curve becomes larger, as the left square has greater height than the right square. From the geometric interpretation it is also apparent that the marginal probability always falls in between the highest probability for the factor of interest (i.e., the probability of Roger winning when it is windy, which is .70) and the lowest probability for the factor of interest (i.e., the probability of Roger winning when it is not windy, which is .50).

The tennis example can be generalized by differentiating between multiple wind strengths (e.g., not windy, a little windy, windy, very windy, and stormy), each associated with a different probability of Roger winning. The Venn diagram would then consist of multiple squares, one for each wind condition. The marginal probability of Roger winning would still be the area under the curve across the wind factor. If the wind factor becomes a continuous variable the curve changes smoothly instead of abruptly.

Marginal probability is not just important in soccer and tennis, but it also plays a key role in Bayes' rule, to which we now turn.

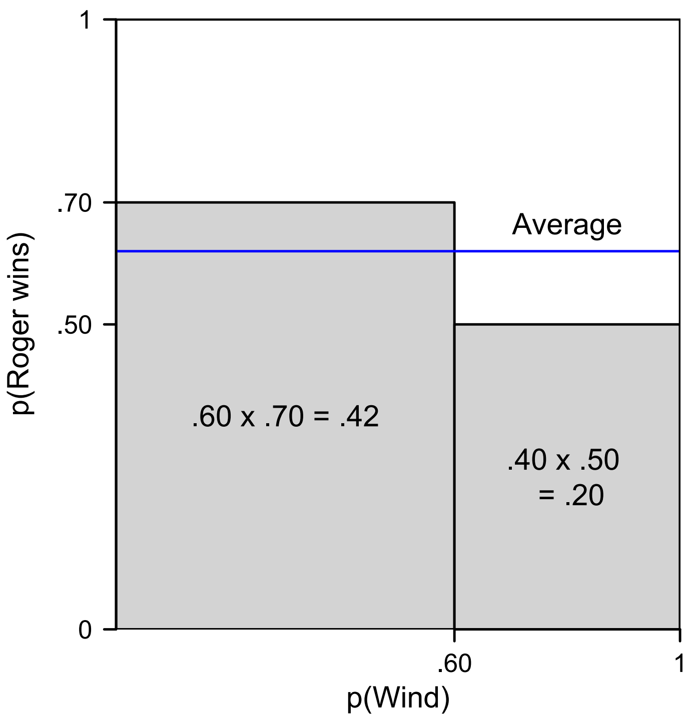


Figure 3.7: Geometric interpretation of marginal probability. The probability that Roger wins the match is the sum of the two grey squares, or the area under the 'Roger wins' curve. Because the x-axis ranges from 0 to 1, this equals the average height of the curve, which is indicated by the blue horizontal line.

Bayes' Rule

A simple consequence of the definition of conditional probability, Bayes' rule shows how we can move from $p(B \mid A)$ to $p(A \mid B)$, and thus move from a purely *deductive* system that makes only predictions (i.e., p(data | state of the world)) to a system that can also achieve *induction* (i.e., p(state of the world | data)). In other words, Bayes' rule inverts the causal arrow from *causes* $\rightarrow$ *consequences* (i.e., p(consequences | causes)) to *consequences* $\rightarrow$ *causes* (i.e., p(causes | consequences)).[11]

[11] This is the reason why, until the 1950s, 'Bayesian inference' was referred to as 'inverse probability'.

Deriving Bayes' rule is straightforward. We have already seen the definition of conditional probability:

$$p(A \cap B) = p(A \mid B) \times p(B).$$

Switching labels A and B yields another valid version:

$$p(B \cap A) = p(B \mid A) \times p(A).$$

The conjunction of events is symmetric (i.e., the probability of A *and* B is the same as the probability of B *and* A):

$$p(A \cap B) = p(B \cap A),$$

and it follows that

$$p(A \mid B) \times p(B) = p(B \mid A) \times p(A).$$

Dividing both sides by $p(B)$ then yields Bayes' rule:

$$p(A \mid B) = \frac{p(B \mid A) \times p(A)}{p(B)}.$$

Bayes' rule is extremely powerful, as becomes clearer when we replace the abstract symbol 'A' with 'θ' ('theta')[12] and 'B' with 'data':

$$p(\theta \mid \text{data}) = \frac{p(\text{data} \mid \theta) \times p(\theta)}{p(\text{data})}.$$

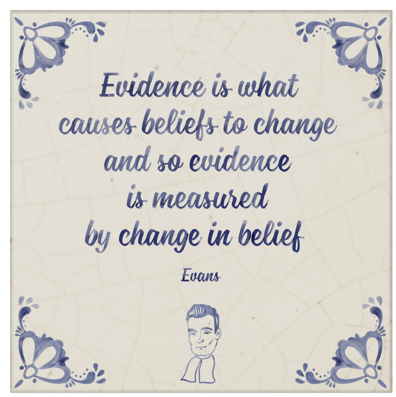
We now move $p(\theta)$ in front and behold, here is the equation that has changed the world (McGrayne 2011), the rule that formalizes the predictive principle of learning from experience:

$$\underbrace{p(\theta \mid \text{data})}_{\substack{\text{Posterior beliefs} \\ \text{about the world}}} = \underbrace{p(\theta)}_{\substack{\text{Prior beliefs} \\ \text{about the world}}} \times \underbrace{\frac{p(\text{data} \mid \theta)}{p(\text{data})}}_{\substack{\text{Predictive} \\ \text{updating factor}}}. \tag{3.1}$$

The equation states that the change from prior to posterior beliefs about the world 'theta' is driven by a *predictive updating factor*. This factor quantifies the relative predictive adequacy of a particular value of θ by comparing its predictive performance to the predictive performance averaged across all values of θ, that is, $p(\text{data})$. Thus, values of θ that predict better than average enjoy a boost in plausibility, whereas values of θ that predict worse than average suffer a decline (Wagenmakers et al. 2016a). But we are getting well ahead of ourselves. For now, note the following aspects about Bayes' rule (Equation 3.1):

- Posterior belief about the world is explicitly a *conditional* probability – it conditions on the observed data.
- Prior belief about the world is also a conditional probability, be it in disguise – prior belief conditions on background knowledge K (as does the posterior belief; Lindley 2006, pp. 43-44). Here we leave this dependence implicit.
- The denominator in the predictive updating factor, $p(\text{data})$, is a marginal probability, commonly known as marginal likelihood, that involves a weighted average or integral across the different values of

[12] The Greek letter θ refers to an unknown aspect of the world that we wish to learn about. Keep in mind that for a statistician, "the world" means "my mathematical abstraction of a microscopically small part of the world".



If accepted as true, this statement by Evans (2015) rules out all non-Bayesian methods of inference as far as the quantification of evidence is concerned. Figure available at `BayesianSpectacles.org` under a CC-BY license.

θ following the law of total probability: $p(\text{data}) = \int p(\text{data} \mid \theta)p(\theta)\,\mathrm{d}\theta$. The integration (or sum, in case θ is discrete) reveals that in order to learn about which state of the world is most plausible, we need to start out by postulating multiple rival states, each of which must make predictions and have a prior plausibility.

- The predictive updating factor quantifies the change in belief brought about by the data, and it is also known as the 'strength of the evidence'.[13]
- When prior beliefs are relatively weak (i.e., the claim at hand is relatively implausible *a priori*), the predictive updating factor needs to produce evidence that is relatively compelling in order for the posterior beliefs to be appreciable. This quantifies the adage 'extraordinary claims require extraordinary evidence'.

[13] For details see Evans (2015) and Etz and Wagenmakers (2017).

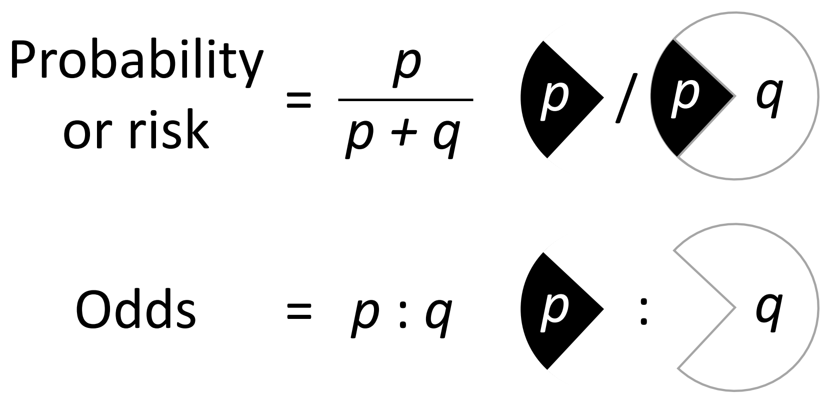


Figure 3.8: Comparison of probability and odds by C. M. G. Lee. Figure available on Wikipedia under a CC BY-SA 4.0 license.

Odds Form of Bayes' Rule

The above version of Bayes' rule is in probability form. We can also entertain an odds form. Start by considering a specific value, say, θ_1. The probability form of Bayes' rule yields:

$$p(\theta_1 \mid \text{data}) = p(\theta_1) \times \frac{p(\text{data} \mid \theta_1)}{p(\text{data})}.$$

For a rival value, θ_2, Bayes' rule yields:

$$p(\theta_2 \mid \text{data}) = p(\theta_2) \times \frac{p(\text{data} \mid \theta_2)}{p(\text{data})}.$$

The odds form of Bayes' rule can be obtained by dividing the above two expressions, with the following result:

From Probability to Odds and Back Again

Uncertainty about an event or a proposition A can be quantified by probability, $p(A)$, but it can just as well be quantified by the *odds*, which is defined as the probability of the event occurring, $p(A)$, divided by the probability of the event *not* occurring, $p(\neg A)$:

$$o(A) = \frac{p(A)}{p(\neg A)} = \frac{p(A)}{1 - p(A)}.$$

Note that when $p(A) = {}^1\!/\!{}_2$, $o(A) = 1$, so that an odds of 1 indicates that an event is just as likely to occur as not. Also note that probabilities range from 0 to 1 but odds range from 0 to infinity. This makes odds better suited to represent extreme probabilities. For instance, $p(A) = .999$ yields an odds of $o(A) = 999$, whereas $p(B) = .999999$ yields $o(B) = 999,999$ – the probabilities are close to 1 and therefore differ only little, but the odds differ a lot. However, one complication with the odds scale is that it is not symmetric. When $p(A) = .999$ then $o(A) = 999$; but $p(\neg A) = .001$ yields $o(\neg A) = 1/999 \approx .001$. In other words, astronomically high odds are well separated (999 is very different from $999,999$), but astronomically low odds are pushed up against the bound of 0. The scale can be made symmetric by using the *logarithm* of the odds:

$$lo(A) = \log \frac{p(A)}{p(\neg A)}.$$

The *log odds* scale is symmetric: $lo(A) = -lo(\neg A)$; for instance, $p(A) = .999$ gives $lo(A) \approx 3$ whereas $p(\neg A) = .001$ gives $lo(\neg A) \approx -3$, that is, high probabilities have the same distance from the point of equivalence as low probabilities (for details see Chapter 23). Finally, when we have the odds we can transform back to probabilities as follows:

$$p(A) = \frac{o(A)}{o(A) + 1}.$$

For example, when $o(A) = 2$ ("the odds are two to one") then $p(A) = 2/3$; when $o(A) = 999$ then $p(A) = 999/1000 = .999$.

$$\underbrace{\frac{p(\theta_1 \mid \text{data})}{p(\theta_2 \mid \text{data})}}_{\text{Posterior odds}} = \underbrace{\frac{p(\theta_1)}{p(\theta_2)}}_{\text{Prior odds}} \times \underbrace{\frac{p(\text{data} \mid \theta_1)}{p(\text{data} \mid \theta_2)}}_{\text{Evidence}}. \tag{3.2}$$

Suppose the evidence is 6; this means that θ_1 predicted the observed data six times better than θ_2. In other words, the observed data were six times more likely to occur under θ_1 than under θ_2. Suppose the prior odds are $1/3$, that is, θ_2 is *a priori* three times more plausible than θ_1.[14] Updating the prior odds with the evidence yields a posterior odds of $1/3 \times 6 = 2$ in favor of θ_1 over θ_2. As the example in the next section will demonstrate, the odds form is often more convenient to work with, especially from the perspective of human intuition.

[14] Such considerations may flow, for instance, from an analysis of previous data.

Example: The Inevitable Base Rate Fallacy

No book on probability is complete without an example on the base rate fallacy.[15] The fallacy concerns the fact that the outcome of a test with fantastic operating characteristics may actually provide a deeply misleading impression of the true state of affairs. It is often suggested that the Bayesian solution is too complicated for mere mortals to wrap their heads around. Indeed, the Bayesian solution is complicated *when it is presented as a single step, in its probability form*. Break it down into its component steps, in its odds form, and the process becomes much simpler.

[15] This example is based on the `BayesianSpectacles.org` blog post "The single most prevalent misinterpretation of Bayes' rule". See also the YouTube video "The medical test paradox: Can redesigning Bayes rule help?" from '3Blue1Brown'.

Consider the same problem as is mentioned on the Wikipedia page for the base rate fallacy[16]:

> "A group of police officers have breathalyzers displaying false drunkenness in 5% of the cases in which the driver is sober. However, the breathalyzers never fail to detect a truly drunk person. One in a thousand drivers is driving drunk. Suppose the police officers then stop a driver at random to administer a breathalyzer test. It indicates that the driver is drunk. We assume you do not know anything else about them. How high is the probability they really are drunk? Many would answer as high as 95%, but the correct probability is about 2%."

[16] `https://en.m.wikipedia.org/wiki/Base_rate_fallacy`, as accessed on September 6th, 2021

"all the sciences would be only unconscious applications of the calculus of probabilities. To condemn this calculus would be to condemn the whole of science." (Poincaré 1913, p. 157)

In the first step of our Bayesian odds-form analysis of this problem, we take stock of our prior information: "one in a thousand drivers is driving drunk". This means that $p(\text{drunk}) = 1/1000$ and $p(\text{sober}) = 999/1000$. So, before we see any data, the prior odds in favor of someone being sober instead of drunk are $p(\text{sober})/p(\text{drunk}) = 999$. In the second step we consider the evidence that is provided by the data. We know that the breathalyzer test is positive. The probability of this happening for drunk drivers is 1, and for sober drivers it is $.05$. The evidence in favor of the driver being drunk rather than sober is therefore: $p(\text{test positive} \mid \text{drunk})/p(\text{test positive} \mid \text{sober}) = 1/.05 = 20$.

In the third step we combine our prior information (i.e., odds of 999 in favor of the driver being sober) with the evidence from the test (i.e., an updating factor of 20 in favor of the driver being drunk[17]) in order to arrive at the posterior odds, that is, $p(\text{sober} \mid \text{test positive})/p(\text{drunk} \mid \text{test positive})$. The odds for the driver being sober were 999 prior to the test result; the test result, however, is positive and this requires a downward adjustment by a factor of 20, so that the posterior odds for the driver being sober have been reduced to $^{999}/_{20} = 49.95$.

These steps are intuitive but they can be formalized by applying Equation 3.2 as follows:

$$\underbrace{\frac{p(\text{sober} \mid \text{test positive})}{p(\text{drunk} \mid \text{test positive})}}_{\substack{\text{Posterior uncertainty} \\ \text{about the driver}}} = \underbrace{\frac{p(\text{sober})}{p(\text{drunk})}}_{\substack{\text{Prior uncertainty} \\ \text{about the driver}}} \times \underbrace{\frac{p(\text{test positive} \mid \text{sober})}{p(\text{test positive} \mid \text{drunk})}}_{\substack{\text{Evidence} \\ \text{from the test}}}.$$

In the final step, we transform the posterior odds of 49.95 for the driver being sober to a posterior probability: $p(\text{sober} \mid \text{test positive}) = 49.95/(49.95+1) \approx 0.98$. This means that even after a positive breathalyzer test outcome, the probability that a given driver is drunk is still only about 2%.

The standard Bayesian solution to the base rate fallacy involves the law of total probability in order to compute $p(\text{positive test})$ as $p(\text{positive test} \mid \text{drunk})p(\text{drunk}) + p(\text{positive test} \mid \text{sober})p(\text{sober})$ and then use this as the denominator in a fraction with $p(\text{positive test} \mid \text{sober})p(\text{sober})$ as the numerator. The end-result is obtained in one step, but requires three simultaneous operations: multiplication, addition, and division. In contrast, the odds form of Bayes' rule is intuitive and immediately clarifies the importance of the prior odds and the separate role of evidence.[18]

[17] This is the same as an updating factor of $^{1}/_{20}$ in favor of the driver being sober; although this interpretation may be more intuitive for this specific calculation, it is generally easier to interpret ratios that are larger than 1.

[18] For a more extensive treatment see John Kruschke's blog post at `http://doingbayesiandataanalysis.blogspot.com/2015/12/lessons-from-bayesian-disease-diagnosis_27.html`.

Exercises

1. Explain the law of conditional probability using Venn diagram and lego (e.g., Kurt 2019).

2. In the left panel of Figure 3.4, explain what the '0.4' on top of the bars means; In the right panel of Figure 3.4, explain what the '0.4' in the grey area means.

3. Consider again the tennis match between Roger and Zita and the tree diagram from Figure 3.6. After the match, what is the probability that it was windy, given that you know that Zita won?

4. If you throw a fair die twice, what is the chance of obtaining at least one six? Plot the sample space as a six-by-six grid, and explain two

Poincaré on the Base Rate Fallacy

"An effect may be produced by the cause A or by the cause B. The effect has just been observed. We ask the probability that it is due to the cause A. This is an *a posteriori* probability of cause. But I could not calculate it, if a convention more or less justified did not tell me *in advance* what is the *a priori* probability for the cause A to come into play; I mean the probability of this event for some one who had not observed the effect.

The better to explain myself I go back to the example of the game of écarté mentioned above [see the box in Chapter 2 – EWDM]. My adversary deals for the first time and he turns up a king. What is the probability that he is a sharper? The formulas ordinarily taught give $8/9$, a result evidently rather surprising. If we look at it closer, we see that the calculation is made as if, *before sitting down at the table*, I had considered that there was one chance in two that my adversary was not honest. An absurd hypothesis, because in that case I should have certainly not played with him, and this explains the absurdity of the conclusion.

The convention about the *a priori* probability was unjustified, and that is why the calculation of the *a posteriori* probability led me to an inadmissible result. We see the importance of this preliminary convention. I shall even add that if none were made, the problem of the *a posteriori* probability would have no meaning. It must always be made either explicitly or tacitly." (Poincaré 1913, p. 169; italics in original) ways of obtaining the answer. Repeat the exercise for the case of three throws (you now need a cube).

5. Figures 3.1 and 3.2 concern two concrete examples in probability. Discuss the extent to which each is either epistemic or aleatory in nature (see previous chapter).

6. This is the chorus of Jeff Wayne's 'The Eve of the War':

   "The chances of anything coming from Mars
   Are a million to one, he said (ah, ah)
   The chances of anything coming from Mars
   Are a million to one, but still, they come..."

   Is the statement "a million to one" really a *chance*?

7. The following fragment is taken from the section 'The Puzzle of the Three Prisoners' in Lindley (1985). First formulated by Martin Gardner (i.e., Gardner 1959a for the problem statement; Gardner 1959b for the solution; see also Gardner 1961), this puzzle anticipates the famous 'Monty Hall problem'. An earlier version of this problem was proposed by French mathematician Joseph Bertrand (1822–1900) in his 1889 book *Calcul des Probabilités* – an English translation can be found in the box that concludes this chapter.

   > "A problem which intrigues many people and also demonstrates the notion of coherence in an interesting way is that of the three prisoners. Alan, Bernard, and Charles are in jail unable to communicate with one another or with anyone besides their respective jailers. Alan knows that two of them are to be executed and the other set free, and after some thinking concludes that he has no reason to think that one of them is more likely to be the lucky one than either of the others. If A denotes the event that Alan will go free, and B and C similarly for Bernard and Charles, this last statement means that $p(A) = p(B) = p(C) = 1/3$ in Alan's opinion. Alan now says to his jailer 'Since either Bernard or Charles is certain to be executed, you will give me no information about my own chances if you give me the name of one man, Bernard or Charles, who is going to be executed.' Accepting this argument the jailer truthfully says 'Bernard will be executed.' Thereupon Alan feels happier because now either he or Charles will go free and, as before, he has no reason to think it is more likely to be Charles, so his chance is now $1/2$, not $1/3$, as before. Which argument is correct, the one that convinced the jailer or the latter one?" (Lindley 1985, pp. 41-42)

8. "The Smiths have exactly two children, and at least one is a girl. Assume for simplicity that boys and girls are equally likely (...) and that children are one or the other (...). Assume also that the sexes of this children are independent random variables (...)."
   (a) "What is the probability that the Smiths have two girls?"

(b) "Now suppose that the *elder* child is a girl. What is the probability that they have two girls?"
(c) "Finally, suppose that at least one is a girl born on a Tuesday. What is the probability that they have two girls? (Assume all days of the week are equally likely – also not true in reality, but not too far off.)" (Stewart 2019, p. 66; pp. 70-75)

9. Consider the British court case of Sally Clark (Dawid 2005, Hill 2005, Nobles and Schiff 2005):

> "Clark had experienced a double tragedy: Her two babies had both died, presumably from cot death or sudden infant death syndrome (SIDS). If the deaths are independent, and the probability of any one child dying from SIDS is roughly $1/8,543$, the probability for such a double tragedy to occur is as low as $1/8,543 \times 1/8,543 \approx 1$ in 73 million. Clark was accused of killing her two children, and the prosecution provided the following statistical argument as evidence: Because the probability of two babies dying from SIDS is as low as 1 in 73 million, we should entertain the alternative that the deaths at hand were due not to natural causes but rather to murder. And indeed, in November 1999, a jury convicted Clark of murdering both babies, and she was sentenced to prison." (Rouder et al. 2016a, p. 521)

Based on the statistical argument alone, was the jury correct in sentencing Sally Clark to prison?

10. de Finetti (1974, pp. 154-155) explained that gamblers often use odds instead of probability. As before, we define the odds for an event A by $o(A) = r = p(A)/p(\neg A)$. The odds "are usually expressed as a fraction or ratio, $r = h/k = h : k$ (h and k integers, preferably small), by saying that the odds are '*h to k on*' the event, or '*k to h against*' the event. Of course, given r, that is the odds, or, as we shall say, the *probability ratio*, the probability can immediately be obtained by

$p = r/(r+1)$, i.e. (if r is written as h/k) $p = h/(h+k)$"

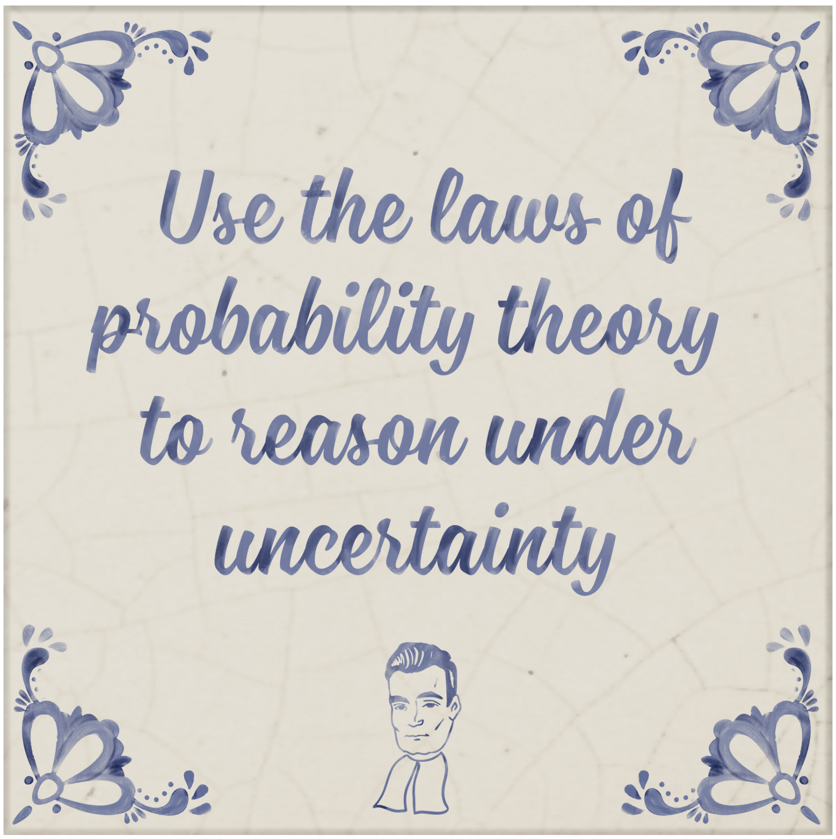
De Finetti then presents a version of Table 3.1 with examples:

Table 3.1: Examples of the correspondence between probabilities and odds, based on de Finetti (1974, p. 155).

Probability	Odds	$= r$	$= h/k$	in words	(check) $h/(h+k) = p$
0.20	$20/80$	$= 0.25$	$= 1/4$	'4 to 1 against'	$1/(1+4) = 0.20$
$2/7 = 0.286$	$28.6/71.4$	$= 0.40$	$= 2/5$	'5 to 2 against'	$2/(2+5) = 0.286$
0.50	$50/50$	$= 1$	$= 1/1$	'evens'	$1/(1+1) = 0.50$
0.75	$75/25$	$= 3$	$= 3/1$	'3 to 1 on'	$3/(3+1) = 0.75$

Finally, the questions: (a) what is a probability of $^5/_7$ 'in words', and how could it have been deduced directly from the information in

Table 3.1? (b) a bookie offers 13/2 odds on Holy Moly to win the Kentucky Derby. This means that if you bet \$2 on Holy Moly, and Holy Moley wins, you gain \$13 (i.e., the total payout equals \$15: \$13 plus your initial \$2 stake). In continental Europe, a popular alternative to the traditional/fractional/British odds are so-called decimal odds. The decimal odds represents the total payout for every unit (dollar, say) that is wagered. What are the decimal odds for Holy Moly, and how can they be obtained from the traditional odds in general?



Chapter Summary

This chapter provided an overview of the elementary laws of probability theory: the sum rule, the multiplication rule, the definition of conditional probability and marginal probability, and Bayes' rule. Bayes' rule was presented both in its probability form and its odds form. The odds form is particularly convenient when it comes to knowledge updating, and it makes it easier to avoid the base rate fallacy.

Richard Feynman on Doubt and Certainty

Nobel-laureate Richard Feynman (1918-1988) is one of the most famous physicists from the 20th century. A brilliant researcher, a gifted communicator, and a devoted advocate of science, Feynman's legacy is now tainted by revelations concerning sexual misconduct and domestic violence. An FBI report on Feynman (`https://cdn.muckrock.com/foia_documents/Feynman_Master_of_Deception.pdf`) states that in 1956, "His ex-wife reportedly testified that on several occasions when she unwittingly disturbed either his calculus or his drums he flew into a violent rage, during which time he choked her, threw pieces of bric-a-brac about and smashed the furniture." Below are two of Feynman's statements about doubt and certainty that are relevant in the context of this book.

> "(…) it is imperative in science to doubt; it is absolutely necessary, for progress in science, to have uncertainty as a fundamental part of your inner nature. To make progress in understanding, we must remain modest and allow that we do not know. Nothing is certain or proved beyond all doubt. You investigate for curiosity, because it is *unknown*, not because you know the answer. And as you develop more information in the sciences, it is not that you are finding out the truth, but that you are finding out that this or that is more or less likely.
>
> That is, if we investigate further, we find that the statements of science are not of what is true and what is not true, but statements of what is known to different degrees of certainty (…) Every one of the concepts of science is on a scale graduated somewhere between, but at neither end of, absolute falsity or absolute truth.
>
> It is necessary, I believe, to accept this idea, not only for science, but also for other things; it is of great value to acknowledge ignorance. It is a fact that when we make decisions in our life, we don't necessarily know that we are making them correctly; we only think that we are doing the best we can–and that is what we should do."
>
> (Feynman 1999, pp. 247-248)

> "You see, one thing is, I can live with doubt and uncertainty and not knowing. I think it's much more interesting to live not knowing than to have answers which might be wrong. I have approximate answers and possible beliefs and different degrees of certainty about different things, but I'm not absolutely sure of anything and there are many things I don't know anything about (…) I don't have to know an answer, I don't feel frightened by not knowing things. (Feynman 1999, pp. 24-25)

Want to Know More?

✓ Grant Sanderson's YouTube channel '3Blue1Brown' presents fascinating visualizations of a wide range of mathematics, including probability theory. To the two videos referenced in the margin of this chapter we would like to add "Bayes theorem, the geometry of changing beliefs". 3Blue1Brown is creative, informative, and fun – do check it out.[19]

[19] At the time of writing, 3Blue1Brown has over 5.5 million followers, so we are not alone in our appreciation.

✓ Bolstad, W. M. (2007). *Introduction to Bayesian Statistics (2nd ed.)*. Hoboken, NJ: Wiley. Chapter 4 provides an accessible and concise overview of key concepts and laws in probability theory.

✓ Blitzstein, J. K., & Hwang, J. (2019). *Introduction to Probability (2nd ed.)*. Taylor & Francis Group. Fabian Dablander: "I recommend this book and online lectures to everybody who wants to get started with probability. The new edition of his book is freely available online, written in great style, and has lots of very good exercises." More information is available at `https://projects.iq.harvard.edu/stat110/home`. The book also comes with a very good cheat sheet.

✓ De Morgan, A. (1838). *An Essay on Probabilities and on Their Application to Life Contingencies and Insurance Offices*. London: Longman. An oldie but a goodie. Contains a number of exercises.

✓ Kurt, W. (2019). *Bayesian Statistics the Fun Way*. San Francisco: No Starch Press. Highly recommended. From a review on `BayesianSpectacles.org`: "As a first introduction to Bayesian inference, this book is hard to beat. It nails the key concepts in a compelling and instructive fashion."

✓ Lindley, D. V. (2006). *Understanding Uncertainty*. Hoboken: Wiley. We should really resist the temptation to recommend this book at the end of every chapter.

✓ Marks, S., & Smith, G. (2011). The two-child paradox reborn? *CHANCE*, 24, 54-59. Just when you think you understand the two-child paradox, this article comes along to make you rethink your entire reasoning process. The authors conclusion: "There is no paradox" (p. 58).

✓ Nickerson, R. S. (1996). Ambiguities and unstated assumptions in probabilistic reasoning. *Psychological Bulletin*, 120, 410–433.

> "The results of a considerable amount of research have been taken as evidence that people's intuitions about probability are faulty. Some of the problems that have been used to study those intuitions, and to study reasoning under uncertainty more generally, are ambiguous and not solvable in the absence of assumptions." (p. 430)

✓ Rouder, J. N., & Morey, R. D. (2019). Teaching Bayes' theorem: Strength of evidence as predictive accuracy. *The American Statistician*, 73, 186-190. Outlines the evidential interpretation of Bayes' theorem.

✓ Stewart, I. (2019). *Do Dice Play God? The Mathematics of Uncertainty*. New York: Basic Books. Ian Stewart is a great writer, and, on pages 70-75, he explains the two-child paradox particularly clearly in terms of restricted sample spaces (for details see the answer to the last exercise above). However, Marks and Smith (2011, p. 59) argue this approach answers the wrong question:

> "A general question is how best to accommodate new information into the evaluation of uncertain situations. Use of the restricted sample space approach for the two-child problem does not yield a proper conditional probability that a family has, say, two girls, given that one has learned that one of the children is a girl. All it offers, in this case, is a hypothetical calculation of the fraction of BG, GB, and GG families that are GG. In the classic two-child problem, it also offers an erroneous illusion of simplicity–that, in general, a two-child family is equally likely to be BG, GB, or GG if we learn one of the children is a girl.
>
> In contrast, the Bayesian approach provides useful conditional probabilities that can be applied directly to a family at hand as we acquire new information about it. It also provides discipline in that it requires us to be clear about the full set of assumptions that enter into our probabilistic inferences."

✓ Taylor, D. G. (2021). *Games, Gambling, and Probability: An Introduction to Mathematics (2nd ed.)*. Boca Raton: CRC Press. An accessible introduction, especially suitable for those who remain confused about the relation between probability and odds.

The next page provides a liberal translation of Bertrand's famous "box paradox", by Nick Brown and EW. A literal translation by Bianca van Rossum is available at `https://tinyurl.com/Bertrandliteral`. Another famous –and much more challenging– Bertrand paradox in probability theory illustrates how subtly different conceptualizations of a seemingly straightforward problem can give dramatically different answers (e.g., Aerts and de Bianchi 2014).[20]

[20] See also `https://en.wikipedia.org/wiki/Bertrand_paradox_(probability)` and two episodes of the YouTube channel 'Numberphile'.

A Liberal Translation of Joseph Bertrand's Box Paradox

"There are three identical-looking boxes. Each box has two drawers and each drawer contains one coin. In the first box, each drawer contains a gold coin; in the second, each drawer contains a silver coin; and in the third, one drawer contains a gold coin and the other contains a silver coin.

One of the three boxes is chosen at random. What is the probability of finding one gold coin and one silver coin?

The answer seems obvious: There are three equally possible cases. Only one case gives the required outcome (one coin of each type). Hence, the probability is $1/3$.

However, now consider what happens if, after choosing the box, we open one of its drawers at random. Let's say we see a gold coin. We now know that we did not get the box with two silver coins. We have chosen either the box with two gold coins, or the box with one gold and one silver coin. The drawer that we have not opened may therefore contain a gold coin or a silver coin, with a probability for either event of $1/2$. But now consider the alternative scenario: the first drawer reveals a silver coin. The same reasoning again leads to a probability of $1/2$ for the unopened drawer to contain either a gold coin or a silver coin. So regardless of whether the first drawer shows a gold coin or a silver coin–and it is certain to show one of the two–the probability of finding a non-matching coin in the second drawer is $1/2$. We therefore conclude that the mere act of opening a drawer changes the probability, increasing it from $1/3$ to $1/2$.

The reasoning cannot be correct. And in fact it is not.

It is true that, after opening the first drawer and seeing a gold coin, two cases (gold-gold and gold-silver) remain possible. It is also true that only one of these two gives us the gold-silver combination, whose probability we were asked to find. But the crucial point here is that these two cases were not equally likely to have happened in the first place.

To make this clearer, imagine that instead of three boxes we have three hundred: A hundred contain two gold coins, a hundred contain two silver coins, and a hundred contain one gold coin and one silver coin. We open one drawer of each box, revealing a total of 300 coins. For the hundred "double-gold" and the hundred "double-silver" boxes, we know that we will always see a gold or a silver coin, respectively. For the other hundred boxes, those with a gold and a silver coin, the proportions will be determined by chance, but we will probably see about 50 of each. However, we know that of the roughly 150 gold coins we see, 100 of them are in a gold-gold box and only 50 are in a gold-silver box. There (50 out of 150) is our correct probability of $1/3$.

You can also see that, if we were asked to choose one of the open boxes in which we see a gold coin and to bet on what color the other coin in that box is, we would be wise to bet on gold, because in two-thirds of cases (100 out of 150) we would be right. Again, this corresponds to the fact that one-third of the boxes in which we can see a gold coin in the open drawer have a silver coin in the other (closed) drawer, whereas two-thirds have a gold coin in the other drawer." (Bertrand 1889, pp. 2-3; see also `https://tinyurl.com/Bertrandliteral`)

4 *Interlude: Leibniz's Blunder*

It is very curious how often the most acute and powerful intellects have gone astray in the calculation of probabilities.

Jevons, 1874

Chapter Goal

This chapter demonstrates that probability theory trips up even mathematical geniuses of the highest order.

Gottfried Wilhelm Leibniz

Gottfried Wilhelm Leibniz was a scientist whose name will never be forgotten. He invented calculus[1], and when we write, for instance, $\int p(y, \theta)\, \mathrm{d}\theta$, we owe him the signs $\int$ and d. In addition, Leibniz proposed that time and space are relative (anticipating Einstein), and argued that the earth has a molten core (a hypothesis confirmed in 1926 by the hero of this book, Sir Harold Jeffreys, before it was corrected to its modern form by Inge Lehmann in 1936, with additional contributions from Arwen Deuss in 2000). Leibniz also made pioneering contributions to psychology (influencing Wilhelm Wundt) and theology (e.g., to retain the notion that God is both omnipotent and benevolent, Leibniz argued that we live in the best of all possible worlds – see the box on *Theodicity* below). He invented the first mechanical calculator to do addition, subtraction, multiplication, and division. Leibnitz wrote in Latin, French, and German, but also in English, Italian, and Dutch. As detailed on Wikepedia, "Leibniz made major contributions to physics and technology, and anticipated notions that surfaced much later in philosophy, probability theory, biology, medicine, geology, psychology, linguistics, and computer science. He wrote works on philosophy, politics, law, ethics, theology, history, and philology."

[1] Independently from Newton, at around the same time.

Portrait of Gottfried Wilhelm Leibniz (1646-1716) by Christoph Bernhard Francke.

In addition to all of these accomplishments, Leibniz raised the spirits of future generations of students who find themselves struggling with probability theory. Leibniz accomplished this by committing a blunder.

The Blunder

Probability theory presents a minefield of mistakes and misconceptions. Is there another discipline in which mathematicians made so many false claims? As summarized by Jevons (1874/1913):

> "The doctrine of probability, though undoubtedly true, requires very careful application. Not only is it a branch of mathematics in which positive blunders are frequently committed, but it is a matter of great difficulty in many cases, to be sure that the formulæ correctly represent the data of the problem. [...]
>
> It is very curious how often the most acute and powerful intellects have gone astray in the calculation of probabilities. Seldom was Pascal mistaken, yet he inaugurated the science with a mistaken solution.[2] Leibnitz fell into the extraordinary blunder of thinking that the number twelve was as probable a result in the throwing of two dice as the number eleven.[3] In not a few cases the false solution first obtained seems more plausible to the present day than the correct one since demonstrated. James Bernouilli candidly records two false solutions of a problem which he at first thought self-evident;[4] and he adds an express warning against the risk of error, especially when we attempt to reason on this subject without a rigid adherence to the methodical rules and symbols.[5] Montmort was not free from similar mistakes,[6] and as to D'Alembert, great though his reputation was, and perhaps is, he constantly fell into blunders which must diminish the weight of his opinions.[7] He could not perceive, for instance, that the probabilities would be the same when coins are thrown successively as when thrown simultaneously.[8] Some men of high ability, such as Ancillon, Moses Mendelssohn, Garve,[9] Auguste Comte[10] and J. S. Mill,[11] have so far misapprehended the theory, as to question its value or even to dispute altogether its validity.
>
> Many persons have a fallacious tendency to believe that when a chance event has happened several times together in an unusual conjunction, it is less likely to happen again. D'Alembert seriously held that if head was thrown three times running with a coin, tail would more probably appear at the next trial.[12] Bequelin adopted the same opinion, and yet there is no reason for it whatever. If the event be really casual, what has gone before cannot in the slightest degree influence it.
>
> As a matter of fact, the more often the most casual event takes place the more likely it is to happen again; because there is some slight empirical evidence of a tendency, as will afterwards be pointed out. The source of the fallacy is to be found entirely in the feelings of surprise with which we witness an event happening by apparent chance, in a manner which seems to proceed from design." (Jevons 1874/1913, pp. 243-245)

[2] Montucla, 'Histoire des Mathématiques,' vol. iii. p. 386

[3] Leibnitz 'Opera,' Dutens' Edition, vol. vi. part i. p. 217. Todhunter's 'History of the Theory of Probability,' p. 48.

[4] Todhunter, pp. 67-69.

[5] Ibid. p. 63.

[6] Ibid. p. 100.

[7] Ibid. pp. 258-59, 286.

[8] Todhunter, p. 279.

[9] Ibid. p. 453.

[10] 'Positive Philosophy,' translated by Martineau, vol. ii. p. 120.

[11] 'System of Logic,' bk. iii. chap. 18. 5th Ed. vol. ii. p. 61.

[12] Montucla, 'Histoire,' vol. iii. p. 405. Todhunter, p. 263.

EWDM: Gorroochurn (2011, p. 250) mentions that d'Alembert was "a man of immense mathematical prowess" and that he had a strong basis for his probabilistic reasoning. D'Alembert's thinking "was partly responsible for later mathematicians seeking a solid theoretical foundation for probability, culminating in its axiomatization by Kolmogorov in 1933 (Kolmogorov 1933)."

Wait, what is this? Did the immortal Leibniz truly suggest that "the number twelve was as probable a result in the throwing of two dice as the number eleven"? We find more details in Todhunter (1865), the absolute authority on early works in probability theory:

> "Leibnitz took great interest in the Theory of Probability and shewed that he was fully alive to its importance, although he cannot be said

himself to have contributed to its advance. There was one subject which especially attracted his attention, namely that of games of all kinds; he himself here found an exercise for his inventive powers. He believed that men had nowhere shewn more ingenuity than in their amusements, and that even those of children might usefully engage the attention of the greatest mathematicians. He wished to have a systematic treatise on games, comprising first those which depended on numbers alone, secondly those which depended on position, like chess, and lastly those which depended on motion, like billiards. This he considered would be useful in bringing to perfection the art of invention, or as he expresses it in another place, in bringing to perfection the art of arts, which is the art of thinking.

See *Leibnitii Opera Omnia, ed. Dutens*, Vol. V. pages 17, 22, 28, 29, 203, 206. Vol. VI. part 1, 271, 304. *Erdmann*, page 175.

See also *Opera Omnia, ed. Dutens*, Vol. VI. part 1, page 36, for the design which Leibnitz entertained of writing a work on estimating the probability of conclusions obtained by arguments.

Leibnitz however furnishes an example of the liability to error which seems peculiarly characteristic of our subject. He says, *Opera Omnia, ed. Dutens*, Vol. VI. part 1, page 217,

...par exemple, avec deux dés, il est aussi faisable de jetter douze points, que d'en jetter onze; car l'un et l'autre ne se peut faire que d'une seule manière; mais il est trois fois plus faisable d'en jetter sept; car cela se peut faire en jettant six et un, cinq et deux, quatre et trois; et une combinaison ici est aussi faisable que l'autre.[13]

It is true that eleven can only be made up of six and five; but the six may be on *either* of the dice and the five on the other, so that the chance of throwing eleven with two dice is twice as great as the chance of throwing twelve: and similarly the chance of throwing seven is six times as great as the chance of throwing twelve." (Todhunter 1865, pp. 47-48)

[13] "...for example, with two dice, it is as feasible to throw twelve as to throw eleven; because the one and the other can be done in only one way; but it is three times more feasible to throw seven; because it can be done by throwing six and one, five and two, four and three; and each combination is as feasible as another." (translation courtesy of Bruno Boutin).

Galileo 1, Leibniz 0

In their 2018 book "Ten Great Ideas About Chance", Persi Diaconis and Brian Skyrms discuss an earlier version of the problem that ensnared Leibniz:

"In the early seventeenth century Galileo composed a short note on dice to answer a question posed to him (by his patron, the Grand Duke of Tuscany). The Duke believed that counting possible cases seemed to give the wrong answer. Three dice are thrown. Counting combinations of numbers, 10 and 11 can be made in 6 ways, as can 9 and 12. '...yet it is known that long observation has made dice-players consider 10 and 11 to be more advantageous than 9 and 12.' How can this be?

Galileo replies that his patron is counting the wrong thing. He counts three 3s as one possibility for making a 9 and two 3s and a 4 as one possibility for making a 10. Galileo points out the latter covers three possibilities, depending on which die exhibits the 4:

$$< 4,3,3 >, < 3,4,3 >, < 3,3,4 > .$$

> For the former, there is only $< 3, 3, 3 >$. Galileo has a complete grasp of permutations and combinations and does not seem to regard it as anything new." (Diaconis and Skyrms 2018, pp. 4-5)

Theodicy

Leibniz was a devout Christian, and he was deeply concerned with the *problem of evil*. Diogenes the Cynic (412/404 BC – 323 BC) already argued that "the prosperity and good fortune of the wicked disprove the might and power of the gods entirely." (Cicero 45BC/1956b, III: xxxvi). Consider the holocaust as example of the ultimate evil. Now there are several options, none of them agreeable: either God did not care about the holocaust, and which case he is malicious; or he did not know about the holocaust, in which case he is not omniscient; or he was unable to prevent the holocaust, in which case he is not omnipotent. It may be argued that the holocaust is people's own fault and God wanted humanity to learn from its mistakes. One would think that the lesson could have been a little less intense. Moreover, this argument does not work for evil that appears haphazard: it is hard to see God's hand in debilitating diseases such as multiple sclerosis or Alzheimer's, and remain convinced that He has humanities best interests at heart.

At any rate, Leibniz' goal was *theodicy*, "the vindication of divine providence in view of the existence of evil." To achieve this, Leibnitz proposed a radical solution, namely to declare that we live in the best of all possible worlds (for details see `https://plato.stanford.edu/entries/leibniz-evil/`). Remove the holocaust, remove multiple sclerosis, remove Alzheimer's, and that world would be *worse* than the one we currently inhabit – perhaps because we lack a proper appreciation of overall "goodness" of the world, or because by eliminating one disease we inadvertently allow some bigger evil to arise. Leibniz's suggestion was lampooned by Voltaire in his famous book *Candide, ou l'Optimisme*.

The Emperor of China

We end with one last remarkable story about Leibniz. At some point, based on an analysis of an infinite series with alternating values of $+1$ and -1,

> "(...) Leibniz believed that creation was mirrored in his binary arithmetic, where he used only the two symbols 0 and 1. He imagined that God could be represented by 1 and Nothing by 0, and that the Supreme Being had created all matter from Nothing, just as 1 and 0 together express all the

numbers in this system of arithmetic. This idea so pleased Leibniz that he told the Jesuit Grimaldi, president of the mathematical council of China, about it, in the hope that this symbolic representation of creation would convert the emperor of that time (who had a particular predilection for the sciences) to Christianity. I recount this incident only to show just to what a degree puerile prejudices may mislead the greatest men." (Laplace 1814/1995, p. 97)

Chapter Summary

Even a scientific demigod such as Gottfried Leibniz faltered when confronted with a simple problem in probability theory. Or perhaps there are no simple problems in probability theory!

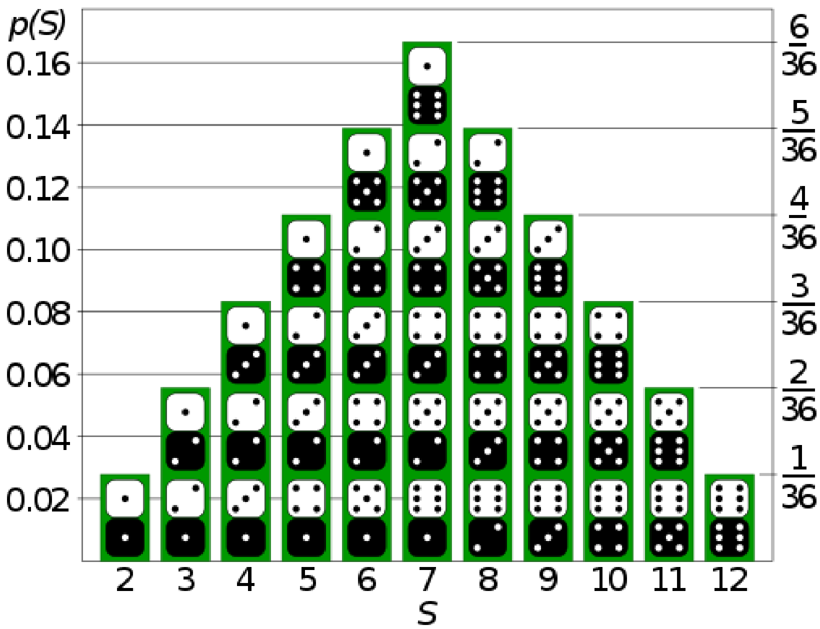


Figure 4.1: "Probability mass function of sum of two regular dice. Bar graph used to portray discrete density function. Labels on the right correspond to the $n/36$ results format." Figure available on Wikipedia (public domain), courtesy of Tim Stellmach.

Want to Know More?

✓ Gorroochurn, P. (2011). Errors of probability in historical context. *The American Statistician*, 65, 246-254. On p. 250 of this fascinating overview, the author emphasizes that, despite Leibniz's blunder, "Nonetheless, this should not in any way undermine some of the contributions Leibniz made to probability theory. For one thing, he was one of the very first to give an explicit definition of classical probability except phrased in terms of an expectation (Leibniz 1969, p. 161)[14]:

[14] EWDM: From *Théodicée*, original published in 1710.

> If a situation can lead to different advantageous results ruling out each other, the estimation of the expectation will be the sum of the possible advantages for the set of all these results, divided into the total number of results.

In spite of being conversant with the classical definition, Leibniz was very interested in establishing a logical theory for different degrees of certainty. He may rightly be regarded as a precursor to later developments in the logical foundations of probability by Keynes, Jeffreys, Carnap, and others. Since Jacob Bernoulli had similar interests, Leibniz started a communication with him in 1703. He undoubtedly had some influence in Bernoulli's *Ars Conjectandi* (Bernoulli 1713)."

✓ Todhunter, I. (1865). *A History of the Mathematical Theory of Probability From the Time of Pascal to That of Laplace*. Cambridge: MacMillan and Co. A comprehensive technical treatment.

✓ In his book *Do Dice Play God?*, mathematician Ian Stewart starts the chapter *Fallacies and Paradoxes* with a pithy remark: "Human intuition for probability is hopeless" (p. 65). Some of the pernicious misunderstandings concern the base rate fallacy (covered in Chapter 3; this is also known as the prosecutor's fallacy or transposing the conditional) and the conjunction fallacy (i.e., deeming the proposition "Linda is a bank teller" as *less* probable than the conjunctive proposition "Linda is a bank teller *and* a feminist"; see Tversky and Kahneman 1983; for a critique see Hertwig and Gigerenzer 1999).

✓ Gigerenzer, G., Multmeier, J., Föhring, A., & Wegwarth, O. (2021). Do children have Bayesian intuitions? *Journal of Experimental Psychology: General*, 150, 1041-1070. A counterweight to the prevailing opinion that people are inherently bad at solving problems in probability theory. When the problem is presented in terms of natural frequencies (i.e., as an 'icon array'), performance is surprisingly good. "A series of experiments demonstrates for the first time that icon arrays elicited Bayesian intuitions in children as young as second-graders for 22% to 32% of all problems; fourth-graders achieved 50% to 60%. Most surprisingly, icon arrays elicited Bayesian intuitions in children with dyscalculia, a specific learning disorder that has been attributed to genetic causes. These children could solve an impressive 50% of Bayesian problems, a level similar to that of children without dyscalculia. By seventh grade, children solved about two thirds of Bayesian problems with natural frequencies alone, without the additional help of icon arrays." (p. 1041).

✓ We recommend you go online to consult information on the 'Stepped reckoner', the mechanical calculator invented by Leibniz in around 1673. According to Leibniz, "It is beneath the dignity of excellent men to waste their time in calculation when any peasant could do the work just as accurately with the aid of a machine." (Martin 1925/1992, p. 38)[15]

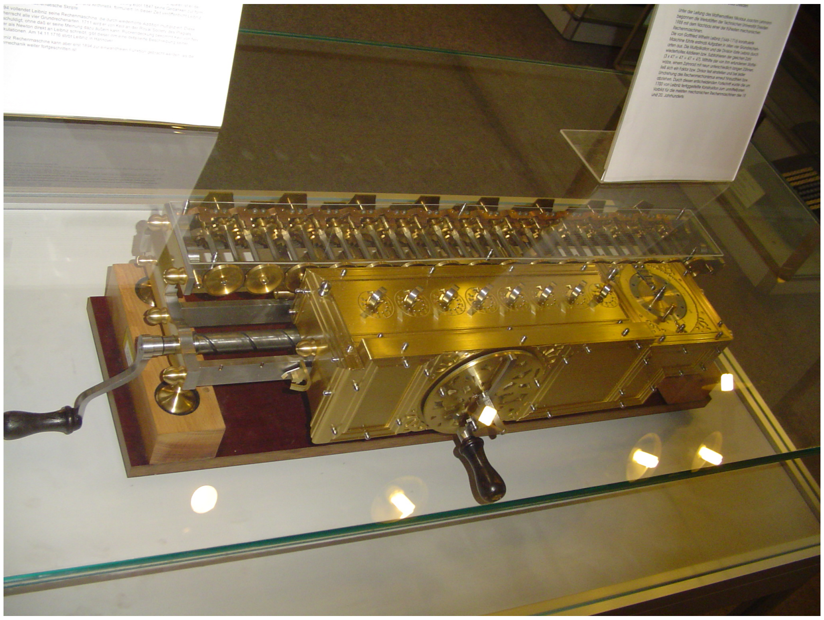


Figure 4.2: "The Staffelwalze, or Stepped Reckoner, a digital calculating machine invented by Gottfried Wilhelm Leibniz around 1672 and built around 1700, on display in the Technische Sammlungen museum in Dresden, Germany. It was the first known calculator that could perform all four arithmetic operations; addition, subtraction, multiplication and division. 67 cm (26 inches) long. The cover plate of the rear section is off to show the wheels of the 16 digit accumulator. Only two machines were made. The single surviving prototype is in the National Library of Lower Saxony (Niedersächsische Landesbibliothek) in Hannover; this is a contemporary replica." Description and photo from Kolossos, available under a CC BY-SA 3.0 license.

[15] The Latin original reads "Indignum enim est excellentium virorum horas servili calculandi labore perire, qui machina adhibita vilissimo cuique secure transcribi posset." This does not feature peasants specifically, but it does refer to "vilissimo cuique", that is, anybody without value whatsoever.

5 *The Measurement of Probability*

Almost the greatest difficulty in this subject consists in acquiring a precise notion of the matter treated. What is it that we number, and measure, and calculate in the theory of probabilities? Is it belief, or opinion, or doubt, or knowledge, or chance, or necessity, or want of art?

Jevons, 1874.

Chapter Goal

Bayesians define probability as 'degree of reasonable belief' or 'intensity of conviction'. Although the concept may seem vague, it is possible –at least in principle– to *measure* belief, that is, to compare it to a standard and assign it a number. This chapter outlines five methods by which this may be accomplished.

How to Measure Belief?

In everyday life, belief and conviction are usually conveyed in words, not in numbers. The statement 'I am pretty sure Luigi's Pizza Palace opens at 6 pm' is unremarkable, whereas the statement 'I am 85% certain that Luigi's Pizza Palace opens at 6 pm' may raise eyebrows. But words are vague and notoriously susceptible to alternative interpretation. For example, Figure 5.1 shows the results of a survey on the use of 23 words that denote various degrees of uncertainty, such as 'always', 'often', and 'possibly'. In their blog post 'If you say something is "likely," how likely do people think it is?', Andrew and Michael Mauboussin argued that some of these probabilistic words are interpreted quite broadly – for instance, some people indicated that the words 'real possibility' refer to an event with a 20% probability, whereas others indicated this to be 80%. The lesson Mauboussin and Mauboussin draw from all this? Simple: "Use probabilities instead of words to avoid misinterpretation" (cf. Mosteller and Youtz 1990, Theil 2002, Willems et al. 2020).

Instead of through words, belief can also be expressed indirectly, by decisions or *actions* – if I leave the house in order to arrive at Luigi's

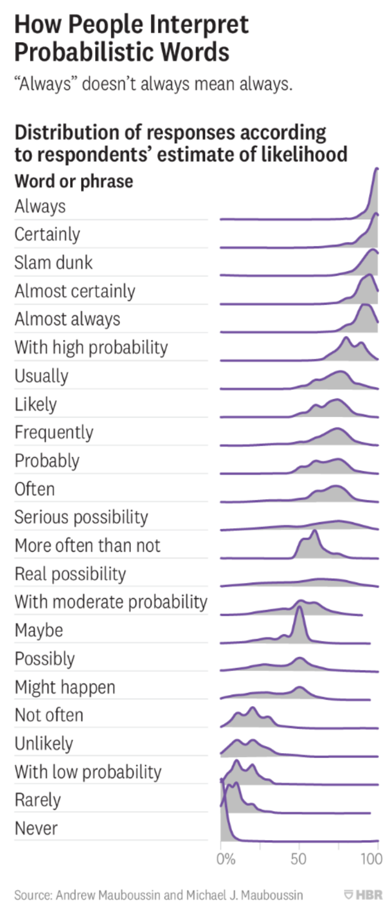


Figure 5.1: Results from a survey (`http://www.probabilitysurvey.com/`) where 1700 people assigned probabilities to 23 words that convey a degree of uncertainty. Data reported by Andrew and Michael Mauboussin. Figure reprinted with permission.

Pizza Palace by 6 pm, this act signals that I have a non-negligible degree of belief that Luigi's Pizza Palace will be open by that time. But decisions and actions are influenced not only by belief, but also by *utility*. For instance, when someone visits the doctor in order to have a mole checked out, this does not signal that the person believes there is a good chance they have skin cancer; instead, the costs of getting it wrong are wildly asymmetric – an unnecessary visit to a doctor presents only a mild inconvenience, but a tumor that goes undiagnosed can prove lethal. The decision to visit the doctor is dominated not by belief, but by utility ('better safe than sorry').[1]

[1] Dennis Lindley's 1985 book 'Making Decisions' is perhaps the clearest exposition of how belief and utility together determine decisions.

So degree of belief and intensity of conviction[2] are often expressed in words, reflected in decisions, but rarely quantified in numbers. Notable exceptions are the betting office, the insurance industry, and the stock market. Here the entire business model is predicated on uncertainty – people speculate on what will happen in the future, and to some degree their financial decisions are a numerical reflection of their beliefs.[3]

[2] Jeffreys (1937b, p. 253) suggests 'degree of knowledge'.

[3] It is perhaps not a coincidence that the study of probability started with applications in gambling and insurance (e.g., Stigler 1986a, Todhunter 1865).

Real-life experience with the vagueness of beliefs and convictions may suggest that the concept is so slippery that it eludes quantitative treatment. But before giving up so soon after we have started, let's consider what a numerical assessment of belief would require. In general, measurement requires comparison to a standard:

> "Any measurement is constructed by reference to a standard. Length is described in terms of the wavelength of sodium light; time by reference to the oscillation of a crystal. It is therefore sensible to attempt the same comparative technique when measuring uncertainty. Before doing this note that actual measurements are not made by using the standard. We do not assess the size of the table by sodium light; a tape-measure or similar device is used. Consequently the reference to a standard for uncertainty is not usually a practical way of measuring it. Rather it provides a definition and, more importantly, enables important properties of the measure to be found. A vital feature of numerical uncertainty is the rules that it has to obey." (Lindley 1985, p. 17)

Let's see how this plays out in five concrete methods.

Method I. De Finitti's Bet

Suppose we wish to measure the intensity of conviction concerning event E. For concreteness, let's say E is 'within the next five years there will be a successful coup in Venezuela'. The most intuitive way to measure belief in E is by having people bet on it. For instance, in a *prediction market*, participants can buy and sell 'shares' of E, and the market price provides a reasonable indication of the shared opinion about how likely E is to transpire. For instance, let's say the price of a share of E stands at \$0.60; this means that when you buy a share of E, this costs you \$0.60, but will pay out \$1 in case E indeed transpires; if E does not transpire, the share loses its value. If people believe that a coup is very likely to happen, \$0.60 is an attractive price and many shares may initially be bought for that price. However, this demand drives up the price until it stabilizes at the value that the market believes to be fair.

The problem with most betting scenarios is that the bettor is risking part of his wealth, and elements of risk and utility pollute the measure. This limitation can be circumvented by the following scheme, also proposed by de Finetti. Suppose there exists a ticket that pays \$1 if event E transpires. You have to determine a fair price for the ticket, but I can then decide whether to buy the ticket from you or sell the ticket to you (for that price). This is similar to two people dividing a cake fairly: one person cuts, the other person chooses.

Method II. Lindley's Urn

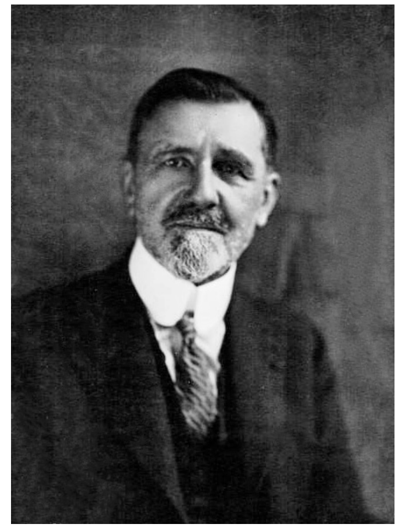
In the section 'Measurement by Reference to a Standard', Lindley (1985) proposed to measure uncertainty with the help of an urn[4]:

> "The contents are 100 balls as near identical as possible except that some are coloured black and the rest white.(...) A ball is drawn from the urn in such a way that you think each of the 100 balls has the same chance of being drawn. (...) Consider the uncertain event B that the withdrawn ball is black. The uncertainty clearly depends on how many black balls are truly in the urn. If b are black, and $100 - b$ white, the probability of the event B is defined to be $b/100$ or b%. Thus, if 50 are black, the probability is ½ or 50%. This is the standard to which all uncertain events will be referred: or rather, the set of standards for differing numbers b of black balls from 0 to 100.
>
> Now consider any uncertain event E. To fix ideas take the event that it will rain tomorrow in London. Now suppose you were to be offered a small prize if the event occurred: if it did not, you would get nothing. No stake is involved. Next, suppose you were to be offered the same prize if a black ball were to be drawn from the urn under the conditions already described. That is, there are two gambles, one contingent on E, rain, the other on B, a black ball, but otherwise identical. Granted that you may only have one gamble, which do you prefer? Again it depends on the number b of black balls. If there are none it would be best to gamble on rain: at the other extreme with all black balls, the urn is better. Generally, the more black balls the better is the urn gamble. It easily follows that there must be a particular number of black balls such that you are indifferent between two gambles: call this number b. Were there $(b + 1)$ balls the urn gamble would improve and be better than the rain one: with $(b - 1)$ it would be worse. The event B has probability $b/100$ or b%. Since the two gambles are now in all respects equivalent we say the probability of E, rain tomorrow in London, is also b%." (Lindley 1985, pp. 17-18)

[4] The following urn scheme is called the 'de Finetti game' by Devlin (2008, pp. 159–164); as discussed below, the essence of this setup dates back at least to 1838.

Dennis Victor Lindley (1923–2013). Photo taken ca. 1964-1968. Included by permission of Janet, Rowan, and Robert Lindley.

The three conceptual ingredients of the urn scheme are: (1) there is not a stake to be risked, but a prize to be gained. This removes complications related to the diminishing returns of money and the fact that people are generally risk-averse (i.e., unwilling to gamble); (2) the standard is itself an uncertain event, but with uncertainty well understood and quantified; (3) the standard is adjusted (i.e., the contents of the urn changed) until a point of indifference is reached. The next two methods –the one mentioned by Borel and the one proposed by De Morgan– echo this idea.



Félix Édouard Justin Émile Borel (1871–1956). Photo taken 1932; public domain, courtesy of Bibliothèque nationale de France.

Method III. Borel's Dice

In the section 'The Probability of an Isolated Case', the great French probabilist Émile Borel discusses how probability may be measured. The procedure is conceptually identical to Lindley's urn. The first edition of the Borel book came out in French as early as 1909 but appears to be missing the following fragment:

> "(...) let us consider a match between two tennis players who have never played against one another; however, each of them has played in many tournaments and an enlightened amateur can appreciate the quality of their play. Suppose now that we ask such an amateur to evaluate the probability that one of the two players will win the match. It is assumed that the match is of sufficient importance so that each player will make a maximum effort to win.
>
> If the amateur does not recognize probabilities referring to isolated events, he might refuse to evaluate this probability, since it refers to an event which (so far as we are concerned) cannot be reproduced a second time. To force him to give us an evaluation we might resort to methods based on betting. One cannot force a person to bet, that is, risk part of his fortune, but few persons would refuse to accept a present offered in exchange for a small intellectual effort. We thus make the amateur the following proposition: We offer him a certain amount which he can win in two different ways, either by rolling at least 10 with three dice or by betting on player A. If he chooses the second alternative, that is, he prefers to bet on player A, we can conclude that he regards the probability of this event as greater than that of betting on the dice, namely, greater than 0.50.[5] Then we could ask him to choose between betting on player A or betting on getting 1, 2, 3, or 4 with a single die. If he chooses the last alternative, which has a probability of 2/3, we can conclude that he considers the probability of player A winning as being less than 2/3. We have thus obtained two limits, 0.50 and 0.67, containing the probability p that player A will win. It would be possible to obtain more stringent limits by analogous means, so that the result would be exact to at least one decimal; for example we might find that the probability is contained between 0.50 and 0.60, It might seem that this result is rather crude, but it often happens in the natural sciences that certain experimental constants are known only very crudely, and such approximate knowledge certainly differs from total ignorance." (Borel 1965, pp. 167-168)

[5] EWDM: The probability of rolling at least 10 with three dice is actually 62.5. Borel must have meant to write "rolling at least 11", which does yield 0.50. Pointed out to us by Arne John.

Borel proposed a similar procedure in a 1924 article, *A propos d'un traité de probabilités*, later translated to English:

> "I can in the same way offer to someone who enunciates a judgment capable of verification a bet on his judgment. If I want to avoid having to account for the attraction or repugnance which inspires the bet, I can offer a choice between two bets procuring the same advantages in case of gain. Paul claims that it will rain tomorrow; I agree that we are in accord on the precise meaning of this claim and I offer him the choice of receiving 100 francs if he is correct or 100 francs if he receives a 5 or a 6 in a throw of dice. In the second case the probability of receiving 100 francs is one third; if he then prefers to receive 100 francs if his meteorological prediction is correct, it is because he attributes to this prediction a probability superior to one third. The same method can be applied to all verifiable judgments; it allows a numerical evaluation of probabilities with a precision quite comparable to that with which one evaluates prices." (Borel 1964, p. 57)

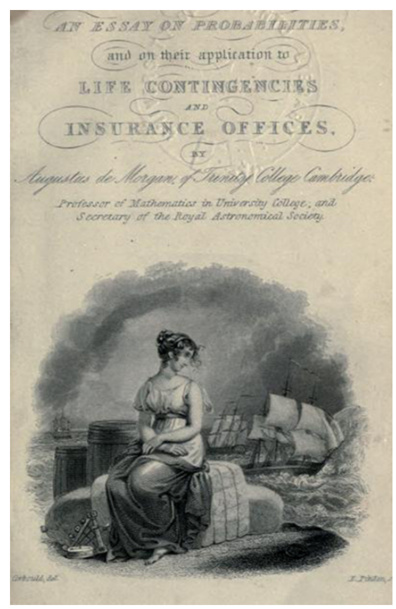
Method IV. De Morgan's Alphabet

The scenarios sketched by Borel and Lindley were anticipated by Augustus De Morgan. First, in De Morgan's 1849 encyplodedia entry 'Theory of Probabilities', De Morgan discusses the measurement problem and offers the urn as a solution:

> "The notion we mean is this; we assert and require it to be granted that the feeling of probability or improbability is of the same kind, whatever may be the event in question; that the probability we attach to one event, say a fact in history, *bears a ratio* to that which we attach to any other of another kind, say the gaining of a prize in a lottery. (...) with regard to probability, or the state of mind which produces it, if we were empowered to put the following question, we conceive that there would be but one answer. "There are two events, one past and one to come, on neither of which are you in possession of total and mathematical certainty. The first is the execution of Charles I.; the second is the drawing of a white ball from an urn which contains one white and ninety-nine black balls. Choose one of these, and let your interest in any way depend on your deciding rightly the one you select: would you rather the safety of your life should depend upon your saying correctly whether Charles I. was or was not executed, or upon your drawing the white ball, and not one of the black ones?" "(De Morgan 1849, p. 395)

Even earlier, in his 1838 book 'An Essay on Probabilities and on Their Application to Life Contingencies and Insurance Offices', De Morgan had proposed a similar but more elaborate scenario. Here we also encounter the crucial remark that the 'feeling of probability' is comparable for different events, and it is this comparability that allows quantitative measurement.

> "On this we remark, firstly, that by it we feel sensible of our assent and dissent to propositions derived in very different ways, being a sort of



Title page of Augustus De Morgan's 1838 book 'An Essay on Probabilities and on Their Application to Life Contingencies and Insurance Offices'. Does the lady who watches the ships perhaps represent Fortuna, the goddess of chance? The names of the artists at the bottom of the page suggests this is an engraving of a Henry Corbould painting – we have been unable to confirm this.

> impression which is of the same kind in all. To make this clearer, observe the following:—A merchant has freighted a ship, which he expects (is not certain) will arrive at her port. Now suppose a lottery, in which it is quite certain that every ticket is marked with a letter, and that all the letters enter in equal numbers. If I ask him, which is most probable, that his ship will come into port, or that he will draw no letter if he draw, he will answer, unquestionably, the first, for the second will certainly not happen. If I ask, again, which is most probable, that his ship will arrive, or that he will, if he draw, draw either a, or b, or c, ... or x, or y, or z, he will answer, the second, for it is quite certain. Now suppose I write the following series of assertions:—
>
> He will draw no letter (a drawing supposed).
> He will draw a.
> He will draw either a or b.
> He will draw either a, or b, or c.
> ....
> ....
> He will draw either a or b or .... or y.
> He will draw either a or b or .... or y or z.
>
> and making him observe that there are, of their kind, propositions of all degrees of probability, from that which cannot be, to that which must be, I ask him to put the assertion that his ship will arrive, in its proper place among them. This he will perhaps not be able to do, not because he feels that there is no proper place, but because he does not know how to estimate the force of his impressions in ordinary cases. If the voyage were from London Bridge to Gravesend, he would (no steamers being supposed) place it between the last and last but one: if it were a trial of the north-west passage, he would place it much nearer the beginning; but he would find difficulty in assigning, within a place or two, where it should be. All this time he is attempting to compare the magnitude of two very different kinds (as to the sources whence they come) of assent or dissent; and he shows by the attempt that he believes them to be of the same sort. He would never try to place the *weight* of his ship in its proper position in a table of *times* of high water." (De Morgan 1838, pp. 4-5)

As already noted in chapter 2, 'Epistemic and Aleatory Uncertainty', it is evident that De Morgan subscribes to a thoroughly subjectivist interpretation of probability.

Method V. Ramsey's Farmer

Despite dying at a young age, Frank Ramsey has had a profound impact on the field of probability and inference. In his book 'Making Decisions', Lindley lionizes Ramsey to the point of hyperbole:

> "The basic ideas discussed in this book were essentially discovered by Frank Ramsey, who worked in Cambridge in the 1920s. To my mind Ramsey's discoveries in the twentieth century are as important to mankind as Newton's made in the same city in the seventeenth. Newton discovered the laws of mechanics, Ramsey the laws of human action." (Lindley 1985, p. 64)

In a famous paper, Ramsey (1926) casually mentions how one could measure degree of uncertainty by means of a farmer. The story, illustrated in Figure 5.2, unfolds as follows. Harriet stands on a T-junction and needs to walk distance d to arrive at her hotel in the village of Rottevalle. Her confidence or belief that the correct way is to the right is indicated by p. If Harriet chooses the wrong direction, however, she will travel distance d and find herself in the village of Eastermar, after which she has to walk back another $2d$ before finally arriving at Rottevalle, for a total distance of $3d$ if she is wrong. Alternatively, Harriet can walk distance f to a friendly Frisian farmer who will point her to Rottevalle for sure; walking to the farmer and back, and then walking to Rottevalle implies a distance of $2f + d$. Harriet's degree of uncertainty $1 - p$ that she needs to go right to end up in Rottevalle can be measured by that distance f between Harriet and the farmer where Harriet is exactly indifferent between (1) guessing the direction and risk going the wrong way; and (2) walking up to the farmer to ask for directions. The larger the distance f that Harriet is willing to walk to obtain the farmer's advice, the larger her uncertainty about the correct direction must be.

Frank Plumpton Ramsey (1903-1930). Source: Wikepedia.

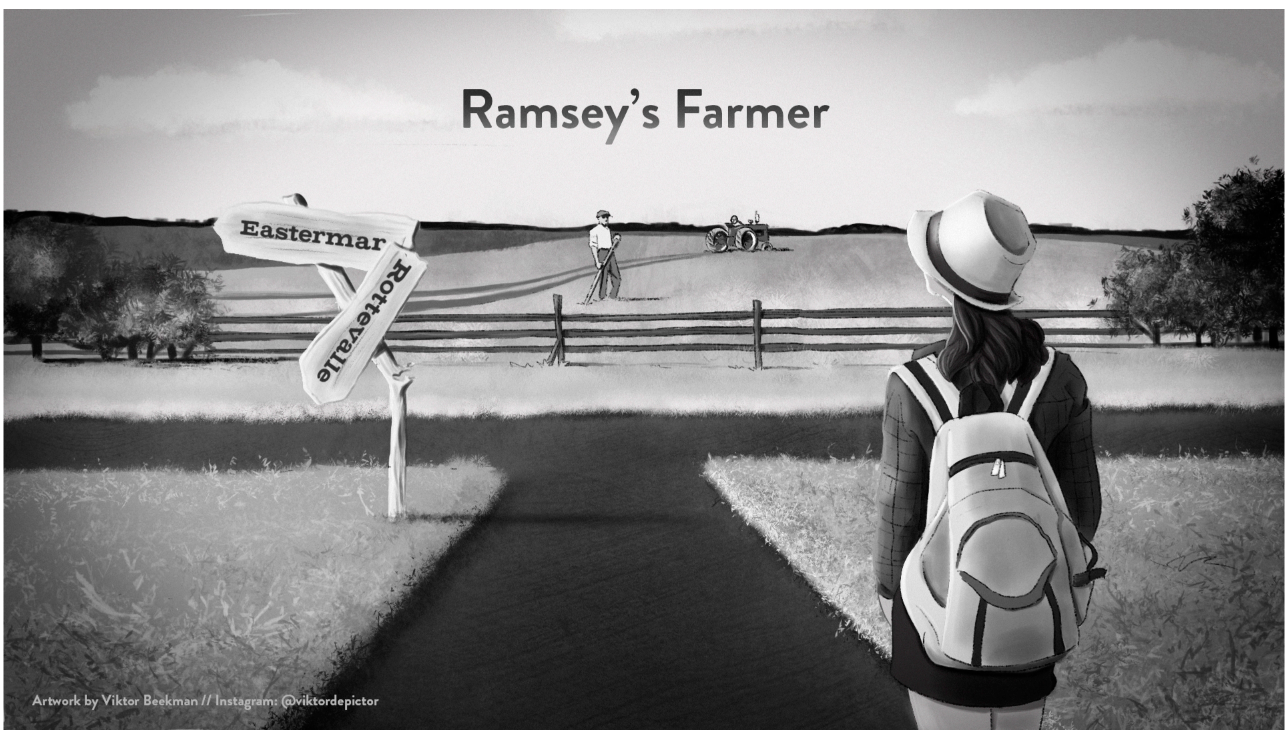


Figure 5.2: Ramsey's farmer. Harriet is not 100% certain about the direction of her hotel. Her degree of uncertainty can be measured by the distance she is just willing to walk in order to obtain the correct information from a friendly Frisian farmer. Figure available at `BayesianSpectacles.org` under a CC-BY license.

Of course, whenever it is useful to quantify uncertainty or elicit probabilities one does not always have easy access to a friendly Frisian farmer, let alone a friendly Frisian farmer who stands perpendicular to a T-section. Ramsey's point is that uncertainty can be quantified as the fair price for information that results in a certain outcome. When Harriet is already very confident that she needs to go right, the added information will be of little value to her, and so she is only willing to 'buy' that information when it is very cheap, that is, when the Frisian farmer is very close.

Exercises

1. Show why the distance to the Frisian farmer f is a measure of uncertainty p.

2. The analysis from the previous exercise implies that when you are perfectly uncertainty about the correct direction (i.e., $p = 1/2$) the distance to the farmer at the point of indifference equals $f = 1/2\, d$. Now imagine you arrive at the intersection in the late afternoon, and you'd like to be at the hotel in time for dinner. You can cover a distance of $2.5\, d$ before dinner service closes. Is $1/2\, d$ still a reasonable point of indifference? What does this say about the Frisian farmer scenario as a *pure* measure of uncertainty?

3. In what fundamental way does the Lindley-Borel-De Morgan setup differ from that of Ramsey?

Chapter Summary

This chapter discussed several ways in which degree of belief could be measured, at least in principle.

Want to Know More?

✓ Borel, E. (1965). *Elements of the Theory of Probability*. Englewood Cliffs, NJ: Prentice-Hall. The famous probabilist Borel appears to have been a staunch Bayesian. This is an English translation of the French original (first edition 1909).

> "There can be no doubt that probabilities, as they are known to us, are creations of the human mind. An omniscient being who knows all the mechanisms of the universe in all details would need no probabilities.[6] Probabilities exist in the human mind and they depend on, and are determined by, the body of knowledge K contained in the mind. This body of knowledge is not always exactly the same for two different minds, nor is it always the same even for one and the same mind at

[6] We shall leave aside all considerations concerning the modern theories of wave mechanics, according to which certain real phenomena can be defined only in terms of probabilities.

two different times. Thus, one should never speak of the probability of an event (say, a particular outcome of a roll of a pair of dice), but of the probability for Peter who rolls the dice, or for Paul who observes the throw, perhaps after having placed a bet." (Borel 1965, p. 165).

✓ Duke, A. (2018). *Thinking in Bets: Making Smarter Decisions When You Don't Have All the Facts.* New York: Portfolio/Penguin. Written by Annie 'The Duchess of Poker' Duke, this popular science book presents various insights on betting. The two-part review on `BayesianSpectacles.org` mentions the following eight:

- Every decision is a bet.
- We bet on our beliefs.
- What makes a decision good or bad is determined by the process, *not* by the final outcome.
- By articulating uncertainty as a bet we avoid black-and-white thinking, we become accountable for our beliefs, and it becomes easier to adjust our opinion.
- By embracing uncertainty we can learn more effectively and hence formulate more accurate beliefs that allow improved bets in the future.
- People are exceptionally poor at updating their beliefs, particularly because of hindsight bias and self-serving bias (and a host of other biases). It takes conscious effort to overcome these biases, but it's worth it.
- Our decision making is improved when we expose ourselves to a diversity of viewpoints rather than dwell in our own echo-chambers.
- Better decisions can be made when we imagine different future scenarios, their plausibilities, and their utilities.

✓ Misak, C. (2020). *Frank Ramsey: A Sheer Excess of Powers.* Oxford: Oxford University Press. A 500-page biography on the great Bayesian probabilist Frank Ramsey, who died at age 26 due to complications after having developed jaundice. Both Ramsey and Jeffreys were members of the Cambridge-based 'PsychAn' ($\psi\alpha$) discussion society on psychoanalysis (see also Strachey and Strachey 1986). On page 221, Misak writes: "But it was only now, through the Psych An Society, that they really got to know each other and discover a mutual interest in the philosophical foundations of induction and statistics." Surprisingly, this claim is false. Howie (2002, p. 117) writes: "though Jeffreys visited him in hospital during his illness, it was only after his death that Jeffreys discovered they had shared an interest in probability as well as psychoanalysis." And this is confirmed by

Jeffreys himself, in an unpublished interview with Dennis Lindley for the Royal Statistical Society on August 25, 1983: "I knew Frank Ramsey well and visited him in his last illness but somehow or other neither of us knew that the other was working on probability theory." ("Transcription of a Conversation between Sir Harold Jeffreys and Professor D.V. Lindley," Exhibit A25, St John's College Library, Papers of Sir Harold Jeffreys).

✓ Mosteller, F., & Youtz, C. (1990). Quantifying probabilistic expressions. *Statistical Science, 5*, 2–12. "Many people say that one cannot put a single number on a qualitative word. Actually one can put many numbers on a qualitative word, and that is one reason for pursuing such studies." (p. 3)

✓ Ramsey, F. P. (1926). Truth and probability. In Braithwaite, R. B. (Ed.), *The Foundations of Mathematics and Other Logical Essays*, pp. 156–198. London: Kegan Paul. One of the most famous essays in probability theory.

✓ Willems, S., Albers, C., & Smeets, I. (2020). Variability in the interpretation of probability phrases used in Dutch news articles – a risk for miscommunication. *Journal of Science Communication, 19*, A03. A Dutch replication of earlier results obtained in English.

6 *Coherence*

If one accepts, in its totality, the subjectivistic interpretation, probability theory constitutes the logic of uncertainty; this complements the logic of certainty and the two together form a unified and complete framework within which to conduct any argument. Those who reject this point of view find themselves without any coherent foundation on which to build.

de Finetti, 1974

Chapter Goal

Bayesians learn about the world in the same way that logicians draw conclusions using syllogisms (e.g., *modus ponens*: if all story-tellers are poor, and Kai Lung is a story-teller, then it follows that Kai Lung is poor). The difference is that in the Bayesian world, propositions are not only true or false, but have an in-between degree of plausibility. And, just like systems of pure logic, Bayesian reasoning ('the logic of partial beliefs') is governed by laws that make it impossible to draw conclusions that are silly, that is, internally inconsistent, contradictory, or *incoherent*. In this chapter we first discuss the importance of coherence and then discuss how the only way to avoid incoherence is to reallocate plausibility assignments using the laws of probability theory.

Against Contradictions

In their quest to better understand the world, researchers generally hate to end up with a contradiction. Contradictions suggest that, at an earlier stage in the reasoning process, something fundamental has gone off the rails. This visceral antipathy for contradictions is particularly pronounced for mathematicians and logicians.[1]

Contradictions in Mathematics

Mathematicians embrace contradictions only insofar as they reveal that a particular assumption must be false. Specifically, the method known as 'proof by contradiction' proceeds as follows[2]:

[1] For robots in the science fiction genre, a contradiction is often simply intolerable – as soon as the artificial intelligence realizes it faces a contradiction, it is just a matter of time before it turns insane or becomes catatonic (e.g., Asimov 1950).

[2] The example below is taken from `https://www.youtube.com/watch?v=jkhKPySIHgY`.

1. Consider a statement one wishes to prove, for instance, 'There are no positive integer solutions to the equation $x^2 - y^2 = 1$'.

2. Assume that the statement is false; that is, assume that there *do* exist positive integer solutions to the equation $x^2 - y^2 = 1$.

3. Demonstrate that assuming the statement to be false leads to nonsense, that is, it results in a contradiction. Rewrite $x^2 - y^2 = 1$ as $(x + y) \cdot (x - y) = 1$, and note that this is true for positive integers x and y only when $x + y = 1$ *and* $x - y = 1$. This in turn implies that $x = 1$ and $y = 0$; but y was supposed to be a positive integer, and this contradicts the solution that $y = 0$.

4. Having thus rejected the possibility that the statement is false, the only viable option is to assume the statement is correct.

One can even go a step further and argue that the absence of contradictions lies at the very heart of mathematics. The great French mathematician Henri Poincaré seems to have felt this way:

> "Mathematics is independent of the existence of material objects; in mathematics the word exist can have only one meaning, it means free from contradiction." (Poincaré 1913, p. 454)

and

> "Be not deceived. What is after all the fundamental theorem of geometry? It is that the assumptions of geometry imply no contradiction (...)." (Poincaré 1913, p. 467)

and finally

> "a definition is acceptable only on condition that it implies no contradiction." (Poincaré 1913, p. 468)

Contradictions in Logic

The tolerance for contradictions is hardly any higher among logicians. For ease of exposition, consider the logic of syllogisms, first outlined by Aristotle in his 350 BC book *Prior Analytics*.

Given two *premises* –statements assumed to be true with absolute certainty– we wish to draw a conclusion that is necessarily true. One valid rule of syllogistic reasoning is known as *modus ponens* ('affirming the antecedent'):

All story-tellers are poor
Kai Lung is a story-teller

Kai Lung is poor

"The general problem of deduction is as follows: —*From one or more propositions called premises to draw such other propositions as will necessarily be true when the premises are true.*" (Jevons 1874/1913, p. 59)

Another valid rule is known as *modus tollens* (‘denying the consequent’):

All story-tellers are poor
Kai Lung is not poor

Kai Lung is not a story-teller

Other such forms of valid logical reasoning exist and go under names such as *Barbara*, *Celarent*, *Darii*, *Ferio*, *Baralipton*, *Celantes*, *Dabitis*, *Fapesmo*, *Frisesomorum*, *Cesare*, *Cambestres*, *Festino*, *Barocho*, *Darapti*, *Felapto*, *Disamis*, *Datisi*, *Bocardo*, and *Ferison* – medieval mnemonics that were invented to make it easier for students to recall the different logical forms (for details see Lagerlund 2008).

There also exist *invalid* rules –logical fallacies– for drawing inferences from the premises. One beguiling logical fallacy is known as ‘affirming the consequent’:

All story-tellers are poor
Kai Lung is poor

Kai Lung is a story-teller [invalid!]

It is evident that this conclusion is not necessarily true, because Kai Lung could be poor for a different reason than being a story-teller; Kai Lung could be a beggar, or a businessman who has just gone bankrupt. Another fallacy is known as ‘denying the antecedent’:

All story-tellers are poor
Kai Lung is not a story-teller

Kai Lung is not poor [invalid!]

Again, the premises do not make the conclusion necessarily true – Kai Lung could be a poor cobbler.

Having introduced the basics of syllogistic logic, one may wonder what happens if the premises contain a *contradiction*. One may correctly anticipate that the method collapses; however, the nature and the totality of the collapse may elicit more surprise: the method collapses because a contradiction allows any statement whatever to be proven. This is known as the *principle of explosion* (i.e., *ex contradictione sequitur quodlibet*, ‘from a contradiction, anything follows’).

The disastrous effects of contradictions on logic and science were emphasized by Sir Karl Popper (1902-1994). For instance, in his book *Conjectures and Refutations* he elaborates:

Aristotle (384-322 BC), as painted in 1811 by Francesco Hayez (1791-1882). Public domain. “Aristotle has been called the most important thinker who has ever lived; he is recognized as the father of science, logic, biology, political science, zoology, embryology, natural law, scientific method, rhetoric, psychology, realism, criticism, individualism, teleology, meteorology and of all philosophers.” (`https://en.wikipedia.org/wiki/Aristotle`)

> "But this means that if we are prepared to put up with contradictions, criticism, and with it all intellectual progress, must come to an end. (...)
>
> For it can easily be shown that if one were to accept contradictions then one would have to give up any kind of scientific activity: it would mean a complete breakdown of science. This can be shown by proving that *if two contradictory statements are admitted, any statement whatever must be admitted*; for from a couple of contradictory statements any statement whatever can be validly inferred.
>
> This is not always realized,[6] and will therefore be fully explained here. It is one of the few facts of elementary logic which are not quite trivial, and deserve to be known and understood by every thinking man. It can easily be explained to those readers who do not dislike the use of symbols which look like mathematics; but even those who dislike such symbols should understand the matter easily if they are not too impatient, and prepared to devote a few minutes to this point." (Popper 1972, p. 317)

[6] See for example H. Jeffreys, 'The Nature of Mathematics', *Philosophy of Science*, **5**, 1938, 449, who writes: 'Whether a contradiction entails any proposition is doubtful.' See also Jeffreys' reply to me in *Mind*, **51**, 1942, p. 90, my rejoinder in *Mind*, **52**, 1943, pp. 47 ff., and *L.Sc.D.*, note *2 to section 23. All this was known, in effect, to Duns Scotus (*ob.* 1308), as has been shown by Jan Lukasiewicz in *Erkenntnis*, **5**, p. 124. [footnote in original – EWDM]

Popper then proceeds to give an example where two contradictory premises –'the sun is shining now' and 'the sun is not shining now'– allow the conclusion of the statement 'Caesar was a traitor'. The example is instructive, but a version that is simpler and shorter can be found on the Wikipedia entry for the *principle of explosion*:

> "As a demonstration of the principle, consider two contradictory statements– "All lemons are yellow" and "Not all lemons are yellow"–and suppose that both are true. If that is the case, anything can be proven, e.g., the assertion that "unicorns exist", by using the following argument:
>
> 1. We know that "Not all lemons are yellow", as it has been assumed to be true.
> 2. We know that "All lemons are yellow", as it has been assumed to be true.
> 3. Therefore, the two-part statement "All lemons are yellow *or* unicorns exist" must also be true, since the first part "All lemons are yellow" of the two-part statement is true (as this has been assumed).
> 4. However, since we know that "Not all lemons are yellow" (as this has been assumed), the first part is false, and hence the second part must be true to ensure the two-part statement to be true, i.e., unicorns exist."
>
> (Wikipedia, obtained from `https://en.wikipedia.org/wiki/Principle_of_explosion` on 19-09-2022)[3]

[3] Almost a millennium earlier, Duns Scotus gave yet another example, identical in structure to that provided by Popper and Wikipedia: "Socrates walks and Socrates does not walk, therefore you are in Rome" ("*Socrates currit et Socrates non currit; igitur tu es Romae*" – full quotation in Lukasiewicz 1935).

Popper then concludes:

> "We see from this that if a theory contains a contradiction, then it entails everything, and therefore, indeed, nothing. A theory which adds to every information which it asserts also the negation of this information can give us no information at all. A theory which involves a contradiction is therefore entirely useless *as a theory*." (Popper 1972, p. 319; see also Popper 1940)

When discussing the impact of contradictions, Sir Ronald Fisher illustrated the problem with the following anecdote[4]:

[4] The anecdote is repeated in Jeffreys 1973, p. 18, who was convinced by Popper that a contradiction implies any proposition (see also Jeffreys 1961, pp. 34-35).

> "There is a story that emanates from the high table at Trinity that is instructive in this regard. G. H. Hardy, the pure mathematician—to whom I owe all that I know of pure mathematics—remarked on this remarkable fact, and someone took him up from across the table and said, "Do you mean, Hardy, if I said that two and two make five that you could prove any other proposition you like?" Hardy said, "Yes, I think so." "Well, then, prove that McTaggart[5] is the Pope." "Well," said Hardy, "if two and two make five, then five is equal to four. If you subtract three, you will find that two is equal to one. McTaggart and the Pope are two; therefore, McTaggart and the Pope are one." (Fisher 1958, p. 269)

[5] John McTaggart (1866-1925) was a lecturer in philosophy at Trinity College, Cambridge – EWDM.

In sum, contradictory premises utterly destroy the kind of deductive logic that underlies syllogistic reasoning. But what is the nature and impact of contradictions if our premises are uncertain, and we wish to learn from noisy data?

The Logic of Partial Beliefs

> The idea of a reasonable degree of belief intermediate between proof and disproof is fundamental. It is an extension of ordinary logic, which deals only with the extreme cases.
>
> Jeffreys, 1955

As indicated by the epigraph to this section, Bayesian inference is a generalization of pure logic[6]; in this generalization, the premises can be probabilistic rather than true with absolute certainty. For example, here is a Bayesian version of the *modus ponens*:

[6] This was also stressed by arch-Bayesians such Ramsey, de Finetti, and Jaynes.

If you were to learn that Kai Lung is a story-teller, the probability that he is poor equals $.60$; if you were to learn that Kai Lung is not a story-teller, the probability that he is poor equals $.30$.
You see Kai Lung walk into the town square and unroll his mat; this behavior is characteristic of story-tellers and consequently you assign a probability of $.80$ to the proposition that Kai Lung is a story-teller

---

The probability that Kai Lung is poor is $(.80 \times .60) + (.20 \times .30) = .54$

The premises now involve probabilistic statements, and the conclusion results from applying the law of total probability. The practical relevance of this style of reasoning –contra that of syllogistic logic– is immediately evident:

> "They say that Understanding ought to work by the rules of right reason. These rules are, or ought to be, contained in Logic; but the actual science of Logic is conversant at present only with things either certain, impossible, or *entirely* doubtful, none of which (fortunately) we have to

> reason on. Therefore the true Logic for this world is the Calculus of Probabilities, which takes account of the magnitude of the probability (which is, or which ought to be in a reasonable man's mind). This branch of Math., which is generally thought to favour gambling, dicing, and wagering, and therefore highly immoral, is the only "Mathematics for Practical Men," as we ought to be." (James Clerk Maxwell, in a 1850 letter to Lewis Campbell; reproduced in Campbell and Garnett 1882, p. 80)

Corroborating the Consequent

The introduction of probabilities and uncertainty also opens the door to *learning from experience*, as incoming information may continually change the relevant probabilities. Hence, instead of conducting a purely deductive analysis we now find ourselves involved in induction. And this means that a logical pitfall is transformed to an inductive principle.[7]

[7] The fragment that follows is based in part on the `BayesianSpectacles.org` blog post "Is Polya's fundamental principle fundamentally flawed?"

As discussed above, a famous fallacy in deductive logic is "affirming the consequent". Another example of a syllogism gone wrong:

When Socrates rises early in the morning, he always has a foul mood
Socrates has a foul mood

Socrates has risen early in the morning [invalid!]

The deduction is invalid because Socrates may also be in a foul mood at other times of the day. What the fallacy does is take the general statement "A $\rightarrow$ B" (A implies B; rising in the morning $\rightarrow$ foul mood), and interpret it as "B $\rightarrow$ A" (B implies A; foul mood $\rightarrow$ rising in the morning).

When we switch from deductive reasoning to inductive learning, however, the fallacy of "affirming the consequent" is transformed to a law, one that might be called "corroborating the consequent". In two brilliant books, the mathematician George Pólya (1887-1985) describes in detail how inductive reasoning is important in mathematics, a field that most people would believe is governed solely by deductive processes and rigorous proof. As Pólya states in a lecture that is available on YouTube[8]: "first guess, then prove". Actually, in his books Pólya proposes that the process by which mathematicians work is slightly more complicated: first guess, then corroborate the guess with examples, then prove. Here we focus on what Pólya called "the fundamental inductive pattern":

[8] `https://www.youtube.com/watch?v=h0gbw-Ur_do`

> There is no demonstrative conclusion: the verification of its consequence B does not prove the conjecture A. Yet such verification renders A more credible. (...) "We have here a pattern of *plausible inference*:
>
> A implies B
> B true

A more credible

The horizontal line again stands for 'therefore.' We shall call this pattern the *fundamental inductive pattern*, or, somewhat shorter, the 'inductive pattern'.

This inductive pattern says nothing surprising. On the contrary, it expresses a belief which no reasonable person seems to doubt: *The verification of a consequence renders a conjecture more credible.* With a little attention, we can observe countless reasonings in everyday life, in the law courts, in science, etc., which appear to confirm to our pattern." (Pólya 1954b, pp. 4–5)

Thus, in the Socrates example we only need to make a small change to go from deductive fallacy to inductive law:

When Socrates rises early in the morning, he always has a foul mood
Socrates has a foul mood

It has now become more credible than before that Socrates has risen early in the morning

This example actually suggests that Pólya's definition has a flaw. When the consequent is predictively irrelevant, the credibility of the conjecture ought to remain unaffected. For instance, suppose we know that Socrates was perpetually in a foul mood, irrespective of the time of day; this invalidates the inference above. To drive the point home, here is another example:

On Mondays, trains from Hilversum to Amsterdam run every 15 minutes
Today, trains from Hilversum to Amsterdam run every 15 minutes

---

It has now become more credible than before that today is a Monday

But what if I tell you that trains from Hilversum to Amsterdam run every 15 minutes *every day of the week*? It becomes clear that the alternative hypotheses (days of the week) also imply the consequent, and the consequent is therefore predictively irrelevant, and the credibility of the proposition is left unchanged.[9]

[9] In the second appendix of Chapter 9 we will take a stronger stance, and argue that Pólya's definition is untenable except in the complete absence of background knowledge (cf. Good 1967, Rosenkrantz 1982). The verification of a consequence may even render a conjecture *less* credible! We were shocked when we first learned about this.

Garbage in, Garbage out

Another similarity to deductive reasoning is that in Bayesian inference, the conclusion is only as good as its premises. In other words, Bayesian inference does not tell you how to define your prior knowledge; instead, Bayesian inference tells you how to update beliefs from a given starting point of background knowledge. Just as in pure logic and deductive reasoning, faulty Bayesian premises may yield faulty Bayesian conclusions, in line with the popular adage *garbage in, garbage out*.[10] Bruno de Finetti expressed the sentiment more eloquently:

> "*The calculus of probability can say absolutely nothing about reality; in the same way as reality, and all sciences concerned with it, can say nothing about the calculus of probability.* The latter is valid whatever use one makes of it, no matter how, no matter where. One can express in terms of it any opinion whatsoever, no matter how 'reasonable' or otherwise, and the consequences will be reasonable, or not, for me, for You, or anyone, according to the reasonableness of the original opinions of the individual using the calculus. As with the logic of certainty, the logic of the probable adds nothing of its own: it merely helps one to see the implications contained in what has gone before (either in terms of having accepted certain facts, or having evaluated degrees of belief in them, respectively)." (de Finetti 1974, p. 182)

[10] An example from syllogistic logic: the premises 'all birds can fly' and 'penguins are birds' leads to the conclusion 'penguins can fly'. The reasoning itself is valid, but because one of the premises is false, the conclusion is also false.

Coherence

(...) the most generally accepted parts of logic, namely, formal logic, mathematics and the calculus of probabilities, are all concerned simply to ensure that our beliefs are not self-contradictory.

Ramsey, 1926

The theory must be self-consistent; that is, it must not be possible to derive contradictory conclusions from the postulates and any given set of observational data.

Jeffreys, 1939

Coherence acts like geometry in the measurement of distance; it forces several measurements to obey the system.

Lindley, 2000

Finally we arrive at the heart of the matter. We have seen that Bayesian inference –the calculus of probability– "can say absolutely nothing about reality". But what then typifies Bayesian inference? Ultimately, it comes down to a single concept: *coherence*.

"The Bayesian theory is about *coherence*, not about right or wrong". (Lindley 1976, p. 359)

In Chapter 2 we mentioned that for a Bayesian, the word 'probability' is synonymous with 'reasonable degree of belief'. This suggests that if we assign degrees of belief to different propositions, we have to obey the rules of probability theory – if these laws are violated, our beliefs are mutually inconsistent or nonsensical. Thus:

> "[The rules of probability] proscribe constraints on your beliefs. While you are free to assign any probability to the truth of the event, once this has been done, you are forced to assign one minus that probability to the truth of the complementary event. If your probability for rain tomorrow is 0.3, then your probability for no rain must be 0.7." (Lindley 2006, p. 40)

Another perspective is that the laws of probability theory protect us from incoherence. These laws dictate that when (a) we learn that Kai Lung is a story-teller, the probability that he is poor equals .60 (whereas it would have equaled .30 if Kai Lung is not a story-teller); and when (b) you see Kai Lung walk into the town square and unroll his mat; this behavior is characteristic of story-tellers and consequently you assign a probability of .80 to the proposition that Kai Lung is a story-teller; then it has to follow that the probability that Kai Lung is poor is $(.80 \times .60) + (.20 \times .30) = .54$. Any other assessment would be incoherent.

It is immediately clear that people are in dire need of the protection that the laws of probability theory provide. Unaided by probability theory, people will find it impossible to specify coherent degrees of beliefs across many propositions of varying complexity. The notion of coherence is therefore *prescriptive*, not descriptive:

> "(...) a formal and consistent theory of inductive processes cannot represent the operation of every human mind in detail; it will represent an ideal mind, but it will also help the actual mind to approximate to that ideal." (Jeffreys 1961, p. 421)

Coherence therefore constrains the assignment of degrees of belief; this holds across related propositions but, crucially, coherence also exerts complete control over how beliefs are updated as additional information becomes available. Let's revisit the example in Chapter 3 on the base rate fallacy. In this example, the prior odds was 999:1 of a driver being sober rather than drunk; a positive breathalyzer test outcome (i.e., the incoming data) is 20 times more likely when the driver is drunk than when they are sober; consequently, the posterior odds for the driver being sober *has to be* $^{999}/_{20} = 49.95$.

In other words, once our prior knowledge has been specified, confrontation with the data will cause a unique, coherent update to posterior knowledge. An apt metaphor is to the laws of geometry, as illustrated by the triangle shown in Figure 6.1. The adjacent side symbolizes the prior knowledge, and the opposite side symbolizes the observed data; with these two sides in place, the location of the hypotenuse (i.e., the posterior knowledge) is defined uniquely.

This implies that if the posterior knowledge is deemed unpalatable or implausible, the fault lies either with our intuition, or with the data (these may have been recorded or reported incorrectly), or with the prior knowledge – the fault most definitely does *not* lie with the updat-

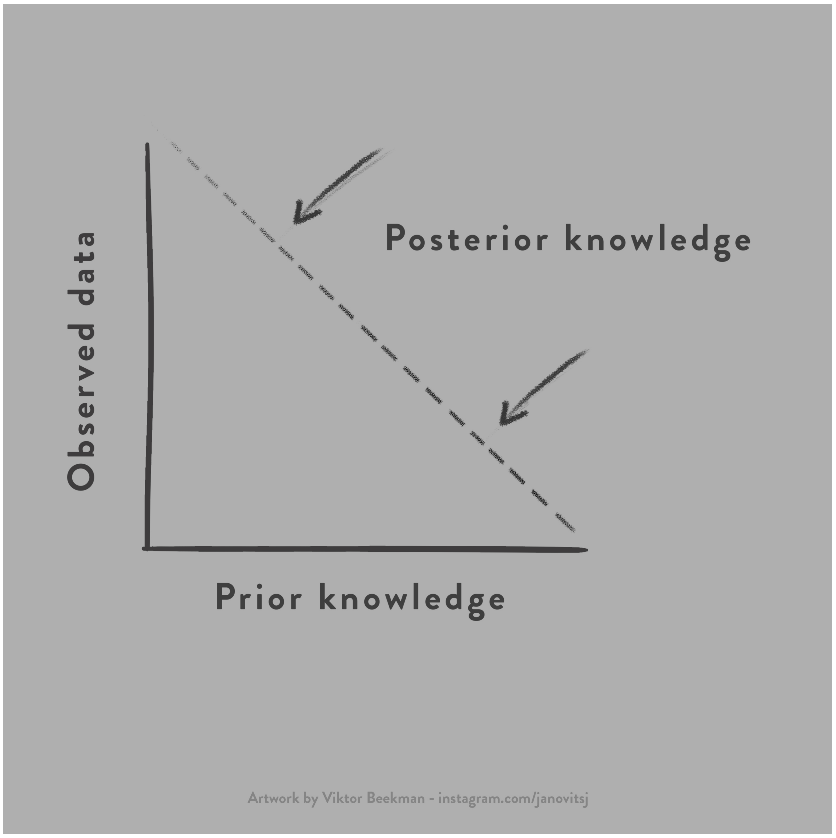


Figure 6.1: With prior knowledge fully specified, incoming data trigger a learning process that results in uniquely defined posterior knowledge, courtesy of Bayes' theorem. "This theorem is to the theory of probability what Pythagoras's theorem is to geometry." (Jeffreys 1931, p. 19). Figure available at `BayesianSpectacles.org` under a CC-BY license.

ing process, which is a mathematical operation to ensure that posterior beliefs cohere with prior beliefs. Imagine a perfect chef who creates the best possible dish (tailored to your tastes) given the available ingredients. If you nevertheless strongly dislike the dish, this can only mean that the ingredients were poor, and it is inappropriate to critique the chef.

To elaborate on this important point, assume one wishes to estimate the proportion θ of first-year psychology students who prefer cats to dogs. We are getting ahead of ourselves, but the standard Bayesian analysis assumes that every value of θ from 0 to 1 is equally likely *a priori*. Suppose the first student we ask indicates that they prefer cats to dogs; an application of the rules of probability theory then transform the prior beliefs about θ to posterior beliefs. Examining these posterior beliefs reveals that the single most likely value of θ equals 1, which corresponds to the assertion that *all* first-year psychology students prefer cats to dogs. If this conclusion appears unreasonable, it signals a problem with the specification of the prior distribution. When sufficient thought is given to the problem, one may discover that it is actually unreasonable to deem every value of θ equally likely *a priori*.

You may remain unconvinced. It may seem unappealing that your beliefs should find themselves shackled and constrained to particular values. Indeed, you could adopt the philosophy of Feyerabend, embrace *epistemological anarchism*, and provocatively state that with respect to your beliefs, "anything goes". What then is the downside of incoherence? First and foremost, we should not forget that 'incoherence' is just a fancy word for 'nonsensical'. For instance, we may assume that the order in which the data come in is irrelevant, but then obtain a different conclusion depending on whether the data are analyzed all at once, batch-by-batch, or one at a time.[11] Hence, incoherence is intellectually disturbing and suggests a hidden flaw in one's reasoning. Second, as mentioned before, coherence is the axiomatic basis for a rational system of learning from experience. "Anything goes" does not provide a firm foundation for any theory, let alone a theory that eliminates all reasoning that is internally inconsistent. The case for coherence can be made in many ways (e.g., Cox 1946, Jaynes 2003, Joyce 1998, Jeffreys 1961; see also Diaconis and Skyrms 2018) but here we pursue a line of attack that is due to de Finetti: if you, as an epistemological anarchist, were forced to act on those incoherent beliefs, your actions would allow a malevolent third party to exploit you with impunity. In other words, acting on incoherent beliefs leads to a sure loss. The next section provides a concrete example.

[11] In contrast, coherent Bayesian inference always draws the same conclusion: "It is self-consistent in the sense that the final probabilities of a set of hypotheses are the same in whatever order the data are taken into account." (Jeffreys 1938d, p. 444; see also Jeffreys 1938a, pp. 191-192)

De Finetti's Bet Revisited

In order to clarify the importance of coherence, Bruno de Finetti proposed a scenario involving betting. The scenario shows that degrees of belief need to be governed by the rules of probability theory. If these rules are flaunted, the beliefs are incoherent, and a third party can exploit this incoherence to obtain a guaranteed profit.

Consider then a ticket that pays \$1 if a particular proposition holds true. Ticket I presents the proposition "At the next summer Olympics, the gold medalist for the women's marathon will have the Kenian nationality". How much money do you believe Ticket I is worth? To ensure that your assessment is fair, we agree that I will have the choice either to buy the ticket from you *or* sell the ticket to you, for the price that you have determined.[12] Let's assume that you believe a fair price is \$0.40. Note that this assessment depends on your knowledge of marathon runners; a person who knows more (or less) about this discipline may set a different price.

[12] De Finetti's scenario was already introduced in Chapter 5.

We continue and examine Ticket II. This ticket presents the proposition "At the next summer Olympics, the gold medalist for the women's

Anything Goes, Except for Incoherence?

In his deliberately provocative book *Against Method*, Austrian-born philosopher Paul Feyerabend (1924-1994) advocated what he termed *epistemological anarchism*:

> "Science is an essentially anarchic enterprise (...) The only principle that does not inhibit progress is: *anything goes*."(Feyerabend 1993, p. 5; first edition 1975)

Militant subjective Bayesians would broadly agree but insist on coherence as a crucial addendum. Hence their amended rule would be: *anything goes, except for incoherence*. Below one of the most militant of subjective Bayesians underscores the point:

> "There are some probabilities that are almost universally accepted. For example, if A includes extensive knowledge about a coin and θ is the event that it falls heads when reasonably tossed, then it would be an unusual person who came up with $p(\theta \mid A)$ anything other than $1/2$. But if John insists that $p(\theta \mid A) = 1/3$ who is to say he is wrong? He will be wrong if he fails to react to data on tosses of the coin by using Bayes' theorem (...) but I can see no sense in which his original curious value is wrong. The only way he can be wrong is in not being coherent." (Lindley 1985, p. 192)

marathon will have the Ethiopian nationality". What is the fair price for this ticket? For the sake of the argument, suppose you set the price to \$0.75. This would be *incoherent* – your evaluation does not respect the laws of probability theory and therefore you can be made a sure loser. In particular, I notice that you have overpriced the tickets – the sum of the prices is \$1.15, more than the amount that can be won. Consequently, I will sell both tickets to you and gain \$1.15, whereas you are left with only a chance to win \$1. You do not fall into this trap, however, and instead you set a price for Ticket II that equals \$0.30.

Now consider Ticket III. This ticket presents the proposition "At the next summer Olympics, the gold medalist for the women's marathon will have *either* the Kenian nationality *or* the Ethiopian nationality". How much is this ticket worth? Coherence allows only one answer: \$0.70. Set any other price and the resulting incoherence allows you to be made a sure loser. For instance, suppose you incoherently set the price of Ticket III to \$0.60. This is cheaper than \$0.70, and so I will buy Ticket III from you and sell Tickets I and II to you; this gives me a \$0.10 pure profit, as our chances to win the \$1 are identical. Alternatively, suppose you incoherently set the price of Ticket III to \$0.80. This is more expensive than \$0.70, and so I will sell Ticket III to you and buy Tickets I and II from you, earning a pure profit of \$0.10 – again, our chances to win the $1 are identical. In both example cases, the incoherence revealed by Ticket III led you to lose $0.10 without the slightest compensation.

The only way to avoid a sure loss is to price Ticket III as $0.40 + $0.30 = $0.70. Note that by assigning beliefs so as to avoid a certain loss, we have in fact reproduced one of the defining rules of probability theory: For mutually exclusive events, probability adds. The other rules of probability theory may be obtained from de Finetti's betting scenatio in similar fashion (e.g., Diaconis and Skyrms 2018, pp. 22-33).

Figure available at `BayesianSpectacles.org` under a CC-BY license.

Rebuttal of the Common Critique on Betting

Some philosophers would sooner participate in a season of *Temptation Island*[13] than admit that Bayesian inference has practical or theoretical merit. This is one of life's great mysteries, as philosophers should be especially keen to embrace a methodology that, by its very construction, weeds out opinions and convictions that are inherently inconsistent.

At any rate, when detractors of the Bayesian gospel are presented with de Finetti's betting scenario, their knee-jerk response is to argue

[13] "Temptation Island is an American reality dating show, in which several couples agree to live with a group of singles of the opposite sex, in order to test the strength of their relationships." (Wikepia, `https://en.wikipedia.org/wiki/Temptation_Island_(TV_series`, consulted 21-09-2022)

that people rarely bet on their beliefs, and that betting introduces complications to do with the utility of money, loss aversion, etc. Hence, the betting scenario is judged to be irrelevant. We believe such a critique is superficial at best and purely rhetorical at worst.

In order to disarm the critique, it should first be stressed again that coherence is prescriptive, not descriptive: it is a framework for how rational agents *ought* to reason under uncertainty, not how people actually fumble about in practice, unaided by probability theory and depending solely on intuition.

Secondly, no actual betting with monetary stakes needs to take place:

> "Aiming for coherence has its roots in a desire for consistency. It applies to logic as well. One of the wisest men we know put it this way: "We all believe inconsistent things. The purpose of rational discussion aims at this: If someone says 'You accept A and B, but by a chain of reasoning, each step of which you accept, it can be shown that A implies not B,' you would think that something is wrong and want to correct it."
>
> It is similar with judgments of uncertainty. Of course, there is no bookie, and no one is betting. Still coherence, like consistency, seems like a worthwhile standard." (Diaconis and Skyrms 2018, pp. 25-26)

Third, the betting scenario is merely a demonstration of the misfortunes that befall anybody who is prepared to act on a set of incoherent beliefs. Finally, even though one may object that people rarely bet on their beliefs, there is an argument to be made that people bet on their beliefs all the time, except not with money:

> "Objections have been raised because the standard involves gambling and some people object to gambling. The confusion here is due to inadequacies in the English language (or in my use of it). We are all faced with uncertain events like 'rain tomorrow' and have to act in the reality of that uncertainty—shall we arrange for a picnic? We do not ordinarily refer to these as gambles but what word can we use? In this sense all of us 'gamble' every day of our lives, and the word is used in this sense. The gambles that people object to are unnecessary gambles on horses, or sport, or cards, usually conducted for monetary gain or excitement. The prize in our case need not be awarded: it is only contemplated. The essential concept is *action* in the face of uncertainty." (Lindley 1985, p. 19)

and

> "Some statisticians have protested that to base opinions on betting is to reduce statistics to the level of a racecourse. However, in a sense any decision in life is a kind of generalized bet. If we go out for a walk without a raincoat, this is a bet with nature that it will be fine. If it is, we have the reward of unencumbered movement; if it rains, we pay the penalty of the discomfort of being soaked or having to take shelter" (Smith 1965, p. 477)

and

> "(...) all our lives we are in a sense betting. Whenever we go to the station we are betting that a train will really run, and if we had not a sufficient degree of belief in this we should decline the bet and stay at home. The options God gives us are always conditional on our guessing whether a certain proposition is true." (Ramsey 1926 as given in Eagle (Ed.) 2011, p. 62)

Closing Remarks

When asked about the benefits of Bayesian inference, few practitioners and theoreticians will mention coherence. This is not because coherence is somehow unimportant – paradoxically, it is exactly because coherence is so important that it does not get mentioned: coherence is automatically achieved whenever prior opinions are updated by the data using Bayes' rule, so Bayesians generally need not worry about it.[14] In this way, coherence is akin to good health; it is usually enjoyed without much thought. Only when it breaks down does it suddenly become apparent that it was in fact crucial all along.

[14] Some Bayesians occasionally use prior knowledge that is informed by the observed data (for examples see Consonni et al. 2018); strictly speaking this practice is incoherent, but the degree of incoherence may be relatively mild.

Coherence is the bedrock of rationality. In a way, it is a minimum requirement for reasoning under uncertainty. Through the laws of probability theory, coherence restricts the beliefs that one can entertain. This is limiting only to the degree that one desires the freedom to be silly. Coherence is rather like a crutch that supports people when they draw inferences from uncertain events. Epistemological anarchists may throw away the crutch of coherence and cry "freedom!", but they will immediately find themselves falling to the floor, unable to make further progress.

One final thought. In real life people are not coherent, and yet most of us get by without our incoherence being ruthlessly exposed and exploited. We suspect that when people operate in the real world, their actions are shaped through continual feedback with the environment[15]: adaptive behavior is rewarded, and inopportune behavior is punished. For some tasks, this results in acceptable performance. When a cognitively limited agent operates under considerable time pressure in a highly complex environment, it may just be a waste of resources and opportunity to strive for perfect coherence. We end with a quotation from the hero of this book:

[15] This learning process takes place at multiple time scales, including the time scale of human evolution.

> "The theory of probability is a formal statement of common-sense. Its excuse for existence is that it gives rules for consistency. It does not try to justify common-sense nor to alter its general practice; it recognizes that the human mind is a useful tool, but that, like other tools, it is not necessarily perfect." (Jeffreys 1936a, p. 337)

Exercises

1. Consider the box "Anything goes, except for incoherence". Lindley argues that someone with peculiar prior beliefs cannot be judged to be wrong. Argue against this view.
2. Explain why it is incoherent to inform prior knowledge by the observed data.

Coherence as a Jigsaw Puzzle

Consider again the simplest example of incoherence: the probability of an event happening is judged to be x, and the probability of that event not happening is judged to be different from $1 - x$. For instance, you may believe that the probability of rain tomorrow in the Atacama Desert is 0.98; given that the Atacama Desert is one of the world's driest places, this is certainly a remarkable belief – but it is not yet incoherent. It only becomes incoherent if you also believe, at the same time, that the probability of it not raining in the the Atacama Desert tomorrow is 4%, 1%, 50%, or really anything different from $100 - 98 = 2\%$.

In this example, our propositions may be likened to a jigsaw puzzle with only two pieces: 'rain' and 'not rain'. When the puzzle pieces fit together, they belong to the same puzzle. When one puzzle piece ('rain') is 98% and the other is, say, 4%, this means that the pieces originate from different puzzles – they are beliefs that may legitimately be entertained, but not simultaneously by the same agent.

Puzzles of just two pieces are easy and few people will hold incoherent beliefs in such cases. But in real life as well as in statistics, the puzzles quickly grow in complexity as new information is added. Some puzzles may have hundreds of pieces, or even infinitely many. Very quickly, it becomes a daunting task to check whether or not the pieces form a single coherent puzzle. And this is perhaps the single outstanding benefit of Bayes' rule: it ensures that initially simple sets of coherent beliefs *remain* coherent when they are updated or sharpened as more information (such as new data or additional background information) becomes available.

Chapter Summary

In syllogistic logic, contradictions allow any statement whatever to be proven. Bayesian inference is the logic of partial beliefs, that is, the coherent way of reasoning in an uncertain world. The Bayesian equivalent of a contradiction is termed an incoherence. In order to reason in a coherent fashion (i.e., remain free from internal inconsistencies) it is required that our beliefs obey the laws of probability. Those who are prepared to act on a set of incoherent beliefs can be exploited with impunity by a malevolent third party. Coherence is the bedrock of rationality; Bayesians rarely ponder the wonders of coherence because Bayes' theorem has coherence built in.

Want to Know More?

✓ Chapter 27 demonstrates the role of coherence in Bayesian evidence updating.

✓ Diaconis, P., & Skyrms, B. (2018). *Ten Great Ideas About Chance*. Princeton: Princeton University Press. Chapter 2, 'Judgment' provides a good discussion of the different aspects of coherence.

✓ Eagle (Ed.), A. (2011). *Philosophy of Probability: Contemporary Readings*. New York: Routledge. A collection of key readings in the philosophy of probability theory. Requires some background in mathematics for its proper appreciation. Our quotations of Ramsey (1926) were taken from this source. The collection also contains an article by Joyce, who proved that "any system of degrees of belief that violates the axioms of probability can be replaced by an alternative system that obeys the axioms and yet is more accurate in every possible world" (Joyce 1998, as given in Eagle (Ed.) 2011, p. 89)

✓ Lindley, D. V. (2000). The philosophy of statistics. *The Statistician*, 49, 293-337. Throughout his work, Lindley hammered home the importance of coherence, up to the point where he proposed to replace the term 'Bayesian statistics' with 'coherent statistics' (Lindley 1985). 'The philosophy of statistics' is one of Lindley's best articles. A background in statistics is recommended.

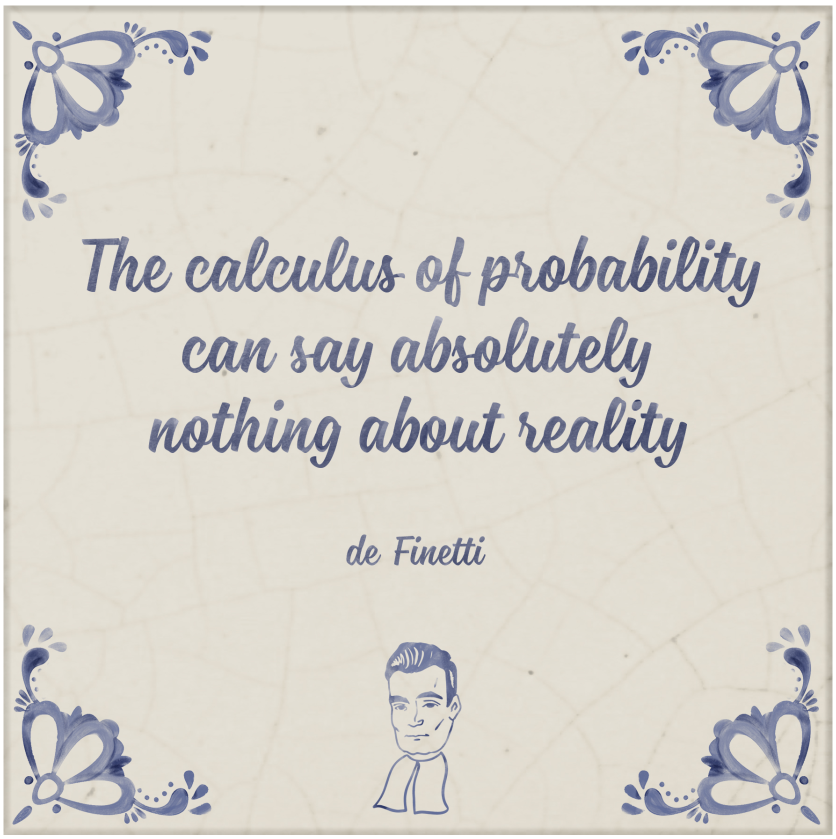


Figure available at `BayesianSpectacles.org` under a CC-BY license.

Part II

Coherent Learning, Laplace Style

7 Learning from the Likelihood Ratio [with Alexandra Sarafoglou and František Bartoš]

The theory comes into play where ignorance begins, and the knowledge we possess requires to be distributed over many cases.

Jevons, 1874

"(...) if you can't do simple problems, how can you do complicated ones?" Lindley (1985, p. 65)

Chapter Goal

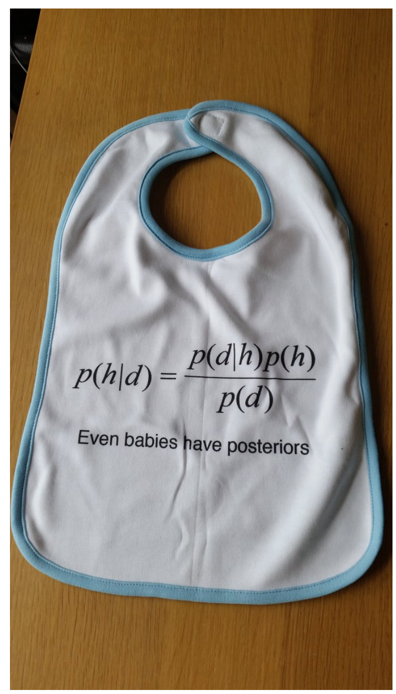
This chapter showcases each of the separate elements of the Bayesian learning cycle in its simplest form. The guiding example has the minimum uncertainty required to get the Bayesian ball rolling.

Two Pressing Questions about Pancakes

Miruna comes home and discovers that it's Dutch pancakes for dinner. Hurray! She knows the pancakes were baked by either of her parents, Andy and Bobbie, but she does not know which one. The only clue as to the identity of the baker is provided by the composition of the pancakes: Andy has a probability of producing a bacon pancake of $\theta_A = 0.40$, whereas that probability is $\theta_B = 0.80$ for Bobbie. We assume that all non-bacon pancakes are plain, that is 'vanilla' type pancakes. We also assume that the stack is produced randomly, that is, any order is as likely as any other.[1]

This is a simple scenario. There are only two candidate bakers, only two types of pancakes, and the probability of Andy and Bobbie producing a bacon pancake (their 'bacon proclivity') is constant over time and known exactly. We can relax these assumptions and consider more realistic scenarios, but for now we keep things simple. Consider two fundamentally different questions:

- After inspection of the pancake stack, what can we say about the probable identity of the baker? Desired here is an *inference about an unobserved cause* or latent data-generating process.



Bayes' rule on a bib. Here d stands for 'data' and h for 'hypothesis'. In the current chapter we will limit ourselves to two hypotheses: did Andy or Bobbie bake the pancakes?

[1] In Bayesian lingo, the pancakes are said to be 'exchangeable' (de Finetti 1974, Zabell 1982).

- After inspection of the pancake stack, what is the probability that the next pancake will have bacon? Desired here is a *prediction about a to-be-observed consequence* or future datum.

We will now address these questions in turn.

A stack of Dutch pancakes, with a bacon pancake on top.

Question 1: Who Baked the Pancakes?

In our example, there are two rival hypotheses, that is, two candidate causes for the pancake stack: either Andy or Bobbie is the baker. Before we can start our Bayesian analysis, we need to specify our prior knowledge: the relative plausibility of the rival hypotheses, reflecting our uncertainty about who baked the pancakes. In this case, Miruna has no information that suggests that either Andy or Bobbie is the baker, and she therefore believes both hypotheses are equally credible *a priori* – hence, $p(\theta_A) = p(\theta_B) = 1/2$; equivalently, we can say that the prior odds is 1: $p(\theta_A)/p(\theta_B) = 1$. Miruna's lack of information concerning the identity of the baker is illustrated in Figure 7.1.

We abuse notation and denote p(Andy is the baker and therefore $\theta = \theta_A$) by $p(\theta_A)$.

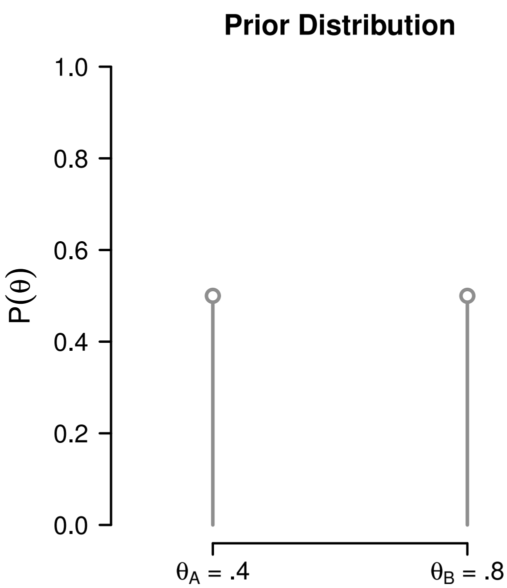


Figure 7.1: Before having seen any of the pancakes, Miruna believes that Andy and Bobbie are equally likely to have baked the stack. This uncertainty is reflected in a prior distribution that assigns Andy and Bobbie equal mass.

We pause here and reflect on a momentous occasion. What you see in Figure 7.1 is a *prior distribution*, the first of many in this book. Note that the distributions you usually encounter are distributions of something you can observe directly, such as height or income. Figure 7.1, however, shows a distribution of something more ephemeral: a *distribution of belief*, expressing the relative plausibility of the different values for the bacon proclivity θ. This prior distribution is very simple, as our belief is distributed across just two discrete values ('atoms'), θ_A and θ_B. Let's see how this distribution is updated as we observe data.

Datum 1: A Bacon Pancake

Now Miruna observes the first pancake and notices that it has *bacon*, an event that we denote as $\{b\}$. This observation has to shift her conviction in the direction of Bobbie being the baker; after all, the probability of a bacon pancake is higher for Bobbie than it is for Andy. To compute how much this information should shift her belief we use Bayes' rule. Here we will apply both the probability form and the odds form (cf. Chapter 3). First, the probability form of Bayes' rule:

$$\begin{aligned} p(\theta_B \mid \{b\}) &= p(\theta_B) \cdot \frac{p(\{b\} \mid \theta_B)}{p(\{b\} \mid \theta_A)\, p(\theta_A) + p(\{b\} \mid \theta_B)\, p(\theta_B)} \\ &= {}^1\!/_2 \cdot \frac{{}^8\!/_{10}}{{}^4\!/_{10} \cdot {}^1\!/_2 + {}^8\!/_{10} \cdot {}^1\!/_2} = {}^2\!/_3. \end{aligned}$$

Second, we can apply the odds form and obtain the same result:

$$\begin{aligned} \overbrace{\frac{p(\theta_B \mid \{b\})}{p(\theta_A \mid \{b\})}}^{\text{Posterior odds}} &= \overbrace{\frac{p(\theta_B)}{p(\theta_A)}}^{\text{Prior odds}} \times \overbrace{\frac{p(\{b\} \mid \theta_B)}{p(\{b\} \mid \theta_A)}}^{\text{Evidence}} \\ &= 1 \times \frac{{}^8\!/_{10}}{{}^4\!/_{10}} = 2. \end{aligned}$$

The 'evidence' term is the extent to which the data mandate a change from prior to posterior odds. Here our rival hypotheses are specified without uncertainty – we know that Andy has θ_A exactly equal to $.40$, and that Bobbie has θ_B exactly equal to $.80$; in such a scenario, the evidence is also known as the *likelihood ratio* (e.g., Royall 1997).[2] The evidence term tells us that the data (i.e., a bacon pancake) are twice as likely under the hypothesis that Bobbie is the baker than they are under the hypothesis that Andy is the baker; that is, the data are twice as surprising under the hypothesis that Andy is the baker than under the hypothesis that Bobbie is the baker. In other words, the Bobbie-is-the-baker hypothesis predicted the data twice as well as the Andy-is-the-baker hypothesis. With a prior odds equal to 1, this means that Miruna should now believe that it is twice as likely that Bobbie is the baker than that Andy is the baker. As explained in Chapter 3, 'The Rules of Probability', in order to transform any odds Ω to a probability, we compute $\frac{\Omega}{\Omega+1}$; an odds of 2 in favor of Bobbie therefore translates to a posterior probability of ${}^2\!/_3$, consistent with the result from the probability form of Bayes' rule. The result is visualized in Figure 32.2.

We have arrived at another moment for solemn contemplation, because Figure 32.2 shows the first *posterior distribution* in this book. The

[2] As we will discuss in more detail later, the statistical term *likelihood* means unsurprise: the extent to which the observed data were expected or predicted under a hypothesized data-generating process θ.

interpretation of the prior and posterior distribution is identical, in the sense that both reflect the relative plausibility of the candidate values of bacon proclivity θ – both distributions quantify the allocation of belief across the different values of θ. The difference is that the 'prior' distribution reflects the relative uncertainty about the values of θ *before* seeing the data, and the 'posterior' distribution reflects the relative uncertainty about the values of θ *after* seeing the data. The 'before' and 'after' refer to our state of knowledge, not to time. For instance, an existing 'prior' opinion about a species of dinosaur may be updated by the discovery of a new set of fossils, resulting in 'posterior' opinion, even though the data were laid down before millions of years before the prior opinion was formed.[3]

[3] Linguistically, we may distinguish 'prediction' (a statement of uncertainty regarding future data that are as yet unknown to the forecaster) from 'retrodiction' (a statement of uncertainty regarding past data that are as yet unknown to the forecaster). There is also 'postdiction' (a statement of uncertainty regarding data that are known to the forecaster), but this comes close to statistical cheating.

Likelihood

In our pancake example, we updated our beliefs about the identity of the baker as a function of how well the rival hypotheses predicted the first datum (i.e., a bacon pancake, $\{b\}$), that is, $p(\{b\} \mid \theta_A)$ and $p(\{b\} \mid \theta_B)$. This measure of predictive success is generally known as the *likelihood*, "the probability that the observations should have occurred, given the hypothesis and the previous knowledge" (Jeffreys 1939, p. 46). Non-Bayesians slightly complicate matters by defining it as anything that is *proportional* to predictive success, such that $c \cdot p(\{b\} \mid \theta_A)$ is also a likelihood, for any non-zero number c (Etz 2018, Myung 2003).

Regardless, Bayesians and non-Bayesians agree on the importance of the likelihood. Our Bayesian hero Sir Harold Jeffreys wrote:

> "The prior probability of the hypothesis has nothing to do with the observations immediately under discussion, though it may depend on previous observations. Consequently, the whole of the information contained in the observations that is relevant to the posterior probabilities of different hypotheses is summed up in the values that they give to the likelihood." (Jeffreys 1939, p. 46; see also Jeffreys 1938c and Jeffreys 1961, p. 57).

In a brief comment to Jeffreys (1938c), his anti-Bayesian nemesis Sir Ronald Fisher actually agreed:

> "It may thus be said as Jeffreys notes, that the likelihood function contains the whole of the information supplied by the observations."

Given its central importance to statistical inference, it is surprising that most introductions to statistics hardly mention likelihood at all.

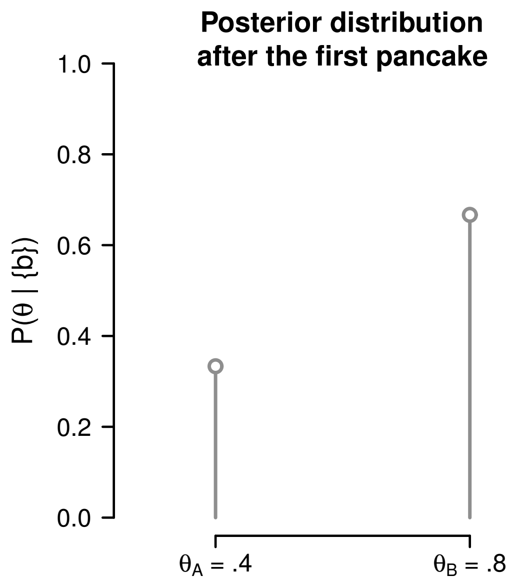


Figure 7.2: Having observed that the first pancake has bacon, Miruna now believes it is twice as likely that Bobbie rather than Andy is the baker.

Datum 2: A Vanilla Pancake

Miruna observes a second pancake and notices that it does *not* have bacon, an event that we denote as $\{v\}$ (for 'vanilla'). This observation has to shift her conviction back in the direction of Andy being the baker. Moreover, the totality of pancakes observed so far (i.e., $\{b, v\}$) has a bacon sample mean of $.50$, closer to Andy's $\theta_A = .40$ than Bobbie's $\theta_B = .80$, so the overall evidence ought to support the hypothesis that Andy is the baker. Let's substantiate this intuition with a Bayesian calculation.

We continue with the odds form of Bayes' rule. Taking into account the knowledge that the first pancake was bacon, we have:

$$\overbrace{\frac{p(\theta_B \mid \{b,v\})}{p(\theta_A \mid \{b,v\})}}^{\text{Posterior odds}} = \overbrace{\frac{p(\theta_B) \mid \{b\})}{p(\theta_A) \mid \{b\})}}^{\text{Prior odds}} \times \overbrace{\frac{p(\{v\} \mid \theta_B)}{p(\{v\} \mid \theta_A)}}^{\text{Evidence}}$$
$$= 2 \times \frac{2/10}{6/10} = 2/3.$$

Transforming odds to probability, we obtain the posterior probability that Bobbie is the baker as $p(\theta_B \mid \{b, v\})$ as $\frac{2/3}{2/3+1} = 2/5 = .40$, and hence the posterior probability that Andy is the baker equals $p(\theta_A \mid \{b, v\}) = 1 - .40 = .60$. The updated posterior distribution after two pancakes is shown in Figure 7.3.

Note that the prior odds had been updated to take into account the knowledge that the first pancake was bacon. We could also have updated differently: what if Miruna had seen the two pancakes at the

"For, evidently, those systems will be regarded as the more probable in which the greater expectation had existed of the event which actually occurred. The estimation of this probability rests upon the following theorem:

If, any hypothesis H being made, the probability of any determinate event E is h, and if, another hypothesis H' being made excluding the former and equally probable in itself, the probability of the same event is h': then I say, when the event E has actually occurred, that the probability that H was the true hypothesis, is to the probability that H' was the true hypothesis, as h to h'." (Carl Friedrich Gauss, 1809, as reported in D'Agostini 2020; italics in original)

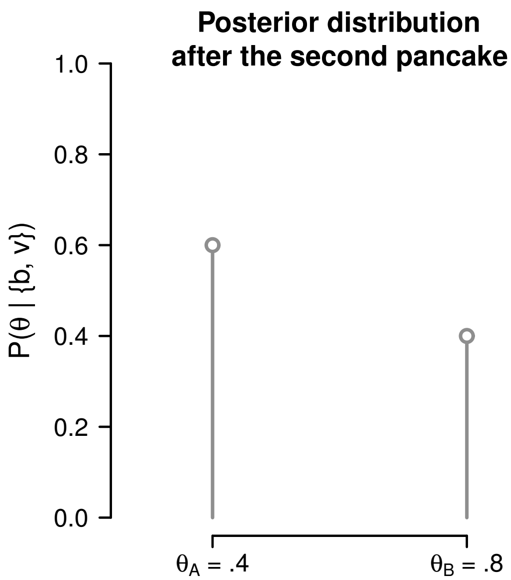


Figure 7.3: Having observed that the first pancake has bacon and the second pancake is vanilla, Miruna now believes the probability is .60 that Andy rather than Bobbie is the baker.

same time, instead of one-by-one? We would then have had:

$$\begin{aligned}\overbrace{\frac{p(\theta_B \mid \{b,v\})}{p(\theta_A \mid \{b,v\})}}^{\text{Posterior odds}} &= \overbrace{\frac{p(\theta_B)}{p(\theta_A)}}^{\text{Prior odds}} \times \overbrace{\frac{p(\{b,v\} \mid \theta_B)}{p(\{b,v\} \mid \theta_A)}}^{\text{Evidence}} \\ &= 1 \times \frac{8/10}{4/10} \times \frac{2/10}{6/10} \\ &= 1 \times 2 \times 1/3 = 2/3,\end{aligned}$$

which gives exactly the same result. In general, it does not matter for our conclusion whether the pancakes come in sequentially, as they are being baked, or simultaneously, as a completed stack.[4] To drive home this important point, notice that every bacon pancake yields a likelihood ratio of 2 in favor of Bobbie (i.e., $\text{LR}_b = p(\{b\} \mid \theta_B)/p(\{b\} \mid \theta_A) = 2$), whereas every vanilla pancake yields a likelihood ratio of 3 in favor of Andy (i.e., $\text{LR}_v = p(\{v\} \mid \theta_B)/p(\{v\} \mid \theta_A) = 1/3$). Every new pancake therefore multiplies the posterior odds by either 2 (if it's bacon) or $1/3$ (if it's vanilla). Symbolically, for just two pancakes, bacon followed by vanilla, we have:

[4] See Chapter 27 for details.

$$\text{Posterior odds} = \text{Prior odds} \times \text{LR}_b \times \text{LR}_v.$$

Updating the prior odds after the first pancake, and then adding the evidence from the second pancake can be represented as

$$\text{Posterior odds} = [\text{Prior odds} \times \text{LR}_b] \times \text{LR}_v,$$

whereas simultaneous updating can be represented as

$$\text{Posterior odds} = \text{Prior odds} \times [\text{LR}_b \times \text{LR}_v].$$

The commutative property of multiplication entails that these operations result in the same outcome. It also follows that the order in which the pancakes are observed does not matter for the end result. Finally, note that as the pancakes accumulate, the associated multiplicative evidence factors keep accumulating as well, such that the influence of the prior odds is increasingly diluted: eventually, the evidence overwhelms the prior opinion. Given that the problem was correctly specified, this overwhelming evidence will identify the best predicting hypothesis with a probability that approaches 1.

"Thus it does not matter in what order we introduce our data; as long as we start with the same data and finish with the same additional data, the final results will be the same. The principle of inverse probability cannot lead to inconsistencies." (Jeffreys 1938a, pp. 191-192).

An Excursion to Stylometry

Before proceeding to the second question ("will the next pancake have bacon?") we will attempt to pacify those readers who feel the pancake scenario lacks gravitas. Consider the following authorship question (Mosteller and Wallace 1963):[5]

> "The *Federalist* papers were published anonymously in 1787-1788 by Alexander Hamilton, John Jay, and James Madison to persuade the citizens of the State of New York to ratify the Constitution. Of the 77 essays, 900 to 3500 words in length, that appeared in newspapers, it is generally agreed that Jay wrote five: Nos. 2, 3, 4, 5, and 64, leaving no further problem about Jay's share. Hamilton is identified as the author of 43 papers, Madison of 14. The authorship of 12 papers (Nos. 49-58, 62, and 63) is in dispute between Hamilton and Madison; finally, there are also three joint papers, Nos. 18, 19, and 20, where the issue is the extent of each man's contribution." (Mosteller and Wallace 1963, p. 276)

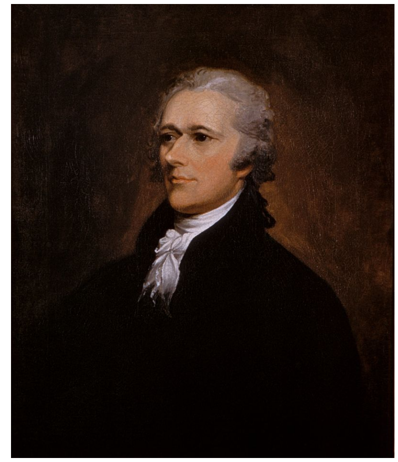
[5] This example is inspired by Donovan and Mickey (2019) and the `https://priceonomics.com` blog post "How Statistics Solved a 175-Year-Old Mystery About Alexander Hamilton".

Remarkably, this authorship dispute can be resolved even hundreds of years after the authors have passed away, and in a way that is statistically similar to the pancake scenario. Instead of asking "who baked the pancakes, Andy or Bobbie?" we ask "who wrote the disputed *Federalist* papers, Hamilton or Madison?"

The general idea is that the authorship dispute can be resolved by considering *writing style*. We first use the undisputed works to analyze and quantify the writing style of each candidate author. For instance, perhaps Hamilton generally used longer words or longer sentences than Madison; this difference in writing style can then be used as a clue about authorship of the disputed papers. Specifically, we could compute the average word-length or sentence-length from the disputed papers and assess whether these features are more Hamilton-like or more Madison-like. The idea may have been first conceived by Augustus De Morgan[6]. In a 1851 letter to a friend, De Morgan wrote:

[6] We already met De Morgan in Chapter 5, when we discussed his 'alphabet' for measuring epistemic probability.

> "I wish you would do this: run your eye over any part of those of St. Paul's Epistles which begin with $\Pi\alpha\upsilon\lambda o\varsigma$—the Greek I mean—and without paying any attention to the meaning. Then do the same with the Epistle to the Hebrews, and try to balance in your own mind the question whether the latter does not deal in longer words than the former. It has always run in my head that a little expenditure of money would settle questions of authorship in this way. The best mode of explaining what I would try will be to put down the results I should *expect* as if I had tried them.
>
> Count a large number of words in Herodotus—say all the first book—and count all the letters; divide the second numbers by the first, giving the average number of letters to a word *in that book*.
>
> Do the same with the second book. I should expect a very close approximation. If Book I. gave 5.624 letters per word, it would not surprise me if Book II. gave 5.619. I judge by other things.
>
> But I should not wonder if the same result applied to two books of Thucydides gave, say 5.713 and 5.728. That is to say, I should expect the slight differences between one writer and another to be well maintained against each other, and very well agreeing with themselves. If this fact were established there, if St. Paul's Epistles which begin with $\Pi\alpha\upsilon\lambda o\varsigma$ gave 5.428 and the Hebrews gave 5.516, for instance, I should feel quite sure that the *Greek* of the Hebrews (passing no verdict on whether Paul wrote in Hebrew and another translated) was not from the pen of Paul.
>
> If scholars knew the law of averages as well as mathematicians, it would be easy to raise a few hundred pounds to try this experiment on a grand scale. I would have Greek, Latin, and English tried, and I should expect to find that one man writing on two different subjects agrees more nearly with himself than two different men writing on the same subject. Some of these days spurious writings will be detected by this test. Mind, I told you so." (De Morgan 1882, pp. 215-216; from a 1851 letter to Rev. W. Heald)



Alexander Hamilton (1755 or 1757 – 1804), one of the authors of the *Federalist* papers and one of the Founding Fathers of the United States of America. Portrait by John Trumbull, 1806.

I told you so, indeed![7] Now well-established, the field of *stylometry* –the computational analysis of writing style– offers a sophisticated statistical methodology to attribute authorship for disputed works. Modern stylometry often depends on machine learning methods such as provided by the Java Graphical Authorship Attribution Program (Juola 2006) or the `R` package `stylo` (Eder et al. 2016).

[7] Not all of De Morgan's ideas proved similarly prophetic. For instance, in an 1853 letter to the same friend, De Morgan wrote: "I remember giving you my experience in regard to clairvoyance. I will now tell you some of my experience in reference to table-turning, spirit-rapping, and so on. (...) I am, however, satisfied of the reality of the phenomenon." De Morgan 1882, pp. 221-222

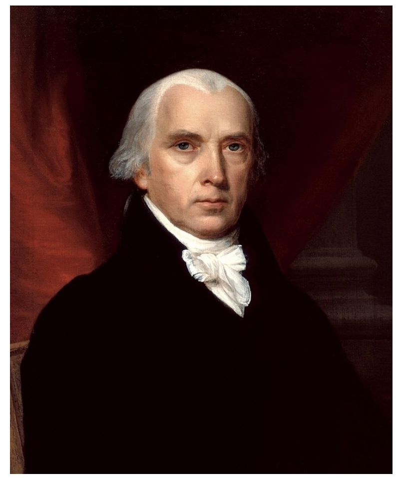
With only limited assistance of computers, however, stylometry can be quite laborious. To begin with, one of the main challenges in the pre-computer era was to discover which aspects of a writing style are *diagnostic* in the first place. And, unfortunately, Hamilton and Madison were stylistically rather similar:

> "The writings of Hamilton and Madison are difficult to tell apart because both authors were masters of the popular *Spectator* style of writing–complicated and oratorical. To illustrate, in 1941 Frederick Williams and Frederick Mosteller counted sentence lengths for the undisputed papers and got means of 34.55 and 34.59 words respectively for Hamilton and Madison, and average standard deviations for papers of 19.2 and 20.3.

> These results show that for some measures the authors are practically twins." (Mosteller and Wallace 1963, p. 276)

Mosteller and Wallace (1963) then proceeded to consider the frequency with which Hamilton and Madison used individual words – they focused their efforts on filler words such as 'an', 'of', 'to', and 'by'; because these are both common and topic-independent, they are potentially ideal candidates for discriminating the writers. After a considerable amount of work, Mosteller and Wallace (1963, p. 278) concluded that "The best single discriminator we have ever discovered is *upon*, whose rate is about 3 per thousand for Hamilton and about 1/6 per thousand for Madison." For educational purposes (and with some trepidation, for we are doing the work of Mosteller and Wallace an injustice), we consider only the discriminator word 'upon'. We follow Donovan and Mickey (2019) and focus on disputed paper no. 54, "The Apportionment of Members Among the States", a document of 2008 words in which the word 'upon' occurs twice.

The similarity to our pancake scenario is now clear: Hamilton is a baker of words with an 'upon' proclivity of $\theta_H = {}^3\!/_{1000} = .003$, whereas Madison has an 'upon' proclivity of $\theta_M = {}^1\!/_{6000} \approx .00017$. We are then presented with a 'stack' of 2008 words, two of them being 'upon'. What evidence does this provide for each man's authorship claim? One of the exercises at the end of this chapter invites the reader to use the *Learn Bayes* module in JASP to find out exactly, but we can already guesstimate the outcome; the observed frequency of occurrence for 'upon' in "The Apportionment of Members Among the States" is about 1 in a 1000 – slightly lower than Hamilton's rate of 3/1000, but higher than Madison's rate of 1/6000. Overall, the sample outcome is closer to what is expected under Hamilton than to what is expected under Madison; the sample ought to provide modest evidence for Hamilton being the author.



James Madison (1751–1836), one of the authors of the *Federalist* papers, and the fourth President of the United States of America.

Question 2: Will the Next Pancake Have Bacon?

Miruna goes through 6 pancakes and finds that 4 have bacon, in the order $\{b, v, b, b, b, v\}$. The likelihood ratio contribution is

$$\frac{p(\{b,v,b,b,b,v\} \mid \theta_B)}{p(\{b,v,b,b,b,v\} \mid \theta_A)} = \left[\frac{p(\{b\} \mid \theta_B)}{p(\{b\} \mid \theta_A)}\right]^4 \times \left[\frac{p(\{v\} \mid \theta_B)}{p(\{v\} \mid \theta_A)}\right]^2$$
$$= 2^4 \times \frac{1}{3}^2 = {}^{16}\!/_9.$$

Transforming the odds to posterior probability we find that $p(\theta_B \mid \{b,v,b,b,b,v\}) = \frac{16/9}{16/9+1} = {}^{16}\!/_{25} = .64$ (i.e., the probability that Bobbie is the baker equals .64), and hence $p(\theta_A \mid \{b,v,b,b,b,v\}) = 1 - .64 = .36$ (i.e., the probability that Andy is the baker is .36).[8] We are now in the situation to quantify

[8] The order of trials may be unknown or irrelevant, in which case we compute not the probability of a specific order, but the probability of *any* order that includes, say, 4 bacon pancakes and 2 vanilla pancakes (see Chapter 33.) This does not affect the outcome of the Bayesian analysis.

Extraordinary Claims Require Extraordinary Evidence

The odds form of Bayes' rule shows that the posterior odds (what we believe after having seen the data) equals the evidence (how the data change our beliefs) when the prior odds is 1; in that case we have:

$$\underbrace{\frac{p(\text{Hypothesis X} \mid \text{data})}{p(\text{Hypothesis Y} \mid \text{data})}}_{\substack{\text{Posterior plausibility} \\ \text{for the rival hypotheses}}} = 1 \times \underbrace{\frac{p(\text{data} \mid \text{Hypothesis X})}{p(\text{data} \mid \text{Hypothesis Y})}}_{\substack{\text{Evidence} \\ \text{from the data}}}.$$

When the prior odds is not 1, however, evidence and posterior belief/knowledge can be quite different, as is conveyed by the adage 'extraordinary claims require extraordinary evidence'. For instance, suppose that, upon entering her house, Miruna is greeted by Bobbie, who is smelling strongly of bacon, has pieces of pancake stuck in her hair, and is wearing a chef's apron with fresh butter stains. These prior observations mean that the prior odds are now massively in favor of Bobbie being the baker. The same stack of pancakes (i.e., the same evidence) that, starting from a position of equipoise, would have made Miruna believe that Andy is the baker, now –when taking this prior knowledge into account– still has her believe that it is in fact Bobbie who is the baker.

The great Pierre-Simon Laplace –the first real 'Bayesian'– often used prior odds of 1 in his work. However, Laplace was well aware of the fact that this practice is correct only if the competing hypotheses are equally likely a priori. In fact, Laplace stated that "The weight of evidence for an extraordinary claim must be proportioned to its strangeness.", a statement that anticipates the popular phrase from the American astronomer Carl Sagan (1934-1996): "extraordinary claims require extraordinary evidence."

our conviction that the seventh pancake will have bacon. Note that, as demonstrated in Chapter 2, this requires that we take into account both our *epistemic* uncertainty ("who baked the pancakes"?) and our *aleatory* uncertainty ("given the identity of the baker, what is the chance of getting a bacon pancake?").

We know that if Andy is the baker, the probability that the seventh pancake (or any other, for that matter) has bacon is $\theta_A = .40$; if Bobbie is the baker, this probability is $\theta_B = .80$. According to the law of total probability (see Chapter 3), the overall probability that the seventh pancake has bacon is an average of these two θ's, with averaging weights given by the posterior probability that Andy (or Bobbie) is the baker:

$$\begin{aligned} p(\{b\} \mid \{b,v,b,b,b,v\}) &= p(\{b\} \mid \theta_A) \cdot p(\theta_A \mid \{b,v,b,b,b,v\}) \\ &\quad + p(\{b\} \mid \theta_B) \cdot p(\theta_B \mid \{b,v,b,b,b,v\}) \\ &= {}^4\!/_{10} \cdot {}^9\!/_{25} + {}^8\!/_{10} \cdot {}^{16}\!/_{25} \\ &= {}^{164}\!/_{250} = .656. \end{aligned}$$

As usual, the law of total probability can be understood by constructing a tree diagram, as in Figure 7.4. The probability that the Andy branch is taken and results in a bacon pancake is $.36 \cdot .40$; for the Bobbie branch this probability is $.64 \cdot .80$. Adding both probabilities yields $.656$.

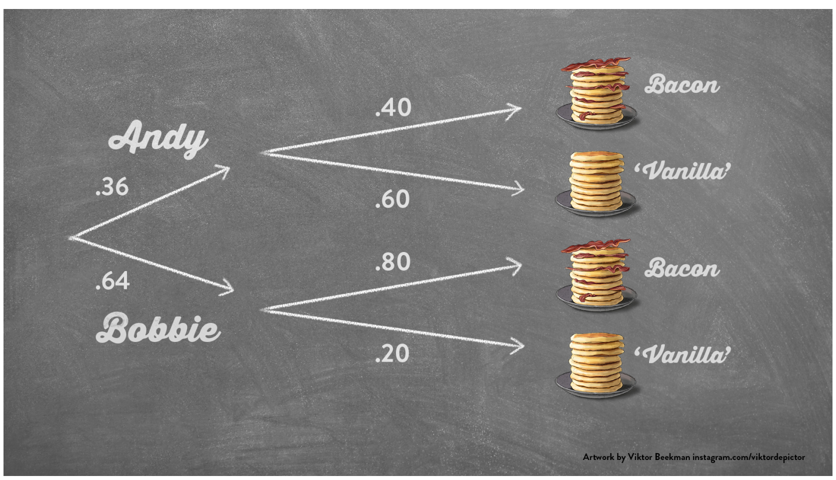


Figure 7.4: To obtain the probability that the seventh pancake has bacon, use the law of total probability and add the probability of the two branches that result in bacon: $.36 \cdot .40 + .64 \cdot .80 = .656$. Note that, in the figure, the first branching factor refers to our *epistemic* uncertainty regarding the identity of the baker, and the second branching factor refers to our *aleatory* uncertainty regarding the nature of the pancake, given that we know the identity of the baker.

When we view bacon proclivity θ as a *parameter* (i.e., a single-process 'dial' that can be set to different values), this application of the law of total probability is called computing a 'posterior predictive'. When instead we view Andy and Bobbie as two rival *models of the world*, this application of the law of total probability is called 'Bayesian model averaging'. The operation is mathematically identical, and only the surface label differs (e.g., Gronau and Wagenmakers 2019).

As a final thought, note the similarity of the averaging process with the phenomenon known as the 'wisdom of the crowd', where the average prediction of a group of people outperforms the majority of the individual predictions. In the Bayesian version, the average is weighted by posterior plausibility, which can be likened to a person's level of expertise (i.e., their prior credentials and the adequacy of their previous predictions).

The Bayesian World is Comparative

Suppose we were to observe that all of $n = 20$ pancakes are of the vanilla variety. The evidence for Andy being the baker is then computed as follows:

$$\text{Evidence that Andy is the baker} = \left[\frac{p(\{v\} \mid \theta_A)}{p(\{v\} \mid \theta_B}\right]^n = \left[\frac{6/10}{2/10}\right]^{20} = 3^{20}.$$

With prior odds of 1, this means that it is now 3,486,784,401 times more likely that Andy rather than Bobbie is the baker, for a posterior probability of $3{,}486{,}784{,}401/3{,}486{,}784{,}402 \approx 0.9999999997$. This looks like a pretty compelling result – but there is a catch. The data are a sequence of 20 consecutive vanilla pancakes, and such a sequence is highly unlikely if Andy is the baker. The reason that the evidence is overwhelmingly in favor of Andy is because the data are virtually impossible under the hypothesis that Bobbie is the baker. So both hypotheses predict the data poorly, but the Bobbie hypothesis is particularly abysmal.

It should therefore be kept in mind that "The Bayesian world is a comparative world in which there are no absolutes." (Lindley 2000, p. 308). Our Bayesian plausibility assessments are always conditional on background knowledge K; hence, we could have written the prior probabilities more elaborately as $p(\theta_A \mid K)$ and $p(\theta_B \mid K)$. The background knowledge may include the fact that we believed we were faced with a choice between Andy and Bobbie. The fact that Andy's sister came to visit, and that she is a fanatic vegetarian, was not part of K. In such a case, the models are said to be *misspecified* (see also Gronau and Wagenmakers 2019 and references therein). Some Bayesians have devised more or less ad-hoc devises to evaluate a model in isolation (e.g., Box 1980) but the royal Bayesian road always involves multiple models – the Bayesian world is comparative.

Exercises

1. Many textbooks present Bayes' rule as follows:

$$\begin{aligned} p(\theta \mid \text{data}) &= \frac{p(\theta)\, p(\text{data} \mid \theta)}{p(\text{data})} \\ &= p(\theta)\, p(\text{data} \mid \theta) \cdot {}^{1}\!/\!_{c} \\ &\propto p(\theta)\, p(\text{data} \mid \theta), \end{aligned}$$

where c is a single non-zero number and the $\propto$ symbol means 'is proportional to'. In words, we have (Jeffreys 1939, p. 46):

$$\text{Posterior} \;\propto\; \text{Prior} \times \text{Likelihood},$$

which means that our updated knowledge of the world ('posterior') is a compromise between our old knowledge ('prior') and the information coming from the data ('likelihood', or 'predictive success'). Show how to use this formulation to go from Figure 7.1 to Figure 32.2.

2. Consider again the authorship question for the *Federalist* papers. As before, assume that Hamilton's rate of using 'upon' equals $\theta_H = {}^{3}\!/\!_{1000}$ whereas Madison's rate equals $\theta_M = {}^{1}\!/\!_{6000}$. Disputed paper no. 54 is 2008 words long, two of which are 'upon'.

2.1. What is the prior probability that Hamilton is the author?

2.2. As was done in the first paragraph of the section 'Question 2: Will the next pancake have bacon?' decompose the likelihood ratio and quantify the contribution of each occurrence of 'upon' versus each occurrence of any other word. Which term is more influential?

2.3. Compute the evidence that the 'upon' data (i.e., 2 out of 2008) provide for the hypothesis that Hamilton wrote paper no. 54.

2.4. Update your prior probability that Hamilton wrote paper no. 54 to your posterior probability.

2.5. Use the *Learn Bayes* module in JASP to confirm your results.

2.6. Consider disputed paper no. 63, "The Senate Continued", which is 3033 words long, without any occurrence of 'upon'.[9] Use the *Learn Bayes* module in JASP to quantify the evidence that these data provide for Madison rather than Hamilton being the author.

2.7. It is striking how rarely the word 'upon' occurs in the disputed papers. What does this suggest?

3. At the start of this chapter, we argued that the questions "who baked the pancakes?" and "will the next pancake have bacon?" are *fundamentally different*. Now argue that we were wrong, and that these questions are in fact intimately connected.

[9] The full text of the *Federalist* papers is available at `https://guides.loc.gov/federalist-papers`.

4. We've established that the probability that the seventh pancake will have bacon is .656.

4.1. What is the probability that the seventh *and* eighth pancakes will both have bacon? (hint: expand the tree diagram in Figure 7.4).

4.2. Confirm your answer using the *Learn Bayes* module in JASP (hint: use the *Binomial Testing* routine).[10]

4.3. Explain why the answer $.656 \times .656$ is both tempting and wrong.

[10] The term 'binomial' refers to the fact that only two outcomes are possible (here: the pancakes are assumed to be of only two types, bacon or vanilla).

Chapter Summary

These are the main lessons from this chapter:

- Prior knowledge about the relative plausibility of rival hypotheses is adjusted by the data to yield posterior knowledge.
- The adjustment brought about by the data is a function of the rival hypotheses' success in predicting those data. Hypotheses under which the data are relatively surprising decrease in plausibility.
- Only when the rival hypotheses are equally plausible a priori is it true that the evidence (i.e., relative predictive success) equals knowledge or belief (i.e., posterior probability).
- Bayes' rule allows one to infer probable causes (e.g., the identity of the baker) from observed consequences (e.g., the composition of the pancake stack).
- Data may be analyzed sequentially or simultaneously: the end result is exactly the same.
- Eventually, the evidence from the data will overwhelm the prior opinion.
- In order to obtain a *prediction* for to-be-observed data one needs to consider all possible causes, and weigh the prediction from each with the posterior plausibility of that cause (i.e. apply the law of total probability).

Want to Know More?

✓ Donovan, T. M., & Mickey, R. M. (2019). *Bayesian Statistics for Beginners: A Step–by–Step Approach*. Oxford: Oxford University Press.

✓ Mosteller, F., & Wallace, D. L. (1963). Inference in an authorship problem. *Journal of the American Statistical Association*, 58, 275–309.

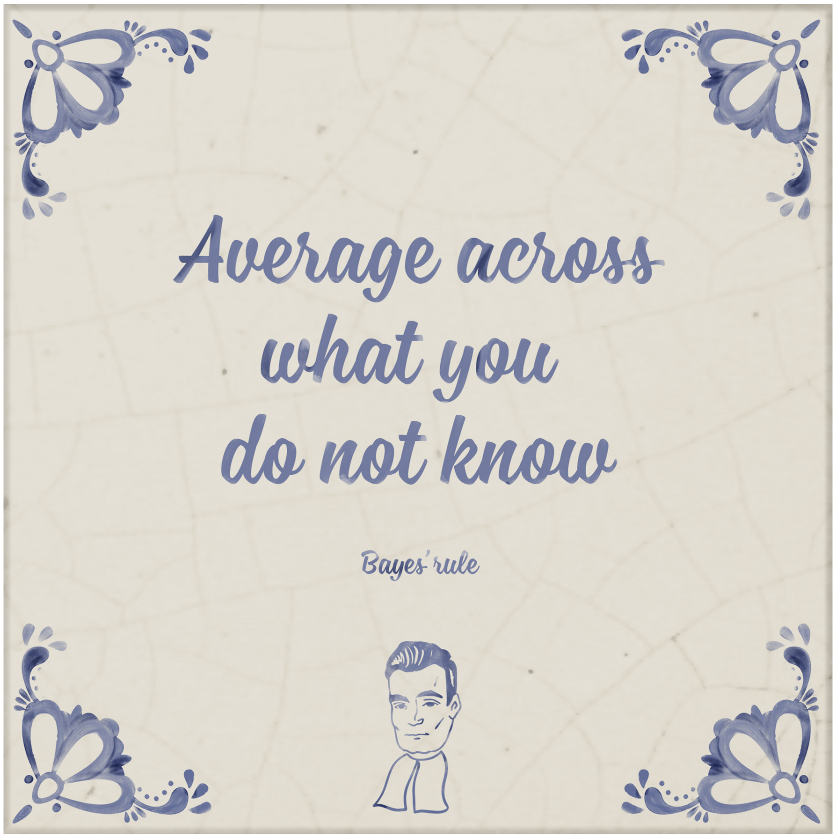


Figure available at `BayesianSpectacles.org` under a CC-BY license.

The paper that energized the field of *stylometry*: the use of statistics to quantify writing style.

✓ Mosteller, F., & Wallace, D. L. (1984). *Applied Bayesian and Classical Inference: The Case of* The Federalist *Papers (2nd ed.)*. New York: Springer. A riveting and comprehensive Bayesian account of the authorship problem, a summary of which was given in the above-referenced 1963 article. The first edition of this book was published in 1964 under the title "Inference and Disputed Authorship: *The Federalist*".

8 An Infinite Number of Hypotheses [with Quentin F. Gronau]

It might seem, indeed, utterly impossible to calculate out a problem having an infinite number of hypotheses, but the wonderful resources of the integral calculus enable this to be done (...) But I may add that though the integral calculus is employed as a means of summing infinitely numerous results, we in no way abandon the principles of combinations already treated.

Jevons, 1874

Chapter Goal

This chapter explains how Bayesians routinely update beliefs about an infinite number of rival hypotheses.

Many Potential Bakers

In the example from Chapter 7 there were only two possible bakers, each with known bacon proclivity: Andy with $\theta_A = .40$, and Bobbie with $\theta_B = .80$. Exactly the same principles of knowledge updating apply when more candidate bakers are introduced. For instance, we can add the following nine: Charly with $\theta_C = 0$; Denver with $\theta_D = .10$; Evan with $\theta_E = .20$; Frankie with $\theta_F = .30$; Jackie with $\theta_J = .50$; Lennon with $\theta_L = .60$; Oakly with $\theta_O = 0.70$; Robin with $\theta_R = 0.90$; and Sidney with $\theta_S = 1$. Note that Charly is a vegetarian and *never* bakes bacon pancakes, whereas Sidney is a carnivore who *always* bakes bacon pancakes. So now the question that Miruna faces, when she comes home to have a pancake dinner with her extended family, is "who baked the pancakes – Andy, Bobbie, Charly, Denver, Evan, Frankie, Jackie, Lennon, Oakly, Robin, or Sidney?"

As before, the probability-form of Bayes' rule shows that the posterior probability of person i being the baker (i.e., $p(\theta_i \mid \text{data})$) is obtained by updating their prior probability (i.e., $p(\theta_i)$) with their relative predictive performance:

$$
\begin{aligned}
p(\theta_i \mid \text{data}) &= p(\theta_i) \cdot \frac{p(\text{data} \mid \theta_i)}{p(\text{data})} \\
&= p(\theta_i) \cdot \frac{p(\text{data} \mid \theta_i)}{\sum_{j=1}^{n} p(\text{data} \mid \theta_j)\, p(\theta_j)}.
\end{aligned}
$$

Note that the average predictive performance, $p(\text{data})$, is obtained by applying the rule of total probability (cf. the tree diagram in Figure 7.4). The knowledge updating term $p(\text{data} \,|\, \theta_i)/p(\text{data})$ can also be interpreted in terms of a change in *surprise*. Averaged across all rival hypotheses, $p(\text{data})$ quantifies the extent to which the observed data are predictable or *unsurprising*: the lower this number, the more surprising the data. Then we consider how unsurprising the observed data are when we assume that person i was the baker (i.e., $p(\text{data} \,|\, \theta_i)$), that is, when we *condition* on person i being the baker. When the act of conditioning on θ_i reduces the surprise (i.e., increases the 'unsurprise'[1]), we have $p(\text{data} \,|\, \theta_i) > p(\text{data})$ and this in turn implies that $p(\theta_i \,|\, \text{data}) > p(\theta_i)$: in words, hypotheses gain credibility when they make the observed data more predictable (i.e., less surprising).[2]

[1] Instead of unsurprise, Rosenkrantz (1983, p. 75) used the term 'expectedness'.

[2] This is the central concept of Bayesian learning, and we will keep bringing it up throughout this book, for instance in Chapter 25. See also Rouder and Morey (2019).

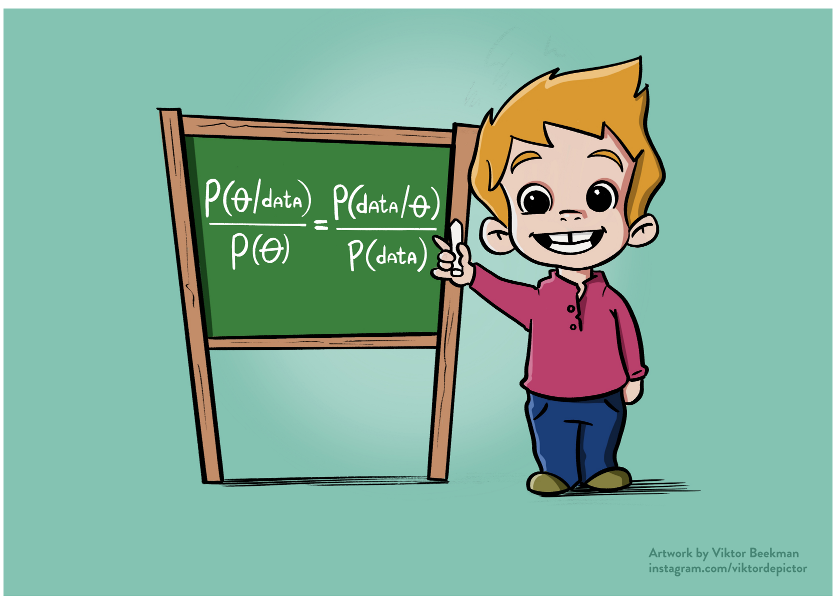


Figure 8.1: A hypothesis θ gains credibility (i.e., $p(\theta \,|\, \text{data}) > p(\theta)$ when it acts to reduce surprise from the data (i.e., $p(\text{data} \,|\, \theta) > p(\text{data})$). *Surprise lost is credibility gained.* Figure available at `BayesianSpectacles.org` under a CC-BY license.

When each person is deemed equally likely *a priori* to be the baker, the factor $p(\theta_i)$ cancels (in our pancake example, $p(\theta_i) = 1/11$, as there are 11 candidate bakers), and the posterior probability is determined solely by relative predictive success, unweighted with prior plausibility:

$$p(\theta_i \mid \text{data}) = \frac{p(\text{data} \mid \theta_i)}{\sum_{j=1}^{n} p(\text{data} \mid \theta_j)}.$$

The posterior probability for person i can then be interpreted as the *proportion* of unsurprise, or the proportion of predictability.[3]

For concreteness, consider that the data consists of a single bacon pancake, data $= \{b\}$. For each baker i, the prediction for this event simply equals their bacon proclivity parameter θ_i. Table 8.1 shows the 11 candidate bakers, the associated prediction that the first pancake will be either vanilla or bacon, the bakers' prior probability, and their resulting posterior probability after observing that the first pancake has bacon. Note that in the equation immediately above, $\sum_{j=1}^{n} p(\text{data} = \{b\} \mid \theta_j) = 0 + .1 + .2 + .3 + .4 + .5 + .6 + .7 + .8 + .9 + 1 = 5.5$ (i.e., the sum of the 'Bacon' column in Table 8.1), such that the posterior probability for each baker is simply the proclivity θ_i divided by 5.5.

[3] "If there is originally no ground to believe one of a set of alternatives rather than another, the prior probabilities are equal. The most probable, when evidence is available, will then be the one that was most likely to lead to that evidence. *We shall be most ready to accept the hypothesis that requires the fact that the observations have occurred to be the least remarkable coincidence.*" (Jeffreys 1961, p. 29; italics ours)

"*Sixth Principle.*—The greater the probability of an observed event given any one of a number of causes to which that event may be attributed, the greater the likelihood of that cause {given that event}. The probability of the existence of anyone of these causes {given the event} is thus a fraction whose numerator is the probability of the event given the cause, and whose denominator is the sum of similar probabilities, summed over all causes. If these various causes are not equally probable *a priori*, it is necessary, instead of the probability of the event given each cause, to use the product of this probability and the possibility of the cause itself. This is the fundamental principle of that branch of the analysis of chance that consists of reasoning *a posteriori* from events to causes." (Laplace 1814/1995, pp. 8-9, italics in original)

Table 8.1: Who baked the pancakes? Eleven candidate bakers, each with known bacon proclivity θ_i, are associated with a prediction for whether or not the first pancake will have bacon. After observing that the first pancake has bacon, the candidate bakers' prior plausibility $p(\theta_i) = 1/11 = 5/55$ is updated to a posterior probability, given in the final column.

		Pancake prediction			
Candidate baker	Bacon proclivity	Vanilla	Bacon	Prior probability	Posterior probability
Charly	$\theta_C = 0$	1	0	$5/55$	0
Denver	$\theta_D = .10$	.90	.10	$5/55$	$1/55 \approx .02$
Evan	$\theta_E = .20$	.80	.20	$5/55$	$2/55 \approx .04$
Frankie	$\theta_F = .30$	.70	.30	$5/55$	$3/55 \approx .05$
Andy	$\theta_A = .40$	.60	.40	$5/55$	$4/55 \approx .07$
Jackie	$\theta_J = .50$	.50	.50	$5/55$	$5/55 \approx .09$
Lennon	$\theta_L = .60$	.40	.60	$5/55$	$6/55 \approx .11$
Oakly	$\theta_O = .70$	.30	.70	$5/55$	$7/55 \approx .13$
Bobbie	$\theta_B = .80$	.20	.80	$5/55$	$8/55 \approx .15$
Robin	$\theta_R = .90$	.10	.90	$5/55$	$9/55 \approx .16$
Sidney	$\theta_S = 1$	0	1	$5/55$	$10/55 \approx .18$

The prior and posterior probabilities from Table 8.1 are shown in Figure 8.2. As explained in Chapter 7, these are distributions of *belief*, *conviction*, *plausibility*, or *uncertainty*, and they reflect our lack of knowledge about the identity of the baker before and after observing a single bacon pancake. Figure 8.2 and Table 8.1 allow the following conclusions:

- The observation of a single bacon pancake has 'irrevocably exploded' (Pólya 1954a, p. 6) the hypothesis that Charly is the baker – Charly

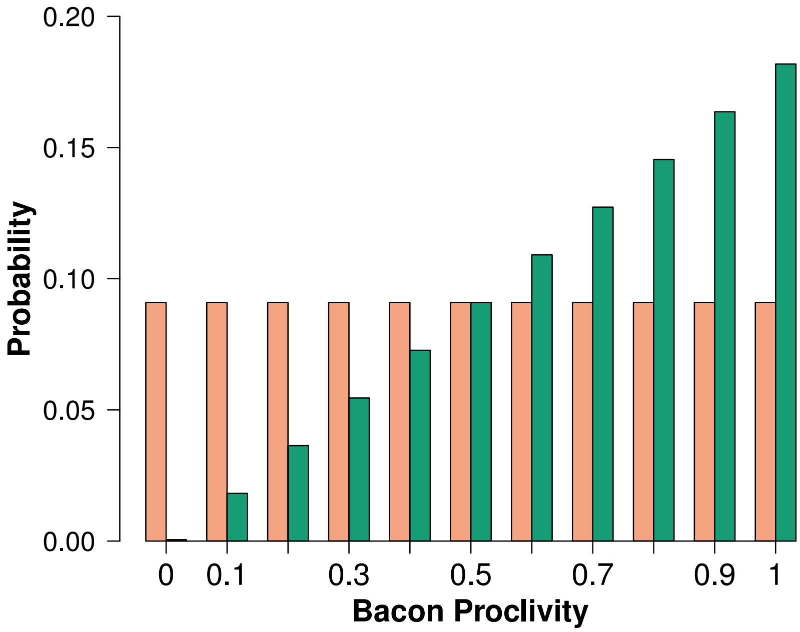


Figure 8.2: Prior distribution (in salmon) and posterior distribution (in green) across 11 possible bakers with known bacon proclivity, after observing a single bacon pancake. Exact numbers shown in Table 8.1.

is a vegetarian and *never* bakes bacon pancakes (i.e., $\theta_C = 0$). So we can be absolutely certain that Charly is not the baker. In this case, Bayesian inference reduces to propositional logic: 'Charly never bakes bacon pancakes' & 'A bacon pancake was baked' $\rightarrow$ 'Charly is not the baker'.

- The observation of a bacon pancake (i.e., a known 'consequence') makes it more likely that the baker (i.e., an unknown 'cause') has a high bacon proclivity rather than a low bacon proclivity. This is because the observation of a bacon pancake is less and less surprising as the bacon proclivity increases. The data are the least surprising under the hypothesis that Sidney is the baker – in fact, Sidney *only* bakes bacon pancakes, so the observation of a bacon pancake elicits no surprise whatsoever. Consequently, based on the observation of a single bacon pancake, the highest posterior probability is for Sidney being the baker.[4]
- Compared to their prior probabilities, high bacon proclivities θ_i have become more credible, and low θ_i's have become less credible; the fulcrum of the posterior distribution is at $\theta_J = .50$; for Jackie, the prior probability is the same as the posterior probability – in other words, the predictive performance of θ_J is exactly equal to the average, and its plausibility is therefore unchanged.

[4] Note the evidential asymmetry between Charly and Sidney: a prediction that is completely correct increases Sidney's plausibility from $5/55 \approx .09$ to $10/55 \approx .18$, whereas a prediction that is completely incorrect decreases Charly's plausibility to zero, from which it is impossible to recover.

- The observation of a new datum leads to an adjustment of beliefs; that is, credibility is re-allocated and flows towards hypotheses that predicted the datum relatively well and flows away from hypotheses that predicted the datum relatively poorly. Note that credibility is not gained or lost overall – the mass of the prior and posterior distribution always sums to 1.[5]

[5] This may be likened to the *conservation of volume* – when water is poured into a differently-sized container, the water level may change but the volume stays the same.

Probability versus Likelihood: It's Complicated

There is a subtle difference between 'probability' and 'likelihood'. Consider first the pancake predictions for each baker shown in Table 8.1. When Evan is the baker, the probability that the next pancake will have bacon is $.20$. Consequently a probability of $1 - .20 = .80$ is assigned to the complementary event that the pancake will be vanilla (in statistical jargon: 'with the parameter fixed and the data variable'). So, in Table 8.1, each row-specific prediction is a *probability*: given a specific account of the world, unknown events are assigned probabilistic predictions. Once the data are in (e.g., once we observe that the first pancake has bacon) it makes sense to consider only the predictions for the event that actually occurred. In Table 8.1, this means that we focus on the 'Bacon' column, and inspect how unsurprising the observed data are under the rival hypotheses (in statistical jargon: 'with the data fixed and the parameter variable'). For the observed data the predictions across the bakers are *not* probabilities – for instance, the numbers in the 'Bacon' column do not sum to 1. Instead, each individual prediction is known as a 'likelihood', and the entire column is known as a 'likelihood function' (e.g., Edwards 1992, Etz 2018, Lindley 2006, Myung 2003). If we want to transform the column of likelihoods $p(\text{data} \mid \theta_i)$ to a posterior probability $p(\theta_i \mid \text{data})$, we need to apply Bayes' rule and multiply each likelihood by a prior probability $p(\theta_i)$ and divide by $p(\text{data})$, the weighted average prediction across all bakers.

In sum, a statistical hypothesis makes predictions for to-be-observed data by assigning *probabilities* to exhaustive events (consequently, the numbers sum to 1 across the space of possible outcomes). With a particular observation in hand, however, we may compare the associated predictive performance across rival hypotheses. Considered as a function of the hypotheses, the numbers to not generally sum to 1; hence they are referred to not as probabilities, but as *likelihoods*. So yeah, it's complicated.

The Pancake Proclivity of Mr. X

In the previous example, we considered 11 possible bakers, each with a unique value of θ. This means we have 11 *discrete* possibilities for θ; each person was equally likely *a priori* to be the baker, that is, $p(\theta_i) = 1/11$. Now imagine an *army* of N possible bakers, each with their own bacon proclivity θ. Figure 8.2 would then consist of N prior and posterior probability bars; the prior probability of each soldier being the baker would equal $1/N$, and decrease towards zero as the army grows larger. In the limit of an infinitely large army, we transition from a discrete distribution to a *continuous* distribution, where the probability of any single baker is zero; the concept of probability now applies to a range of bakers, that is, to an area under the curve (cf. Figure 3.4).

To see why a continuous distribution would be useful, consider the following situation. Miruna comes home and is informed that the pancakes have been baked by Mr. X, a family friend whose bacon proclivity θ_X is *unknown*. Every value of θ_X from 0 to 1 represents a hypothesis about Mr. X's preference for bacon pancakes, and there is an infinite number of them. Before we consider the statistical details, let's consider what happens when we assume that, *a priori*, all values of θ_X are equally plausible – the resulting prior distribution is shown in Figure 8.3.

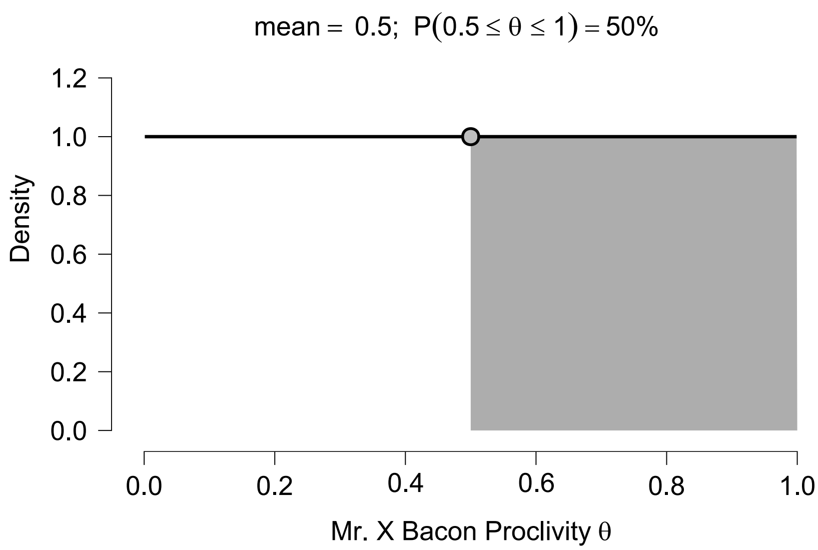


Figure 8.3: Prior distribution for the unknown bacon proclivity of Mr. X. Figure from the JASP module *Learn Bayes*.

The horizontal line indicates that all values for θ_X are deemed equally likely *a priori* (cf. the shape of the salmon-colored prior distribution across the eleven values of θ shown in Figure 8.2). The prior mean of θ_X is indicated by a dot and equals $1/2$. The prior probability that Mr.

X prefers bacon pancakes over vanilla pancakes (i.e., $p(\theta_X) > 1/2$) equals $1/2$ – the gray area under the curve.

The first pancake that Mr. X bakes has bacon, and this yields an update for all values of θ_X. The resulting posterior distribution is shown in Figure 8.4. The observation has tilted the distribution towards higher values of θ (cf. the shape of the green-colored posterior distribution across the eleven values of θ shown in Figure 8.2). The posterior mean equals $2/3$ (as we will see later, the value 0.667 is due to rounding). The most likely value of θ_X –the *mode*, where the posterior reaches its highest point– is 1.0. The posterior *median* –that value for θ_X below which lies 50% of posterior mass– equals .707. Finally, the posterior probability that Mr. X prefers bacon pancakes over vanilla pancakes equals $.75$ – the size of the gray area under the curve.

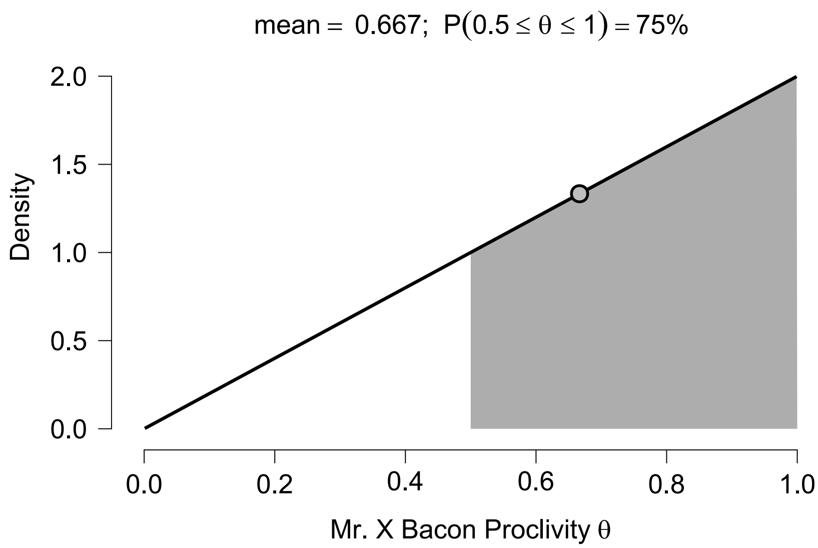


Figure 8.4: Posterior distribution for the unknown bacon proclivity of Mr. X, after observing a single bacon pancake. Figure from the JASP module *Learn Bayes*.

A posterior distribution can be summarized and queried in a myriad ways. One may report the posterior mean, mode, or median; one may report the posterior mass that lies in any interval of interest;[6] or one may specify a target amount of posterior mass and request an interval that contains that mass. One of the most popular posterior summary measures is the "95% credible interval", an interval that contains 95% of the posterior mass.

[6] Above, we were interested in $p(0.5 \leq \theta \leq 1)$, but we may enquire about $p(a \leq \theta \leq b)$ for any a and b as long as $0 \leq a < b \leq 1$. Note that $p(a \leq \theta \leq b)$ can also be written $p(\theta \in [a, b])$.

There are two popular types of 95% credible intervals. The first one is the *central* 95% credible interval, which is obtained by excluding 2.5% of posterior mass from both ends of the distribution, left and right. By construction, θ is just as likely to fall below the interval as it is to lie above it. Central credible intervals are sometimes called ‘equal-tailed intervals’. Figure 8.5 shows the 95% credible interval method as applied to the example of Mr. X. The interval ranges from $.158$ to $.987$ and contains 95% of the posterior mass. Note that the interval *excludes* that part of the posterior distribution which contains the most likely values of θ_X, namely the slice from $.987$ to 1.

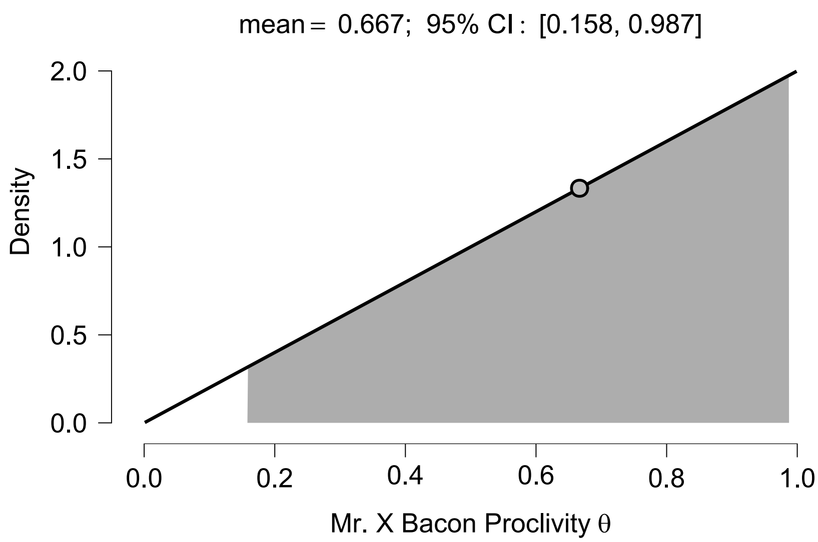


Figure 8.5: Central 95% credible interval of the posterior distribution for the unknown bacon proclivity of Mr. X, after observing a single bacon pancake. Figure from the JASP module *Learn Bayes*.

The second popular type of credible interval is the 95% *highest posterior density* (HPD) interval, which is defined as the smallest interval that contains 95% of posterior mass. Figure 8.6 shows the 95% HPD method as applied to the example of Mr. X. The interval ranges from $.224$ to 1 and contains 95% of the posterior mass. Note that this interval *includes* the part of the posterior distribution which contains the most likely values of θ.

Which type of 95% interval should you use? We don’t have a strong preference, and in most practical applications it does not matter much. When the two intervals do give very different results, it is prudent to *display the entire posterior distribution* rather than summarize it by a few numbers. When summary measures are used, no matter their sophistication or rationale, information is inevitably lost.

Regardless of what type of x% credible interval is being reported, its interpretation is the same: x% of the posterior mass falls in the specified interval from a to b. Hence, under the statistical model that is being entertained, and with the data in hand, you can be 95% certain that the parameter of interest lies between a and b. This is a direct, intuitive

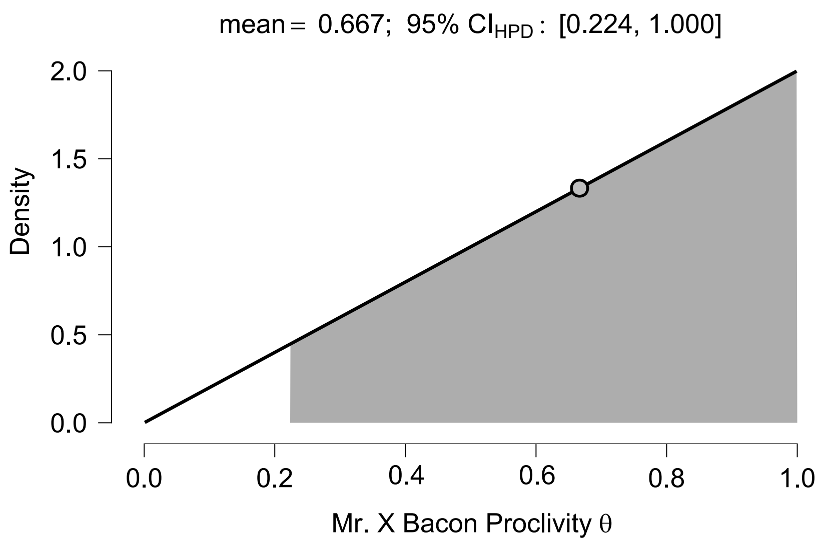


Figure 8.6: 95% highest posterior density interval of the posterior distribution for the unknown bacon proclivity of Mr. X, after observing a single bacon pancake. Figure from the JASP module *Learn Bayes*.

interpretation that is inappropriate for a frequentist 'confidence interval' (Morey et al. 2016a).[7]

[7] Briefly, a frequentist 95% confidence interval is generated by a procedure that, in repeated use across different data sets, encloses the true data-generating parameter value 95% of the time. Note that no reference can be made to the actual end-points of the interval. For frequentists, confidence refers to an evaluation of performance in repeated use, not to an assessment of plausibility for the individual case.

A Second Pancake from Mr. X

Mr. X now produces a second pancake and it's vanilla. We can now update our knowledge in two ways, which lead to exactly the same end result. The first method is to retain the uniform prior, and pretend that the two pancakes $\{b, v\}$ were seen at the same time. Doing this leads to the dome-shaped posterior distribution shown in Figure 8.7. The observation of a vanilla pancake has considerably reduced the previous enthusiasm for high values of θ_X, and the posterior mean is reduced to $.5$, the same value it had before any pancakes were observed. The posterior distribution is now symmetric around the value of $\theta = .5$ (which means that $.5$ is the posterior median); the posterior distribution also peaks on $\theta = .5$ (which means that $.5$ is the posterior mode). As can be seen from the size of the gray area, the posterior probability that Mr. X prefers bacon pancakes over vanilla pancakes equals $.50$. So in many ways we appear to be back where we started before any pancake was observed. However, a comparison between the flat prior distribution to the dome-shaped posterior distribution shows that, after two pancakes, middle values of θ_X have become more credible than they were before, whereas values lower than about $.20$ and higher than about $.80$ have become less credible.

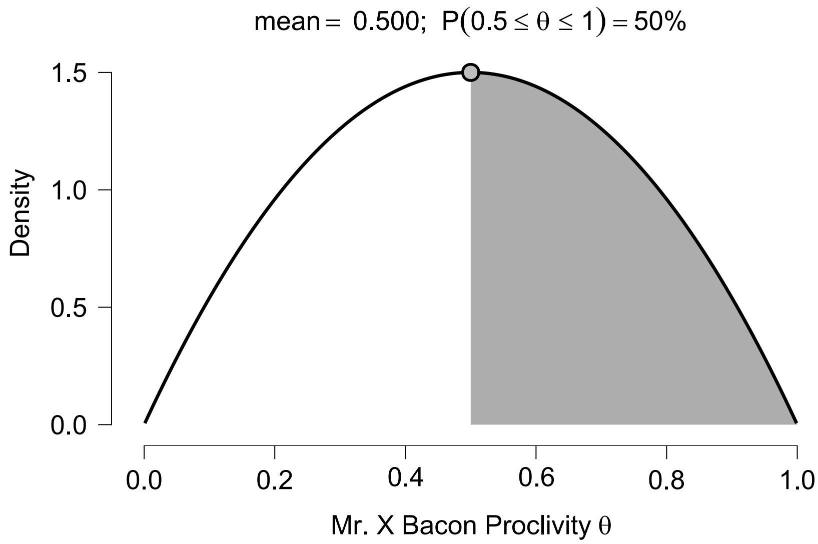


Figure 8.7: Posterior distribution for the unknown bacon proclivity of Mr. X, after observing one bacon pancake and one vanilla pancake. Figure from the JASP module *Learn Bayes*.

The second way of updating is more elegant. Instead of pretending to have observed the two pancakes simultaneously, we stay true to the sequential nature of how the data were obtained. Specifically, we first update our knowledge about θ_X based on having observed the bacon pancake, yielding the posterior distribution shown in Figure 8.4 (i.e., the ramp). Next, this posterior distribution then becomes our prior distribution for the second knowledge update, based on having observed the vanilla pancake. The end result is exactly the posterior distribution shown in Figure 8.7; it does not matter whether the data were analyzed simultaneously or sequentially. But how exactly do we set up the sequential analysis? In particular, how can we specify a prior distribution (prior to the observation of the second pancake) to be equal to a posterior distribution (posterior to the observation of the first pancake)?

The Beta Prior

In principle, θ –the unknown chance that any specific pancake comes with bacon– can be assigned a prior distribution at will, no matter how erratic, haphazard and idiosyncratic, just as long as it has area 1, and as long as it respects the fact that θ is defined on the interval from 0 to 1.

In practice, it is convenient to select a prior distribution from a flexible family of distributions whose shape can be adjusted by changing one or two parameters. And, as mentioned in the previous section, for

The Principle of Insufficient Reason

The Principle of Insufficient Reason, a term due to Laplace, is also known as the Principle of Indifference (Keynes 1921) or the Principle of Non-sufficient Reason (Jeffreys 1933a, p. 528). The principle holds that, when we have no ground for preferring one alternative over the other (i.e., when we are indifferent), the prior probabilities are taken to be equal. An example is to assign prior probability $1/11$ to each of the 11 candidate bakers in our pancake scenario. The Principle may appear self-evident. As stated by Jeffreys (1933a, p. 528): "The fundamental rule is the Principle of Non-sufficient Reason, according to which propositions mutually exclusive on the same data must receive equal probabilities if there is nothing to enable us to choose between them. This principle (…) seems to me so obvious as hardly to require statement" (see also Howie 2002, 148-150; Jeffreys 1931, p. 20). It is obvious in part because any other assignment of prior probabilities seems indefensible. Specifically, "if we do not take the prior probabilities equal we are expressing confidence in one rather than another before the data are available, and this must be done only from definite reason. *To take the prior probabilities different in the absence of observational reason for doing so would be an expression of sheer prejudice.*" (Jeffreys 1961, p. 33, italics ours; see also Jeffreys's 1934 letter to Fisher presented in Bennett 1990, p. 154).

Nevertheless, it has been argued that the blind application of the Principle of Insufficient Reason results in paradoxes (e.g., Eva 2019; Keynes 1921; Van Fraassen 1989, Chapter 12). For instance, when we are indifferent about a standard deviation σ we might be tempted to assign it a uniform distribution from 0 to ∞, such that every value of σ is deemed equally likely *a priori*. However, not only is this distribution *improper* (i.e., it does not have area 1), it also induces a non-uniform distribution on the variance σ^2, a quantity about which we might likewise be indifferent. These challenges were addressed by Jeffreys (1961, Chapter 3), a discussion of which would lead us too far afield. In modern Bayesian analysis, data analysts have adopted a more pragmatic approach, and this has reduced the relevance of philosophical debates concerning the Principle of Insufficient Reason.

sequential updating is desirable that the prior distribution for the nth observation can be specified to equal the posterior distribution after the $(n-1)$th observation.

For our problem concerning the chance θ, the standard choice is to select a prior distribution from the *beta* family. Beta distributions have two parameters; these are traditionally called a and b, but in this book we refer to them as α and β, in line with the convention to use the Greek alphabet for unobserved quantities and the Latin alphabet for observed quantities. Figure 8.8 shows four examples of beta distributions. The flat *green* line is the beta$(1, 1)$ distribution that we already encountered in Figure 8.3; this distribution indicates that every value of θ is equally plausible *a priori*. The *red* line is a beta$(1/2, 1/2)$ distribution, whose U-shape indicates that extreme values are more likely *a priori* than values in the middle of the range.[8] The *yellow* line is a beta$(10, 1)$ distribution, whose J-shape indicates that relatively high values of θ are deemed much more plausible than low values; values of θ lower than $1/2$ are relatively unlikely. Finally, the *blue* line is a beta$(10, 10)$ distribution. Its inverted-U shape indicates that values of θ in the middle of the range are more plausible than those in the extremes; specifically, values of θ lower than $.20$ and larger than $.80$ are relatively unlikely. We encourage the reader to explore different values for α and β and their effect on the shape of the beta distribution. In JASP, this can be done both from the *Learn Bayes* module ('Binomial Estimation') and from the *Distributions* module ('Continuous' $\rightarrow$ 'Beta').[9]

[8] The beta$(1/2, 1/2)$ distribution is known as 'Jeffreys's prior', but a discussion on its rationale is well beyond the scope of this textbook. Curious readers can find a tutorial-style explanation in Ly et al. (2017).

[9] A Shiny app to examine the shape of different beta distributions is available at `http://shinyapps.org/`, under "A first lesson in Bayesian inference".

In general, the following regularities can be observed about the shape of beta priors as parameters α and β are varied:

- Beta priors with $\alpha = \beta$ are symmetric around $\theta = 1/2$, and thus do not encode a prior preference for successes (e.g., bacon pancakes) over failures (e.g., vanilla pancakes).
- As α and β increase, the beta prior becomes more peaked, indicating more prior certainty about the plausible values of θ.
- When α and β are both large, the beta distribution is peaked around the value $\alpha/(\alpha+\beta)$, which is also the distribution's mean.
- When $\alpha > \beta$ (e.g., the yellow line in Figure 8.8), the prior distribution assigns more mass to values of θ greater than $1/2$, reflecting a prior preference for successes over failures; when $\beta > \alpha$, the prior distribution assigns more mass to values of θ lower than $1/2$, reflecting a prior preference for failures over successes.

These regularities concerning the beta prior suggest that parameter α can be interpreted as the hypothetical number of prior successes and parameter β can be interpreted as the hypothetical number of prior

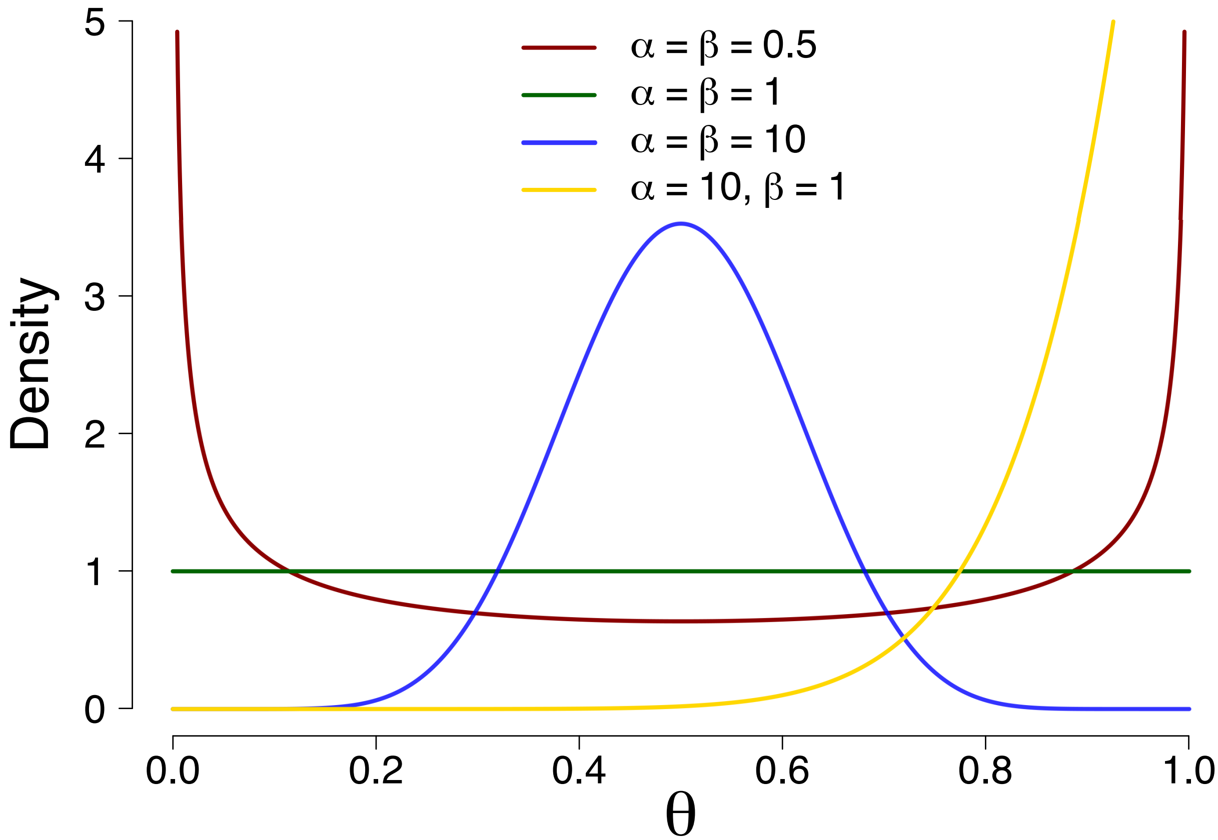


Figure 8.8: Example of four beta distributions that could be specified to capture one's uncertainty about the chance θ in advance of data collection. Parameter α can be interpreted as the hypothetical prior number of successes, and parameter β can be interpreted as the hypothetical prior number of failures (Jaynes 2003, pp. 385-386).

failures. To demonstrate that this suggestion is correct we now turn to the underlying mathematics.[10]

[10] There is an ongoing debate on whether α and β ought to be interpreted as the number of hypothetical prior success and failures, or as these numbers *minus one*. See the last exercise in this chapter.

Knowledge Updating with the Beta Prior

Having specified our prior knowledge about θ by means of a beta distribution, we are now ready to update this knowledge by means of the data. By Bayes' rule:

$$\begin{aligned} p(\theta \mid \text{data}) &= p(\theta) \cdot \frac{p(\text{data} \mid \theta)}{p(\text{data})} \\ &\propto p(\theta) \cdot p(\text{data} \mid \theta), \end{aligned}$$

where $\propto$ stands for 'is proportional to'.[11] As mentioned in Chapter 7,

[11] Recall that $p(\text{data})$ is a constant: a marginal likelihood that does not depend on θ.

Parameter or Hypothesis?

In the example of the 11 candidate bakers, it is intuitive to view each proclivity θ_i as a separate, rival *hypothesis* concerning the baker's identity. But when the number of bakers grows infinitely large and θ becomes continuous, convention dictates that θ is then called a *parameter*, not a space for an infinite number of hypotheses. Although the difference is linguistically convenient, it should be kept in mind that the distinction is merely that – a matter of linguistics (e.g., Good 1983, p. 126; Gelman 2011, p. 76; Gronau and Wagenmakers 2019). In particular, the Bayesian rules for updating knowledge do not depend on whether θ called a hypothesis (in the discrete case) or a parameter (in the continuous case).

this means (Jeffreys 1939, p. 46):

$$\text{Posterior} \propto \text{ Prior} \times \text{Likelihood.} \tag{8.1}$$

Firstly, consider the beta prior:

$$\begin{aligned} p(\theta) &\sim \text{beta}(\alpha, \beta) \\ &\propto \theta^{\,\alpha-1}(1-\theta)^{\,\beta-1}. \end{aligned} \tag{8.2}$$

The complete expression for the beta distribution contains an additional term, but because this term is a constant that does not involve θ we can omit it from the equation – for the current explanation we only need the result in proportional form. Note that entertaining $\alpha = \beta = 1$ produces the flat prior (i.e., the green line in Figure 8.8).

Secondly, consider the *binomial* likelihood, that is, the predictive performance of particular θ for the observed number of bacon and vanilla pancakes. For example, consider again our pancake sequence from Chapter 7: $\{b, v, b, b, b, v\}$. The probability of this exact sequence is $\theta \times (1-\theta) \times \theta \times \theta \times \theta \times (1-\theta) = \theta^4 \times (1-\theta)^2$. In general, the probability of the exact observed sequence containing s successes and f failures is $\theta^s \times (1-\theta)^f$.

At this point it may be tempting to define the binomial likelihood as $p(s, f \mid \theta) = \theta^s \times (1-\theta)^f$. But this is not quite correct. That probability is for the exact sequence $\{b, v, b, b, b, v\}$; but the data summary $s = 4, f = 2$ is also consistent with 14 *other* sequences, including $\{b, b, v, b, b, v\}$, $\{b, b, b, b, v, v\}$, and so forth. Hence, for the case of $s = 4, f = 2$ the likelihood is given by $p(s = 4, f = 2 \mid \theta) = 15 \times \theta^4 \times (1-\theta)^2$, where 15 represents the number of possible sequences. But because that single number does not involve the parameter θ, we can write the binomial likelihood as follows:

$$\begin{aligned} p(s, f \mid \theta) &= \text{binomial}(s, f \mid \theta) \\ &\propto \theta^{s}(1-\theta)^{f}. \end{aligned} \tag{8.3}$$

This likelihood is clearly of a form similar to the beta prior shown in Equation 8.2. Multiplying beta prior and binomial likelihood we obtain a posterior distribution proportional to $\theta^{\alpha-1} \times (1-\theta)^{\beta-1} \times \theta^{s} \times (1-\theta)^{f} = \theta^{a+s-1} \times (1-\theta)^{b+f-1}$. This posterior quantity can be recognized as proportional to another beta distribution – specifically, a beta$(\alpha+s, \beta+f)$ distribution.

Consequently, the tinkering above has provided the following helpful rule: *if* we define our prior beliefs about a binomial chance parameter θ by a beta(α, β) distribution, and *if* we observe binomial data constituting of s successes and f failures, *then* our updated beliefs are quantified by a posterior distribution which is also a beta distribution, just like the prior, but now with parameters beta$(\alpha+s, \beta+f)$. This is so convenient, and so important, that it deserves a separate equation:

$$\underbrace{p(\theta \mid s, f)}_{\substack{\text{Posterior for } \theta: \\ \text{beta}(\alpha+s, \beta+f)}} \propto \underbrace{p(\theta)}_{\substack{\text{Prior for } \theta: \\ \text{beta}(\alpha, \beta)}} \times \underbrace{p(s, f \mid \theta)}_{\substack{\text{Probability for } s, f \\ \text{given } \theta}} \tag{8.4}$$

This property –that the prior distribution and the posterior distribution are in the same family, making the updating process intuitive and convenient– is called *conjugacy*.[12] Unfortunately, more complicated models are rarely conjugate.

Reflecting on the fact that a beta(α, β) prior distribution, updated with s successes and f failures, yields a beta$(\alpha+s, \beta+f)$ posterior distribution produces a number of insights:

- The order in which the observations have arrived does not influence the inference. Ultimately all that matters is the number of successes and failures. Their order is of no import (Jeffreys 1938d, p. 444; Jeffreys 1938a, pp. 191-192).

- It does not matter whether data are analyzed simultaneously or sequentially. Again, all that matters is the final number of successes and failures.

- As s and f increase, they will start to dominate α and β. This means that, as far as the location and shape of the posterior distribution is concerned, the impact of the prior distribution is increasingly watered down as the data accumulate. This is sometimes described by the phrase 'the data overwhelm the prior'.[13]

[12] Although few people are familiar with the concept of conjugacy ('connected'; literally: 'yoked together'), many more will be familiar with the term 'conjugal visit'.

[13] Wrinch and Jeffreys (1919).

- Suppose there exists a true value for θ, denoted $\theta^\star$. As the data accumulate the posterior will be increasingly peaked, and the mean of the posterior distribution, which is $(\alpha+s)/(\alpha+s+\beta+f)$ will become arbitrarily close to $s/(s+f)$, the value corresponding with $\theta^\star$. This suggests that the posterior distribution will converge to $\theta^\star$ (a suggestion that was proven by Laplace 1774/1986).[14]

[14] In statistical jargon, this property is called *consistency*.

Figure available at `BayesianSpectacles.org` under a CC-BY license.

Mr. X Revisited

Armed with newfound knowledge about the beta prior and about conjugacy, we briefly return to the scenario of estimating the bacon proclivity θ_X of Mr. X. We started with a uniform prior distribution (cf. Figure 8.3) and after the first pancake (which was bacon) our knowledge was updated to a posterior distribution that resembled a ramp (cf. Figure 8.4). We now know that the uniform prior distribution is a beta$(1, 1)$, and that the posterior distribution is a beta$(2, 1)$. We then observed a second pancake (which was vanilla) and updated our beta$(1, 1)$ distribution all at once with both observations, yielding a dome-shaped posterior (cf. Figure 8.7). We now know that this dome-shaped posterior is a beta$(2, 2)$. In addition, we now have an answer to the question how we can analyze the data from Mr. X sequentially, one pancake after the other. After the first pancake is observed, our knowledge is reflected in a beta$(2, 1)$ posterior. It is this posterior that should be our prior distribution as we await the second pancake. When that second pancake arrives, we update to a beta$(2, 2)$ distribution, and we end up with the same inference that we did when the data were analyzed all at once. Figure 8.9 visualizes the second sequential updating step.

Exercises

1. Based on the information in Table 8.1, compute the likelihood ratio for Denver versus Lennon.

2. Construct Figure 8.2 (i.e., the 11-baker plot) with the *Learn Bayes* module (under Binomial Testing).

3. Imagine that instead of 1 bacon pancake, we observe a stack of 20 pancakes, 10 of which are vanilla and 10 of which have bacon. What general conclusion can we draw about the relative plausibility of the bakers? Confirm your intuition with the *Learn Bayes* module.

4. Suppose we entertain a large number of plausible hypotheses. One of the hypotheses provides the best prediction for the observed data. Explain how the Bayesian paradigm tempers the enthusiasm for this best-predicting model.

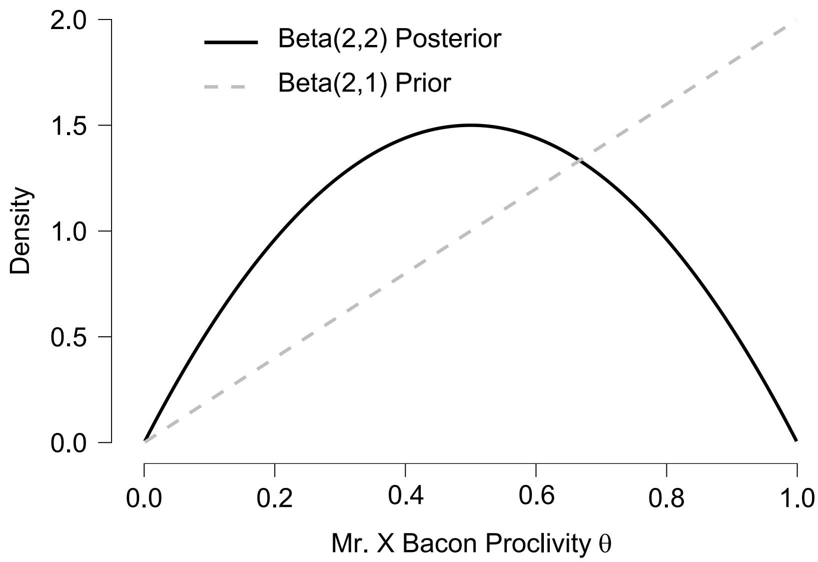


Figure 8.9: Sequential analysis of the unknown bacon proclivity of Mr. X. The dotted gray line represents a beta(2,1) distribution, which is posterior to the occurrence of the first pancake but prior to the occurrence of the second pancake. After observing the second pancake, the beta$(2, 1)$ distribution is updated to a beta$(2, 2)$ distribution, represented by the black line. Figure from the JASP module *Learn Bayes*.

5. Consider again Figure 8.4. Use the *Learn Bayes* module to confirm that the posterior median is $.707$. For further confirmation, what credible interval would you need to show?

6. Consider Figure 8.3 (i.e., the uniform prior) and Figure 8.4 (i.e., the posterior ramp). What is the evidence, obtained from observing a single bacon pancake, that $\theta_X > .50$?

7. Suppose we start with the beta$(1, 1)$ prior distribution for the bacon proclivity for a Mr. Y (the green line in Figure 8.8), and we end up with a beta$(10, 1)$ posterior distribution (the yellow line). What pancakes did Mr. Y produce?

8. After observing one bacon and one vanilla pancake, we wrote that "middle values of θ_X have become more credible than they were before, whereas values lower than about $.20$ and higher than about $.80$ have become less credible". Use the *Learn Bayes* module to obtain the exact numbers. [hint: use the support interval (Wagenmakers et al. 2022)].

9. The statistical framework outlined in the previous chapters can be applied widely. Describe how you would apply it to the following problems[15]:

[15] NB. Neither problem involves any pancakes.

9.1. How much of the earth's surface is covered by water? The only objects at your disposal are a globe, a pencil, and piece of paper.[16]

9.2. What is the median speed of flowing traffic on the highway closest to where you live? You have at your disposal a car, a driver (who obeys your instructions), a pencil, and a piece of paper.

10. "**Bayesian**: One who, vaguely expecting a horse and catching a glimpse of a donkey, strongly concludes he has seen a mule." (Senn 2007, p. 46). Discuss.

11. Assume you update a beta($1/2, 1/2$) prior distribution for θ with a single success and a single failure. What does the posterior distribution look like?

12. Amy assigns a beta($\alpha = 8, \beta = 2$) prior distribution to a chance θ. What number of hypothetical previously seen successes and failures does this prior distribution correspond to?

[16] To the best of our knowledge, this example application was first suggested by Richard McElreath.

Chapter Summary

In this chapter we demonstrated how to update beliefs about an infinite number of hypotheses. We first expanded our set of candidate bakers (i.e., rival hypotheses or possible *causes*) from 2 to 11. In the limit of an infinite number of candidate bakers, each associated with a unique value for their bacon proclivity parameter θ, we obtain a continuous distribution. This continuous distribution may be summarized by a central tendency (e.g., the mean) and a measure of its spread or width (e.g., an x% credible interval, which contains x% of the distribution mass). For inference concerning chances, a convenient choice is the beta distribution: a beta(α, β) prior distribution, when updated with s successes and f failures, yields a beta($\alpha + s, \beta + f$) posterior distribution. This shows that the order of the observations is irrelevant, as is the choice of whether to analyze the data sequentially or all at once.

Want to Know More?

✓ Albert, J. M. (2007). *Bayesian Computation with R*. New York: Springer. This book interweaves conceptual explanation with concrete application – and all analyses are supported with concise R scripts.

✓ Bolstad, W. M. (2007). *Introduction to Bayesian Statistics (2nd ed.)*. Hoboken, NJ: Wiley. Prior to writing the book you are reading now, Bolstad was our go-to reference for students needing a gentle introduction to Bayesian inference.

✓ Etz, A. (2018). Introduction to the concept of likelihood and its applications. *Advances in Methods and Practices in Psychological Science, 1*, 60-69. Alexander Etz is an exceptionally clear writer.

✓ Kruschke, J. K. (2015). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan (2nd ed.).* Academic Press/Elsevier. This book has greatly helped popularize Bayesian inference, especially in the field of psychology. It has puppies on the cover.

✓ Kurt, W. (2019). *Bayesian Statistics the Fun Way.* San Francisco: No Starch Press. We have recommended this introductory treatment in an earlier chapter, and we are re-issuing our recommendation here.

✓ Stone, J. V. (2016). *Bayes' Rule with R: A Tutorial Introduction to Bayesian Analysis.* Sebtel Press. A concise, well-presented introduction, with R code.

Appendix: A Simple Illustration of Bayesian Inference, by Jevons (1874)

This appendix is also presented, with minor changes, in Gronau and Wagenmakers (2019).

Jevons' 1874 masterpiece *The Principles of Science* contains the section 'Simple Illustration of the Inverse Problem' that showcases Bayesian updating and posterior prediction for the case of multiple discrete hypotheses. For historical interest, and out of respect for the clarity of Jevons' writing, we present the section in full:[17]

[17] For a modern-day account, see D'Agostini (1999) and other works by the same author.

> "Suppose it to be known that a ballot-box contains only four black or white balls, the ratio of black and white balls being unknown. Four drawings having been made with replacement, and a white ball having appeared on each occasion but one, it is required to determine the probability that a white ball will appear next time. Now the hypotheses which can be made as to the contents of the urn are very limited in number, and are at most the following five:–
>
> 4 white and 0 black balls
> 3 ,, ,, 1 ,, ,,
> 2 ,, ,, 2 ,, ,,
> 1 ,, ,, 3 ,, ,,
> 0 ,, ,, 4 ,, ,,
>
> The actual occurrence of black and white balls in the drawings renders the first and last hypotheses out of the question, so that we have only three left to consider.
>
> If the box contains three white and one black, the probability of drawing a white each time is $\frac{3}{4}$, and a black $\frac{1}{4}$; so that the compound event observed, namely, three white and one black, has the probability $\frac{3}{4} \times \frac{3}{4} \times \frac{3}{4} \times \frac{1}{4}$, by the rule already given (p. 233).[18] But as it is indifferent to us in what order the balls are drawn, and the black ball might come

[18] The relevant text on p. 233 reads: "When the component events are independent, a simple rule can be given for calculating the probability of the compound event, thus—*Multiply together the fractions expressing the probabilities of the independent component events.*" [italics in original]

first, second, third, or fourth, we must multiply by four, to obtain the probability of three white and one black in any order, thus getting $\frac{27}{64}$.

Taking the next hypothesis of two white and two black balls in the urn, we obtain for the same probability the quantity $\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times 4$, or $\frac{16}{64}$, and from the third hypothesis of one white and three black we deduce likewise $\frac{1}{4} \times \frac{1}{4} \times \frac{1}{4} \times \frac{3}{4} \times 4$, or $\frac{3}{64}$. According, then, as we adopt the first, second, or third hypothesis, the probability that the result actually noticed would follow is $\frac{27}{64}$, $\frac{16}{64}$, and $\frac{3}{64}$. Now it is certain that one or other of these hypotheses must be the true one, and their absolute probabilities are proportional to the probabilities that the observed events would follow from them (see p. 279).[19] All we have to do, then, in order to obtain the absolute probability of each hypothesis, is to alter these fractions in a uniform ratio, so that their sum shall be unity, the expression of certainty. Now since $27 + 16 + 3 = 46$, this will be effected by dividing each fraction by 46 and multiplying by 64. Thus the probability of the first, second, and third hypotheses are respectively—

$$\frac{27}{46}, \quad \frac{16}{46}, \quad \frac{3}{46}.$$

The inductive part of the problem is now completed, since we have found that the urn most likely contains three white and one black ball, and have assigned the exact probability of each possible supposition. But we are now in a position to resume deductive reasoning, and infer the probability that the next drawing will yield, say a white ball.[20] For if the box contains three white and one black ball, the probability of drawing a white one is certainly $\frac{3}{4}$; and as the probability of the box being so constituted is $\frac{27}{46}$, the compound probability that the box will be so filled and will give a white ball at the next trial, is

$$\frac{27}{46} \times \frac{3}{4} \text{ or } \frac{81}{184}.$$

Again, the probability is $\frac{16}{46}$ that the box contains two white and two black, and under those conditions the probability is $\frac{1}{2}$ that a white ball will appear; hence the probability that a white ball will appear in consequence of that condition, is

$$\frac{16}{46} \times \frac{1}{2} \text{ or } \frac{32}{184}.$$

From the third supposition we get in like manner the probability

$$\frac{3}{46} \times \frac{1}{4} \text{ or } \frac{3}{184}.$$

Now since one and not more than one hypothesis can be true, we may add together these separate probabilities, and we find that

$$\frac{81}{184} + \frac{32}{184} + \frac{3}{184} \text{ or } \frac{116}{184}$$

is the complete probability that a white ball will be next drawn under the conditions and data supposed." (Jevons 1874/1913, pp. 292-294)

In the next section, *General Solution of the Inverse Problem*, Jevons points out that in order for the procedure to be applied to natural phenomena, an infinite number of hypotheses need to be considered:

[19] Note from the authors: this assumes that the hypotheses are equally likely a priori. The relevant text on p. 279 reads: "*If an event can be produced by any one of a certain number of different causes, the probabilities of the existence of these causes as inferred from the event, are proportional to the probabilities of the event as derived from these causes.*" [italics in original]

[20] EWDM: Note that when the possible content of each ballot-box is considered a *parameter*, this forecast is known as a 'posterior prediction'; when the possible content is interpreted as a competing hypothesis, the same forecast is known as 'Bayesian model averaging' (e.g., Hinne et al. 2020, Gronau and Wagenmakers 2019), see Chapter 7.

"When we take the step of supposing the balls within the urn to be infinite in number, the possible proportions of white and black balls also become infinite, and the probability of any one proportion actually existing is infinitely small. Hence the final result that the next ball drawn will be white is really the sum of an infinite number of infinitely small quantities. It might seem, indeed, utterly impossible to calculate out a problem having an infinite number of hypotheses, but the wonderful resources of the integral calculus enable this to be done with far greater facility than if we supposed any large finite number of balls, and then actually computed the results. I will not attempt to describe the processes by which Laplace finally accomplished the complete solution of the problem. They are to be found described in several English works, especially De Morgan's 'Treatise on Probabilities,' in the 'Encyclopædia Metropolitana,' and Mr. Todhunter's 'History of the Theory of Probability.' The abbreviating power of mathematical analysis was never more strikingly shown. *But I may add that though the integral calculus is employed as a means of summing infinitely numerous results, we in no way abandon the principles of combinations already treated*.[italics ours]" (Jevons 1874/1913, p. 296)

9 *The Rule of Succession*

If there have been m occasions on which a certain event might have been observed to happen, and it has happened on all those occasions, then the probability that it will happen on the next occasions of the same kind is $\frac{m+1}{m+2}$.

Jevons, 1874

Chapter Goal

The goal is to derive Laplace's Rule of Succession and set up the proper understanding for the next chapter.

The Beta Prediction Rule

Suppose a binomial chance θ has a beta distribution, that is, $\theta \sim \text{beta}(\alpha, \beta)$. An example of a beta distribution with parameters $\alpha = 4$, $\beta = 6$ is shown in Figure 9.1. Using the information in the beta distribution, we now wish to predict the outcome of the next binomial trial – what is the probability that it will be a success?[1]

[1] In the pancake example, successes and failures were defined as occurrences of bacon and vanilla pancakes, respectively. The term 'success' and 'failure' is more generic. In the following, we denote a success by '1' and a failure by '0'.

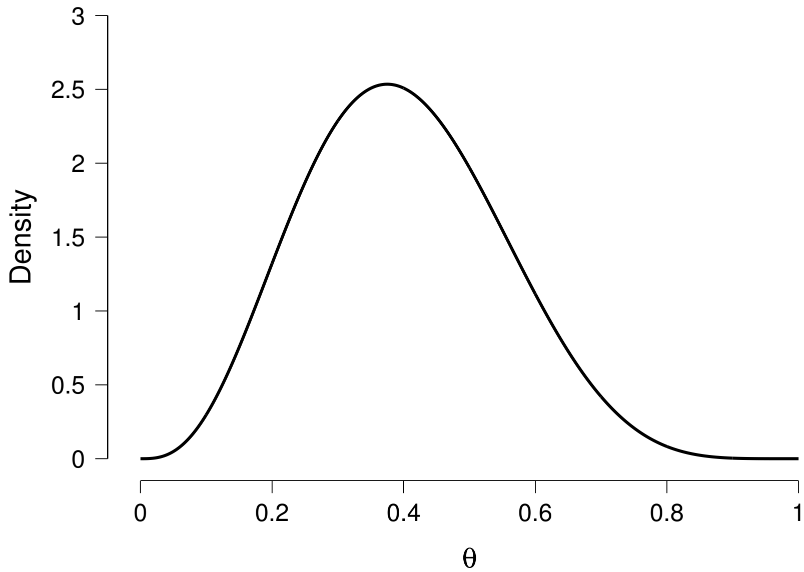


Figure 9.1: A beta(α=4, β=6) distribution for a binomial success parameter θ.

What we know is the probability of a success *given* a particular value of θ: this is simply θ. For instance, if we know that Andy has a proclivity for producing bacon pancakes that equals $\theta = .40$, then the probability that the next pancake contains bacon is .40. Therefore, $p(y = 1 \mid \theta) = \theta$, where $y = 1$ stands for the next observation y being a success (i.e., a bacon pancake). But we wish to make an overall statement, a prediction that takes into account all possible values of θ, weighted with the plausibility as provided by the beta distribution. In other words, we need to average out θ according to the *law of total probability*, as explained in Chapter 3.

Now if θ were composed of n discrete possibilities, we would obtain our prediction by computing a weighted average:

$$p(y = 1) = \sum_{i=1}^{n} p(y = 1 \mid \theta_i)\, p(\theta_i).$$

This process is essential, so we will drive this point home. Suppose θ is composed of just $n = 2$ discrete possibilities: $\theta_A = .40$ and $\theta_B = .80$. Furthermore, suppose the prior distribution on θ assigns probability .36 to θ_A and probability .64 to θ_B. This discrete, two-point prior distribution across θ is displayed in Figure 9.2.

Stamp "Laplace" (N° Yvert & Tellier 1031) by Paul-Pierre Lemagny. Reproduced with permisson of ©La Poste and Rosine Gosset-Lemagny.

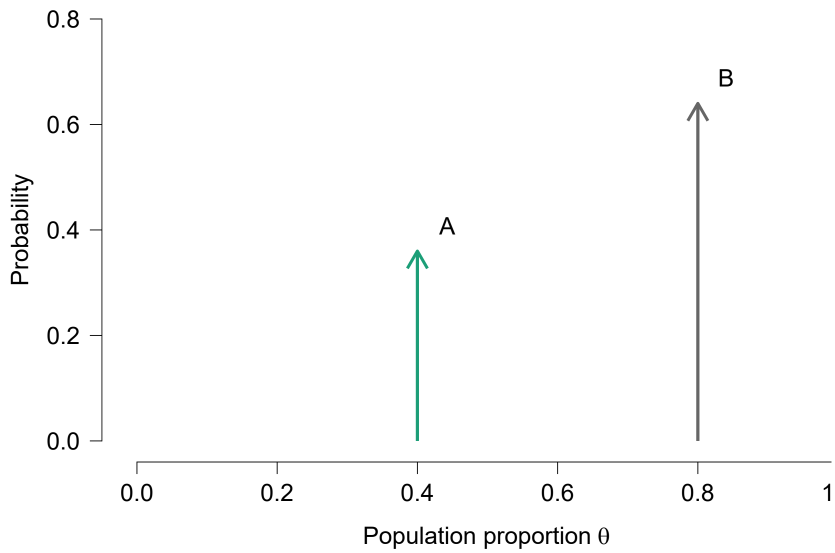


Figure 9.2: A discrete, two-point prior distribution for a chance θ, assigning prior mass .36 and .64 to $\theta_A = .40$ and $\theta_B = .80$, respectively. Applying the law of total probability yields the probability that the next observation will be a success. See text for details. Figure from the JASP module *Learn Bayes*.

By applying the law of total probability we can issue a prediction that accounts for our uncertainty about the possible values of θ:

$$\begin{aligned} p(y=1) &= \sum_{i=1}^{n} p(y=1 \mid \theta_i)\, p(\theta_i) \\ &= p(y=1 \mid \theta_A)\, p(\theta_A) + p(y=1 \mid \theta_B)\, p(\theta_B) \\ &= p(y=1 \mid \theta=.40)\, .36 + p(y=1 \mid \theta=.80)\, .64 \\ &= .40 \times .36 + .80 \times .64 \approx .656. \end{aligned}$$

These numbers are in fact *identical* to those used in Chapter 7, when we predicted whether or not the seventh pancake would have bacon, averaging across the uncertainty about the identity of the baker (i.e., either Andy, with a bacon proclivity of $\theta = .40$, or Bobbie, with a bacon proclivity of $\theta = .80$). The associated tree diagram was presented as Figure 7.4.

However, the beta distribution is continuous and this means that we need to compute an integral instead of a sum, as follows:

$$\begin{aligned} p(y=1) &= \int_0^1 p(y=1 \mid \theta)\, p(\theta)\, \mathrm{d}\theta \\ &= \frac{\alpha}{\alpha+\beta}. \end{aligned} \tag{9.1}$$

As it turns out, the integral across θ yields a surprisingly simple result: the required probability is $\alpha/(\alpha+\beta)$, which is in fact just the mean of a beta(α, β) distribution.[2] For example, for the beta(α=4, β=6) distribution shown in Figure 9.1, weighted predictions across the different values of θ integrate to $4/(4+6) = .40$. This shortcut can be used to solve a series of historically important problems with relative ease.

[2] The appendix to this chapter provides three related ways to derive the result mathematically.

Example 1: Update & Predict

Suppose we assign θ a beta prior distribution with parameters $\alpha = 2$ and $\beta = 2$; we then observe $s = 2$ successes and $n - s = 4$ failures. What is the probability of a success on the seventh trial?

The solution proceeds in two steps. First, we use conjugacy to update our beta prior, resulting in a beta posterior: $p(\theta \mid s, n) \sim$ beta$(\alpha + s, \beta + n - s) =$ beta$(4, 6)$. Not coincidentally, it is this posterior distribution that is shown in Figure 9.1. Second, we apply the prediction rule from Equation 9.1 and this yields

$$p(y = 1 \mid s, n) = \frac{\alpha + s}{\alpha + s + \beta + n - s} = \frac{\alpha + s}{\alpha + \beta + n}, \tag{9.2}$$

showing that when the information in the sample (i.e., s and n) dominates the information in the prior (i.e., α and β), the prediction will be relative close to the sample proportion s/n. Plugging in our prior values $\alpha = \beta = 2$ and our sample values $s = 2, n = 6$ yields a prediction that the seventh trial is a success of $4/10 = .40$.

Example 2: Laplace's Rule of Succession

Laplace's famous Rule of Succession, stated by Jevons in the epigraph to this chapter, follows from Equation 9.1 when θ is assigned a uniform prior distribution (i.e., $\alpha = \beta = 1$) and the sample consists of only successes (i.e., $s = n$). In this case, we obtain:

$$p(y = 1 \mid s = n) = \frac{s + 1}{s + 2}.$$

Jevons (1874/1913, pp. 299-300) describes the relevance of the Rule of Succession as follows:

> "When an event has happened a very great number of times, its happening once again approaches nearly to certainty. Thus if we suppose the sun to have risen demonstratively one thousand million times, the probability that it will rise again, on the ground of this knowledge merely, is $\frac{1{,}000{,}000{,}000+1}{1{,}000{,}000{,}000+1+1}$. But then the probability that it will continue to rise for as long a period as we know it to have risen is only $\frac{1{,}000{,}000{,}000+1}{2{,}000{,}000{,}000+1}$, or almost exactly 1/2. The probability that it will continue so rising a thousand times as long is only about $\frac{1}{1001}$. The lesson which we may draw from these figures is quite that which we should adopt on other grounds, namely that experience never affords certain knowledge, and that it is exceedingly improbable that events will always happen as we observe

Will the Sun Rise Tomorrow?

In 'Philosophical essay on probabilities', Pierre-Simon Laplace provides a famous example of his Rule of Succession:

> "Thus one finds that when an event has happened any number of times running, the probability that it will happen again next time is equal to this number increased by 1, divided by the same number increased by 2. For example, if we place the dawn of history at $5{,}000$ years before the present date, we have $1{,}826{,}213$ days on which the sun has constantly risen in each 24 hour period. We may therefore lay odds of $1{,}826{,}214$ to 1 that it will rise again tomorrow. But this number would be incomparably greater for one who, perceiving in the coherence {or totality} of phenomena the principle regulating days and seasons, sees that nothing at the present moment can check the sun's course." (Laplace 1814/1995, p. 11)

This example is easy to critique, but only if one conveniently forgets Laplace's final sentence, and the fact that it is likely inspired by Hume, who repeatedly brought up the example of the sun rising (Diaconis and Skyrms 2018, p. 103; Zabell 1989).

The example of the sun rising was also discussed by Richard Price, in an appendix to Thomas Bayes' famous 1763 article 'An Essay towards Solving a Problem in the Doctrine of Chances'. After going over an example calculation, Price cautions:

> "It should be carefully remembered that these deductions suppose a previous total ignorance of nature. After having observed for some time the course of events it would be found that the operations of nature are in general regular, and that the powers and laws which prevail in it are stable and parmanent [sic]. The consideration of this will cause one or a few experiments often to produce a much stronger expectation of success in further experiments than would otherwise have been reasonable; just as the frequent observation that things of a sort are disposed together in any place would lead us to conclude, upon discovering there any object of a particular sort, that there are laid up with it many others of the same sort. It is obvious that this, so far from contradicting the foregoing deductions, is only one particular case to which they are to be applied." (Richard Price, 1763, in the appendix to Bayes 1763, p. 410)

> them. Inferences pushed far beyond their data soon lose any considerable probability."

Example 3: Laplace's Rule of Succession for Series

Given a uniform prior on θ, and an unbroken sequence of past successes, the Rule of Succession provides the probability that the next single event is again a success. But what if we wish to know the probability that the next k trials are also an unbroken sequence of successes? This generalizes the Rule from predicting a single success to a string of k successes. As summarized by Jevons (1874/1913, pp. 297-298)[3]:

> "*To find the probability that an event which has not hitherto failed will not fail for a certain number of new occasions, divide the number of times the event has happened increased by one, by the same number increased by one and the number of times it is to happen.* An event having happened s times without fail, the probability that it will happen k more times is $\frac{s+1}{s+k+1}$."

[3] We have changed Jevons's notation to be consistent with that used in this book.

Thus, the probability for an unbroken string of k successes is

$$\frac{s+1}{s+k+1},$$

a probability that decreases towards zero as the desired sequence k grows large (cf. Jeffreys 1973, Appendix II). This reveals that the Laplace method of inference is built on the assumption that no general law can be absolutely true, and exceptions are certain to arise if the observer is sufficiently patient. But, as Hume already wrote decades before Laplace:

> "One wou'd appear ridiculous, who wou'd say, that 'tis only probable the sun will rise to-morrow, or that all men must dye; tho' 'tis plain we have no further assurance of these facts, than what experience affords us." (Hume 1739)

In other words, is it really 'common sense expressed in numbers' –as Laplace liked to describe his method– to assume that we believe that we will eventually discover a person who is in fact immortal, if only we search long enough? This conundrum remained unaddressed for almost 150 years, until Dorothy Wrinch and Harold Jeffreys proposed a way to adapt the Laplacean system to overcome this limitation. But this will be the topic of future chapters in this book.

Example 4: Laplace's Rule of Succession from Mixed Past Experience

Another way to generalize the Rule of Succession that yields a clean result is to assume that the past is not an unbroken series of successes, but a mix of s successes and f failures. As summarized by Jevons (1874/1913, p. 298):

> "*An event having happened and failed a certain number of times, to find the probability that it will happen the next time, divide the number of times the event has happened increased by one, by the whole number of times the event has happened or failed to happen increased by two.* Thus, if an event has happened s times and failed f times, the probability that it will happen on the next occasion is $\frac{s+1}{s+f+2}$."

Thus, the probability that the next trial is a success after having experienced s successes and f failures is

$$\frac{s+1}{s+f+2}.$$

Comparison to Equation 9.2 shows that this rule is, again, based on assuming a uniform distribution on θ (i.e., $\alpha = \beta = 1$).

More intricate prediction problems can be proposed; for instance, one might wish to obtain the probability, from mixed past experience, of an unbroken sequence of k successes. More generally still, one might seek the probability, from the combination of any beta(α, β) prior and mixed past experience (i.e., s successes and f failures), of a mixed sequence consisting of k successes out of m future trials. As described in the appendix to this chapter, these probabilities follow from the beta-binomial distribution.[4]

[4] Specifically, given any beta$(\alpha + s,$ $\beta + f)$ posterior distribution on θ, the probability of future k successes out of m trials is a ratio of beta functions, $\binom{m}{k}\mathrm{B}(\alpha + s + k, \beta + f + m - k)/\mathrm{B}(\alpha + s, \beta + f)$, as discovered already by Laplace (e.g., Laplace 1774/1986, p. 365; Stigler 1986b).

We can conveniently analyze such problems with the *Learn Bayes* module in JASP. For instance, suppose we assign the chance θ a beta$(\alpha = 2, \beta = 2)$ prior distribution and observe $s = 2$ successes and $f = 4$ failures, yielding a beta(4,6) posterior distribution for θ. Desired is the predicted number of successes in the next 100 trials. To obtain these predictions from JASP, open the *Learn Bayes* module and select 'Counts' $\rightarrow$ 'Binomial Estimation'. Enter the observed data and specify the prior distribution. Then open the 'Posterior prediction' tab and enter '100' in the field 'Future observations'. The result is shown in Figure 9.3 by the wide gray predictive distribution labeled 'Epistemic + Aleatory'. For comparison, the narrow green predictive distribution labeled 'Aleatory' yields the predictions from a model in which the chance parameter θ is assumed to equal $.40$ exactly. With a relatively wide beta(4,6) posterior distribution for θ, there is considerable epistemic uncertainty; this uncertainty propagates to the predictive distribution, making it much wider than the one that reflects only aleatory uncertainty (cf. Chapter 2).

Exercises

1. Prove Laplace's Rule of Succession for series (Example 3 above).

2. A coin is tossed twice. The uncertainty about the chance θ of the coin landing heads is quantified by a beta(α, β) distribution. What

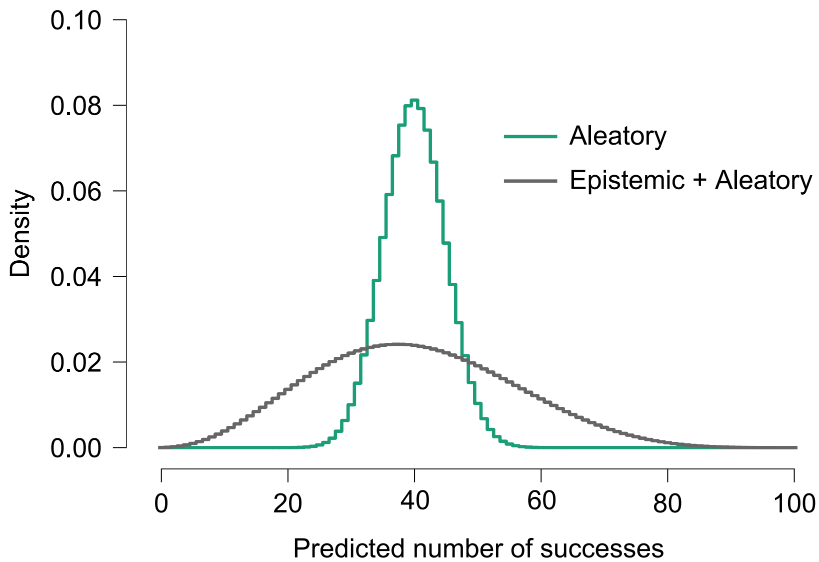


Figure 9.3: Predictions for the number of successes in the next 100 trials, based on the mixed past experience scenario described in the main text. The 'aleatory' curve is based on the assumption that the binomial chance θ equals .40 exactly. The 'epistemic + aleatory' curve includes epistemic uncertainty about θ as expressed in a beta(4,6) posterior distribution. This added uncertainty is reflected in predictions that are more spread out. Figure from the JASP module *Learn Bayes*.

is the probability that the coin comes up heads on both tosses? (cf. Jevons 1874/1913, p. 301; Laplace 1774/1986, p. 378; Todhunter 1865, p. 472)

3. A chance θ is assigned a prior beta(α, β) distribution. A single datum is observed, and the resulting posterior distribution is either a beta$(\alpha+1, \beta)$ distribution (when the observation shows a success) or a beta$(\alpha, \beta+1)$ distribution (when the observation shows a failure). Both posterior distributions intersect the prior distribution once, at the point where $\theta = {}^{\alpha}/_{(\alpha+\beta)}$. Confirm this visually with a concrete example, and use the beta prediction rule to explain why this has to be the case.

Chapter Summary

"The grand object of seeking to estimate the probability of future events from past experience, seems to have been entertained by James Bernouilli and De Moivre, at least such was the opinion of Condorcet; and Bernouilli may be said to have solved one case of the problem.[5] The English writers Bayes and Price are, however, undoubtedly the first who put forward any distinct rules on the subject.[6] Condorcet and several other eminent mathematicians advanced the mathematical theory of the subject; but it was reserved to the immortal Laplace to bring to the sub-

[5] Todhunter's 'History,' pp. 378, 79.

[6] 'Philosophical Transactions' [1763], vol. liii. p. 370, and [1764], vol. liv. p. 296. Todhunter, pp. 294-300.

ject the full power of his genius, and carry the solution of the problem almost to perfection." (Jevons 1874/1913, p. 302)

Want to Know More?

✓ Bayes, T. (1763). An Essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London, 53*, 370–418.

This essay, which was published posthumously by initiative of Bayes' friend Richard Price, unquestionably marks the birth of Bayesian inference. At the start, Bayes states his main objective:

> "*Given* the number of times in which an unknown event has happened and failed: *Required* the chance that the probability of its happening in a single trial lies somewhere between any two degrees of probability that can be named." (Bayes 1763, p. 376)

In other words, Bayes aims to find a rule of succession. His efforts were frustrated by the fact that the computations involve an integral –the incomplete beta function– that is difficult to evaluate. As summarized by Stigler (1986a, pp. 130–131):

> "This much is clear, however: Through an exceedingly painstaking and tortured analysis, Bayes sought to bound the incomplete beta above and below. His solution was no more than a Pyrrhic victory because his bounds were far too complex for practical evaluation and were not even very close. (...)
>
> Bayes's treatment of the binomial may be regarded as mathematically incomplete. Whether or not that is accepted as the reason for his reluctance to publish, it is a sufficient explanation for the lack of attention his work received after it was published. A decade later Laplace was led to the same problem; but he was armed with a far greater analytic skill and consequently his solution was richer and more complete."

✓ Diaconis, P., & Skyrms, B. (2018). *Ten Great Ideas About Chance*. Princeton: Princeton University Press.

✓ Laplace, P.–S. (1774/1986). Memoir on the probability of the causes of events. *Statistical Science, 4*, 364-378. A solid contender for Most Impressive Paper on Statistics of All Time, this 1774 article (translated by Stephen Stigler in 1986) was published when Laplace was only 25 years old.

✓ Rosenkrantz, R. D. (1982). Does the philosophy of induction rest on a mistake? *The Journal of Philosophy, 79*, 78–97. Impressed by the scientific reputation of Laplace, Jevons, Jeffreys, and Pólya, we had unthinkingly accepted the fundamental inductive pattern mentioned in Chapter 6: *"The verification of a consequence renders a conjecture* *more credible*" (Pólya 1954b, p. 5). This pattern also forms the basis for Laplace's Rule of Succession.[7] The Rosenkrantz article blew our preconceptions out of the water (see also Berent 1972, Good 1967, Gardner 1976). Rosenkrantz demonstrates that background knowledge plays a vital role, and that, generally speaking, "*conforming observations need not be confirming*." (p. 84). In fact, the verification of a consequence can render a conjecture *less* likely. We elaborate on this surprising insight in the second appendix of this chapter.

[7] And for *Haldane*'s Rule of Succession outlined in Chapter 16.

✓ Stigler, S. M. (1986). Laplace's 1774 memoir on inverse probability. *Statistical Science, 4*, 359-378.

✓ Todhunter, I. (1865). *A History of the Mathematical Theory of Probability From the Time of Pascal to That of Laplace*. Cambridge: MacMillan and Co.

✓ Zabell, S. L. (1989). The rule of succession. *Erkenntnis, 31*, 283-321. All of Sandy Zabell's papers are scholarly, informative, and highly recommended; this one is a must-read for anybody who wishes to understand the Rule of Succession in more detail. "This paper will trace the evolution of the rule, from its original formulation at the hands of Bayes, Price, and Laplace, to its generalizations by the English philosopher W. E. Johnson, and its perfection at the hands of Bruno de Finetti. By following the debate over the rule, the criticisms of it that were raised and the defenses of it that were mounted, it is hoped that some insight will be gained into the achievements and limitations of the probabilistic attempt to explain induction." (p. 283).

✓ Zabell, S. L. (2005). *Symmetry and Its Discontents: Essays on the History of Inductive Probability*. Cambridge: Cambridge University Press. What holds for Zabell's papers also holds for his books: scholarly, informative, and highly recommended.

Appendix A: Deriving the Beta Prediction Rule

This chapter was concerned with the following prediction rule:

$$\begin{aligned} p(y=1) &= \int_0^1 p(y=1 \mid \theta)\, p(\theta)\, \mathrm{d}\theta \\ &= \frac{\alpha}{\alpha+\beta}, \end{aligned}$$

in other words, the probability that the next binomial trial results in a success, given that the uncertainty across parameter θ is described by a beta(α, β) distribution. Here we provide three different ways to obtain the result.

First, we may use the fact that $p(y = 1 \mid \theta) = \theta$ and obtain:

$$\begin{aligned} p(y=1) &= \int_0^1 p(y=1 \mid \theta)\, p(\theta)\, \mathrm{d}\theta \\ &= \int_0^1 \theta\, p(\theta)\, \mathrm{d}\theta, \end{aligned}$$

which is easily recognized as the expression for a mean, and we know that the mean of a beta(α, β) distribution is $\alpha/(\alpha+\beta)$. This solution is intuitive, but it is mathematically less satisfactory than computing the integral.

Second, we may use the properties of the beta integral:

$$\begin{aligned} p(y=1) &= \int_0^1 \theta\, p(\theta)\, \mathrm{d}\theta \\ &= \int_0^1 \theta\, \frac{\theta^{(\alpha-1)}(1-\theta)^{(\beta-1)}}{\mathrm{B}(\alpha,\beta)}\, \mathrm{d}\theta \\ &= \frac{1}{\mathrm{B}(\alpha,\beta)} \int_0^1 \theta^{\alpha}(1-\theta)^{(\beta-1)}\, \mathrm{d}\theta \\ &= \frac{\mathrm{B}(\alpha+1,\beta)}{\mathrm{B}(\alpha,\beta)} \\ &= \frac{\alpha}{\alpha+\beta}. \end{aligned}$$

Here B is the beta function; for integer values of x and y, we have $\mathrm{B}(x,y) = {}^{(x-1)!\,(y-1)!}/_{(x+y-1)!}$. The last step above follows from the identity $\mathrm{B}(\alpha+1,\beta) = \mathrm{B}(\alpha,\beta) \times \frac{\alpha}{\alpha+\beta}$.

Third, we can use the expression for the probability mass function for the *beta-binomial*, that is, the distribution of the number of predicted successes s out of n attempts when $\theta \sim$ beta(α,β):

$$p(s \mid n) = \binom{n}{s} \frac{\mathrm{B}(\alpha+s, \beta+n-s)}{\mathrm{B}(\alpha,\beta)}.$$

Entering $s = 1$ and $n = 1$ simplifies the formula to

$$p(s=1 \mid n=1) = \frac{\mathrm{B}(\alpha+1,\beta)}{\mathrm{B}(\alpha,\beta)} = \frac{\alpha}{\alpha+\beta}.$$

The astute reader will realize that the mass that the beta-binomial distribution assigns to a specific predicted outcome (i.e., s successes out of n attempts) equals the marginal probability for that outcome (i.e., the integral from the second method).

Appendix B: ‘Conforming Observations Need Not Be Confirming’

In Chapter 6, we briefly mentioned the idea of ‘corroborating the consequent’. The famous mathematician George Pólya termed this the fundamental inductive pattern: *The verification of a consequence renders a conjecture more credible.* Moreover, Pólya considered this to be self-evident to the point of triviality; the fundamental inductive pattern "says nothing surprising. On the contrary, it expresses a belief which no reasonable person seems to doubt" (Pólya 1954b, p. 5). However, closer inspection reveals that the verification of a consequence does not always renders a conjecture more credible – it may even render it *less* credible. We will not do a deep dive into the relevant theory, but instead content ourselves with a few concrete examples.[8]

[8] One such example was already given in Chapter 6.

Example 1: Two Worlds

Statistician and World War II code breaker Jack Good was particularly adept at providing counterexamples to the fundamental inductive pattern.[9] Here is the first one:

[9] We will revisit Good's work in Chapter 23.

> "Suppose that we know we are in one or other of two worlds, and the hypothesis, H, under consideration is that all the crows in our world are black. We know in advance that in one world there are a hundred black crows, no crows that are not black, and a million other birds; and that in the other world there are a thousand black crows, one white one, and a million other birds. A bird is selected equiprobably at random from all the birds in our world. It turns out to be a black crow. This is strong evidence (...) that we are in the second world, wherein not all crows are black. Thus the observation of a black crow, in the circumstances described, undermines the hypothesis that all the crows in our world are black." (Good 1967, p. 322)

Example 2: The White Crow

The hypothesis under consideration holds that all crows are black. Now suppose we observe a white *raven*. Even though this observation adheres to the rule, it is intuitively obvious that it actually undercuts it, because crows and ravens are biologically similar (Good 1960, p. 149; see also Rosenkrantz 1982, pp. 82-83). The knowledge that ravens can be white greatly increases the probability that the same holds for crows.

Example 3: The Baby

The hypothesis under consideration holds that all crows are black. The observation of a white shoe seems to conforms to this hypothesis (or at least not violate it), but...

> "(...) in very special circumstances, the sight of a white shoe can actually undermine the hypothesis that all crows are black. Suppose that a child had seen black crows, black shoes, and no other black objects, and that all the crows and shoes had been black. He now sees a white shoe and he says, 'How surprising! Apparently objects that are supposed to be black

can sometimes be white instead.' On the information available to the child this may be a very rational thing for him." (Good 1961, p. 64)[10]

[10] Note the conceptual similarity between the baby example and the white crow example.

Example 4: Convicts

Here is yet another one of Good's counterexamples to the fundamental inductive pattern:

> "Suppose we are told that all men in Ealing whose surnames end with the letter z are escaped convicts. We take a random sample of the citizens of Ealing, and, after a very short time, we find one whose surname ends with z. Then the fact that we found such a one so quickly tends to undermine the hypothesis, for this evidence by itself suggests that there are more people whose surnames end with z than we had previously supposed." (Good 1961, p. 64)

Example 5: Rosenkrantz's Hats

Howson and Urbach (2006) summarize an example introduced by Rosenkrantz (1977, p. 35):

> "Three people leave a party, each with a hat. The hypothesis that none of the three has his own hat is confirmed, according to Nicod[11], by the observation that person I has person 2's hat and by the observation that person 2 has person 1's hat. But since the hypothesis concerns only three, particular people, the second observation must *refute* the hypothesis, not confirm it." (Howson and Urbach 2006, p. 102; italics in original)

[11] "Hypotheses of the form 'All Rs are B' are confirmed by evidence of something that is both R and B. (Hempel called this *Nicod' Condition*, after the philosopher Jean Nicod.)" (Howson and Urbach 2006, p. 100; italics in original).

Example 6: Grasshoppers

The hypothesis states that "All grasshoppers are located in parts of the world other than Pitcairn Island." (Swinburne 1971, p. 326). Consider then the following:

> "Finding by chance a grasshopper somewhere else than on Pitcairn Island as such (that is, in the absence of further information, e.g., that it was found in a region where grasshoppers were already known to abound) only suggests that grasshoppers are more abundant than we supposed and so in view of the similarities between things located and things not located on Pitcairn Island, more likely than we supposed to be located on Pitcairn Island. We can see the point yet more clearly if we consider the effect on the hypothesis of the discovery of a large number of (...)instances. Discovery that the rest of the world was swarming with grasshoppers clearly casts grave doubt on the hypothesis. But the discovery of a large number n of grasshoppers can be represented as the discovery of n individual grasshoppers in succession. Either each discovery disconfirms slightly or at some stage there is a sudden large increment of disconfirmation. The latter is implausible, for any choice of m, such that although observation of m grasshoppers did not disconfirm, observation of the $m + 1$th grasshopper discontinued substantially, would seem arbitrary. Hence, I conclude, each instance is separately disconfirmatory." (Swinburne 1971, p. 326)[12]

[12] Note the conceptual similarity between Swinburne's grasshopper example and Good's convict example.

Example 7: Giants and Other Surprises

For 25 years, the mathematician Martin Gardner (1914–2010) kept the readership of *Scientific American* spellbound through his 'Mathematical Games' columns, which were subsequently bundled into several books. In the 1988 book 'Time Travel and Other Mathematical Bewilderments', the chapter 'Induction and Probability' (a reprint of the original Gardner 1976 column) presents a series of exceptions to the fundamental inductive pattern. Here we highlight the simplest case:

> "(...) there are situations in which confirmations make a hypothesis less likely. Suppose you turn the cards of a shuffled deck looking for confirmations of the guess that no card has green pips. The first ten cards are ordinary playing cards, then suddenly you find a card with blue pips. It is the eleventh confirming instance, but now your confidence in the guess is severely shaken." (Gardner 1988, p. 244)

The crux of the example is that the critical observation is conforming but also *surprising*. In this context Gardner refers to a one-page article by Paul Berent from which we quote at length:

> "Consider the following example: the statement 'All men are less than 100 feet tall' would decrease in probability upon discovery of a man 99 feet tall (*almost* a negative instance). If subsequent men were found to be either normal or else *exactly* 99 feet tall then new giants would disconfirm the generalization less and less until a low point would be reached (when?) whereupon the probability would increase and eventually reach the original level (when?). At this point a new giant would confirm, whereas had he been the first giant he would have disconfirmed, although the probability given the old evidence would have been the same.
>
> A second way a positive instance can disconfirm is by being in an unsuspected place, e.g. a normal man on Mars. A third way is by breaking a pattern, e.g. a man 98 feet tall after a long sequence of normal men and men exactly 99 feet tall. A fourth way is by disconfirming a background theory which supports the generalization. An example of this type of case would be given by a normal size yogi with ability to get by on little oxygen; for this would render less plausible an important biological argument against the occurrence of *giants*: volume increases more rapidly than surface area (we breathe on the surface of our lungs)." (Berent 1972, p. 522; italics in original)

As an aside, the Berent article does not seem to get the recognition it deserves.[13] For instance, Rosenkrantz presents the following example, but without crediting Berent:

[13] Apparently it has been cited just once (Google Scholar, June 13, 2023).

> "the existence of a man 199 years old and in perfect health is consistent with the hypothesis that no man (past, present, or future) attains the age of 200, but can hardly be thought to confirm that hypothesis." (Rosenkrantz 1982, p. 84)

In addition, Jack Good mentions the example hypothesis "that no man weighs more than 2000lbs" in several of his writings (e.g., Good

1986; 1989) – he cites the 1976 Gardner column but mentions the Berent article only indirectly: "Essentially this example was attributed to Paul Berent by Gardner (1976)" (Good 1989, p. 121), which falsely suggests that Berent conveyed the example to Gardner in conversation.

At any rate, a similar example can be constructed for the Laplacean hypothesis that the sun will rise tomorrow. Suppose we wake up to find the sun has risen. This is a conforming observation. However, the sun is ten times its usual size, dark-blue in color, and pulsating rapidly. Even though the observation is conforming, it also signals impending solar doom, and therefore severely undercuts our confidence that the sun will rise again tomorrow: it *disconfirms* the hypothesis.

Both Gardner and Rosenkrantz believe that the philosopher Rudolf Carnap was well aware of the exceptions to the fundamental inductive pattern (cf. Carnap 1950, Chapter 6). Unfortunately, a serious study of Carnap demands considerable time and effort. To paraphrase Napoleon when Laplace handed him his monograph on celestial mechanics: we will study Carnap as soon as we have six months of free time at our disposal.

Example 8: Mathematics

George Pólya concerned himself with induction as relevant for mathematics. It seems appropriate therefore to present a counterexample in that discipline.

Suppose you are tasked to evaluate the hypothesis 'The function $f(x)$ is non-negative, that is, $f(x) \geq 0$ for any real number x.' You are not given $f(x)$ directly, but you can issue queries – in other words, you may provide a number of input values and observe the resulting output values. You decide to input five values, $x = \{0, 60, 90, 150, 180\}$, and you are then informed that $f(x)$ is non-negative for all of them. In other words, you obtain a sequence of five conforming observations for $f(x)$. This may increase your confidence that $f(x)$ is indeed non-negative. But now consider that you are given additional information – not just whether or not $f(x) \geq 0$, but the precise outcome. The outcome values are $f(x) = \{0, \sqrt{3}/2, 1, 1/2, 0\}$. A mathematician will recognize that these are exactly the output values of the sine function (with the input x provided in arc degrees). However, the sine function ranges from -1 to 1 and hence conflicts with the hypothesis that $f(x)$ is non-negative.

Note that the precise outcomes are consistent with an infinite number of hypotheses. For instance, $f(x)$ may just be zero outside of the $0 - 180$ interval, and this function would be non-negative. Or the function may be $\cos(x) + 1$ for all values of x other than $0, 60, 90, 150$, and 180. But such hypotheses seem much less plausible than the simple sine function. We suspect that even the single input-output pair $f(60) = \sqrt{3}/2$ will prompt mathematicians to assign the sine function a relatively high probability: the *conforming* value of $\sqrt{3}/2$ acts to *disconfirm* the hypothesis.

As the examples above demonstrate, the fundamental inductive pattern does not hold across the board. In particular, background knowledge may play a decisive role. Arch-Bayesians Harold Jeffreys, Dennis Lindley, and Ed Jaynes realized the importance of prior knowledge and explicitly conditioned on it in their notation – they would never write $p(\theta \mid \text{data})$, but always $p(\theta \mid \text{data}, K)$, where K represents prior knowledge. Throughout this book we omit this conditioning in order to keep the notation succinct, but it is important to keep in mind that our probabilistic inference is based on a web of background assumptions on how the data may have been generated – as the examples in this appendix serve to underscore.

Jack Good has suggested that the fundamental inductive pattern may hold when the observations are reported through a 'stooge'. The stooge reports not the actual observation, but merely whether or not it is conforming.[14] It is obvious from the examples that such stoogian observations sometimes omit crucial information and may be highly misleading (Good 1989).

This appendix has underscored the conclusion drawn by Rosenkrantz in 1982: "In short, from a Bayesian point of view, *conforming observations need not be confirming* (Rosenkrantz 1982, p. 84; italics in original).

[14] "If he gives any other information he will be shot dead and knows it." (Good 1960, p. 148)

10 *The Problem of Points [with Jiashun Wang]*

Neglecting the trifling hints which may be found in preceding writers we may say that the Theory of Probability really commenced with Pascal and Fermat; and it would be difficult to find two names which could confer higher honour on the subject.

Todhunter, 1865

Chapter Goal

This chapter illustrates the difference between aleatory and epistemic uncertainty with the iconic 'Problem of Points': given that a game has been interrupted and cannot be resumed, how should the stakes be divided?

Blaise Pascal (1623–1662). Portrait painted in 1691 by François II Quesnel.

Interrupting a Game of Chance

The field of statistics and probability theory was born around 1654, in a famous correspondence between Blaise Pascal and Pierre de Fermat. These two French mathematicians concerned themselves with a problem in gambling: suppose players A and B are engaged in a match – for concreteness, suppose they are repeatedly tossing a fair coin. Whenever the coin lands heads, player A wins a point; whenever it lands tails, player B wins a point. It is agreed that the first player to reach six points wins the match and receives a stake of $100. When the score is $5-3$ in favor of player A the match is interrupted, never to be resumed. How can the stakes be divided fairly?

This 'Problem of Points' had been studied previously, but without resulting in a satisfactory answer. Some mathematicians even concluded that the problem was unsolvable! A detailed history of the problem can be found elsewhere (cf. Devlin 2008, pp. 16-18; Edwards 1987/2019; Todhunter 1865; see also Diaconis and Skyrms 2018, Chapter 1); here we proceed straight to the solution. The key idea is that the stake should be divided according to the probability of winning the match in case it had continued. For our present scenario, the computation is simple: the only way in which player B could win is when the coin lands tails on three consecutive tosses. This probability is $1/2 \times 1/2 \times 1/2 = 1/8$, so player B should receive $100 \times 1/8 =$ \$12.5 and player A should receive $100 \times 7/8 =$ \$87.5. In hindsight, it seems mysterious that this straightforward idea totally escaped the mathematicians who studied the problem before Pascal and de Fermat.

The Problem of Points becomes more complicated, however, when there are multiple ways for player B to win. For instance, the score could be 4-3 rather than 5-3. Now player B wins in the following sequences of outcomes (H stands for heads, T for tails):

TTT (probability $1/2 \times 1/2 \times 1/2 = 1/8$)
TTHT (probability $1/2 \times 1/2 \times 1/2 \times 1/2 = 1/16$)
THTT (probability $1/2 \times 1/2 \times 1/2 \times 1/2 = 1/16$)
HTTT (probability $1/2 \times 1/2 \times 1/2 \times 1/2 = 1/16$)

The sum of these four sequences is $5/16$, so player B ought to receive $100 \times 5/16 =$ \$31.25, with the remaining $100 \times 11/16 =$ \$68.75 going to player A.

Enumerating the winning sequences is tedious, and Pascal invented his famous 'triangle' to facilitate the computation. A discussion of Pascal's triangle will lead us too far afield, and instead we refer the interested reader to Chapters 33 and 34 for details. Here we will focus on a different method to obtain the solution: JASP.[1] After opening JASP and activating the *Learn Bayes* module, navigate to *The Problem of Points* and select *Game of Chance*. We then set up the scenario described above: 'Points needed to win the game' equals 6, and 'Points gained' is 4 for player A and 3 for player B.[2] The corresponding JASP output is shown in Figure 10.1.

[1] See also the blog post "Teaching the problem of points with JASP" on `https://jasp-stats.org`.

[2] In the JASP input panel, the fields for 'p(win 1 point) are set to 1 for both players; these numbers are normalized (i.e., divided by their sum) to yield the corresponding probabilities. Here the probability of a fair coin landing heads equals 0.5, so the default '1-1' setting need not be changed.

The Summary Table confirms the result obtained earlier, that is, the probability of winning the match is $11/16 = 0.6875$ for player A and $5/16 = 0.3125$ for player B. In addition, the table also reports the results of a simulation, the details of which are presented in the lower panel of Figure 10.1. For the simulation, a set of 500 synthetic matches are played, of which 347 were won by player A, for a winning percentage of 0.6940. In the figure, the wiggly black line shows how the proportion of wins by player A fluctuates as the number of simulated matches increases. To quantify the uncertainty in this proportion, the steel blue area shows the 95% (highest posterior density) credible interval. The horizontal red line shows the analytical result; as the number of simulated matches increases, the win percentage increasingly approximates the theoretical result.

Game of Chance ▾

Summary Table

			p(win the game)	
Players	p(win 1 point)	Points Gained	Analytical	Simulated
A	0.5000	4	0.6875	0.6940
B	0.5000	3	0.3125	0.3060

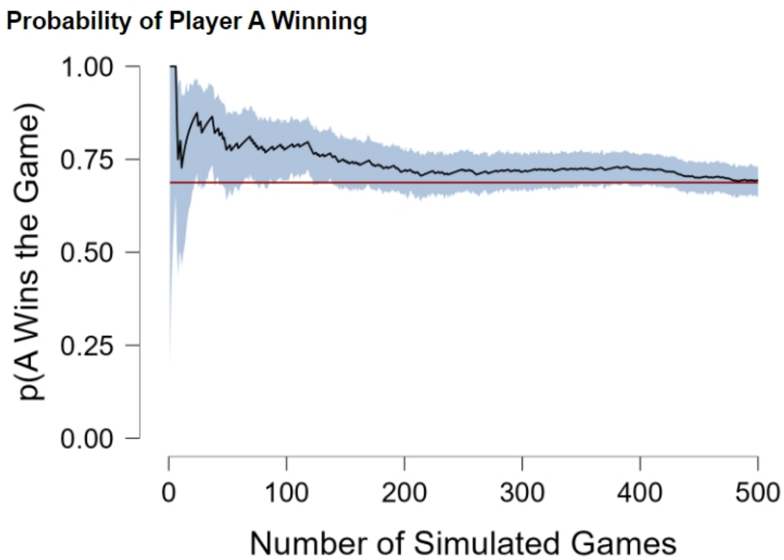


Figure 10.1: Screenshot from the JASP module *Learn Bayes* → *The Problem of Points* → *Game of Chance*, for the scenario where the score is 4—3 for player A in a race to six. See text for details.

In the highly recommended book ‘Do dice play God?’, mathematician Ian Stewart provides a birds-eye view of the work by Pascal and de Fermat:

> “Their key insight is that what matters is not the past history of the play – aside from setting up the numbers – but what might happen over the remaining rounds. If the agreed target is 20 wins and the game is interrupted with the score 17 to 14, the money ought to be divided in exactly the same way as it would be for a target of 10 and scores 7 to 4. (In both cases, one player needs 3 more points and the other needs 6. *How they reached that stage is irrelevant*.) The two mathematicians analysed this set-up, calculating what we would now call each player’s expectation – the average amount they would win if the game were to be repeated many times. The answer for this example is that the stakes should be divided in the ratio 219 to 37, with the player in the lead getting the larger part. Not something you’d guess.” (Stewart 2019, p. 31; italics added for emphasis – we will return to this statement in the next section)

We can confirm Stewart's answer with JASP; entering the suggested number of points and the target value yields a probability of about 0.1445 that the player who trails ends up winning the match. This is in line with Stewart's answer, as $^{37}/_{(37+219)} \approx 0.1445$.

The scenario above may be generalized in several ways. For instance, one may consider a game that has more than two players, or a game where the probability of winning a point is not the same for each player. Another generalization is to consider not a game of *chance* (such as tossing coins), but a game of *skill* (such as tennis). As we will see in the next section, this changes the nature of the results in a fundamental way.

Pierre de Fermat (1607–1665), a French lawyer and mathematician who contributed to number theory, analytic geometry, optics, and probability theory. Fermat sometimes teased his fellow mathematicians by omitting the proofs of his propositions. 'Fermat's Last Theorem' holds that, for positive integers n, a, b, and c, the equation $a^n + b^n = c^n$ has no solution for $n > 2$. Around 1637, Fermat wrote in the margin of a copy of Diophantus's *Arithmetica* that "I have discovered a truly marvelous proof of this, which this margin is too narrow to contain." ("(...) cuius rei demonstrationem mirabilem sane detexi. Hanc marginis exiguitas non caperet"). It took until 1994 before Andrew Wiles first presented a correct proof, using modern mathematical techniques that were unavailable to Fermat at the time. It is generally considered unlikely that Fermat actually had a correct proof: he never published his 'truly marvelous proof' during his lifetime, and we know of it only because Fermat's note was published posthumously by his son. Portrait by an unknown artist.

Interrupting a Game of Skill

In the previous section we considered a simple game –tossing a fair coin– where the uncertainty about the outcome is fully *aleatory*, that is, solely the result of sampling variability. In other words, the binomial success probability θ was known with absolute certainty and therefore stayed constant throughout the duration of play.

However, now consider a scenario in which points are earned in a game of skill, and the players' relative skill level θ is not known exactly. For instance, let's revisit the scenario in which we have a race to six and the score is $5-3$ in favor of player A; the game at hand is a version of pocket billiards known as *pool*. The fact that the score is $5-3$ suggests that A is the better player, so more likely to win the match than if it were a game of chance; consequently, the fair share of the stake for player B should be lower. From a Bayesian perspective, the lack of knowledge concerning the relative skill of the players is usually expressed by means of a beta distribution. In other words, the game of skill features not only *aleatory* uncertainty, but also *epistemic* uncertainty.

Inserting epistemic uncertainty complicates the problem, and it was Pierre-Simon Laplace who presented the solution at 25 years of age (Laplace 1774/1986, p. 369). Here we approach the problem conceptually, making use of two important rules:

1. Conjugacy: Observing s successes and f failures updates a beta(α, β) prior distribution to a beta$(\alpha + s, \beta + f)$ posterior distribution (cf. Chapter 8).

2. The Beta Prediction Rule: Given a beta(α, β) distribution, the probability that the next observation is a success equals the mean of that distribution, that is, $p(y = 1) = {}^{\alpha}/_{\alpha+\beta}$ and $p(y = 0) = {}^{\beta}/_{\alpha+\beta}$ (cf. Chapter 9).

Now suppose that the probability of A beating B on any one game is θ, and that θ is assigned a beta$(1, 1)$ prior distribution. When the score is $5-3$ in favor of player A, the prior distribution is updated to a beta$(5+1, 3+1) =$ beta$(6, 4)$ posterior distribution (by conjugacy).

Given this posterior distribution and the fact that the score is $5-3$ is a race to six, what is the probability of player B winning the match? Let's take things one step at a time. First, the probability that player B wins the *ninth* game is $4/10$ (by the Beta Prediction Rule).

Next suppose player B wins that hypothetical game, narrowing the scores to $5-4$. This would yield a beta$(6, 5)$ distribution for θ (by conjugacy), and the associated probability of player B winning the *tenth* game is $5/11$ (by the Beta Prediction Rule).

Finally, supposing that player B also wins the tenth game, evening the scores at $5-5$. This yields a beta$(6, 6)$ distribution for θ (by conjugacy), and the associated probability of player B winning the decisive *eleventh* game is $6/12 = 1/2$ (by the Beta Prediction Rule and according to intuition[3]).

In order for player B to win the match, all three successive points need to be won, so this gives $4/10 \times 5/11 \times 6/12 = 1/11$. To summarize, with a beta$(1, 1)$ prior distribution on the probability θ of player A beating player B on any single game of pool, and with the score $5-3$ in favor of player A in a race to six, we have that:

$$p(\text{Player B wins ninth point} \mid \theta \sim \text{beta}(6, 4)) = 4/10.$$

$$p(\text{Player B wins tenth point} \mid \theta \sim \text{beta}(6, 5)) = 5/11.$$

$$p(\text{Player B wins eleventh point} \mid \theta \sim \text{beta}(6, 6)) = 6/12.$$

which then yields:

$$\begin{aligned} p(\text{Player B wins the match} \mid \theta \sim \text{beta}(6, 4)) &= 4/10 \times 5/11 \times 6/12 \\ &= 120/1320 = 1/11. \end{aligned}$$

This means player B, who was trailing player A by a score of $5-3$ in a 'first to six' game of pool, stands to receive $1/11 \times \$100 \approx \$9.09$ when the game is interrupted. Note that this fair share of the stakes is somewhat less than what player B would have been entitled to if engaged in a game of chance, which would have resulted in a payout of $1/8 \times \$100 = \$12.50$. This confirms our earlier remark that the very fact that B is trailing suggests that B might be the inferior player and hence less likely to win future points, more likely to lose the match, and consequently not deserving the $12.50 cut that would be fair if we knew with certainty that the players were exactly evenly matched.

We continue our pool playing scenario and now consider a scoreline of $4-3$ rather than $5-3$. As explained in the section on the game of chance, there are four sequences that result in player B winning the match: {B,B,B}, {B,B,A,B}, {B,A,B,B}, and {A,B,B,B} (where A and B stand for a point gained by player A and B, respectively). The associated

[3] If the prior distribution does not express a preference for either player and the scores are tied, both players must have the same probability of winning the next point.

probabilities for these sequences can be obtained sequentially, using conjugacy and the Beta Prediction Rule, as we demonstrated for the case of the $5{-}3$ score – we leave this as an exercise for the reader.

After opening JASP and activating the *Learn Bayes* module, we navigate to *The Problem of Points* and now select *Game of Skill*. We set up the scenario described above: 'Points needed to win the game' equals 6, and 'Points gained' is 4 for player A and 3 for player B. For two players, the 'Prior skill parameter' refers to the corresponding parameters of the beta distribution for θ. The default setting is to assign θ a beta$(1, 1)$ prior distribution. The result is shown in Figure 10.2.

Game of Skill

Summary Table

			p(win the game)	
Players	Prior Skill	Points Gained	Analytical	Simulated
A	1	4	0.7273	0.7180
B	1	3	0.2727	0.2820

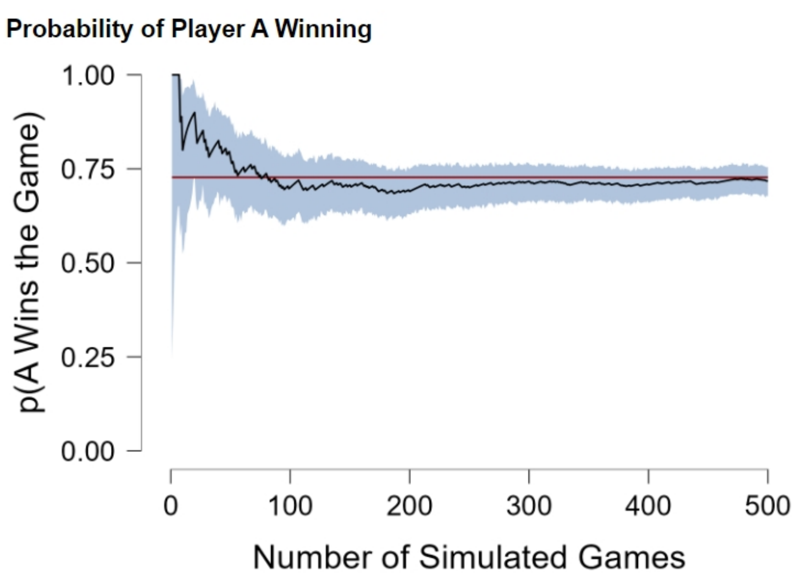


Figure 10.2: Screenshot from the JASP module *Learn Bayes* $\rightarrow$ *The Problem of Points* $\rightarrow$ *Game of Skill*, for the scenario where the score is $4{-}3$ for player A in a race to six. See text for details.

The Summary Table indicates that player B has a probability of winning the game that equals about 0.2727 – as expected, this is somewhat lower than the probability of 0.3125 from the game of chance (cf. Figure 10.1). The lower panel of Figure 10.2 confirms the analytical result with a small simulation of 500 synthetic matches, 141 of which were won by player B, for a percentage of 0.2820.

We now return to the quotation by Ian Stewart at the end of the section on the game of chance. Steward stressed that the past history of play is irrelevant, in the sense that it does not matter whether the score is $7-4$ in a race to 10 or $17-14$ in a race to 20; in both cases player A is three points ahead of player B, and three points away from the target number. However, this 'key insight' by Pascal and de Fermat is fallacious as soon as we consider the game of skill. The reason is that for the game of skill, the history of past outcomes provides valuable information about θ. This is illustrated in Table 10.1, which features five possible scorelines; for each scoreline, player A is ahead by three points and requires three more points to reach the target number. The right-most column confirms that for the game of chance (with $\theta = 1/2$) the probability that player A wins the match is about 0.8555 (i.e., $219/256$) which does not depend on the number of points that were played in the past. For the game of skill, however, the past number of plays does matter. With a scoreline of $3-0$ and a beta$(1, 1)$ prior distribution for θ, the predictions about future play follow from a beta$(4, 1)$ posterior, which reflects the opinion that player A might well be superior, and the most likely outcome is a 'sweep', that is, $6-0$. With a scoreline of $997-994$, on the other hand, the predictions about future play follow from a beta$(998, 995)$ posterior, which is highly peaked around $\theta = 1/2$, reflecting the opinion that players A and B are equally strong. In this scenario, the probability that player A wins the match is almost the same as if it was a game of chance and θ was known to equal $1/2$ exactly.

Table 10.1: In a game of skill, the history of past outcomes is informative about the skill difference θ, and this affects the expectation that the player in the lead will win the match. This is not the case for a game of chance, where θ is known. See text for details.

		p (A wins match)	
Score A–B	Race to	Game of skill	Game of chance
$3-0$	6	0.9697	0.8555
$7-4$	10	0.9151	0.8555
$17-14$	20	0.8824	0.8555
$97-94$	100	0.8605	0.8555
$997-994$	1000	0.8560	0.8555

Exercises

1. This is the 'exercise for the reader' mentioned above: consider a game of pool where player A leads player B by a score of $4-3$ when the game is interrupted. What is the fair proportion of the stake that

should go to player B? NB. Four outcome sequences result in player B winning the match: {B,B,B}, {B,B,A,B}, {B,A,B,B}, and {A,B,B,B}. Use conjugacy and the Beta Prediction Rule to obtain the relevant probability.

2. Consider a game of skill. Player A has 3 points, Player B has 5 points, and both require 6 points to win. With a uniform distribution on θ, the fair proportion of the stake for player A is $1/11$. In the game of chance, the fair proportion is $1/8$. Use the *Learn Bayes* module and adjust your prior assumptions about the relative skill level θ such that the fair proportion of the stake approximates $1/8$.

3. Someone approaches you and proposes to throw a fair die; when the die lands five or six, you win a point, else you lose a point. The game is a race to 10. How many points do you think you would need as a head start to make this a fair game? And how about when the game is a race to 100, or a race to 1000? Test your intuition with the *Learn Bayes* module.

4. Bonus question, generalizing the previous one: suppose your chance of winning any point is $\theta < 1/2$. In a race to n points, what proportion of points do you need as a head start to make the game fair? [hint: try out some values in JASP first, and then try to guess or derive the general result]

5. Consider a game of skill, with player A having two points and player B having four points. The winner either has to obtain six points, or 60 points. What game is B more likely to win? Can you explain why?

6. The Problem of Points may or may not be relevant for law, as illustrated by two fictitious court cases:

6.1. Don and Harriet find themselves in a car collision. Harriet gets a whiplash which temporarily prevents her from working. Her total damages are estimated to be around \$ 150,000. What proportion of Harriet's damages should Don's insurance company be obliged to cover? Note that there is an 80% probability that the collision was caused by Don; there is a 95% probability that Harriet's complaints were caused by the collision. Moreover, Harriet was considering to switch jobs. There is a 60% probability that she would have stayed in her current job (which pays a net annual salary of \$150,000), and a 40% probability that she would have taken a less stressful job (which pays a net annual salary of \$60,000). In light of this information, what do you consider to be fair compensation for Harriet?

6.2. John was walking his dog when he was hit in the head by an iron ball that came flying over a hedge. On the other side of the hedge,

> Olympic athletes Don and Bob had been practicing their hammer throws. No witnesses were present to identify who threw the fatal hammer, and both Don and Bob claim that the other one was the culprit. The judge rules that the penalty for negligent homicide in this case would be 6 months in jail and a fine of \$10,000. Should Don and Bob each get 3 months in jail, and a fine of \$5,000?

Chapter Summary

The original Problem of Points featured two players engaged in a game of chance. For instance, a fair coin is tossed – 'heads' yields a point for player A, 'tails' yields a point for player B. Play continues until one of the players first reaches a target number of points. At some stage the game is interrupted, never to be resumed – how should the stakes be divided?

Through a correspondence between Blaise Pascal and Pierre de Fermat, the Problem of Points gave birth to probability theory and statistics. The main idea is that the stakes ought to be divided in proportion to the probability of each player winning the game. For instance, with the scoreline $5-3$ for player A in a race to six, and the probability of player A winning a point equal to $\theta = 1/2$, player B can only win if successful on three consecutive plays, such that the fair proportion of the stake which should go to player B equals $1/2 \times 1/2 \times 1/2 = 1/8$. Note that in the game of chance, θ is known precisely, and all uncertainty is therefore aleatory (i.e., sampling variability).

This is different in the game of skill, where the true value of θ (i.e., the probability of player A beating player B on any one play) is unknown. Hence, the game of skill *also* has epistemic uncertainty. For instance, players A and B may be engaged in a game of pool. When player A leads player B by a score of $5-3$ in a race to six, this may be because player A is simply better than player B, and hence more likely to win the match. Under a beta$(1, 1)$ prior on θ, the probability that player B comes back from $5-3$ to win the match is only $1/11$.

One paradoxical feature of the game of skill is that *adding* the epistemic uncertainty about the players' relative skill acts to *reduce* the uncertainty about the identity of the likely winner.

Want to Know More?

✓ Devlin, K. (2008). *The Unfinished Game: Pascal, Fermat, and the Seventeenth-Century Letter that Made the World Modern.* New York: Basic Books.

> "Opening the final section of his letter, Pascal makes it clear that he fully realizes Fermat is by far the better mathematician. Although he

> himself solved the problem of the points, much of his long letter is devoted to his attempt to understand Fermat's clearly superior (because simpler and more insightful) method. He appreciates that whereas he labored long and hard to find a solution, Fermat almost certainly saw at once how to set about it. Such is the mark of a truly great mathematician, of which history has seen but a handful." (Devlin 2008, p. 85)

✓ Todhunter, I. (1865). *A History of the Mathematical Theory of Probability From the Time of Pascal to That of Laplace*. Cambridge: MacMillan and Co. The Problem of Points was studied by Pascal and Fermat, but also later by James Bernoulli, Lagrange, Trembley, and Laplace. The Todhunter book is a classic text that provides an in-depth and authoritative overview. "The history of the theory of probability, from the celebrated question as to the equitable division of the stakes between two players on their game being interrupted, proposed to Pascal by the Chevalier de Méré in 1654, embracing, as it does, contributions from almost all the great names of Europe during the period, down to Laplace and Poisson, is elaborately and admirably given by Mr Todhunter in his *History* of the subject, now a classical work." (Crofton 1885, p. 769).

> "We see then that the Problem of Points was the principal question discussed by Pascal and Fermat, and it was certainly not exhausted by them. For they confined themselves to the case in which the players are supposed to possess equal skill; and their methods would have been extremely laborious if applied to any examples except those of the most simple kind. Pascal's method seems the more refined (...) (Todhunter 1865, p. 17)

✓ Edwards, A. W. F. (1987/2019). *Pascal's Arithmetical Triangle: The Story of a Mathematical Idea*. Mineola, NY: Dover Publications.

Appendix I, "Pascal and the Problem of Points" provides an in-depth overview. Reprint of Edwards1982. From abstract 1982 paper:

> " The Pascal-Fermat correspondence and Pascal's Traité du triangle arithmétique are re-examined with special reference to the Problem of Points. It is concluded that, contrary to the views of some modern commentators, Pascal was responsible for the modern solution to the Problem, and that, in demonstrating it, he made use not only of mathematical induction, but of the concepts of expectation and of the binomial distribution for equal chances." (Edwards 1982, p. 259)

11 Interlude: Buffon's Needle [with Quentin F. Gronau and Jiashun Wang]

The mathematical ability evinced by Buffon may well excite surprise; that one whose life was devoted to other branches of science should have had the sagacity to discern the true mathematical principles involved in a question of so entirely novel a character, and to reduce them correctly to calculation by means of the integral calculus, thereby opening up a new region of inquiry to his successors, must move us to admiration for a mind so rarely gifted.

Crofton, 1869

Georges-Louis Leclerc, Comte de Buffon (1707–1788). Portrait by François-Hubert Drouais. "This famous portrait of Buffon has been copied and engraved time and time again. The naturalist is shown here in all his glory, at the age of 53. In his rich embroidered clothes, he breathes dignity, opulence, self-confidence, and a certain good-heartedness all at the same time. To see him, it is understandable that his contemporaries had spoken of the "imposing" air of the naturalist, and it is easy to forget that this athlete stood barely five feet five [1.65 m – EWDM]. Diderot greatly admired this portrait, "where the nobility and the vigor of the truly picturesque head of this philosopher can be seen." " (Roger 1997, p. 222)

Chapter Goal

Take a needle, toss it randomly on a floor with parallel planks, and keep track of whether or not the needle crosses one of the cracks. Surprisingly, this procedure can be used to estimate π, the ratio of a circle's circumference to its diameter. For instance, when the needle is half as long as the plank is wide, one point estimate of π is simply the number of *tosses* divided by the number of *crosses*. In this chapter we cast this procedure in a Bayesian light. We translate the posterior distribution for the proportion of crosses θ to the corresponding posterior distribution for π. Application to previously collected data underscore the value of reporting the entire posterior distribution instead of only a point estimate.

Buffon's Natural History

Before we turn to his needle, we should say a few words about the Count of Buffon himself. Early in life, Buffon inherited a small fortune, allowing him to dedicate his time to the pursuit of his scientific interests. And these interests concerned a wide range of topics. Buffon is remembered mostly as an ecologist, a zoologist, and an anthropologist, but initially, Buffon was fascinated by mathematics[1] and the mechanical properties of wood (for the construction of ships). Buffon translated

[1] As witness the needle problem below, and his correspondence with the mathematician Gabriel Cramer (1704-1752).

Isaac Newton's *Method of Fluxions and Infinite Series* into French, and speculated that our Solar System was created when a comet collided with our sun, a hypothesis that bears similarity to the *tidal theory* which was proposed much later.[2] Buffon conducted experiments on gravitational pull, pendulum movements, ballistics, and optical phenomena – in fact, in an experiment on human color perception Buffon irreparably damaged his own eyesight (Fellows and Milliken 1972, p. 80).

[2] The tidal theory holds that the planets were created through interaction between the sun and another star passing nearby. It was first proposed by Sir James Jeans (1877-1946) and further developed by Sir Harold Jeffreys, the hero of this book, who explicitly acknowledged the similarity: "These considerations led both Jeans and me to abandon any idea of gradual development and to examine a tidal theory on the lines of that of Buffon" (Jeffreys 1952, p. 282)

Buffon's theory of the earth "freed geology from the Bible and opened an unfathomable past to the imagination." (Roger 1997, p. 105)

Buffon also suggested that the earth was much older than 4004 BC, the date of creation calculated by the Archbishop of Armagh, James Ussher (1581-1656):

> "The first attempt at measurement [of the earth's age] that could be called remotely scientific was made by the Frenchman Georges-Louis Leclerc, Comte de Buffon, in the 1770s. It had long been known that the Earth radiated appreciable amounts of heat – that was apparent to anyone who went down a coal mine – but there wasn't any way of estimating the rate of dissipation. Buffon's experiment consisted of heating spheres until they glowed white-hot and then estimating the rate of heat loss by touching them (presumably very lightly at first) as they cooled.[3] From this he guessed the Earth's age to be somewhere between 75,000 and 168,000 years old. This was of course a wild underestimate; but it was a radical notion nonetheless, and Buffon found himself threatened with excommunication for expressing it. A practical man, he apologized at once for his thoughtless heresy, then cheerfully repeated the assertions throughout his subsequent writings." (Bryson 2004, p. 105)[4]

[3] See also Fellows and Milliken (1972, p. 74), who cite a "scandalous" account by the Chevalier Aude: "To determine the epoch of the formation of the planets and to calculate the cooling time of the terrestrial globe, he had resort to four or five pretty women, with very soft skin; he had several balls, of all sorts of matters and all sorts of densities, heated red hot, and they held these in turns in their delicate hands, while describing to him the degrees of heat and cooling."

[4] Buffon himself had said, "It is better to be humble than hung." (Roger 1997, p. 188)

Buffon was admitted to the prestigious French Academy of Sciences in 1734, and to the literary Académie française in 1753. In 1739 Buffon was appointed intendent of the *Jardin du Roi* –the Royal Botanical Garden– in Paris, which now goes under the name of *Jardin des Plantes*. Buffon enlarged the Jardin du Roi and gradually transformed it to a research center and a museum. The zeal with which Buffon expanded the Jardin du Roi can be appreciated from the following anecdote:

> "In Paris, one rainy morning early in September, 1782, the monks of the Abbey of Saint-Victor, who had refused to vacate a building Buffon wanted to demolish, as part of his plan for enlarging the Jardin du Roi, awoke to find that Buffon's laborers were busily ripping their roof off." (Fellows and Milliken 1972, p. 144)

Buffon's magnum opus was an encyclopedia titled *Histoire naturelle générale et particulière avec la description du Cabinet du Roi*.[5] During Buffon's life, this encyclopedia consisted of 36 volumes – with 8 more published after his death. The topics covered in *Histoire naturelle* mostly dealt with minerals, birds, and quadrupeds. The entries often came with detailed tables of measurements, lively descriptions, and beautiful engravings. The *Histoire* was a big hit. As summarized by one of Buffon's biographers:

[5] "Buffon soon added to his duties the project of publishing a descriptive catalogue of the reorganized and enlarged Cabinet du Roi, and this proposed catalogue quickly developed into his monumental *Histoire Naturelle*" (Fellows and Milliken 1972, p. 55).

> "The first three volumes of the *Natural History* were an immediate and resounding success in sales. (...) This success continued during the entire time the work was published; we know that the *Natural History* was the most widespread work of the eighteenth century, beating the abbé Pluche's *Spectacle of Nature*, Diderot's and d'Alembert's *Encyclopédie*, and even the better-known works of Voltaire and Rousseau. Buffon had wanted to touch the general public; he had succeeded completely." (Roger 1997, p. 184)

The popularity of the *Histoire* was arguably driven by two main factors: Buffon's writing style and the nearly 2,000 engravings that enliven the work. An example set of engravings is shown in Figure 11.1.

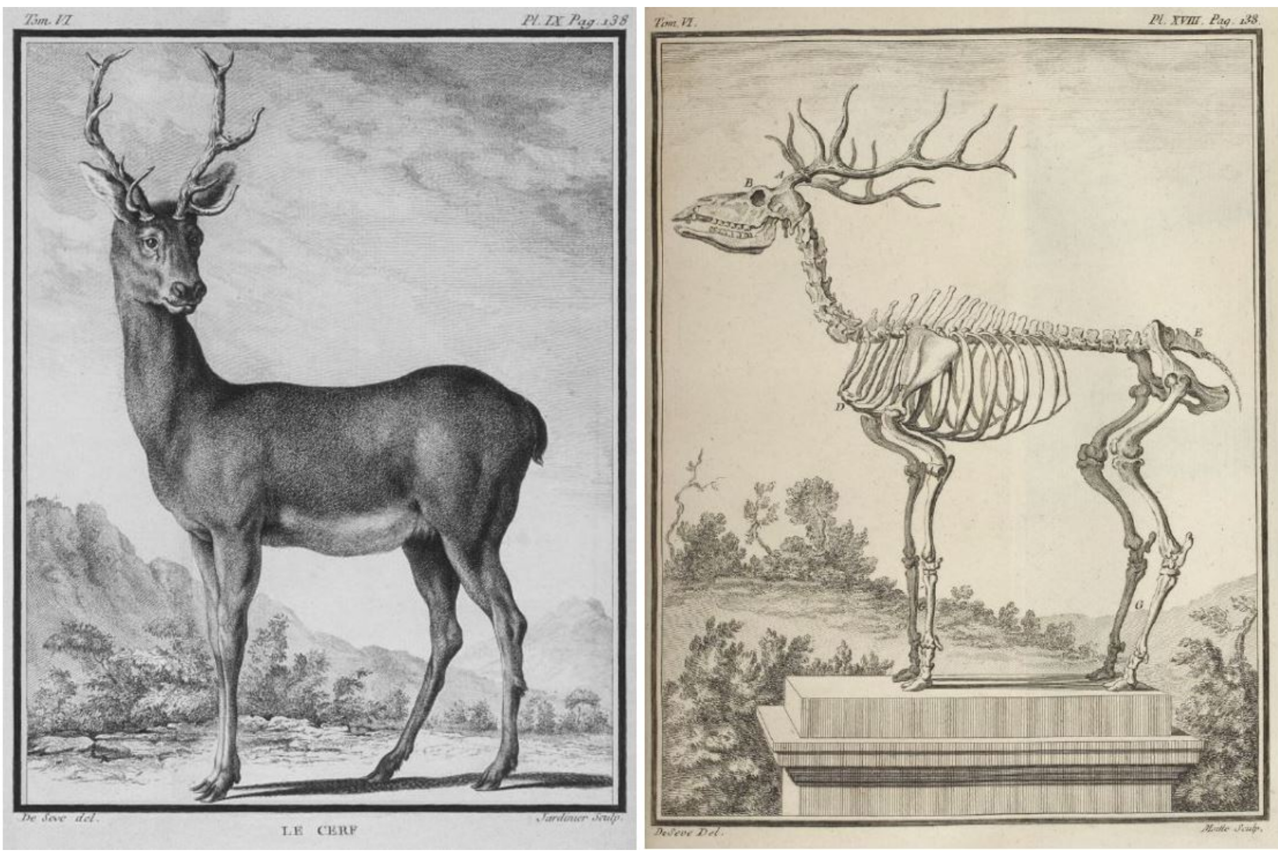


Figure 11.1: Two example illustrations from the sixth volume of Buffon's magnum opus *Histoire naturelle générale et particulière avec la description du Cabinet du Roi* (1756, p. 138). Left panel: 'Le cerf' (stag red deer). This retouched version was obtained from `https://en.wikipedia.org/wiki/Histoire_Naturelle`; the original source is `http://gallica.bnf.fr/ark:/12148/btv1b2300253d/f11.item`. Red panel: the deer skeleton. Note the letters that identify different parts. Source: `https://gallica.bnf.fr/ark:/12148/bpt6k10672421/f187.item`. Both illustrations were designed by Jacques De Sève; the left panel was engraved by Claude Donat Jardinier; the right panel was engraved by Pierre-Etienne Moitte.

Buffon's writing style was considered flowery and unscientific by some of his colleagues; we present a few examples and have the reader decide for themselves. Firstly, here is how Buffon introduces the domestic cat, at the start of the sixth volume of *Histoire naturelle générale et particulière*:

> "The cat is a faithless domestic, and only kept through necessity to oppose to another domestic which incommodes us still more, and which we cannot drive away; for we pay no respect to those who, being fond of all beasts, keeps cats for amusement. Though these animals are gentle and frolicksome when young, yet they even then possess an innate cunning, and perverse disposition, which age increases, and which education only serves to conceal. They are naturally inclined to theft, and the best education only converts them into servile and flattering robbers; for they have the same address, subtilty [sic], and inclination for mischief or rapine. Like all knaves they know how to conceal their intentions, to watch, wait, and choose opportunities for seizing their prey; to fly from punishment, and to remain away until the danger is over and they can return with safety.
>
> They readily conform to the habits of society, but never acquire its manners; they have only the appearance of attachment, as may be seen by the obliquity of their motions, and the duplicity of their looks; they never look in the face of those who treat them best and of whom they seem to be the most fond, but either through fear, or falsehood, they approach him by windings to seek for those caresses they have no pleasure in but only to flatter those from whom they receive them. Very different from that faithful animal the dog, whose sentiments are all directed to the person of his master, the cat appears only to feel for himself, only to love conditionally, only to partake of society that he may abuse it; and by this disposition he has more affinity to man than the dog, who is all sincerity."

Le chat domestique – the domestic cat. Illustration from the sixth volume of Buffon's *Histoire naturelle générale et particulière avec la description du Cabinet du Roi* (1756, p. 48). Design by Jacques De Sève, engraving by Pierre Charles Baquoy. Source: `https://gallica.bnf.fr/ark:/12148/bpt6k10672421/f65.item`.

Secondly, the fragment below concerns the state of a pristine nature, a wilderness unspoiled by human intervention. In contrast to what one may expect from a 'naturalist' today, Buffon is less than enthusiastic:

> "Enormous serpents trace wide furrows on this swampy earth: crocodiles, toads, lizards, and a thousand other reptiles with broad feet knead the mire; millions of insects multiplied by the humid heat lift up the sludge from it, and this entire corrupt population slithers in the silt or hums in the air that it obscures; all this vermin with which the earth swarms attracts flocks of voracious birds whose raucous cries, multiplied by and mixed with the croakings of the reptiles, trouble the silence of these awful wastes and seem to add fear to the horror in order to repel man and forbid the entry of other sentient beings." (as cited in Roger 1997, p. 239)

Almost automatically Buffon's words spawn an image in the reader's mind, painting a scene of a world that lies beyond personal experience. A drier, more scientific style would only have served to blur that image.

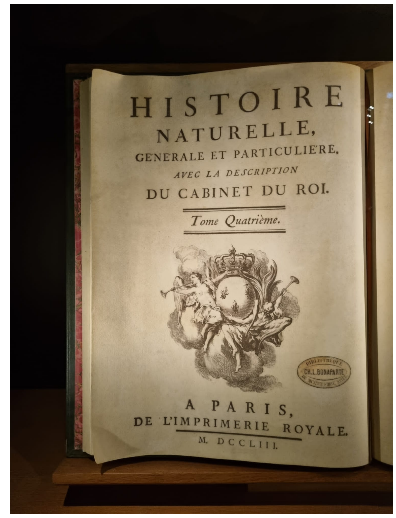
Much more can be said about Buffon, and the interested reader is referred to two biographies for details (i.e., Fellows and Milliken 1972, Roger 1997).[6] We cannot restrain ourselves and present one more example about Buffon's scientific exploits before moving to his needle.

[6] Unfortunately, both biographies are currently out of print.

Buffon's Demonstration of the Archimedes Death Ray

Early in his career Buffon had successfully carried out a single experiment that instantly made him famous. As described by Fellows and Milliken (1972),

> "But Buffon's fame was also due in part to his remarkable public relations sense. He had reached the height of fame very early in his career, in 1747, prior to the publication of the first volumes of the *Histoire Naturelle*, on the strength of a single experiment, artfully chosen for its dramatic possibilities.
>
> To disprove Descartes' theoretical demonstration of the impossibility of constructing a burning lens or mirror capable of setting fires at a considerable distance by concentrating the sun's rays on a target area, the device which Archimedes was said to have used against the Roman fleet at Syracuse, Buffon set out to construct such a machine, and succeeded. After a number of failures, he hit upon the device of an upright wooden grid on which a large number of small, flat mirrors were attached by adjustable screws that permitted each individual mirror to be aimed by hand, and with this device he was able to ignite wood at a distance of more than two hundred feet [61 meter – EWDM]. The spectacle of a modern scientist recreating one of the fabled marvels of antiquity, in defiance of a theoretical pronouncement by the great Descartes himself, stirred imaginations across all of Europe. Spectators flocked to the demonstrations, and even King Louis XV condescended to view the new marvel in operation. Frederick the Great of Prussia sent the hitherto little known French physicist his personal congratulations. Buffon had made his name, the name he had chosen for its simplicity and euphony, for the ease with which it could be remembered, a household word throughout Europe." (Fellows and Milliken 1972, pp. 56-57; see Buffon 1747 for the original paper and Vol. 10, pp. 193–244 in Buffon 1797–1807 for an English rendition)



Cover page of the Histoire Naturelle as displayed in the Grande Galerie de l'Évolution at the Jardin des Plantes, Paris, France (October 2022).

Roger (1997) describes Buffon's death ray as follows:

> "In Greek history, Archimedes set fire to Roman vessels that were attacking Syracuse by using concave mirrors that concentrated the sun's rays. According to Descartes, these mirrors "had to be extremely large, or more likely mythical."[7] Not allowing that opinion to influence him, Buffon built several square concave mirrors made up of smaller, slightly curved mirrors. The largest mirror, which measured 6 feet on one side (about 1.8 meters) was made of 360 small mirrors. With it, Buffon was indeed able to set fire to buildings made of wood at a distance of 10 to 200 feet (from 3 to about 65 meters). At a distance of 10 feet, he could melt iron." (Roger 1997, p. 52)

[7] La Dioptrique, Discours huitième, in Descartes 1987, p. 119.

This is not, however, where the story ends. The Archimedes death ray has continued to capture the imagination, but attempts to recreate it have met with mixed success. In particular, the 'Mythbuster' show in the USA failed multiple times to construct a death ray of mirrors that could set a wooded ship ablaze.[8] However, Mythbusters ignored the fact

[8] The last time that Mythbusters tried was at the suggestion of US president Barack Obama.

that in antiquity the wood of ships would have been waterproofed with tar, and that setting fire to the sails would also have been an effective strategy.[9] Nevertheless, modern consensus seems to be that Archimedes' death ray –if it was ever really employed– would not have constituted a serious military deterrent. Details on the Archimedes death ray can be found both online and in the literature (e.g., Africa 1975, Knowles Middleton 1961, Kreyszig 1994, Mills and Clift 1992, Scott 1869).

[9] As an aside, it is strange that Mythbusters did not seek to rebuild Buffon's apparatus, which was well documented to work.

Buffon's Vanity

Buffon was never shy about his accomplishments. Fellows and Milliken (1972) describe the account by one of Buffon's guests at Montbard castle, Marie-Jean Hérault de Séchelles:

> "From the first, Hérault was impressed by his host's singularly frank vanity. Asked immediately which of Buffon's writings he had most recently read, Hérault named the *Vues sur la nature*, and Buffon remarked, "There are in it passages of the most sublime eloquence." Vanity was the shortcoming Hérault commented upon most often in his account of Buffon. Again and again he was frankly flabbergasted by his host's serene confidence in his own immortal genius. Advising Hérault to confine his reading largely to the few, truly great writers that mankind has produced, Buffon listed the five greatest as follows: "Newton, Bacon, Leibnitz, Montesquieu, and Myself." In the end Hérault was more dazzled than amused by this trait. Buffon received a great deal of fan mail from an admiring literary public, kept it all, and showed much of it to Hérault. Confronted by several letters written to Buffon by Catherine the Great of Russia, filled with such assurances as "Newton took the first step, you have taken the second" and "You haven't yet emptied your pockets on the subject of Man," and a similar letter from Prince Henry of Prussia, Hérault enthused, "Glory seemed to take on visible form before my eyes; I felt that I could reach out and touch it, lay my hands upon it, and this admiration from Crowned Heads, compelled to bow down in this way before a greatness in no way specious, pierced my heart, homage of superhuman proportions…." " (Fellows and Milliken 1972, pp. 32–33; fragments taken from Hérault's 'Voyage à Montbard')

Buffon's Needle

After a long introduction on Buffon the man, we have now arrived at the topic of this chapter: Buffon's needle. Just as the Problem of Points, the Problem of the Needle originates from gambling:

> "I suppose that in a room where the floor is simply divided by parallel joints one throws a stick in the air, and that one of the players bets that the stick will not cross any of the parallels on the floor, and that the other in contrast bets that the stick will cross some of these parallels; one

asks for the chances of these two players. One can play this game on a checkerboard with a sewing needle or a headless pin." (Hey et al. 2010, p. 277, translated from Buffon 1777b).

For concreteness, Figure 11.2 shows an example of fictitious results where 100 tosses of a needle result in 41 crosses (in brown), with a needle length that is two-thirds of the distance between two seams (i.e., the width of the plank).[10]

[10] To follow along the reader may activate the *Learn Bayes* JASP module and select 'Buffon's Needle' → 'Simulating Buffon's Needle' and adjust the default settings to match those in the text.

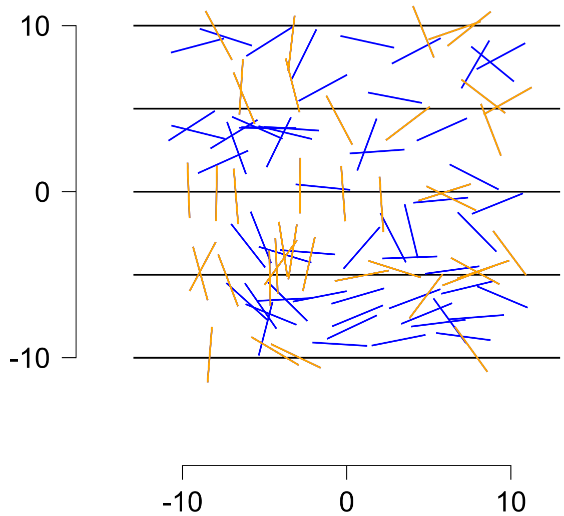


Figure 11.2: One hundred needles are thrown onto a planked floor. The length of each needle equals two-thirds of the distance between two seams. The 41 needles that cross a seam are colored brown, and the 59 needles that do not cross a seam are colored blue. Figure from the JASP module *Learn Bayes*.

Let ℓ be the length of the needle, and $d \geq \ell$ be the distance between two seams. Let θ be the probability that the randomly tossed needle crosses a seam. Buffon showed that

$$\theta = \frac{2 \cdot \ell}{\pi \cdot d}, \tag{11.1}$$

with $\pi \approx 3.14159$ the ratio of a circle's circumference to its diameter.

"What hurt Buffon's mathematical career was surely not a lack of competence or imagination but more likely a certain impatience that did not adapt itself well to the meticulousness of the discipline." (Roger 1997, p. 19)

Laplace (1812, p. 360) later suggested that by actually carrying out the experiment it is possible to obtain an estimate of π. Let $\hat{\theta}$ denote the maximum likelihood point estimate for θ, that is, the fraction of needles that cross a seam. Then the corresponding point estimate $\hat{\pi}$ is obtained as follows:

$$\hat{\pi} = \frac{2 \cdot \ell}{\hat{\theta} \cdot d}. \tag{11.2}$$

If the length of the needle is half of the distance between the seams (i.e., $\ell = 1/2 \cdot d$), the point estimate $\hat{\pi}$ is simply $1/\hat{\theta}$, that is, the total number of tosses divided by the total number of crosses.

At this stage, three misconceptions should be cleared up:

- The express purpose of Buffon was to demonstrate that problems in probability could be solved using geometry (cf. Gorroochurn and Levin 2013, Kendall and Moran 1963). His goal was therefore loftier and more abstract than the solution to the gambling problem may suggest. In fact, Buffon can rightly be considered the father of geometric probability.
- Buffon himself did not estimate π using needle-tossing.[11]
- Buffon's derivation differs from the ones that are usually given in textbooks (Gorroochurn and Levin 2013).

[11] As an aside, Buffon did conduct what is possibly the first experiment in statistics, when he had a child simulate the St. Petersburg paradox by tossing a coin for 2,048 uninterrupted sequences of 'heads'.

The appeal of Buffon's needle is partly in its surprise value: "The fact that π can be approximated from a technique as crude as dropping a needle on the floor will amaze the students every time!" (Schroeder 1974, p. 184). However, the needle also finds practical application. This was already anticipated by Buffon himself:

> "These examples suffice to give an idea of the games that one can imagine on the relationships of size; one could propose several other problems of this type, which do not cease to be interesting and even useful: if one asked, for example, how much one risks passing a river on a more or less narrow plank; what must be the fear one must have of lightning or of a bomb drop, and a number of other problems of conjecture where one must consider only the ratio of the size, and that consequently belong to geometry as much as to analysis." (Hey et al. 2010, p. 279, translated from Buffon 1777a).

In the modern era of science, it has been suggested that Buffon's needle algorithm is used by ants:

> "(...) ants can measure the size of potential nest sites. Nest size assessment is by individual scouts. (...) Experiments indicated that individual scouts use the intersection frequency between their own paths to assess nest areas. These results are consistent with ants using a 'Buffon's needle algorithm' to assess nest areas." (Mallon and Franks 2000, p. 765)

In another example, Newman (1966) showed that the length of a root can be estimated by the number of intersections with random lines:

> "(...) a rectangular area within which some straight lines lie at random. If a root is laid within the area, we should expect that the longer the root the more intersections it will make, on average, with the straight lines. Thus the number of intersections can be used to estimate the length of the root." (Newman 1966, p. 139)

Tossing the Needle: Foul Play?

Neither Buffon nor Laplace actually tossed any needles. However, several later scientists did. Table 11.1 provides an overview, updated from Gridgeman (1960, p. 190). A quick glance at the table suggests that many attempts were relatively successful in approximating π.

Table 11.1: Results from several needle-throwing experiments. NB. $\pi = 3.1415926\ldots$. The data from Wolf (1850) are reported in Edgeworth (1911, p. 387); those of Smith (1855) are reported in De Morgan (1915, p. 283); those of De Morgan (*c.* 1860) are reported in De Morgan (1915, p. 284); those of Fox (1884) are reported in Hall (1872); those of Reina (1925) are reported in Gridgeman (1960) (with the earliest reference to a 1925 work by Castelnuovo); those of Mathematica (2008) are reported by Siniksaran (2008) who used his 'BuffonNeedle' Mathematica program; Padilla (2012) refers to the 'Numberphile' YouTube channel, episode 'Pi and Buffon's Matches'; JASP (2023) refers to the outcome of a computer simulation conducted with the *Learn Bayes* module. The value of $\hat{\pi}$ is computed through Equation 11.2.

Experimenter	Needle length	Tosses	Crosses	$\hat{\pi}$
Wolf (1850)	0.8	5000	2532	3.1596
Smith (1855)	0.6	3204	1218.5	3.1553
De Morgan (*c.* 1860)	1.0	600	382.5	3.137
Fox (1884)	0.75	1030	489	3.1595
Lazzarini (1901)	5/6	3408	1808	3.1415929
Reina (1925)	0.5419	2520	859	3.1795
Gridgeman (*c.* 1960)	0.7857	2	1	3.143
Schroeder (1974)	2/3	100	41	3.3
Mathematica (2008)	0.91	10,000	5855	3.10845
Padilla (2012)	0.5	163	52	3.1346
JASP (2023)	0.75	99,999	47,961	3.1275

However, Gridgeman (1960) was skeptical of some of these earlier tossing experiments, finding their results suspiciously close to the true value. To lampoon these "malodorous" experiments, Gridgeman (1960) proposes the following method to obtain a close approximation with only two tosses:

> "When Laplace wrote, the concept of probability as a limiting frequency was unknown, and the theory of errors was still in parturition. Today we can see that the commonly cited needlecasting trials were not heuristic but teleologic. Out of the casters' zeal has emerged a zero. The sole value remaining in their work is its furnishing material to illustrate paralogy, humbug, and gullibility. But, as H. L. Mencken found when he tried to kill his own bathtub hoax, legend dies hard.[12] Fox and Lazzerini [*sic*] will continue, we may be reasonably sure, to attract laudatory attention for years to come. I can only hope that my own Buffon-Laplace trial will be treated with similar esteem; and, as it is not yet on record, it may appropriately serve as a finale:

[12] EWDM: In 1917, the journalist H. L. Mencken published a history of the American bathtub ("A Neglected Anniversary"). The article was entirely false, but this did not prevent it from being widely cited.

> Handing my pupil a needle, I explained the problem to him. An able and willing youth, he at once bared some floor space and threw the needle down. It fell clear of the edges of the floorboards. He threw again, and this time it fell athwart two boards. Then he measured the boards, which were $3^1/_2$ inches wide, and the needle, which was $2\,^3/_4$ inches long, fetched his slide rule, and presently announced: "I estimate $P = {}^1/_2$, and therefore π to be 3.143." " (Gridgeman 1960, pp. 194-195)

The most suspicious result is that by Mario Lazzarini, probably an Italian math teacher, whose approximation to π is almost spot on.[13] Lazzarini's result has been met with widespread disbelief (e.g., Coolidge 1925, p. 82; Gridgeman 1960; Mantel 1953), and was subjected to a detailed statistical take-down by Badger (1994). Based on Badger's analysis, *Nature* editor John Maddox issued a stern verdict:

[13] As indicated by Mantel (1953, p. 675), Lazzarini's estimate had "an error of only 0.0000003. Terminating the experiment one fall sooner or later would inevitably have lost half the decimal places of accuracy."

> "The truth is that if Lazzarini's result had been published in 1994 and not in 1901, it would be called a barefaced fraud. Indeed, Badger himself, after elegantly demonstrating that Lazzarini's good luck must somehow have been contrived, himself uses the word "hoax" to describe how an even better approximation to π might be obtained. In short, Badger's tale should be a warning to all those who pollute the literature that their misdeeds will follow them to the grave." (Maddox 1994)

Recently, Dutch journalist Hans van Maanen has suggested that Lazzarini was not being serious when he presented his results:

> "Surely it is inconceivable that any of Lazzarini's colleagues took this result seriously? Everything, but everything, points toward a joke, perfectly usable in math classes. Especially when students have just learned the miraculous approximation of pi found by the Chinese mathematician Zu Chongzhi, 355/113, back in the fifth century." (van Maanen 2018; translated to English by DeepL)[14]

[14] According to van Maanen, the suggestion that Lazzarini was joking is supported not just by the extreme precision of the outcome, but also by the fact that Lazzarini claimed to have obtained the data with help of a machine whose operation is physically impossible.

In order to demonstrate that the Lazzarini approximation is too good to be true, we may consider *in advance* how much needles need to be tossed in order to obtain an accurate result. Laplace already showed that the optimal needle length is $\ell{=}d$; thus, if the goal is to determine the value of π as accurately as possible, it is best to select a needle that is as just as long as the distance between the seams is wide (cf. Crofton 1885, p. 784; Todhunter 1865, p. 591; Santaló 1976, p. 72).

Now suppose we toss a needle with optimal length, that is, $\ell{=}d$. Then Gridgeman (1960) approximates the number of tosses required to correctly attain the D^{th} decimal of π in 95% of the cases as 90×10^{2D}. A reasonable shot at correctly identifying the *first* decimal of π therefore already requires about $9{,}000$ tosses:

> "Evidently as many as 10,000 casts could do no more than establish the first decimal place of π with reasonable confidence. We can now tell our waiting needlecaster that if he works at a continuous day-and-night rate of one cast per second for 3 years, his final [results] will yield π to the third decimal." (Gridgeman 1960, pp. 190-191)

Clearly the massive effort required to reach accurate results stands in stark contrast to the modest number of tosses that populate Table 11.1.

Bayesian Inference with Buffon's Needle

Those researchers who conducted a needle-throwing experiment usually report only $\hat{\pi}$, the maximum likelihood point estimate as computed using Equation 11.2. Such a report ignores the uncertainty that accompanies the point estimate. More fundamentally, the report is not Bayesian.

Here we outline a Bayesian analysis as instantiated in JASP. In order to follow along the reader may activate the *Learn Bayes* module and select *Buffon's Needle* $\rightarrow$ *Manipulating Buffon's Needle*. We will first analyze the data reported by Schroeder (1974) (cf. Figure11.2 and Table 11.1): with a needle length of $\ell = 2/3$ Schroeder observed 41 crosses out of 100 tosses, for a point estimate of $\hat{\pi} = 3.3$.

In the JASP interface, we set 'Proportion of needle length to interline distance' to 67%, the 'Number of tosses' to 100, and the 'Number of crosses' to 41. We assign a prior distribution to θ, the probability of any needle crossing a seam. For illustrative purposes, we assign θ a uniform beta(α=1, β=1) prior distribution.[15] The data then cause an update of knowledge that yields a beta$(42, 60)$ posterior distribution for θ, as shown in Figure 11.3.

[15] Additional reflection may suggest prior distributions that are more reasonable.

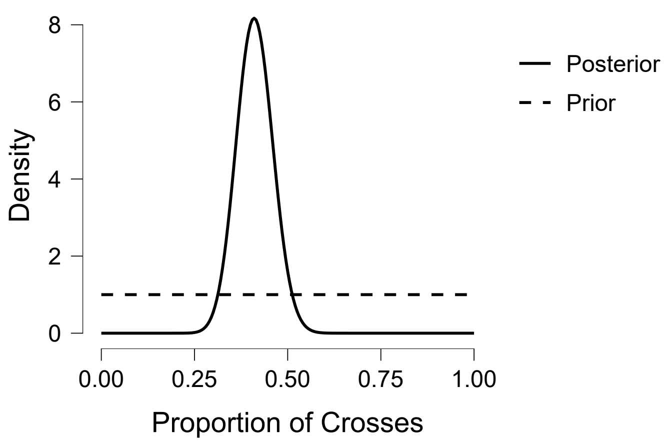


Figure 11.3: Data from the Schroeder (1974) needle-tossing experiment cause an update of beliefs for the proportion θ of needles that cross a seam (i.e., from a uniform beta$(1, 1)$ prior distribution to a beta$(42, 60)$ posterior distribution). A 95% posterior credible interval for θ ranges from 0.32 to 0.51. Figure from the JASP module *Learn Bayes*.

Now assume that we have no knowledge concerning π except for its relation to θ as given by Equation 11.2. This means that our uncertainty about θ translates completely to our uncertainty about π – and this holds both for the prior and for the posterior distribution. These induced distributions of uncertainty for π are shown in Figure 11.4. A 95% posterior credible interval for π extends from 2.64 to 4.24, which is so wide as to render the results almost completely uninformative.

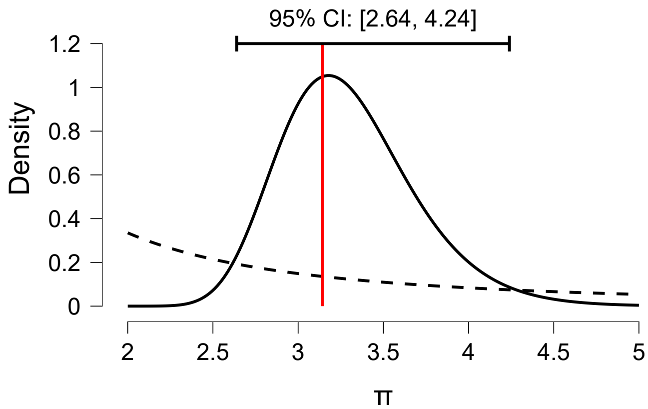


Figure 11.4: Prior and posterior beliefs for π induced by the prior and posterior beliefs for θ shown in Figure 11.3. A 95% posterior credible interval for π ranges from 2.64 to 4.24. The red line indicates the true value of π. Figure from the JASP module *Learn Bayes*.

We may now examine the other needle-tossing results reported in Table 11.1 in similar fashion. For every experiment, we carried out a Bayesian analysis where the proportion of crosses θ was assigned a uniform beta distribution, which was then updated by the data and transformed to the matching posterior distribution for π. Table 11.2 shows the results. By and large, these results confirm the pattern shown in Figure 11.4: the uncertainty is much larger than is suggested by the close correspondence between the point estimates and the true value. Consistent with the analysis of Gridgeman (1960), the JASP simulation with 99,999 virtual tosses is the *only* result that nails the first digit, in the sense that a 95% credible interval falls entirely inside the range from $3.0999\ldots$ to $3.1999\ldots$ (so that we can be more than 95% certain that the true value of π starts with 3.1).

We conclude this chapter with two remarks. Firstly, the preceding analyses assume that you know nothing about π other than its relation to θ given by Equation 11.2; your prior knowledge was therefore expressed in terms of θ – specifically, we assumed that each value of θ was equally likely *a priori*. We designed the inference problem this way in order to demonstrate how uncertainty about one unknown (i.e., parameter θ) can be transformed into uncertainty about a related unknown (i.e., 'parameter' π). However, it may well be that there is advance knowledge about π, and therefore you may wish to assign a prior distribution directly to π (e.g., a uniform distribution from 2 to 4). The appendix to this chapter shows how this can be accomplished using

Table 11.2: Bayesian inference for the needle-throwing experiments listed in Table 11.1. Shown are the maximum likelihood point estimate $\hat{\pi}$, the posterior median for π, and the lower and upper bound of a 95% credible interval for π. The analysis is conducted with a uniform beta prior distribution on the proportion of crosses θ. Fractional outcomes were handled by averaging. Needle proportions were rounded to the nearest integer percentage, which is the main source of discrepancy between $\hat{\pi}$ and the posterior median.

Experimenter	$\hat{\pi}$	Posterior Median	Lower 95% CI	Upper 95% CI
Wolf (1850)	3.1596	3.1596	3.0754	3.2485
Smith (1855)	3.1553	3.1556	3.0213	3.3006
De Morgan (*c.* 1860)	3.137	3.1365	2.9609	3.34085
Fox (1884)	3.1595	3.1596	2.9687	3.3759
Lazzarini (1901)	3.1415929	3.1290	3.0333	3.2312
Reina (1925)	3.1795	3.1687	3.0042	3.3489
Gridgeman (*c.* 1960)	3.143	3.1600	1.6205	63.200
Schroeder (1974)	3.3	3.2731	2.6409	4.2405
Mathematica (2008)	3.10845	3.1084	3.0581	3.1607
Padilla (2012)	3.1346	3.1419	2.5488	4.0015
JASP (2023)	3.1275	3.1275	3.1074	3.1478

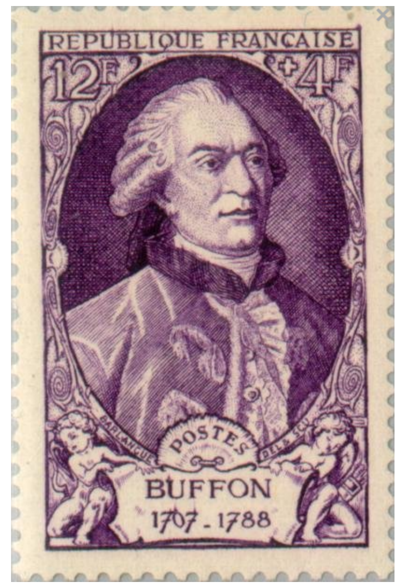
sampling-based inference techniques that we will not cover in the rest of this book.

Secondly, suppose you find yourself confronted with a posterior distribution for π that is as wide as the one shown in Figure 11.4. What should you conclude? Well, the most obvious conclusion is that you are left with a considerable amount of uncertainty about the true value of π. This may prompt you to toss the needle many more times, causing the posterior distribution to become more narrow. One of the wonderful (and often poorly understood) properties of Bayesian inference is that you may quantify your uncertainty at any time during the needle tossing process, and you may stop whenever your uncertainty is sufficiently reduced or you run out of time, money, or patience (whichever comes first; Berger and Wolpert 1988, Edwards et al. 1963, Wagenmakers et al. 2018b).

Exercises

1. You are given money to bet on whether or not a needle, tossed at random, will cross a seam. Find the line length ℓ, expressed as a proportion of the distance between the seams d, which makes you indifferent between betting on the needle crossing vs. not crossing a seam.

2. A needle, half the length of the distance between the seams, is tossed n times, and crosses a seam k times. Using a flat prior on the proportion of crosses, use the *Learn Bayes* module to obtain the probability that the true value of π falls in between 3.130 and 3.150.
3. In the setup discussed in the previous exercise, what beta prior distribution on the proportion of crosses roughly corresponds to the prior knowledge that π is likely to fall in the interval from 3.0 to 3.2?
4. Perlman and Wichura (1975) examine how the data from a Buffon's needle experiment should be analyzed to provide the best (non-Bayesian) estimate of π. Specifically, Perlman and Wichura (1975) "apply the concepts of sufficiency and completeness, efficiency, and ancillarity, in the guise of the Rao-Blackwell-Lehmann-Scheffe theorems [4, 12], the Cramer-Rao lower bound [15], and the principle of conditionality [1, 2, 3, 5], to obtain alternate estimators which utilize the available statistical information as fully as possible." But we are Bayesians, and for Bayesians there is only a single estimator that is possible – which one is it?
5. Consider the needle tossing data from Fox (1884), as reported in Table 11.1. Under consideration is the hypothesis 'did Fox cheat to obtain these results?' Sketch the elements of a Bayesian answer to this question.



Stamp "Comte de Buffon" (N° Yvert & Tellier 856) by George Louis Leclerc. Reproduced with permisson of ©La Poste.

Chapter Summary

It is surprising how many lessons can be learned when tossing a needle on a floor with parallel planks. First, we have learned that π is omnipresent in nature; second, we have learned that researchers are often unable to withstand the Siren song of selective reporting, even when throwing needles on a floor; third, we have learned that when reporting a result, it is crucial to go beyond a point estimate and instead report all of the uncertainty – *the Bayesian estimate is the entire posterior distribution*; fourth, we have learned that uncertainty can be quantified and updated even though the target of inference is itself certain (i.e., it is easy to obtain the first 100 digits of π with the help of a computer; however, from the point of view of uncertainty reduction, this fact is irrelevant if you do not have access to a computer – see also Gronau and Wagenmakers 2018); fifth, we have learned that uncertainty in one unknown can be transformed into uncertainty in a related unknown.

Want to Know More?

✓ When in Paris, we recommend a visit to the beautiful Jardin des Plantes. The grounds cover 28 hectares and includes gardens, a zoo, and four large galleries: The Grande Galerie de l'Évolution, the Galerie de Minéralogie et de Géologie, the Galerie de Paléontologie et d'Anatomie comparée, and the Galerie de Botanique (which contains close to eight million samples of plants).

✓ Buffon spent most of his life in his native village of Montbard, which he much preferred over Paris. Buffon's castle in Montbard is now a museum. The nearby village of Buffon features giant ironworks established by Buffon, the 'Forges de Buffon'.

"Buffon assis dans son fauteuil" (Buffon seated in his armchair). This statue in bronze and stone was created in 1907 by Jean-Marius Carlus (1852–1930) and stands opposite the Grande Galerie de l'Évolution of the Jardin des Plantes, Paris, France (October 2022). The Grande Galerie itself houses yet another statue of Buffon – commissioned by Louis XVI in 1776, it is an exuberant, Greco-Roman marble sculpture by Augustin Pajou. The base of the statue features the inscription "Majestati Naturae par Ingenium" (a genius equal to the majesty of nature) and contains...Buffon's cerebellum! Pajou was also responsible for the bust of Buffon that can be seen overlooking the Rue Buffon.

Poor Joseph

"In Montbard the day started early. Buffon, however, enjoyed his sleep, and early mornings were painful to him. "I loved sleep in my youth," he said of himself, "it relieved me of a lot of time." He tells how, because he was "unhappy with himself," he had asked Joseph, an elderly servant, to wake him before six o'clock, promising him a crown each time he succeeded. One morning, having run out of arguments, Joseph pulled off the bedclothes and poured a bowl of cold water on his master. He received his crown, and Buffon ends the story by saying, "I owe ten to twelve volumes of my works to poor Joseph."[*Corr.*, 1971, I, p. 34.] Buffon was, and would remain until his death, a formidable machine for work: fourteen hours a day for forty years." (Roger 1997, p. 28)

✓ A Shiny app that tosses Buffon's needle and conducts Bayesian inference is available at `https://qfgronau.shinyapps.io/BuffonsNeedle/`.

✓ Another unexpected way to estimate π is presented on the YouTube channel of 3Blue1Brown: "The most unexpected answer to a counting puzzle" (`https://youtu.be/HEfHFsfGXjs`).

✓ Buffon, G.–L. (1749-1788). *Histoire Naturelle Générale et Particulière (Vols. 1-36).* Paris: Imprimerie Royale. With eight additional volumes published posthumously, this encyclopedia is the result of a herculean effort. Verbal descriptions are accompanied by tables with measurements and by detailed engravings.

✓ Buffon, G.–L. (1797-1807). *Buffon's Natural History (Vols. 1-10).* London: T. Gillet. The English translation of the French original.

A Paris metro station is named after Louis Jean-Marie Daubenton (1716–1800), a co-author and close collaborator of Buffon (Roger 1997, p. 337).

✓ Todhunter, I. (1865). *A History of the Mathematical Theory of Probability From the Time of Pascal to That of Laplace.* Cambridge: MacMillan and Co. The go-to reference, authoritative and complete. On p. 347, Todhunter mentions that Buffon solves the parallel lines problem correctly, but provides an incorrect solution for the tiles problem. Buffon also gives the incorrect result for throwing a cube instead of a needle.

✓ Laplace, P.–S. (1812). *Théorie Analytique des Probabilités*. Paris: Courcier. On pp. 359-362, Laplace solves Buffon's needle problem (without mentioning Buffon). Todhunter (1865, pp. 590-591) mentions that in the 1812 first edition, Laplace presents the correct analysis of efficiency (p. 360): the estimation of π is most efficiently achieved when the needle length ℓ equals the distance between the seams d. Curiously, Laplace replaced this correct analysis by an incorrect analysis in the two later editions of his book, "thus causing a change from truth to error" (Todhunter 1865, p. 591).

"Paris is hell" (Buffon, in a letter from 1738, as cited in Roger 1997, p. 30)

✓ Wohl, R. (1960). Buffon and his project for a new science. *Isis, 51*, 186-199. This article clarifies Buffon's vision on science.

> "Probably no figure in the history of the natural sciences is more shrouded in ambiguity than Georges Louis Leclerc de Buffon. The uncertainty of his present reputation stands in all the greater contrast with the eminence he attained in his own age. Philosopher of Nature, biologist, anthropologist, mathematician, translator of Newton, entrepreneur and builder of iron forges, haughty administrator of the *Jardin du Roi*, austere academician – Buffon was one of the most famous of savants in a century that esteemed intellect above all other virtues. Yet even in his own time Buffon's claim to scientific stature was severely questioned. Despite the vast popularity of his work – or, even more likely, because of it – many of Buffon's fellow scientists thought his forty-four volume *Histoire naturelle* more a romance for young ladies than a serious contribution to natural history." (p. 186)

✓ Doron, C.–O. (2012). Race and genealogy. Buffon and the formation of the concept of "race". *Humana. Mente Journal of Philosophical Studies, 22*, 75–109. Doron argues that racism was an integral part of Buffon's philosophy. Additional sources include Roger (1997, p. 178 and pp. 181-182), Fellows and Milliken (1972, p. 140-141), and (for instance) Buffon (1797–1807, pp. 38-39, Vol. 7 and pp. 6-7, Vol. 8).

✓ Dugatkin, L. A. (2019). Buffon, Jefferson and the theory of New World degeneracy. *Evolution: Education and Outreach, 12*, 15. A modern reader may be puzzled by Buffon's strong opinion on animals, habits, places, and people with whom he was almost entirely unfamiliar. A good example is the *Theory of New World Degeneracy*.[16] The main premise of this 'theory' is perhaps best explained by Buffon himself:

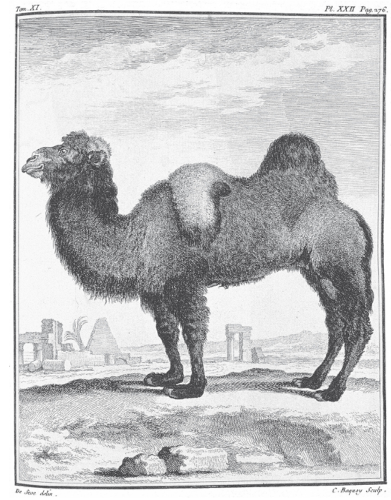
[16] See also `https://www.americanscientist.org/article/jefferson-buffon-and-the-moose`.

> "Horses have multiplied nearly as much in the hot as in the cold countries throughout America; but have diminished in size, a circumstance which is common to all animals transported from Europe to America; and what is still more singular, all the native animals of America are much smaller in general than those of the old continent. Nature in their formation seems to have adopted a smaller scale, and to have formed man alone in the same mould." (Buffon 1797–1807, p. 15, Vol. 7)

and

> "Animated nature, therefore, is in this portion of the globe less active, less varied, and even less vigorous; for by the enumeration of the American animals we shall perceive, that not only the number of species is smaller, but that in general they are inferior in size to those of the old continent; not one animal throughout America can be compared to the elephant, rhinoceros, hippopotamus, dromedary, buffalo, tiger, lion, &c." (Buffon 1797–1807, p. 27, Vol. 7)

Buffon's put down of the New World greatly annoyed Thomas Jefferson, who issued a rebuttal of Buffon's claims in his 'Notes on the State of Virginia' (Jefferson 1787). Conversations with Franklin and Jefferson ultimately had Buffon abandon the theory of New World degeneracy (see also Fellows and Milliken 1972, p. 146).



Le chameau – the camel. Illustration from the eleventh volume of Buffon's *Histoire naturelle générale et particulière avec la description du Cabinet du Roi* (1764, p. 276). Design by Jacques De Sève, engraving by Pierre Charles Baquoy. Source: `https://books.openedition.org/mnhn/3079`.

✓ Eymard, P., & Lafon, J.-P. (2004). *The Number π*. Providence, Rhode Island: American Mathematical Society. Everything you always wanted to know about π.

✓ Hey, J. D., Neugebauer, T. M., & Pasca, C. M. (2010). George-Louis Leclerc de Buffon's 'Essays on moral arithmetic'. In Ockenfels, A., & Sadrieh, A. (Eds.), *A Collection of Essays in Honor of Reinhard Selten*, pp. 245-282. Berlin: Springer. An English translation of Buffon's published needle work (Buffon 1777b). The relevant part is article XXIII, pp. 275-279.

✓ Velasco, S., Román, F. L., González, A., & White, J. A. (2006). Statistical estimation of some irrational numbers using an extension of Buffon's needle experiment. *International Journal of Mathematical Education in Science and Technology, 37*, 735-740. "(...) replacing the needle by a square, a regular pentagon and a regular hexagon in Buffon's experiment will give an estimate of $\sqrt{2}$, the golden ratio, $\Phi = (1+\sqrt{5})/2$, and $\sqrt{3}$, respectively."

✓ We include this note for completeness. If the uncertainty for $\theta = 2 \cdot \ell/(\pi \cdot d)$ is quantified by a beta(α, β) distribution, then the corresponding uncertainty for π is quantified by a beta-prime(β, α) distribution (shifted and scaled). In particular, we have that

$$p(\pi \mid \ell, d) = \frac{2\ell}{\pi^2 d} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \left(\frac{2\ell}{\pi d}\right)^{\alpha-1} \left[1 - \frac{2\ell}{\pi d}\right]^{\beta-1},$$

where Γ denotes the gamma function ; for integer n, $\Gamma(n) = (n-1)!$.

✓ Many articles and books discuss Buffon's needle (and geometric probability more generally). For instance, Crofton (1885, p. 784) contains an early summary; Ellenberg (2014, pp. 202-212) provides an intuitive derivation; Gorroochurn and Levin (2013) provides historical background; Arnow (1994) describes Laplace's solution to an extension where the needle is tossed on a floor with rectangular tiles; Kendall and Moran (1963, pp. 70-77) discuss various extensions; Ramaley (1969) considers tossing a wet noodle instead of a straight needle; Schroeder (1974) presents a clear proof and some example data; Solomon (1978) describes a series of estimators of varying efficiency; Perlman and Wichura (1975) compare different estimators and designs; they conclude that for estimating π, a tiled, double-grid floor is more efficient than a parallel plank, single-grid floor, but less efficient than a diamond-patterned, triple-grid floor; Wood and Robertson (1998) expand on the previous result by adding the honeycomb, hexagonal grid; after enforcing equal grid density, the single grid turns out to yield the most efficient estimate of π when $\ell \geq 0.9\,d$; finally, Siniksaran (2008) introduces the Mathematica program 'BuffonNeedle' that tosses a digital needle in different scenarios.

Honor

"Western civilization has known few men of letters and even fewer scientists who were as singularly honored during their lives as France's Georges-Louis Leclerc, comte de Buffon, scientist and belletrist, whose published work seemed to lay claim to science as a new branch of the humanities. So unimpeded was his rise to fame, so great the weight of his authority, so imposing his very presence, that few among his contemporaries dared to attack him frankly and openly. It seemed far easier to honor him, and he was showered with honors." (Fellows and Milliken 1972, p. 15)

Appendix: An Excursion to MCMC

In this chapter we assigned a prior distribution to the proportion of crosses θ, updated it by means of the data, and then translated the posterior uncertainty to π, which was the target of inference. But what if we wanted to assign a prior to π directly? This prior may take on all sorts of shapes, but for simplicity let's say that all we are happy to assume is that π falls in the interval from 2 to 4, and that every value inside this interval is equally likely. This knowledge does not translate to a beta prior on the proportion of crosses θ exactly. But it is nevertheless possible to use JASP and assign a prior to π directly – we just cannot do it using the *Learn Bayes* module. Instead, we have to use the *JAGS* module.

The *JAGS* module is based on the 'JAGS' program (Plummer 2003), which itself was inspired by the 'BUGS' program (Lunn et al. 2012). Both JAGS and BUGS are probabilistic programming languages. They allow users to specify how the data are generated, and what the prior distributions are on the model parameters. With the model specified and the data given, JAGS and BUGS are then able to obtain the resulting posterior distributions – not by deriving them analytically, but by repeatedly drawing samples from them. The histogram of those samples approximates the analytical result to any desired degree of accuracy (i.e., more accurate results can always be obtained by drawing more values). This sampling process is known as 'Markov chain Monte Carlo' (MCMC), and it has transformed the field of Bayesian statistics from the 1990s onward.[17] At first glance, MCMC may come across as dark magic: if we cannot express the posterior distribution analytically, how can we draw samples from it?

[17] It is often remarked that with MCMC, Bayesian model specification is limited only by the user's imagination.

Here we take the dark magic for granted and illustrate the flexibility of MCMC with an example. For concreteness, we will analyze the Schroeder (1974) needle-tossing experiment with $\ell = {}^2\!/_3 \cdot d$ that yielded 41 crosses out of 100 tosses, for a point estimate of $\hat{\pi} = 3.3$. We open the *JAGS* module in JASP and specify the following model code in the syntax window:

```
model{
L <- 2; d <- 3;
mypi    ~ dunif(2,4)
crosses ~ dbinom(theta,tosses)
theta  <- 2*L/(d*mypi)
}
```

The first line of this code[18] specifies the line length as $\ell = {}^2\!/_3 \cdot d$; the second line assigns π (called 'mypi' in the code, to avoid confusion with the true value of π) a prior distribution that is uniform from 2 to 4, as desired; the third line indicates that 'crosses' follows a binomial distribution dictated by chance parameter θ and 'tosses'; the fourth line provides the relation between the binomial chance parameter θ and π. Note that the assignment operator $\leftarrow$ specifies a deterministic relationship (i.e., 'is given by') whereas the tilde operator $\sim$ specifies a stochastic relationship (i.e., 'is distributed as').

[18] Consistent with common coding practice, we write the letter 'l' in upper case to avoid visual confusion with the digit '1'.

After specifying the syntax, JAGS needs to be informed about the values for 'crosses' and 'tosses'. Open the tab 'Observed Values', define 'crosses' as 41 and 'tosses' as 100. Return to the syntax window and press 'control $+$ enter' to run the analysis. Then go to the box 'Parameters in model' and select 'mypi' as the parameter for which results should be shown. We then open the tab 'Plots' and select 'Histogram'. The results ought to be similar to those displayed in Figure 11.5. The output table gives the posterior median as 3.2944, and a 95% credible interval ranging from 2.6885 to 3.9213. The histogram is based on the $6{,}000$ MCMC draws from the posterior distribution. Note that all samples obey the prior restriction that π lies in between 2 and 4.

As a fun aside, the JAGS code can easily be adjusted to address a slightly different (and arguably more useful) problem: suppose we know the value of π exactly, but we wish to learn ℓ/d, the needle length ℓ ex-

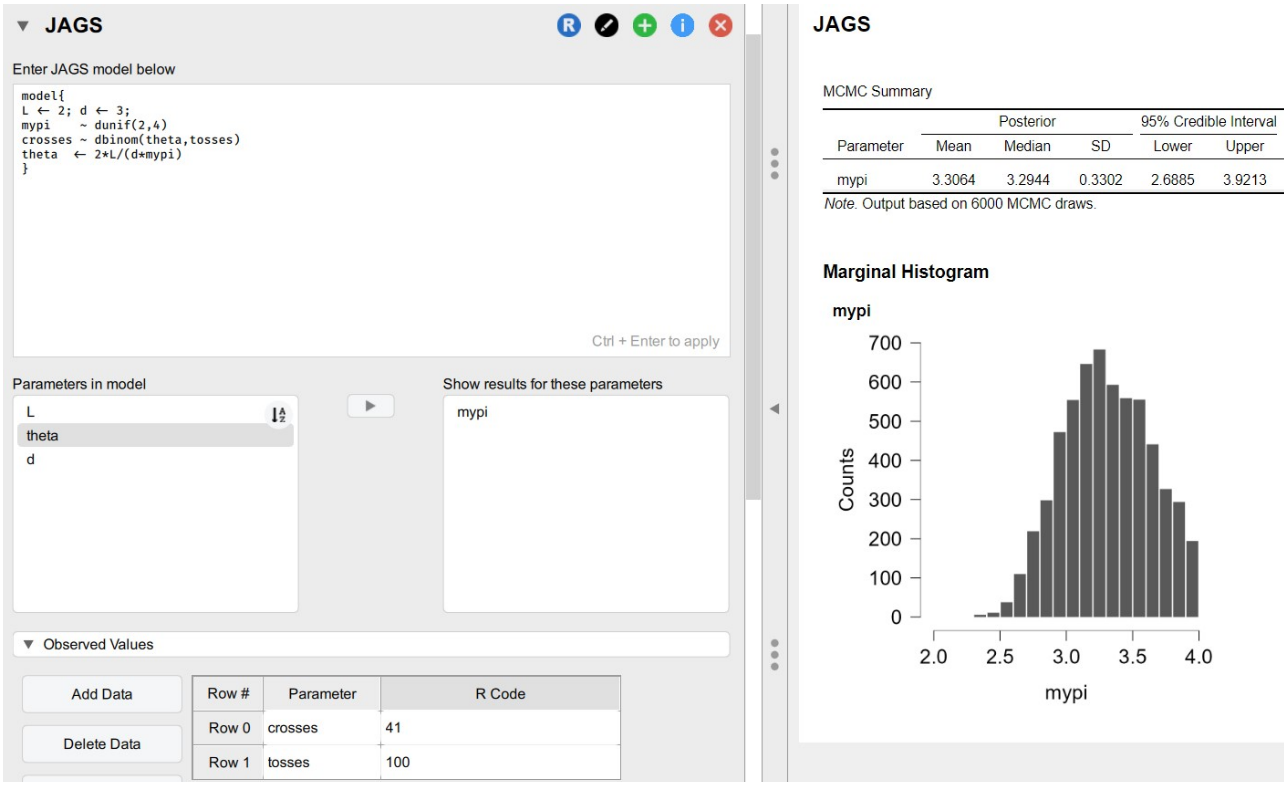


Figure 11.5: Bayesian MCMC-style inference for π based on the needle-tossing experiment from Schroeder (1974). Left input panel: the JAGS model syntax assigns a uniform prior directly to π. Right output panel: the samples are plotted as a histogram. Note that the samples respect the restriction imposed by the prior distribution (i.e., there are no samples that exceed 4). Screenshot from the *JAGS* module in JASP.

pressed as a proportion of the interseam distance d. In order to achieve our goal we can arbitrary set $d = 1$ and assign ℓ a prior distribution on the $0-1$ interval, which is then updated to a posterior distribution based on the observed number of tosses and crosses:

```
model{
d <- 1
L ~ dunif(0,1)
pi <- 3.14159265359
crosses ~ dbin(theta,tosses)
theta <- (2*L)/(d*pi)
}
```

The second line assigns ℓ a uniform prior; if strong prior knowledge is available we might prefer an informed beta distribution instead. The third line yields an approximate value for π; if a more precise value is needed we can use the following expression instead:

```
pi <- 4 * atan(1)
```

Executing this code will yield posterior samples for ℓ. The examples from this appendix serve to illustrate how probabilistic programming languages allow users to change their models almost at will, without first having to do the mathematical derivations.[19]

We have hardly scratched the surface of MCMC sampling, and we will not return to it in this book. Excellent resources on MCMC are available both online and in the literature.

[19] This should not be interpreted as an invitation to spurn mathematics.

12 The Pancake Puzzle [with Charlotte Tanis]

When two persons who consider themselves equally competent assign different subjective probabilities to certain gambles and one can observe them a sufficient number of times, it is often possible to decide which of the two is superior so far as their judgement is concerned.

Borel, 1909/1965

Chapter Goal

This chapter showcases the predict-update Bayesian learning cycle for a real-life binomial data set involving eight pancakes. We emphasize the predictive aspect of the learning cycle by first having individual people assign a prior beta distribution to the chance θ that any pancake will come with bacon. Each individual person therefore acts as a probabilistic bacon forecaster, with their beta prior as the quantitative device to formalize the forecasts. As the pancakes accumulate, consecutive prediction errors drive a continual adjustment of beliefs, such that the posterior distribution after the nth pancake becomes the prior distribution for pancake $n + 1$. The predict-update cycle is first shown for a single forecaster, and then for several rival forecasters. Bayes' rule specifies how the relative adequacy of the individual forecasters can be quantified, and how one may arrive at a joint prediction by computing a weighted average across all forecasters.

Data collection in action.

The Problem

One of us [EJ] was going to bake pancakes for his family. From the sample proportion of bacon pancakes we wish to learn about EJs *bacon proclivity* θ_{EJ}, that is, the probability that any one of his pancakes will have bacon. We also wish to predict whether future pancakes will have bacon.

A Standard Solution

The observed sequence of pancakes was as follows: $y = \{v, v, v, b, b, v, b, v\}$, where '$v$' stands for a 'vanilla' pancake and b stands for a bacon pancake. So EJ baked eight pancakes, three of them with bacon. We may adopt Laplace's Principle of Insufficient Reason (see Chapter 8) and assign a uniform prior distribution to the chance θ_{EJ} that any pancake comes with bacon (i.e., $\theta \sim$ beta$(1, 1)$). Updating this prior distribution with the observed data y yields a beta$(4, 6)$ posterior distribution, which is depicted in Figure 12.1. The mean of this posterior distribution is $^4/_{10}$, which also equals the probability that the next pancake will come with bacon (see the 'beta prediction rule' outlined in Chapter 9). To summarize the posterior distribution we may, for instance, report that the 95% central credible interval ranges from $.14$ to $.70$. We may also compute the posterior probability that θ_{EJ} lies in any interval of interest (e.g., $p(\theta_{EJ} \in [.4, .6] \,|\, y) \approx .38$) or the posterior probability that θ_{EJ} is larger than $^1/_2$ (i.e., $p(\theta_{EJ} > {}^1/_2 \,|\, y) \approx .25$).

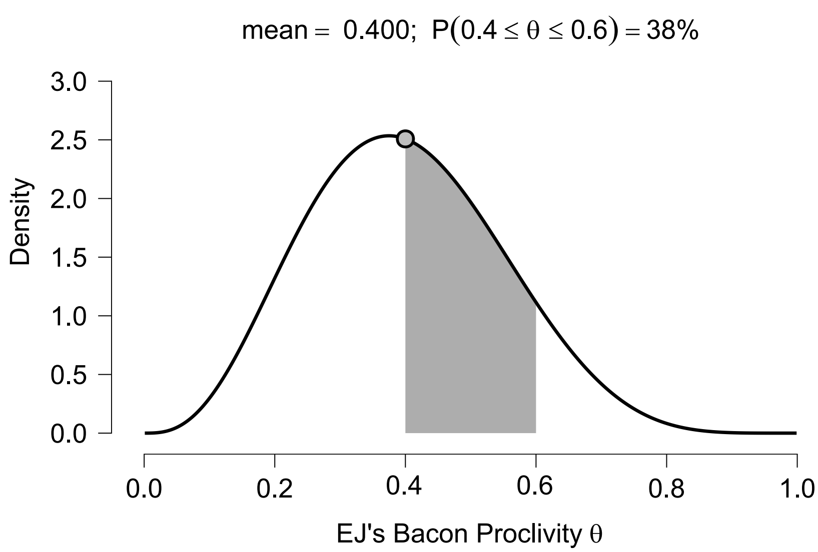


Figure 12.1: Standard solution for Bayesian inference about EJ's bacon proclivity θ_{EJ}. A uniform beta(1,1) prior has been updated by the data (i.e., three bacon pancakes, five vanilla pancakes) to a beta(4,6) posterior distribution. The posterior mean is $^4/_{10}$, which, by the beta prediction rule outlined in Chapter 9, is also the probability that the next pancake will have bacon. The gray area visualizes the posterior probability that θ_{EJ} is in between .40 and .60. Figure from the JASP module *Learn Bayes*.

Below we explore the consequences of (1) assigning θ_{EJ} an informed beta prior distribution rather than the Laplacean flat beta(1,1) distribution; (2) updating the informed prior distribution one pancake at a time; (3) contrasting and combining multiple rival informed prior distributions, which may be considered as competing forecasting systems.

Sequentially Updating an Informed Prior

As part of a course assignment, all 34 students (henceforth *forecasters*) in our 2019 Research Master class 'Bayesian inference for psychological science' each had to specify and motivate their own 'informed' beta prior for EJ's bacon proclivity θ_{EJ}, before learning the outcome of his pancake dinner. The 34 informed beta priors are listed in Appendix A of this chapter. For educational purposes, here we focus on just four forecasters: Tabea, Sandra, Elise, and Vukasin. Their beta priors and posteriors are listed in Table 12.1 and shown in Figure 12.2.

Table 12.1: Informed beta priors for EJ's bacon proclivity θ_{EJ}, and their associated posteriors after updating with the data (i.e., three bacon pancakes out of eight total), for four forecasters.

	Beta prior		Beta posterior	
Forecaster	α	β	α	β
Tabea	4	4	7	9
Sandra	4	7	7	12
Elise	9	3	12	8
Vukasin	10	1	13	6

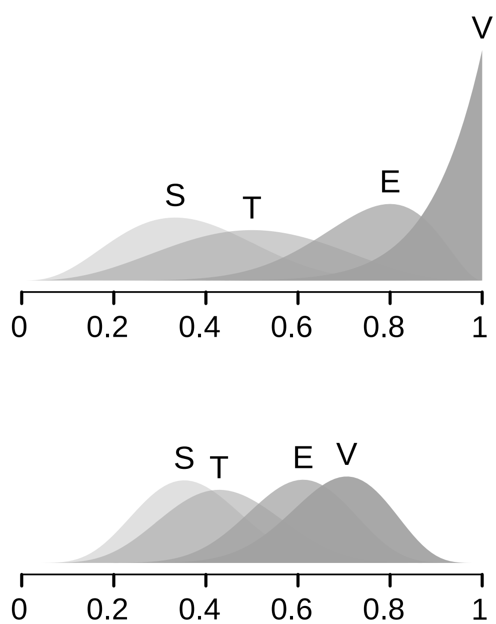


Figure 12.2: Prior and posterior beta distributions for EJ's bacon proclivity θ_{EJ}. The top panel shows the beta priors for Sandra ('S'), Tabea ('T'), Elise ('E'), and Vukasin ('V'). The bottom panel shows the beta posteriors based on updating the priors with the information in the sample (i.e., three bacon pancakes and five vanilla pancakes, for a bacon sample proportion of $3/8 = .375$). See also Table 12.1.

Here we first demonstrate the details of the sequential updating process with one of the prior distributions, the beta$(4, 4)$ prior by Tabea. The Tabea-prior pancake-by-pancake updating process is shown in Figure 12.3 and it proceeds from top to bottom. The top distribution is Tabea's beta(4,4) prior, and the bottom distribution is her beta(7,9) posterior distribution after having observed all eight pancakes. The rows in between visualize the intermediate beta distributions that obtain when the observed pancake sequence $y = \{v, v, v, b, b, v, b, v\}$ is encountered and analyzed one pancake after the other. For instance, the second row shows a beta(4,5) distribution: Tabea's posterior distribution after learning that the first pancake is vanilla. Note that each vanilla pancake pulls the distribution to the left, whereas each bacon pancake pulls it to the right. Also note that, as the pancakes accumulate, the distributions tend to become more narrow, signifying increased confidence about the most plausible values of θ_{EJ}.

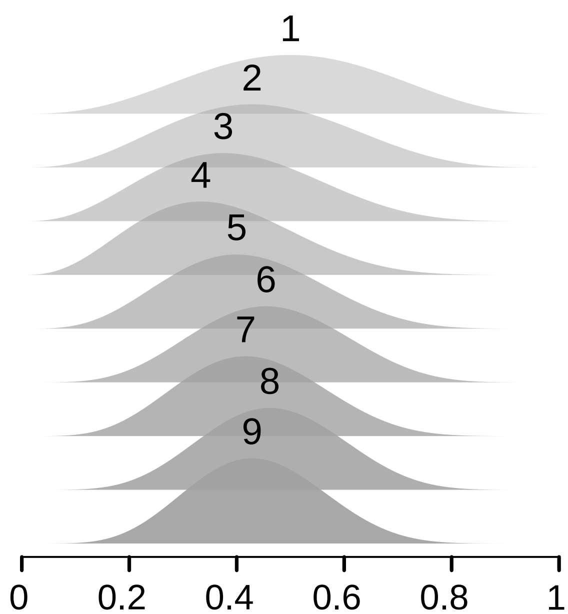


Figure 12.3: The Tabea-prior pancake-by-pancake updating process. The distribution on top is Tabea's beta(4,4) prior. The rows below show the updated beta distributions when going through the observed pancake sequence $y = \{v, v, v, b, b, v, b, v\}$ one pancake at a time. For instance, the second row gives the beta(4,5) posterior distribution after observing that the first pancake was vanilla, and the bottom row is the final beta(7,9) distribution after having observed all eight pancakes.

The same updating process is shown in Table 12.2, but here we also show the predictive success for Tabea at each step. For instance, before observing the first pancake, Tabea's belief about θ_{EJ} was quantified by a beta(4,4) prior distribution. From the beta prediction rule (Chapter 9) it follows that the predicted probability is $^4/_8$ for the occurrence of a bacon pancake and $4/8$ for the occurrence of a vanilla pancake. A vanilla pancake is observed, and this means the predictive success for the observed data is $1/2$ (i.e., right-most column, 'Probability'). The observation that the first pancake is vanilla also leads to an update of the beta(4,4) prior distribution to a beta(4,5) posterior distribution. This posterior distribution is the prior distribution before the arrival of the second pancake. From this beta(4,5) prior distribution it follows that the predicted probability is $4/9$ for the occurrence of a bacon pancake and $5/9$ for the occurrence of a vanilla pancake. The second pancake turns out to be vanilla, and this means the predictive success for the observed data is $5/9$. This process is repeated until all eight pancakes have been observed. The total predictive score is $1/2 \times 5/9 \times 6/10 \times 4/11 \times 5/12 \times 7/13 \times 6/14 \times 8/15 = 4/1287 \approx .0031$.

Table 12.2: The predict-update sequential analysis of Tabea's beta prior based on the pancake order $\{v, v, v, b, b, v, b, v\}$. Predictions for the next pancake are based on the beta prediction rule outlined in Chapter 9. Eight pancakes were baked, so the row for the ninth pancake contains a prediction but no outcome.

Pancake	Prior	Prediction	Outcome	Probability
1	beta(4,4)	$p(\{b\}) = 4/8$		
		$p(\{v\}) = 4/8$	vanilla	$1/2$
2	beta(4,5)	$p(\{b\}) = 4/9$		
		$p(\{v\}) = 5/9$	vanilla	$5/9$
3	beta(4,6)	$p(\{b\}) = 4/10$		
		$p(\{v\}) = 6/10$	vanilla	$6/10$
4	beta(4,7)	$p(\{b\}) = 4/11$	bacon	$4/11$
		$p(\{v\}) = 7/11$		
5	beta(5,7)	$p(\{b\}) = 5/12$	bacon	$5/12$
		$p(\{v\}) = 7/12$		
6	beta(6,7)	$p(\{b\}) = 6/13$		
		$p(\{v\}) = 7/13$	vanilla	$7/13$
7	beta(6,8)	$p(\{b\}) = 6/14$	bacon	$6/14$
		$p(\{v\}) = 8/14$		
8	beta(7,8)	$p(\{b\}) = 7/15$		
		$p(\{v\}) = 8/15$	vanilla	$8/15$
9	beta(7,9)	$p(\{b\}) = 7/16$	?	
		$p(\{v\}) = 9/16$	?	

We now compute the predictive score for all pancakes at once, using the beta-binomial distribution. The beta-binomial distribution gives the probability of observing k successes out of n trials, given that the binomial chance parameter θ follows a beta distribution with parameters α and β. Applying the beta-binomial with $k = 3$, $n = 8$, and $\alpha = \beta = 4$, we find that the probability that is returned equals .174, much larger than the value of $.0031$ obtained from Table 12.2.[1] The discrepancy occurs because the beta-binomial takes into account that the three bacon pancakes and five vanilla pancakes could be arranged in any order. As explained in Chapter 33, 'Jevons Explains Permutations', the possible number of different orders is 56.[2] When we multiply the number of orders with Tabea's predictive score, we obtain $56 \times {}^{4}/_{1287} = {}^{224}/_{1287} \approx .174$, which matches the result from the beta-binomial.

[1] Example R code:
```
library(extraDistr);N.bacon<-3;
N.total<-8; alpha<-4; beta<-4;
dbbinom(N.bacon,N.total,alpha,beta).
```

[2] That is, ${}^{8!}/_{(3!\,5!)}$.

The result can also be obtained from the JASP *Learn Bayes* module. Go to 'Counts' → 'Binomial Testing'. Enter the observed data and specify Tabea's beta(4,4) prior under 'Hypothesis'. Then, under 'Predictive Performance', select 'Prior predictive distribution'. To highlight the data that were actually observed, also tick 'Observed number of successes'. The result is shown in Figure 12.4.

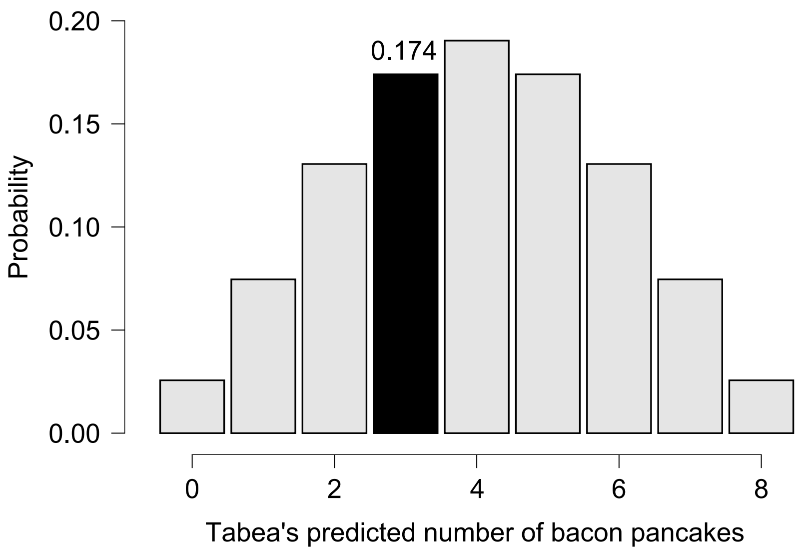


Figure 12.4: Tabea's predicted number of pancakes that come with bacon, out of a total of eight. The beta-binomial predictions are based on Tabea's beta(4,4) prior distribution on θ_{EJ}. The highlighted bar corresponds to the observed data and its height, 0.174, quantifies Tabea's predictive success. Figure from the JASP module *Learn Bayes*.

As we have discussed in previous chapters, the end-result of the Bayesian updating process does not depend on the specific order of the observations. This can be seen immediately from the fact that s successes and f failures update a beta(α, β) prior distribution for a binomial chance θ to a beta$(\alpha + s, \beta + f)$ posterior distribution – the end result depends only on the total numbers s and f, not their order. A concrete demonstration of this fact is offered in Table 12.3, which shows the sequential updating steps for an alternative pancake order, namely $\{b, b, v, v, v, v, v, b\}$. We note that the final posterior is a beta$(7, 9)$ distribution, as was the case for the original order. Also, for the original order the overall predictive success was $1/2 \times 5/9 \times 6/10 \times 4/11 \times 5/12 \times 7/13 \times 6/14 \times 8/15 = 4/1287 \approx .0031$. For the shuffled order, the total predictive score is $1/2 \times 5/9 \times 4/10 \times 5/11 \times 6/12 \times 7/13 \times 8/14 \times 6/15 = 4/1287 \approx .0031$: many individual elements in the multiplication differ, but the end result is identical.

Table 12.3: The predict-update sequential analysis of Tabea's beta prior based on a different pancake order, namely $\{b, b, v, v, v, v, v, b\}$. The end-result is identical to that of the original order.

Pancake	Prior	Prediction	Outcome	Probability
1	beta(4,4)	$p(\{b\}) = 4/8$	bacon	$1/2$
		$p(\{v\}) = 4/8$		
2	beta(5,4)	$p(\{b\}) = 5/9$	bacon	$5/9$
		$p(\{v\}) = 4/9$		
3	beta(6,4)	$p(\{b\}) = 6/10$		
		$p(\{v\}) = 4/10$	vanilla	$4/10$
4	beta(6,5)	$p(\{b\}) = 6/11$		
		$p(\{v\}) = 5/11$	vanilla	$5/11$
5	beta(6,6)	$p(\{b\}) = 6/12$		
		$p(\{v\}) = 6/12$	vanilla	$6/12$
6	beta(6,7)	$p(\{b\}) = 6/13$		
		$p(\{v\}) = 7/13$	vanilla	$7/13$
7	beta(6,8)	$p(\{b\}) = 6/14$		
		$p(\{v\}) = 8/14$	vanilla	$8/14$
8	beta(6,9)	$p(\{b\}) = 6/15$	bacon	$6/15$
		$p(\{v\}) = 9/15$		
9	beta(7,9)	$p(\{b\}) = 7/16$	?	
		$p(\{v\}) = 9/16$	?	

A Rival Forecaster

We now consider a rival forecaster, Elise, who had assigned θ_{EJ} a beta(9,3) prior (cf. Figure 12.2). Similar to our pancake-by-pancake analysis of Tabea, Table 12.4 shows the updating process for Elise's prior. As the table shows, we start with a beta(9,3) prior and finish with a beta(12,8) posterior distribution. This updating process is accompanied by a total predictive score of $3/12 \times 4/13 \times 5/14 \times 9/15 \times 10/16 \times 6/17 \times 11/18 \times 7/19 = 2494800/3047466240 = 55/67184 \approx .0008$. As was the case for Tabea, this result is for a specific pancake order; because there are 56 different orders of three bacon pancakes and five vanilla pancakes, the predictive score for Elise in terms of the number of bacon pancakes, irrespective of the pancake order, is $56 \times 55/67184 = 385/8398 \approx .046$.

This result can be confirmed using the JASP *Learn Bayes* module. As before, go to 'Counts' $\rightarrow$ 'Binomial Testing'. Enter the observed data and specify Elise's beta(9,3) prior under 'Hypothesis'. Under 'Predictive Performance', select 'Prior predictive distribution' and also tick 'Observed number of successes'. The result is shown in Figure 12.5.

Table 12.4: The predict-update sequential analysis of Elise's beta prior based on the pancake order $\{v, v, v, b, b, v, b, v\}$. Predictions for the next pancake are based on the beta prediction rule outlined in Chapter 9. Eight pancakes were baked, so the row for the ninth pancake contains a prediction but no outcome.

Pancake	Prior	Prediction	Outcome	Probability
1	beta(9,3)	$p(\{b\}) = 9/12$		
		$p(\{v\}) = 3/12$	vanilla	$3/12$
2	beta(9,4)	$p(\{b\}) = 9/13$		
		$p(\{v\}) = 4/13$	vanilla	$4/13$
3	beta(9,5)	$p(\{b\}) = 9/14$		
		$p(\{v\}) = 5/14$	vanilla	$5/14$
4	beta(9,6)	$p(\{b\}) = 9/15$	bacon	$9/15$
		$p(\{v\}) = 6/15$		
5	beta(10,6)	$p(\{b\}) = 10/16$	bacon	$10/16$
		$p(\{v\}) = 6/16$		
6	beta(11,6)	$p(\{b\}) = 11/17$		
		$p(\{v\}) = 6/17$	vanilla	$6/17$
7	beta(11,7)	$p(\{b\}) = 11/18$	bacon	$11/18$
		$p(\{v\}) = 7/18$		
8	beta(12,7)	$p(\{b\}) = 12/19$		
		$p(\{v\}) = 7/19$	vanilla	$7/19$
9	beta(12,8)	$p(\{b\}) = 12/20$	?	
		$p(\{v\}) = 8/20$	?	

Who Predicted Better?

So far we have considered two forecasters, Tabea and Elise, and it may be of interest to compare their predictive performance. Similar to the scenario discussed in Chapter 10, The Problem of Points, there may be a stake to divide –a prize for the best bacon forecaster– and it seems fair to divide that stake in proportion to the forecasters' relative predictive success for the past pancakes. Also, we might need to hire a single bacon forecaster – whom should we pick, and how confident should we be about our choice? Finally, as we will elaborate upon later, we might desire a forecast for unseen pancakes that is a weighted average of the individual forecasts from Tabea and Elise, with averaging weights determined by past predictive performance (cf. Figure 7.4).

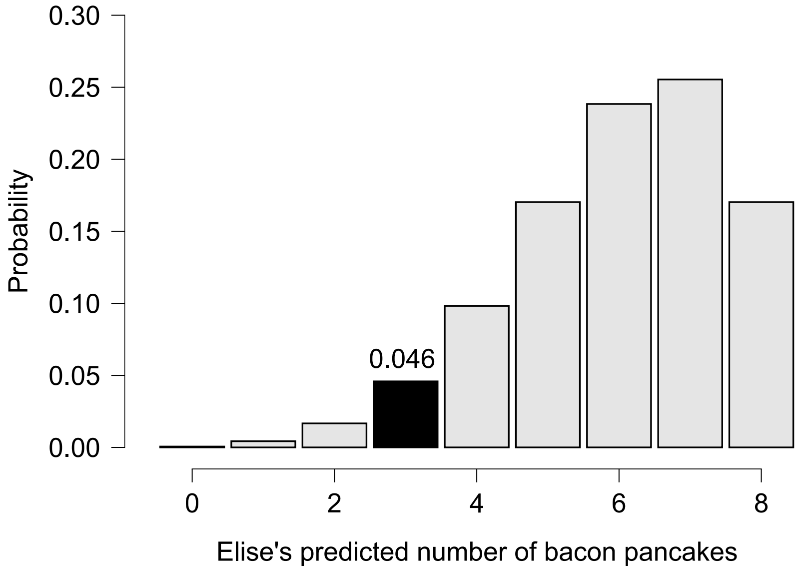


Figure 12.5: Elise's predicted number of pancakes that come with bacon, out of a total of eight. The beta-binomial predictions are based on Elise's beta(9,3) prior distribution on θ_{EJ}. The highlighted bar corresponds to the observed data and its height, 0.046, quantifies Elise's predictive success. Figure from the JASP module *Learn Bayes*.

As indicated above, the predictive score for Tabea is $.174$ (cf. Figure 12.4), whereas the predictive score for Elisa is $.046$ (cf. Figure 12.5). We conclude that Tabea outpredicted Elise by a factor of $^{.174}/_{.046} = 3.78$. Formally, we can use the odds form of Bayes' rule and write

$$\underbrace{\frac{p(\text{Tabea} \mid y)}{p(\text{Elise} \mid y)}}_{\text{Posterior odds}} = \underbrace{\frac{p(\text{Tabea})}{p(\text{Elise})}}_{\text{Prior odds}} \times \underbrace{\frac{p(y \mid \text{Tabea})}{p(y \mid \text{Elise})}}_{\text{Evidence}}. \tag{12.1}$$

The 'Evidence' in this equation is the degree to which the data change our beliefs about the relative ability of the rival forecasters: the change from prior to posterior odds. This change is generally known as the *Bayes factor* and here it equals the extent to which Tabea outpredicted Elise.[3] In the present example, each forecaster's predictive performance is obtained by averaging predictive performance over the possible values of the binomial chance parameter, with the prior distributions providing the averaging weights.[4] For this particular example we therefore have

$$\begin{aligned} \underbrace{\frac{p(y \mid \text{Tabea})}{p(y \mid \text{Elise})}}_{\text{Evidence}} &= \frac{\int p(y \mid \theta)\, p(\theta)\, \mathrm{d}\theta}{\int p(y \mid \zeta)\, p(\zeta)\, \mathrm{d}\zeta}, \quad \theta \sim \text{beta}(4,4), \quad \zeta \sim \text{beta}(9,3) \\ &\approx \frac{0.174}{.046} = 3.78. \end{aligned}$$

[3] When the forecasters base their predictions on a single value for EJ's bacon proclivity, the Bayes factor reduces to the likelihood ratio.

[4] As explained in Chapter 9, the averaging step is the statistical underpinning for the beta-binomial predictions shown in Figure 12.4 and 12.5.

Four Forecasters

We now return to our initial scenario, summarized in Table 12.1, which features *four* rival forecasters: Tabea, Sandra, Elise, and Vukasin. For completeness, Figure 12.6 shows the beta-binomial predictions from Sandra, and Figure 12.7 shows the beta-binomial predictions from Vukasin. Because Vukasin's beta prior assigned a lot of mass to relatively high values of θ_{EJ}, Vukasin predicted that many pancakes would have bacon. This did not happen, however, and therefore Vukasin's predictions were relatively poor.

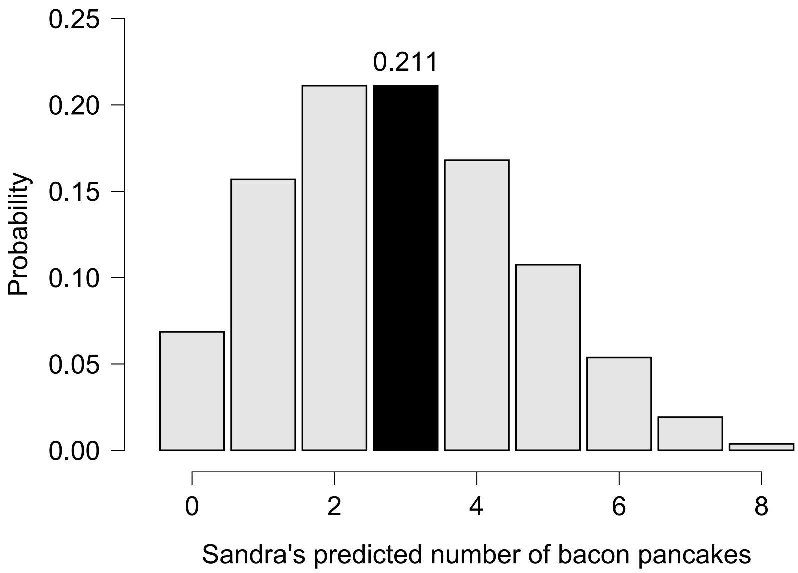


Figure 12.6: Sandra's predicted number of pancakes that come with bacon, out of a total of eight. The beta-binomial predictions are based on Sandra's beta(4,7) prior distribution on θ_{EJ}. The highlighted bar corresponds to the observed data and its height, 0.211, quantifies Sandra's predictive success. Figure from the JASP module *Learn Bayes*.

The results for all four forecasters are summarized in Table 12.5. A comparison between prior and posterior probability shows that Tabea (i.e., $.25 \rightarrow .40$) and Sandra (i.e., $.25 \rightarrow .48$) both gain credibility, whereas Elise (i.e., $.25 \rightarrow .11$) and especially Vukasin (i.e., $.25 \rightarrow .01$) both lose credibility. This is a direct consequence of the fact that Tabea and Sandra predicted the data relatively well, whereas Elise and Vukasin predicted the data relatively poorly.

Despite the fact that Sandra predicted the data best, and therefore has the highest posterior probability, this probability is still a modest $.48$. This means that if an all-or-none decision were made to award Sandra the title 'best bacon forecaster', there is a $1 - .48 = .52$ probability that this decision is wrong.[5] Alternatively, imagine there is a $100 prize for the best bacon forecaster; one may award the entire prize to Sandra, but

[5] Note that a Bayesian posterior probability may be interpreted as the probability of not making an error, if the associated hypothesis were selected as being the best. The error probability is conditional on the observed data and applies to the specific case at hand, in contrast to the error rates in frequentist statistics. For details see the blog post "Error rate schmerror rate" on `BayesianSpectacles.org`.

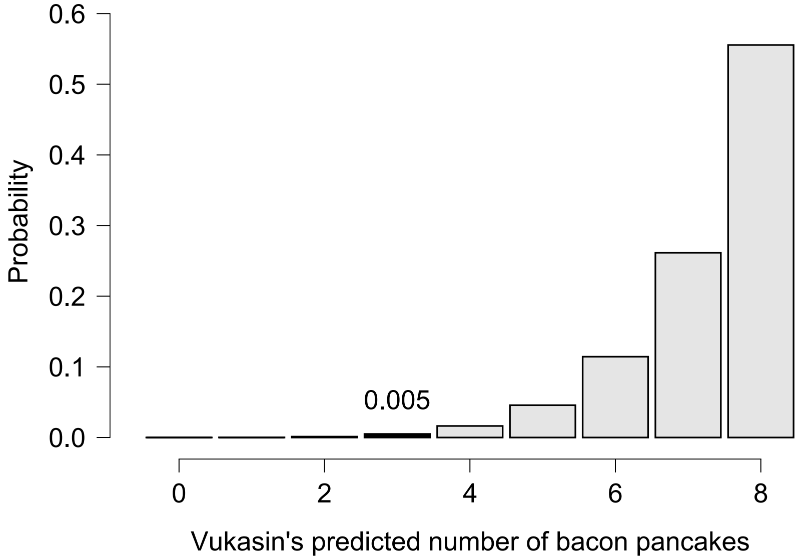


Figure 12.7: Vukasin's predicted number of pancakes that come with bacon, out of a total of eight. The beta-binomial predictions are based on Vukasin's beta(10,1) prior distribution on θ_{EJ}. The highlighted bar corresponds to the observed data and its height, 0.005, quantifies Vukasin's predictive success. Figure from the JASP module *Learn Bayes*.

this decision seems rash (it is more likely to be incorrect than correct). One way to respect the remaining uncertainty is to 'chop' the prize according to the posterior probability. Thus, Tabea would receive $40, Sandra $48, Elise $11, and Vukasin $1. This procedure is similar in spirit to the Problem of Points discussed in Chapter 10.

Of course, the posterior probabilities for the forecasters may also be computed sequentially, one pancake after the other. Table 12.6 shows how the posterior probabilities unfold as the pancakes accumulate.

Table 12.5: Prior probability, predictive success, and resulting posterior probability for bacon forecasters Tabea, Sandra, Elise, and Vukasin. The 'F' denotes 'forecaster', and 'y' denotes the observed data.

Forecaster	Prior $p(F)$	Predictive success $p(y \mid F)$	Posterior $p(F \mid y)$
Tabea	.25	.174	.40
Sandra	.25	.211	.48
Elise	.25	.046	.11
Vukasin	.25	.005	.01

Table 12.6: Sequential analysis of the pancake sequence $\{v, v, v, b, b, v, b, v\}$. Top row: prior model probabilities for each of the four forecasters; bottom row: posterior model probabilities after having observed all eight pancakes.

Pancake	Tabea	Sandra	Elise	Vukasin
0	0.250	0.250	0.250	0.250
1 (v)	0.338	0.431	0.169	0.062
2 (v)	0.350	0.534	0.097	0.019
3 (v)	0.339	0.598	0.056	0.007
4 (b)	0.371	0.513	0.101	0.015
5 (b)	0.386	0.428	0.158	0.028
6 (v)	0.387	0.497	0.103	0.013
7 (b)	0.401	0.424	0.153	0.022
8 (v)	0.399	0.484	0.105	0.012

Bacon Forecasting: Silly?

The example of bacon forecasting is admittedly silly. However, the core Bayesian concepts involved carry over to forecasts that are of great societal importance: election forecasting, economic growth forecasting, climate change forecasting, etc. More generally, *all* Bayesian statistical models may be conceived of as probabilistic forecasting systems (Dawid 1984). This is not immediately obvious when a Bayesian model is specified in a probabilistic programming language such as WinBUGS (Lunn et al. 2012), JAGS (Plummer 2003), or Stan (Carpenter et al. 2017) and is then fit to the data in a single step. But behind the scenes, Bayes' rule governs the knowledge updates with an iron first, and dictates that these updates are driven by relative predictive success: hypotheses and parameters that predict the data well enjoy a boost in credibility, whereas hypotheses and parameters that predict the data poorly suffer a decline (Wagenmakers et al. 2016a).

Figure available at `BayesianSpectacles.org` under a CC-BY license.

Will the Ninth Pancake Have Bacon?

The previous section focused on the relative predictive performance of the rival forecasters. Now suppose we are interested in predicting the identity of the next pancake. For our prediction, it is perhaps tempting to select forecaster Sandra, who predicted the past pancakes best, and forget about her competitors. Sandra has a beta(7,12) posterior distribution for θ_{EJ} after having seen the first eight pancakes, so by the beta prediction rule Sandra assigns probability $7/19 \approx .37$ to the proposition that the ninth pancake will have bacon. However, by basing our predictions solely on Sandra we *throw away information*: we ignore the fact that her rivals Tabea, Elise, and Vukasin also have posterior credibility, and make predictions that differ from that of Sandra.

In order to take into account all uncertainty in our predictions we use the law of total probability and 'model-average' across the four rival forecasters. Figure 12.8 shows a tree diagram with all four forecasters and their predictions for the ninth pancake (cf. Figure 7.4). To obtain the probability that the ninth pancake will have bacon we simply sum the probability of all four branches that result in a bacon pancake. For the data at hand this results in $.40 \cdot 7/16 + .48 \cdot 7/19 + .11 \cdot 12/20 + .01 \cdot 13/19 \approx .42$. Compared to Sandra's prediction of .37, the overall prediction that the ninth pancake will have bacon is slightly higher, as it is driven upwards by the more bacon-enthusiastic predictions from the other forecasters.

In general terms, the *marginal* prediction that the next pancake has bacon is $p(\{b\}) = p(\{b\} \mid \text{Tabea})\, p(\text{Tabea}) + p(\{b\} \mid \text{Sandra})\, p(\text{Sandra}) + p(\{b\} \mid \text{Elise})\, p(\text{Elise}) + p(\{b\} \mid \text{Vukasin})\, p(\text{Vukasin})$.[6] This shows that the overall prediction is a combination of the predictions from each forecaster, weighted by their posterior credibility. The posterior credibility, in turn, is determined by a combination of their prior credibility and their predictive success for the first eight pancakes. This is reminiscent of the 'wisdom of crowds' phenomenon, where the averaged prediction across many forecasters is superior to that of most individual forecasters. In its Bayesian formulation, the averaging across the 'crowd' does not occur blindly; instead, individual forecasts are weighted by expertise, an assessment of which is based on a combination of prior knowledge and previously established predictive success.

[6] For readability, this notation omits to condition on the fact that eight pancakes were already observed. For instance, it is implied that $p(\text{Tabea})$ is not the prior probability for Tabea (i.e., .25), but the posterior probability (i.e., .40).

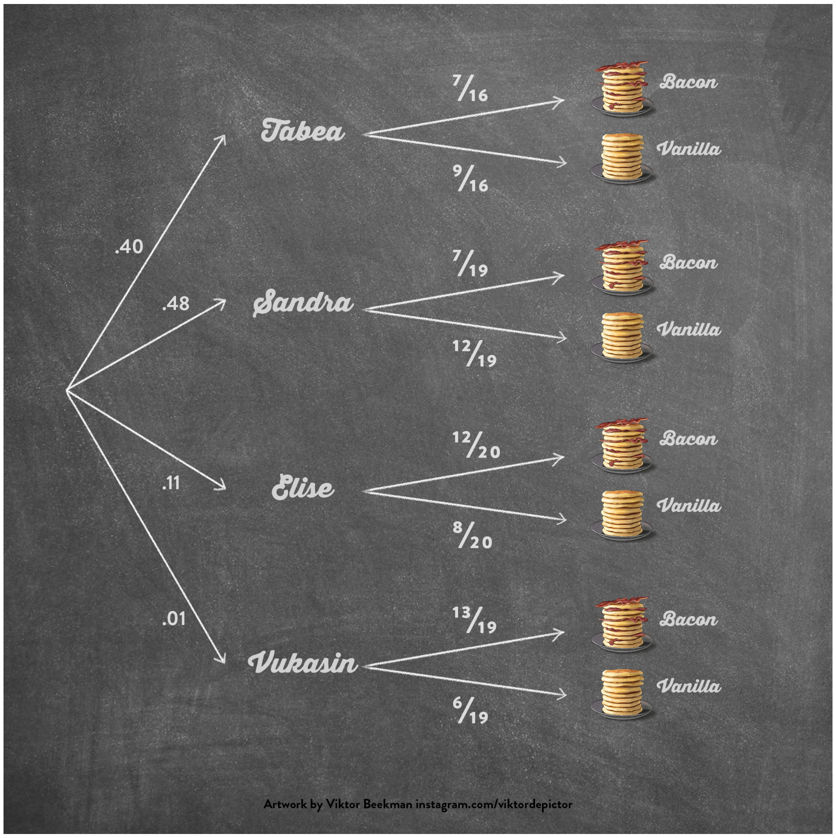


Figure 12.8: To obtain the probability that the ninth pancake has bacon, use the law of total probability and add the probability of the four branches that result in bacon: $.40 \cdot {}^{7}\!/_{16} + .48 \cdot {}^{7}\!/_{19} + .11 \cdot {}^{12}\!/_{20} + .01 \cdot {}^{13}\!/_{19} \approx .42$.

A Trio of Priors

In this chapter we have used the terms 'prior distribution' and 'posterior distribution' in three different ways, and it is important to distinguish between them sharply.

Case I: Bacon Proclivity (i.e., Parameters)

Consider Tabea and forget about the other forecasters for a moment. Tabea's initial uncertainty about EJ's bacon proclivity θ_{EJ} was quantified by a beta(4,4) prior distribution, and the observation of three bacon pancakes and five vanilla pancakes requires that her prior distribution was updated to a beta(7,9) posterior distribution (cf. Figure 12.3). Because of its continuous nature, θ_{EJ} is usually considered a *parameter*.

Case II: Forecaster Quality (i.e., Models and Hypotheses)

Consider our four forecasters and forget about specific values of θ_{EJ} for a moment. The prior credibility of the forecasters is quantified by a uniform prior distribution (i.e., $.25$ for each). This prior distribution is updated by the forecasters' relative predictive success to a posterior distribution (i.e., $.40, .48, .11$, and $.01$ for Tabea, Sandra, Elise, and Vukasin, respectively). Because of its discrete nature, the forecasters are usually considered *rival models or hypotheses*.

Case III: Pancakes (i.e., Data)

Predictions about data can be issued in several ways. We can focus on a specific forecaster such as Tabea and obtain her prior predictive distribution (cf. Figure 12.4). This prior predictive distribution depends on the desired number of hypothetical observations and on the prior distribution for bacon proclivity θ_{EJ}: together with the intended sample size, the prior beta distribution gives rise to a prior predictive beta-binomial distribution. Depending on the specifics of the data-generating process, the prior predictive distribution can be discrete (as it is here) or continuous.[7] In the same way, predictions about future data can be made from the posterior distribution, giving rise to a posterior predictive distribution.

[7] Continuous prior predictive distributions will feature in later chapters.

Predictions can also be made across all forecasters, as demonstrated above in Figure 12.8. Predictions that average over one or more nuisance factors are called 'marginal'[8] For example, Figure 12.9 shows a 'marginal posterior predictive distribution': it is *marginal* because it does not refer to any specific forecaster – this is a nuisance factor that has been averaged out according to the law of total probability; it is *posterior* because it is based on the posterior distributions for θ_{EJ} from the four forecasters, taking into account the knowledge gained from the observed eight pancakes; finally, it is *predictive* because it concerns the predicted number of bacon pancakes out of a total of 20 new, unobserved pancakes.

[8] The terminology comes from 2×2 contingency tables, where the column and row sums are known as the 'table margins'.

Thus, there is uncertainty at different levels. We do not know who has the most knowledge about EJ's bacon proclivity, and this induces epistemic uncertainty on the level of forecasters. In turn, each forecaster is uncertain about the value of the bacon proclivity θ_{EJ}, and this is reflected in a forecaster-specific beta prior distribution for θ_{EJ}. This epistemic uncertainty propagates to predictions, where it is augmented with aleatory uncertainty (cf. Chapter 2). Depending on what we are interested in, we may zoom in on a particular factor and use the law of total probability to average out the nuisance factors. Even though there are various levels of uncertainty, the Bayesian principles stays the same: parameters and hypotheses that predict the data relatively well

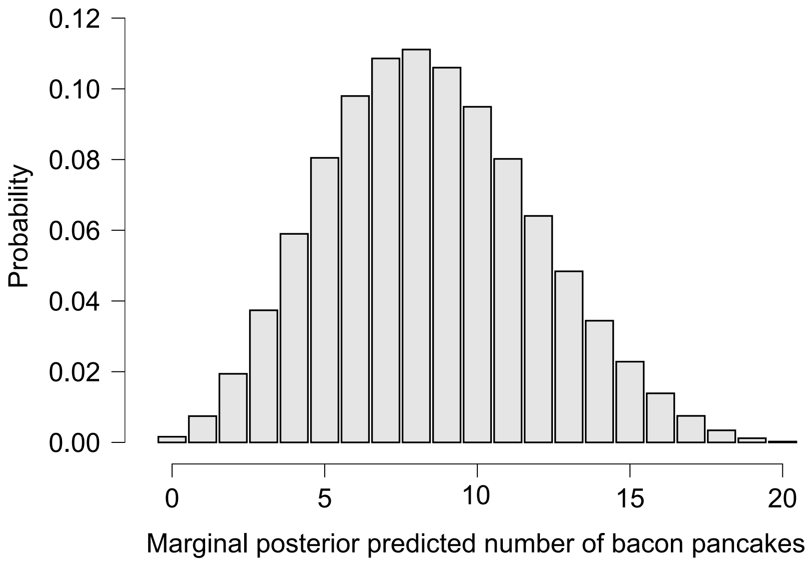


Figure 12.9: Posterior predictive distribution for the number of pancakes that come with bacon, out of a requested total of 20 unobserved pancakes. Predictions are based on the forecasters' posterior distributions for θ_{EJ} and weighted by each forecaster's posterior probability. Figure from the JASP module *Learn Bayes*.

experience a gain in credibility, whereas parameters and hypotheses that predict the data relatively poorly suffer a decline.

Prior Distributions as Bets

When a forecaster assigns the binomial chance θ a relatively narrow prior distribution, this induces a relatively precise prediction for to-be-observed data (i.e., a relatively narrow prior predictive distribution). When the incoming data are consistent with this precise prediction, this empirical validation will generally enhance the forecaster's credibility. However, when the incoming data are inconsistent with the precise prediction, this often greatly undermines the forecaster's credibility.[9]

"There are practical difficulties in assessing the prior probability in many cases as they actually arise. This is not a situation to evade, but one to face." (Jeffreys 1931, p. 34)

An informed prior distribution can therefore be conceived of as an indirect *bet*, a way to distribute prior resources across a range of possible data-generating processes θ with the goal to maximize expected reward (i.e., maximize the predictive score).[10] Conservative forecasters hedge their bets and assign θ a vague prior distribution that gives rise to a broad prior predictive distribution. Aggressive forecasters, on the other hand, use prior knowledge to specify a narrow prior distribution on θ that gives rise to a narrow prior predictive distribution. The aggressive forecaster will outpredict the conservative forecaster whenever the data validate the riskier prediction. This occurs because the aggressive forecaster did not have to waste prior resources by 'betting' on values

[9] These regularities are not universally true, as the reallocation of credibility for any particular forecaster depends on the predictive performance of the rival forecasters.

[10] The bet is *indirect* because the payout is determined by the predictive mass that is assigned to the observed data; in other words, the *direct* bet is in the space of possible data, not in the space of parameters.

of θ with a low probability of generating the observed data. This theme will become increasingly prominent in the next chapters.

Exercises

1. Consider the list of all 34 priors shown in Appendix A. Select an interesting subset and then (1) compute the posterior probabilities for all forecasters in your subset; (2) obtain the associated marginal posterior predictive distribution for 20 new pancakes. How does it compare to Figure 12.9?

2. The text states, "However, the core Bayesian concepts involved carry over to forecasts that are of great societal importance: election forecasting, economic growth forecasting, climate change forecasting, etc." Mention some of these core Bayesian concepts.

3. Consider Equation 12.1. How would you interpret p(Tabea) and p(Elise)? Would this interpretation be helpful for statistical models in general?

4. The text mentions that the fictitious $100 prize for 'best bacon forecaster' can be divided according to the posterior probability. "Thus, Tabea receives $40, Sandra $48, Elise $11, and Vukasin $1. This procedure is similar in spirit to the Problem of Points discussed in Chapter 10." Nevertheless, there is a difference – what is it?

5. From Figure 12.8 it follows that the probability is $.42$ that the ninth pancake will have bacon. Confirm this result with the *Learn Bayes* module.

6. The text states "The aggressive forecaster will outpredict the conservative forecaster whenever the data validate the riskier prediction." Convince yourself that this is true by constructing a concrete example in the *Learn Bayes* module in JASP.

7. In 2022, EJ produced a sequence of five vanilla pancakes: $y = \{v, v, v, v, v\}$. Four Research Master students assigned different prior beta distributions to θ_{EJ}: Lisa specified a beta$(70, 30)$ prior, Seymour a beta$(1, 1)$ prior, Moe a beta$(2, 8)$ prior, and Krusty a beta$(4, 20)$ prior. Assuming the four students are deemed equally good at pancake forecasting *a priori*, compute the resulting posterior probability for each forecaster. Then compute the probability that the sixth pancake is a bacon pancake.

Chapter Summary

This chapter provided a perspective on Bayesian inference as probabilistic sequential forecasting. When data accumulate, prediction errors drive a continual adjustment of beliefs, as was illustrated with the case of eight pancakes with or without bacon. The predict-update cycle of learning holds on all levels – it holds within each forecaster individually (i.e., prior distributions for pancake proclivity θ are updated to posterior distributions for θ in a pancake-by-pancake fashion; see Figure 12.3 and Table 12.2) but also across rival forecasters (i.e., prior probabilities concerning relative forecasting ability are updated to posterior probabilities in a pancake-by-pancake fashion; see Table 12.6). Predictions concerning new pancakes ought to take into account both the uncertainty about pancake proclivity within a specific forecaster, and uncertainty about the relative predictive prowess of the rival forecasters.

Want to Know More?

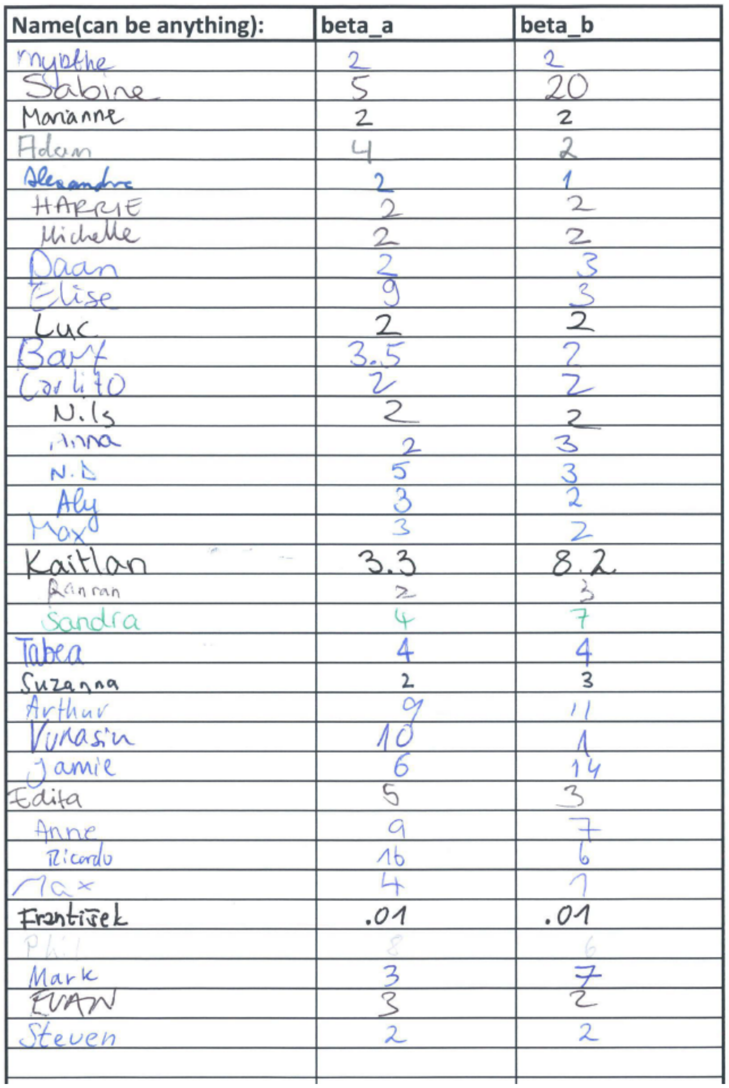
✓ An informative post by Fabian Dablander: `https://fabiandablander.com/r/Bayes-Potter.html`.

✓ Dawid, A. P. (1984). Present position and potential developments: Some personal views: Statistical theory: The prequential approach (with discussion). *Journal of the Royal Statistical Society Series A, 147*, 278-292. This classic paper is inspired by the work of both Bruno de Finetti and Harold Jeffreys. "The prequential approach is founded on the premiss that the purpose of statistical inference is to make sequential probability forecasts for future observations, rather than to express information about parameters."

✓ Hinne, M., Gronau, Q. F., van den Bergh, D., & Wagenmakers, E.–J. (2020). A conceptual introduction to Bayesian model averaging. *Advances in Methods and Practices in Psychological Science, 3*, 200-215. Worth looking up if only for the drawing of the pandemonium.

✓ Veen, D., Stoel, D., Schalken, N., Mulder, K., & van de Schoot, R. (2018). Using the data agreement criterion to rank experts' beliefs. *Entropy, 20*, 592. "By letting experts specify their knowledge in the form of a probability distribution, we can assess how accurately they can predict new data, and how appropriate their level of (un)certainty is."

Appendix A: Prior Distributions From the 2019 Class



Name(can be anything):	beta_a	beta_b
myrthe	2	2
Sabine	5	20
Marianne	2	2
Adam	4	2
Alexandre	2	1
HARRIE	2	2
Michelle	2	2
Daan	2	3
Elise	9	3
Luc	2	2
Bart	3.5	2
Carlito	2	2
Nils	2	2
Anna	2	3
N.D	5	3
Aly	3	2
Max	3	2
Kaitlan	3.3	8.2
Ranran	2	3
Sandra	4	7
Tabea	4	4
Suzanna	2	3
Arthur	9	11
Vukasin	10	1
Jamie	6	14
Edita	5	3
Anne	9	7
Ricardo	16	6
Max	4	7
František	.01	.01
Phil	8	6
Mark	3	7
EVAN	3	2
Steven	2	2

Figure 12.10: The list of 34 beta prior distributions for EJ's bacon proclivity θ_{EJ}. Low values for beta parameters α and β indicate large uncertainty (i.e., a wide prior). Students were informed that their prior choices could be used for this book; they were free to use pseudonyms.

Appendix B: Mixture Distributions

The section 'Will the Ninth Pancake Have Bacon?' illustrated how the predictions of the four forecasters (i.e., Tabea, Sandra, Elise, and Vukasin) may be combined to yield a single overall prediction for the upcoming pancake – a weighted average of the individual predictions, with the averaging weights informed by the forecasters performance on pancakes from the past (cf. Figure 12.8). The interest was on the prediction, and the identity of the forecaster is a nuisance factor that was averaged out using the law of total probability.

Essentially the same process can be used when interest centers on the prior and posterior distribution for bacon proclivity θ_{EJ}, with the forecasters averaged out. The results are easily obtained in the *Binomial Testing* routine of the *Learn Bayes* module in JASP. For simplicity we will take into consideration only the four forecasters Tabea, Sandra, Elise, and Vukasin. Figure 12.11 shows a screenshot of the input GUI, with the data specified in the top panel (i.e., three bacon pancakes and five vanilla pancakes) and the four forecasters specified in the bottom panel, both in terms of their prior probabilities (in this case, $1/4$) and in terms of the beta prior distributions they assign to θ_{EJ}.

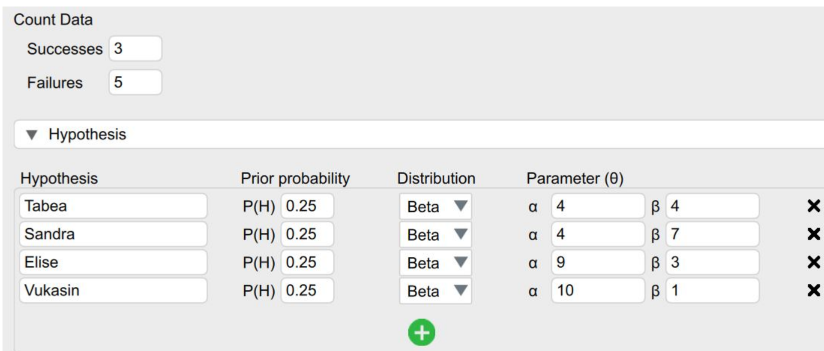


Figure 12.11: JASP screenshot of two input panels from the *Binomial Testing* routine of the *Learn Bayes* module. The input panels control the inference across four pancake forecasters. Top panel: specification of the data; bottom panel: specification of the four forecasters. See text for details.

The resulting 'marginal' prior distribution for θ_{EJ} is a four-component mixture of beta distributions, with the prior probabilities for the individual forecasters acting as mixture weights. This mixture distribution represents the knowledge of the four forecasters combined. Figure 12.12 displays the mixture prior distribution; the multimodal shape[11] is a clear indication of the underlying mixture.

[11] A *multimodal* distribution has more than one maximum or 'bump'.

This mixture prior distribution is then updated by means of the data to yield a mixture *posterior* distribution. The mixture weights for the components in the posterior distribution are the posterior probabilities for the individual forecasters; just as for the prediction of the ninth pancake, the shape of the posterior for θ_{EJ} is determined mostly by those forecasters that proved to be most reliable in the past. The mixture posterior is shown in Figure 12.13.

It is noteworthy that –in contrast to the prior distribution– the posterior distribution shows little outward sign of actually being based on a mixture; it is unimodal and (somewhat) bell-shaped. In general, *all*

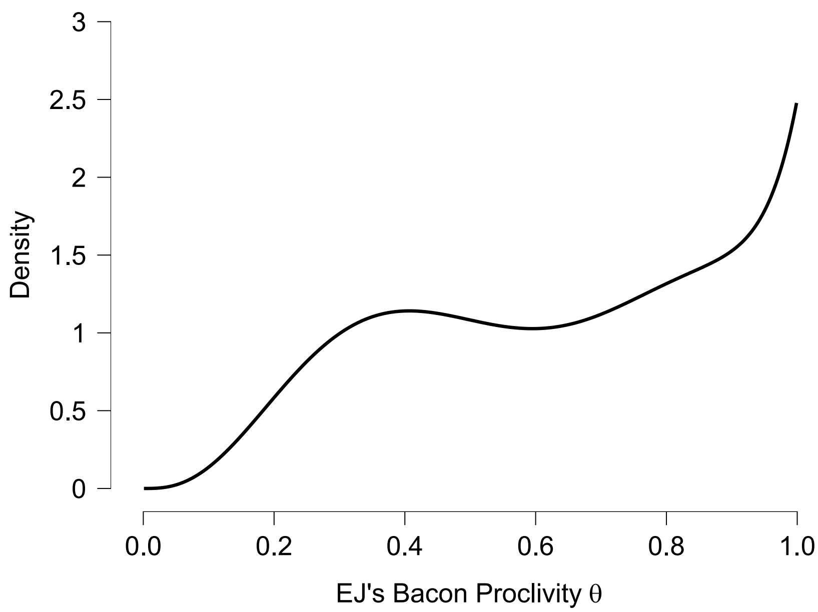


Figure 12.12: Marginal prior distribution for EJ's bacon proclivity θ_{EJ} across the four forecasters as specified in Figure 12.11. Figure from the JASP module *Learn Bayes*.

posterior distributions will become bell-shaped (and symmetric around the maximum likelihood estimator) as sample size increases – this is known as the *Bayesian central limit theorem* or the *Bernstein-von Mises theorem* (e.g., van der Vaart 1998).[12] The theorem holds under 'regularity conditions' and these imply that the true parameter is not located on the boundary of the space. For instance, if the data are generated from $\theta = 1$ or $\theta = 0$ then the posterior will obviously *not* be bell-shaped.[13]

[12] This observation dates back to Laplace.

[13] The doubtful reader may convince themselves by using JASP to analyze a large data set comprised of only successes or only failures.

Even though the posterior distribution shown in Figure 12.13 looks much more bell-shaped than the prior distribution, it is still noticeably asymmetric: the lingering impact of the prior is reflected in a right-skew, which expresses a preference for relatively high values of θ_{EJ}. Foreshadowing the material from the next chapter, we will now pretend that ten times more pancakes were observed, for a total of 30 bacon pancakes and 50 vanilla pancakes. The resulting posterior distribution is shown in Figure 12.14. The additional observations have caused the posterior distribution to narrow and to become more symmetric around the maximum likelihood estimate (i.e., the sample proportion).

For a summary of the ways in which the opinion of different forecasters (or experts) may be combined we refer the interested reader to Wilson and Farrow (2018) and Stefan et al. (2022). The idea of a mixture prior distribution will resurface in Chapter 30.

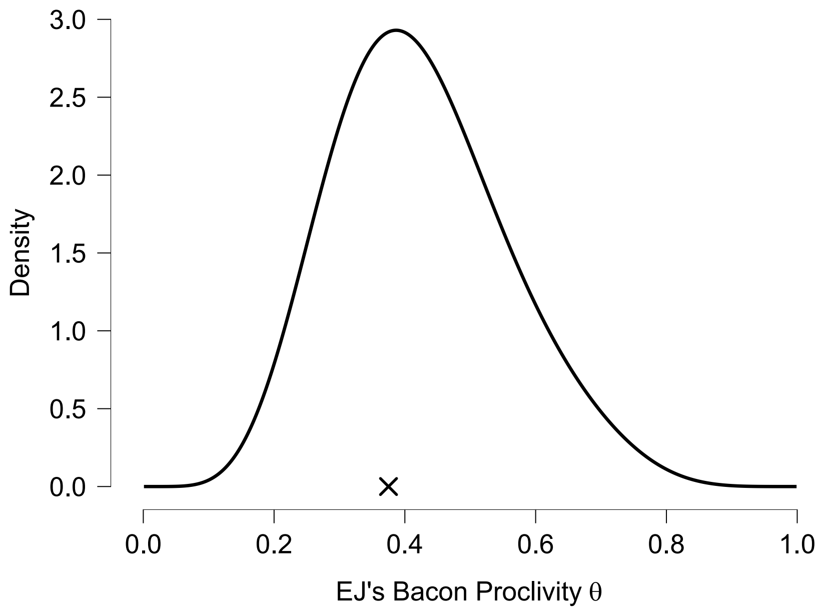


Figure 12.13: Marginal posterior distribution for EJ's bacon proclivity θ_{EJ} across the four forecasters as specified in Figure 12.11. The cross denotes the sample proportion of $3/8 = .375$. Figure from the JASP module *Learn Bayes*.

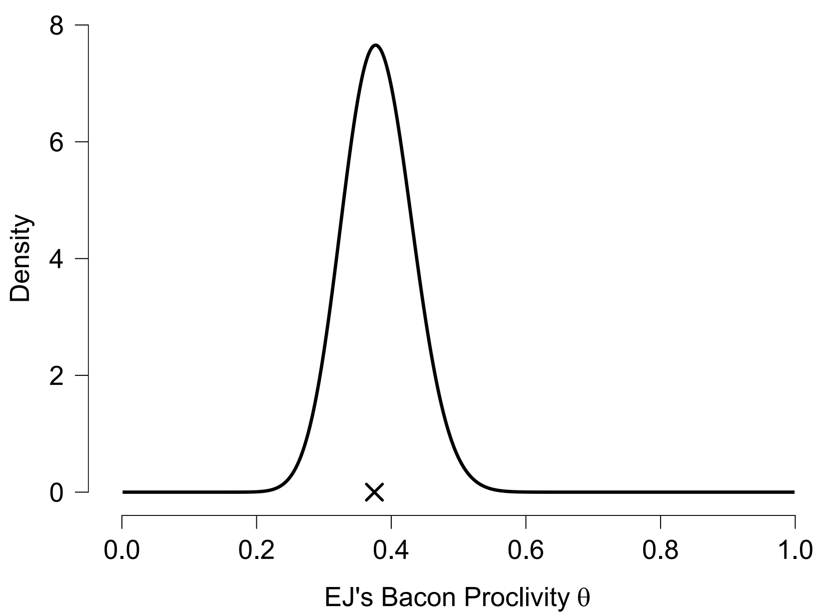


Figure 12.14: Marginal posterior distribution for EJ's bacon proclivity θ_{EJ} across the four forecasters as specified in Figure 12.11, with the exception that the number of bacon and vanilla pancakes has increased tenfold (i.e., to 30 and 50, respectively). The cross denotes the sample proportion of $30/80 = .375$. Figure from the JASP module *Learn Bayes*.

13 A Plethora of Pancakes [with Charlotte Tanis and Alexander Ly]

An accurate statement of the prior probability is not necessary in a pure problem of estimation when the number of observations is large.

Jeffreys, 1939

Chapter Goal

We continue the example from the previous chapter and add more pancake observations. Three facts are demonstrated: (1) As the pancakes accumulate, the posterior distributions become increasingly peaked around the value of θ that predicts the data best, which equals the sample proportion: 'the data overwhelm the prior' (e.g., Wrinch and Jeffreys 1919); (2) A forecaster's overall predictive performance can be obtained by multiplying their performance for separate batches, but only when the beta distributions are updated appropriately after each batch (e.g., Jeffreys 1961, pp. 332-334); (3) As the pancakes accumulate, the difference in predictive performance between the rival forecasters is bounded – even an infinite number of pancakes does not suffice to identify the best bacon forecaster with certainty.

The Data Overwhelm the Prior

The analysis from the previous chapter involved forecasters Tabea, Sandra, Elise, and Vukasin, who each expressed their prior uncertainty about EJ's bacon proclivity θ_{EJ} by their own beta distribution. The observed data consisted of three bacon pancakes and five vanilla pancakes.

We decide to collect more information, and force EJ to bake another few hundred pancakes. For educational purposes, we fix the sample ratio of bacon to vanilla pancakes at 3:5; our extended (fictional) data set now has 300 bacon pancakes and 500 vanilla pancakes. Figure 13.1 and

Table 13.1 show the prior and posterior beta distributions for each of the four forecasters.

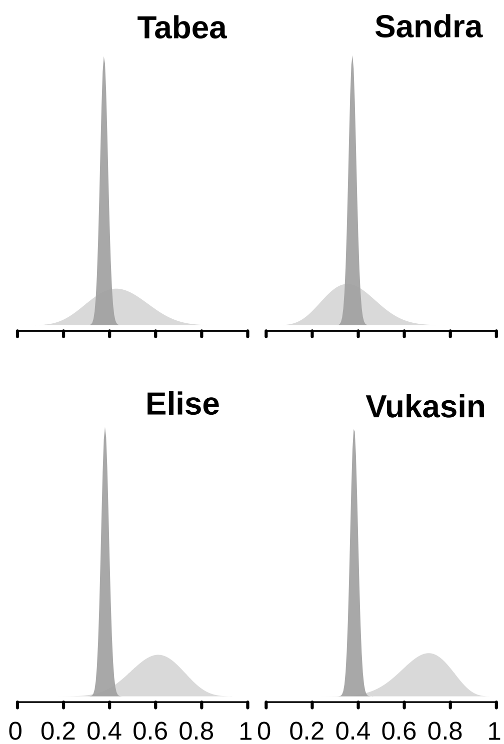


Figure 13.1: Prior and posterior beta distributions for EJ's pancake proclivity θ_{EJ}, for four forecasters. The 'prior' distributions, shown in light gray, have already been updated to include the information from the previous chapter (i.e., the fact that EJ baked three bacon pancakes and five vanilla pancakes). The posterior distributions, shown in dark gray, are based on a fictitious new pancake stack consisting of 297 bacon pancakes and 495 vanilla pancakes. The sample proportion of bacon pancakes is $3/8 = .375$.

In Figure 13.1, the light-gray distributions represent the priors that were obtained by updating the forecasters' initial beliefs with the information from the earlier eight pancakes. In other words, the light-gray distributions represent each forecaster's belief after having seen the results from the eight pancakes discussed in the previous chapter. In general, these prior distributions are relatively wide, indicating considerable uncertainty on the part of the forecasters. Also, the prior distributions are markedly different across the forecasters: Tabea and Sandra assign most prior belief to low and middle values of bacon proclivity θ_{EJ}, whereas Elise and Vukasin assign more belief to higher values of θ_{EJ}.

The dark-gray distributions in Figure 13.1 represent the posteriors obtained from updating each forecaster's initial belief with the information from 800 pancakes, 300 of which have bacon and 500 of which are vanilla. The posterior distributions are relatively peaked, indicat-

Table 13.1: Prior and posterior beta distributions for EJ's pancake proclivity θ_{EJ}, for four forecasters. The 'prior' distributions have already been updated to include the information from the previous chapter (i.e., the fact that EJ baked three bacon pancakes and five vanilla pancakes). The posterior distributions are based on a fictitious new pancake stack consisting of 297 bacon pancakes and 495 vanilla pancakes.

	Beta prior		Beta posterior	
Forecaster	α	β	α	β
Tabea	7	9	304	504
Sandra	7	12	304	507
Elise	12	8	309	503
Vukasin	13	6	310	501

ing a high level of certainty about θ_{EJ}. In addition, the four posterior distributions are relatively similar to one another. That this should be the case is apparent from Table 13.1: the α and β parameters that define the beta posteriors are dominated by the fact that hundreds of pancakes have been observed, and prior differences between forecasters are drowned out by the impact of the data. In other words, Tabea's beta$(7, 9)$ prior distribution may be noticeably different from Vukasin's beta$(13, 6)$ prior distribution, but Tabea's beta$(304, 504)$ posterior distribution is virtually identical to Vukasin's beta$(310, 501)$ posterior distribution.

Intuitively, the posterior distribution is a compromise between the forecasters' prior convictions and the information coming from the data, as described in Chapter 7 (Jeffreys 1939, p. 46):

$$\text{Posterior} \;\propto\; \text{Prior} \times \text{Likelihood}.$$

Each forecaster may have prior beliefs that are unique, but the data are common property. With every observation that comes in, the 'posterior compromise' will be influenced more by the data and less by the prior. Eventually, the deluge of data will cause the posterior to concentrate near the θ_{EJ} value that corresponds to the proportion of bacon pancakes in the sample, $^{300}/_{800} = .375$. This can also be explained from a predictive perspective. Recall that every time an observation arrives, the prior distribution is updated such that values for θ_{EJ} that predict that observation relatively well receive a boost in plausibility, whereas values for θ_{EJ} that predict that observation relatively poorly suffer a decline. Now consider a value such as $\theta_{EJ} = {}^{1}/_{2}$. This value assigns considerable mass to the outcome of three bacon pancakes and five vanilla pancakes; such data are not surprising under $\theta_{EJ} = {}^{1}/_{2}$, and hence it retains a reasonable degree of credibility. Specifically, the predictive probability of three bacon pancakes and five vanilla pancakes is .22 under $\theta_{EJ} = {}^{1}/_{2}$ and .28 under $\theta_{EJ} = {}^{3}/_{8}$ – a minute predictive advantage of $^{.28}/_{.22} = 1.3$

for the value that was cherry-picked to provide the best predictive performance.[1] However, the situation changes dramatically when we consider the larger data set. Under the best predicting value, $\theta_{EJ} = 3/8$, the probability of observing 300 bacon pancakes and 500 vanilla pancakes is .03; under $\theta_{EJ} = 1/2$, the predictive probability is a shockingly low .00000000000031; that is, $\theta_{EJ} = 3/8$ outpredicted $\theta_{EJ} = 1/2$ by a factor of $.03/.00000000000031 = 96,774,193,548$. Thus, $\theta_{EJ} = 1/2$ does an abysmal job in predicting 300 bacon pancakes and 500 vanilla pancakes; such data would be highly surprising under $\theta_{EJ} = 1/2$, and compared to values of θ_{EJ} close to $300/800$, $\theta_{EJ} = 1/2$ loses almost all credibility.

[1] In frequentist statistics, this is known as the *maximum likelihood estimate* (MLE), the value of θ that predicts the data best (i.e., it assigns the largest probability to the observed data).

The continual impact of the data therefore pushes forecasters with clearly different prior beliefs towards an almost identical posterior belief, centered on the sample proportion (i.e., the MLE). This *posterior convergence* is emphasized in almost every Bayesian textbook, and the associated adage is 'the data overwhelm the prior'. This idea goes back at least to Wrinch and Jeffreys (1919), who concluded: "Thus, unless the distribution of prior probability (...) is very remarkable, its precise form does not produce much effect on the probability that the true value lies within a certain range determined wholly by the constitution of the sample itself." (p. 728).[2] In later work, Jeffreys argued that it was this Bayesian regularity that provided a firm foundation for maximum likelihood estimation, ironically the main method advocated by the thoroughly anti-Bayesian Sir Ronald Fisher:

[2] As summarized by Jeffreys (1933b, p. 84), "When the sample is large the variation of $f(r)$ [the prior distribution] produces no important disturbance of the theory, as has already been pointed out, since it is *overwhelmed* by the variation of $h(r)$ [the likelihood], but for small samples the difference is considerable." (italics ours) Also, Jeffreys (1955, p. 280) concluded: "Wrinch and I showed in 1919 that in the estimation of a chance, where the possible values form a continuous set the precise form of the prior probability distribution taken for it has very little effect on the posterior probability, and consequently quite crude forms are quite good enough. This can be extended to most estimation problems."

> "The whole reason for attaching any importance to Fisher's "likelihood" is that it is proportional to the posterior probability given by Laplace's theory, and it has no meaning outside the original sample except in terms of this theory." (Jeffreys 1933b, p. 87)

and

> "Professor Fisher seems to set up his use of likelihood in opposition to the theory of probability. I cannot see why he does this, since the theory of probability provides the use of likelihood with its best justification." (Jeffreys 1935b, p. 70)

and

> "Again, provided the number of observations is large and the prior probability is not very unevenly distributed with the parameters to be found, the posterior probability in any range where it is appreciable is distributed nearly in proportion to the likelihood. This was proved for sampling by Wrinch and me in 1919, but the argument is obviously capable of wide extension. Thus subject to one condition Fisher's principle of maximum likelihood is an immediate consequence of my theory."(Jeffreys 1937b, p. 258)

and

> "Subject to a negligible correction, therefore, the posterior probability density (...) is proportional to the likelihood (...)
>
> This result was given for sampling by Wrinch and me in 1919; we did not extend it in the above way, thinking that the extension would be obvious and that the method of maximum likelihood was already in general use, though Fisher did not introduce the name till 1921 ; and indeed it was in use for the problems of sampling and estimates for normal distributions that interested us at the time." (Jeffreys 1938c, p. 147)

and

> "The method of maximum likelihood has been vigorously advocated by Fisher; the above argument [i.e., the data overwhelm the prior] shows that in the great bulk of cases its results are indistinguishable from those given by the principle of inverse probability [i.e., Bayesian inference], which supplies a justification of it. An accurate statement of the prior probability is not necessary in a pure problem of estimation when the number of observations is large. What the result amounts to is that unless we previously know so much about the parameters that the observations can tell us little more, we may as well use the prior probability distribution that expresses ignorance of their values (...)" (Jeffreys 1961, p. 194)

and

> "In the same paper [Wrinch & Jeffreys, 1919] we (...) showed that if n [sample size] is large the posterior probabilities are nearly in the ratios of the direct probabilities (...). This was in fact the method of maximum likelihood, first given that name by Fisher a few years later. We did not think it at all remarkable at the time, thinking that all statisticians used it already." (Jeffreys 1974, p. 1)

and finally, for good measure:

> "It is shown that in a wide class of problems where there are many observations the posterior probability depends almost entirely on the observations and very little on the prior probability. This justifies the method of maximum likelihood, given that name later by R. A. Fisher." (Jeffreys and Swirles 1977, p. 251)

"The likelihood takes us a long way, but the theory of probability finishes the job." (Jeffreys 1935b, p. 71)

Pancakes Galore

Not satisfied with a mere 800 pancakes, you up the ante and force EJ to increase the stack to a total of 8000 pancakes. We retain the 3:5 bacon to vanilla ratio, which means that our stack now consists of 3,000 bacon pancaked and 5,000 vanilla pancakes. Figure 13.2 and Table 13.2 show the prior and posterior beta distributions for each of the four forecasters.

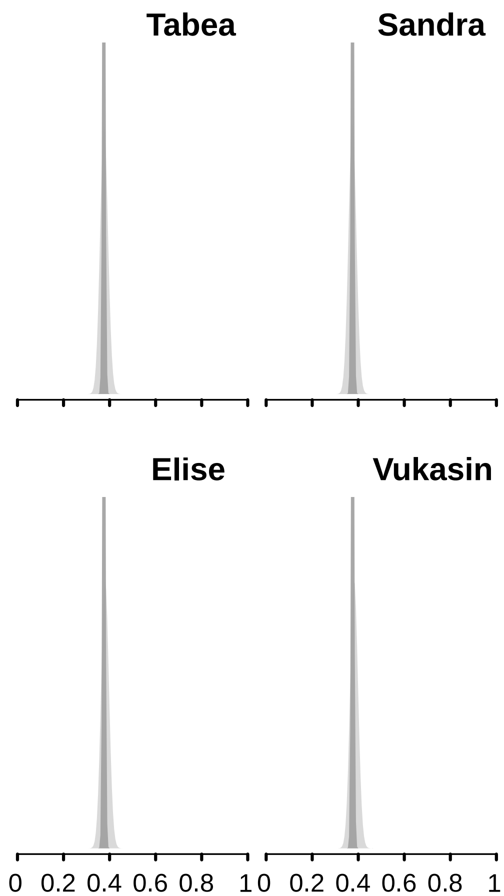


Figure 13.2: Prior and posterior beta distributions for EJ's pancake proclivity θ_{EJ}, for four forecasters. The 'prior' distributions, shown in light gray, have already been updated to include the information from the previous stack (i.e., the fact that EJ baked 300 bacon pancakes and 500 vanilla pancakes). The posterior distributions, shown in dark gray, are based on a fictitious new pancake stack consisting of 2700 bacon pancakes and 4500 vanilla pancakes. The sample proportion of bacon pancakes is $3/8 = .375$. The posterior distributions are so peaked that they do not fit on the graph.

As expected, the effect of the additional pancakes is to increase the forecasters' certainty about θ_{EJ} still further. The dark gray posterior distributions are now so narrow that their peaks do not fit on the graph, like the top of a mountain hidden from view above the clouds. One key difference with respect to the first update in this chapter (shown in Figure 13.1) is that this time, the light gray 'prior' distributions are highly similar between the forecasters. After a few hundred pancakes had been observed, the forecasters had already converged to the same opinion. This may prompt the speculation that the new set of pancakes does little to discriminate the good forecasters from the poor forecasters, even though this set is thousands of pancakes in size. We elaborate on this speculation in the next sections.

Table 13.2: Prior and posterior beta distributions for EJ's pancake proclivity θ_{EJ}, for four forecasters. The 'prior' distributions have already been updated to include the information from the previous stack (i.e., the fact that EJ baked 300 bacon pancakes and 500 vanilla pancakes). The posterior distributions are based on a fictitious new pancake stack consisting of 2700 bacon pancakes and 4500 vanilla pancakes.

	Beta prior		Beta posterior	
Forecaster	α	β	α	β
Tabea	304	504	3004	5004
Sandra	304	507	3004	5007
Elise	309	503	3009	5003
Vukasin	310	501	3010	5001

Combining the Evidence

At this point we have collected a stack of 8000 pancakes, and we wish to compare the predictive performance of Tabea (who assigned θ_{EJ} a beta(4,4) prior) against that of Elise (who assigned θ_{EJ} a beta(9,3) prior). Recall that the evidence, that is, the data-induced change from prior to posterior odds, is generally known as the *Bayes factor*, which we abbreviate as 'BF'. The *Learn Bayes* module informs us that $\text{BF}_{te} \approx 23.73$, that is, Tabea predicted the composition of the 8000 pancakes almost 24 times better than Elise.[3]

However, the complete stack arrived in three separate batches. The first batch consisted of three bacon and five vanilla pancakes; the second batch consisted of 297 bacon and 495 vanilla pancakes (for a running total of 800 pancakes); and the third batch consisted of 2700 bacon pancakes and 4500 vanilla pancakes, bringing the total up to 8000. Let's assume that we wish to combine the evidence across the three batches – how should this be accomplished?

A tempting, but incorrect procedure to combine the evidence works as follows. For the first batch, we compare predictive performance of the Tabea beta(4,4) prior versus the Elise beta(9,3) prior and find that $\text{BF}_{te}^{\text{batch1}} \approx 3.80$. For the second batch, we also compare predictive performance of the Tabea beta(4,4) prior versus the Elise beta(9,3) prior and find that $\text{BF}_{te}^{\text{batch2}} \approx 22.76$. For the third batch, we again compare predictive performance of the Tabea beta(4,4) prior versus the Elise beta(9,3) prior and find that $\text{BF}_{te}^{\text{batch3}} \approx 23.71$. To obtain the overall evidence across all three batches, we then multiply the batch-specific Bayes factors and obtain $3.80 \times 22.76 \times 23.71 \approx 2051$. This is clearly wrong – from analysing all 8000 pancakes simultaneously we already know that the correct answer is approximately 23.73.

What went wrong here is that the priors were used three times, once of each batch. For the first batch, this was correct; so it is true that

[3] The subscript 'te' conveys that 'Tabea' is the forecaster in the numerator and 'Elise' is the forecaster in de denominator of the Bayes factor; hence, BF_{te} stands for $p(\text{data} \mid \text{Tabea})/p(\text{data} \mid \text{Elise})$.

$\text{BF}_{te}^{\text{batch1}} \approx 3.80$. For the analysis of the second batch, however, the initial prior beta distributions are no longer relevant. Instead, the relevant prior distributions are now a beta(7,9) for Tabea and a beta(12,8) for Elise. Comparing predictive performance of these updated priors on the data from the second batch yields $\text{BF}_{te}^{\text{batch2}} \approx 6.0$. The same updating principle applies to the third batch. We now compare predictive performance of Tabea's updated beta(304,504) distribution versus Elise's updated beta(309,503) distribution for the data from the third batch, which yields $\text{BF}_{te}^{\text{batch3}} \approx 1.04$. Notice that, in contrast to the incorrect computation, the successive Bayes factors become increasingly smaller, reflecting the forecasters' converging opinion. After the data from the second batch have been accounted for, Tabea and Elise make highly similar predictions, such that additional data are hardly diagnostic. Multiplying the three updated Bayes factors we find that $3.80 \times 6.0 \times 1.04 \approx 23.71$, which recovers the result from analyzing all 8000 pancakes at once.[4] Therefore, in the words of Harold Jeffreys:

[4] The slight remaining numerical difference is due to rounding.

> "We cannot therefore combine tests by simply multiplying the values of K [the Bayes factor]. This would assume that the posterior probabilities are chances, and they are not. The prior probability when each subsample is considered is not the original prior probability, but the posterior probability left by the previous one. We could proceed by using the subsamples in order in this way, but we already know (...) what the answer must be. The result of successive applications of the principle of inverse probability [Bayesian inference] is the same as that of applying it to the whole of the data together, using the original prior probability (...) Thus if the principle is applied correctly, the probabilities being revised at each stage in accordance with the information already available, the result will be the same as if we applied it directly to the complete sample (...)" (Jeffreys 1961, p. 334; see also Jeffreys 1938a, pp. 190-192)

In order to drive the point home, consider a scenario involving the following two hypotheses: $\mathcal{H}_x$ holds that a stack of ten pancakes is baked either by the vegetarian Charly (i.e., $\theta_C = 0$) or by the carnivore Sidney (i.e., $\theta_S = 1$), with both candidates equally likely *a priori* to be the baker. The competing hypothesis, $\mathcal{H}_y$, holds that the pancakes are baked by Jackie, whose pancake proclivity is $\theta_J = 1/2$. The first pancake in the stack is observed, and it has bacon. The probability of this datum is $1/2$ under both hypotheses, and consequently $\text{BF}_{xy} = 1$: the datum is completely uninformative with respect to the relative predictive performance of the rival hypotheses. Now assume that we examine the entire stack and observe that *all* ten pancakes have bacon. If we multiply evidence without updating, and apply the same prior ten consecutive times, once for each pancake, then $\text{BF}_{xy} = 1$ for every pancake, and the overall result would be $1 \times 1 \times 1 \times 1 \times 1 \times 1 \times 1 \times 1 \times 1 \times 1 = 1$. Clearly something is amiss, because a stack of ten bacon pancakes should provide evidence in support of $\mathcal{H}_x$.

The correct analysis proceeds as follows. After the first pancake, which yields $\mathrm{BF}_{xy} = 1$, the hypothesis $\mathcal{H}_x$ is updated: we now know that Charly cannot be the baker, so all posterior probability is now on Sidney being the baker. For the second pancake, therefore, we compare $\mathcal{H}_x$: Sidney is the baker (i.e., $\theta_S = 1$) versus $\mathcal{H}_y$: Jackie is the baker (i.e., $\theta_J = 1/2$). A bacon pancake is twice as likely to be produced by Sidney than by Jackie, and hence, after two pancakes, $\mathrm{BF}_{xy} = 1 \times 2 = 2$. Each consecutive pancake is twice as likely under $\mathcal{H}_x$ than under $\mathcal{H}_y$, and the total Bayes factor across all ten pancakes therefore equals $1 \times 2 \times 2 \times 2 \times 2 \times 2 \times 2 \times 2 \times 2 \times 2 = 2^9 = 512$.[5]

Finally, another intuition is provided by the law of conditional probability. Let y_1 and y_2 denote two observations. We wish to obtain the predictive performance of a given model for the complete data set, that is, we desire the probability $p(y_1, y_2)$. But by the law of conditional probability this is the same as $p(y_1) \times p(y_2 \mid y_1)$, that is, the probability for the first observation multiplied by the probability for the second observation, *given that the knowledge of the first observation has been properly taken into account*. Chapter 27 examines this important issue in more detail.

[5] Note that after the first pancake, our competing hypotheses consist of *chances* (i.e., fixed beliefs that are not subject to updating: $\theta_S = 1$ for Sidney and $\theta_J = 1/2$ for Jackie) so that we are allowed to multiply the likelihood ratios.

A Bound on the Evidence

In the previous section we showed that the predictive performance of Tabea and Elise was virtually identical for the final batch of $2700 + 4500 = 7200$ pancakes (i.e., $\mathrm{BF}_{te}^{\text{batch3}} \approx 1.04$). In other words, after the first 800 pancakes were in, the remaining 7200 did almost nothing to change our opinion on who is the better bacon forecaster. This suggests that there may be an upper bound on the evidence in Tabea's favor. We first explore this possibility by systematically increasing the number of pancakes while retaining the 3:5 bacon to vanilla ratio. The results are shown in Table 13.3.

The left two columns of Table 13.3 show how the number of bacon and vanilla pancakes increase; the column 'Evidence' shows the corresponding Bayes factor in favor of Tabea, and the rightmost column shows the associated posterior probability that Tabea is a better bacon forecaster than Elise.[6] The table provides support for our intuition that the evidence is bounded. For example, after 80,000 pancakes the Bayes factor in favor of Tabea is 23.83, whereas after 800,000 pancakes it is 23.84: a minuscule increase after adding 720,000 pancakes.

[6] This is calculated under the assumption that Tabea and Elise are equally likely to be the better bacon forecaster *a priori*.

From a mathematical perspective, however, the demonstration in Table 13.3 means little: who is to say that the evidence will not continue to increase, albeit very slowly? As demonstrated in the Appendix Chapter 35, the intuition from Table 13.3 is in fact correct. That is, when the predictive performance of two beta distributions are compared, there is

Table 13.3: Relative predictive performance for Tabea's beta(4,4) prior distribution on θ_{EJ} versus Elise's beta(9,3) prior distribution as the number of pancakes increases while maintaining a 3:5 bacon to vanilla ratio. The column 'Evidence' refers to the Bayes factor in favor of Tabea over Elise, and the column 'Posterior probability' refers to the associated posterior probability that Tabea is a better bacon forecaster than Elise.

Bacon	Vanilla	Evidence	Posterior probability
3	5	3.80	0.79
30	50	15.96	0.94
300	500	22.77	0.96
3000	5000	23.73	0.96
30000	50000	23.83	0.96
300000	500000	23.84	0.96

an upper bound for the evidence. For the scenario involving Tabea and Elise, Equation 35.16 produces an upper limit of 23.84, consistent with the largest value from Table 13.3. The upper bound on the evidence implies an upper bound on the posterior probabilities. This upper bound is visualized in Figure 13.3, which shows how the posterior probabilities for each of four bacon forecasters approaches an asymptotic value as the number of pancakes increases.

In conclusion, the evidence for the comparison of any number of beta distributions is necessarily limited. Posterior convergence means that, after the data have overwhelmed the prior, forecasters with different initial opinions will have come to agree with one another *a posteriori*. From this point onward, the rival forecasters will make indistinguishable predictions, and consequently no amount of additional data has any diagnostic value whatsoever. This then is the price of vagueness: by assigning mass across all values of θ_{EJ}, as any beta distribution does, each forecaster hedges their bets to some degree – even when their initial prior distribution is wildly inconsistent with the data, this distribution, when updated with incoming information, will eventually transform to a posterior distribution that is highly peaked on the value that is most consistent with the observed data. Consequently, in the case of competing beta distributions, the question who is the better forecaster cannot be answered to any desired degree of certainty, even when the data accumulate indefinitely.[7] In the next chapters we will see that, in order that infinite data may provide infinite evidence, the forecasters need to be willing to make riskier predictions.

[7] In statistical jargon, this means that the procedure is *inconsistent*: as sample size increases, the best option cannot be identified with certainty. For details see Ly and Wagenmakers (2022).

Exercises

1. Figure 13.3 shows some initial noisy fluctuations. What could explain these fluctuations?

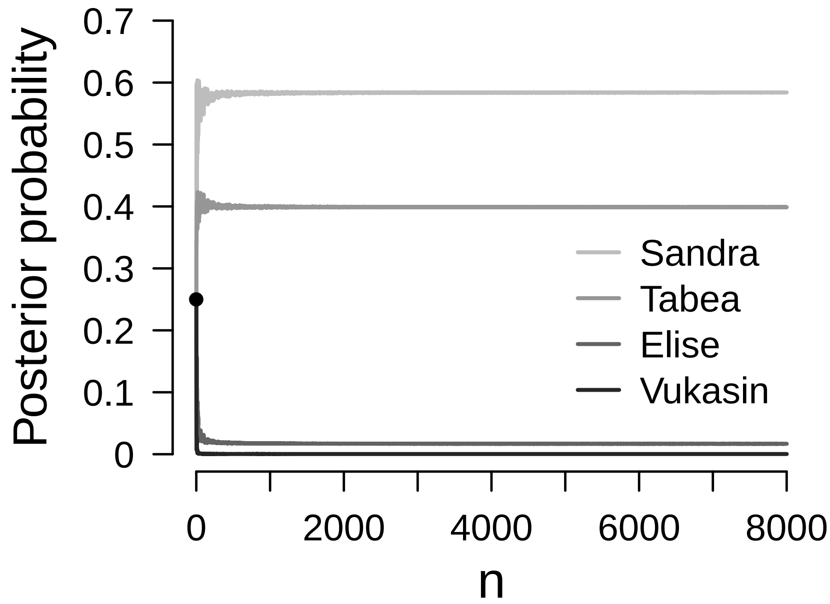


Figure 13.3: Posterior probability of four bacon forecasters as the total number of pancakes n increases while keeping the bacon to vanilla ratio fixed at 3:5 (i.e., every new batch of eight pancakes has three bacon and five vanilla pancakes). In this particular scenario, the posterior probabilities stabilize after a few hundred pancakes.

2. It is the 21st of September, 2021. All across the University of Amsterdam a mask mandate is in place to curtail the COVID-19 pandemic. What concerns us here is θ, the probability that any one student inside the main building on the *Roeterseiland* campus is wearing their face mask correctly (i.e., covering both mouth and nose). (a) Propose three beta prior distributions for θ. Have the first distribution be relatively uninformative, have the second distribution reflect your knowledge as you read these lines, and then create the third distribution to incorporate the additional information that three stewards were present at the building entrance to monitor mask-wearing compliance. (b) Download the mask data at `https://osf.io/4yevk/` and use the *Learn Bayes* module in JASP to conduct a comprehensive Bayesian analysis along the lines sketched in the last two chapters. What is the evidence bound?[8]

[8] A note for teachers: this general exercise type lends itself well to an in-class activity. Divide students in to a few groups and have each group construct their own beta prior for a particular phenomenon of interest. Then analyse the data sequentially and monitor relative predictive performance.

Chapter Summary

This chapter illustrated how the data overwhelm the prior, that is, how data force initial divergent opinions towards posterior agreement. This chapter also showed that the quantification of overall predictive success may occur simultaneously, for a complete data set at once, or it may occur sequentially, batch by batch. In the latter case, in order to obtain the correct result it is essential that the posterior distribution after batch n becomes the prior distribution for the assessment of predictive performance on batch $n + 1$. Finally, the comparison of predictive performance for rival beta distributions may never give a decisive result, even when sample size grows infinitely large – the convergence of posterior opinion implies a bound on the evidence.

Want to Know More?

✓ Ly, A., & Wagenmakers, E.-J. (2022). Bayes factors for peri-null hypotheses. *TEST, 31*, 1121–1142. This paper presents a proof that the Bayes factor for overlapping distributions is bounded: this is the price of vagueness.

Appendix: A *Learn Bayes* Demonstration

The main message of this chapter –the data overwhelm the prior– can be experienced more directly by using the *Learn Bayes* module in JASP. The reader is encouraged to open JASP and follow along. We start by activating the *Learn Bayes* module and selecting *Binomial Estimation*.

Figure 13.4 shows how to specify the data (top panel: three bacon pancakes and five five vanilla pancakes, in the order in which they were baked) and the four models (middle panel: the beta prior distributions for Tabea, Sandra, Elise, and Vukasin). The bottom panel shows that the tab 'Sequential Analysis' contains several options for visualizing how knowledge is updated as the pancakes accumulate.

Ticking the option 'stacked distributions' produces the output shown in Figure 13.5. In each panel, the top row visualizes the prior distribution of θ_{EJ} and the bottom row visualizes the posterior distribution after all pancakes have been taken into account. The change across the rows –from top to bottom– reflect how incoming pancakes gradually update the forecaster's knowledge about the relative plausibility of the different values of θ_{EJ}. For instance, the panels show that as more pancakes are observed, the distributions generally become more narrow, indicating an increase in knowledge about θ_{EJ}.

A comparison across the four panels illustrates how the data drive together opinions that are initially highly divergent. This effect where the 'data overwhelm the prior' is not so clearly present with strong prior opinions and only eight pancakes. Although the forecasters' posteriors are more similar to one another than their priors, the posterior distributions for Tabea and Sandra (top two panels, centered near 0.4) are still markedly different from those of Elise (centered near 0.6) and Vukasin (centered near 0.7).

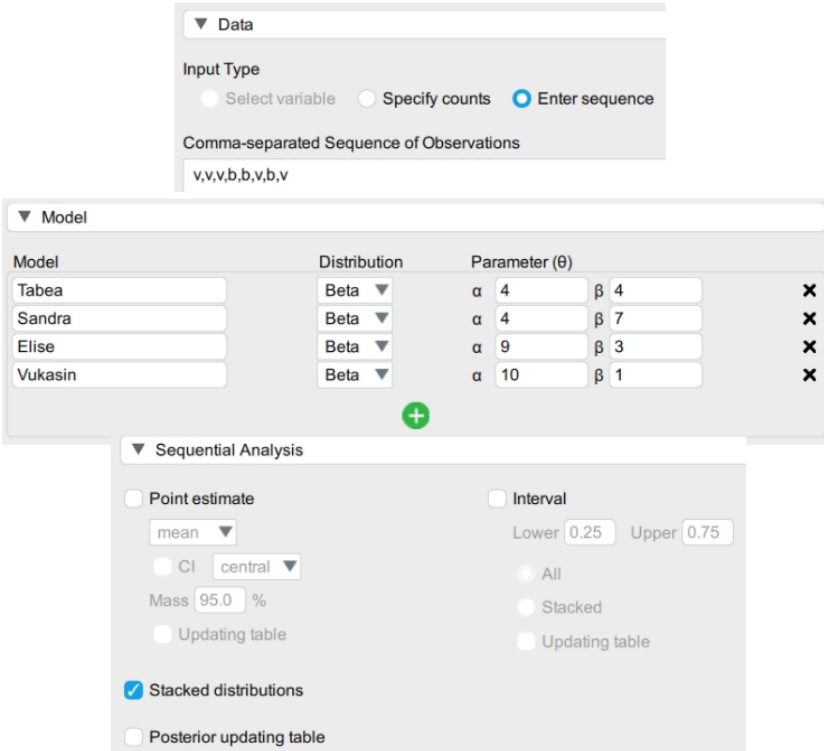


Figure 13.4: JASP screenshot of three input panels from the *Binomial Estimation* routine of the *Learn Bayes* module. The input panels control the sequential estimation of pancake proclivity θ_{EJ} under four different models. Top panel: EJ's pancake data, in order; middle panel: the prior distributions from Tabea, Sandra, Elise, and Vukasin; bottom panel: the options for a sequential analysis.

To highlight the convergence in opinion with increasing data we copy-paste the data row with the original data set nine times, resulting in a total of 80 pancakes, 50 of which are vanilla. The associated sequential analysis with stacked distributions is shown in Figure 13.6. The posterior distributions are now relatively similar across the four pancake forecasters, despite the fact that the prior distributions were relatively dissimilar. The 80 pancakes provide information that is sufficiently strong to drive together the initially divergent beliefs, and these data can therefore be said to have overwhelmed these priors.

The *Sequential Analysis* tab offers additional options that the reader is encouraged to explore. For instance, Figure 13.7 below shows how the posterior mean for θ_{EJ} changes as the pancakes accumulate. The figure confirms that the mean of the distribution converges – the prior means vary considerably between the forecasters, but the posterior means are relatively similar: the data overwhelm the prior. Note that the change in the posterior mean is more pronounced for Vukasin and for Elise than it is for Tabea and Sandra; the reason is that the prior distributions

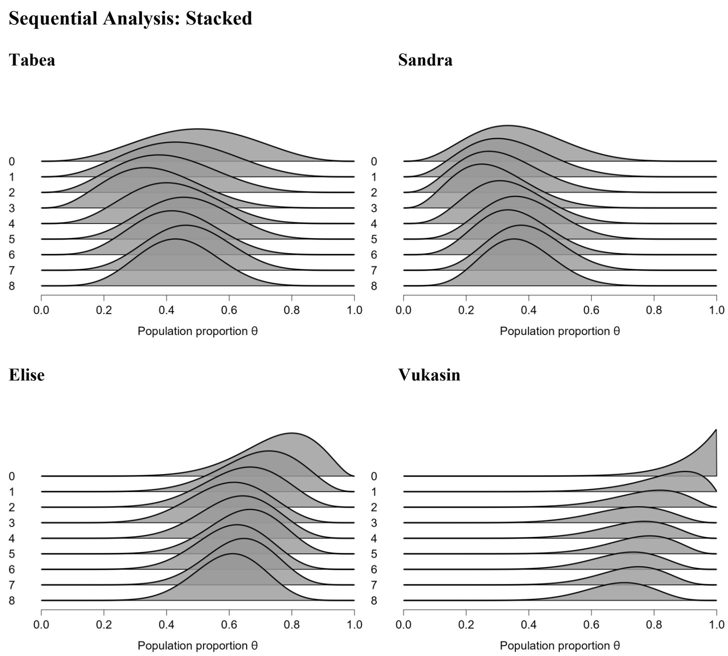


Figure 13.5: Sequential analyses for four forecasters of pancake proclivity θ_{EJ}. After eight pancakes, the posterior distributions still show the impact of the prior distribution. The data were not sufficiently informative to overwhelm these particular priors. Figure from the JASP module *Learn Bayes*.

of Vukasin and Elise put relatively much mass on high values of θ_{EJ}, values that are unlikely in light of the data.

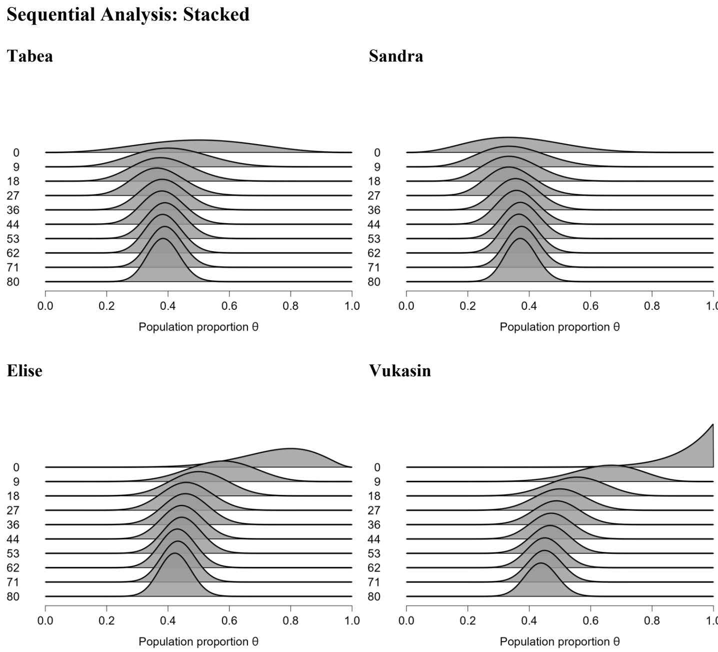


Figure 13.6: Sequential analyses for four forecasters of pancake proclivity θ_{EJ}. After 80 pancakes (of which the last 72 are fictitious), the posterior distributions no longer show much impact of the prior distribution. These particular data can be said to have overwhelmed these particular priors. Figure from the JASP module *Learn Bayes*.

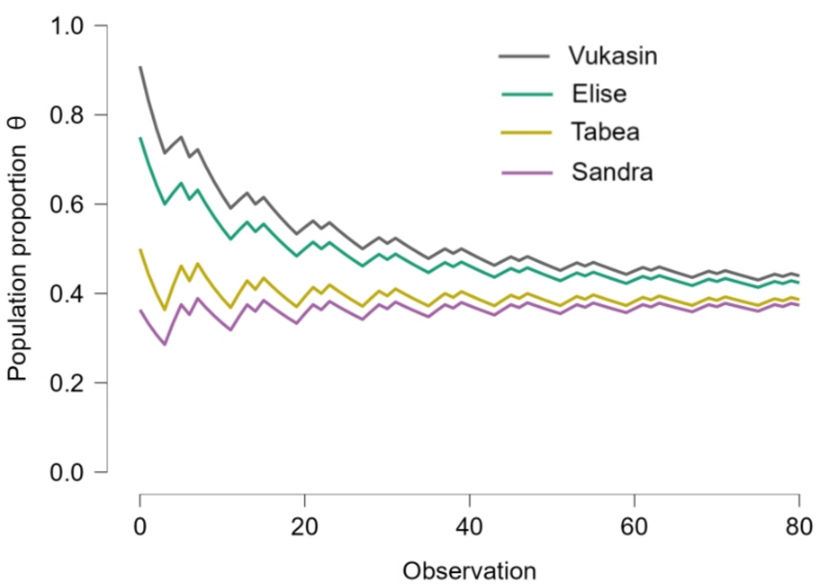


Figure 13.7: Sequential analyses for four forecasters of pancake proclivity θ_{EJ}. After 80 pancakes (of which the last 72 are fictitious), the posterior means for θ_{EJ} have converged and are relatively close. Note that the effect of repeating the original data set nine times is visible in the repeated sawtooth pattern with which the posterior mean changes. Figure from the JASP module *Learn Bayes*.

Part III

Coherent Learning, Jeffreys Style

14 *A Crack in the Laplacean Edifice*

[The Laplace rule] therefore expresses a violent prejudice against any general law, a totally unacceptable description of the scientific attitude.

Jeffreys, 1974

Chapter Goal

This chapter exposes the Achilles heel of Laplacean inference: the Principle of Insufficient Reason, also known as the Principle of Indifference. Although this principle appears neutral and innocuous –probability mass is divided evenly across all parameter values and events– it implies a denial without evidence that a general law is ever true. Universal generalizations that involve a necessary cause (e.g., "all AIDS patients have been exposed to HIV") are deemed false from the outset, in violation of both common sense and scientific practice.

Problems with the Principle of Indifference

For historical and educational reasons, we first consider the Principle of Indifference as it applies to binomial data governed by an unknown chance θ. The Principle of Indifference dictates that θ be assigned a uniform prior distribution, indicating that all possible values for θ are deemed equally plausible *a priori*.

For instance, suppose that, as discussed in earlier chapters, θ_{EJ} represents EJ's tendency to bake his pancakes with bacon. The uniform prior distribution on θ_{EJ} (cf. Figure 8.3) induces a prior predictive distribution that assigns equal probability to each possible number of bacon pancakes (out of a total of n to-be-observed pancakes).[1] For a to-be-observed stack of four pancakes, Figure 14.1 shows that the uniform distribution on θ_{EJ} produces five equally likely outcomes for the number of pancakes that have bacon.[2]

At first sight, the uniform prior assignment across θ_{EJ} appears neutral and 'objective', untarnished by prior knowledge that may push

[1] One of the exercises from the next chapter is to prove this result.

[2] NB. Four pancakes yield five possible outcomes, as the outcome that none of the four pancakes has bacon is also in the cards.

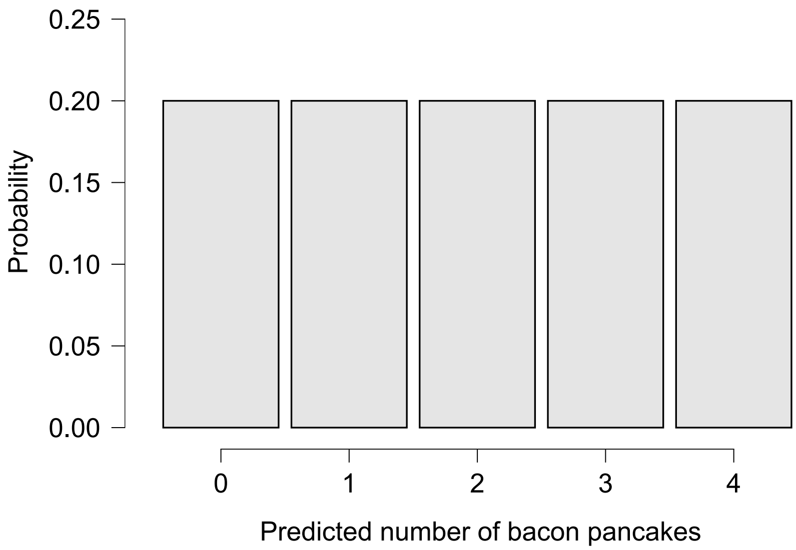


Figure 14.1: Predicted number of pancakes that come with bacon, out of a total of four. The beta-binomial predictions are based on the uniform beta(1,1) prior distribution on bacon proclivity θ_{EJ} motivated by the Principle of Indifference. Figure from the JASP module *Learn Bayes*.

the posterior distribution in the direction of the analyst's expectations. However, deeper reflection reveals that the uniform assignment harbors an extreme bias: it rules out the possibility of universal generalizations such as 'all ravens are black'.

In particular, the uniform $\theta_{EJ} \sim \text{beta}(1, 1)$ distribution assigns probability zero to any specific value of θ_{EJ}, including the value $\theta_{EJ} = 1$ (i.e., 'All of EJ's pancakes come with bacon'). As a result, when the stack of to-be-observed pancakes increases, the prior predictive probability that all pancakes have bacon decreases, as it equals $1/(n+1)$: the prior probability that all pancakes will have bacon approaches zero as the stack grows large.

This prejudice against $\theta_{EJ} = 1$ is also evident from Laplace's Rule of Succession. Recall from Chapter 9 that if $\theta \sim \text{beta}(1, 1)$ and an unbroken string of s successes has been observed, the probability of a further unbroken string of k successes equals

$$\frac{s+1}{s+k+1}.$$

It is clear that, as k increases and the sequence of predicted successes lengthens, its probability decreases towards zero. Thus, no matter how long the initial unbroken sequence of s successes, one would remain firmly convinced that, with sufficient patience (i.e., sufficiently high k), an exception would eventually occur. This firm conviction is unshaken by changing the parameters that define the shape of the beta prior distribution. For general α and β, the probability of a future unbroken string of k successes, after having observed s successes in the past, is

$$\prod_{i=0}^{k-1} \frac{\alpha + s + i}{\alpha + s + i + \beta}, \tag{14.1}$$

a product where each successive term represents the probability of observing another success in the predicted sequence of k successes. When k grows large the product of probabilities inevitably approaches zero, irrespective of the values for s, α, and β.[3]

[3] As long as $\beta > 0$ and $s < \infty$. See the exercises for mathematical details.

Thus, the Principle of Indifference denies the possibility that a general law or universal generalization can ever be true. Irrespective of the extent of previous experience, an exception is deemed certain to occur at some point in the future. Deviating from the 'indifferent' beta$(1, 1)$ prior by changing α and β does nothing to alter the belief that exceptions are inevitable.

In *pure induction*, however, an unbroken sequence of confirmatory instances has been observed, and a key question of interest is how much evidence the observed instances offer in support of the general law that *all* instances will be confirmatory. For instance, a mathematician may observe that several even integers greater than four can be decomposed as the sum of two odd prime numbers. For instance, $6 = 3 + 3$, $8 = 3 + 5$, $10 = 3 + 7 = 5 + 5$, $12 = 7 + 5$, etc. After working through enough instances, the mathematician may feel sufficiently confident to conjecture that *all* instances follow the rule. The problem above is the famous Goldbach conjecture, a puzzle in number theory that remains unsolved to this day. Despite the fact that a mathematical proof has remained elusive, the conjecture has been confirmed for all integers up to 4×10^{18}, a relatively strong level of inductive support.[4] One may apply Laplace's Principle of Indifference to the Goldbach conjecture and assign a beta$(1, 1)$ prior distribution to θ, the chance that any even number greater than four can be decomposed as the sum of two odd primes. However, this implies a denial without evidence that the Goldbach conjecture may be true. According to the Principle of Indifference, an exception is sure to arise if only sufficient numbers are subjected to inspection, an opinion that is manifestly absurd.[5]

[4] `http://sweet.ua.pt/tos/goldbach.html`

[5] Readers interested in learning more about the role of induction in mathematics are referred to Pólya (1954a) and Gronau and Wagenmakers (2018).

Similarly, a team of medical doctors may hypothesize that Alzheimer's disease is caused by a fungal infection of the central nervous system (e.g., Pisa et al. 2015). This hypothesis entails that every patient who has died of Alzheimer's should have traces of the fungus in their brains. Clearly, every new Alzheimer's patient found to have such a fungus infection provides support for the doctors' hypothesis. Indeed, if the fungus is a *necessary condition* for Alzheimer's to develop, then *all* patients with Alzheimer's will have the fungus – a possibility that the Laplacean Principle of Indifference steadfastly denies. Likewise, the

Principle of Indifference would have one believe that if only enough patients with AIDS were examined, it is inevitable that in due time an AIDS patient is found who has *not* been infected with HIV. Because HIV is the virus that actually causes AIDS, this opinion is again manifestly absurd.

Finally, suppose that under regular circumstances (e.g., room temperature, normal air pressure) you drop a small cube of sugar into a large, boiling cup of tea. You stir the cup with a spoon. The sugar cube will dissolve – every single time. By rejecting this notion, the Principle of Indifference denies the validity of physical laws of nature, while remaining silent on the mysterious processes that would produce such a remarkable exception. It is safe to say that even the staunchest proponents of the Principle of Indifference were uneasy about the implicit denial of any general law. For instance, De Morgan considered it "at variance with all our notions":

> "If as before, the first m Xs observed have all been Ys, and we ask what probability thence, and thence only, arises that the next n Xs examined shall all be Ys, the answer is that the odds in favour of it are $m + 1$ to n, and against it n to $m + 1$. No induction then, however extensive, can by itself, afford much probability to a universal conclusion, if the number of instances to be examined be very great compared with those which have been examined. If 100 instances have been examined, and 1000 remain, it is 1000 to 101 against all the thousand being as the hundred.
>
> This result is at variance with all our notions; and yet it is demonstrably as rational as any other result of the theory. The truth is, that our notions are not wholly formed on what I have called the *pure induction*. In this it is supposed that we know no reason to judge, except the mere mode of occurrence of the induced instances. Accordingly, the probabilities shown by the above rules are merely *minima*, which may be augmented by other sources of knowledge. For instance, the strong belief, founded upon the most extensive previous induction, that phenomena are regulated by uniform laws, makes the first instance *of a new case*, by itself, furnish as strong a presumption as many instances would do, independently of such belief and reason for it." (De Morgan 1847/2003, pp. 214-215)

In sum, when the goal is to address a general law or a universal generalization (e.g., by quantifying the empirical support in its favor) one cannot use the Laplacean Principle of Indifference, because its point of departure is to deny that such laws exist at all.

"[Jeffreys's theory] takes as a fact of human thought that we are willing to accept a general law on amounts of observational evidence that are available in practice, and as this contradicts results derivable from Laplace's assessment of prior probabilities and its natural extension to quantitative laws, we infer that Laplace's assessment does not represent our state of mind when we begin an investigation." (Jeffreys 1937b, p. 245)

The Finite Version of Pure Induction

Up to now we have considered a uniform distribution on the chance θ (say EJ's bacon proclivity θ_{EJ}) which induces a uniform distribution on the number of pancakes with bacon (e.g., Figure 14.1). The total number of to-be-observed pancakes is potentially infinite.

Alternatively, we may entertain a *finite* version of the Principle of Indifference, as already suggested by De Morgan's quotation above. For instance, suppose you are confronted with a stack of four pancakes. What is the probability that all of them have bacon? Instead of defining a prior distribution on θ, the finite version of the problem of pure induction directly assigns each possible composition of the stack an equal probability. Denoting by $\mathcal{H}_{ib,jv}$ the hypothesis that the stack consists of i bacon pancakes and j vanilla pancakes, we have

$$\begin{aligned} p(\mathcal{H}_{4b,0v}) &= 1/5 \\ p(\mathcal{H}_{3b,1v}) &= 1/5 \\ p(\mathcal{H}_{2b,2v}) &= 1/5 \\ p(\mathcal{H}_{1b,3v}) &= 1/5 \\ p(\mathcal{H}_{0b,4v}) &= 1/5. \end{aligned}$$

This is the same assumption that was made in the infinite version (cf. Figure 14.1), but there it was a consequence of assigning a uniform distribution to θ.

We then observe, say, one bacon pancake. This observation is most likely under $\mathcal{H}_{4b,0v}$, whereas $\mathcal{H}_{0b,4v}$ is eliminated from contention. Crucially, this observation also changes the nature of the hypotheses – because the pancakes are inspected *without* replacement, the updated hypotheses about the remaining three pancakes are

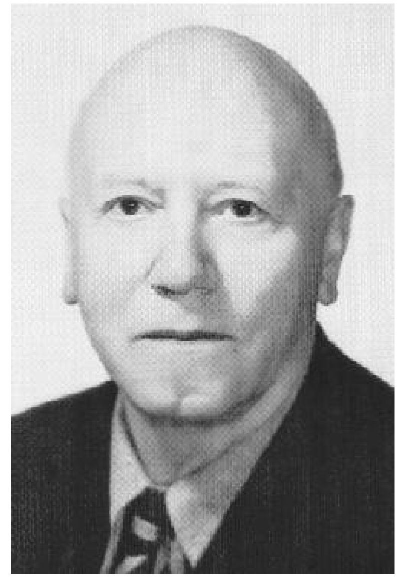
$$\begin{aligned} p(\mathcal{H}_{3b,0v}) &= 4/10 \\ p(\mathcal{H}_{2b,1v}) &= 3/10 \\ p(\mathcal{H}_{1b,2v}) &= 2/10 \\ p(\mathcal{H}_{0b,3v}) &= 1/10. \end{aligned}$$

As the stack dwindles and all pancakes inspected so far have come with bacon, the hypothesis is increasingly plausible that all remaining pancakes will also come with bacon.

For the finite version of pure induction, analyzed according to the Principle of Indifference, Broad (1918) found that with uniform prior assignment on the composition of a stack of n pancakes, and after having observed an unbroken sequence of s bacon pancakes, the probability that the remaining $n - s = k$ pancakes will also have bacon equals

$$\frac{s+1}{n+1}.$$

This result is *identical* to that of the infinite version, a correspondence that some found surprising and others found obvious.[6] Regardless, the finite version highlights the bias inherent in the Principle of Indifference even more than the infinite version. Suppose the number of instances of interest n is very large – the number of birds in England,



Charlie Dunbar Broad (1887–1971). "Broad used Laplace's theory of sampling, which supposes that if we have a population of n members, r of which may have a property φ, and we do not know r, the prior probability of any particular value of r(0 to n) is $1/(n+1)$. Broad showed that on this assessment, if we take a sample of number m and find all of them with φ, the posterior probability that all n are φ's is $(m+1)/(n+1)$. A general rule would never acquire a high probability until nearly the whole of the class had been sampled. We could never be reasonably sure that apple trees would always bear apples (if anything). The result is preposterous, and started the work of Wrinch and myself in 1919–1923. Our point was that giving prior probability $1/(n+1)$ to a general law is that for n large we are already expressing strong confidence that no general law is true." (Jeffreys 1980, p. 452).

[6] Broad was not the first to derive this result. An in-depth discussion is provided by Zabell 1989, p. 286 and Todhunter 1865, pp. 454-457.

the number of electrically neutral atoms in the Milky Way, etc. Suppose s, the number of instances already observed and found to be confirmatory, is also large, but small compared to n. Then, the probability that all $n - s$ non-observed instances are also confirmatory is close to the proportion of inspected samples, $^{s}/_{n}$. Observe half of the electrically neutral atoms in the Milky Way, and find that all of them have as many protons as electrons – according to the Principle of Indifference, this should instill a level of confidence worth no more than an even bet that the same regularity will hold for the remaining half.

"What Laplace's rule says, in fact, is that the prior probability of the general rule is $1/(N + 1)$, and it amounts to a denial without evidence that there are any general laws." Jeffreys (1950, p. 315)

Similarly, if you find a bag of 100 coins, and the first 50, randomly drawn without replacement, are either double-heads or double-tails, the Principle of Indifference holds that your confidence that the remaining 50 coins are of the same type ought to be no higher than $^{51}/_{101} \approx .505$.

In the words of Jeffreys,

> "The last result [i.e., the $^{s+1}/_{n+1}$ rule for the finite scenario] was given by Broad (...) and was the first clear recognition, I think, of the need to modify the uniform assessment if it was to correspond to actual processes of induction. It was the profound analysis in this paper that led to the work of Wrinch and myself.† We showed that Broad had, if anything, understated his case, and indicated the kind of changes that were needed to meet its requirements. The rule of succession had been generally appealed to as a justification of induction; what Broad showed was that it was no justification whatever for attaching even a moderate probability to a general rule if the possible instances of the rule are many times more numerous than those already investigated. (...) Thus I may have seen 1 in 1,000 of the 'animals with feathers' in England; on Laplace's theory the probability of the proposition, 'all animals with feathers have beaks', would be about $1/1000$. This does not correspond to my state of belief or anybody else's. (...)
>
> The fundamental trouble is that the prior probabilities $1/N + 1$ attached by the theory to the extreme values are so utterly small that they amount to saying, without any evidence at all, that it is practically certain that the population is not homogenous in respect of the property to be investigated; so nearly certain that no conceivable amount of observational evidence could appreciably alter this position." (Jeffreys 1961, pp. 128-129)

†*Phil. Mag.* **42**, 1921, 369-90; **45**, 1923, 368-74.

This, then, is the key problem: the Principle of Indifference treats all hypotheses the same, and spreads out its prior mass evenly among them. But some hypotheses deserve special attention. Principle of Indifference does not recognize this, thereby preventing general laws from ever reaching appreciable plausibility. This procedure violates both common sense and scientific practice.

The solution to this conundrum was devised by a series of papers by Dorothy Wrinch and Harold Jeffreys, the main message of which is outlined in the next chapter.

Exercises

1. Apply the Principle of Indifference to inference of temperature. What prior distribution is implied? Are the predictions from this prior distribution reasonable?

2. Consider again Equation 14.1. Derive this equation using the material from the appendix of Chapter 9, and then prove that when $k \to \infty$, the product goes to zero.

3. The main text states, "Clearly, every new Alzheimer's patient found to have such a fungus infection provides support for the doctors' hypothesis." Assume that 1000 Alzheimer's patients are examined and all have traces of the fungus. Argue against the doctors' hypothesis that the fungus causes Alzheimer's.

4. In the section 'The Finite Version of Pure Induction', the prior probability for each of five hypotheses is being updated by the observation that the first pancake from the stack has bacon. Confirm that that the updated probabilities are correct.

Chapter Summary

The Laplacean Principle of Indifference is not indifferent at all, but embodies a denial without evidence that all universal generalizations are false.[7]

[7] It is perhaps ironic that this denial itself is a universal generalization.

Want to Know More?

✓ Broad, C. D. (1918). On the relation between induction and probability (Part I.). *Mind, 27*, 389-404.

✓ Jeffreys, H. (1961). *Theory of Probability (3rd ed.).* Oxford: Oxford University Press. Pages 125-129 offer a good summary of the problem with the Laplacean Principle of Indifference.

✓ Pearson, K. (1892/1937). *The Grammar of Science.* London: J. M. Dent & Sons.

✓ Perks, W. (1947). Some observations on inverse probability including a new indifference rule. *Journal of the Institute of Actuaries, 73*, 285–334. "At one time, the rule of succession was regarded as a logical justification for induction, for scientific inference. But Pearson's result of $.5$ for the probability that the next $(n+1)$ trials will be successes, after n successes in n trials, is clearly too low and unacceptable as a representation of the scientific process of experimentation to test

a proposed scientific law. As Jeffreys says (p. 102), the result does not correspond with anybody's way of thinking." (p. 295)

✓ Polya, G. (1954). *Mathematics and Plausible Reasoning: Vol. I. Induction and Analogy in Mathematics*. Princeton, NJ: Princeton University Press. Highly recommended for those who wish to learn more about the role of induction in mathematics.

✓ Zabell, S. L. (1989). The rule of succession. *Erkenntnis, 31*, 283-321. Essential reading.

✓ Zabell, S. L. (2005). *Symmetry and Its Discontents: Essays on the History of Inductive Probability*. Cambridge: Cambridge University Press. Scholarly, informative, and highly recommended.

15 *Wrinch and Jeffreys to the Rescue*

Dorothy Maud Wrinch (1894–1976). In collaboration with Harold Jeffreys, Dorothy Wrinch was the first to propose a Bayes factor (Wrinch and Jeffreys 1921). Together with Harold Jeffreys she also demonstrated the importance of assigning probability to point null hypotheses – an important lesson that many statisticians continue to ignore at their peril (Etz and Wagenmakers 2017, Howie 2002).

The theory we are attempting to construct is one that includes the processes actually employed by scientific workers; since psychology is by definition the study of behaviour, this work may perhaps be regarded as a part of psychology.

Wrinch & Jeffreys, 1923

Chapter Goal

As discussed in the previous chapter, the main problem with the Laplacean Principle of Indifference is that it 'expresses a violent prejudice against any general law'. This chapter outlines how Dorothy Wrinch and Harold Jeffreys overcame this problem by assigning the general law its own prior probability. Consequently, the Wrinch-Jeffreys proposal allows data to support the general law.

Jeffreys's Oven

Ever since its inception, Bayesian inference (originally known as 'inverse probability') had almost always involved uniform priors. When Broad and others highlighted that such priors had undesirable consequences, this could be interpreted to mean that there is something undesirable about Bayesian inference in general. In response, Harold Jeffreys presented a compelling analogy:

> "Bayes and Laplace, having got so far, unfortunately stopped there, and the weight of their authority seems to have led to the idea that the uniform distribution of the prior probability was a final statement for all problems whatever, and also that it was a necessary part of the principle of inverse probability.[1] *There is no more need for the latter idea than there is to say that an oven that has once cooked roast beef can never cook anything but roast beef.*" (Jeffreys 1961, p. 118; emphasis added)

As outlined in the previous chapter, the problem with the uniform prior distribution on a chance θ is that it expresses a denial without evidence that a universal generalization is true. Broad (1918) showed

[1] EWDM: Laplace did not always recommend the uniform distribution. For instance, at the end of his 1774 essay he discusses the chance of observing a particular number of pips from a regular die. He argues that there is always some deviation from $1/6$ but that this deviation is very small.

that for a large but finite set of instances, the probability that all these instances follow the general law is about equal to the proportion of instances that have been inspected so far. Suppose the entire zombie population counts 5,000,000 members. Of these, 500,000 have been observed, and all are known to be hungry. According to the Principle of Indifference, the probability that all of the remaining 4,500,000 zombies are also hungry equals only $^{500{,}001}/_{5{,}000{,}001} \approx {}^{1}/_{10}$. This cannot be right.

But how should the uniform distribution be adjusted to obtain a result that is in line with common sense and with statistical practice? Dorothy Wrinch and Harold Jeffreys (1921,1923) suggested a straightforward solution: respect the general law and assign it a separate prior probability. That is, "If we are ever to attach a high probability to a general rule, on any practicable amount of evidence, it is necessary that it must have a moderate probability to start with." (Jeffreys 1961, p. 128). In the zombie example, the universal generalization $\theta = 1$ ('all zombies are hungry') may for instance be deemed equally likely *a priori* as its denial (i.e., the Laplacean assumption $\theta \sim \text{beta}(1, 1)$).

"Any result we offer must agree with common-sense and with results that can be logically or mathematically deduced from common-sense." (Wrinch and Jeffreys 1921, p. 378)

Thus, one way to view the Wrinch-Jeffreys setup is as involving two competing hypotheses: the general law and the denial of the general law. The general law provides a relatively simple account of the world; in statistics it is termed the 'null hypothesis', $\mathcal{H}_0$, and its key parameter is fixed to a specific value of interest. In terms of concepts discussed in Chapter 2, there is no epistemic uncertainty for the fixed parameter. The restriction imposed by $\mathcal{H}_0$ is relaxed under the more complicated hypothesis that allows θ to take on any value within a certain range – θ is not 'fixed', but 'free', and the associated epistemic uncertainty is quantified by a prior distribution. In statistics, the more complicated hypothesis is termed the 'alternative hypothesis', $\mathcal{H}_1$.[2] With these rival hypotheses in play, the learning process can then be formalized as follows (Wrinch and Jeffreys 1921, p. 387):

[2] Because $\mathcal{H}_1$ allows θ to take on different values, it is also known as a 'composite' hypothesis.

$$\underbrace{\frac{p(\mathcal{H}_1 \mid \text{data})}{p(\mathcal{H}_0 \mid \text{data})}}_{\substack{\text{Posterior beliefs}\\ \text{about hypotheses}}} = \underbrace{\frac{p(\mathcal{H}_1)}{p(\mathcal{H}_0)}}_{\substack{\text{Prior beliefs}\\ \text{about hypotheses}}} \times \underbrace{\frac{p(\text{data} \mid \mathcal{H}_1)}{p(\text{data} \mid \mathcal{H}_0)}}_{\text{Bayes factor}}. \tag{15.1}$$

Another way to view the Wrinch-Jeffreys setup is as a prior distribution on chance θ that consists of a mixture between a Laplacean 'slab' where $\theta \sim \text{beta}(1, 1)$ and a Wrinchean 'spike' at $\theta = 1$ (e.g., Mitchell and Beauchamp 1988). Figure 15.1 shows the spike-and-slab distribution where the probability on the spike equals $^1/_2$. The model comparison view and the spike-and-slab view are mathematically identical, but are used for different purposes. The model comparison view is preferred by those who wish to assess the extent to which the data support $\mathcal{H}_0$

or $\mathcal{H}_1$ (i.e., the primary interest is on the models and the competition between them), whereas the spike-and-slab view is preferred by those who wish to estimate the parameter θ while taking into account the fact that the general law may be true (i.e., the primary interest is on θ and the models are a nuisance factor that is to be integrated out using the law of total probability).

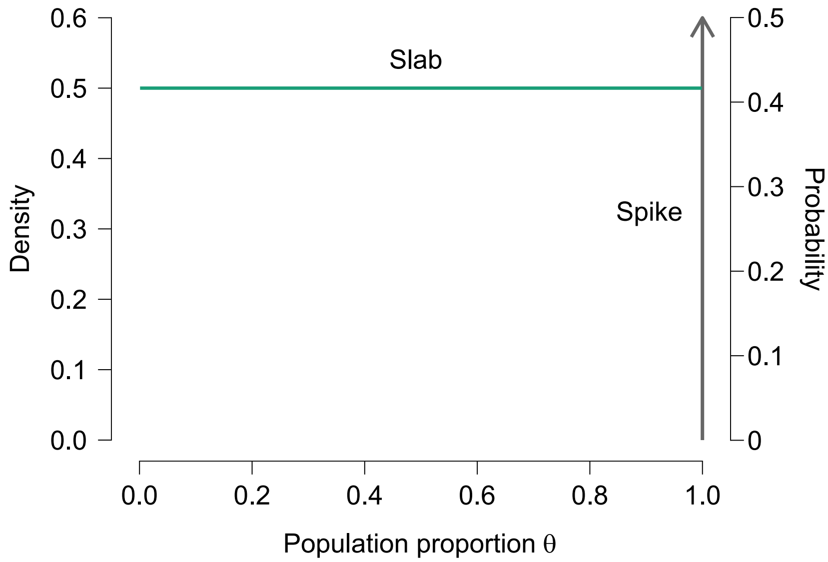


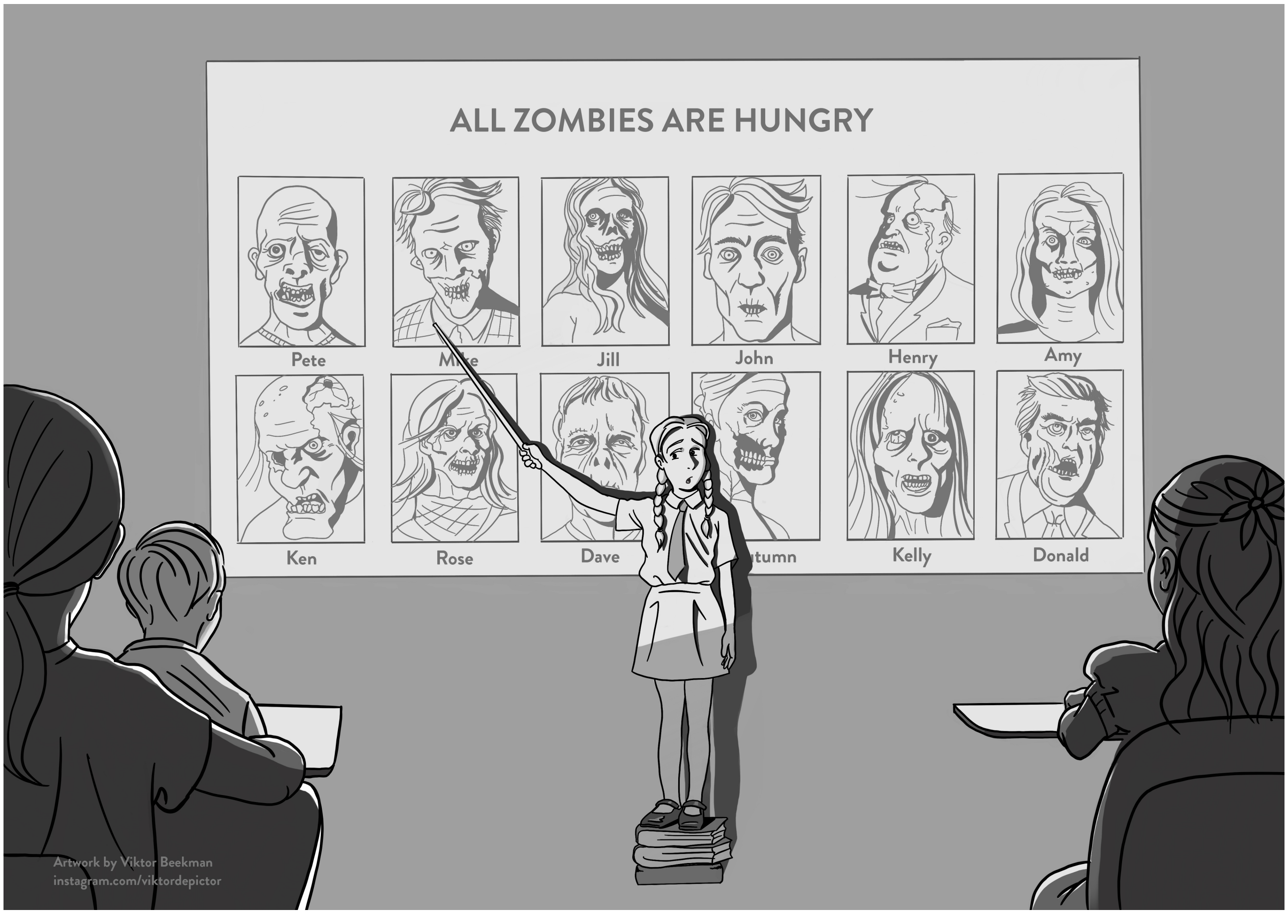
Figure 15.1: The Wrinch-Jeffreys 'spike-and-slab' proposal features probability mass concentrated at a single point. Here, the spike is located at $\theta = 1$, the universal generalization; the height of the spike equals .50 (second y-axis) and represents its prior probability. The slab corresponds to the Laplacean uniform prior distribution on θ, and the area under the slab equals .50, the prior probability of the slab component. Figure from the JASP module *Learn Bayes*.

Below we provide a concrete example of how the Wrinch-Jeffreys proposal successfully overcomes the limitations of the Laplacean Principle of Indifference that is based on assigning θ a continuous distribution.[3]

[3] As discussed in Chapter 16, the concrete implementation of this setup was pioneered by J. B. S. Haldane in 1932.

Are all Zombies Hungry?

Kate is a goth girl fascinated by bats, medieval torture instruments, and the undead. Next week, Kate has to give an in-class presentation with the preliminary title "Hangry? The Quintessential Zombie PR Problem". As part of the assignment, she needs to discuss whether or not all zombies are hungry. Lacking the relevant biological background to address this question theoretically, Kate decides to approach the issue empirically, by visiting zombies and keeping track of how many are hungry and how many are satiated.



Kate presents her school project. Figure available at `BayesianSpectacles.org` under a CC-BY license.

Our example data set features the first 12 zombies that Kate visited. All of them were undeniably hungry.[4] How much evidence is this for the universal generalization that 'all zombies are hungry'? Clearly this law gains plausibility with every hungry zombie that is encountered, whereas the presence of a single satiated zombie refutes the law decisively.[5] Let's make this more concrete by a Bayesian analysis.[6]

[4] Ravenous, even.

[5] In the words of Pólya (1954a, p. 6), the law would be "irrevocably exploded".

[6] More mundane scenarios that allow a similar analysis include 'all ravens are black', 'all electrically neutral electrons have the same numbers of positrons and electrons', and 'all positive even integers ≥ 4 can be expressed as the sum of two odd primes' (i.e., the Goldbach conjecture). See also Berger and Jefferys (1992).

Data Analysis

Kate wants to know the extent to which the data support the proposition that "all zombies are hungry". Statistically, this proposition corresponds to a null hypothesis that assigns a fixed value of 1 to the binomial chance θ – the probability that any given zombie is hungry. In other words, $\mathcal{H}_0 : \theta = 1$. The alternative hypothesis $\mathcal{H}_1$ relaxes the constraint on θ and allows it to take on values lower than 1. For historical and educational purposes, we assume a uniform prior distribution for θ under $\mathcal{H}_1$, that is, $\mathcal{H}_1 : \theta \sim \text{beta}(1, 1)$, such that every value of θ is deemed equally likely *a priori*.

We also assume that, *a priori*, both hypotheses are equally plausible, such that $p(\mathcal{H}_0) = p(\mathcal{H}_1) = 1/2$. The joint prior on θ across the two hypotheses therefore corresponds to the situation depicted in Figure 15.1.

In contrast to the setup that is entertained by Kate, a Laplacean analysis would focus solely on $\mathcal{H}_1$ and ignore $\mathcal{H}_0$. The result of such a Laplacean analysis is shown in Figure 15.2. After having seen 12 hungry zombies, the beta(1,1) prior distribution on θ has been updated to a beta(13,1) posterior distribution. This posterior distribution is concentrated on high values for θ. Laplace's Rule of Succession states that the probability that the next zombie is hungry equals $13/14 \approx .93$.

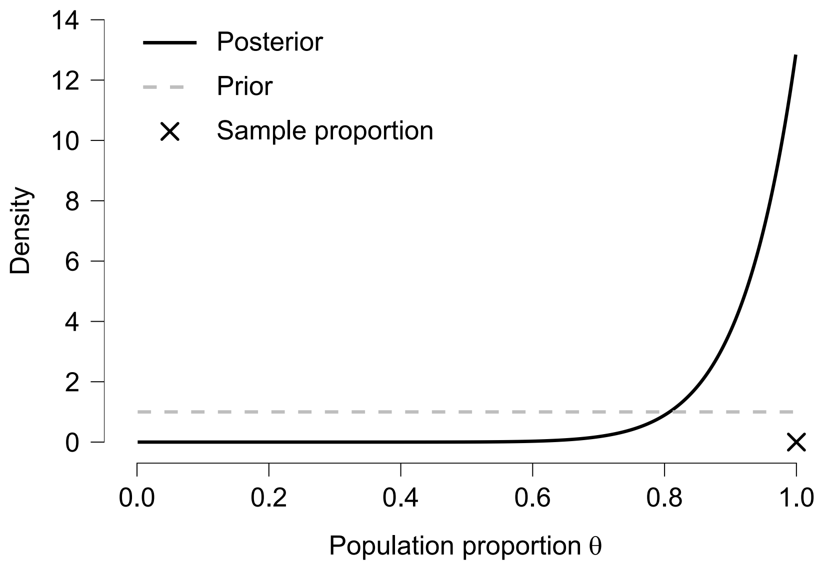


Figure 15.2: A Laplacean analysis of the zombie data. A beta(1,1) prior distribution is updated to a beta(13,1) posterior distribution after having observed that all of 12 zombies are hungry. Figure from the JASP module *Learn Bayes*.

This Laplacean analysis, however, is unable to address Kate's key question, which is 'are *all* zombies hungry?'. As explained in the previous chapter, the Laplacean analysis will answer this question with 'no, absolutely not' irrespective of how many hungry zombies have already been observed.[7] Kate could eye-ball the posterior distribution for θ that was obtained under the implicit Laplacean assumption that 'not all zombies are hungry' – but this is not something that Kate wants to *assume*; it is something that she wants to *test*.

[7] This assumes that the number of observed zombies is finite, and the zombie population is infinite.

In order to test $\mathcal{H}_0 : \theta = 1$ versus $\mathcal{H}_1 : \theta \sim \text{beta}(1, 1)$ we need to consider the predictive adequacy of the two hypotheses for the data at hand. Kate observed $s = 12$ hungry zombies out of a total of $n = 12$. Given that 12 zombies are observed, the null hypothesis can make no other prediction. That is, under $\mathcal{H}_0$ the probability of observing $s = 12$ equals 1 – no other data are possible. In other words, $\mathcal{H}_0$ makes a highly specific and daring prediction. The prediction of $\mathcal{H}_0$ for the data obtained is shown by the highlighted bar in Figure 15.3.

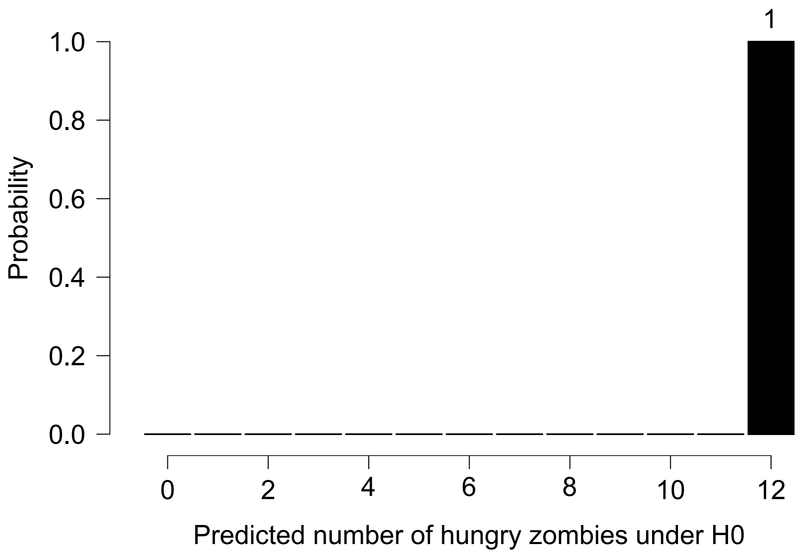


Figure 15.3: The universal generalization $\mathcal{H}_0 : \theta = 1$, 'all zombies are hungry', makes only a single, precise prediction for Kate's data set of 12 zombies. Figure from the JASP module *Learn Bayes*.

The situation is dramatically different for the alternative hypothesis $\mathcal{H}_1$. This hypothesis states that every value of θ is equally likely; the previous chapter showed that, predictively, this means every possible value for s out of $n = 12$ is deemed equally likely to occur.[8] There are 13 values for s (the count starts at $s = 0$ hungry zombies), and therefore the alternative hypothesis assigns probability $1/13$ to the observed data $s = 12$. The predictions of $\mathcal{H}_1$ are shown in Figure 15.4.

[8] One of the exercises for this chapter is to prove this result.

In contrast to $\mathcal{H}_0$, the alternative hypothesis $\mathcal{H}_1$ has hedged its bets, dividing its predictive resources evenly across all possible 13 outcomes. In Bayesian inference, such statistical cowardice comes at a price. Under the daring $\mathcal{H}_0 : \theta = 1$, the probability of the observed data (i.e., $s = 12$) equals 1; under the cowardly $\mathcal{H}_1 : \theta \sim \text{beta}(1, 1)$, the probability of the observed data equals only $1/13$. The ratio of these predictions equals the Bayes factor shown in Equation 15.1. Specifically, this Bayes factor equals

"Thus the more precise the inferences given by a law are, the more its probability is increased by a verification, even if the contradictory law also gives a prediction consistent with the observation. (...) We may say that to make predictions with great accuracy increases the probability that they will be found wrong, but in compensation they tell us much more if they are found right." (Jeffreys 1973, p. 39)

$$\text{BF}_{10} = \frac{p(s = 12 \mid n = 12, \mathcal{H}_1)}{p(s = 12 \mid n = 12, \mathcal{H}_0)} = \frac{1/13}{1} = 1/13.$$

This is the Bayes factor in favor of $\mathcal{H}_1$ over $\mathcal{H}_0$; for ease of interpretation, it is customary to switch numerator and denominator whenever the Bayes factor is lower than 1. Here this means that instead of $\text{BF}_{10} = 1/13$, we prefer the equivalent expression $\text{BF}_{01} = 13$.[9] We can interpret this Bayes factor in multiple ways:

[9] NB. The first subscript to the Bayes factor indicates the model in the numerator; the second subscript indicates the model in the denominator.

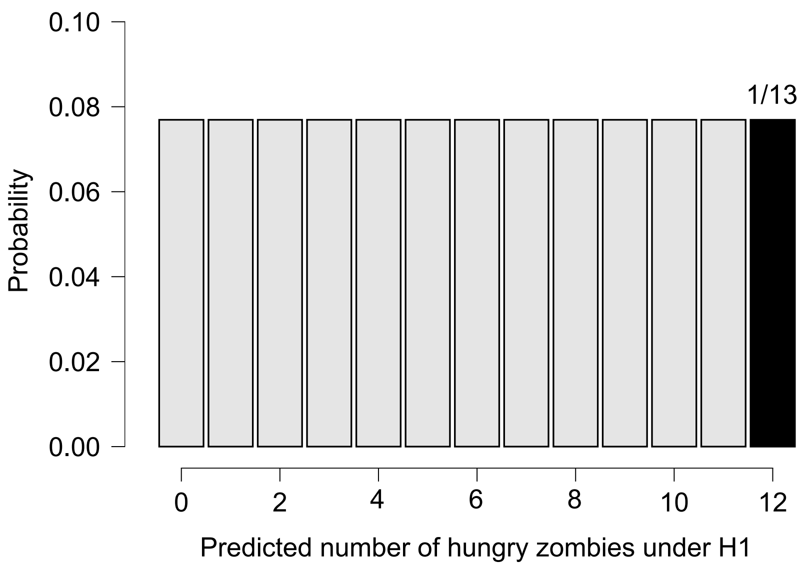


Figure 15.4: The Laplacean hypothesis $\mathcal{H}_1 : \theta = \text{beta}(1, 1)$, 'all values for the chance θ of observing a zombie who is hungry are equally likely' predicts that, for Kate's data set of 12 zombies, all possible numbers of hungry zombies are equally likely to occur. Figure from the JASP module *Learn Bayes*.

- The observed data are 13 times more likely under $\mathcal{H}_0$ than under $\mathcal{H}_1$.
- $\mathcal{H}_0$ predicted the observed data 13 times better than $\mathcal{H}_1$.
- The data have increased the odds in favor of $\mathcal{H}_0$ over $\mathcal{H}_1$ by a factor of 13.
- If the prior probabilities for the rival hypotheses are equal (i.e., $p(\mathcal{H}_0) = p(\mathcal{H}_1) = 1/2$) then the posterior probability for $\mathcal{H}_0$ equals $13/14 \approx .93$.

A common pitfall is to interpret the Bayes factors directly as a posterior odds: 'If the Bayes factor is $\text{BF}_{01} = x$, this means that $\mathcal{H}_0$ is x times more likely than $\mathcal{H}_1$' (cf. Chapter 3, section 'Example: The Inevitable Base Rate Fallacy'). As Equation 15.1 shows, such an interpretation is warranted only when the prior odds are 1, that is, when the prior probability for each of the two rival models equals $1/2$.[10]

It is worth emphasizing that the result, $\text{BF}_{01} = 13$, represents evidence in favor of the null hypothesis $\mathcal{H}_0$.[11] As demonstrated by the zombie example, this happens because $\mathcal{H}_0$ makes precise predictions that are then validated by the data; the forecasts of $\mathcal{H}_1$ are less impressive because it assigns equal probability to all possible outcomes.[12] The underlying principle, as with all of Bayesian inference, is that hypotheses that predict the data relatively well enjoy a boost in credibility,

[10] See also the blog post "The single most prevalent misinterpretation of Bayes' rule" on `BayesianSpectacles.org`.

[11] No other statistical approach that we are aware of is able to quantify evidence for a point-null hypothesis, at least not for a reasonable definition of evidence (i.e., something that ought to affect an opinion).

[12] Note that observing a single satiated zombie results in $\text{BF}_{01} = 0$ or $\text{BF}_{10} = \infty$, that is, infinite evidence against $\mathcal{H}_0$. Daring predictions are rewarded when they come true, but heavily punished where they fall flat.

whereas hypotheses that predict the data relatively poorly suffer a decline (Wagenmakers et al. 2016a).

The updated results may also be presented as a posterior spike-and-slab distribution, as shown in Figure 15.5. The posterior distribution under the slab has the same shape as the beta(13,1) posterior from Figure 15.2, but the area under the curve does not equal 1. Instead, the area equals $^1/_{14}$, the posterior probability for $\mathcal{H}_1$. The remaining posterior probability, $^{13}/_{14} \approx .93$, goes to $\mathcal{H}_0$ and is represented in Figure 15.2 by the height of the posterior spike at $\theta = 1$.

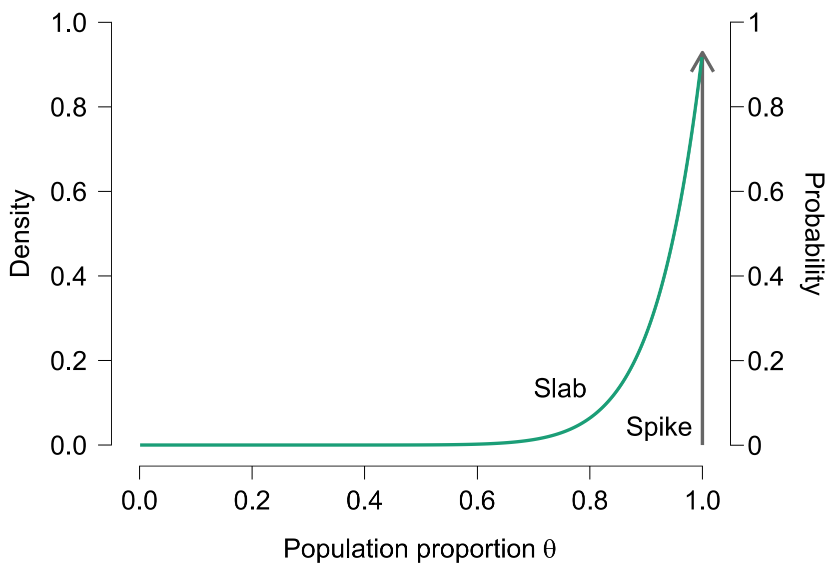


Figure 15.5: The Wrinch-Jeffreys 'spike-and-slab' posterior distribution after having observed 12 hungry zombies. The spike at $\theta = 1$ has height $^{13}/_{14} \approx .93$ (second y-axis), which is the posterior probability for $\mathcal{H}_0$. The area under the posterior slab equals $^1/_{14} \approx .07$, the posterior probability for $\mathcal{H}_1$. Figure from the JASP module *Learn Bayes*.

General Solution

At the end of the day, the inclusion of the spike at $\theta = 1$ has allowed Kate to answer her original question and quantify the evidence that the observed data provide for the universal generalization that all zombies are hungry. Specifically, after comparing the predictive performance of $\mathcal{H}_0 : \theta = 1$ versus that of $\mathcal{H}_1 : \theta \sim \text{beta}(1, 1)$ Kate concludes that the occurrence of 12 hungry zombies is 13 times more likely under $\mathcal{H}_0$ than it is under $\mathcal{H}_1$. Assuming $\mathcal{H}_0$ and $\mathcal{H}_1$ to be equally likely *a priori*, this means the posterior probability for $\mathcal{H}_0$ equals $^{13}/_{14} \approx .93$.

Kate's result for 12 zombies can be easily generalized to an observed unbroken hungry zombie sequence of any length. Figure 15.4 shows that a uniform prior on θ induces a uniform prior on the predicted number of hungry zombies. Hence, under $\mathcal{H}_1 : \theta \sim \text{beta}(1, 1)$ the probability that all n zombies are hungry equals $1/(n+1)$. Under $\mathcal{H}_0 : \theta = 1$, the probability of an unbroken sequence of hungry zombies equals 1, for any length n. Consequently, the Bayes factor in favor of $\mathcal{H}_0$ over $\mathcal{H}_1$ equals $\text{BF}_{01} = n + 1$. Under equal prior model probabilities, the posterior probability for $\mathcal{H}_0$ equals $(n+1)/(n+2)$ (Jeffreys 1973, p. 55).[13] Thus, every confirmatory instance offers support for the general law; specifically, it increases the Bayes factor by 1. "This is in accordance with the principle that a high probability can be attached to a general law by a moderate amount of evidence." (Jeffreys 1973, p. 55).

[13] This equation should look eerily familiar. The next subsection goes into detail.

To drive home the contrast to the Laplacean analysis using the Principle of Indifference (cf. Figure 15.2), assume that, from an infinite zombie population, 100,000 participants are sampled, all of whom indicate to be hungry. Based on these data, what is the probability that all zombies are hungry? The Laplacean answer is that this probability is *zero*. On the other hand, the Wrinch-Jeffreys answer is that this probability is $100001/100002 = 0.99999$.

Two Sequential Analyses

As we have already seen many times throughout this book, it does not matter whether the data are analyzed simultaneously or sequentially: the end result is identical. We now explore two ways in which the data from Kate may be analyzed sequentially: one zombie at a time, or in two batches of six zombies each.

First, assume that $\mathcal{H}_1 : \theta \sim \text{beta}(\alpha, \beta)$, and we desire the probability that the very next zombie is hungry. By the beta prediction rule (Chapter 9) this equals $\alpha/(\alpha+\beta)$. For a single hungry zombie, the Bayes factor in favor of $\mathcal{H}_0$ therefore equals

$$\text{BF}_{01}(s = 1) = \frac{1}{\alpha/(\alpha+\beta)} = \frac{\alpha + \beta}{\alpha}.$$

For $\alpha = \beta = 1$, this yields $\text{BF}_{01}(s = 1) = 2$, confirming the $n + 1$ rule outlined above.

The probability that the second zombie is hungry, given that the first zombie is hungry, is $(\alpha+1)/(\alpha+1+\beta)$, and the corresponding Bayes factor equals $(\alpha+1+\beta)/(\alpha+1)$. For $\alpha = \beta = 1$, this yields $3/2$; multiplying these two probabilities yields $2/1 \times 3/2 = 3$, again confirming the $n + 1$ rule.

When we go through the entire sequence of 12 hungry zombies this way, we obtain:

$$\text{BF}_{01}(s = 12) = \frac{2}{1} \cdot \frac{3}{2} \cdot \frac{4}{3} \cdot \frac{5}{4} \cdot \frac{6}{5} \cdot \frac{7}{6} \cdot \frac{8}{7} \cdot \frac{9}{8} \cdot \frac{10}{9} \cdot \frac{11}{10} \cdot \frac{12}{11} \cdot \frac{13}{12}.$$

As the numerator of the nth term equals the denominator of the $n+1$th term, this series telescopes and the end result is 13, again confirming the $n + 1$ rule.[14]

[14] This sequential analysis provides another way to prove the $n + 1$ rule.

Second, assume that $\mathcal{H}_1 : \theta \sim \text{beta}(1,1)$ and analyze the 12 zombies in two successive bathes of size six. We know that the first batch gives $\text{BF}_{01} = 7$, as dictated by the $n + 1$ rule. What is the Bayes factor for the second batch, given that we have already observed the first batch? To answer this question easily we can use the law of conditional probability to combine evidence (cf. the section 'Combining the Evidence' in Chapter 13). That is, we know that the overall Bayes factor for all 12 zombies equals the Bayes factor for the first batch, multiplied by the Bayes factor for the second batch (after having properly updated the parameter priors based on the information from the first batch), that is, $\text{BF}_{01}(s_1 + s_2 = 12) = \text{BF}_{01}(s_1 = 6) \times \text{BF}_{01}(s_2 = 6 \mid s_1 = 6)$. We know that $\text{BF}_{01}(s_1 + s_2 = 12) = 13$ and that $\text{BF}_{01}(s_1 = 6) = 7$, which means that $\text{BF}_{01}(s_2 = 6 \mid s_1 = 6) = {}^{13}/_{7} \approx 1.86$. More generally, for the first batch, $\text{BF}_{01}(s_1) = s_1 + 1$, and for the total data set $\text{BF}_{01}(s_1 + s_2) = s_1 + s_2 + 1$; consequently, the Bayes factor for the second batch, given the first, equals $\text{BF}_{01}(s_2 \mid s_1) = {}^{(s_1+s_2+1)}/_{(s_1+1)}$.[15]

[15] See Chapter 27 for a more extensive discussion on this topic.

This result can also be obtained by applying Laplace's Rule of Succession for Series (cf. Chapter 9): the probability of an unbroken sequence of k successes, given that an unbroken sequence of s successes has already been observed, equals ${}^{(s+1)}/_{(s+k+1)}$. Because the probability of the data equals 1 under $\mathcal{H}_0 : \theta = 1$, the Bayes factor is $\text{BF}_{01} = {}^{(s+k+1)}/_{(s+1)}$, confirming the result obtained by applying the law of conditional probability.

A Curious Coincidence

At this point, the attentive reader may have noticed something peculiar. When we were discussing the Laplacean 'slab-only' analysis of Kate's zombie data (cf. Figure 15.2), we mentioned that according to the Rule of Succession, the probability that the next zombie is hungry equals ${}^{(n+1)}/_{(n+2)} = {}^{13}/_{14} \approx .93$. A little later, we applied the 'spike-and-slab' Wrinch-Jeffreys approach and concluded that, when $p(\mathcal{H}_0) = p(\mathcal{H}_1) = {}^{1}/_{2}$, the posterior probability for the general law equals ${}^{(n+1)}/_{(n+2)} = {}^{13}/_{14} \approx .93$. This is the probability that all zombies from an infinite zombie population are hungry. The key probability from the 'spike-and-slab' Wrinch-Jeffreys approach equals exactly the key probability from the 'slab-only' Laplace approach, even though these probabilities are based on different assumptions and address a very different question.

Intuition may suggest that this correspondence is maintained for any $\text{beta}(\alpha, \beta)$ prior on θ under $\mathcal{H}_1$, but this is not true. Miraculously, if $p(\mathcal{H}_0) = p(\mathcal{H}_1) = {}^{1}/_{2}$ the correspondence holds only when $\alpha = \beta = 1$, the most popular default prior specification. To realize that the identity breaks down for values of α and β other than 1, consider $\mathcal{H}_1 : \theta \sim$

beta(α, α), a prior distribution symmetric around $\theta = 1/2$. Assume we observe a single success.

First we consider the setup where a general law (here $\theta = 1$) is assigned separate prior mass, and we answer the question "what is the probability that all future observations will be successes?". Under $\mathcal{H}_1$, the symmetric beta(α, α) prior does not encode a preference for successes or failures, and hence the prior predictive probability that the first trial is a success equals $1/2$. This also follows from the beta prediction rule (cf. Chapter 9): $p(s = 1 \mid \theta) = \alpha/(\alpha+\alpha) = 1/2$. Under the general law $\mathcal{H}_0 : \theta = 1$ the probability that the first trial is a success equals 1. Consequently, $\text{BF}_{01} = 2$ for any value of α that defines the symmetric prior beta(α, α) distribution under $\mathcal{H}_1$. Assuming both hypotheses to be equally likely *a priori* (i.e., $p(\mathcal{H}_0) = p(\mathcal{H}_1) = 1/2$), the posterior probability for $\mathcal{H}_0$, that is, the posterior probability that *all* future trials will be successes, equals $2/3$.

Next we consider the setup where the general law (here $\theta = 1$) is *not* assigned separate prior mass, and we answer the question "what is the probability that the next observation will also be a success?" The observation of a single success updates the beta(α, α) prior distribution to a beta$(\alpha + 1, \alpha)$ posterior distribution. The beta prediction rule then gives the probability that the next trial is also a success as $(\alpha+1)/(2\alpha+1)$. This equals $2/3$, the probability that *all* future trials will be successes, only when $\alpha = 1$. A more in-depth discussion on the differences between the Laplacean answer and the one by Wrinch and Jeffreys is presented in the appendix to this chapter.

Exercises

1. Let θ denote the chance that any one zombie is hungry. You entertain two hypotheses, $\mathcal{H}_0 : \theta = 1$ (i.e., all zombies are hungry), and $\mathcal{H}_1 : \theta \sim \text{beta}(1, 1)$ (i.e., every value for the chance θ is equally likely *a priori*). Let $p(\mathcal{H}_0) = p(\mathcal{H}_1) = 1/2$, that is, both hypotheses are equally likely *a priori*. You observe four zombies, and all of them are hungry. What is the probability that the fifth one will be hungry too?

2. Figure 15.4 shows that under $\mathcal{H}_1 : \theta \sim \text{beta}(1, 1)$, all possible number of hungry zombies are equally likely. Prove this mathematically (hint: simplify the expression for the beta-binomial distribution).

3. Explore the robustness of Kate's Bayes factor by examining the results for several alternative prior beta distributions for θ under $\mathcal{H}_1$. Explain why and how the shape of the prior beta distribution influences the Bayes factor.

4. Repeat the previous exercise but increase the number of hungry zombies. Do the data overwhelm the prior? Why or why not?

Recalling the Trio of Priors

Because it is so important, we reiterate the distinction between the three main uses for the word 'prior' in Bayesian inference outlined earlier in Chapter 12. First, *prior model probabilities* indicate the relative plausibility for each member of a set of discrete models before observing the data. For instance, in the zombie example we assumed that $p(\mathcal{H}_0) = p(\mathcal{H}_1) = 1/2$. Second, *prior parameter distributions* indicate the relative plausibility of a set of parameter values before observing the data. Usually the set of parameter values is continuous. For instance, in the zombie example we assumed that under $\mathcal{H}_1$, the chance θ was assigned a uniform prior distribution, $\theta \sim$ beta$(1, 1)$. When the parameter can only take on a finite set of discrete values, the difference between prior model probabilities and prior parameter distributions becomes blurred (e.g., Gronau and Wagenmakers 2019). Third, *prior predictive distributions* refer to the predictions for to-be-observed data that are generated from a model as defined by its likelihood and its prior parameter distributions. For instance, in the zombie example the uniform prior distribution on θ induced a uniform prior predictive distribution for the number of hungry zombies (cf. Figure 15.4). Relatedly, the word 'prior' also occurs in the term *prior predictive likelihood*, which refers to the mass that the prior predictive distribution assigns to the data that actually occurred. For instance, in the zombie example the prior predictive under $\mathcal{H}_1$ is indicated by the highlighted bar in Figure 15.4.

5. Return to the example of the 10 possible bakers discussed in the introduction of Chapter 8. Can you translate the slab-only approach and the spike-and-slab approach to the discrete case? What insights does this bring?

6. To solidify your understanding, dissect and summarize the fragment below in your own words:

   > "Philosophers often argue that induction has so often failed in the past that Laplace's estimate of the probability of a general law is too high, whereas the main point of the present work is that scientific progress demands that it is far too low. Philosophers, for instance, appeal to exceptions found to such laws as 'all swans are white' and 'all crows are black'. Now if Laplace's rule is adopted and we have a pure sample of m members, there is a probability $\frac{1}{2}$ that the next $m + 1$ will have the property. If this is applied to many different inductions, these probabilities should be nearly independent as any we know of, and Bernoulli's theorem should hold; therefore in about half of the cases where an induction has been based on a pure sample, an exception should have been found when the size of the sample was slightly more than doubled. This seems to be glaringly false. The original propounder of 'all swans are white' presumably based it on a sample of hundreds or thousands; but the verifications before the Australian black swan was discovered must have run into millions. According to the modification (...) the number of the fresh sample before the probability that it contains no exception sinks to $\frac{1}{2}$ is of order m^2, and this is much more in accordance with experience." (Jeffreys 1961, p. 132)

Chapter Summary

In order for data to be able to support a universal generalization, the associated general law needs to be assigned its own prior probability. By doing so, the Laplacean framework of parameter estimation –which reflects a denial without evidence that any general law could be true– is transformed to a framework of model comparison or hypothesis testing, where the null hypothesis $\mathcal{H}_0$ represents the general law that fixes a key parameter to a specific value of interest, and the alternative hypothesis $\mathcal{H}_1$ relaxes the restriction and allows the key parameter to take on other values. The fact that $\mathcal{H}_0$ is assigned definite prior mass accords with the principle of parsimony, which is the topic of Chapter18.

Want to Know More?

✓ A comprehensive summary of the academic work of Harold Jeffreys is available online at `http://www.economics.soton.ac.uk/staff/aldrich/jeffreysweb.htm`, courtesy of John Aldrich. "Jeffreys was a noted physical scientist who re-established the statistical theory of his time on Bayesian foundations. This page is a guide to literature and websites which may be useful to anyone interested in Jeffreys's statistical work and its background. The emphasis is on Jeffreys's own writings and on the older literature."

✓ Aldrich, J. (2005). The statistical education of Harold Jeffreys. *International Statistical Review, 73*, 289-307.

✓ Etz, A., & Wagenmakers, E.-J. (2017). J. B. S. Haldane's contribution to the Bayes factor hypothesis test. *Statistical Science, 32*, 313-329.

✓ Howie, D. (2002). *Interpreting Probability: Controversies and Developments in the Early Twentieth Century*. Cambridge: Cambridge University Press. An in-depth overview of the debate between the Bayesian Harold Jeffreys and the frequentist Ronald Fisher. Some background knowledge of statistics is required to understand the finer details. Fragment, related to Figure 15.6: "The collaboration with Wrinch was uncharacteristic: Jeffreys was reserved by nature, and awkward in company, and had chosen research fields and methods that allowed him to work almost entirely alone – typically with his typewriter on his knees, his hand-cranked Marchant calculating machine on the floor in front, and the room ankle-deep in research papers and works-in-progress." (Howie 2002, p. 110)

✓ Jeffreys, H. (1936). The problem of inference. *Mind, 45*, 324–333.

✓ Miyake, T. (2017). Scientific Inference and the Earth's Interior: Dorothy Wrinch and Harold Jeffreys at Cambridge. In Stadler, F. (Ed.), *Integrated History and Philosophy of Science*, Vol. 20, pp. 81-91. Cambridge: Springer.

✓ Senechal, M. (2012). *I Died for Beauty: Dorothy Wrinch and the Cultures of Science*. New York: Oxford University Press.

✓ Smith?, R. (2014). *Mathematical Modelling of Zombies*. Canada: University of Ottawa Press. Convinced that many –if not all– zombies have a ravenous appetite? Worried that an apocalypse will quickly reduce you to zombie döner kebab? This book might help you survive! The question mark that follows the author's name is not a typo.

✓ van den Bergh, D., Haaf, J. M., Ly, A., Rouder, J. N., & Wagenmakers, E.-J. (2021). A cautionary note on estimating effect size. *Advances in Methods and Practices in Psychological Science, 4*, 1–8. Advocates the spike-and-slab model for estimating effect size.

Appendix: A Dialogue on the Curious Coincidence

This appendix continues the discussion from the subsection "A Curious Coincidence" and focuses on the question whether or not the Laplacean Rule of Succession is fundamentally different from the Wrinch-Jeffreys 'Rule of Pure Induction'.

EJ: "Dora, another way to see that the Laplacean answer differs from the one by Wrinch and Jeffreys is to consider the relative importance of the beta prior distribution and the data. Consider the setup where $\mathcal{H}_1 : \theta \sim \text{beta}(\alpha, 1)$. The Rule of Succession states that the probability that the next trial is a success, based on a previous unbroken string of s successes, equals ${}^{(\alpha+s)}/_{(\alpha+s+1)}$. This shows that there is a perfect trade-off relationship between α and s: all that matters in the Laplacean formulation is $\alpha + s$. For the posterior distribution it does not matter whether, say, $\alpha = 1$ and $s = 100$, or $\alpha = 100$ and $s = 1$. From a posterior point of view, the data have been combined with the information in the prior; this updating process occurred in the past and, as far as the prediction for the next observation is concerned, it is no longer relevant.

This is arguably different from the approach where we wish to assess the posterior probability in favor of the general law $\mathcal{H}_0 : \theta = 1$ based on the previous observation of s successes. Assuming that $\mathcal{H}_0$ and $\mathcal{H}_1$ are equally likely *a priori*, this posterior probability is identical to the Bayes factor – the degree to which $\mathcal{H}_0$ outpredicted $\mathcal{H}_1$ for the observed data s. In order to evaluate the relative predictive adequacy of $\mathcal{H}_0$ versus $\mathcal{H}_1$, we need to consider the prior distribution under $\mathcal{H}_1$.

For instance, consider the scenario where $\alpha = 1$ and $s = 100$. The means that the alternative hypothesis hedges its bets; it states that "all values of θ are equally likely a priori", which means that in the prior predictive distribution, all numbers of successes from 0 to 100 are equally likely. In contrast, $\mathcal{H}_0$ puts all its predictive mass on $s = 100$ – it makes the precise and highly falsifiable prediction that all trials will be successes. The precise prediction comes true and, with a substantial number of $s = 100$ confirmatory instances, $\text{BF}_{01} = s + 1 = 101$, with a corresponding posterior probability of ${}^{101}/_{102} \approx .99$. In the alternative scenario we have $\alpha = 100$ and $s = 1$. The situation here is dramatically different. The alternative hypothesis now states that "high values of θ are much more plausible than low values of θ". The posterior mean is $\theta = .99$, and the 95% HPD interval ranges from .97 to 1. In other words, the alternative hypothesis predicts that a very high proportion of future trials will be successes. This prediction is relatively similar to that of $\mathcal{H}_0$, which holds that *all* future trials will be successes. For discriminating such similar predictions we need a lot of data. But, to make matters worse, we do not have a lot of data – we have only a single confirmatory observation, $s = 1$. The combination of these two unfortunate factors (i.e., similar model predictions and sparse data) means that the Bayes factor will be close to 1. Specifically, $\mathcal{H}_1$ assigns the observed data $s = 1$ a prior predictive probability of .99 (i.e., $^{100}/_{101}$), and $\mathcal{H}_0$ assigns the observed data $s = 1$ a prior predictive probability of 1. This results in a Bayes factor $\text{BF}_{01} = {}^{101}/_{100} \approx 1.01$, a smidgen of evidence for $\mathcal{H}_0$, which results in a posterior probability of $^{1.01}/_{2.01} \approx .502$ that all future trials will be successes. To underscore the difficulty of discriminating among hypotheses that make highly similar predictions, we may entertain the possibility of observing a larger number of $s = 100$ confirmatory instances. This provides more evidence in favor of $\mathcal{H}_0 : \theta = 1$ over $\mathcal{H}_1 : \theta \sim \text{beta}(100, 1)$, but at $\text{BF}_{01} = 2$, the degree of support is weak at best.

In sum, the question "Given an unbroken string of successes observed in the past, what is the probability that the next trial will also be a success, given that no special attention is given to any particular value of θ?" is radically different from the question "Given an unbroken string of successes observed in the past, what is the probability that all future trials will also be successes, given that we deem it plausible, *a priori*, that a general law (e.g., $\theta = 1$) is true?" For the former question, the answer depends only on the shape of the posterior distribution, and the degree to which it is determined by prior knowledge or observed data is irrelevant. For the latter question, the answer depends on predictive performance for the past data, and to assess this predictive performance we need to separate what is used to make the prediction (i.e., the prior distribution) from what is predicted (i.e., the data). It cannot come as a surprise, therefore, that such different questions generally yield highly different answers – what is surprising is the fact that they yield the same answer for the most common scenario (i.e., $\alpha = \beta = 1$, $p(\mathcal{H}_0) = p(\mathcal{H}_1) = {}^1/_2$): a curious mathematical coincidence."

Dora: Thanks for *mansplaining* this to me in so much detail, EJ. However, I believe you may be mistaken when you argue that the Wrinch-Jeffreys setup depends on predictive performance whereas the Laplacean setup does not. This reminds me of the common critique that the prior distribution under $\mathcal{H}_1$ affects the Bayes factor much more than it affects the posterior distribution. Let me offer the following observations:

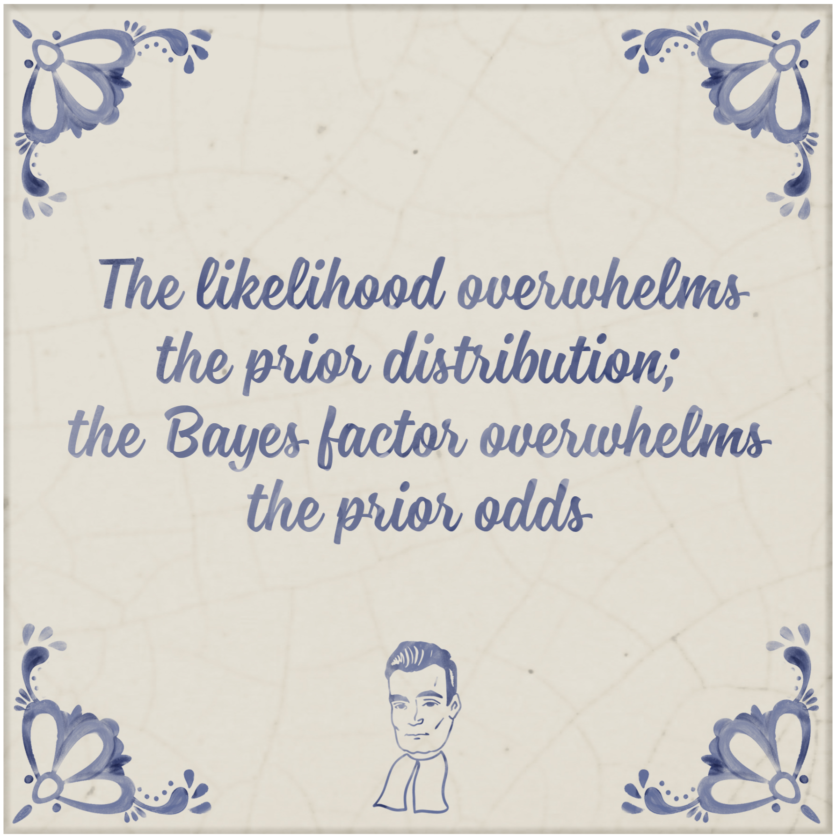
- Consider the spike-and-slab representation from Figures 15.1 and 15.5. As always in Bayesian learning, values of θ that predicted the observed data better than average have *gained* plausibility, whereas values of θ that predicted worse than average have *lost* plausibility. This predictive updating principle holds irrespective of whether or not the distribution consists (a) only of spikes (as in the pancake examples from Chapters 7 and 8), (b) of a mixture of spike and a slab, or (c) only of a slab.

- We need to discriminate sharply between evidence and posterior belief. Evidence is the extent to which the data change our opinion: therefore it represents the difference between prior and posterior conviction. Hence, it is natural, desirable, and inevitable that evidence depends on our prior beliefs. At the same time, however, the accumulation of evidence will gradually come to dominate our prior beliefs, in the sense that divergent prior beliefs will converge to highly similar posterior beliefs: "the data overwhelm the prior" (e.g., Wrinch and Jeffreys 1919).

- The data overwhelm the prior regardless of whether the prior distribution includes spikes. Specifically, for spikes one may state that "The Bayes factor overwhelms the prior odds".

- You mention that, when it comes to determining the shape of the posterior distribution under $\mathcal{H}_1$, all that matters is $\alpha + s$, whereas for the evidence it is important to treat these separately. As mentioned above, however, evidence and posterior beliefs are different concepts – it is only for the quantification of *evidence*, not posterior belief, that it is important to treat α and s separately. Also, the shape of the spike-and-slab prior includes the height of the spike (i.e., $p(\mathcal{H}_0)$) and the area of the slab (i.e., $p(\mathcal{H}_1) = 1 - p(\mathcal{H}_0)$). The posterior height of the spike in the spike-and-slab model is based on a combination of the prior height and the evidence from the data; for the spike-and-slab posterior it is irrelevant whether the spike is high because it had relatively large prior probability or relatively large support from the data, just as it is irrelevant for the shape of the slab whether α is high and s is low or vice versa.

For concreteness, consider the task of discriminating between a bent coin with unknown chance θ (i.e., $p(\mathcal{H}_1 : \theta \sim \text{beta}(1,1))$ and a magician's coin (i.e., a coin constructed to be double-heads or double-tails, with the two options equally likely: $\mathcal{H}_0 : p(\theta = 0) = p(\theta = 1) = {}^1\!/_2$). Suppose the coin is tossed n times, and all tosses land heads. The Bayes factor BF_{01} equals $\frac{1}{2}(n+1)$: as in the zombie example, the probability of the data under $\mathcal{H}_1$ equals ${}^1\!/_{(n+1)}$, but, unlike the zombie example, the probability of the data under $\mathcal{H}_0$ equals ${}^1\!/_2$ – this is the probability for the very first toss, after which the 'magician's coin' is updated and uniquely identified as 'double-heads', with probability 1 for the remaining sequence of tosses. Thus, adding the option of 'double-tails' (i.e., $\theta = 0$) in the magician's coin hypothesis halves the Bayes factor, even though that option can be discarded after the very first toss.



Data and evidence cause initially divergent opinions to converge. As a loose physical analogy, consider two metal balls positioned on a smooth table. At time zero, the balls may occupy a very different position. When a sufficiently strong magnet is placed anywhere on the table, however, the magnetic pull draws the balls to the same location. Here the initial position represents the prior opinion, the magnetic pull represents the information coming from the data, and the position of the magnet represent the point of posterior convergence. The data overwhelm the prior, but at the same time it is true that for each ball the distance travelled (i.e., the evidence) depends on its initial position relative to the position of the magnet. Figure available at `BayesianSpectacles.org` under a CC-BY license.

This example shows that the height of the spike matters – stipulating a second spike at $\theta = 0$ halved the Bayes factor. When the impact of the prior distribution on hypothesis testing is concerned, it may therefore be reasonable to employ a spike-and-slab representation and discuss the impact of the prior distribution under $\mathcal{H}_1$ as well as the impact of the prior probability for $\mathcal{H}_0$.

Figure 15.6: Sir Harold Jeffreys (1891-1989) with laptop typewriter in New Court, St John's College, Cambridge, 1928. (Photographer unknown, included by permission of the Master and Fellows of St John's College, Cambridge). See also Swirles (1992). The top right frame shows the sculpture 'Hercules and Lichas' by Antonio Canova (1795). In the frame to the left of the door, the man with the hat is probably the Austrian geologist Edward Suess; the man in the leftmost frame could be the Scottish geologist Charles Lyell (both suggested to us by Benjamin Deonovic).

16 Haldane's Rule of Succession [with Sandy Zabell and Quentin Gronau]

The content of this chapter is based on Wagenmakers et al. (in press).

The essential point is that when we consider a general law we are supposing that it may possibly be true, and we express this by concentrating a positive (non-zero) fraction of the initial probability in it. Before my work on significance tests, the point had been made by J. B. S. Haldane (1932).

Jeffreys, 1977

J. B. S. Haldane (right; 1892-1964) in the Black Watch. "At the beginning of the War our Idol [Haldane – EWDM] received a commission in the 3rd Battalion of the Black Watch, served in France and in Mesopotamia with the 1st and 2nd Battalions of that Regiment, and was twice wounded. Whilst he was in France, he was one of the first persons on whom they experimented with Chlorine Gas in the funny crude old gas-mask devices, a piece of unshowy and cold-blooded gallantry which commands everyone's admiration." (from the Oxford student magazine *Isis*, as reported on `https://skipperswar.com/tag/jbs-haldane/`). Photo taken circa 1915, public domain.

Chapter Goal

This chapter highlights the forgotten work on Bayesian inference by the famous geneticist and polymath J. B. S. Haldane. In 1932, Haldane was the first to calculate a Bayes factor hypothesis test; subsequently, Haldane also derived the probability that an unbroken string of $s = n$ successes will be followed by another success. Recall that Laplace's Rule of Succession states that this probability is ${}^{s+1}/_{s+2}$; in Haldane's setup, where the general law is given a prior probability of ${}^{1}/_{2}$, this probability instead equals $[{}^{s+1}/_{s+2}] \times [{}^{s+3}/_{s+2}]$. This elegant adjustment of the Laplacean analysis we term *Haldane's Rule of Succession*.

A Muddled Narrative

Up to this point in the book, the narrative may seem relatively simple and straightforward. Let's take a moment to recapitulate. Chapter 9 introduced Laplace's Rule of Succession: when all $s = n$ instances observed so far are of a particular type, the probability that the next instance will also be of that type equals ${}^{s+1}/_{s+2}$. For instance, having observed that each member of a group of 12 zombies is hungry, the Laplacean probability that the next, 13$^{\text{th}}$ zombie will also be hungry equals ${}^{13}/_{14} \approx .93$. As s grows large, this probability approaches 1, which seems perfectly reasonable. However, Laplace's Rule of Succession implicitly assumes that the general law is false. The Laplacean prediction follows from assigning the latent proportion θ a continuous uniform distribution from 0 to 1, that is, $\mathcal{H}_1 : \theta \sim \text{beta}(1, 1)$, and this does not acknowledge that $\theta = 1$ (i.e., the value stipulated by the general law) is worthy of special attention. It becomes clear that this is problematic when we use the Laplacean setup to derive the probability that the next k instances will all be confirmatory:

$$p(k \mid s = n) = \frac{s+1}{s+k+1}. \tag{16.1}$$

As k increases, this probability goes to zero. In other words, no matter how long the unbroken series $s = n$ of hungry zombies you already encountered may be, you should remain fully certain that an exception is bound to occur sooner or later. This accords neither with intuition nor experience. The impression that something is amiss is reinforced by considering the scenario where $k = s + 1$. For instance, suppose that, after observing a group of $s = 12$ hungry zombies, you urgently wish to know whether each member of an approaching group of $k = 13$ zombies is hungry. An application of Equation 16.1 yields a probability of only 1/2. In general, when the observed number of confirmatory instances equals s, and the predicted sequence is $k = s + 1$ long, the probability that all k members are of the same type is 1/2. Therefore Laplace's Rule of Succession expresses a "violent prejudice" against the general law that all instances are of a particular type:

> "This shows that the analysis of sampling procedure given so far is quite inadequate to account for the high probability that we often attach to a general law." (Jeffreys 1973, p. 53; see also Jeffreys 1961, pp. 127-128).

Many scholars –arguably including Laplace himself!– recognized early on that Laplace's rule did not apply to the scenario where background knowledge suggests the general law could be true (e.g., Zabell 1989, and references therein). In order to explain the glaring discrepancy between Laplace's rule and common sense, Wrinch and Jeffreys suggested the general law needed to be taken seriously and perhaps be given separate prior mass (e.g., Wrinch and Jeffreys 1921; Jeffreys 1931, pp. 29-31). So far so good.

At this stage, however, the narrative becomes decidedly muddled. It is tempting to conclude that, after suggesting in the early 1920s that a general law deserves separate prior mass, Wrinch and Jeffreys followed up with a concrete analysis such as the one outlined in the previous chapter. This is the standard interpretation, and –in the interest of simplicity– it is also the interpretation that we have adopted throughout this book. This interpretation is even more tempting because such a concrete analysis would later form a cornerstone of his work in statistics (e.g., Jeffreys 1939). However, this conclusion appears to be incorrect, or at least incomplete. Although Wrinch and Jeffreys provided the conceptual basis for a concrete analysis, they never actually carried it out. And when Jeffreys did carry out the analysis, well over a decade after his work with Wrinch, he had already been scooped – by the famous geneticist John Burdon Sanderson 'JBS' or 'Jack' Haldane (1892-1964).

Haldane's Remarkable Anticipation of Harold Jeffreys

"If I am not forgotten completely a hundred years hence, I shouldn't wonder if I should be remembered for something which I have not mentioned today. It might be something like, let us say, a letter to *The Observatory* entitled, 'Is space-time simply connected?' I am not going to try to explain to you what that means. It is a rather abstract geometrical idea. It might be the clue to new approaches to cosmology, though I should think it is more than twenty to one that it will not be: it might be—but, still more likely, it will be something which I have completely forgotten now. Some little remark I made in some paper which perhaps someone will dig out and say: 'Oh, but that explains what I found last year'. Or perhaps some historian will find out and say: 'Haldane's remarkable anticipation of Chew Wong', or something like that. We do now know. But to take an example, the estimation of human mutation rates was, so to speak, a footnote to what then seemed to me more important.

But I don't really very much care what people think about me, especially a hundred years hence. I should not like them to be too critical of me as long as my widow and a few friends survive me. But the greatest compliment made to me today, I believe, is when people refer to something which I discovered (...) without mentioning me at all. To have got into the tradition of science in that way is to me more pleasing than to be specially mentioned. But what matters, in my opinion, is what I have done, good or evil, and not what people think of me." (Haldane, 1964, self-obituary; taken from Tredoux 2018, pp. 310-311)

Haldane's Forgotten Rule

In 1932, J. B. S. Haldane published a remarkable seven-page article – titled *A note on inverse probability*– that contained two main advances. First, Haldane computed the Bayes factor as the ratio of two marginal likelihoods, one for a point-null hypothesis $\mathcal{H}_0$ that assigns θ a single value, and one for an alternative hypothesis $\mathcal{H}_1$ in which θ is assigned a prior distribution (Etz and Wagenmakers 2017). Second, Haldane then used this Bayes factor to obtain an alternative to Laplace's Rule of Succession. We can never know for certain, but it is entirely possible that these two advances jolted Jeffreys into action, and motivated him to pursue a similar agenda throughout the 1930s. More remarkable than Haldane's article is perhaps the fact that it has been almost entirely forgotten. The mystery deepens when one realizes that both Haldane and Jeffreys were among the foremost researchers of their day, and knew one another well (Etz and Wagenmakers 2017). Haldane subsequently abandoned this line of work, and Jeffreys mentioned Haldane only occasionally, and in passing.

At any rate, let's now turn to Haldane's line of reasoning. As in the previous chapter, Haldane assumed the presence of two hypotheses: the Laplacean hypothesis $\mathcal{H}_1 : \theta \sim \text{beta}(1,1)$ and the general law $\mathcal{H}_0 : \theta = 1$.[1] For simplicity we assign both $\mathcal{H}_1$ and the general law $\mathcal{H}_0$ equal probability, such that the prior odds is 1 and the posterior odds equals the Bayes factor. Assume we observe $s = n$ confirmatory instances. Then Chapter 9 tells us that the Bayes factor BF_{01} is $s + 1$, and the corresponding posterior probability for $\mathcal{H}_0$ is ${}^{s+1}/_{s+2}$ and for $\mathcal{H}_1$ is the complement ${}^{1}/_{s+2}$. Now suppose we wish to determine the probability that the next observation also confirms the general law. Under $\mathcal{H}_0$, this probability is 1 (i.e., $\mathcal{H}_0 : \theta = 1$ can only predict the occurrence of confirmatory instances); under $\mathcal{H}_1$, this probability is ${}^{s+1}/_{s+2}$ (i.e., the Laplacean answer).

[1] Haldane actually considered the case of $\mathcal{H}_0 : \theta = 0$, but this yields the same results when we switch the data labels (i.e., "all zombies are hungry" is the same as "no zombies are non-hungry").

To obtain the desired predictive probability, we have to *average* out the hypothesis; in other words, we *model-average* using the law of total probability. This can be graphically represented by a tree diagram similar to Figures 3.6, 7.4, and 12.8). The posterior model probabilities act as averaging weights for the predictions from the respective hypotheses. Statistically, the probability that the next instance is confirmatory, $p(y = 1 \mid s = n)$, is given by

$$\begin{aligned}
p(y = 1 \mid s = n) &= \overbrace{\frac{s+1}{s+2}}^{p(\mathcal{H}_0 \mid s=n)} \times \overbrace{1}^{p(y=1 \mid \mathcal{H}_0, s=n)} + \overbrace{\frac{1}{s+2}}^{p(\mathcal{H}_1 \mid s=n)} \times \overbrace{\frac{s+1}{s+2}}^{p(y=1 \mid \mathcal{H}_1, s=n)} \\
&= \frac{s+1}{s+2} \times \frac{s+3}{s+2} \\
&= \left[1 - \frac{1}{s+2}\right] \times \left[1 + \frac{1}{s+2}\right] \\
&= 1 - \frac{1}{(s+2)^2},
\end{aligned} \tag{16.2}$$

an expression we term *Haldane's Rule of Succession*.[2] The first line of the equation is given in the book of answers for Exercise 1 in the previous chapter. The second line of the equation highlights that including the hypothesis that the general law is correct yields a particularly elegant result: Laplace's Rule of Succession, $(s+1)/(s+2)$, needs to be adjusted by a multiplicative factor of $(s + 3)/(s + 2)$. The third line of the equation

[2] This rule is also discussed explicitly in Tuyl (2019) and Tuyl et al. (in press).

shows that the Laplacean first factor and the Haldanean second factor are symmetric about 1. The fourth line of the equation underscores that as n grows, Haldane's Rule of Succession is associated with an increase in confidence that is more pronounced than it is for Laplace's Rule of Succession – a fact that becomes apparent when rewriting Laplace's Rule ${}^{n+1}/_{n+2}$ as $1 - \frac{1}{s+2}$. Thus, the probability of finding an exception

> "(...) is clearly of the order n^{-2}, rather than n^{-1}. This seems to be a more reasonable estimate of the validity of an induction than that generally given." (Haldane 1932, p. 59).

As an example, consider having observed that all of 12 zombies are hungry. What is the probability that the 13th will also be hungry? According to Haldane's Rule of Succession, this equals ${}^{13}/_{14} \times {}^{15}/_{14} = {}^{195}/_{196} \approx .99$, clearly higher than the Laplacean probability of $.93$.

Haldane's Rule of Succession for Series

The difference between Laplace's Rule of Succession and Haldane's Rule of Succession becomes more pronounced as the number of to-be-predicted instances increases. Consider the probability that an entire sequence of k new instances are all confirmatory. As indicated above, Laplace's Rule gives $p(k \mid s = n) = (s+1)/(s+k+1)$, which goes to zero as k grows large. Under Haldane's setup, in contrast, we obtain

$$
\begin{aligned}
p(k \mid s = n) &= \overbrace{\frac{s+1}{s+2}}^{p(\mathcal{H}_0 \mid s=n)} \times \overbrace{1}^{p(k \mid \mathcal{H}_0, s=n)} + \overbrace{\frac{1}{s+2}}^{p(\mathcal{H}_1 \mid s=n)} \times \overbrace{\frac{s+1}{s+k+1}}^{p(k \mid \mathcal{H}_1, s=n)} \\
&= \frac{s+1}{s+k+1} \times \frac{s+k+2}{s+2},
\end{aligned}
$$

where the second factor represents the Haldanean adjustment. As k grows large, this probability goes to ${}^{s+1}/_{s+2}$ – the posterior probability for the general law $\mathcal{H}_0$ after having observed $s = n$ confirmatory instances. Let's return to the scenario where you observe $s = 12$ hungry zombies, and you wish to know whether all of $k = 13$ incoming zombies are likewise hungry. We have already seen that the Laplacean probability equals ${}^{1}/_{2}$, in violation of common sense; in contrast, the Haldanean probability equals ${}^{27}/_{28} \approx 0.96$. In general, when $k = s + 1$ the Laplacean analysis gives $p(k \mid s = n) = {}^{1}/_{2}$ whereas the Haldanean analysis gives $p(k \mid n = 2) = {}^{1}/_{2} + {}^{1}/_{2} \cdot \frac{s+1}{s+2}$, an upward adjustment equal to half of the Laplacean probability that the single next observation is confirmatory.[3]

[3] Unfortunately the result as provided by Haldane (1932) is not completely correct. The mistake is obvious and most likely due to a typographical error (for details see Wagenmakers et al. in press).

J. B. S. Haldane was a precocious child. One anecdote has it that the four-year old Haldane, when inspecting the blood that trickled out of a cut on his forehead, asked "Is it oxyhaemoglobin or carboxyheamoglobin?" (Subramanian 2019, p. 45)

Exercises

1. You observe 20 hungry zombies. What is the probability that the next 2 zombies will also be hungry (a) according to the Laplace setup;

On Being the Right Size

"The most obvious differences between different animals are differences of size, but for some reason the zoologists have paid singularly little attention to them. In a large textbook of zoology before me I find no indication that the eagle is larger than the sparrow, or the hippopotamus bigger than the hare, though some grudging admissions are made in the case of the mouse and the whale. But yet it is easy to show that a hare could not be as large as a hippopotamus or a whale as small as a herring. For every type of animal there is a most convenient size, and a large change in size inevitably carries with it a change of form. (...)

To the mouse and any smaller animal it [gravity – EWDM] presents practically no dangers. You can drop a mouse down a thousand-yard mine shaft; and, on arriving at the bottom it gets a slight shock and walks away, provided that the ground is fairly soft. A rat is killed, a man is broken, a horse splashes. For the resistance presented to movement by the air is proportional to the surface of the moving object. Divide an animal's length, breadth, and height each by ten; its weight is reduced to a thousandth, but its surface only a hundredth. So the resistance to falling in the case of the small animal is relatively ten times greater than the driving force.

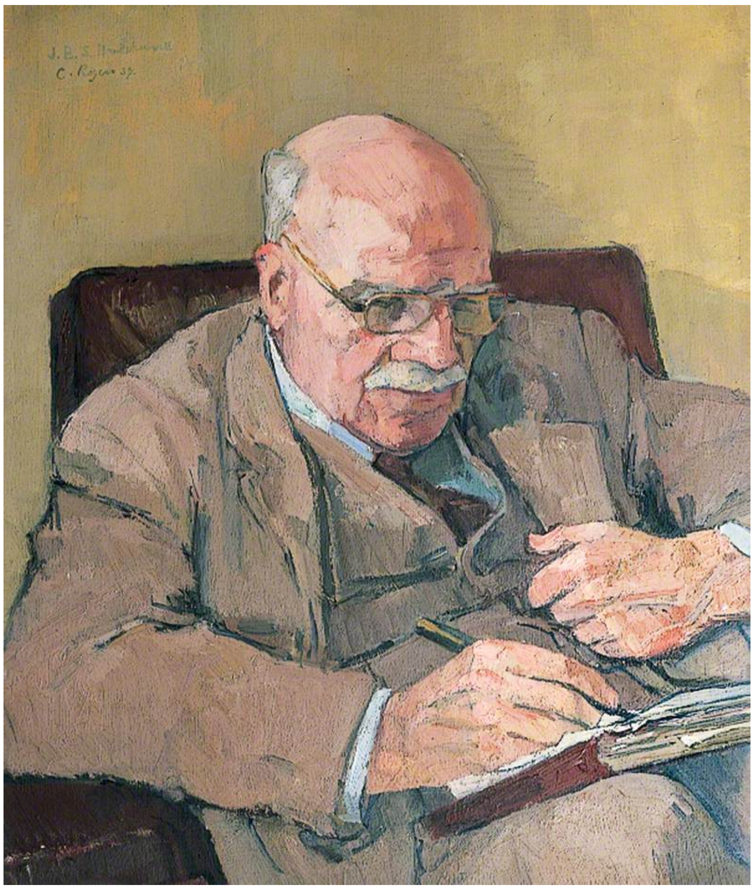
An insect, therefore, is not afraid of gravity; it can fall without danger, and can cling to the ceiling with remarkably little trouble. It can go in for elegant and fantastic forms of support like that of the daddy-longlegs. But there is a force which is as formidable to an insect as gravitation to a mammal. This is surface tension. A man coming out of a bath carries with him a film of water about one-fiftieth of an inch in thickness. This weighs roughly a pound. A wet mouse has to carry about its own weight of water. A wet fly has to lift many times its own weight and, as everyone knows, a fly once wetted by water or any other liquid is in a very serious position indeed. An insect going for a drink is in a great danger as man leaning out over a precipice in search of food. If it once falls into the grip of the surface tension of the water –that is to say, gets wet– it is likely to remain so until it downs. A few insects, such as water-beetles, contrive to be unwettable; the majority keep well away from their drink by means of a long proboscis. (...)

Such are a very few of the considerations which show that for every type of animal there is an optimum size. Yet although Galileo demonstrated the contrary more than three hundred years ago, people still believe that if a flea were as large as a man it could jump a thousand feet into the air. As a matter of fact the height to which an animal can jump is more nearly independent of its size than proportional to it. A flea can jump about two feet, a man about five. To jump a given height, if we neglect the resistance of air, requires an expenditure of energy proportional to the jumper's weight. But if the jumping muscles form a constant fraction of the animal's body, the energy developed per ounce of muscle is independent of the size, provided it can be developed quickly enough in the small animal. As a matter of fact an insect's muscles, although they can contract more quickly than our own, appear to be less efficient; as otherwise a flea or grasshopper could rise six feet into the air." (Haldane 1926)

(b) according to the Haldane setup. Draw the tree diagram for the Haldane setup.

2. Choose some values for s and k and (a) apply the key equations in this chapter; (b) use the *Learn Bayes* $\rightarrow$ *Binomial Testing* functionality in JASP. Do the results match?

3. Haldane's Rule of Succession entails (at least) two important assumptions. What are they?



Portrait of J. B. S. Haldane, by Claude Rogers (1907-1979). Reproduced with permission of ©Crispin Rogers, who added: "I believe that this painting was done by my father Claude Rogers when our family lived at 13 Taviton Street in central London, very near to London University where my father was a lecturer at the Slade School of Fine Art. He knew Haldane well and was given an antique cupboard by him, which we have and call the Haldane cupboard. So the story goes Haldane stored his experiments in it."

Chapter Summary

Despite widely-felt dissatisfaction with Laplace's Rule of Succession, it took until 1932 before J. B. S. Haldane first proposed the mixture prior representation in which a Laplacean ‘slab’ is combined with a Wrinchean ‘spike’. Haldane computed the Bayes factor and applied model-averaging to obtain an alternative Rule of Succession in which the probability of finding an exception decreases as $1/n^2$ rather than $1/n$.

Unfortunately, Haldane’s result contains a typographical error and was not presented in the elegant form of Eq. 16.2. The citation record suggests that as far as the Rule of Succession is concerned, Haldane’s contribution has been almost entirely forgotten.[4] Nevertheless, Haldane’s work possibly motivated Jeffreys to start his extensive studies on Bayes factor hypothesis testing that culminated in his magnum opus *Theory of Probability* – a book that inspired generations of Bayesians including your authors.[5]

[4] We have added the qualifier ‘almost’ because of the work by Frank Tuyl and colleagues (i.e., Tuyl 2019, Tuyl et al. in press).

[5] The book you are now reading may be considered an accessible summary of the material from *Theory of Probability*.

Want to Know More?

✓ Clark, R. (1968/2013). *J. B. S. The Life and Work of J. B. S. Haldane*. London: Bloomsbury Reader. A gripping biography of one of the most interesting scientists of all time.

✓ Subramanian, S. (2019). *A Dominant Character: How J. B. S. Haldane Transformed Genetics, Became a Communist, and Risked his Neck for Science*. New York: W. W. Norton & Company. Another gripping biography.

✓ Devitt, D. (2022). *The Skipper’s War: Dragon School, Oxford & the Great War*. London: Scala Arts Publishers Inc. Contains several fragments on Haldane’s wartime heroism. See also `https://skipperswar.com/book/`.

✓ Haldane, J. B. S. (1926). On being the right size. *Harper’s Magazine, 152*, 424-427. Haldane wrote several popular-science books and short articles for the general public. “On being the right size” is one of his best-known works – the box below provides several characteristic excerpts.

✓ Haldane, J. B. S. (1932). A note on inverse probability. *Mathematical Proceedings of the Cambridge Philosophical Society, 28*, 55-61. In this short paper Haldane presents the first Bayes factor for a point null hypothesis versus an alternative hypothesis that involves a continuous (beta) distribution for θ. Based on this Bayes factor Haldane also proposes the concrete alternative to Laplace’s Rule of Succession that is outlined in this chapter. Although his derivation contain a typographical error, Haldane’s work clearly anticipates the later contributions by Jeffreys.

✓ Etz, A., & Wagenmakers, E.–J. (2017). J. B. S. Haldane's contribution to the Bayes factor hypothesis test. *Statistical Science, 32*, 313-329. The abstract: "This article brings attention to some historical developments that gave rise to the Bayes factor for testing a point null hypothesis against a composite alternative. In line with current thinking, we find that the conceptual innovation–to assign prior mass to a general law–is due to a series of three articles by Dorothy Wrinch and Sir Harold Jeffreys (1919, 1921, 1923a). However, our historical investigation also suggests that in 1932, J. B. S. Haldane made an important contribution to the development of the Bayes factor by proposing the use of a mixture prior comprising a point mass and a continuous probability density. Jeffreys was aware of Haldane's work and it may have inspired him to pursue a more concrete statistical implementation for his conceptual ideas. It thus appears that Haldane may have played a much bigger role in the statistical development of the Bayes factor than has hitherto been assumed." (p. 313)

✓ Wagenmakers, E.–J., Zabell, S., & Gronau, Q.F. (in press). J. B. S. Haldane's rule of succession. *Statistical Science*. This article contains the material from this chapter, but presents several generalizations as well, some of which will be covered in Chapter 17.

✓ Jeffreys occasionally gave credit to Haldane. One example is given in this chapter's epigraph. Another one is here:

> "Everybody in fact believes a large number of general laws, and as the function of the theory is to give a consistent statement of commonsense, and not to alter it in a fundamental respect, it appears that the estimate of Bayes and Laplace needs modification for the extreme cases. (...) for the case of sampling J. B. S. Haldane and I have pointed out that general laws can be established with reasonable probabilities if their prior probabilities are moderate and independent of the whole number of members of the class sampled. These rules have been called "simplicity postulates"; they do not say that any particular simple law must be true, or even that some simple law must be true, but they do say that when we consider a simple law seriously an assessment of the prior probability that will make it impossible ever to establish it even if it happens to be true is not a correct representation of our state of knowledge." (Jeffreys 1936a, p. 344)

Appendix A: "Stalin Was a Very Great Man Who Did a Very Good Job"

In many ways, J. B. S. Haldane was a hero. He spoke truth to power, fought fearlessly on the front in multiple wars (i.e., World War I in France and Iraq, and the Spanish civil war), and experimented on himself to find the most effective gas mask:

> "The Germans had attacked with chlorine north of Ypres. My father had been sent out to tackle the menace. I met him at Hazebrouck, and we started trying respirators of various kinds in a room in the college there in which chlorine was liberated. The concentration was not sufficient to cause fatal injury to the lungs in less than 2 minutes or so. But it made one cough very much sooner. About half a dozen of us went in, trying a different type of respirator; and another would take his place when he had inhaled enough gas to incapacitate him for a few hours, or in one case, for several days." (Haldane, unpublished autobiographical remarks, as reported in Tredoux 2018, p. 252)

Haldane was also characterized by "a combination of aristocratic self-assurance, intellectual integrity and almost endearing bloody-mindedness" (Clark 1968/2013, p. 3). That bloody-mindedness meant that Haldane was slow to acknowledge his mistakes:

> "Traditionally, there was one field in which no doubt could be allowed—that when a Haldane made up his mind that it was right to act, then action would follow as a duty, ignoring all obstacles or any suggestion that the proposed course could be anything other than the ideal. Like the aristocrat down the ages, he responded to opposition by not giving a damn for anyone (...) (Clark 1968/2013, p. 5)

At one point Haldane had become an active member of the communist party in the UK. As the atrocities of the Soviet regime became ever more visible, Haldane found himself unable to speak up publicly against the deportation of scientists, against the influence of Lysenko, against the Molotov-Ribbentrop Pact, and against Stalin in general. This uncharacteristic meekness was not born out of fear or out of malevolence – it probably arose because Haldane could not bring himself to acknowledge that he had been wrong; in other words, it root cause was sheer bloody-mindedness.

There has even been a suggestion that Haldane was a Soviet spy. In a polemic thinly disguised as a biography, Tredoux (2018) corrects the rumor that Haldane was the Soviet spy INTELLIGENTSIA (which was Ivor Montagu). Tredoux (2018) does argue that Haldane was a Soviet spy, but the only evidence for this claim appears to be the fact that Montagu passed on a 1940 army report by Haldane on how long a man could remain underwater.[6] It appears to us that a mountain gave birth to a mouse. More damning is Haldane's continued, bloody-minded support for Stalin:

[6] From the Venona intercepts: "INTELLIGENTSIA has handed over a copy of Professor HALDANE's report to the Admiralty on his experience relating to the length of time a man can stay underwater." (Tredoux 2018, p. 319).

> "As for Haldane, he never let go of Stalin. The Society for Cultural Relations with the USSR passed him a letter of condolence to co-sign when Stalin, "one of the great men of world history," died in 1953. He was glad to do so. A letter to a friend written during his final days in India shows that he did not even accept Khrushchev's renunciation of Stalin in his secret speech of 1956 (...). "I certainly don't go all the way with Khrushchev. As you know, I disagreed, during Stalin's lifetime, with

> some of his actions. But I thought, and think, that he was a very great man who did a very good job. And as I did not denounce him then, I am not going to do so now." (Tredoux 2018, pp. 156-157; see also Clark 1968/2013, p. 326)

This is yet another unfortunate demonstration of how difficult it is for a person to change their long-held opinions, especially when this person is a genius and bloody-mindedness runs in the family.

Appendix B: "Cancer's a Funny Thing"

Haldane's wit, writing skills, and combative nature are all on full display in the famous poem "Cancer's a funny thing". The poem was composed in a London hospital bed, as Haldane was recovering from surgery. The surgery was declared a success, but the cancer would soon return. Haldane died December 1st of the same year in Bhubaneswar, India. "The poem, which was reprinted in a number of countries, brought great praise, caused great offence, and in some ways crystallises both Haldane's attitude to the world and the world's reaction." (Clark 1968/2013, p. 340)

Cancer's a Funny Thing
I wish I had the voice of Homer
To sing of rectal carcinoma,
Which kills a lot more chaps, in fact,
Than were bumped off when Troy was sacked.
Yet, thanks to modern surgeon's skills,
It can be killed before it kills
Upon a scientific basis
In nineteen out of twenty cases.
I noticed I was passing blood
(Only a few drops, not a flood).
So pausing on my homeward way
From Tallahassee to Bombay
I asked a doctor, now my friend,
To peer into my hinder end,
To prove or to disprove the rumour
That I had a malignant tumour.
They pumped in $BaSO_4$.
Till I could really stand no more,
And, when sufficient had been pressed in,
They photographed my large intestine,
In order to decide the issue
They next scraped out some bits of tissue.
(Before they did so, some good pal
Had knocked me out with pentothal,
Whose action is extremely quick,
And does not leave me feeling sick.)
The microscope returned the answer

That I had certainly got cancer,
So I was wheeled into the theatre
Where holes were made to make me better.
One set is in my perineum
Where I can feel, but can't yet see 'em.
Another made me like a kipper
Or female prey of Jack the Ripper,
Through this incision, I don't doubt,
The neoplasm was taken out,
Along with colon, and lymph nodes
Where cancer cells might find abodes.
A third much smaller hole is meant
To function as a ventral vent:
So now I am like two-faced Janus
The only[1] god who sees his anus.
I'll swear, without the risk of perjury,
It was a snappy bit of surgery.
My rectum is a serious loss to me,
But I've a very neat colostomy,
And hope, as soon as I am able,
To make it keep a fixed time-table.
So do not wait for aches and pains
To have a surgeon mend your drains;
If he says "cancer" you're a dunce
Unless you have it out at once,
For if you wait it's sure to swell,
And may have progeny as well.
My final word, before I'm done,
Is "Cancer can be rather fun".
Thanks to the nurses and Nye Bevan
The NHS is quite like heaven
Provided one confronts the tumour
With a sufficient sense of humour.
I know that cancer often kills,
But so do cars and sleeping pills;
And it can hurt one till one sweats,
So can bad teeth and unpaid debts.
A spot of laughter, I am sure,
Often accelerates one's cure;
So let us patients do our bit
To help the surgeons make us fit."
(J. B. S. Haldane, first printed in *The New Statesman*, 21 February 1964)

[1] In India there are several more
With extra faces, up to four,
But both in Brahma and in Shiva
I own myself an unbeliever.

[Aneurin "Nye" Bevan (1897-1960) was a Welsh Labour Party politician who had helped create the British National Health Service – EWDM].

17 Jeffreys's Platitude

The most beneficial result that I can hope for as a consequence of this work is that more attention will be paid to the precise statement of the alternatives involved in the questions asked. It is sometimes considered a paradox that the answer depends not only on the observations but on the question; it should be a platitude.

Jeffreys, 1961

Chapter Goal

This chapter emphasizes that (1) *prior* distributions on model parameters partly determine the model *predictions*; (2) the relative adequacy of the model predictions define the *evidence* (i.e., the *Bayes factor*), that is, the extent to which the data change our beliefs; (3) consequently, different prior distributions result in different Bayes factors. This tautology needs to be understood and exploited rather than bemoaned and avoided.

Predictions, Evidence, and Prior Distributions

Throughout this book we stress a key principle of Bayesian inference: hypotheses that predicted observed data successfully enjoy a boost in plausibility, whereas hypotheses that predicted the data poorly suffer a decline. The change in plausibility brought about by the data –the *evidence*– is known as the Bayes factor. We repeat the updating rule:

$$\underbrace{\frac{p(\mathcal{H}_0 \mid \text{data})}{p(\mathcal{H}_1 \mid \text{data})}}_{\substack{\text{Posterior beliefs}\\ \text{about hypotheses}}} = \underbrace{\frac{p(\mathcal{H}_0)}{p(\mathcal{H}_1)}}_{\substack{\text{Prior beliefs}\\ \text{about hypotheses}}} \times \underbrace{\frac{p(\text{data} \mid \mathcal{H}_0)}{p(\text{data} \mid \mathcal{H}_1)}}_{\text{Bayes factor BF}_{01}}. \tag{17.1}$$

In the following our focus remains on the case of pure induction, such that $\mathcal{H}_0$ represents the general law according to which the population proportion θ equals 1 (i.e., *all* zombies are hungry). This general law is pitted against an alternative hypothesis $\mathcal{H}_1$ that relaxes the restriction imposed on θ. As in the previous chapter, we consider the case where all instances accord with the general law, so $s = n$. With only confirmatory instances observed, we can already draw three qualitative conclusions:

- The evidence favors $\mathcal{H}_0$ over $\mathcal{H}_1$.[1] This has to be the case because the general law makes only a single prediction (e.g., 'the next zombie will certainly be hungry') and hence $p(s = n \mid \mathcal{H}_0) = 1$. By relaxing the restriction that $\theta = 1$, the alternative hypothesis $\mathcal{H}_1$ also predicts other outcomes, and hence $p(s = n \mid \mathcal{H}_1) < 1$.
- Every new confirmatory instance that is observed increases the evidence for the general law $\mathcal{H}_0$.[2] Intuitively, this happens because even after many confirmatory instances have been observed, the alternative hypothesis $\mathcal{H}_1$ still does not assign probability 1 to the next instance being confirmatory, whereas $\mathcal{H}_0$ does.
- The degree to which the data support $\mathcal{H}_0$ over $\mathcal{H}_1$ depends directly on how close $p(s = n \mid \mathcal{H}_1)$ is to 1. When the data '$s = n$' (i.e., all observed instances are confirmatory) are highly likely under $\mathcal{H}_1$), then the evidence in favor of $\mathcal{H}_0$) will be relatively modest; but when the data '$s = n$' are highly unlikely under $\mathcal{H}_1$), the evidence in favor of $\mathcal{H}_0$) will be relatively compelling. Thus, the strength of evidence that the data provide for $\mathcal{H}_0$) depends critically on the predictive adequacy of $\mathcal{H}_1$). This adequacy is determined by the prior distribution for θ under $\mathcal{H}_1$).

[1] One exception that proves the rule is given by case 1 below. For other exceptions based on background knowledge see Chapter 9, Appendix B: 'Conforming observations need not be confirming'.

[2] The exception that proves the rule is given by case 2 below.

Before starting in earnest, consider three cases in which $\mathcal{H}_1$ is specified by a point-prior (i.e., a spike) at a particular value of θ.[3] For concreteness, we continue the example from Chapter 15: based on an observed sequence of 12 hungry zombies we wish to quantify the evidence for $\mathcal{H}_0 : \theta = 1$ ('all zombies are hungry') versus $\mathcal{H}_1$).

[3] In these cases, the Bayes factor reduces to a likelihood ratio, cf. Chapter 7.

1. Consider $\mathcal{H}_1 : \theta = 1$. This specification means that $\mathcal{H}_1$ is identical to $\mathcal{H}_0$; just as $\mathcal{H}_0$, $\mathcal{H}_1$ predicts that all instances are confirmatory. The question that is being asked is, 'Are the data predicted better by the hypothesis that all zombies are hungry or by the hypothesis that all zombies are hungry?' The Bayes factor equals 1 regardless of the value of $s = n$:

$$\mathrm{BF}_{01} = 1 \qquad \text{if } \begin{aligned} \mathcal{H}_0 &: \theta = 1, \\ \mathcal{H}_1 &: \theta = 1, \\ \text{data} &: s = n. \end{aligned}$$

2. Consider $\mathcal{H}_1 : \theta = 0$. This specification means that $\mathcal{H}_1$ is maximally different from $\mathcal{H}_0$; in diametric opposition to $\mathcal{H}_0$, $\mathcal{H}_1$ predicts that all instances are non-confirmatory (e.g., all zombies are satiated). The question that is being asked is, 'Are the data predicted better by the

hypothesis that all zombies are hungry or by the hypothesis that no zombie is hungry?' A single zombie suffices to obtain a certain answer: $\text{BF}_{01} = \infty$ if the first zombie is hungry (as in our example), and $\text{BF}_{10} = \infty$ if the first zombie is not hungry:

$$\text{BF}_{01} = \infty \qquad \begin{aligned} \text{if } \mathcal{H}_0 &: \theta = 1, \\ \mathcal{H}_1 &: \theta = 0, \\ \text{data} &: s = n \geq 1. \end{aligned}$$

3. Consider $\mathcal{H}_1 : \theta = 1/2$. This specification means that exactly half of the instances in the population are assumed to accord with the general law. The question that is being asked is, 'Are the data predicted better by the hypothesis that all zombies are hungry or by the hypothesis that half of the zombie population is hungry?' Every new hungry zombie is twice as likely to occur under $\mathcal{H}_0 : \theta = 1$ than under $\mathcal{H}_1 : \theta = 1/2$. Therefore we have:

$$\text{BF}_{01} = 2^s \qquad \begin{aligned} \text{if } \mathcal{H}_0 &: \theta = 1, \\ \mathcal{H}_1 &: \theta = 1/2, \\ \text{data} &: s = n. \end{aligned}$$

For the example featuring 12 hungry zombies, $\text{BF}_{01} = 2^{12} = 4096$.

These three cases form extreme examples in the sense that $\mathcal{H}_1$ is specified as a single value of θ. Hence there can be no learning and the data cannot overwhelm the prior, because the prior cannot budge from its initial value. We now consider several scenarios in which $\mathcal{H}_1$ is characterized by a beta prior on θ. In these scenarios the prior distribution on θ is updated by the data such that $\mathcal{H}_1$ 'learns' that θ is near 1 as the number of confirmatory instances increases. Nevertheless, the scenarios below demonstrate that the evidence remains highly dependent on the prior distribution.[4]

[4] See also the assessment of the pancake forecasters in Chapters 12 and 13, and see exercise 3 from Chapter 15.

Scenario 1: 'All Options Open'

Consider $\mathcal{H}_1 : \theta \sim \text{beta}(1, 1)$. Detailed in Chapter 15, this specification means that all possible values for θ are deemed equally likely *a priori*. Colloquially one may term this the 'all options open' model. The question that is being asked is, 'Are the data predicted better by the hypothesis that all zombies are hungry or by the hypothesis that every proportion of hungry zombies is *a priori* equally likely?' The uniform distribution on θ induces a predictive distribution on the $n + 1$ possible outcomes (i.e., from 0 to n confirmatory instances) that is likewise uniform (cf. Figures 14.1 and Figures 15.4). This means that the prior predictive mass on the result '$s = n$' is $^1/_{(s+1)}$. Hence we have:

$$\mathrm{BF}_{01} = s + 1 \qquad \text{if } \mathcal{H}_0 : \theta = 1, \quad \mathcal{H}_1 : \theta \sim \text{beta}(1,1), \quad \text{data} : s = n.$$

For the example featuring 12 hungry zombies, $\mathrm{BF}_{01} = 13$.

Scenario 2: 'Most Instances Are Confirmatory'

Consider $\mathcal{H}_1 : \theta \sim \text{beta}(\alpha, 1)$, with $\alpha > 1$. This specification means that values for θ are deemed more likely the closer they are to $\theta = 1$. The higher the value of α, the more the prior distribution is concentrated near $\theta = 1$. Figure 17.1 gives an example of a beta$(12, 1)$ prior distribution.

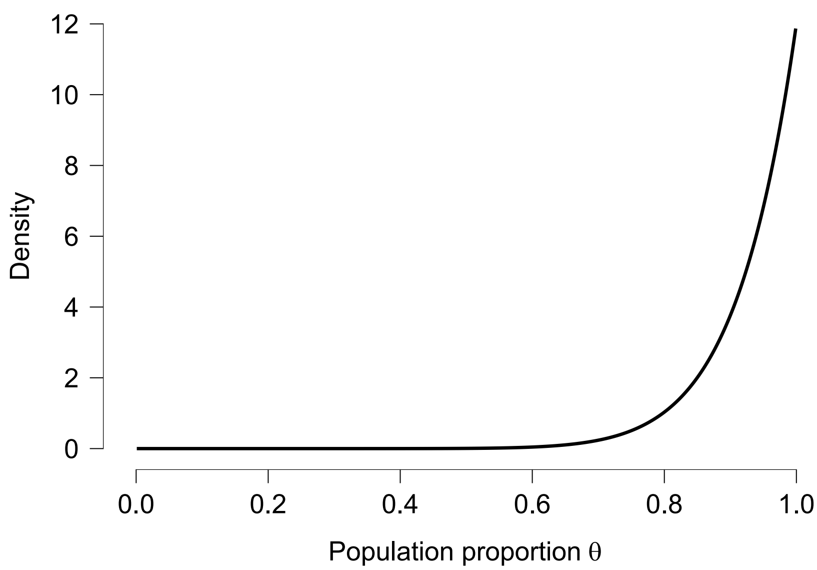


Figure 17.1: The beta(12,1) prior distribution for θ under $\mathcal{H}_1$. Values of θ near 1 are deemed relatively likely. Figure from the JASP module *Learn Bayes*.

The question that is being asked is, 'Are the data predicted better by the hypothesis that all zombies are hungry or by the hypothesis that most hungry zombies are hungry?' Note that this question is more difficult to answer than the question from the previous scenario. This is underscored by the fact that the monotonically increasing beta distribution on θ induces a predictive distribution on the $n + 1$ possible outcomes that is likewise monotonically increasing. For example, Figure 17.2 shows the predictions for a data set of 12 zombies that follow from the beta$(12, 1)$ distribution. The figure suggests that the prior mass on $s = n = 12$ equals about 0.5, which would mean that the Bayes factor in favor of $\mathcal{H}_0$ is about 2.

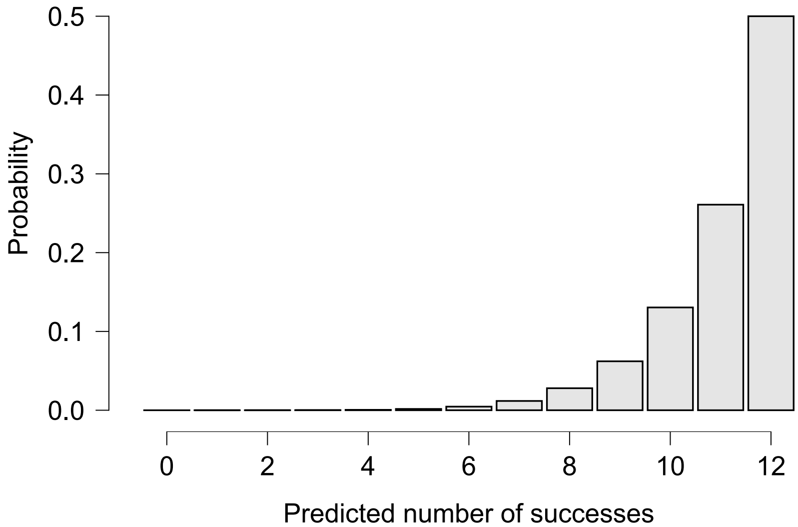


Figure 17.2: Predictions for a data set of 12 zombies, as induced by the beta(12,1) prior distribution for θ shown in Figure 17.1. Figure from the JASP module *Learn Bayes*.

This suggestion is correct. The general expression for the Bayes factor equals:

$$\mathrm{BF}_{01} = \frac{s}{\alpha} + 1 \qquad \begin{aligned} &\text{if } \mathcal{H}_0 : \theta = 1, \\ &\mathcal{H}_1 : \theta \sim \text{beta}(\alpha, 1), \alpha \geq 1 \\ &\text{data} : s = n. \end{aligned}$$

For the example featuring $\mathcal{H}_1 : \theta \sim \text{beta}(12, 1)$ and 12 hungry zombies, $\mathrm{BF}_{01} = (12/12) + 1 = 2$. It is important to recognize the crucial impact of α on the Bayes factor for the comparison to $\mathcal{H}_1 : \theta \sim \text{beta}(\alpha, 1)$. Essentially α quantifies the degree of similarity between $\mathcal{H}_0$ and $\mathcal{H}_1$; the higher α, the more prior mass is allocated to the event that $s = n$, and the less diagnostic are the data. Concretely, if α is doubled, the number of confirmatory instances needs to be doubled as well in order to attain the same level of evidence.[5]

[5] A reassuring note: for models that are commonly used in scientific practice, different prior distributions often do not cause the Bayes factor to change so much, unless the prior distributions are deeply implausible.

Scenario 3: Most Instances Are Not Confirmatory

Consider $\mathcal{H}_1 : \theta \sim \text{beta}(1, \beta)$, with $\beta > 1$. This specification means that values for θ are deemed more likely the closer they are to $\theta = 0$. The higher the value of β, the more the prior distribution is concentrated near $\theta = 0$. Figure 17.3 gives an example of a beta$(1, 4)$ prior distribution.

The question that is being asked is, 'Are the data predicted better by the hypothesis that all zombies are hungry or by the hypothesis that most hungry zombies are not hungry?' Note that this question is

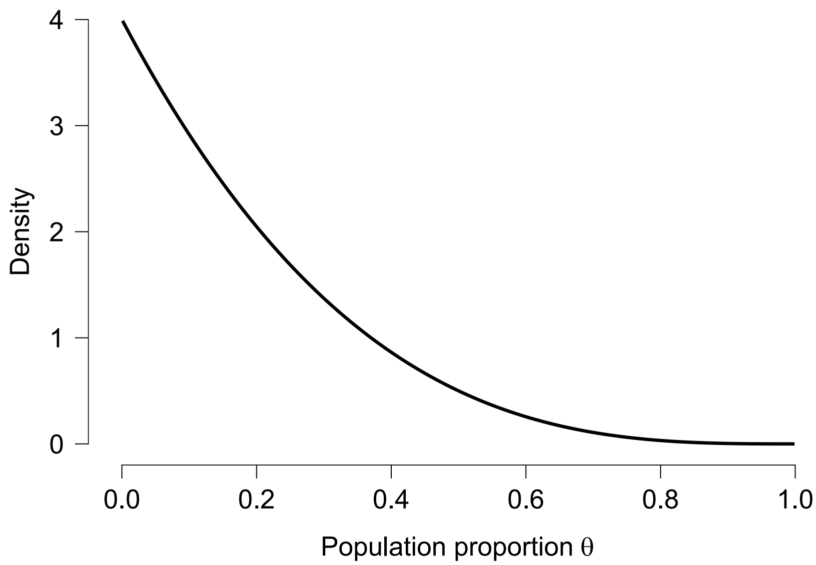


Figure 17.3: The beta(1,4) prior distribution for θ under $\mathcal{H}_1$. Values of θ near 0 are deemed relatively likely. Figure from the JASP module *Learn Bayes*.

relatively easy to answer, because the hypotheses make very different predictions. Specifically, the monotonically decreasing beta distribution on θ induces a predictive distribution on the $n+1$ possible outcomes that is likewise monotonically decreasing. For example, Figure 17.4 shows the predictions for a data set of 12 zombies that follow from the beta$(1,4)$ distribution. The figure suggests that the prior mass on $s = n = 12$ is very low, which would mean that the Bayes factor in favor of $\mathcal{H}_0$ is very high.

This suggestion is again correct. The general expression for the Bayes factor equals:

$$\begin{aligned} \text{BF}_{01} = \frac{(s+\beta)!}{s!\,\beta!} \qquad\qquad & \text{if } \mathcal{H}_0 : \theta = 1, \\ & \quad \mathcal{H}_1 : \theta \sim \text{beta}(1,\beta), \beta \geq 1 \\ & \text{data} : s = n. \end{aligned}$$

For the example featuring $\mathcal{H}_1 : \theta \sim \text{beta}(1,4)$ and 12 hungry zombies, $\text{BF}_{01} = 16!/(12!\,4!) = 1820$. As was the case for α in the previous scenario, β exerts a powerful impact on the Bayes factor for the comparison of $\mathcal{H}_1 : \theta \sim \text{beta}(1,\beta)$ to $\mathcal{H}_0 : \theta = 1$. Here β quantifies the degree of *dissimilarity* between $\mathcal{H}_0$ and $\mathcal{H}_1$; the higher β, the less prior mass is allocated to the event that $s = n$, and the more diagnostic are the data. To appreciate the role of β, notice that when $\beta = 1$ and $s = n = 1000$, this gives $\text{BF}_{01} = 1001$ – a thousand confirmatory instances yield a Bayes factor of 1001 when $\mathcal{H}_1$ stipulates a uniform prior distribution on θ. The same evidence is obtained when the roles of $s = n$ and β are switched,

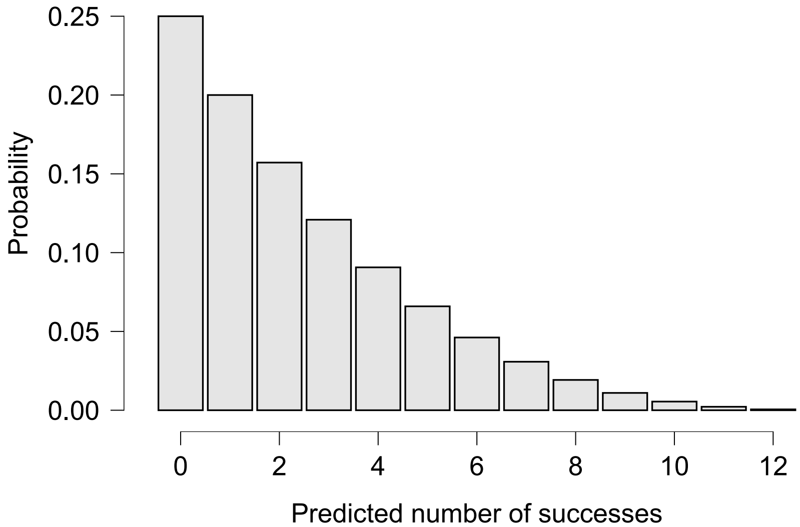


Figure 17.4: Predictions for a data set of 12 zombies, as induced by the beta(1,4) prior distribution for θ shown in Figure 17.3. Figure from the JASP module *Learn Bayes*.

that is, when $s = n = 1$ and $\beta = 1000$. Thus, a single confirmatory instance yields a Bayes factor of 1001 when $\mathcal{H}_1$ stipulates a beta$(1, 1000)$ prior distribution on θ. When $\beta \to \infty$, the comparison approximates a test between $\mathcal{H}_0 : \theta = 1$ versus $\mathcal{H}_1 : \theta = 0$ (case 2 discussed at the beginning of this chapter), and a single outcome is decisive.

Scenario 4: About Half of the Instances are Confirmatory

Consider $\mathcal{H}_1 : \theta \sim \text{beta}(\alpha, \alpha)$, with $\alpha > 1$. This specification means that values for θ are deemed more likely the closer they are to $\theta = 1/2$. The higher the value of α, the more the prior distribution is concentrated near $\theta = 1/2$. Figure 17.3 gives an example of a beta$(2, 2)$ prior distribution.

The question that is being asked is, 'Are the data predicted better by the hypothesis that all zombies are hungry or by the hypothesis that about half of the zombie population is hungry?' This question is again relatively easy to answer, because the rival hypotheses make very different predictions. The dome-shaped beta distribution on θ induces a predictive distribution on the $n + 1$ possible outcomes that is also dome-shaped, and therefore assigns the least mass to extreme outcomes such as $s = n$. For example, Figure 17.6 shows the predictions for a data set of 12 zombies that follow from the beta$(2, 2)$ distribution. There is modest prior mass on $s = n = 12$, and this means that the Bayes factor in favor of $\mathcal{H}_0$ should be relatively high.

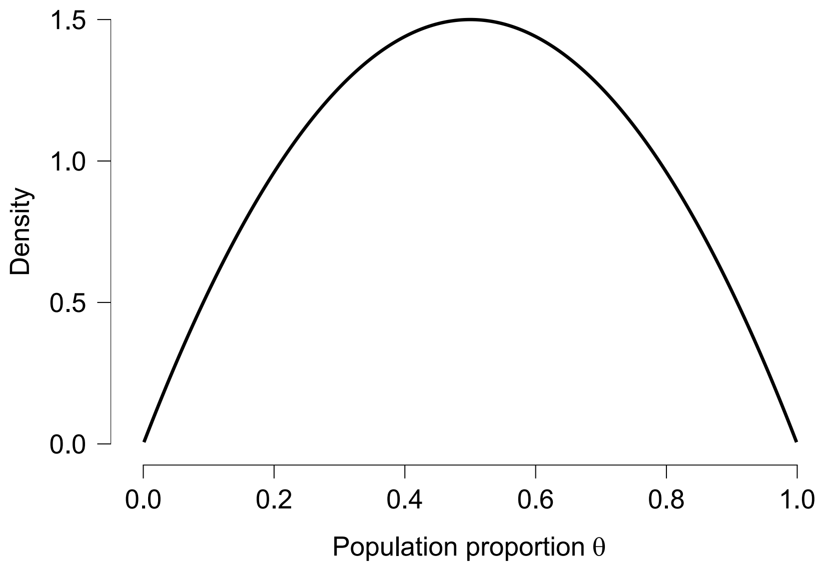


Figure 17.5: The beta(2,2) prior distribution for θ under $\mathcal{H}_1$. Values of θ near $1/2$ are deemed relatively likely. Figure from the JASP module *Learn Bayes*.

The associated analytical expression for the Bayes factor equals:

$$\begin{aligned}\mathrm{BF}_{01} &= \frac{(\alpha-1)!\ (2\alpha+s-1)!}{(2\alpha-1)!\ (\alpha+s-1)!} \\ &= \prod_{\alpha}^{2\alpha-1}\left[\frac{s+\alpha}{\alpha}\right] = \prod_{\alpha}^{2\alpha-1}\left[\frac{s}{\alpha}+1\right] \qquad \text{if } \mathcal{H}_0 : \theta = 1, \\ &\qquad\qquad \mathcal{H}_1 : \theta \sim \text{beta}(\alpha,\alpha), \alpha \geq 1 \\ &\qquad\qquad \text{data} : s = n.\end{aligned}$$

The elegance of this equation can be appreciated better when it is written out for a number of different values of α:

$$\begin{aligned}&\text{if } \alpha = 1 : \mathrm{BF}_{01} = s+1 \\ &\text{if } \alpha = 2 : \mathrm{BF}_{01} = \frac{s+2}{2} \times \frac{s+3}{3} \\ &\text{if } \alpha = 3 : \mathrm{BF}_{01} = \frac{s+3}{3} \times \frac{s+4}{4} \times \frac{s+5}{5} \\ &\text{if } \alpha = 4 : \mathrm{BF}_{01} = \frac{s+4}{4} \times \frac{s+5}{5} \times \frac{s+6}{6} \times \frac{s+7}{7} \\ &\text{if } \alpha = 5 : \mathrm{BF}_{01} = \frac{s+5}{5} \times \frac{s+6}{6} \times \frac{s+7}{7} \times \frac{s+8}{8} \times \frac{s+9}{9}.\end{aligned}$$

Note that the Bayes factors in favor of the general law $\mathcal{H}_0$ increase with α, that is, the evidence becomes more compelling when the prior distribution for θ under $\mathcal{H}_1$ is more peaked around the value of $\theta = 1/2$.[6] For the example featuring $\mathcal{H}_1 : \theta \sim \text{beta}(2,2)$ and 12 hungry zombies, $\mathrm{BF}_{01} = 1/6(12+2)(12+3) = 35$.

[6] Also noteworthy is that the first factor in the series, $s/\alpha + 1$, equals the Bayes factor for $\mathcal{H}_0$ against $\mathcal{H}_1 : \theta \sim \text{beta}(\alpha, 1)$ (i.e, scenario 2 above).

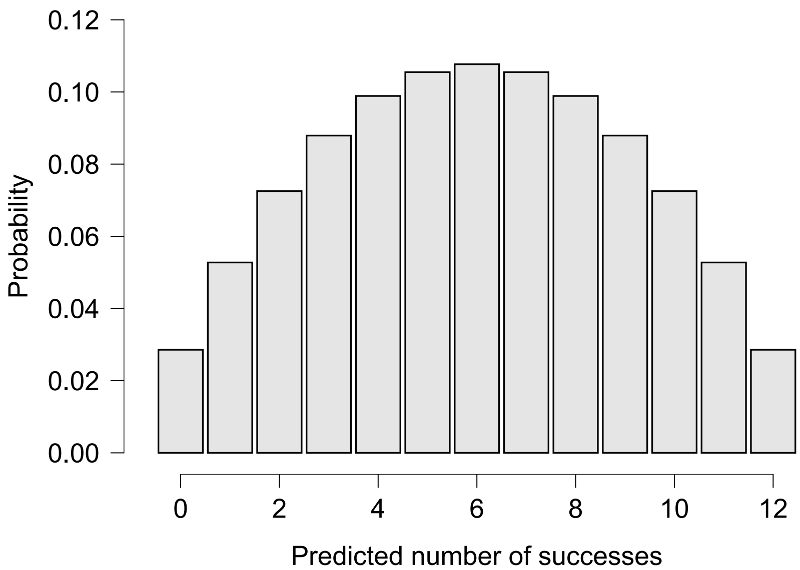


Figure 17.6: Predictions for a data set of 12 zombies, as induced by the beta(2,2) prior distribution for θ shown in Figure 17.5. Figure from the JASP module *Learn Bayes*.

An Inconvenient Truth

The scenarios above reveal a truth that many statisticians find highly inconvenient: when it comes to quantifying evidence for competing hypotheses, the prior distribution on the model parameters matters – and as we have seen it may matter a great deal. Of course, Bayes' rule tells us the prior distribution *should* matter: it partly determines the model predictions, and the evidence is given by the models' relative predictive performance. A carefully chosen prior distribution will result in a meaningful assessment of the evidence (i.e., the extent to which the data change our opinion) and we know of no other statistical methodology that is able to achieve this goal.

But what if you don't 'know' the prior distribution for the parameter under $\mathcal{H}_1$? In the above example you may even refuse to specify what scenario is relevant. If you find yourself in this situation, then:

1. You are unable to specify the predictions under the alternative hypothesis $\mathcal{H}_1$.

2. More generally, you do not know what question to ask.

3. Consequently, you are not in the position to quantify evidence, that is, determine the degree to which the data ought to change your beliefs concerning $\mathcal{H}_0$.

4. You are advised to collect more information so that you may then put forward a specific question, that is, an alternative hypothesis that makes predictions.

5. You may try out several prior distributions and use these to generate synthetic data – that is, you may inspect the *prior predictive distribution*. These prior predictive data may provide more concrete guidance as to what prior distributions are reasonable.

On the other hand, in the above example, you may know what scenario applies but you do not know *exactly* what prior distribution is reflects your background knowledge best (i.e., do I specify $\mathcal{H}_1 : \theta \sim$ beta$(2, 1)$ or do I specify $\mathcal{H}_1 : \theta \sim$ beta$(3, 1)$?). In such cases it is prudent simply to try them all, and see whether it matters. This is termed a *sensitivity analysis* or a *robustness analysis*. When the conclusions from the various plausible prior distributions differ substantially then this is something that needs to be acknowledged; perhaps more data need to be collected. In our experience with standard statistical models, the Bayes factor is actually surprisingly robust to reasonable changes in the prior distribution.

We conclude this chapter with a corollary to Jeffreys's platitude: *If you don't know the question, you are in no position to demand an answer.*

Exercises

1. Suppose $\mathcal{H}_0 : \theta = 1$ and $\mathcal{H}_1 : \theta = 0$. The first zombie is hungry, but the second zombie is not. What do you conclude?

2. Consider another scenario: $\mathcal{H}_1 : \theta \sim$ beta(α, α) and $\alpha \to \infty$. What is the Bayes factor in favor of $\mathcal{H}_0 : \theta = 1$ when s confirmatory instances are observed?

3. You observe $s = n$ confirmatory instances. What is the Bayes factor for $\mathcal{H}_A : \theta \sim$ beta$(\alpha, 1)$ versus $\mathcal{H}_B : \theta \sim$ beta(α, α) [hint: exploit the fact that Bayes factors are transitive]. Confirm your answer with the *Learn Bayes* module in JASP, using the case of $n = 12$ and $\alpha = 2$.

4. Consider the Bayes factor for $\mathcal{H}_1 : \theta \sim$ beta(α, α) against $\mathcal{H}_0 : \theta = 1$. When a single confirmatory instance is observed (i.e., $s = n = 1$), the Bayes factor equals 2 regardless of the value of α. Confirm this with the equations, and provide an intuition as to why this must be the case.

Chapter Summary

The prior distribution for the model parameters partly governs the model predictions, and the relative adequacy of the predictions in turn defines the *evidence*. Hence it cannot come as a surprise that the prior partly determines the evidence – that is, the Bayes factor. Each prior distribution in fact defines a different model, and effectively poses a different question.

We highlighted the fact that radically different questions (i.e., radically different prior distributions) yield radically different answers. We should therefore not expect an answer if we do not know the question.

Want to Know More?

✓ Etz, A., Haaf, J. M., Rouder, J. N., & Vandekerckhove, J. (2018). Bayesian inference and testing any hypothesis you can specify. *Advances in Methods and Practices in Psychological Science, 1*, 281–295. This article echoes the main message from this chapter. The authors discuss Jeffreys's platitude and demonstrate how different models instantiate different questions, that then yield different answers.

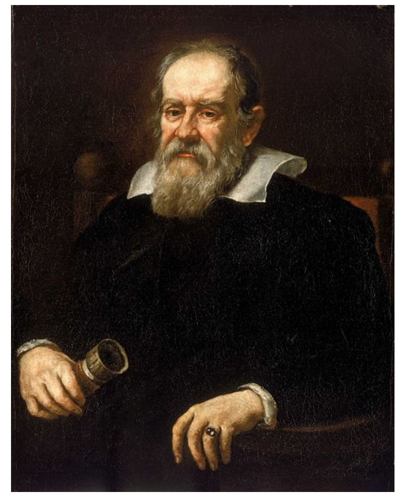
> "Critical in the model-selection endeavor is the specification of the models. In the case of hypothesis testing, it is of the greatest importance that the researcher specify exactly what is meant by a "null" hypothesis as well as the alternative to which it is contrasted, and that these are suitable instantiations of theoretical positions. Here, we provide an overview of different instantiations of null and alternative hypotheses that can be useful in practice, but in all cases the inferential procedure is based on the same underlying method of likelihood comparison." (p. 281).

✓ Rouder, J. N., Haaf, J. M., & Aust, F. (2018). From theories to models to predictions: A Bayesian model comparison approach. *Communication Monographs, 85*, 41–56.

✓ Vanpaemel, W. (2010). Prior sensitivity in theory testing: An apologia for the Bayes factor. *Journal of Mathematical Psychology, 54*, 491–498.

> "A commonly voiced concern with the Bayes factor is that, unlike many other Bayesian and non-Bayesian quantitative measures of model evaluation, it is highly sensitive to the parameter prior. This paper argues that, when dealing with psychological models that are quantitatively instantiated theories, being sensitive to the prior is an attractive feature of a model evaluation measure. (...) Because the prior is a vehicle for expressing psychological theory, it should, like the model equation, be considered as an integral part of the model. It is argued that the combined practice of building models using informative priors, and evaluating models using prior sensitive measures advances knowledge." (p. 491)

18 The Principle of Parsimony



Galileo Galilei (1564-1642), father of modern science. "When, therefore, I observe a stone initially at rest falling from an elevated position and continually acquiring new increments of speed, why should I not believe that such increases take place in a manner which is exceedingly simple and rather obvious to everybody?" (Galileo 1638/1914, p. 161). Portrait from 1636 by Justus Sustermans.

We consider it a good principle to explain the phenomena by the simplest hypotheses possible.

Ptolemy

Chapter Goal

As outlined in the previous chapters, Wrinch, Jeffreys, and Haldane avoided the Laplacean prejudice against a universal generalization by assigning it a separate prior mass. This way they solved the problem of pure induction, and quantified how every confirmatory instance provides evidence in favor of the universal generalization.

However, the Wrinch-Jeffreys-Haldane proposal applies to a broad range of scenarios that involve learning from data, as it formalizes the common scientific practice of retaining the simpler hypothesis until the data provide evidence against it: "The onus of proof is always on the advocate of the more complicated hypothesis." (Jeffreys 1961, p. 343)

This chapter introduces the principle of parsimony in scientific learning. The next chapters will describe two Bayesian *simplicity postulates* that jointly explain the scientific attitude towards parsimonious models.

Galileo's Experiment

We introduce the principle of parsimony by closely following the example outlined in Jeffreys (1973, pp. 61–64): "We consider an experiment that is done in first year physics classes. A solid of revolution can roll down an inclined plane, and its displacement is observed every fifth second after it starts from rest." The first such experiment was conducted by Galileo Galilei, who let a bronze ball roll down a ramp to measure the time t it took for the ball to reach particular distances x.[1] The outcome of the experiment supported Galileo's hypothesis that a falling object picks up equal speed in equal intervals of time; in other words, the rate of acceleration is constant. Jeffreys provides the following ex-

[1] To measure time Galileo used a water-clock or *klepsydra*.

"A motion is said to be uniformly accelerated, when starting from rest, it acquires, during equal time-intervals, equal increments of speed." (Galileo 1638/1914, p. 162).

ample data (for an extended discussion with empirical data see Jeffreys 1936a, pp. 351-353; see also Jeffreys 1961, pp. 3-4, 46-47):

time t (sec.)	0	5	10	15	20	25	30
displacement x (cm.)	0	5	20	45	80	125	180

For this ramp the displacement is related to time by the equation $5x = t^2$. However, Jeffreys notes, "the facts would be fitted equally well if the displacement was really connected with the time by the formula

$$5x = t^2 + t(t-5)(t-10)(t-15)(t-20)(t-25)(t-30)f(t),$$

where $f(t)$ might be any function whatever that is finite at

$$t = 0, 5, 10, ...30 \text{ sec.}$$

The law $5x = t^2$ is not the only description that fits the data; it is only one of an infinite number of descriptions that would fit the data equally well."[2]

[2] Note of clarification: the observed times are either $0, 5, 10, 15, 20, 25$ or 30 sec. This means that one of the multiplicative terms in the expression $t(t-5)(t-10)(t-15)(t-20)(t-25)(t-30)$ equals zero, and for these observed time points the relation therefore simplifies to $5x = t^2$.

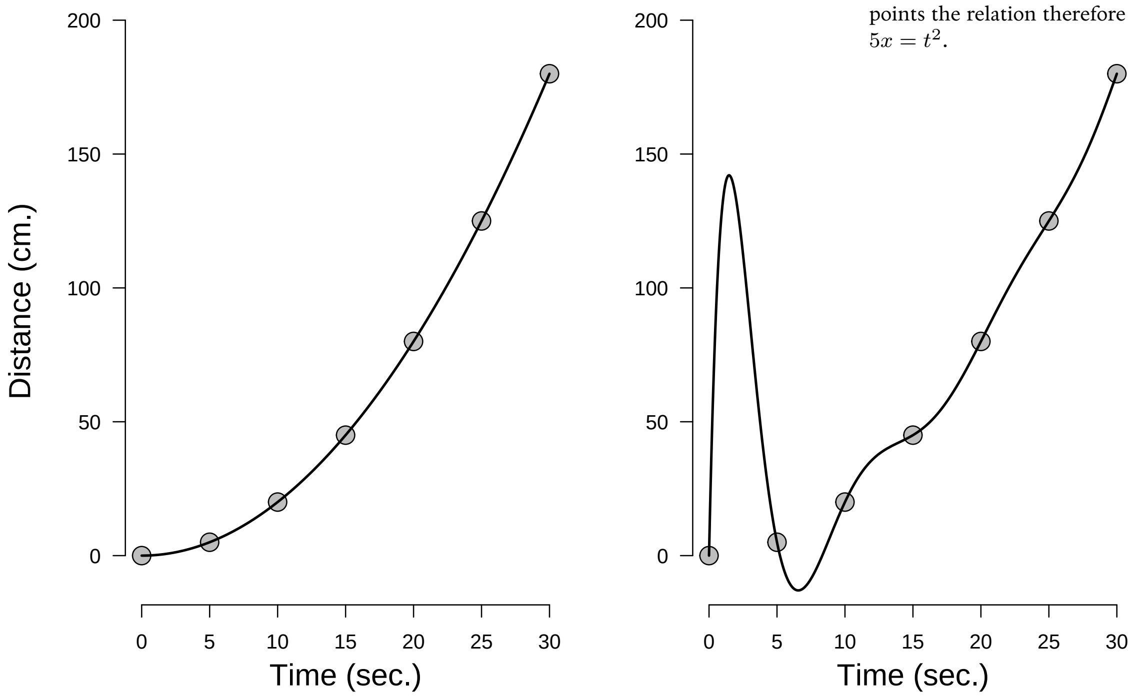


Figure 18.1: Preference for parsimony in a fictitious physics experiment described by Jeffreys (1973). Balls roll down a ramp and the displacement x is measured every 5 seconds. Left panel: The observations obey the simple equation $5x = t^2$. Right panel: a less parsimonious equation fits the observations equally well. Scientists have a strong preference for the simple equation.

As an illustration of Jeffreys's point, the left panel of Figure 18.1 shows the simple $5x = t^2$ relation, whereas the right panel shows a much more complicated relation between time and displacement that also captures the data exactly. Confronted with a possible choice between the two relations, scientists will select the simple model without any hesitation. Jeffreys concludes:

> "An infinite number of laws agree with previous experience, and an infinite number that have agreed with previous experience will inevitably be wrong in the next instance. What the applied mathematician does, in fact, is to select one form out of this infinity; and his reason for doing so has nothing whatever to do with traditional logic. He chooses the simplest." (Jeffreys 1961, pp. 3-4)

In fact, the preference for parsimony is so strong that scientists will adopt simple models even when these models describe the data *less well* than their more complex competitors. To show this, Jeffreys (1973) introduces a new example data set, where the displacement is now subject to a small measurement error:

time t (sec.)	0	5	10	15	20	25	30
displacement x (cm.)	0	5	19	44	81	124	178

For this data set, the fit of the square law model $5x = t^2$ will be slightly off, whereas

> "we could find a polynomial of seven terms
>
> $$x = a_0 + a_1 t + a_2 t^2 + a_3 t^3 + a_4 t^4 + a_5 t^5 + a_6 t^6$$
>
> that would fit the observations exactly. Nevertheless the physicist would still use the square law. (...) [the physicist's] predilection for the simple law is so strong that he will retain it when it does not satisfy the observations exactly, in spite of the existence of more complex laws that do satisfy them exactly. He would apply the law to predict the value of x for $t = 60$ sec. and would expect the result to be right within a few centimetres, provided the plane was long enough to permit the displacement required. He would, on the other hand, expect the polynomial of seven terms to give a seriously wrong result when extrapolated to such an extent." (Jeffreys 1973, pp. 62-63)

The above considerations suggest that there is a trade-off between *goodness-of-fit* and *model complexity*. If we prefer the model that fits the sample data best, we will always select the most complex model. For instance, a model with as many free parameters as there are data points will be able to describe the sample data perfectly. But we do not want a model that perfectly *fits* the present data. Instead, we want a model that best *predicts* future data: we want to *extrapolate* and *generalize* (e.g., Myung and Pitt 1997, Myung 2000, Pitt and Myung 2002). Schematically, we have

$$\underbrace{\text{Generalizability}}_{\text{Fit to future data}} = \underbrace{\text{Goodness-of-Fit}}_{\text{Fit to present data}} - \underbrace{\text{Model Complexity}}_{\text{Data-fitting capacity}}. \tag{18.1}$$

This 'equation' conveys that generalizability is highest when a good fit to the present data is achieved with a model that is relatively simple. It will be always possible to achieve an even better fit with a more complex model, but when the gain in fit is smaller than the increase in complexity, generalizability suffers. As we will see in the next chapters, the Wrinch-Jeffreys methodology allows us to navigate the fit-complexity trade-off as an automatic by-product of Bayesian inference.

The Goldilocks Fit

Empirical data are usually understood to consist of a mix of signal and noise (Silver 2012). The *signal* is the part that is structural, replicable, systematic, and predictable. The *noise* is the part that is idiosyncratic, that is, an unknown consequence of the specific setting in which the experiment was conducted. For instance, when Galileo operated the *klepsydra* his observations will have been determined to some extent by momentarily lapses of attention. This is a source of measurement error – its effects have nothing to do with the forces of gravity. By definition, fluctuations due to noise are not replicable and not predictable. To drive the point home:

$$\text{Data} = \underbrace{\text{Signal}}_{\text{Replicable}} + \underbrace{\text{Noise}}_{\text{Idiosyncratic}}.$$

The trade-off between goodness-of-fit and parsimony implies that there is a sweet spot (the so-called *Goldilocks fit*) where a statistical model is sufficiently complex to extract most of the replicable patterns in the data while sufficiently simple to ignore the idiosyncratic noise. This way the *Goldilocks model* achieves optimal predictive performance. Margin-figure 18.2 provides an example using Jeffreys's fictitious data set with measurement error. The top panel shows the fit of a linear model. This linear model is parsimonious but it fails to account for systematic, replicable patterns in the data. The model fails – it is too simple and *underfits* the data. The middle panel shows the fit of a high-order polynomial model. This model is not parsimonious but it does account for the sample data perfectly. Unfortunately, the model is so flexible that it tunes its many parameters not just to the systematic, replicable patterns, but also to the idiosyncratic measurement noise. This model also fails – it is too complex and *overfits* the data. The bottom panel shows the quadratic model. This model is more complex than the linear model, allowing it to capture the systematic effect of

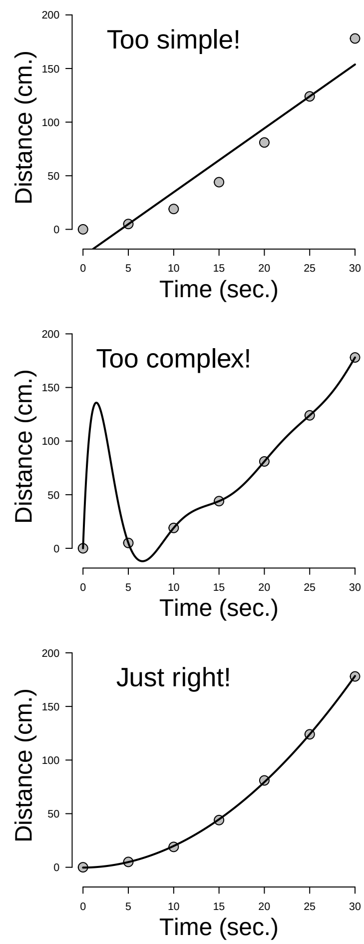


Figure 18.2: A Goldilocks fit to the noisy data from the fictitious physics experiment described by Jeffreys (1973). In the top panel, the model is too simple (i.e., it underfits the data and misses replicable signal); in the middle panel, the model is too complex (i.e., it overfits the data and mistakes idiosyncratic noise for replicable signal); in the bottom panel, the model is as complex as it needs to be to separate noise from signal to thereby achieve optimal predictive performance.

constant acceleration; at the same time, the model is less complex than the high-order polynomial, allowing it correctly to treat measurement error as irreproducible noise (Vandekerckhove et al. 2015).

Overfitting in Practice

In practical applications, underfitting may be easier to detect than overfitting. Models that underfit are incapable of accounting for important aspects of the data, as is demonstrated in the top panel of Figure 18.2. In contrast, models that overfit rarely produce the wild wiggliness that is on display in the middle panel of Figure 18.2. Instead, models that overfit the data usually mimic the Goldilocks model by producing a similar fit within the range of the data.

This phenomenon is illustrated in Figure 18.3, again with the noisy data from the fictitious physics experiment reported by Jeffreys (1973).

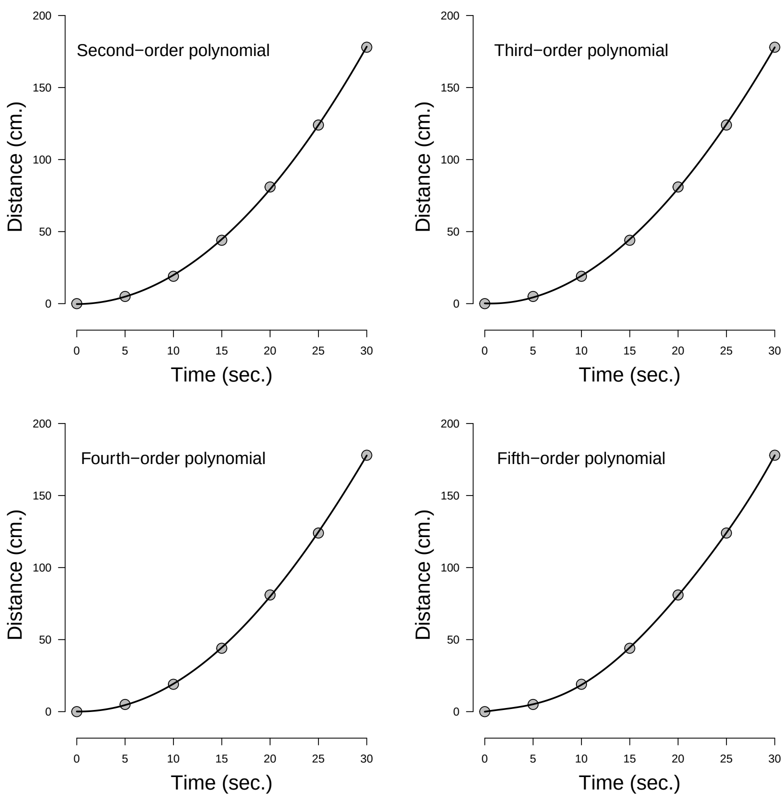


Figure 18.3: The problem with detecting overfitting as illustrated with the fictitious physics experiment described by Jeffreys (1973). Noisy data originate from the quadratic law $5x = t^2$. The top left panel shows the best fit of a second-order polynomial (i.e., $x = a_0 + a_1 t + a_2 t^2$), the top right panel shows the best fit of a third-order polynomial (i.e., $x = a_0 + a_1 t + a_2 t^2 + a_3 t^3$), the bottom left panel shows the best fit of a fourth-order polynomial (i.e., $x = a_0 + a_1 t + a_2 t^2 + a_3 t^3 + a_4 t^4$), and the bottom right panel shows the best fit of a fifth-order polynomial (i.e., $x = a_0 + a_1 t + a_2 t^2 + a_3 t^3 + a_4 t^4 + a_5 t^5$).

Each panel shows the fit of a polynomial: a second-order polynomial for the top left panel, a third-order polynomial for the top right panel, a fourth-order polynomial for the bottom left panel, and a fifth-order polynomial for the bottom right panel. It is immediately clear that even the fifth-order polynomial –which is much more complex than needed– provides an account that closely resembles that of the second-order polynomial.

From a Bayesian perspective, there is a good reason why overly complex models such as the fifth-order polynomial can mimic the performance of the Goldilocks model (i.e., the second-order polynomial): the concept of 'fit' is misleading, at least when it comes to model comparison. In the example from Figure 18.3, the 'fit' does not refer to the overall or average ability of the models to account for the data. Instead, the fit shown is for a single set of parameter values (within each of the models) that were cherry-picked because they produced the *best* account of the data. Specifically, the best-performing parameter values were determined by a 'least-squares' fitting routine that finds the single parameter vector with the smallest squared deviation between the observed data and the prediction. The 'predictions' from this parameter vector are then singled out and presented as 'the' fit of the model, conveniently ignoring the earlier parameter selection process. It is not surprising that the resulting performance is not representative of the model's overall predictive performance (cf. Pitt and Myung 2002).[3]

To stress this important point, suppose you are an investor and you are uncertain whether to do business with stockbroker firm *Monkey Business* or *Win-Win*. The firm *Monkey Business* employs 20 brokers, whereas *Win-Win* employs 100 brokers; your goal is to identify the firm with the most expertise. Both companies agree to provide you with information on the predictive performance of their brokers over the past year. *Win-Win* proposes that, as 'goodness-of-fit' for the entire firm, you consider the predictive performance of their single best-predicting stockbroker. *Monkey Business* disagrees and argues that a fairer assessment of a firm's success is obtained by *averaging* the predictions across all brokers under employ. We hope you agree with *Monkey Business*. With enough brokers under employ, the performance of the single best broker –selected after the fact– will simultaneously be spectacularly good and spectacularly unrepresentative.[4]

Table 18.1 shows the best-fitting parameter values of the four polynomials (as per usual, these values are denoted by placing a 'hat' above the parameter names, so $\hat{a}_0$ represents the best-fitting parameter value for the intercept). The true relationship, $5x = t^2$, is shown in the top row. Ideally, the rival polynomials would yield $\hat{a}_2 = 0.20$, and estimate the remaining (redundant) parameters to be zero exactly. To interpret these estimates correctly, Table 18.1 also shows the *standard errors* associated

[3] It can nonetheless be informative to inspect the best fit. For instance, if even the best fit is poor then this implies that the model is misspecified and may underfit the data. And if the best fit is excellent this implies that at least some parameter values are able to provide a good account of the data.

[4] We will later see that the Bayesian solution to the trade-off between fit and complexity basically involves the solution proposed by *Monkey Business*, that is, to determine success by averaging over all brokers of a particular firm (i.e., all parameter values of a particular model).

with each estimate. Briefly, a standard error indicates the precision associated with a parameter estimate; it is the frequentist equivalent of the standard deviation of the posterior distribution.

Table 18.1: Parameter point estimates $\hat{a}_i$ (and their associated standard errors underneath, in brackets) for four polynomial models fit to the data from Jeffreys's fictitious physics experiment. Corresponding model fits are displayed in Figure 18.3. The true model is $5x = t^2$, so $a_2 = 0.20$. Model $\mathcal{M}_j$ denotes a polynomial of order j.

	$\hat{a}_0$	$\hat{a}_1$	$\hat{a}_2$	$\hat{a}_3$	$\hat{a}_4$	$\hat{a}_5$
Truth	–	–	0.20	–	–	–
$\mathcal{M}_2$	−0.29	0.04	0.20	–	–	–
	(0.87)	(0.14)	(0.00)	–	–	–
$\mathcal{M}_3$	0.21	−0.29	0.23	−0.00	–	–
	(0.88)	(0.28)	(0.02)	(0.00)	–	–
$\mathcal{M}_4$	0.12	−0.10	0.19	0.00	0.00	–
	(1.07)	(0.61)	(0.09)	(0.00)	(0.00)	–
$\mathcal{M}_5$	−0.03	1.05	−0.14	0.03	−0.00	0.00
	(0.79)	(0.82)	(0.21)	(0.02)	(0.00)	(0.00)

Note. All values are rounded to two decimals, including 0.00 and -0.00.

Jeffreys's scenario features a straightforward signal accompanied by very little idiosyncratic noise. With so little noise, the complex model does not have much to overinterpret, and it will therefore closely mimic the Goldilocks model. But this mimicry does come at a cost. To see this, consider the column for $\hat{a}_2$ in Table 18.1. The true value is 0.20, and the quadratic model $\mathcal{M}_2$ correctly recovers it (i.e., $\hat{a}_2 = 0.20$), and does so with great precision – the standard error is 0.004. However, as the number of polynomial parameters grows, the standard error gradually increases (i.e., 0.02 for $\mathcal{M}_3$, 0.09 for $\mathcal{M}_4$, and 0.21 for $\mathcal{M}_5$). In other words, the inclusion of redundant parameters decreases the precision with which the relevant parameters can be estimated.[5] When the true value is 0.20, it is obviously better to report an estimate of 0.20 with a standard error of 0.004 than it is to report an estimate of -0.14 with a standard error of 0.21.

[5] This is often referred to as the *bias-variance trade-off*.

There are other problems with needlessly complex models as well. For instance, if we adopt $\mathcal{M}_5$, why not adopt a model that is even more complex? Ultimately we end up with an infinitely complex model (or at least a model with as many parameters as there are data points) which makes the model meaningless – it neither summarizes the data nor allows good predictions. Moreover, the generalization of the complex model will fail when the predictions are extrapolated far enough outside the range of the observed data. This reflects the fact that the correct model for Galileo's experiment is simply *not* a fifth-order polynomial. Finally, choosing a needlessly complex model exposes the inexperienced scientist to ridicule. Scientists prefer the simple model whenever the data do not provide strong grounds for adopting a more complex one.[6]

The situation changes when we add measurement error to Jeffreys's data. Specifically, consider the following fictitious series of observations:

time t (sec.)	0	5	10	15	20	25	30
displacement x (cm.)	0	5	5	30	95	110	150

The data and associated polynomial best-fits are shown in Figure 18.4. In contrast to the low-noise scenario discussed earlier, the more complex models no longer mimic the behavior of the second-order polynomial. With more noise in play, the complex models are able to describe the idiosyncratic fluctuations in terms of their best-fitting parameter values. Because these best-fitting parameter values are based on pure noise the complex models will generalize poorly, even if they are tested on new data that fall within the range of the observed data. For instance, consider a replication experiment that measures displacement for times $t = \{1, 2, 3, 4, 5\}$ seconds. Models $\mathcal{M}_3$ and $\mathcal{M}_4$ predict the ball to move *up* the ramp, whereas model $\mathcal{M}_5$ predicts the ball to move down the ramp first, and then up again. There predictions are preposterous.

[6] "We could thus see no reason why we should not solve DNA in the same way. All we had to do was to construct a set of molecular models and begin to play–with luck, the structure would be a helix. Any other type of configuration would be much more complicated. *Worrying about complications before ruling out the possibility that the answer was simple would have been damned foolishness.* (Watson 1968, pp. 47-48; italics added for emphasis).

Table 18.2: Parameter point estimates $\hat{a}_i$ (and their associated standard errors underneath, in brackets) for four polynomial models fit to the data from Jeffreys's fictitious physics experiment with extra measurement noise. Corresponding model fits are displayed in Figure 18.4. The true model is $5x = t^2$, so $a_2 = 0.20$. Model $\mathcal{M}_j$ denotes a polynomial of order j. R^2 denotes the proportion of variance explained (i.e., a measure of goodness-of-fit).

	$\hat{a}_0$	$\hat{a}_1$	$\hat{a}_2$	$\hat{a}_3$	$\hat{a}_4$	$\hat{a}_5$	R^2
Truth	–	–	0.20	–	–	–	
$\mathcal{M}_2$	-4.29	0.64	0.16	–	–	–	0.96
	(13.09)	(2.04)	(0.07)	–	–	–	
$\mathcal{M}_3$	3.21	-4.36	0.61	-0.01	–	–	0.97
	(13.19)	(4.17)	(0.34)	(0.01)	–	–	
$\mathcal{M}_4$	1.75	-1.48	0.10	0.02	-0.00	–	0.98
	(16.11)	(9.15)	(1.41)	(0.07)	(0.00)	–	
$\mathcal{M}_5$	-0.39	15.76	-4.85	0.49	-0.02	0.00	0.99
	(11.84)	(12.36)	(3.16)	(0.29)	(0.01)	(0.00)	

Note. All values are rounded to two decimals, including 0.00 and -0.00.

Table 18.2 shows the parameter estimates and associated standard errors. Compared to the low-noise results shown in Table 18.1 it is evi-

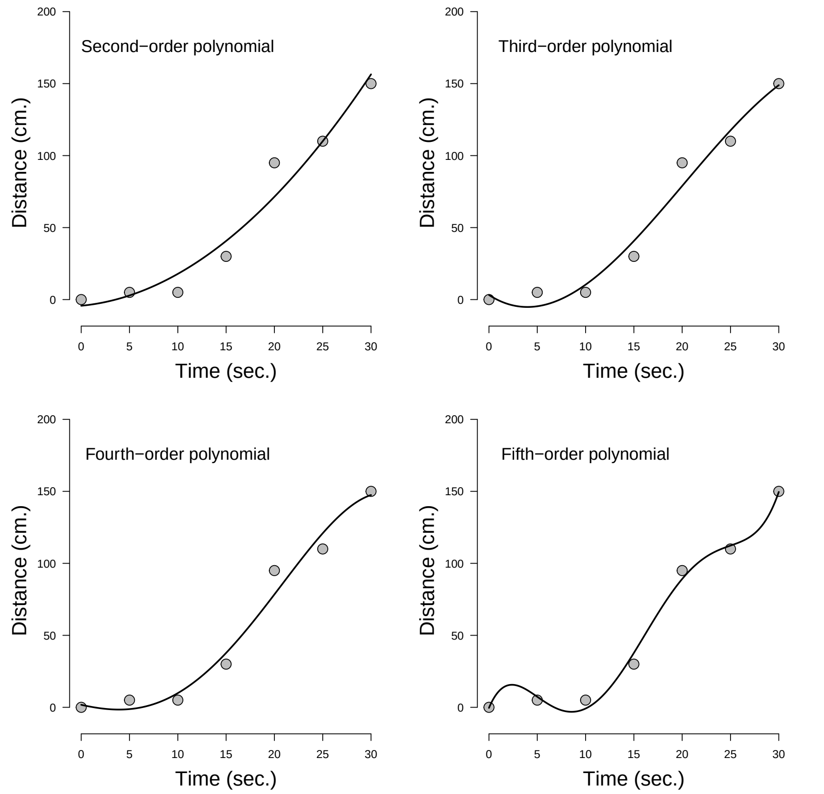


Figure 18.4: Polynomial fits to data from the fictitious physics experiment described by Jeffreys (1973), but with extra measurement noise. Noisy data originate from the quadratic law $5x = t^2$. The top left panel shows the best fit of a second-order polynomial (i.e., $x = a_0 + a_1 t + a_2 t^2$), the top right panel shows the best fit of a third-order polynomial (i.e., $x = a_0 + a_1 t + a_2 t^2 + a_3 t^3$), the bottom left panel shows the best fit of a fourth-order polynomial (i.e., $x = a_0 + a_1 t + a_2 t^2 + a_3 t^3 + a_4 t^4$), and the bottom right panel shows the best fit of a fifth-order polynomial (i.e., $x = a_0 + a_1 t + a_2 t^2 + a_3 t^3 + a_4 t^4 + a_5 t^5$).

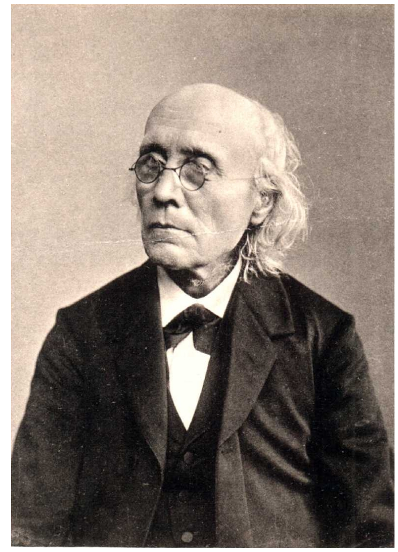
dent that the addition of measurement noise has decreased the precision of the estimates (i.e., the standard errors have increased considerably). The estimate of $\hat{\alpha}_2$ under $\mathcal{M}_2$ is still within one standard error of the true value of 0.20 (i.e., 0.16 ± 0.07). As before, more complex models have higher standard errors for $\hat{\alpha}_2$ (i.e., 0.34 for $\mathcal{M}_3$, 1.41 for $\mathcal{M}_4$, and 3.16 for $\mathcal{M}_5$). Also note that R^2, the proportion of explained variance, *increases* as the models become more complex: $R^2 = 0.96$ for $\mathcal{M}_2$ which steadily increases to $R^2 = 0.99$ for $\mathcal{M}_5$. In other words, the more complex the model, the more impressive its best-fit to the sample data.[7] This is also visually apparent from Figure 18.4: in terms of its deviation from the sample observations, the fifth-order polynomial does better than the second-order polynomial. This underscores the fact that when we evaluate the performance of rival statistical models we need to go beyond best-fit to the sample data and consider generalizability instead.

[7] This is also the case for the low-noise scenario discussed earlier. We did not show the R^2 values then because they were nearly 1, indicating a perfect fit.

Two Examples from Psychology

Across the empirical sciences, researchers attach great importance to parsimony. To demonstrate this point we leave Galileo's bronze balls and turn to psychology instead.

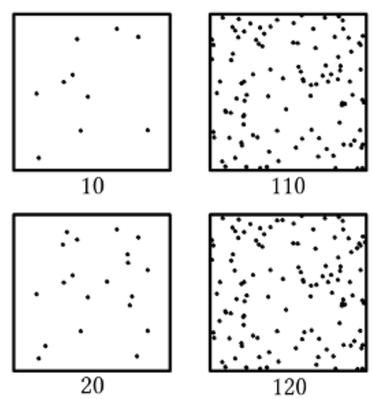
As a first example we consider the relation between physical intensity I and subjective experience Ψ. For instance, participants in a psychophysical experiment may be asked to judge the subjectively experienced intensity of a briefly flashed light. As the physical intensity I of the flash increases, so does the subjective experience Ψ – but what is the function that relates I to Ψ?



Gustav Theodor Fechner (1801–1887), experimental psychologist avant la lettre. His 1860 book *Elemente der Psychophysik* (Elements of Psychophysics) created the field of psychophysics.

The most famous proposal for the relation between I and Ψ is known as the Weber-Fechner law, or just *Fechner's law*. Fechner's law states that $\Psi = k \ln(I - a)$; in words, subjective experience Ψ is a negatively accelerating (i.e., logarithmic) function of physical intensity I. As mathematician Ian Stewart eloquently explains:

> "If we look at a light, the brightness that we perceive varies as the logarithm of the actual energy output. If one source is ten times as bright as another, then the difference we perceive is constant, however bright the two sources really are. The same goes for the loudness of sounds: a bang with ten times a much energy sounds a fixed amount louder.
>
> (...) Evolution pretty much had to come up with something like a logarithmic scale, because the external world presents our senses with stimuli over a huge range of sizes. A noise may be a little more than a mouse scuttling in the hedgerow, or it may be a clap of thunder; we need to be able to hear both. But the range of sound levels is so vast that no biological sensory device can respond in proportion to the energy generated by the sound. If an ear that could hear the mouse did that, then a thunderclap would destroy it. If it tuned the sound levels down so that the thunderclap produced a comfortable signal, it wouldn't be able to hear the mouse. The solution is to compress the energy levels into a comfortable range, and the logarithm does exactly that. Being sensitive to proportions rather than absolutes makes excellent sense, and makes for excellent senses." (Stewart 2012, pp. 33-34)



"An illustration of the Weber–Fechner law. On each side, the lower square contains 10 more dots than the upper one. However the perception is different: On the left side, the difference between upper and lower square is clearly visible. On the right side, the two squares look almost the same." Text and figure from MrPomidor.

The left panel of Figure 18.5 shows three instances of Fechner's law. It is clear that Fechner's law is relatively simple. Despite the fact that the law features the two free parameters k and a, it can only ever account for curves that are negatively accelerating. Fechner's law is parsimonious because it makes daring predictions.

In the 1950's, Stanley Smith Stevens (1906–1973) proposed a rival psychophysical law. *Stevens's law* also relates I to Ψ, but through a power function: $\Psi = k\, I^b$. Stevens's law is considered less parsimonious than Fechner's law (cf. Lee and Wagenmakers 2013, Myung and Pitt 1997, Townsend 1975). The reason is obviously not in the number of free parameters (both laws have two), but in the effect that the parameters can exert on the shape of the function – that is, the effect on

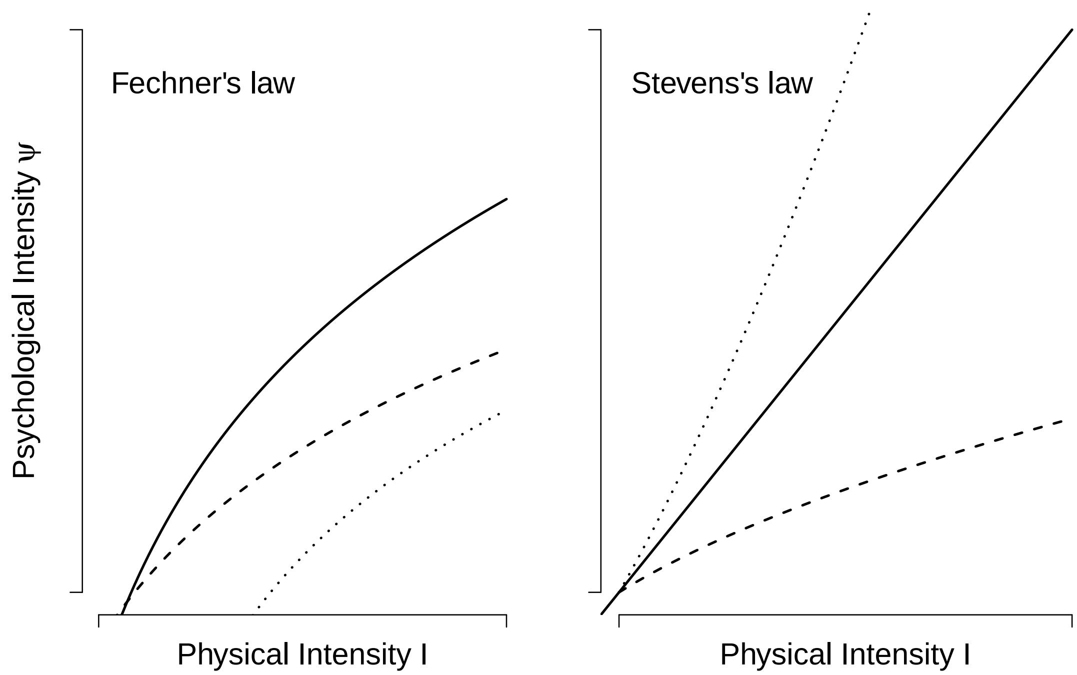


Figure 18.5: Parsimony in psychophysics. The left panel shows three examples of Fechner's law, according to which subjectively experienced intensity Ψ is a negatively accelerated function of physical intensity I. The right panel shows three examples of Stevens's law, according to which subjectively experienced intensity Ψ relates to physical intensity I either as a negatively accelerated function (i.e., the dashed line), a constantly accelerating function (i.e., the solid line), or a positively accelerating function (i.e., the dotted line), depending on the parameter values. Fechner's law is less flexible than Stevens's law, because it can only account for one particular pattern of results – in other words, the predictions from Fechner's law are riskier and less vague.

predictions. Specifically, when $b < 1$ Stevens's law produces negatively accelerating curves; when $b = 1$ Stevens's law produces a constantly accelerating curve (i.e., a straight line); and when $b > 1$ Stevens' law produces positively accelerated curves. This is illustrated in the right panel of Figure 18.5 (cf. Stevens 1975, Figure 5; Stevens 1961).

Townsend (1975, p. 213) remarks "With regard to degree of precision, Fechner's predicted psychophysical function makes a stronger statement about the world than does that relationship described by Stevens. (...) by choosing b greater than or less than 1, one can make the function positively or negatively accelerated without affecting the sign of the first derivative, whereas we are constrained to a negatively accelerated function with the logarithmic expression as long as we demand (as we must) that the function be monotonic increasing". Similarly, Myung and Pitt (1997, p. 82) write "(...) psychological and physical dimensions are assumed to be related by a power function in Stevens's law, making it capable of fitting data that have negative, positive, and zero curvature. Fechner's law assumes a logarithmic relationship, which can fit data patterns with a negative curvature only." In other words, Fechner's law is more parsimonious than Stevens's law.[8]

How strong is the preference that scientists have for Fechner's relatively simple logarithmic law over Stevens's relatively complex power law? To gauge this, imagine that the only data sets at our disposal show a negatively accelerated curve.[9] In this hypothetical scenario, the following would be true:

- If Fechner's law had already been proposed, no serious scientist would ever propose Stevens's law as a rival hypothesis. There would simply be no point.
- If a serious scientist were nonetheless to propose Stevens's law as a rival to Fechner's law, this would have to be because of a strong expectation that data violating Fechner's law can be demonstrated in a concrete experiment.
- Most scientists would nevertheless retain Fechner's law until such a concrete experiment had actually been conducted and the results were shown to be inconsistent with Fechner's law but consistent with Stevens's law. And in fact, Stevens proposed his law only because the empirical data suggested it. For instance, Stevens found that a value of $b = 0.33$ is typical for the assessment of brightness and yields a negatively accelerating curve, consistent with Fechner's law. But the value of $b = 1$ yields a straight line –inconsistent with Fechner's law– and is characteristic for the assessment of repetition rate; furthermore, the value of $b = 3.5$ yields a positively accelerating curve –even more inconsistent with Fechner's law– and is typical for assessment of electric current running through the fingers (for these and other examples see Stevens 1961, Table 1).
- If Stevens's law had been proposed first – well, the immediate question is whether this would even happen. A serious scientist, confronted exclusively with negatively accelerating psychophysical curves, would not turn first to the power functions. Or if the scientist would propose a power function form, it would be under the implicit or explicit restriction that $b < 1$.

To further underscore the importance of parsimony in the field of psychology we turn to the drift diffusion model (DDM; Ratcliff 1978). The DDM provides an account of how people process noisy information in order to make a speeded decision between two response options. Figure 18.6 shows an application of the DDM to the popular *lexical decision task* (Meyer and Schvaneveldt 1971). In this task, participants are confronted with letter strings that they have to categorize quickly –usually

[8] Fechner's law is in fact a special case of Stevens's law (Kvålseth 1992). Additional theoretical reflections can be found in MacKay (1963). See the final exercise in this chapter for a Bayesian warning against the blanket statement that Fechner's law is less parsimonious than Stevens's law.

[9] This situation is analogous to that shown in Figure 18.3, where the data are consistent with the simple second-order polynomial.

by pressing one of two response buttons on a computer keyboard with their index finger– as being either words (e.g., *table*) or 'nonwords' (e.g., *drapa*). The speed and accuracy of the classifications are thought to measure how efficiently participants can access lexical representations stored in memory. For instance, words that occur relatively often (i.e., high-frequency words such as *grass*) are classified faster and with fewer mistakes than low-frequency words such as *harpy*.

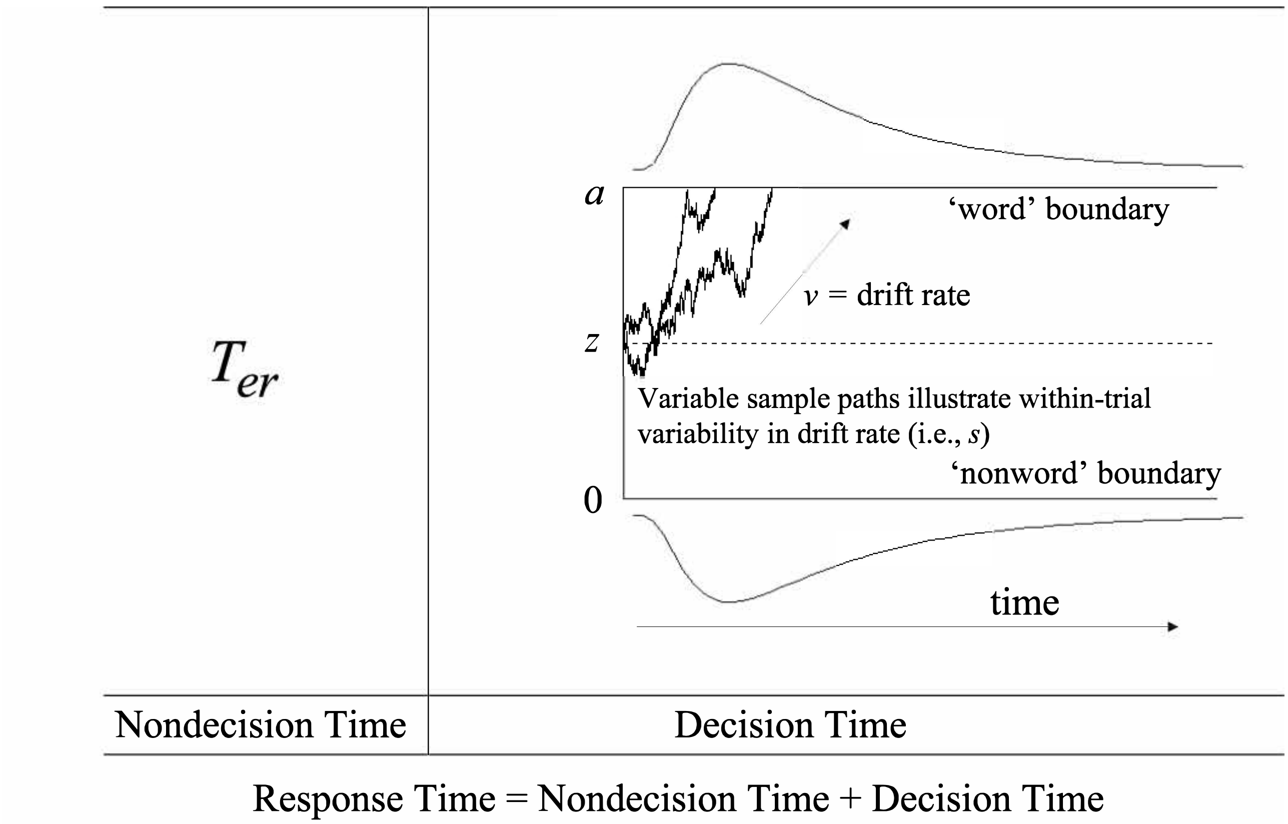


Figure 18.6: A simplified drift diffusion model as applied to lexical decision (cf. Wagenmakers 2009). Noisy information is accumulated until a threshold level of evidence is reached, which then triggers the associated response. The quality of information processing is measured by drift rate v, whereas response caution is quantified by the distance between the response boundaries. The right-skewed densities near the two response boundaries visualize the shape of the predicted response time distributions. Bias favoring the 'word' or 'nonword' response is accounted for by starting point z, and nondecision time (i.e., encoding and response execution) is given by T_{er}.

However, the interpretation of performance on the lexical decision task is frustrated by the fact that participants can *trade speed for accuracy*. That is, participants can choose to adopt a more cautious attitude and collect more information before committing to a decision – and by doing so, they will slow down but also make fewer mistakes. It would be desirable to have a measure of cognitive processing that is independent of such strategic behavior, and this is exactly what the DDM delivers.

The basic structure of the DDM is shown in Figure 18.6. For every individual decision, the DDM assumes that the observed response time is given by the sum of a nondecision component (i.e., T_{er}, the time associated with encoding and response processes that take place regardless of what choice is made) and a decision component, which is the main focus of the DDM. The decision component is characterized by the accumulation of noisy information until a threshold of evidence is reached, after which the corresponding decision is initiated. High absolute values of drift rate v result in low-noise accumulation processes – a quick march to the correct boundary. On the other hand, low absolute values of v result in high-noise accumulation processes – a slow, meandering trajectory that often terminates at the incorrect boundary. The DDM parameter v therefore captures the efficacy of the information accumulation process. In contrast, the DDM parameter a –the distance between the two response boundaries– governs the strategic tradeoff between speed and accuracy. Specifically, participants who are relatively cautious will adopt a boundary separation that is relatively high, making responses slow but relatively accurate (because relatively insensitive to chance fluctuations). Prior preference for either the 'word' or 'nonword' decision is quantified by the starting point parameter z (Mulder et al. 2012). Finally, Figure 18.6 shows the predicted response time densities next to the response threshold.[10]

[10] NB. These are predictions for data, not prior or posterior distributions of uncertainty about a model parameter.

In sum, the DDM can be used to decompose observed performance (i.e., response speed and accuracy) into hypothesized psychological processes such as the quality of information processing and response caution. Across numerous applications, Roger Ratcliff and Gail McKoon demonstrated that (a) the DDM often provides an excellent account of the data; (b) the DDM offers insights that go beyond what can be accomplished with a direct evaluation of response time and accuracy.

The DDM model shown in Figure 18.6 makes a number of risky predictions (cf. Ratcliff 2002). For instance, the model predicts that response time distributions are *always* right-skewed, and that the skew will *always* increase when z decreases toward zero. When the starting point is unbiased (i.e., $z = a/2$), the DDM from Figure 18.6 makes another risky prediction: correct responses are just as fast as errors, that is, the predicted response time distribution is the same for corrects and errors.

Now consider an alternative to the simple DDM which posits that (a) starting point z varies from one trial to the next, which leads to the prediction that errors are *faster* than correct responses; (b) drift rate v varies from one trial to the next, which leads to the prediction that errors are *slower* than correct responses (for an explanation see Ratcliff and Rouder 1998, Figure 2). Let's call the model that adds these two across-trial variabilities the 'complex DDM'. By changing its parameter values, the complex DDM can account for slow errors, for fast errors, and for errors and correct responses that are equally fast. It therefore makes predictions that are more vague that those from the simple DDM shown in Figure 18.6.

Similar to our discussion of Fechner's law vs. Stevens's law above, let's assume that real data would consistently show that error responses are about as fast as correct responses. This would mean the same as before:

- No serious scientist would dare propose the complex DDM.
- The only reason for entertaining the complex DDM would be the strong expectation that data can be found that go against the simple DDM and can be accounted for by the complex DDM.
- Until these data are reported, many researchers would retain the simple DDM. In fact, the simple DDM would receive compelling support from the data, as rival models of response time generally cannot account for the phenomenon that errors and corrects are equally fast. The complex DDM with its across-trial variability is now accepted as the standard model of response time, but –just as in the case of Stevens's law– this has happened because the empirical data effectively necessitated the addition of the across-trial variabilities. For instance, errors are usually slower than correct responses in the lexical decision task; the reverse holds in simple perceptual tasks, especially when speed is stressed. And even within the same task, errors can be either slow or fast depending on the level of speed stress (e.g., Wagenmakers et al. 2008).

The examples on psychophysics and speeded decision making both underscore that researchers strongly prefer simple models: they are the first models that are proposed and evaluated, and researchers demand compelling empirical evidence before they feel forced to make their models more complex by adding processes or parameters. No serious scientist would propose a complex model as a worthwhile alternative when the data are consistent with the simple model. The progression from simple to complex models is one that scientists engage in reluctantly, and only because they feel the data leave them no choice.

Ockham's Razor

"Everything should be made as simple as possible, but no simpler." (Albert Einstein).

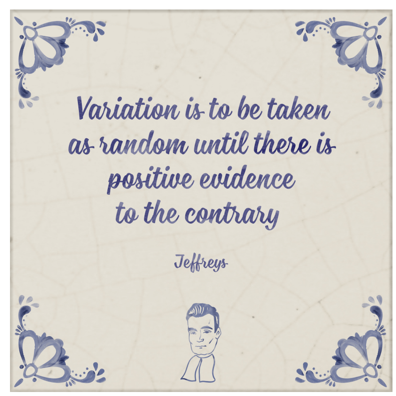
No treatment of parsimony is complete without a discussion of Ockham's razor. Ockham's razor is virtually synonymous with the principle of parsimony. The metaphorical razor cuts away all theorizing that is needlessly complex; the razor therefore embodies a preference for assumptions, theories, and hypotheses that are as simple as possible without being false. The razor is named after the English logician and Franciscan friar Father William of Ockham (c.1288-c.1348), who stated

"Pluralitas non est ponenda sine necessitate" (Plurality should not be assumed without necessity), and "Frustra fit per plura quod potest fieri per pauciora" (It is futile to do with more what can be done with fewer). Indeed, it is not an exaggeration to state that the crucial difference between Laplacean learning (the topic of Part II of this book) and Jeffreyian learning is that only the latter respects Ockham's razor. Indeed, Jeffreys was quite explicit about the importance of Ockham's razor:

> "The best way of testing differences from a systematic rule is always to arrange our work so as to ask and answer one question at a time. Thus William of Ockham's rule,‡ 'Entities are not to be multiplied without necessity' achieves for scientific purposes a precise and practically applicable form: *Variation is random until the contrary is shown; and new parameters in laws, when they are suggested, must be tested one at a time unless there is specific reason to the contrary.* [italics in original] (Jeffreys 1961, p. 342; see also Jeffreys 1937c, pp. 489-490 and Jeffreys 1938e, p. 716; cf. Poincaré 1913)

‡William of Ockham (d. 1349 ?), known as the Invincible Doctor and the Venerable Inceptor, was a remarkable man. He proved the reigning Pope guilty of seventy errors and seven heresies, and apparently died at Munich with so little attendant ceremony that there is even a doubt about the year. (...) The above form of the principle, known as Ockham's Razor, was first given by John Ponce of Cork in 1639. Ockham and a number of contemporaries, however, had made equivalent statements. A historical treatment is given by W. M. Thorburn, *Mind*, **27**, 1918, 345-53.

Ockham, however, was far from the first to articulate the razor. Indeed, the central idea goes back to Aristotle and Ptolemy. For instance, Aristotle stated "Altogether it is better to make your basic things fewer and limited, like Empedocles." (Aristotle 350BC/1970, p. 10), and Ptolemy wrote "We consider it a good principle to explain the phenomena by the simplest hypotheses possible." Readers curious to learn more about William Ockham may consult the 1402-page tome *William Ockham* (Adams 1987). We summarize some of the highlights here:



Jeffreys's razor. Figure available at `BayesianSpectacles.org` under a CC-BY license.

1. Ockham fell victim to his own razor: "Ultimately, Ockham gave up the objective-existence theory–both where thoughts of particulars and thoughts of universals are concerned–because Walter Chatton convinced him that the objective-existence theory violated the principle of parsimony better known now as Ockham's Razor." (p. 102)

2. Ockham's most explicit description of his razor is: "No plurality should be assumed unless it can be proved by reason, or by experience, or by some infallible authority" (pp. 156-157; p. 1008), or, in the original Latin: "Nulla pluralitas est ponenda nisi per rationem vel experientiam vel auctoritatem illius, qui non potest falli nec errare, potest convinci." The overlap between this statement and those by Jeffreys is striking.

3. Despite the fact that (a) the principle of parsimony goes back at least to Aristotle; (b) other medieval scholars invoked the principle of parsimony before Ockham (e.g., John Duns Scotus, Peter Auriol, and Thomas Aquinas; see Ariew 1977); (c) Ockham did not justify the principle of parsimony;[11] (d) Ockham primarily used other arguments – despite these considerations, Adams argues that the association of the razor with Ockham is nevertheless appropriate because

[11] Adams remarks that this is not really surprising, because "contemporary philosophers of science are convinced that simplicity is a legitimate criterion against which to judge the adequacy of theories, but they are hard pressed to explain why or even to say what they mean by simplicity!" (p. 160)

"in comparison with his predecessors, Ockham's metaphysical conclusions are what one would expect from a philosopher who let (D)-(G) [Ockham's statements about parsimony – EWDM] be his guide." (p. 157; but see Ariew 1977 for the opposite opinion)

4. Adams argues that according to Ockham, "So far as the order of salvation is concerned, God does not abide by the principle of parsimony" (p. 159)

5. Ockham uses his razor to provide a "persuasive argument" that the matter of the heavens is of the same kind as the matter of things on earth: "...plurality should never be assumed without necessity, as has often been said. But now there is no apparent necessity in supposing that the matter here and there are of different kinds. For whatever can be saved by different kinds of matter can be saved equally well or better by matter of the same kind." (pp. 160-161)

Figure 18.7: William of Ockham (c.1288-c.1348) as depicted on a stained glass window at a church in Surrey.

Note the similarity to a general principle of law known as *affirmanti incumbit probatio*: the onus of proof is on the person who makes an assertion.

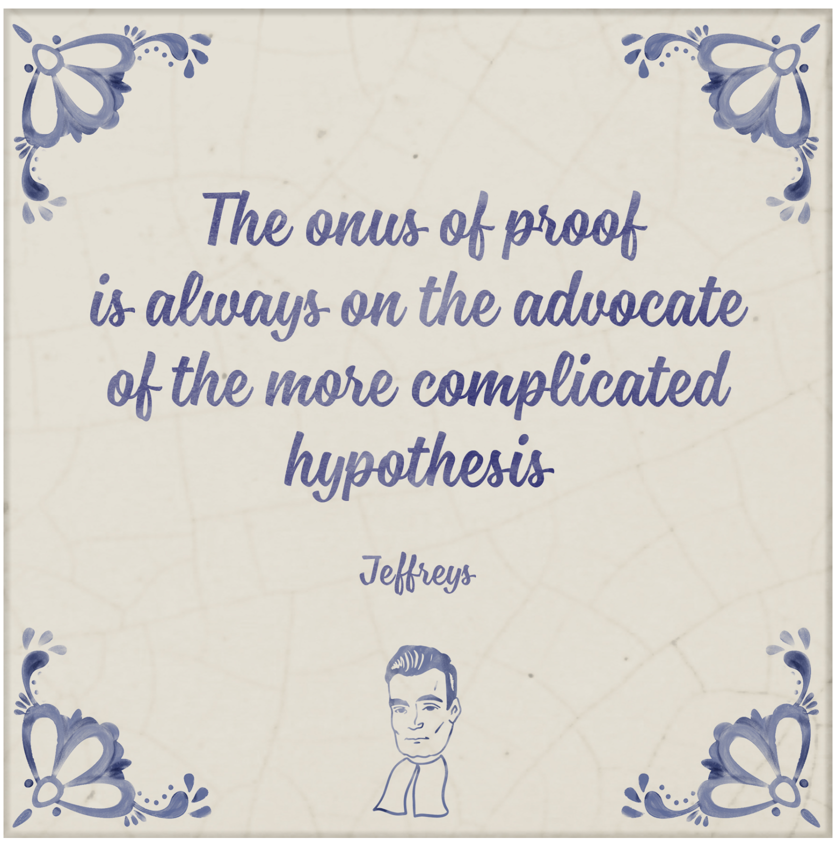


Figure available at `BayesianSpectacles.org` under a CC-BY license.

Exercises

1. When discussing the right panel in Figure 18.1 we stated that "Scientists have a strong preference for the simple equation." This was an understatement – the complex equation violates the laws of the universe. Why?

2. Out of the models listed in Table 18.1, which one provides the best fit to the data?

3. Consider again the stockbroker firms *Monkey Business* (with 20 brokers) and *Win-Win* (with 100 brokers). *Monkey Business* argued that a firm's success should be assessed by averaging performance across all brokers, not by singling out the one broker who happened to perform best. *Win-Win* argues that they distribute the work according to past performance, such that more work will be performed by brokers that do well. At the end of the year, almost all of the work will be done by the single broker that outperformed the others, so that this broker is in fact representative for the entire firm.[12] Pretend that you are the CEO of *Monkey Business* and write a short response.

4. Revisit Fechner's law and Stevens's law of psychophysics and (1) explain why data qualitatively consistent with both Fechner's law and Stevens's law increase the plausibility of the former and decrease the plausibility of the latter; (2) explain why Stevens's law is not necessarily less parsimonious than Fechner's law; (3) draw a comparison between models of psychophysics and the Goldilocks demonstration from Margin-figure 18.2.

5. Consider again the drift diffusion model shown in Figure 18.6. What qualitative similarities do you see with the process of Bayesian inference?

[12] This is analogous to the process of Bayesian estimation, where parameter values that predict relatively well gain plausibility at the expense of those that predict poorly.

Chapter Summary

We demonstrated the appeal of parsimonious models by fitting fictitious data from a simple physics experiment in which a ball rolls down a ramp. The relation between time and distance is of interest, and we considered the account provided by several polynomial models. The example may have appeared trivial in the sense that scientists would prefer the simple second-order polynomial model over the more complex higher-order polynomial models, *without any hesitation whatsoever*, even when these complex models provide a better fit to the sample data.[13] Two examples from psychology reinforced the general message: researchers are reluctant to make their models more complex, and only do so when the data leave them no other choice. How can we account for this preference for parsimony within a Bayesian framework? The next

[13] Or, more accurately, even when a cherry-picked parameter value from the complex models provides a better fit to the sample data.

chapters highlight two complementary mechanisms in turn: *adjustment of prior model probability* and *assessment of predictive performance*. In line with Jeffreys, we term these mechanisms *simplicity postulates*.

"The theory of probability explains Ockham's razor" (Jeffreys 1937b, p. 265)

Want to Know More?

✓ 'Nullius in verba' is the motto of the Royal Society, the UK national science academy whose roots date back to 1660. Inspired by a poem from Horace (65BC-8BC), the meaning of 'Nullius in verba' is 'take nobody's word for it'. According to the Society website, "It is an expression of the determination of Fellows to withstand the domination of authority and to verify all statements by an appeal to facts determined by experiment." Around the time that the Society was founded, authority may have referred to the writings of the Greek philosophers from antiquity (particularly Aristotle) whose claims were sometimes speculative, unsupported by experiment, and yet stood unchallenged for over a thousand years. For details see Sutton (1994).

Coat of arms of the *Royal Society*.

✓ Adams, M. M. (1987). *William Ockham*. Notre Dame, IN: University of Notre Dame Press. A 1402-page tome. Some insight about Ockham's razor are mentioned in the main text above.

✓ Etz, A., Haaf, J. M., Rouder, J. N., & Vandekerckhove, J. (2018). Bayesian inference and testing any hypothesis you can specify. *Advances in Methods and Practices in Psychological Science, 1*, 281-295. Explains why Bayesian inference comes with an automatic Ockham's razor. Also includes a discussion of Russell's celestial teapot (see the appendix to this chapter for details).[14]

[14] Yes, we also recommended this article in the chapter on Jeffreys's platitude.

✓ Jefferys, W. H., & Berger, J. O. (1992). Ockham's razor and Bayesian analysis. *American Scientist, 80*, 64-72. Highly recommended as a general introduction to the role of parsimony in Bayesian inference. Includes many concrete examples from a broad range of disciplines.

✓ Jeffreys, H. (1931). *Scientific Inference*. Cambridge: Cambridge University Press. The second-best book on statistics ever written. This first edition includes the Galileo example to demonstrate the influence of parsimony in scientific reasoning, which was introduced earlier by Wrinch and Jeffreys (1921).

✓ Jeffreys, H. (1936). On some criticisms of the theory of probability. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 22*, 337-359. One of several riveting (and mostly ignored) articles published by Harold Jeffreys in the 1935-1939 period. This article includes an extended example, with real data, on the Galileo experiment (pp. 351-353).

✓ Jeffreys, H. (1961). *Theory of Probability* (3rd ed.). Oxford: Oxford University Press. The best book on statistics ever written, and by a landslide. The principle of parsimony is one of the unifying themes of Jeffreys's work.

✓ Gelman, A. (2009). Bayes, Jeffreys, prior distributions and the philosophy of statistics. *Statistical Science, 24*, 176-178. In this provocative article Andrew Gelman argues against the use of Ockham's razor in the statistical modeling of social science data: "In the social science problems I've seen, Ockham's razor is at best an irrelevance and at worse can lead to acceptance of models that are missing key features that the data could actually provide information on."

✓ Hudson, T. E. (2021). *Bayesian Data Analysis for the Behavioral and Neural Sciences*. Cambridge: Cambridge University Press. In Chapter 6, "Model Comparison" (pp. 359-506), the author uses a polynomial regression example to highlight the need for a statistical Ockham's razor. The chapter then demonstrates how Ockham's razor is an automatic by-product of Bayesian inference.

✓ Kalish, M. L., Griffiths, T. L., & Lewandowsky, S. (2007). Iterated learning: Intergenerational knowledge transmission reveals inductive biases. *Psychonomic Bulletin & Review, 14*, 288-294. An experimental setup akin to the children's game of *telephone* (in the UK: *Chinese whispers*) reveals that people have an inductive bias for simplicity. For a complementary line of research see Blanchard et al. (2018).

✓ McFadden, J. (2023). Razor sharp: The role of Occam's razor in science. *Annals of the New York Academy of Sciences, 1530*, 8-17. A recent article that reinforces the main message from this chapter.[15] Its abstract reads:

> "Occam's razor–the principle of simplicity–has recently been attacked as a cultural bias without rational foundation. Increasingly, belief in pseudoscience and mysticism is growing. I argue that inclusion of Occam's razor is an essential factor that distinguishes science from superstition and pseudoscience. I also describe how the razor is embedded in Bayesian inference and argue that science is primarily the means to discover the simplest descriptions of our world." (p. 8)

✓ Sober, E. (2015). *Ockham's Razors: A User's Manual*. Cambridge: Cambridge University Press. Elliott Sober is not a Bayesian but has nonetheless managed to write an informative and entertaining book about parsimony.

✓ Thorburn, W. M. (1918). The myth of Occam's Razor. *Mind, 27*, 345-353.

✓ Vandekerckhove, J., Matzke, D., & Wagenmakers, E.–J. (2015). Model comparison and the principle of parsimony. In Busemeyer, J., Townsend, J., Wang, Z. J., & Eidels, A. (Eds.), *Oxford Handbook of*

[15] The overlap in content is coincidental: the article was published after this chapter had already been completed.

Computational and Mathematical Psychology, pp. 300-319. Oxford: Oxford University Press.

✓ Villarreal, J. M., Etz, A. J., & Lee, M. D. (2023). Evaluating the complexity and falsifiability of psychological models. *Psychological Review, 130*, 853–872.

✓ Wagenmakers, E.–J., van der Maas, H. J. L., & Grasman, R. P. P. P. (2007). An EZ-diffusion model for response time and accuracy. *Psychonomic Bulletin & Review, 14*, 3-22. One of the take-away points is that a model that is manifestly wrong may nonetheless be useful.

Appendix: Teapots, Donkeys, and Dragons

Sir Bertrand Russell was an intellectual giant who worked mainly in mathematics and philosophy. In 1950 Russell was awarded the Nobel Prize in Literature "in recognition of his varied and significant writings in which he champions humanitarian ideals and freedom of thought." During World War I, Russell was imprisoned for his pacifism. Here we limit our discussion of Russell's work to his introduction of a teapot:

> "Many orthodox people speak as though it were the business of sceptics to disprove received dogmas rather than of dogmatists to prove them. This is, of course, a mistake. If I were to suggest that between the Earth and Mars there is a china teapot revolving about the sun in an elliptical orbit, nobody would be able to disprove my assertion provided I were careful to add that the teapot is too small to be revealed even by our most powerful telescopes. But if I were to go on to say that, since my assertion cannot be disproved, it is intolerable presumption on the part of human reason to doubt it, I should rightly be thought to be talking nonsense. If, however, the existence of such a teapot were affirmed in ancient books, taught as the sacred truth every Sunday, and instilled into the minds of children at school, hesitation to believe in its existence would become a mark of eccentricity and entitle the doubter to the attentions of the psychiatrist in an enlightened age or of the Inquisitor in an earlier time. It is customary to suppose that, if a belief is widespread, there must be something reasonable about it. I do not think this view can be held by anyone who has studied history. Practically all the beliefs of savages are absurd." (Russell 1952/1997, pp. 547-548)

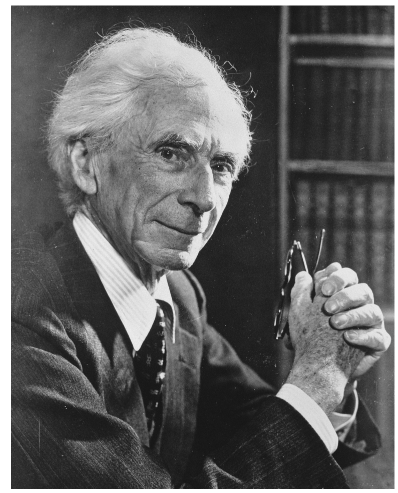


Figure 18.8: British philosopher, mathematician, and pacifist Bertrand Russell (1872-1970) in 1957. Dorothy Wrinch, the heroine of this book, was a pupil of Russell and introduced him to his later wife Dora Black. In one of his letters, Russell refers to her as "the elusive little Wrinch" (Russell 1975/2009, p. 356). For a discussion of Russell's view on probability see Jeffreys (1950).

Russell introduced the teapot as an argument against religion, but it can be considered a more general argument in favor of Ockham's razor and the principle of parsimony. In the above fragment, note the correspondence with Jeffreys's maxim: "the onus of proof is always on the advocate of the more complicated hypothesis" (Jeffreys 1961, p. 343).

Also note that it does not matter whether the teapot theory could be quickly and decisively confirmed or falsified. Suppose that one year from now we stand to gain access to an advanced technology that could tell us in an instant whether or not a celestial teapot orbits the sun. This would be irrelevant to the current epistemic status of the teapot theory. It is not the fact that the teapot theory cannot be falsified; it is that the teapot theory provides an account of the world that *adds complexity without proof*. For this reason, and this reason alone, the teapot theory violates the canon of scientific procedure. As will be detailed in the next chapter, the first simplicity postulate states that complex hypotheses are *a priori* less plausible than simple hypotheses.

Russell was not the first to suggest that religious dogma violates scientific procedure:

> "It may be objected that there is a legitimate domain for authority, consisting of doctrines which lie outside human experience and therefore cannot be proved or verified, but at the same time cannot be disproved. Of course, any number of propositions can be invented which cannot be disproved, and it is open to any one who possesses exuberant faith to believe them; but no one will maintain that they all deserve credence so long as their falsehood is not demonstrated. And if only some deserve credence, who, except reason, is to decide which? If the reply is, Authority, we are confronted by the difficulty that many beliefs backed by authority have been finally disproved and are universally abandoned. Yet some people speak as if we were not justified in rejecting a theological doctrine unless we can prove it false. But the burden of proof does not lie upon the rejecter. I remember a conversation in which, when some disrespectful remark was made about hell, a loyal friend of that establishment said triumphantly, "But, absurd as it may seem, you cannot disprove it." If you were told that in a certain planet revolving round Sirius there is a race of donkeys who talk the English language and spend their time in discussing eugenics, you could not disprove the statement, but would it, on that account, have any claim to be believed? Some minds would be prepared to accept it, if it were reiterated often enough, through the potent force of suggestion. This force, exercised largely by emphatic repetition (the theoretical basis, as has been observed, of the modern practice of advertising), has played a great part in establishing authoritative opinions and propagating religious creeds." (Bury 1913, pp. 19-20)

More recently, the American astronomer and skeptic Carl Sagan (1934-1996) made a similar point. He invited the reader to imagine him making the claim "a fire-breathing dragon lives in my garage". The following hypothetical conversation between Sagan and the reader then unfolds:

> "Show me," you say. I lead you to my garage. You look inside and see a ladder, empty paint cans, an old tricycle–but no dragon.
>
> "Where's the dragon?" you ask.
>
> "Oh, she's right here," I reply, waving vaguely. "I neglected to mention that she's an invisible dragon."
>
> You propose spreading flour on the floor of the garage to capture the dragon's footprints.

> "Good idea," I say, "but this dragon floats in the air."
> Then you'll use an infrared sensor to detect the invisible fire.
> "Good idea, but the invisible fire is also heatless."
> You'll spray-paint the dragon and make her visible.
> "Good idea, except she's an incorporeal dragon and the paint won't stick."
> And so on. I counter every physical test you propose with a special explanation of why it won't work.
> Now, what's the difference between an invisible, incorporeal, floating dragon who spits heatless fire and no dragon at all? If there's no way to disprove my contention, no conceivable experiment that would count against it, what does it mean to say that my dragon exists? Your inability to invalidate my hypothesis is not at all the same thing as proving it true. Claims that cannot be tested, assertions immune to disproof are veridically worthless, whatever value they may have in inspiring us or in exciting our sense of wonder. What I'm asking you to do comes down to believing, in the absence of evidence, on my say-so." (Sagan 1995, p. 171)

We strongly agree with the part of the Bury-Russell-Sagan argument which holds that the onus of proof is on the advocate of the more complicated hypothesis. At the same time, however, we strongly *disagree* that it is the openness to empirical falsification that characterizes a scientific hypothesis.

To clarify, the mere fact that an assertion is falsifiable does not make it scientific. For instance, the Egyptian-American biochemist Rashad Khalifa (1935-1990) concluded that the Quran contains the prediction that the word will end in 2280: "Thus the world ends in 1710 AH, 19×90, which coincides with 2280 AD, 19×120. For the disbelievers who do not accept these powerful Quranic proofs, the end of the world will come suddenly" (Khalifa 2010, p. 1481 in his Appendix 25, 'End of the World', pp. 1479-1482). Such precise doomsday predictions are highly falsifiable –and so far all of them have been falsified– but predictions derived from holy scripture are certainly not scientific.

Khalifa was assassinated by Sunni Islamic extremists on January 31, 1990.

The reverse also holds: a scientific assertion need not be falsifiable. This goes for most claims about events that have happened in the past about which no more information will be forthcoming. For instance, based on an evaluation of all historical information available, one may make the following claim: "The philosopher Leucippus, inventor of atomism, truly existed." When backed up by a comprehensive analysis of ancient Greek and Latin texts, this claim strikes us as eminently scientific, and certainly not "veridically worthless". What is essential is that the claim is supported by evidence.[16] For a similar view see the box below.

[16] Consider the Aesop (c. 620–564 BC) fable 'The Fox and the Monkey': "A Fox and a Monkey were travelling together on the same road. As they journeyed, they passed through a cemetery full of monuments. "All these monuments which you see," said the Monkey, "are erected in honour of my ancestors, who were in their day freed-men, and citizens of great renown." The Fox replied, "You have chosen a most appropriate subject for your falsehoods, as I am sure none of your ancestors will be able to contradict you." " (Townsend 1887, p. 131)

Josiah Royce on the Sciences of Past History

In the introduction to Poincaré's trilogy *The Foundations of Science*, the American philosopher Josiah Royce (1855-1916) elaborates on Poincaré's notion that scientific hypotheses can be valuable even when they cannot be confirmed or falsified by experience:

> "Unverifiable and irrefutable hypotheses in science are indeed, in general, indispensable aids to the organization and to the guidance of our interpretation of experience. (...)
>
> The historical sciences, and in fact all those sciences such as geology, and such as the evolutionary sciences in general, undertake theoretical constructions which relate to past time. Hypotheses relating to the more or less remote past stand, however, in a position which is very interesting from the point of view of the logic of science. Directly speaking, no such hypothesis is capable of confirmation or of refutation, because we can not return into the past to verify by our experience what then happened. (...)
>
> (...) whenever a science is mainly concerned with the remote past, whether this science be archeology, or geology, or anthropology, or Old Testament history, the principal theoretical constructions always include features which no appeal to present or to accessible future experience can ever definitely test. Hence the suspicion with which students of experimental science often regard the theoretical constructions of their confrères of the sciences that deal with the past. The origin of the races of men, of man himself, of life, of species, of the planet; the hypotheses of anthropologists, of archeologists, of students of 'higher criticism'—all these are matters which the men in the laboratory often regard with a general incredulity as belonging not at all to the domain of true science. Yet no one can doubt the importance and the inevitableness of endeavoring to apply scientific method to these regions also. Science needs theories regarding the past history of the world. And no one who looks closer into the methods of these sciences of past time can doubt that verifiable and unverifiable hypotheses are in all these regions inevitably interwoven (...)" (Royce, in Poincaré 1913, pp. 17-20; cf. Poincaré 1913, p. 343)

Part IV

Appendices

33 Jevons Explains Permutations

Certain it is that life demands incessant novelty, and that nature though it probably never fails to obey the same fixed laws, yet presents to us an apparently unlimited series of varied combinations of events.

Jevons, 1874

Chapter Goal

This chapter describes the basic concepts of permutations. One of the best explanations of permutations was provided by Jevons in his 1874 masterpiece *The Principles of Science*, and instead of bumbling through the topic ourselves and withholding from the reader the pleasure of digesting a superior explanation we decided to extract the most relevant sections from Jevons, and offer them here. A modern explanation can be be found for instance in Blitzstein and Hwang (2019).

The Art or Doctrine of Combinations

In the chapter 'The Variety of Nature, or the Doctrine of Combinations and Permutations,' Jevons provides a lively and clear exposition of permutations and combinations. At the start of the chapter, Jevons seeks to establish the importance of the topic by including a lengthy citation from *De Arte Conjectandi* by Jacob Bernoulli (pp. 198-200). In the cited fragment, Jacob Bernoulli[1] first claims that the intuitive assessment of permutations leads to errors in reasoning:

[1] Written by Jevons as James Bernouilli.

> "*the insufficient or imperfect enumeration of parts or causes* (...) is the chief, and almost the only, source of the vast number of erroneous opinions, and those too very often in matters of great importance, which we are apt to form on all the subjects we reflect upon, whether they relate to the knowledge of nature or the merits and motives of human actions."

Bernoulli continues to argue that the doctrine of combinations affords a cure to this weakness, and therefore:

> "...that art [*the doctrine of combinations*]...deserves to be considered as most eminently useful and worthy of our highest esteem and attention. (...) Nor is this art or doctrine to be considered merely

as a branch of the mathematical sciences. For it has a relation to almost every species of useful knowledge that the mind of man can be employed upon. It proceeds indeed upon mathematical principles, in calculating the number of the combinations of the things proposed: but by the conclusions that are obtained by it, the sagacity of the natural philosopher, the exactness of the historian, the skill and judgment of the physician, and the prudence and foresight of the politician may be assisted; because the business of all these important professions is but to *form reasonable conjectures* concerning the several objects which engage their attention, and all wise conjectures are the results of a just and careful examination of the several different effects that may possibly arise from the causes that are capable of producing them." James Bernouilli, 'De Arte Conjectandi,' translated by Baron Maseres. London, 1795, pp. 35-36.

Rarely if ever has the theory of combinations and permutations been introduced more eloquently or more passionately.[2] The importance of the topic thus established, Jevons' first order of business is to establish some terminology.

[2] Perhaps James Bernouilli spent little too much time on the study of permutations.

Distinction of Combinations and Permutations

"We must at once consider the deep difference which exists between Combinations and Permutations; a difference involving important logical principles, and influencing the form of all our mathematical expressions. In *permutation* we recognize varieties of order or arrangement, treating AB as a different group from BA. In *combination* we take notice only of the presence or absence of a certain thing, and pay no regard to its place in order of time or space. Thus the four letters *a*, *e*, *m*, *n* can form but one combination, but they occur in language in several permutations, as *name*, *amen*, *mean*, *mane*. " (Jevons 1874/1913, p. 200)

Next, Jevons describes how to compute permutations without restrictions.

Unrestricted Permutations

"Permutations of certain things are far more numerous than combinations of those things, for the obvious reason that each distinct thing is regarded differently according to its place. Thus the letters A, B, C, will make different permutations according as A stands first, second, or third; having decided the place of A, there are two places between which we may choose for B; and then there remains but one place for C. Accordingly the permutations of these letters will be altogether $3 \times 2 \times 1$ or 6 in number. With four things or letters, A, B, C, D, we shall have four choices of place for the first letter, three for the second, two for the third, and one for the fourth, so that there will be altogether $4 \times 3 \times 2 \times 1$, or 24 permutations. The same simple rule applies in all cases; beginning with the whole number of things we multiply at each step

> by a number decreased by a unit, In general language, if n be the number of things in a combination, the number of permutations is $n(n-1)(n-2) \cdot \ldots \cdot 4 \cdot 3 \cdot 2 \cdot 1$. Thus, if we were to re-arrange the names of the days of the week, the possible arrangements out of which we should have to choose the new order, would be no less than $7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1$, or 5040, or, excluding the existing order, 5039." (Jevons 1874/1913, p. 201)

Jevons goes on to mention that "the product of all integer numbers, from unity up to any number n, is the *factorial* of *n*." (p. 202) The modern notation for this is $n!$, or '*n* factorial'.

Restricted Permutations

In many cases, however, there are important restrictions on the permutations that are to be distinguished:

> "In some questions the number of permutations may be restricted and reduced by various conditions. Some things in a group may be undistinguishable [sic] from others, so that change of order will produce no difference. Thus if we were to permutate [sic] the letters of the name *Ann*, according to our previous rule, we should obtain $3 \times 2 \times 1$, or 6 orders; but half of these arrangements would be identical with the other half, because the interchange of the two *n*'s has no effect. The really different orders will therefore be $\frac{3 \cdot 2 \cdot 1}{1 \cdot 2}$ or 3, namely *Ann*, *Nan*, *Nna*. In the word *utility* there are two *i*'s and two *t*'s, in respect of both of which pairs the number of permutations must be halved. Thus we obtain $\frac{7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{1 \cdot 2 \cdot 1 \cdot 2}$ or 1260, as the number of permutations. The simple rule evidently is that when some things or letters are undistinguished, proceed in the first place to calculate all the possible permutations as if all were different, and then divide by the number of possible permutations of those series of things which are not distinguished, and of which the permutations have therefore been counted in excess. Thus since the word *Utilitarianism* contains fourteen letters, of which four are *i*'s, two *a*'s, and two *t*'s, the number of distinct arrangements will be found by dividing the factorial of 14, by the factorials of 4, 2, and 2, the result being 908,107,200. From the letters of the word *Mississippi* we can get in like manner[3] $\frac{11!}{4! \times 4! \times 2!}$ or 34,650 permutations, or not one-thousandth part of what we should obtain were all the letters different." (Jevons 1874/1913, pp. 203-204)

[3] Here we use the modern notation for the factorial instead of that used by Jevons.

Calculation of Number of Combinations

Finally, Jevons then describes how many ways there are to select m units from a total of n:

> "Suppose that we wish to determine the number of ways in which we can select three letters out of the alphabet, without allowing the same letter to be repeated. At the first choice we can take any one of 26 letters; at the next step there remain 25 letters, any one

of which may be joined with that already taken; at the third step there will be 24 choices, so that apparently the whole number of ways of choosing is $26 \times 25 \times 24$. But the fact that one choice succeeded another has caused us to obtain the same combinations of letters in different orders; we should get, for instance, *a*, *p*, *r* at one time, and *p*, *r*, *a* at another, and every three distinct letters will appear six times over, because three things can be arranged in six permutations. Thus the true number of combinations will be $\frac{24\times23\times22}{1\times2\times3}$, or 2024.[4]

It is apparent that we need the doctrine of permutations in order that we may in many questions counteract the exaggerating effect of successive selection. If out of a senate of 30 persons we have to choose a committee of 5, we may choose any of 30 first, any of 29 next, and so on, in fact there will be $30\times29\times28\times27\times26$ selections; but as the actual character of the members of the committee will not be affected by the accidental order of their selection, we divide by $1\times2\times3\times4\times5$, and the possible number of different committees will be $142,506$. (...)

In general algebraic language, we may say that a group of *m* things may be chosen out of a total number of *n* things, in a number of combinations denoted by the formula

$$\frac{n\cdot(n-1)(n-2)(n-3)....(n-m+1)}{1\cdot2\cdot3\cdot4...m}$$

The extreme importance and significance of this formula seems to have been first adequately recognised by Pascal, although its discovery is attributed by him to a friend, M. de Ganières.[5] We shall find it perpetually recurring in questions both of combinations and probability, and throughout the formulæ of mathematical analysis traces of its influence will be noticed." (Jevons 1874/1913, pp. 204-205)

[4] This is an error that Jevons, in a later edition, corrected to $\frac{26\times25\times24}{1\times2\times3}$, or 2600.

[5] 'Œuvres Complètes de Pascal' (1865), vol. iii. p. 302. Montucla states the name as De Gruières, 'Histoire des Mathématiques,' vol. iii. p. 389.

Binomial Likelihood

Given a binomial success parameter θ, what is the probability mass function of the number of successes s out of n attempts, and the remaining f attempts resulting in failure? For instance, given a particular value of θ we might wish to know the probability of obtaining exactly 6 successes (i.e., $s = 6$) out of 10 trials (i.e., $n = 10$, $f = 4$). Denoting successes by '1' and failures by '0', we could entertain the sequence $(1,1,1,1,1,1,0,0,0,0)$. For this exact sequence, the probability of obtaining it is given by $\theta^6 \times (1-\theta)^4$. But the sequence order is irrelevant, and other sequences exist that have the same probability, for instance $(0,0,0,0,1,1,1,1,1,1)$ or $(0,1,0,1,0,1,0,1,1,1)$. How many of these orderings exist? As explained by Jevons earlier, we start by computing all permutations, that is, $n! = 10! = 3,628,800$. However, the orderings of the successes are irrelevant, and there are $s! = 6! = 720$ of them; the orderings of the failures are likewise irrelevant, and they number $f! = 4! = 24$. These irrelevant permutations correct the total relevant permutations to[6]

$$\binom{n}{s} = \frac{n!}{s!\,f!} = \frac{10!}{6!\,4!} = 210.$$

[6] Note that this is the same equation as given by Jevons above.

In other words, there are 210 relevant sequences that consists of 6 successes and 4 failures. The probability of finding any single sequence may be $\theta^6 \times (1-\theta)^4$, but there are 210 of them, so the overall probability equals $210 \times \theta^6 \times (1-\theta)^4$. In general then, given a particular value of θ the probability of obtaining exactly s successes out of n trials equals

$$\binom{n}{s}\theta^s \times (1-\theta)^{(n-s)}.$$

Want to Know More?

✓ Blitzstein, J. K., & Hwang, J. (2019). *Introduction to Probability (2nd ed.)*. Taylor & Francis Group.

✓ Jevons, W. S. (1874/1913). *The Principles of Science: A Treatise on Logic and Scientific Method*. London: MacMillan.

34 *Pascal's Arithmetical Triangle*

The Arithmetical Triangle is the most famous of all number patterns. Apparently a simple listing of the binomial coefficients, it contains the triangular and pyramidal numbers of ancient Greece, the combinatorial numbers which arose in the Hindu studies of arrangements and selections, and (barely concealed) the Fibonacci numbers from medieval Italy. It reveals patterns which delight the eye, raises questions which tax the number-theorists, and amongst the coefficients "There are so many relations present that when someone finds a new identity, there aren't many people who get excited about it any more, except the discoverer!" [1]

Reference [1] is to Knuth (1973, pp. 52-53).

Edwards, 2019

Chapter Goal

This chapter describes Pascal's arithmetical triangle, a simple yet fascinating mathematical construction that has played a key role in the development of probability theory.

The City Block

You find yourself in a recently constructed city whose roads form a perfect grid, as illustrated in Figure 34.1.[1] Your goal is to travel from the starting position indicated by the blue dot to the end position indicated by the red dot. The shortest path always involves exactly five moves to the east ('E') and three moves to the north ('N'), for a total of eight moves. The order of the moves is irrelevant, that is, any order will get you to your final position. In Figure 34.1, the journey consists of the move sequence $\{E, N, E, E, N, E, E, N\}$. How many ways can you travel from the blue position to the red position? In other words, how many different sequences exist that have exactly five 'E' moves and three 'N' moves? From Chapter 33 we know the answer. Let $n = 8$ be the total number of moves, $s = 5$ equal the number of moves to the east, and $f = 3$ equal the number of moves to the north. We then have

$$\binom{n}{s} = \frac{n!}{s!\,f!} = \frac{8!}{5!\,3!} = 56.$$

Figure 34.1 provides a geometric representation of the number of different ways in which two elements (i.e., 'E' and 'N') may be ordered.

[1] This is also called a *lattice diagram*, see Edwards (1987/2019, p. 73).

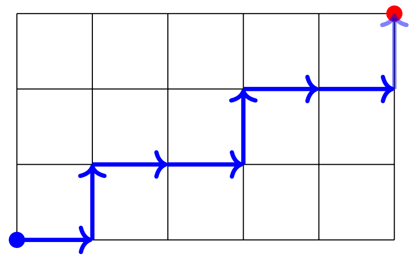


Figure 34.1: A grid city in which the shortest route from the blue dot to the red dot takes exactly five moves to the east and three moves to the north. There are 56 possible paths.

This representation suggests a more difficult question: suppose we start at the blue dot, and we take eight random moves east or north, where will we end up, and with what probability? The associated grid city is shown in Figure 34.2.

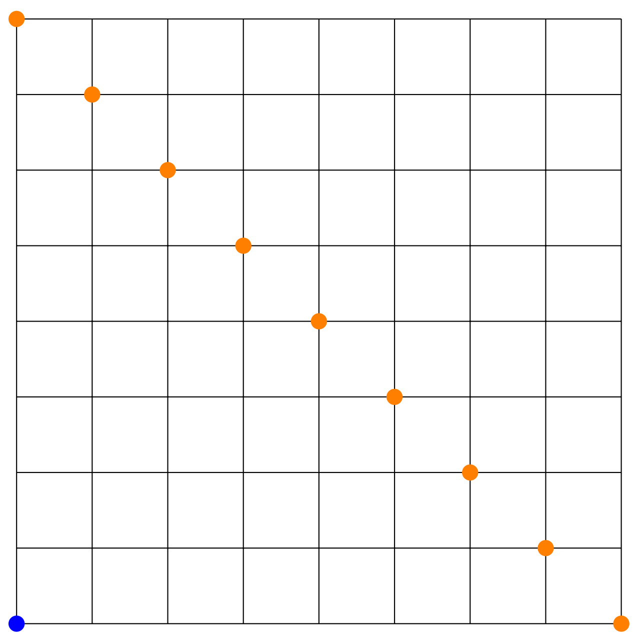
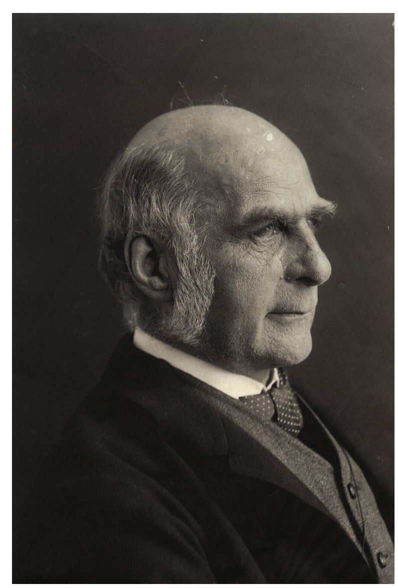


Figure 34.2: A grid city in which each of the orange dots marks the potential end of a journey that starts at the blue dot and involves eight random moves east or north.

Note that relatively many paths lead to end points in the center of the city. The end point at the edges, however, can only be reached by a few paths. For instance, the rightmost orange dot can only be reach by a single path: $\{E, E, E, E, E, E, E, E\}$. This feature is brought out more clearly by a physical process – the Galton board or *quincunx*.

The Galton Board aka the Quincunx

The English polymath Sir Francis Galton (1822-1911) was brilliant, energetic, and highly influential. Among many other contributions, Galton coined the phrase 'nature versus nurture', he initiated the statistical study



Sir Francis Galton (1822–1911), shown here at 73 years of age. Photograph taken by Eveleen Myers (née Tennant).

of correlation and regression, he devised the first weather map, and he founded the field of psychometrics (i.e, the measurement of individual differences in cognitive ability). His disciple Karl Pearson –a phenomenally influential statistician himself– wrote a four-volume, 1786-page (!) biography on Galton in which he called him "perhaps the greatest scientist of the nineteenth century" (Pearson 1930a, p. vi).

Unfortunately for Galton's legacy, he also invented the word 'eugenics' and obsessively promoted scientific racism. This is something that should *not* be swept under the rug, and for those readers who wonder 'but how bad can it really be?' we have added an appendix that provides a few characteristic quotations – by Galton, but also by fellow statisticians and eugenicists Karl Pearson and Ronald Fisher. The reader should be warned: the fragments in the appendix are abhorrent, callous, and could, if advocated nowadays, even result in jail time.

For now we leave the topic of eugenics and consider the section in Galton's 1889 book *Natural Inheritance* where he introduces his 'quincunx' – the Galton board, illustrated by margin Figure 34.3. The relevant section is titled *Mechanical Illustration of the Cause of the Curve of Frequency* and we quote from it liberally:

> "[The apparatus] is a frame glazed in front, leaving a depth of about a quarter of an inch behind the glass. Strips are placed in the upper part to act as a funnel. Below the outlet of the funnel stand a succession of rows of pins stuck squarely into the backboard, and below these again are a series of vertical compartments. A charge of small shot [i.e., small lead or steel pellets – EWDM] is inclosed. When the frame is held topsy-turvy, all the shot runs to the upper end; then, when it is turned back into its working position, the desired action commences. Lateral strips, shown in the diagram, have the effect of directing all the shot that had collected at the upper end of the frame to run into the wide mouth of the funnel. The shot passes through the funnel and issuing from its narrow end, scampers deviously down through the pins in a curious and interesting way; each of them darting a step to the right or left, as the case may be, every time it strikes a pin. The pins are disposed in a quincunx fashion [i.e., as five pips on a die: ⚄ – EWDM], so that every descending shot strikes against a pin in each successive row. The cascade issuing from the funnel broadens as it descends, and, at length, every shot finds itself caught in a compartment immediately after freeing itself from the last row of pins. The outline of the columns of shot that accumulate in the successive compartments approximates to the Curve of Frequency (...), and is closely of the same shape however often the experiment is repeated. The outline of the columns would become more nearly identical with the Normal Curve of Frequency, if the rows of pins were much more numerous, the shot smaller, and the compartments narrower; also if a larger quantity of shot was used.
>
> The principle on which the action of the apparatus depends is, that a number of small and independent accidents befall each shot in its career. In rare cases, a long run of luck continues to favour the course of a particular shot towards either outside place, but in the large majority of instances the number of accidents that cause

FIG. 7.

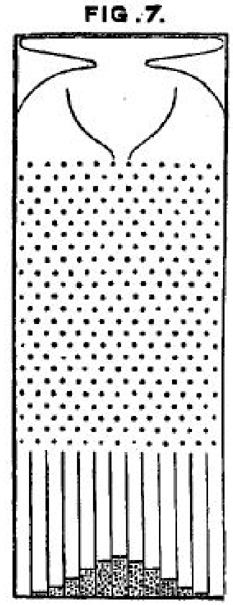


Figure 34.3: Galton's original illustration of his 'quincunx' (Galton 1889, p. 63).

Deviation to the right, balance in a greater or less degree those that cause Deviation to the left. Therefore most of the shot finds its way into the compartments that are situated near to a perpendicular line drawn from the outlet of the funnel, and the Frequency with which shots stray to different distances to the right or left of that line diminishes in a much faster ratio than those distances increase. This illustrates and explains the reason why mediocrity is so common." (Galton 1889, pp. 63-65)

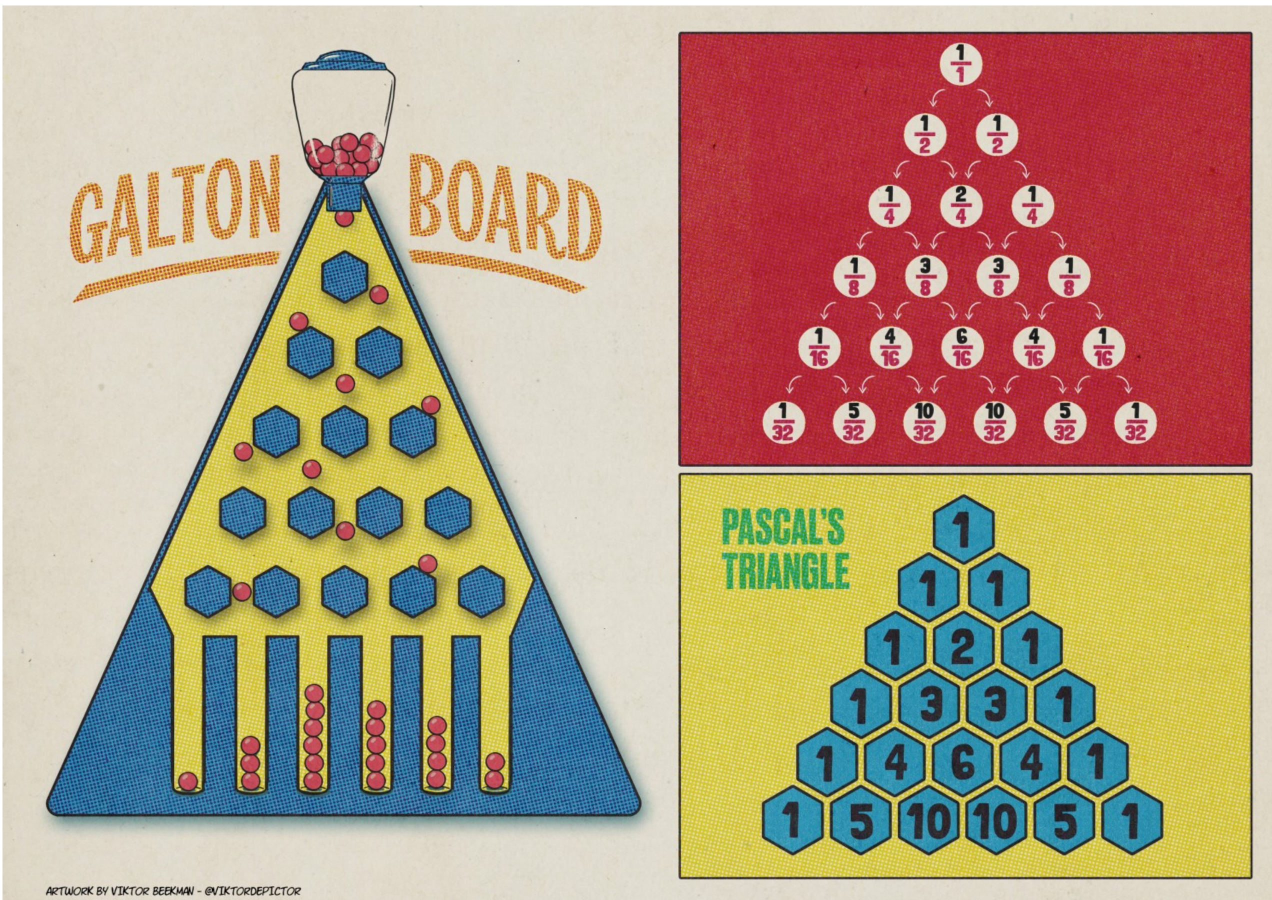


Figure 34.4: The regularities of randomness. Left panel: the Galton board or *quincunx*; top right panel: the probabilities associated with each position on the Galton board; bottom right panel: Pascal's triangle. Each number is the sum of the two parent numbers in the row above it. The behavior of a single process is random and unpredictable, but the behavior of the group is highly regular.

A modern rendition of the quincunx is shown in the left panel of Figure 34.4. Instead of a person wandering aimlessly in a grid city we now consider a falling pallet that, whenever it hits a pin, bounces to the left or to the right with equal probability, continuing its downward journey until it comes to rest in a container at the bottom.

For a pallet to end up in the leftmost container, it needs to have made five consecutive left turns, meaning that only a single path is possible: $\{L, L, L, L, L\}$. For the pallet to land in the adjacent container, it needs to have made four left turns and one right turn, which could occur at any pin; thus, there are a total of five possible paths. In general, the number of paths to the s^{th} column from the left (starting at $s = 0$ and ending at $s = 5$, where s can also be interpreted as the number of times the pallet bounced to the right) equals $\binom{n}{s}$, where $n = 5$ is the number of bounces before the pallet lands in a container. For the six containers in Figure 34.4 this yields $\{1, 5, 10, 10, 5, 1\}$ possible paths for $s = 0, ..., 5$. The total number of paths across all containers is 2^n (i.e., every pin row doubles the number of paths), so that the probability that a pallet will finish in the s^{th} column from the left equals $\binom{n}{s}/2^n$ (i.e., the proportion of the total number of paths that lead to the s^{th} column). This is echoed by the top right panel of Figure 34.4.

Consistent with Galton's description, relatively many paths terminate at the middle containers, and relatively few paths terminate at containers toward the edges. As the number of rows increases, the distribution of pellets across the containers is approximated increasingly well by a bell-curve, widely known as the *Gaussian* or *normal* distribution. This approximation was a crucial step in the development of statistics, but its history and derivation are outside of the scope of this appendix.[2]

[2] For a detailed technical account see Todhunter (1865); for an accessible overview (albeit with a consistent mistake in the equation for the normal distribution!) see Stewart (2012, Chapter 7).

The Galton board illustrates several statistical ideas. Firstly, as indicated above, processes that are the result of an accumulation of many small impacts tend to be normally distributed. Secondly, the behavior of a single pellet may appear haphazard but the *ensemble* of pellets shows a highly predictable pattern. Thirdly, a more detailed study of the Galton board in action reveals that this predictable pattern arises even when individual pellets behave anomalously:

> "(...) consider how the balls bounce around. According to the binomial model, each time a ball hits a peg, it should cleanly drop either to the left or to the right. But this is not what happens in our real-world Galton board. There, the balls bounce around wildly: they hit one another, they bounce upward, they hop to the side and hit the next peg in the same row, they ricochet off the walls, they skip several rows; a brief glance at the demonstration[3] should convince anybody that the abstraction offered by the binomial model is not warranted – that is, the abstraction is clearly wrong and the model is misspecified. Nevertheless, the histograms at the bottom appear to be consistent with the binomial model – the normal distribution provides a good description of the end result. So there is considerable value to the use of a parametric model (e.g., the binomial model, or its normal approximation) even though we can be certain that the model is dead wrong in the details." (Wagenmakers, 2018)[4]

[3] See the `BayesianSpectacles.org` blog post "A Galton board demonstration of why all statistical models are misspecified" for a movie featuring 3,000 pellets traveling downward in slow-motion – EWDM.

[4] Quotation taken from the `BayesianSpectacles.org` blog post "A Galton board demonstration of why all statistical models are misspecified".

Pascal's Triangle

After a long introduction we have finally arrived at "the most famous of all number patterns": *Pascal's triangle*. The triangle was known long before the famous French mathematician Blaise Pascal (1623-1662) wrote *Traité du triangle arithmétique, avec quelques autres petits traitez sur la mesme matière* (published in 1665, composed in 1654; see Edwards 1987/2019, p. 58). As noted by Edwards:

> "Pascal was, as we shall see, a little forgetful about his sources. Practically everything in the *Traité* except the solution to the important "Problem of Points" will have been known to Mersenne's circle[5] by 1637. It seems likely that Pascal absorbed most of this as a young man, and then, more than a decade later, his correspondence with Fermat stimulated him to compose the *Traité*, which he did in the space of a few weeks. The evidence is that, with the passage of time, he had lost most of the details whilst retaining the outline. (...) His novel theme was to view the properties of the Arithmetical Triangle as *pure mathematics*, to be demonstrated from the fundamental addition relation independently of any binomial or combinatorial application." (Edwards 1987/2019, p. 58)

[5] Founded in 1635, Marin Mersenne's informal *Academia Parisiensis* was a hub for mathematical discourse in Europe – EWDM.

The triangle is displayed in the bottom right panel of Figure 34.4. Its construction is simple: other than the entries '1' that form the triangle flanks, each number is the sum of the two numbers just above it. By convention the top number, '1' is considered row $n = 0$; consider then row $n = 4$, with entries $\{1, 4, 6, 4, 1\}$. The leftmost '4' arises because $1 + 3 = 4$, the center '6' because $3 + 3 = 6$, and the rightmost '4' because $3 + 1 = 4$. In row $n = 5$, the leftmost '10' arises because $4 + 6 = 10$, and the rightmost '10' because $6 + 4 = 10$. The triangle can be expanded indefinitely.

A comparison of the top and bottom right panels of Figure 34.4 shows that the path numbers that lead to a particular position on the Galton board are *identical* to the entries in Pascal's triangle. This occurs because the mathematical method of construction for Pascal's triangle is mimicked by the physical action on the Galton board. Consider for instance a pallet that ended up in the third container from the left, a position marked as $^{10}/_{32}$ in the top right panel of Figure 34.4. This pellet arrived there either from the left 'parent path' (i.e., through the position marked as $^{4}/_{16}$) or from the right 'parent path' (i.e., through the position marked as $^{6}/_{16}$). There are no other possibilities. The total number of pellet paths that lead to a given position is therefore the sum of the number of paths for its two potential parents.

Each entry in Pascal's triangle can therefore be given a Galton-board interpretation as the number of possible paths that lead to it. In turn this implies that the numbers in the triangle quantify the ways in which a given number of 'left' and 'right' movements can be ordered. In other words, the entry in the n^{th} row and s^{th} column in Pascal's triangle is given by $\binom{n}{s}$. For instance, the $n = 5$, $s = 2$ entry (i.e., lowest row, third number from the left) equals $\binom{5}{2} = 10$.

Remarkably, the entries of Pascal's triangle also provide the coefficients for the different factors in the binomial expansion of $(a + b)^n$. For instance, for $n = 0 \ldots 5$ we have:

$$
\begin{aligned}
(a+b)^{0} &= 1\\
(a+b)^{1} &= 1 \cdot a + 1 \cdot b\\
(a+b)^{2} &= 1 \cdot a^2 + 2 \cdot ab + 1 \cdot b^2\\
(a+b)^{3} &= 1 \cdot a^3 + 3 \cdot a^2 b + 3 \cdot ab^2 + 1 \cdot b^3\\
(a+b)^{4} &= 1 \cdot a^4 + 4 \cdot a^3 b + 6 \cdot a^2 b^2 + 4 \cdot ab^3 + 1 \cdot b^4\\
(a+b)^{5} &= 1 \cdot a^5 + 5 \cdot a^4 b + 10 \cdot a^3 b^2 + 10 \cdot a^2 b^3 + 5 \cdot ab^4 + 1 \cdot b^5.
\end{aligned}
$$

The red exponent indicates the row number n, and the blue numbers provide the values for the coefficients – identical to the entries in Pascal's triangle. The binomial theorem states that $(a+b)^n = \sum_{s=0}^{n} \binom{n}{s} a^{n-s} b^s$, which of course features the $\binom{n}{s}$ term explicitly. Laplace explains:

> "Suppose that an urn contains a white and b black balls, and that after one ball has been extracted it is replaced in the urn. What is the probability that, in n such draws, one will get m white and $n-m$ black balls? It is clear that the number of possible outcomes or cases on each draw is $a+b$. Each outcome of the second draw may be combined with all outcomes of the first, and so the number of possible outcomes in two draws will be the square of the binomial $a+b$ {i.e. $(a+b)^2$}. In the expansion of this square, a^2 denotes the number of cases in which two white balls are drawn, $2ab$ denotes the number of cases in which one white and one black ball are drawn, and finally b^2 denotes the number of cases in which two black balls are drawn. Continuing in this way we find in general that the nth power of the binomial $(a+b)$ {i.e. $(a+b)^n$} denotes the number of all possible outcomes in n draws, and that, in the expansion of this expression, the term multiplied by a^m [see note i below, EWDM] denotes the number of cases in which m white and $n-m$ black balls are drawn. Then, on dividing this term by the whole power of the binomial {i.e. $(a+b)^n$}, we get the probability of drawing m white and $n-m$ black balls. The ratio of the numbers a to $a+b$ is the probability of getting a white ball in one draw, and the ratio of the numbers b to $a+b$ is the probability of drawing a black ball: if one calls these probabilities p and q, the probability of getting m white balls in n draws will be the coefficient of the mth power of p in the expansion of the binomial $(p+q)^n$ (notice that $p+q=1$). This remarkable property of the binomial is very useful in probability theory. [see note ii below – EWDM]" (Laplace 1814/1995, p. 16)

The translator, Andrew I. Dale, added the following notes:
i."More correctly, the coefficient of $a^m b^{n-m}$."
ii. "This Pollaczek-Geiringer (see also von Mises [1932, p. 191]) sees as an example of the solution of the so-called Bernoulli problem, wherein the probability $\omega_n(m)$ that, from an urn containing a white and b black balls, m white balls are drawn in n draws (with replacement) is

$$\omega_n(m) = \binom{n}{m} p^m q^{n-m},$$

where $p = a/(a+b)$ and $q = b/(a+b)$. The number of cases in which this result obtains is then $(a+b)^n \omega_n(m)$, or $\binom{n}{m} a^m b^{n-m}$. Laplace's contribution to probability, in connexion [sic] with this matter, was the limiting form as $n \to \infty$, that is, the $\exp(-x^2)$ law. De Moivre's approximation to the binomial distribution is discussed in Hald [1990, chap. 24] and Stigler [1986b, pp. $70-88$], while Laplace's extension of de Moivre's theorem is examined in Hald [op. cit. §24.6]."

As suggested in this chapter's epigraph, Pascal's triangle hides many more mathematical treasures. Exploring these treasures is well beyond the scope of this book, but guidance is easily found online.

Exercises

1. How can Pascal's triangle be used to obtain an estimate of π? [hint: consider the normal approximation to the binomial distribution]
2. A coin is assumed to be fair. It is tossed six times. Scenario A yields {H,H,H,H,H,H} (i.e., all heads), and scenario B yields {H,T,T,T,H,H} (i.e., three heads, three tails). Scenario A produces more surprise and suspicion than scenario B. However, both sequences are equally likely – under the hypothesis that the coin is fair, the probability for each sequence is $1/2^6 = 1/64$. What's going on?
3. Let's return to the Problem of Points discussed in Chapter 10. Consider a game of chance where player A requires 2 points to win and player B requires 3 points to win (e.g., a score of $4 - 3$ in a race to 6). (a) use the *Learn Bayes* module to obtain the probability that A wins the game; (b) how can this probability be obtained using Pascal's triangle?
4. Consider again a score of $4 - 3$ in a race to 6. In JASP, activate the *Distributions* module, navigate to the *Discrete* distributions and try to recover the correct result (a) using the binomial distribution; (b) using the negative binomial distribution.

Want to Know More?

✓ Edwards, A. W. F. (1987/2019). *Pascal's Arithmetical Triangle: The Story of a Mathematical Idea*. Mineola, NY: Dover Publications. Essential reading for those who wish to learn more about the history of Pascal's triangle.

✓ Kunert, J., Montag, A., & Pöhlmann, S. (2001). The quincunx: History and mathematics. *Statistical Papers, 42*, 143–169.

✓ Pearson, K. (1914,1924,1930a,1930b). *The Life, Letters and Labours of Francis Galton*. Cambridge: Cambridge University Press. A multi-volume, 1786-page biography written by friend and admirer Karl Pearson. If the biography was not permeated with eugenics and scientific racism, it may have been one of the most impressive and interesting biographies ever composed. A sample fragment: "Civilisation has gained nothing from rivalry in destructive warfare; It can gain enormously from the rivalry of nations in rearing their future generations from the most efficient of their citizens. Galton was the first to realise this great truth, to preach it as a moral code, and to lay the foundations of the new science which it demands of man. In the centuries to come, when the principles of Eugenics shall be commonplaces of social conduct and of politics, men, whatever their race, will desire to know all that is knowable about one of the greatest, perhaps the greatest scientist of the nineteenth century." (Pearson 1930a, p. vi)

✓ The internet offers many excellent resources on Pascal's triangle. Example are `https://www.theochem.ru.nl/~pwormer/Knowino/knowino.org/wiki/Pascal's_triangle.html`, `https://www.mathsisfun.com/pascals-triangle.html`, and `https://www.mathsisfun.com/algebra/binomial-theorem.html`; the relevant Wikipedia pages (e.g., `https://en.wikipedia.org/wiki/Binomial_theorem`) are also informative.

Appendix: The Taint of Eugenics

We mentioned earlier that we do not wish to praise the scientific contributions of Sir Francis Galton without openly discussing the scientific racism that he and his followers advocated. These eugenicists did not 'merely' promote scientific racism as an abstract hypothesis, but also encouraged the associated political action and its real-world consequences.

Below are a few statements that are certain to make a modern-day reader recoil. It is likely that a more thorough reading could have unearthed quotations that are even more shocking, but the point will be clear and we can only stomach so much.

"Francis Galton (right), aged 87, on the stoep at Fox Holm, Cobham, with the statistician Karl Pearson." (`https://en.wikipedia.org/wiki/Francis_Galton`) Public domain.

The Eugenicism of Sir Francis Galton

Galton was the cousin of Charles Darwin and was greatly influenced by *The Origin of Species*. Galton was not only convinced that nature trumps nurture, but he also believed that some races were genetically superior to others. Galton in fact coined the term 'eugenics'. For those who believe that Galton meant well, behold his 1873 letter to *the Times*:

> "average negroes possess too little intellect, self-reliance and self-control to make it possible for them to sustain the burden of any respectable form of civilisation without a large measure of external guidance and support. The Chinaman is a being of another kind, who is endowed with a remarkable aptitude for a high material civilisation. (...) one population continually drives out another. We note how Arab, Tuarick, Fellatah, Negroes of uncounted varieties, Caffre and Hottentot surge and reel to and fro in the struggle for existence. It is into this free fight among all present that I wish to see a new competitor introduced—namely the Chinaman. The gain would be immense to the whole civilised world if he were to outbreed and finally displace the negro, as completely as the latter has displaced the aborigines of the West Indies. The magnitude of the gain may be partly estimated by making the converse supposition –namely the loss that would ensue if China were somehow to be depopulated and restocked by negroes." (Francis Galton, letter to *the Times* of June 6, 1873, as cited in Pearson 1924, p. 33).

The Eugenicism of Karl Pearson

Karl Pearson was a highly influential researcher, a brilliant statistician, and a gifted writer. His book *The Grammar of Science* is a classic that features phrases such as the following:

> "The field of science is unlimited; its material is endless, every group of natural phenomena, every phase of social life, every stage of past or present development is material for science. *The unity of all science consists alone in its method, not in its material.* The man who classifies facts of any kind whatever, who sees their mutual relation and describes their sequences, is applying the scientific method and is a man of science. The facts may belong to the past history of mankind, to the social statistics of our great cities, to the atmosphere of the most distant stars, to the digestive organs of a

> worm, or to the life of a scarcely visible bacillus. It is not the facts themselves which form science, but the method in which they are dealt with." (Pearson 1892/1937, p. 16)

Unfortunately, Karl Pearson was completely on board with Galton's eugenics agenda.[6] Below are three hair-raising quotations.[7] The first one is from Pearson's 1901 book *National life from the standpoint of science*:

[6] Egon Pearson –Karl's son and a highly influential statistician on his own account– did not endorse eugenics.

[7] Content based partly on the `BayesianSpectacles.org` blog post "Karl Pearson's worst quotation?".

> "History shows me one way, and one way only, in which a high state of civilization has been produced, namely, the struggle of race with race, and the survival of the physically and mentally fitter race. If you want to know whether the lower races of man can evolve a higher type, I fear the only course is to leave them to fight it out among themselves, and even then the struggle for existence between individual and individual, between tribe and tribe, may not be supported by that physical selection due to a particular climate on which probably so much of the Aryan's success depended." (Pearson 1901, pp. 19-20)

At the time, Pearson certainly wasn't the only academic who felt this way, and the Holocaust lay hidden in the future, but such statements nevertheless have a spine-chilling effect. In his book Pearson continues in the same style for a couple of pages more, discussing the inferiority of the negro race and the dangers of cross-racial relationships – "if the bad stock be raised the good is lowered". Nausea prevented us from reading further.

With this background in mind, dear readers, hold on to your hats for quotation number two. This quotation requires some background, provided by Wikipedia:

> "In *The Myth of the Jewish Race* Raphael and Jennifer Patai cite Karl Pearson's 1925 opposition (in the first issue of the journal *Annals of Eugenics* which he founded) to Jewish immigration into Britain. Pearson alleged that these immigrants "will develop into a parasitic race. (...) taken *on the average*, and regarding both sexes, this alien Jewish population is somewhat inferior physically and mentally to the native population." (entire citation: Wikipedia; last quotation: Pearson and Moul 1925, pp. 125-126).

This is nothing short of callous of course. But there is more. We were attended to a speech from Pearson in 1934.[8] Judge for yourself quotation number three:

[8] We thank David Colquhoun for bringing this to our attention. For more references please see the website of Dr. Joe Cain, starting with **https://profjoecain.net/karl-pearson-praised-hitler-nazi-race-hygiene/**.

> "The climax culminated in Galton's preaching of Eugenics, and his foundation of the Eugenics Professorship. Did I say "culmination"? No, that lies rather in the future, perhaps with Reichskanzler Hitler and his proposals to regenerate the German people. In Germany a vast experiment is in hand, and some of you may live to see its results. If it fails it will not be for want of enthusiasm, but rather because the Germans are only just starting the study of mathematical statistics in the modern sense!". (Karl Pearson, 1934; in Filon et al. 1934, p. 23)

So here we stand. Karl Pearson –brilliant scientist, phenomenal writer, convinced socialist and freethinker– was about as racist as they come.

The Eugenicism of Sir Ronald Fisher

Sir Ronald Aylmer Fisher (1890-1962) was one of the greatest statisticians of all time.[9] However, Fisher was also stubborn, belligerent, and a eugenicist. When it comes to shocking remarks, one does not need to dig deep. We start with a remark from 1948, so *after* the Holocaust:

> "I have no doubt also that the [Nazi] Party sincerely wished to benefit the German racial stock, especially by the elimination of manifest defectives, such as those deficient mentally, and I do not doubt that von Verschuer gave, as I should have done, his support to such a movement." (Fisher, 1948; for details see Weiss 2010)

Moreover, in a dissenting opinion on the 1950 UNESCO report "The race question", Fisher argued that "Available scientific knowledge provides a firm basis for believing that the groups of mankind differ in their innate capacity for intellectual and emotional development".[10]

Galton, Pearson, and Fisher were unfortunately not the only prominent statisticians who supported eugenics. For instance, famous economist and Bayesian John Maynard Keynes still believed, in 1946 (!), that eugenics was "the most important, significant and, I would add, genuine branch of sociology which exists". Such statements permanent stain otherwise brilliant academic legacies.

Sir Ronald Aylmer Fisher (1890–1962) at 23 years of age. Public domain.

[9] Content partly based on the `BayesianSpectacles` blog post "This statement by Sir Ronald Fisher will shock you".

[10] See `http://unesdoc.unesco.org/images/0007/000733/073351eo.pdf`.

35 Statistical Analysis of the Binomial Distribution [with Quentin F. Gronau and Alexander Ly]

The binomial distribution is the Drosophila of statistics.

EJ and Dora, 2020

Chapter Goal

This chapter presents a brief statistical overview of Bayesian inference for a binomial chance parameter θ. The contents of this chapter can be safely skipped by pragmatic readers who care mostly about correct execution and proper interpretation rather than mathematical detail.

Overview

Below we first concentrate on parameter estimation and derive the posterior distribution for θ under the alternative hypothesis $\mathcal{H}_1$ that assigns θ a beta(α, β) prior distribution. Next we turn to hypothesis testing and derive the Bayes factor for the binomial test under various scenarios.

Posterior Distribution of θ under $\mathcal{H}_1$

Here we derive the posterior distribution for θ under the alternative hypothesis $\mathcal{H}_1$ which assigns θ a beta(α, β) prior. As shown in earlier chapters, after observing s successes out of n attempts (and $f = n - s$ failures) the posterior distribution of θ is given by:

$$\underbrace{p(\theta \mid s, f)}_{\substack{\text{Posterior for } \theta\text{:} \\ \text{beta}(\alpha+s, \beta+f)}} \propto \underbrace{p(\theta)}_{\substack{\text{Prior for } \theta\text{:} \\ \text{beta}(\alpha, \beta)}} \times \underbrace{p(s, f \mid \theta)}_{\substack{\text{Probability for } s, f \\ \text{given } \theta}} . \qquad (35.1)$$

In this chapter we take a closer look at how this result can be obtained. Recall that $p(s, f \mid \theta)$ is the binomial likelihood given by

$$p(s, f \mid \theta) = \binom{n}{s} \theta^s \, (1 - \theta)^f, \qquad (35.2)$$

where $n = s + f$ and $\binom{n}{s}$ is known as the *binomial coefficient* which gives the number of ways that s successes and f failures can be arranged

Howard Raiffa (1924–2016). In their book "Applied Statistical Decision Theory", Howard Raiffa and Robert Schlaifer introduced the concept of *conjugate* prior distributions. The beta prior for θ is conjugate to the binomial likelihood, because their combination produces a posterior for θ that is also a beta distribution. Harvard Business School Archives Photograph Collection.

in sequence. Specifically, $\binom{n}{s} = \frac{n!}{s!(n-s)!}$, where the exclamation mark denotes the factorial function: $k! = k \times (k-1) \times (k-2) \ldots \times 2 \times 1$.[1]

By $p(\theta)$ we denote the prior distribution for θ which in our case is a beta(α, β) distribution:

$$p(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\,\theta^{\alpha-1}\,(1-\theta)^{\beta-1}. \tag{35.3}$$

Here $\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$ is the *normalizing constant* of the beta(α, β) distribution that was omitted in the main text. $\Gamma(x)$ denotes the gamma function; for a positive integer k, $\Gamma(k)$ simplifies to $(k-1)!$.[2]

The normalizing constant ensures that the beta distribution integrates to one so that it is a proper probability density function. This means that we know that

$$1 = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \underbrace{\int_0^1 \theta^{\alpha-1}\,(1-\theta)^{\beta-1}\mathrm{d}\theta}_{=\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}}. \tag{35.4}$$

This integral –known as the Beta-integral, or the Beta function– will become important later.[3]

Returning to the derivation of the posterior distribution, we now only need to combine the binomial likelihood with the beta prior distribution, rearrange, and drop the terms that are constant with respect to θ to see that the posterior distribution is proportional to a beta$(\alpha+s, \beta+f)$ distribution as mentioned in the earlier chapters:

$$\begin{aligned} p(\theta \mid s, f) &\propto \underbrace{\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\,\theta^{\alpha-1}\,(1-\theta)^{\beta-1}}_{p(\theta)} \times \underbrace{\binom{n}{s}\,\theta^{s}\,(1-\theta)^{f}}_{p(s,f\mid\theta)} \\ &\propto \theta^{\alpha+s-1}\,(1-\theta)^{\beta+f-1}. \end{aligned} \tag{35.5}$$

Evidence

To assess the evidence that the data provide for rival hypotheses, we need to compute their predictive performance. Below we consider three scenarios: point versus point (i.e., the likelihood ratio), point versus distribution (i.e., the standard Bayesian hypothesis test), and distribution versus distribution.

Case I. Point versus point: The likelihood ratio

As stated in earlier chapters, the Bayes factor is defined as

$$\mathrm{BF}_{10} = \frac{p(s, f \mid \mathcal{H}_1)}{p(s, f \mid \mathcal{H}_0)}. \tag{35.6}$$

The probability of the data given the point null hypothesis $\mathcal{H}_0$ is simply the binomial likelihood where we insert the test value θ_0 for θ. Hence,

$$p(s, f \mid \mathcal{H}_0) = \binom{n}{s}\,\theta_0^{s}\,(1-\theta_0)^{f}. \tag{35.7}$$

[1] For details see the earlier chapter 'Jevons Explains Permutations'.

[2] In general, the gamma function interpolates the factorial function and is defined as $\Gamma(t) = \int_0^\infty x^{t-1}e^{-x}\mathrm{d}x$. For more details see `https://en.wikipedia.org/wiki/Gamma_function`.

[3] The Beta-integral occurs relatively often. "This standard result should be learnt if not already known, as it is frequently needed in statistical calculations." (Lindley 1965, p. 39)

Similarly, when $\mathcal{H}_1$ is defined as a rival point value θ_1, we have

$$p(s, f \mid \mathcal{H}_1) = \binom{n}{s} \theta_1^s \, (1-\theta_1)^f. \tag{35.8}$$

In the case of two point hypotheses, the Bayes factor BF_{10} is known as the likelihood ratio LR_{10}. Dividing the probabilities that $\mathcal{H}_1 : \theta = \theta_1$ and $\mathcal{H}_0 : \theta = \theta_0$ assign to the observed data we obtain

$$\text{LR}_{10} = \left[\frac{\theta_1}{\theta_0}\right]^s \times \left[\frac{1-\theta_1}{1-\theta_0}\right]^f, \tag{35.9}$$

such that the occurrence of any single success multiplies the likelihood ratio by θ_1/θ_0, whereas the occurrence of any single failure multiplies the likelihood ratio by $(1-\theta_1)/(1-\theta_0)$. For a demonstration see Chapter 7.

Andrew Gelman (1965–). A frequent blogger and arguably the world's most influential statistician, Andrew Gelman is not known for mincing words. A footnote to a paper that we have co-authored with him reads: "Andrew Gelman wishes to state that he hates Bayes factors". In contrast, we love Bayes factors; throughout this book we will use concrete examples to demonstrate their worth.

Case II. Point versus distribution: The standard hypothesis test

In this subsection we consider three scenarios of increasing generality: the simplest scenario features a test between the null hypothesis $\mathcal{H}_0 : \theta = 1/2$ versus an alternative hypothesis $\mathcal{H}_1$ that assigns θ a uniform prior distribution; the intermediate scenario features a test between the null hypothesis $\mathcal{H}_0 : \theta = 1/2$ against an alternative hypothesis $\mathcal{H}_1$ that assigns θ a beta(α, β) prior distribution; the most general scenario features a test between a null hypothesis $\mathcal{H}_0 : \theta = \theta_0$ (where θ_0 corresponds to any test value in the interval from 0 to 1) versus an alternative hypothesis $\mathcal{H}_1$ that assigns θ a beta(α, β) prior distribution.

Now we derive the Bayes factor for the three scenarios. It is easiest to start with the most general case, that is, the Bayes factor for testing whether $\theta = \theta_0$ where the alternative hypothesis $\mathcal{H}_1$ specifies a beta(α, β) prior distribution for θ; afterwards, we will outline the simplifications that can be made for the other two cases.

In the previous subsection we defined the Bayes factor and gave the probability of the data under a point null hypothesis $\mathcal{H}_0 : \theta = \theta_0$. In order to obtain the probability of the data under the alternative hypothesis $\mathcal{H}_1 : \theta \sim$ beta(α, β), we use the *law of total probability*, as described in Chapter 3, 'The Rules of Probability'. Lindley called this theorem an *extension of the conversation*. "Let E_1 and E_2 be two events which are exclusive and exhaustive, and let A be any event. Then (...) $p(A) = p(A \mid E_1)p(E_1) + p(A \mid E_2)p(E_2)$." (Lindley 1985, p. 39). Applying the law of total probability, we obtain

$$\begin{aligned} p(s, f \mid \mathcal{H}_1) &= \int_0^1 p(s, f \mid \theta, \mathcal{H}_1) \, p(\theta \mid \mathcal{H}_1) \mathrm{d}\theta \\ &= \int_0^1 \underbrace{\binom{n}{s} \theta^s \, (1-\theta)^f}_{p(s,f \mid \theta, \mathcal{H}_1)} \, \underbrace{\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1} \, (1-\theta)^{\beta-1}}_{p(\theta \mid \mathcal{H}_1)} \mathrm{d}\theta. \end{aligned} \tag{35.10}$$

Here $p(s, f \mid \theta, \mathcal{H}_1)$ is simply the binomial likelihood and $p(\theta \mid \mathcal{H}_1)$ denotes the beta prior distribution for θ under $\mathcal{H}_1$.

Next, we use our knowledge about the integral (as shown in Equation 35.4) to simplify the expression for $p(s, f \mid \mathcal{H}_1)$ as follows:

$$\begin{aligned} p(s, f \mid \mathcal{H}_1) &= \binom{n}{s} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 \theta^{\alpha+s-1} \, (1-\theta)^{\beta+f-1} \mathrm{d}\theta \\ &= \binom{n}{s} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+s)\Gamma(\beta+f)}{\Gamma(\alpha+\beta+n)}. \end{aligned} \tag{35.11}$$

Hence, the Bayes factor for testing the hypothesis $\mathcal{H}_0 : \theta = \theta_0$ where θ_0 corresponds to any test value in the interval [0,1] against an alternative hypothesis $\mathcal{H}_1$ that specifies a beta(α, β) prior distribution for θ is given by:

$$\mathrm{BF}_{10} = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+s)\Gamma(\beta+f)}{\Gamma(\alpha+\beta+n)} \frac{1}{\theta_0^s \, (1-\theta_0)^f}. \tag{35.12}$$

The Bayes factor for testing the hypothesis $\mathcal{H}_0 : \theta = 1/2$ against an alternative hypothesis $\mathcal{H}_1$ that specifies a beta(α, β) prior distribution for θ is obtained by setting $\theta_0 = 1/2$ in Equation 35.12, resulting in:

$$\mathrm{BF}_{10} = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+s)\Gamma(\beta+f)}{\Gamma(\alpha+\beta+n)} \, 2^n. \tag{35.13}$$

The Bayes factor for testing the hypothesis $\mathcal{H}_0 : \theta = 1/2$ against an alternative hypothesis $\mathcal{H}_1$ that specifies a uniform prior distribution for θ is obtained by setting the two parameters α and β of the beta prior distribution equal to 1. For positive integer k we replace $\Gamma(k)$ by $(k-1)!$ and obtain the following Bayes factor:

$$\mathrm{BF}_{10} = \frac{s!f!}{(n+1)!} \, 2^n. \tag{35.14}$$

Case III. Distribution versus distribution: Ly's limit

In Chapter 12, 'The Pancake Puzzle', we pitted against one another several forecasters who each quantified their prior beliefs about θ by means of a beta distribution. Let $\theta_1 \sim$ beta(α_1, β_1) be the prior distribution for forecaster 1, and $\theta_2 \sim$ beta(α_2, β_2) the prior distribution for forecaster 2. The Bayes factor for forecaster 1 over forecaster 2 is then

$$\mathrm{BF}_{12} = \frac{B(\alpha_1+s, \beta_1+f)}{B(\alpha_2+s, \beta_2+f)} \frac{B(\alpha_2, \beta_2)}{B(\alpha_1, \beta_1)}, \tag{35.15}$$

where B is the beta integral encountered earlier. We may wonder what happens to the evidence when the data increase in size (i.e., $n \to \infty$) but the sample proportion s/n stays the same and equals a true value $\theta^\star$. In other words, $s = \theta^\star n$ and $n \to \infty$. Intuitively, as the data accumulate, the two beta distributions converge to a highly similar posterior distribution, and from that point onward the models will make virtually identical predictions. This suggests that there is a bound on the evidence that can be obtained when the rival hypothesis both allow θ to vary across the same range (cf. Chapter 13). The specific limit is:

$$\begin{aligned} \lim_{n\to\infty} \mathrm{BF}_{12}(s, n) &= \lim_{n\to\infty} \frac{B(\alpha_1+s, \beta_1+f)}{B(\alpha_2+s, \beta_2+f)} \frac{B(\alpha_2, \beta_2)}{B(\alpha_1, \beta_1)} \\ &= \theta^{\alpha_1-\alpha_2}(1-\theta)^{\beta_1-\beta_2} \frac{B(\alpha_2, \beta_2)}{B(\alpha_1, \beta_1)}, \end{aligned} \tag{35.16}$$

as follows from Stirling's approximation to the factorial: $\log n! = (n + \frac{1}{2}) \log n - n + \frac{1}{2} \log 2\pi + \frac{1}{12n} - O\left(\frac{1}{n^3}\right)$.

Ly's limit can also be given a visual interpretation (cf. Ly and Wagenmakers 2022; Morey and Rouder 2011, pp. 411-412; see also Jeffreys 1961, p. 367; Jeffreys 1973, p. 39). Specifically, the limit equals the ratio of the prior ordinates at the true value $\theta^\star$, that is,

$$\lim_{n \to \infty} \mathrm{BF}_{12}(s, n) = \frac{p(\theta^\star \mid \text{beta}(\alpha_1, \beta_1))}{p(\theta^\star \mid \text{beta}(\alpha_2, \beta_2))}. \tag{35.17}$$

An exception to this rule occurs when all parameters (i.e., $\alpha_1, \beta_1, \alpha_2, \beta_2$) are 2 or larger and $\theta^\star = 1$ or $\theta^\star = 0$, that is, only successes or only failures are observed. Without loss of generality we consider the case of $\theta^\star = 1$. Then the posterior for θ equals $\theta_1 \sim$ beta$(\alpha_1 + s, \beta_1)$ under forecaster 1 and $\theta_2 \sim$ beta$(\alpha_2 + s, \beta_2)$ under forecaster 2. The data 's' affect the α parameter but not the β parameter. Consequently, a difference in the β parameters leads the Bayes factor to increase indefinitely: if $\beta_1 < \beta_2$, then $\mathrm{BF}_{12} \to \infty$ as $s = n \to \infty$; if $\beta_1 > \beta_2$, then $\mathrm{BF}_{21} \to \infty$ as $s = n \to \infty$; only if $\beta_1 = \beta_2$ is there a limit on the Bayes factor. For example, consider the case where $s = n = 1{,}000{,}000$. If forecaster 1 specifies $\alpha_1 = 2$, $\beta_1 = 3$ and forecaster 2 specifies $\alpha_2 = 2$, $\beta_2 = 4$ then $\mathrm{BF}_{12} = 200{,}001$ (this keeps increasing as $s = n$ grows). When forecaster 2 specifies $\alpha_2 = 2$, $\beta_2 = 2$, however, then $\mathrm{BF}_{21} = 250{,}001$ (again, this keeps increasing as $s = n$ grows). And when forecaster 2 specifies $\alpha_2 = 3$, $\beta_2 = 3$ then $\mathrm{BF}_{21} = 2.5$ (which does not increase as $s = n$ grows).

Exercises

1. Ly's limit equals the ratio of the prior ordinates at the true value $\theta^\star$. Use the Savage-Dickey density ratio to argue why this must be the case.

Want to Know More?

✓ Ly, A., & Wagenmakers, E.-J. (2022). Bayes factors for peri-null hypotheses. *TEST, 31*, 1121–1142.

36 Recommended Readings

[Edwards et al., 1963] proposed that experimenters use Bayesian statistics (...) [this] was a complete flop, since the experimenters already had their statistics.

Gigerenzer et al., 1989

Chapter Goal

This chapter presents a lightly annotated list of Bayesian books and articles that we find particularly insightful or inspiring. The selection is heavily biased towards the inclusion of works that can be understood by those without a degree in mathematical statistics.[1]

[1] If your institution does not carry access to a particular scientific article, you may be tempted to visit the illegal website "Sci-Hub". In our opinion, Sci-Hub is righting a moral wrong. Their adage is "to remove all barriers in the way of science".

Recommendations

We start our reading list with an article that itself presents an annotated reading list:

✓ Etz, A., Gronau, Q. F., Dablander, F., Edelsbrunner, P. A., & Baribault, B. (2018). How to become a Bayesian in eight easy steps: An annotated reading list. *Psychonomic Bulletin & Review, 25*, 219-234. All of Alexander Etz's articles on Bayesian inference are exceptionally clear and we recommend beginning Bayesians browse his blog posts at `https://alexanderetz.com/understanding-bayes/`.

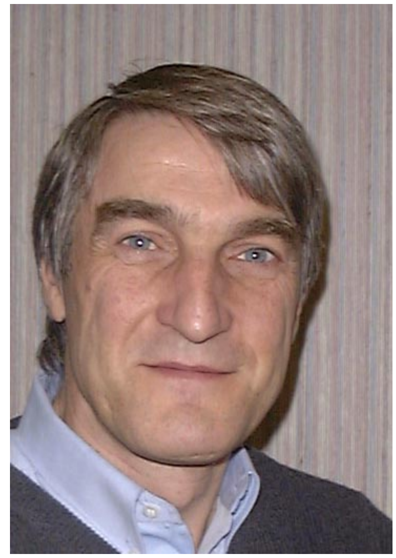
For a historical introduction we suggest the following two works:

✓ Howie, D. (2002). *Interpreting Probability: Controversies and Developments in the Early Twentieth Century*. Cambridge: Cambridge University Press. An in-depth overview of the debate between the Bayesian Harold Jeffreys and the frequentist Ronald Fisher. Some background knowledge of statistics is required to understand the finer details.

✓ McGrayne, S. B. (2011). *The Theory that Would not Die: How Bayes' Rule Cracked the Enigma Code, Hunted Down Russian Submarines, and Emerged Triumphant from Two Centuries of Controversy*. New Haven, CT: Yale University Press. The title says it all. Highly recommended.

For a discussion of foundational issues our list of recommended readings is relatively long:

- ✓ Edwards, W., Lindman, H. & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review, 70*, 193-242. A classic article that is even more relevant today than when it was first published. Unfortunately a full understanding of the article does require a background in statistics. Consider skipping the first sections and persist – it is worth it.
- ✓ O'Hagan, A. (2004). Dicing with the unknown. *Significance, 1*, 132-133. O'Hagan explains the difference between aleatory uncertainty (due to randomness) and epistemic uncertainty (due to lack of knowledge). Highly recommended.
- ✓ Eagle, A. (Ed.) (2011). *Philosophy of Probability: Contemporary Readings*. New York: Routledge. All you ever wanted to know about probability, and much, much more.
- ✓ Dienes, Z. (2008). *Understanding Psychology as a Science: An Introduction to Scientific and Statistical Inference*. New York: Palgrave Macmillan. An easy-to-understand introduction to inference that summarizes the differences between the various schools of statistics. No knowledge of mathematical statistics is required.
- ✓ Royall, R. M. (1997). *Statistical Evidence: A Likelihood Paradigm*. London: Chapman & Hall. Similar in spirit to the Dienes book, this book requires a little more knowledge of statistics to be properly understood.
- ✓ Lindley, D. V. (2000). The philosophy of statistics. *The Statistician, 49*, 293-337. The general rule is to read anything that Lindley has written. Appreciation of the content does require background knowledge.
- ✓ Lindley, D. V. (1993). The analysis of experimental data: The appreciation of tea and wine. *Teaching Statistics, 15*, 22-25. Whenever students ask us for accessible articles on Bayesian versus frequentist statistics, this one tops our list.
- ✓ Pek, J., & and Van Zandt, T. (2020). Frequentist and Bayesian approaches to data analysis: Evaluation and estimation. *Psychology Learning & Teaching, 19*, 21-35. "This article reviews frequentist and Bayesian approaches such that teachers can promote less well-known statistical perspectives to encourage statistical thinking. Within the frequentist and Bayesian approaches, we highlight important distinctions between statistical evaluation versus estimation using an example on the facial feedback hypothesis." (p. 21)
- ✓ Lindley, D. V. (2004). That wretched prior. *Significance, 1*, 85-87. "Objectivity is merely subjectivity when nearly everyone agrees" (p. 87).
- ✓ Berger, J. O., & Wolpert, R. L. (1988). *The Likelihood Principle* (2nd edn.). Hayward, CA: Institute of Mathematical Statistics. The contents of this book is as terrific as its typesetting is terrible. Does require a solid background in mathematical statistics.
- ✓ Berger, J. O., & Berry, D. A. (1988). Statistical analysis and the illusion of objectivity. *American Scientist, 76*, 159-165. An accessible article on the inherent subjectivity of statistical analysis.



Anthony O'Hagan (1948–). "Every statistician needs to understand the difference between the frequentist and Bayesian theories of statistics, and every practising statistician must (at least implicitly) choose between them. And whether something is unknown or unknowable, whether its uncertainty is due to fundamentally unpredictable randomness or to potentially resolvable lack of knowledge, turns out to lie at the heart of the debate".

Jim Berger (1950–).

✓ Bayarri, M. J., & Berger, J. O. (2013). Hypothesis testing and model uncertainty. In Damien, P., Dellaportas, P., Polson, N. G., & Stephens, D. A. (Eds.), *Bayesian Theory and Applications*, pp. 361-400. Oxford: Oxford University Press. When we interviewed Jim Berger in 2017, we asked "If you could give an applied researcher (say in biology or psychology) a single one of your papers to read, which one would that be, and why?" Berger then pointed to this book chapter[2] and explained: "This was written to explain the key issues in testing and model uncertainty, using the best approaches and examples I had seen or developed over many years. So I think it is a good introduction to these issues for someone who actually cares."[3]

[2] Unfortunately, the chapter is difficult to find online.

[3] The complete interview is at `https://jasp-stats.org/2017/07/27/jimberger/`.

✓ Rosenkrantz, R. D. (1977). *Inference, Method and Decision*. Dordrecht: Reidel.

✓ Rouder, J. N., Morey, R. D., Verhagen, A. J., Province, J. M., & Wagenmakers, E.-J. (2016). Is there a free lunch in inference? *Topics in Cognitive Science, 8*, 520-547. The answer is 'no'.

✓ Etz, A., Haaf, J. M., Rouder, J. N., & Vandekerckhove, J. (2018). Bayesian inference and testing any hypothesis you can specify. *Advances in Methods and Practices in Psychological Science, 1*, 281–295.

✓ Howson, C., & Urbach, P. (2006). *Scientific Reasoning: The Bayesian Approach* (3rd edn.). Chicago, IL: Open Court. An informative and entertaining introduction to Bayesian reasoning. Highly recommended.

✓ Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review, 14*, 779-804. Summarizes the statistical problems with p values as indicated in Berger and Wolpert (1988) and proposes the BIC (Bayesian Information Criterion; an approximation to the Bayes factor hypothesis test) as a solution.

✓ Wagenmakers, E.–J., Marsman, M., Jamil, T., Ly, A., Verhagen, A. J., Love, J., Selker, R., Gronau, Q. F., Šmíra, M., Epskamp, S., Matzke, D., Rouder, J. N., & Morey, R. D. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review, 25*, 35-57. An update to the 2007 paper, with a role for JASP.

✓ Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E.-J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review, 23*, 103-123. A confidence interval may be even more difficult to interpret than a p value.

Richard D. Morey (1978–). " confidence intervals should not be used as modern proponents suggest".

For an accessible introduction to Bayesian methods more generally we recommend:

✓ Lindley, D. V. (1985). *Making Decisions* (2nd edn.). London: Wiley. Simple, straightforward, and compelling. A must-read.

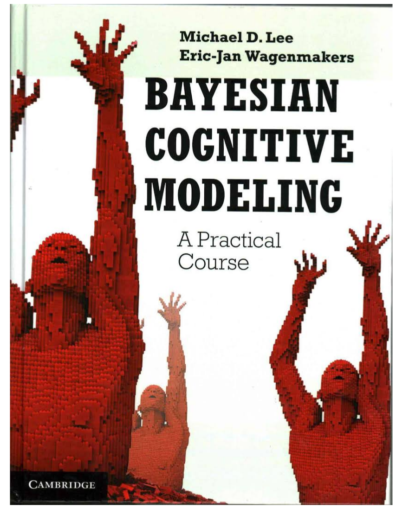
✓ Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review, 4*, 79-95. A breakthrough article for psychology, explaining how Bayesian model selection balances the conflicting demands of parsimony and goodness-of-fit.

✓ Lindley, D. V. (2006). *Understanding Uncertainty*. Hoboken: Wiley. If every student had to read this book, the world would be a better place.

✓ Bolstad, W. M. (2007). *Introduction to Bayesian Statistics* (2nd edn.). Hoboken, NJ: Wiley. This is a real introduction, not a pretend one.

✓ Albert, J. (2009). *Bayesian Computation with R* (2nd edn.). Dordrecht, The Netherlands: Springer. This introductory text is supported by the R package 'LearnBayes' (not to be confused with the eponymous JASP module).

✓ Lee, M. D., & Wagenmakers, E.-J. (2013). *Bayesian Cognitive Modeling: A Practical Course*. Cambridge: Cambridge University Press. A hands-on book with many examples.



The cover of Bayesian Cognitive Modeling, featuring "red" by lego-artist Nathan Sawaya (for more examples see `http://www.brickartist.com/`).

✓ Gelman, A., & Hill, J. (2014). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press. The standard introductory text to hierarchical modeling. It is still worth reading after pouring a cup of coffee over it and then leaving it outside in the rain for a night. Robust stuff.

✓ Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian Data Analysis* (3rd edn.). Boca Raton, FL: Chapman & Hall/CRC. A modern-day 650+ page classic on Bayesian parameter estimation.

✓ Kruschke, J. K. (2015). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan* (2nd edn.). Academic Press/Elsevier. Many students find John Kruschke's style appealing and helpful. Consistent with this conjecture, the first student who borrowed the book from the JASP team has never returned it.[4]

[4] We disagree with Kruschke about Bayes factors (we like them, he dislikes them), and his "ROPE" alternative (we dislike it, he likes it). However, we do agree with Kruschke about the fundamentals and we appreciate what he has done to popularize Bayesian inference in psychology.

✓ McElreath, R. (2016). *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. Boca Raton, FL: Chapman & Hall/CRC Press. Hailed by Rasmus Bååth as a "pedagogical masterpiece". In the style of Gelman and Kruschke, the book prioritizes parameter estimation over model selection.

✓ Vandekerckhove, J., Rouder, J. N., & Kruschke, J. K. (2018). Editorial: Bayesian methods for advancing psychological science. *Psychonomic Bulletin & Review, 25*, 1-4. Most articles in this special issue are tutorial-style works of art.

Jeffrey N. Rouder (1966–). "Progress in science often comes from discovering invariances in relationships among variables; these invariances often correspond to null hypotheses."

✓ Donovan, T. M., & Mickey, R. M. (2019). *Bayesian Statistics for Beginners: A Step–by–Step Approach*. Oxford: Oxford University Press.

✓ Kurt, W. (2019). *Bayesian Statistics the Fun Way*. San Francisco: No Starch Press. As the title suggests, this book sparks joy. A more detailed review can be found on BayesianSpectacles.org.

✓ Hudson, T. E. (2021). *Bayesian Data Analysis for the Behavioral and Neural Sciences*. Cambridge: Cambridge University Press.

✓ Clayton, A. (2021). *Bernoulli's Fallacy: Statistical Illogic and the Crisis of Modern Science*. New York: Columbia University Press. "Consider this, instead, a piece of wartime propaganda, designed to be printed on leaflets and dropped from planes over enemy territory to win the hearts and minds of those who may as yet be uncommitted to one side or the other. My goal with this book is not to broker a peace treaty; my goal is to win the war." (p. xv)

✓ Bozza, S., Taroni, F., & Biedermann, A. (2022). *Bayes Factors for Forensic Decision Analyses with R*. New York: Springer. "The assessment of the value of scientific evidence involves subtle forensic, statistical, and computational aspects that can represent an obstacle in practical applications. The purpose of this book is to provide theory, examples, and elements of R code to illustrate a variety of topics pertaining to value of evidence assessments using Bayes factors in a decision-theoretic perspective." (p. 1) The book is freely available online.

✓ Lambert, B. (2018). *A Student's Guide to Bayesian Statistics*. London: Sage.

✓ Ma, W. J., Kording, K. P., & Goldreich, D. (in press). *Bayesian Models of Perception and Action: An Introduction*. Cambridge, MA: MIT Press. Freely available at `https://www.cns.nyu.edu/malab/bayesianbook.html`.

✓ van Doorn, J. B. (2023). *A Brief Introduction to Bayesian Inference: From Tea to Beer*. Freely available at `https://johnnydoorn.github.io/BayesBookQuarto/`

> "This booklet offers an introduction to Bayesian inference. We look at how different models make different claims about a parameter, how they learn from observed data, and how we can compare these models to each other. We illustrate these ideas through an informal beer-tasting experiment conducted at the University of Amsterdam."

✓ Sprenger, J., & Hartmann, S. (2019). *Bayesian Philosophy of Science*. Oxford: Oxford University Press. A philosophical perspective on Bayesian inference.

✓ Schupbach, J. N. (2022). *Bayesianism and Scientific Reasoning*. Cambridge: Cambridge University Press. Another philosophical perspective on Bayesian inference.

✓ Wagenmakers, E.–J. (2020). *Bayesian Thinking for Toddlers*. Freely available at `https://psyarxiv.com/w5vbp/`.

Finally, we succumb to temptation and provide three recommended readings that, for their proper appreciation, may actually require that degree in mathematical statistics:

✓ Jeffreys, H. (1961). *Theory of Probability* (3rd ed.). Oxford: Oxford University Press. The most impressive work on statistical inference published in the 20th century.

✓ O'Hagan, A., & Forster, J. (2004). *Kendall's Advanced Theory of Statistics Vol. 2B: Bayesian Inference* (2nd ed.). London: Arnold. An invaluable and timeless resource.

✓ Jaynes, E. T. (2003). *Probability Theory: The Logic of Science*. Cambridge: Cambridge University Press. Jaynes's expressive writing style and clarity of thought has resulted in somewhat of a cult following. There are worse cults one could belong to.

Finally, now that we have succumbed anyway, the following two influential books deserve to be mentioned as well:

✓ Marin, J.–M., & Robert, C. P. (2007). *Bayesian Core: A Practical Approach to Computational Bayesian Statistics*. Springer: New York.

✓ Robert, C. P. (2007). *The Bayesian Choice: From Decision–Theoretic Foundations to Computational Implementation* (2nd ed.). Springer: New York.

We apologize if your favorite Bayesian resource is not listed – please attend us to this omission and we may include it in a next edition.

37 Figure Listing

Preface

Figure "(Not) Thomas Bayes": Image on Wikipedia, taken from https://nl.wikipedia.org/wiki/Thomas_Bayes#/media/Bestand:Thomas_Bayes.gif under a CC-PD license (https://creativecommons.org/publicdomain/mark/1.0/). 9

Figure "Laplace Portrait": Image by Paulin Guérin, taken from https://fr.wikipedia.org/wiki/Fichier:Pierre-Simon,_marquis_de_Laplace_(1745-1827)_-_Gu%C3%A9rin.jpg under a CC-PD license (https://creativecommons.org/publicdomain/mark/1.0/). 11

Figure "Viktor Beekman": Image by Viktor Beekman, taken from https://www.bayesianspectacles.org under a CC-BY license (https://creativecommons.org/licenses/by/4.0/legalcode). 12

Figure "Erasmus+ Programme": Image provided by the European Union, taken from https://tinyurl.com/ErasmusLogo. "The European Commission's support for the production of this publication does not constitute an endorsement of the contents, which reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein." 12

Synopsis

Figure 1: Image by Viktor Beekman, taken from https://www.bayesianspectacles.org under a CC-BY license (https://creativecommons.org/licenses/by/4.0/legalcode). 14

Figure "Never assert absolutely": Image by Viktor Beekman, taken from https://www.bayesianspectacles.org under a CC-BY license (https://creativecommons.org/licenses/by/4.0/legalcode). 16

Figure "Bayesian Thinking for Toddlers": Image by Viktor Beekman, taken from https://psyarxiv.com/w5vbp/ under a CC-BY license (https://creativecommons.org/licenses/by/4.0/legalcode). 17

JASP

Figure "JASP Logo": Taken from https://jasp-stats.org/jasp-materials/. 19

Figure "Bayesian Inference Is Hard": Image by Viktor Beekman, taken from https://www.bayesianspectacles.org under a CC-BY license (https://creativecommons.org/licenses/by/4.0/legalcode). 16

Figure "JASP Website": Screenshot taken from https://jasp-stats.org/. 21

Figure "JASP Coat of Arms": Image by Viktor Beekman, taken from https://www.bayesianspectacles.org under a CC-BY license (https://creativecommons.org/licenses/by/4.0/legalcode). 21

Figure 2: Screenshot taken from JASP, available at https://jasp-stats.org/. 23

Figure "Anscombosaurus": Image taken from https://osf.io/m6bi8/ under a CC-BY license (https://creativecommons.org/licenses/by/4.0/legalcode). 23

Figure "JASP World Map": Figure taken from https://jasp-stats.org/teaching-with-jasp/. 25

Figure "Trojan Horse": Image by Viktor Beekman, taken from https://www.bayesianspectacles.org under a CC-BY license (https://creativecommons.org/licenses/by/4.0/legalcode). 27

Probability Belongs Wholly to the Mind?

Figure "Lambert Wilson": Image by Georges Biard, taken from https://commons.wikimedia.org/wiki/File:Lambert_Wilson_Avp_2015.jpg under a CC-BY-SA 3.0 license (https://creativecommons.org/licenses/by-sa/3.0/deed.en). 31

Figure "Young Jevons": Image by National Portrait Gallery London, taken from https://www.npg.org.uk/collections/search/portrait/mw135455/William-Stanley-Jevons under a CC-BY-ND 2.0 license (https://creativecommons.org/licenses/by-nd/2.0/). 33

Figure "Logic Piano": Image from History of Science Museum, University of Oxford. Usage granted until 2031. 33

Figure 1.1: Image by Viktor Beekman, taken from https://www.bayesianspectacles.org/ under a CC-BY license (https://creativecommons.org/licenses/by/4.0/legalcode). 36

Figure "Probability Belongs Wholly to the Mind": Image by Viktor Beekman, taken from https://www.bayesianspectacles.org under a CC-BY license (https://creativecommons.org/licenses/by/4.0/legalcode). 39

Figure "Photo Hossenfelder": Photo copyright by Dr. Sabine Hossenfelder, taken from https://upload.wikimedia.org/wikipedia/commons/thumb/0/0c/Sabine_Hossenfelder.jpg/330px-Sabine_Hossenfelder.jpg under a CC BY-SA 4.0 license (https://creativecommons.org/licenses/by-sa/4.0/). No changes were made. 40

Figure "Portrait Jevons": Image by G. J. Stodart, taken from https://commons.wikimedia.org/wiki/File:William_Stanley_Jevons.jpg under a CC-PD license (https://creativecommons.org/publicdomain/mark/1.0/). 40

Figure "Schopenhauer": Image by J. Schäfer, taken from https://commons.wikimedia.org/wiki/File:Arthur_Schopenhauer_by_J_Sch%

C3%A4fer,_1859b.jpg under a CC-PD license (https://creativecommons.org/publicdomain/mark/1.0/). 41

Epistemic and Aleatory Uncertainty

Figure "De Morgan": Image taken from https://commons.wikimedia.org/wiki/File:Augustus_De_Morgan.jpg under a CC-PD license (https://creativecommons.org/publicdomain/mark/1.0/). 46

Figure "de Finetti": Image taken from http://www.brunodefinetti.it/ with permission from Fulvia de Finetti. 46

Figure 2.1: Figure generated by the JASP module 'Learn Bayes', available at https://jasp-stats.org/. 49

Figure 2.2: Screenshot taken from the JASP module 'Learn Bayes', available at https://jasp-stats.org/. 50

The Rules of Probability

Figure 3.1: Image by Viktor Beekman, taken from https://www.bayesianspectacles.org/ under a CC-BY license (https://creativecommons.org/licenses/by/4.0/legalcode). 58

Figure 3.2: Image by Viktor Beekman, taken from https://www.bayesianspectacles.org/ under a CC-BY license (https://creativecommons.org/licenses/by/4.0/legalcode). 59

Figure 3.3: Image by Viktor Beekman, taken from https://www.bayesianspectacles.org/ under a CC-BY license (https://creativecommons.org/licenses/by/4.0/legalcode). 61

Figure 3.4:Graph created in R, code taken from http://shinyapps.org/apps/RGraphCompendium. 64

Figure 3.5: Image on Project Gutenberg, taken from https://www.gutenberg.org/files/33283/33283-pdf.pdf under Gutenberg-TM License (https://www.gutenberg.org/license). 65

Figure 3.6: Image by Viktor Beekman, taken from https://www.bayesianspectacles.org/ under a CC-BY license (https://creativecommons.org/licenses/by/4.0/legalcode). 66

Figure 3.7: Figure created using R. 67

Figure "Evidence": Image by Viktor Beekman, taken from https://www.bayesianspectacles.org under a CC-BY license (https://creativecommons.org/licenses/by/4.0/legalcode). 68

Figure 3.8: Image by C. M. G. Lee, taken from https://commons.wikimedia.org/wiki/File:Probability_vs_odds.svg under a CC BY-SA 4.0 license (https://creativecommons.org/licenses/by-sa/4.0/deed.en). Figure resolution enhanced by Henrik Godmann. 69

Figure "Laws of Probability": Image by Viktor Beekman, taken from https://www.bayesianspectacles.org/ under a CC-BY license (https://creativecommons.org/licenses/by/4.0/legalcode). 76

Interlude: Leibniz's Blunder

Figure "Portrait Leibniz": Image by Christoph Bernhard Francke, taken from https://commons.wikimedia.org/wiki/File:Christoph_Bernhard_Francke_-_Bildnis_des_Philosophen_Leibniz_(ca._1695).jpg under a CC-PD license (https://creativecommons.org/publicdomain/mark/1.0/). 81

Figure 4.1: Image by Tim Stellmach, taken from https://commons.wikimedia.org/wiki/File:Dice_Distribution_(bar).svg, "Released into the public domain to the fullest extent legally possible." 85

Figure 4.2: Image by Kolossos, taken from https://commons.wikimedia.org/wiki/File:Leibnitzrechenmaschine.jpg under a CC-BY-SA 3.0 license (https://creativecommons.org/licenses/by-sa/3.0/deed.en). 87

The Measurement of Probability

Figure 5.1: Image by Andrew Mauboussin and Michael Mauboussin. Reprinted with permission. 89

Figure "Dennis Lindley": Included by permission of Janet, Rowan, and Robert Lindley. 91

Figure "Borel": Image from Bibliothèque nationale de France under a CC-PD license (https://creativecommons.org/publicdomain/mark/1.0/). 92

Figure "De Morgan's Title Page": Image taken from https://archive.org/details/essayonprobabili00demo under a CC-PD license (https://creativecommons.org/publicdomain/mark/1.0/). 93

Figure "Frank Plumpton Ramsey": Image by Volsav, taken from https://commons.wikimedia.org/wiki/File:30._Frank_Ramsey.jpg under a CC-BY-SA 4.0 license (https://creativecommons.org/licenses/by-sa/4.0/deed.en). 95

Figure 5.2: Image by Viktor Beekman, taken from https://www.bayesianspectacles.org/ under a CC-BY license (https://creativecommons.org/licenses/by/4.0/legalcode). 95

Coherence

Figure "Aristotle": Image taken from https://en.wikipedia.org/wiki/Aristotle#/media/File:Francesco_Hayez_001.jpg under a CC-PD license (https://creativecommons.org/publicdomain/mark/1.0/). 101

Figure 6.1: Image by Viktor Beekman, taken from https://www.bayesianspectacles.org/ under a CC-BY license (https://creativecommons.org/licenses/by/4.0/legalcode). 108

Figure "Be Coherent": Image by Viktor Beekman, taken from https://www.bayesianspectacles.org/ under a CC-BY license (https://creativecommons.org/licenses/by/4.0/legalcode). 111

Figure "The Calculus of Probability": Image by Viktor Beekman, taken from https://www.bayesianspectacles.org/ under a CC-BY license (https://creativecommons.org/licenses/by/4.0/legalcode). 116

Learning from the Likelihood Ratio

Figure "Bayes' Rule Bib": Image by EJ. The source of the bib is unknown. 119

Figure "Pancake Stack": Pancakes (and image) by EJ. 120

Figure 7.1: Figure created using R. 120

Figure 32.2: Figure created using R. 501

Figure 7.3: Figure created using R. 124

Figure "Alexander Hamilton": Portrait by John Trumbull, taken from https://commons.wikimedia.org/wiki/File:Alexander_Hamilton_portrait_by_John_Trumbull_1806.jpg under a CC-PD license (https://creativecommons.org/publicdomain/mark/1.0/). 126

Figure "James Madison": Portrait by John Vanderlyn, taken from https://commons.wikimedia.org/wiki/File:James_Madison(cropped)(c).jpg under a CC-PD license (https://creativecommons.org/publicdomain/mark/1.0/). 127

Figure 7.4: Image by Viktor Beekman, taken from https://www.bayesianspectacles.org/ under a CC-BY license (https://creativecommons.org/licenses/by/4.0/legalcode). 129

Figure 7: Image by Viktor Beekman, taken from https://www.bayesianspectacles.org/ under a CC-BY license (https://creativecommons.org/licenses/by/4.0/legalcode). 133

An Infinite Number of Hypotheses

Figure 8.1: Image by Viktor Beekman, taken from https://www.bayesianspectacles.org/ under a CC-BY license (https://creativecommons.org/licenses/by/4.0/legalcode). 136

Figure 8.2: Figure created using R. 138

Figure 8.3: Figure generated by the JASP module 'Learn Bayes', available at https://jasp-stats.org/. 140

Figure 8.4: Figure generated by the JASP module 'Learn Bayes', available at https://jasp-stats.org/. 141

Figure 8.5: Figure generated by the JASP module 'Learn Bayes', available at https://jasp-stats.org/. 142

Figure 8.6: Figure generated by the JASP module 'Learn Bayes', available at https://jasp-stats.org/. 143

Figure 8.7: Figure generated by the JASP module 'Learn Bayes', available at https://jasp-stats.org/. 144

Figure "Today's Posterior": Image by Viktor Beekman, taken from https://www.bayesianspectacles.org under a CC-BY license (https://creativecommons.org/licenses/by/4.0/legalcode). 150

Figure 8.9: Figure generated by the JASP module 'Learn Bayes', available at https://jasp-stats.org/. 151

The Rule of Succession

Figure 9.1: Figure created using R. 158

Figure "Stamp Laplace": Figure taken from https://www.laposte.fr/toutsurletimbre/connaissance-du-timbre/dicotimbre/timbres/laplace-1031, permission to reproduce granted by ©La Poste and Rosine Gosset-Lemagny. 158

Figure 9.2: Figure generated by the JASP module 'Learn Bayes', available at https://jasp-stats.org/. 159

Figure 9.3: Figure generated by the JASP module 'Learn Bayes', available at https://jasp-stats.org/. 164

The Problem of Points

Figure "Blaise Pascal": Painting by François II Quesnel, taken from https://commons.wikimedia.org/wiki/File:Blaise_Pascal_Versailles.JPG under a CC-BY-3.0. license (https://creativecommons.org/licenses/by/3.0/). 173

Figure 10.1: Figure generated by the JASP module 'Learn Bayes', available at https://jasp-stats.org/. 175

Figure "Pierre de Fermat": Portrait by an unknown artist, taken from https://commons.wikimedia.org/wiki/File:Pierre_de_Fermat.jpg, public domain. 176

Figure 10.2: Figure generated by the JASP module 'Learn Bayes', available at https://jasp-stats.org/. 178

Interlude: Buffon's Needle

Figure "Georges-Louis Leclerc, Comte de Buffon": Image by François-Hubert Drouais, taken from https://commons.wikimedia.org/wiki/File:Georges-Louis_Leclerc_de_Buffon.jpg under a CC-PD license (https://creativecommons.org/publicdomain/mark/1.0/). The online information indicates that the portrait was painted in 1753, but Roger (1997, p. 222) –a highly reliable source– dates the painting to 1760 or 1761. 183

Figure 11.1: Two example illustrations from the sixth volume of Buffon's *Histoire naturelle générale et particulière avec la description du Cabinet du Roi* (1756, p. 138). Left panel: 'Le cerf'. This retouched version was obtained from https://en.wikipedia.org/wiki/Histoire_Naturelle; the original source is http://gallica.bnf.fr/ark:/12148/btv1b2300253d/f11.item. Red panel: the deer skeleton. Source: https://gallica.bnf.fr/ark:/12148/bpt6k10672421/f187.item. Both illustrations were designed by Jacques De Sève; the left panel was engraved by Claude Donat Jardinier; the right panel was engraved by Pierre-Etienne Moitte. 185

Figure "Le chat domestique": Illustration from the sixth volume of Buffon's *Histoire naturelle générale et particulière avec la description du Cabinet du Roi* (1756, p. 48). Design by Jacques De Sève, engraving by Pierre Charles Baquoy. Source: https://gallica.bnf.fr/ark:/12148/bpt6k10672421/f65.item. 186

Figure "Cover page of the Histoire Naturelle": Photo by Eric-Jan Wagenmakers, October 2022. 187

Figure 11.2: Figure generated by the JASP module 'Learn Bayes', available at `https://jasp-stats.org/`. 189

Figure 11.3: Figure generated by the JASP module 'Learn Bayes', available at `https://jasp-stats.org/`. 193

Figure 11.4: Figure generated by the JASP module 'Learn Bayes', available at `https://jasp-stats.org/`. 194

Figure "Stamp Buffon": Figure taken from `https://www.laposte.fr/toutsurletimbre/connaissance-du-timbre/dicotimbre/timbres/buffon-856`, permission to reproduce granted by ©La Poste. 196

Figure "Buffon assis dans son fauteuil": Photo by Eric-Jan Wagenmakers, October 2022. 197

Figure "Le chameau": Illustration from the eleventh volume of Buffon's *Histoire naturelle générale et particulière avec la description du Cabinet du Roi* (1764, p. 276). Design by Jacques De Sève, engraving by Pierre Charles Baquoy. Source: `https://books.openedition.org/mnhn/3079`. 199

Figure 11.5: Screenshot taken from the JASP module 'Learn Bayes', available at `https://jasp-stats.org/`. 202

The Pancake Puzzle

Figure "Data collection in action": Included by permission of Nataschja Wagenmakers. 205

Figure 12.1: Figure generated by the JASP module 'Learn Bayes', available at `https://jasp-stats.org/`. 206

Figure 12.2: Figure created using `R`. 207

Figure 12.3: Figure created using `R`. 208

Figure 12.4: Figure generated by the JASP module 'Learn Bayes', available at `https://jasp-stats.org/`. 210

Figure 12.5: Figure generated by the JASP module 'Learn Bayes', available at `https://jasp-stats.org/`. 213

Figure 12.6: Figure generated by the JASP module 'Learn Bayes', available at `https://jasp-stats.org/`. 214

Figure 12.7: Figure generated by the JASP module 'Learn Bayes', available at `https://jasp-stats.org/`. 215

Figure "Do Not Throw Away": Image by Viktor Beekman, taken from `https://www.bayesianspectacles.org/` under a CC-BY license (`https://creativecommons.org/licenses/by/4.0/legalcode`). 217

Figure 12.8: Image by Viktor Beekman, taken from `https://www.bayesianspectacles.org/` under a CC-BY license `https://creativecommons.org/licenses/by/4.0/legalcode`. 218

Figure 12.9: Figure generated by the JASP module 'Learn Bayes', available at `https://jasp-stats.org/`. 220

Figure 12.10: Image by EJ. Students were informed that their prior choices could be used for this book; they were free to use pseudonyms. 223

Figure 12.11: Screenshot taken from the JASP module 'Learn Bayes', available at https://jasp-stats.org/. 224

Figure 12.12: Figure generated by the JASP module 'Learn Bayes', available at https://jasp-stats.org/. 225

Figure 12.13: Figure generated by the JASP module 'Learn Bayes', available at https://jasp-stats.org/. 226

Figure 12.14: Figure generated by the JASP module 'Learn Bayes', available at https://jasp-stats.org/. 226

A Plethora of Pancakes

Figure 13.1: Figure created using R. 228

Figure 13.2: Figure created using R. 232

Figure 13.3: Figure created using R. 237

Figure 13.4: Screenshot taken from the JASP module 'Learn Bayes', available at https://jasp-stats.org/. 239

Figure 13.5: Figure generated by the JASP module 'Learn Bayes', available at https://jasp-stats.org/. 240

Figure 13.6: Figure generated by the JASP module 'Learn Bayes', available at https://jasp-stats.org/. 241

Figure 13.7: Figure generated by the JASP module 'Learn Bayes', available at https://jasp-stats.org/. 241

A Crack in the Laplacean Edifice

Figure 14.1: Figure generated by the JASP module 'Learn Bayes', available at https://jasp-stats.org/. 246

Figure "Broad": Image by Paul Arthur Schilpp, taken from https://en.wikipedia.org/wiki/C._D._Broad#/media/File:C._D._Broad_philosopher.png under a CC-PD license (https://creativecommons.org/publicdomain/mark/1.0/). 249

Wrinch and Jeffreys to the Rescue

Figure "Dorothy Wrinch": Image by Gallica Digital Library, taken from https://en.wikipedia.org/wiki/File:Dorothy_Maud_Wrinch_1921.jpg under a CC-PD license (https://creativecommons.org/publicdomain/mark/1.0/). 253

Figure "Jeffrey N. Rouder": Included by permission of Jeffrey Rouder. 538

Figure 15.1: Figure generated by the JASP module 'Learn Bayes', available at https://jasp-stats.org/. 255

Figure "Kate presents": Image by Viktor Beekman, taken from https://www.bayesianspectacles.org/ under a CC-BY license (https://creativecommons.org/licenses/by/4.0/legalcode). 256

Figure 15.2: Figure generated by the JASP module 'Learn Bayes', available at https://jasp-stats.org/. 257

Figure 15.3: Figure generated by the JASP module 'Learn Bayes', available at https://jasp-stats.org/. 258

Figure 15.4: Figure generated by the JASP module 'Learn Bayes', available at https://jasp-stats.org/. 259

Figure 15.5: Figure generated by the JASP module 'Learn Bayes', available at https://jasp-stats.org/. 260

Figure "The Likelihood Overwhelms": Image by Viktor Beekman, taken from https://www.bayesianspectacles.org/ under a CC-BY license (https://creativecommons.org/licenses/by/4.0/legalcode). 270

Figure 15.6: Photographer unknown. Included by permission of the Master and Fellows of St John's College, Cambridge. 271

Haldane's Rule of Succession

Figure "J. B. S. Haldane in the Black Watch": Image obtained from https://jbshaldane.org/ under a CC-PD license (https://creativecommons.org/publicdomain/mark/1.0/). 273

Figure "J. B. S. Haldane at Work": Image obtained from https://artuk.org/discover/artworks/professor-j-b-s-haldane-18921964-42374. Painting by Claude Rogers (1907-1979). Reproduced with permission of ©Crispin Rogers. 279

Jeffreys's Platitude

Figure 17.1: Figure generated by the JASP module 'Learn Bayes', available at https://jasp-stats.org/. 288

Figure 17.2: Figure generated by the JASP module 'Learn Bayes', available at https://jasp-stats.org/. 289

Figure 17.3: Figure generated by the JASP module 'Learn Bayes', available at https://jasp-stats.org/. 290

Figure 17.4: Figure generated by the JASP module 'Learn Bayes', available at https://jasp-stats.org/. 291

Figure 17.5: Figure generated by the JASP module 'Learn Bayes', available at https://jasp-stats.org/. 292

Figure 17.6: Figure generated by the JASP module 'Learn Bayes', available at https://jasp-stats.org/. 293

The Principle of Parsimony

Figure "Galileo Galilei": Image by Justus Sustermans, taken from https://commons.wikimedia.org/wiki/File:Justus_Sustermans_-_Portrait_of_Galileo_Galilei,_1636.jpg under a CC-PD license (https://creativecommons.org/publicdomain/mark/1.0/). 297

Figure 18.1: Figure created using R. 298

Figure 18.2: Figure created using R. 300

Figure 18.3: Figure created using R. 301

Figure 18.4: Figure created using R. 305

Figure "Fechner": Image by Smithsonian Libraries, taken from https://commons.wikimedia.org/wiki/File:Gustav_Fechner.jpg under a CC-PD license (https://creativecommons.org/publicdomain/mark/1.0/). 306

Figure "Weber-Fechner Dots": Image by MrPomidor, taken from https://en.wikipedia.org/wiki/Weber-Fechner_law under a CC0 license (https://creativecommons.org/publicdomain/zero/1.0/deed.en). 306

Figure 18.5: Figure created using R. 307

Figure 18.6: Figure created using R. 309

Figure "Jeffrey's Razor": Image by Viktor Beekman, taken from https://www.bayesianspectacles.org/ under a CC-BY license (https://creativecommons.org/licenses/by/4.0/legalcode). 393

Figure 18.7: Image by Moscarlop, taken from https://commons.wikimedia.org/wiki/File:William_of_Ockham.png under a CC-BY-SA license (https://creativecommons.org/licenses/by-sa/3.0/). 313

Figure "Onus of Proof": Image by Viktor Beekman, taken from https://www.bayesianspectacles.org/ under a CC-BY license https://creativecommons.org/licenses/by/4.0/legalcode 313

Figure "Royal Society": Image by Royal Society, taken from https://en.wikipedia.org/wiki/File:The_Royal_Society_Coat_of_Arms.svg under fair use agreement: educational purpose. 315

Figure 18.8: Image by Anefo, taken from https://commons.wikimedia.org/wiki/File:Bertrand_Russell_cropped.jpg under a CC-PD license (https://creativecommons.org/publicdomain/mark/1.0/). 317

Pascal's Arithmetical Triangle

Figure 34.1: Figure created using TikZ. 518

Figure 34.2: Figure created using TikZ. 518

Figure "Sir Francis Galton": Photograph taken by Eveleen Myers (née Tennant). Image taken from `https://www.npg.org.uk/collections/search/portrait/mw127193` under a CC-PD license (`https://creativecommons.org/publicdomain/mark/1.0/`). 518

Figure 34.3: Original illustration of Galton's 'quincunx' (Galton 1889, p. 63), image extracted from `https://galton.org/books/natural-inheritance/pdf/galton-nat-inh-1up-clean.pdf` under a CC-PD license (`https://creativecommons.org/publicdomain/mark/1.0/`). 519

Figure 34.4: Image by Viktor Beekman, taken from `https://www.bayesianspectacles.org` under a CC-BY license (`https://creativecommons.org/licenses/by/4.0/legalcode`). 520

Figure "Pearson and Galton": Image taken from `https://en.wikipedia.org/wiki/Francis_Galton#/media/File:Karl_Pearson;_Sir_Francis_Galton.jpg` under a CC-PD license (`https://creativecommons.org/publicdomain/mark/1.0/`). 525

Figure "Ronald Aylmer Fisher": Image taken from `https://en.wikipedia.org/wiki/Ronald_Fisher#/media/File:Youngronaldfisher2.JPG` under a CC-PD license (`https://creativecommons.org/publicdomain/mark/1.0/`). 527

Statistical Analysis of the Binomial Distribution

Figure "Howard Raiffa": Harvard Business School Archives Photograph Collection. Baker Library. Harvard Business School (olvwork376291). Reprinted with permission. 529

Figure "Andrew Gelman": Image by Schutz, taken from `https://en.wikipedia.org/wiki/Andrew_Gelman#/media/File:Andrew_Gelman_2012.jpg` under a CC-BY-SA license (`https://creativecommons.org/licenses/by-sa/3.0/`). 531

Recommended Reading on Bayesian Inference

Figure "Tony O'Hagan": Reprinted with permission from Dr. O'Hagan. 536

Figure "Jim Berger": Reprinted with permission from Dr. Berger. ??

Figure "Richard Morey": Reprinted with permission from Dr. Morey. 537

Figure "Cover *Bayesian Cognitive Modeling*" 538

Figure "Jeff Rouder": Reprinted with permission from Dr. Rouder. 538

Bibliography

R. Ackerman. A neglected proposal concerning simplicity. *Philosophy of Science*, 30:228–235, 1963.

M. M. Adams. *William Ockham*. University of Notre Dame Press, Notre Dame, IN, 1987.

D. Aerts and M. S. de Bianchi. Solving the hard problem of Bertrand's paradox. *Journal of Mathematical Physics*, 55:083503, 2014.

T. W. Africa. Archimedes through the looking–glass. *The Classical World*, 68:305–308, 1975.

J. Aitchison and I. R. Dunsmore. *Statistical prediction analysis*. Cambridge University Press, Cambridge, 1975.

M. Aitkin. Posterior Bayes factors. *Journal of the Royal Statistical Society. Series B (Methodological)*, 53:111–142, 1991.

J. Albert. *Bayesian Computation with R*. Springer, Dordrecht, 2 edition, 2009.

J. Aldrich. The statistical education of Harold Jeffreys. *International Statistical Review*, 73:289–307, 2005.

S. M. Anglin. Do beliefs yield to evidence? Examining belief perseverance vs. change in response to congruent empirical findings. *Journal of Experimental Social Psychology*, 82:176–199, 2019.

R. Ariew. Did Ockham use his razor? *Franciscan Studies*, 37:5–17, 1977.

Aristotle. *Physics (Books I and II)*. (W. Charlton, Trans.). Oxford University Press Inc., New York, 350BC/1970.

B. J. Arnow. On Laplace's extension of the Buffon needle problem. *The College Mathematics Journal*, 25:40–43, 1994.

I. Asimov. *I, Robot*. Gnome Press, New York, 1950.

B. Autzen. Bayesian Ockham's razor and nested models. *Economics & Philosophy*, 35:321–338, 2019.

L. Badger. Lazzarini's lucky approximation of π. *Mathematics Magazine*, 67:83–91, 1994.

L. Barrett and M. Connell. Jevons and the logic 'piano'. *The Rutherford Journal*, 1, 2005. URL `http://rutherfordjournal.org/article010103.html`.

P. L. Bartlett, P. M. Long, G. Lugosi, and A. Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences of the United States of America*, 117:30063–30070, 2020.

P. L. Bartlett, A. Montanari, and A. Rakhlin. Deep learning: A statistical viewpoint. *Acta Numerica*, 30:87–201, 2021.

F. Bartoš, A. Sarafoglou, H. R. Godmann, A. Sahrani, D. Klein Leunk, P. Y. Gui, D. Voss, K. Ullah, M. J. Zoubek, F. Nippold, F. Aust, F. F. Vieira, C.-G. Islam, A. J. Zoubek, S. Shabani, J. Petter, I. B. Roos, A. Finnemann, A. B. Lob, M. F. Hoffstadt, J. Nak, J. de Ron, K. Derks, K. Huth, S. Terpstra, T. Bastelica, M. Matetovici, V. L. Ott, A. S. Zetea, K. Karnbach, M. C. Donzallaz, A. John, R. M. Moore, F. Assion, R. van Bork, T. E. Leidinger, X. Zhao, A. Karami Motaghi, T. Pan, H. Armstrong, T. Peng, M. Bialas, J. Y.-C. Pang, B. Fu, S. Yang, X. Lin, D. Sleiffer, M. Bognar, B. Aczel, and E.-J. Wagenmakers. Fair coins tend to land on the same side they started: Evidence from $350{,}757$ flips. *Manuscript submitted for publication*, 2023. URL `https://doi.org/10.48550/arXiv.2310.04153`.

M. J. Bayarri and J. O. Berger. Hypothesis testing and model uncertainty. In P. Damien, P. Dellaportas, N. G. Polson, and D. A. Stephens, editors, *Bayesian Theory and Applications*, pages 361–400. Oxford University Press, Oxford, 2013.

T. Bayes. An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, 53:370–418, 1763.

A. Becker. *What is Real? The Unfinished Quest for the Meaning of Quantum Mechanics*. John Murray, London, 2018.

M. Belkin. Fit without fear: Remarkable mathematical phenomena of deep learning through the prism of interpolation. *Acta Numerica*, 30: 203–248, 2021.

M. Belkin, D. Hsu, S. Ma, and S. Mandal. Reconciling modern machine–learning practice and the classical bias–variance trade–off. *Proceedings of the National Academy of Sciences of the United States of America*, 116: 15849–15854, 2019.

J. H. Bennett, editor. *Statistical Inference and Analysis: Selected Correspondence of R. A. Fisher*. Clarendon Press, Oxford, 1990.

P. Berent. Disconfirmation by positive instances. *Philosophy of Science*, 39: 522, 1972.

J. O. Berger. Bayes factors. In S. Kotz, N. Balakrishnan, C. Read, B. Vidakovic, and N. L. Johnson, editors, *Encyclopedia of Statistical Sciences, vol. 1 (2nd ed.)*, pages 378–386. Wiley, Hoboken, NJ, 2006.

J. O. Berger and D. A. Berry. Statistical analysis and the illusion of objectivity. *American Scientist*, 76:159–165, 1988.

J. O. Berger and W. H. Jefferys. The application of robust Bayesian analysis to hypothesis testing and Occam's razor. *Journal of the Italian Statistical Society*, 1:17–32, 1992.

J. O. Berger and L. R. Pericchi. The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association*, 91: 109–122, 1996.

J. O. Berger and L. R. Pericchi. Objective Bayesian methods for model selection: Introduction and comparison (with discussion). In P. Lahiri, editor, *Model Selection*, pages 135–207. Institute of Mathematical Statistics Lecture Notes—Monograph Series, volume 38, Beachwood, OH, 2001.

J. O. Berger and R. L. Wolpert. *The Likelihood Principle (2nd ed.)*. Institute of Mathematical Statistics, Hayward (CA), 1988.

J. O. Berger, S. E. Fienberg, A. E. Raftery, and C. P. Robert. Incoherent phylogeographic inference. *Proceedings of the National Academy of Sciences*, 107:E157, 2010.

J. M. Bernardo and A. F. M. Smith. *Bayesian Theory*. Wiley, New York, 1994.

J. Bertrand. *Calcul des Probabilités*. Gauthier-Villars et Fils, Paris, 1889.

N. Best, R. G. Price, I. J. Pouliquen, and O. N. Keene. Assessing efficacy in important subgroups in confirmatory trials: An example using Bayesian dynamic borrowing. *Pharmaceutical Statistics*, 20:551–562, 2021.

M. Bilalić, K. Smallbone, P. McLeod, and F. Gobet. Why are (the best) women so good at chess? Participation rates and gender differences in intellectual domains. *Proceedings of the Royal Society B*, 276:1161–1165, 2009.

T. Blanchard, T. Lombrozo, and S. Nichols. Bayesian Occam's razor is a razor of the people. *Cognitive Science*, 42:1345–1359, 2018.

J. K. Blitzstein and J. Hwang. *Introduction to Probability (2nd ed.)*. Taylor & Francis Group, 2019.

W. M. Bolstad. *Introduction to Bayesian Statistics (2nd ed.)*. Wiley, Hoboken, NJ, 2 edition, 2007.

B. A. Bolt. Inge Lehmann (1888–1993). *Biographical Memoirs of Fellows of the Royal Society*, 43:286–301, 1997.

E. Borel. Apropos of a treatise on probability (the original 1924 French version was published in *Revue Philosophique*). In H. E. Kyburg Jr. and H. E. Smokler, editors, *Studies in Subjective Probability*, pages 47–60. John Wiley & Sons, New York, 1964.

E. Borel. *Elements of the Theory of Probability*. Prentice-Hall, Englewood Cliffs, NJ, 1965.

S. Bose. Bayesian robustness with mixture classes of priors. *Annals of Statistics*, 22:652–667, 1994.

G. E. P. Box. Sampling and Bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society, Series A*, 143:383–430, 1980.

S. Bozza, F. Taroni, and A. Biedermann. *Bayes Factors for Forensic Decision Analyses with R*. Springer, New York, 2022.

D. Brewster. *The Life of Sir Isaac Newton*. J. J. Harper, New York, 1835.

C. D. Broad. On the relation between induction and probability (part I.). *Mind*, 27:389–404, 1918.

B. Bryson. *A Short History of Nearly Everything*. Black Swan, London, 2004.

G.-L. Buffon. Invention de miroirs ardens, pour brusler à une grande distance. *Mémoires de l'Académie Royale des Sciences*, pages 82–101, 1747.

G.-L. Buffon. *Histoire Naturelle Générale et Particulière (Vols. 1–36)*. Imprimerie Royale, Paris, 1749–1788.

G.-L. Buffon. *Histoire Naturelle Générale et Particulière (Vols. 1–36)*. Imprimerie Royale, Paris, 1777a.

G.-L. Buffon. Essai d'arithmétique morale. *Supplément à l'Histoire Naturelle*, 4:article XXIII, 1777b.

G.-L. Buffon. *Buffon's Natural History (Vols. 1–10)*. T. Gillet, London, 1797–1807.

V. Buranelli. *The Wizard from Vienna*. Coward, McCann and Geoghegan, New York, 1975.

P. C. Bürkner. brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, 80:1–28, 2017.

J. B. Bury. *A History of Freedom of Thought*. The University Press, Cambridge, 1913.

L. Campbell and W. Garnett. *The Life of James Clerk Maxwell With a Selection From His Correspondence and Occasional Writings and a Sketch of His Contributions to Science*. MacMillan and Co., London, 1882.

R. Carnap. *Logical Foundations of Probability*. The University of Chicago Press, Chicago, 1950.

B. Carpenter, A. Gelman, M. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. A. Brubaker, J. Guo, P. Li, and A. Riddell. Stan: A probabilistic programming language. *Journal of Statistical Software*, 76, 2017.

C. D. Chambers. *The Seven Deadly Sins of Psychology: A Manifesto for Reforming the Culture of Scientific Practice*. Princeton University Press, Princeton, 2017.

C. P. Chamley. *Rational Herds: Economic Models of Social Learning*. Cambridge, Cambridge University Press, 2004.

S. Chandramouli and R. Shiffrin. Extending Bayesian induction. *Journal of Mathematical Psychology*, 72:38–42, 2016.

S. H. Chandramouli and R. M. Shiffrin. Commentary on Gronau and Wagenmakers. *Computational Brain & Behavior*, 2:12–21, 2019.

R. A. Chechile. *Bayesian Statistics for Experimental Scientists: A General Introduction Using Distribution–Free Methods*. The MIT Press, Cambridge, MA, 2020.

M. T. Cicero. *de Divinatione*. (W. A. Falconer, Trans.). Harvard University Press, London, 44BC/1923.

M. T. Cicero. *Academica*. (H. Rackham, Trans.). William Heinemann LTD, London, 45BC/1956a.

M. T. Cicero. *de Natura Deorum*. (H. Rackham, Trans.). William Heinemann LTD, London, 45BC/1956b.

R. Clark. *J. B. S.: The Life and Work of J. B. S. Haldane*. Bloomsbury Reader, London, 1968/2013.

A. Clayton. *Bernoulli's Fallacy: Statistical Illogic and the Crisis of Modern Science*. Columbia University Press, New York, 2021.

M. A. Clyde. *BAS: Bayesian Adaptive Sampling for Bayesian Model Averaging*, 2016. R package version 1.4.1.

M. A. Clyde, J. Ghosh, and M. L. Littman. Bayesian adaptive sampling for variable selection and model averaging. *Journal of Computational and Graphical Statistics*, 20:80–101, 2011.

G. Consonni and P. Veronese. Compatibility of prior specifications across linear models. *Statistical Science*, 23:332–353, 2008.

G. Consonni, D. Fouskakis, B. Liseo, and I. Ntzoufras. Prior distributions for objective Bayesian analysis. *Bayesian Analysis*, 13:627–679, 2018.

J. Cook and S. Lewandowsky. Rational irrationality: Modeling climate change belief polarization using Bayesian networks. *Topics in Cognitive Science*, 8:160–179, 2016.

S. R. Cook, A. Gelman, and D. B. Rubin. Validation of software for Bayesian models using posterior quantiles. *Journal of Computational and Graphical Statistics*, 15:675–692, 2006.

J. L. Coolidge. *An Introduction to Mathematical Probability*. Oxford University Press, London, 1925.

R. T. Cox. Probability, frequency and reasonable expectation. *The American Journal of Physics*, 14:1–13, 1946.

M. W. Crofton. Probability. In T. S. Baynes and W. R. Smith, editors, *Encyclopædia Britannica, Vol. XIX*, pages 768–789. Adam and Charles Black, Edinburgh, 9 edition, 1885.

G. D'Agostini. Teaching statistics in the physics curriculum: Unifying and clarifying role of subjective probability. *American Journal of Physics*, 67:1260–1268, 1999.

G. D'Agostini. The Gauss' Bayes factor. *Manuscript available online*, 2020. URL `https://arxiv.org/abs/2003.10878`.

S. R. Dalal and W. J. Hall. Approximating priors by mixtures of natural conjugate priors. *Journal of the Royal Statistical Society, Series B (Methodological)*, 45:278–286, 1983.

Y. Dar, V. Muthukumar, and R. G. Baraniuk. A farewell to the bias–variance tradeoff? An overview of the theory of overparameterized machine learning. *ArXiv*, 2021. URL `https://arxiv.org/abs/2109.02355`.

R. Darnton. *Mesmerism and the End of the Enlightenment in France*. Harvard University Press, Cambridge, MA, 1968.

A. P. Dawid. Present position and potential developments: Some personal views: Statistical theory: The prequential approach (with discussion). *Journal of the Royal Statistical Society Series A*, 147:278–292, 1984.

A. P. Dawid. Fisherian inference in likelihood and prequential frames of reference. *Journal of the Royal Statistical Society B*, 53:79–109, 1991.

A. P. Dawid. Statistics on trial. *Significance*, 2:6–8, 2005.

A. P. Dawid. Posterior model probabilities. In P. S. Bandyopadhyay and M. Forster, editors, *Handbook of The Philosophy of Science: Philosophy of Statistics*, pages 607–630. Elsevier, Amsterdam, 2011.

P. De Blasi and S. G. Walker. Bayesian asymptotics with misspecified models. *Statistica Sinica*, 23:169–187, 2013.

B. de Finetti. Foresight: Its logical laws, its subjective sources. In H. E. Kyburg and H. E. Smokler, editors, *Studies in Subjective Probability*, pages 57–118. Krieger, Huntington, N.Y., 1937/1980.

B. de Finetti. Foresight: Its logical laws, its subjective sources (translated from the original 1937 French monograph). In H. E. Kyburg and H. E. Smokler, editors, *Studies in Subjective Probability*, pages 97–158. John Wiley & Sons, New York, 1964.

B. de Finetti. *Theory of Probability, Vol. 1 and 2*. John Wiley & Sons, New York, 1974.

A. D. De Groot. *Methodology: Foundations of Inference and Research in the Behavioral Sciences*. Mouton, The Hague, 1969.

A. De Moivre. *The Doctrine of Chances*. A. Millar, London, 3 edition, 1718/1756.

A. De Morgan. *An Essay on Probabilities and on Their Application to Life Contingencies and Insurance Offices*. Longman, London, 1838.

A. De Morgan. *Formal Logic: The Calculus of Inference, Necessary and Probable*. University Press of the Pacific, Honolulu, 1847/2003.

A. De Morgan. Theory of probabilities. In E. Smedley, H. J. Rose, and H. J. Rose, editors, *Encyclopædia Metropolitana*, pages 393–490. John Joseph Griffin and Company, London, 1849.

A. De Morgan. *A Budget of Paradoxes: Volume I*. The Open Court Publishing Co., Chicago, 2 edition, 1915.

S. E. De Morgan. *Memoir of Augustus De Morgan*. Longmans, Green, and Co., London, 1882.

M. H. DeGroot. A conversation with Persi Diaconis. *Statistical Science*, 1: 319–334, 1986.

J. Del Vecchio Junior. Chance and probability in Poincaré's epistemology. *Philosophia Scientiæ*, 20:177–196, 2016.

K. Derks, J. de Swart, E.-J. Wagenmakers, J. Wille, and R. Wetzels. JASP for audit: Bayesian tools for the auditing practice. *Journal of Open Source Software*, 6:2733, 2021.

D. Devitt. *The Skipper's War: Dragon School, Oxford & The Great War*. Scala Arts Publishers Inc., London, 2022.

K. Devlin. *The Unfinished Game: Pascal, Fermat, and the Seventeenth–Century Letter that Made the World Modern*. Basic Books, New York, 2008.

P. Diaconis and B. Skyrms. *Ten Great Ideas About Chance*. Princeton University Press, Princeton, 2018.

P. Diaconis, S. Holmes, and R. Montgomery. Dynamical bias in the coin toss. *SIAM Review*, 49:211–235, 2007.

J. M. Dickey. The weighted likelihood ratio, linear hypotheses on normal location parameters. *The Annals of Mathematical Statistics*, 42: 204–223, 1971.

J. M. Dickey and B. P. Lientz. The weighted likelihood ratio, sharp hypotheses about chances, the order of a Markov chain. *The Annals of Mathematical Statistics*, 41:214–226, 1970.

Z. Dienes. *Understanding Psychology as a Science: An Introduction to Scientific and Statistical Inference*. Palgrave MacMillan, New York, 2008.

T. M. Donovan and R. M. Mickey. *Bayesian Statistics for Beginners: A Step–by–Step Approach*. Oxford University Press, Oxford, 2019.

C.-O. Doron. Race and genealogy. Buffon and the formation of the concept of "race". *Humana. Mente Journal of Philosophical Studies*, 22: 75–109, 2012.

A. Douventzidis and E. Landquist. Logarithms are hot stuff: A new rating scale for chili peppers. *PRIMUS*, 32:650–660, 2022.

F. Dudbridge. A scale of interpretation for likelihood ratios and Bayes factors. *ArXiv*, 2022. URL `https://arxiv.org/abs/2212.06669`.

L. A. Dugatkin. Buffon, Jefferson and the theory of New World degeneracy. *Evolution: Education and Outreach*, 12:15, 2019.

A. Duke. *Thinking in Bets: Making Smarter Decisions When You Don't Have All the Facts*. Portfolio/Penguin, New York, 2018.

F. Dyson. A meeting with Enrico Fermi. *Nature*, 427:297, 2004.

A. Eagle (Ed.). *Philosophy of Probability: Contemporary Readings*. Routledge, New York, 2011.

J. Earman. *A Primer on Determinism*. Reidel, Dordrecht, 1986.

A. Eddington. *The Philosophy of Physical Science*. Cambridge University Press, Cambridge, 1939.

M. Eder, J. Rybicki, and M. Kestemont. Stylometry with R: A package for computational text analysis. *R Journal*, 8:107–121, 2016.

F. Y. Edgeworth. Probability. In H. Chisholm, editor, *Encyclopædia Britannica*, pages 376–403. Cambridge University Press, London, 11 edition, 1911.

A. W. F. Edwards. Pascal and the problem of points. *International Statistical Review*, 50:259–266, 1982.

A. W. F. Edwards. *Pascal's Arithmetical Triangle: The Story of a Mathematical Idea*. Dover Publications, Mineola, NY, 1987/2019.

A. W. F. Edwards. *Likelihood*. The Johns Hopkins University Press, Baltimore, MD, 1992.

W. Edwards, H. Lindman, and L. J. Savage. Bayesian statistical inference for psychological research. *Psychological Review*, 70:193–242, 1963.

J. Ellenberg. *How Not To Be Wrong: The Hidden Maths of Everyday Life*. Allen Lane, London, 2014.

E. Ernst. *Chiropractic: Not All That It's Cracked Up to Be*. Springer, New York, 2020.

A. Etz. Introduction to the concept of likelihood and its applications. *Advances in Methods and Practices in Psychological Science*, 1:60–69, 2018.

A. Etz and E.-J. Wagenmakers. J. B. S. Haldane's contribution to the Bayes factor hypothesis test. *Statistical Science*, 32:313–329, 2017.

A. Etz, Q. F. Gronau, F. Dablander, P. A. Edelsbrunner, and B. Baribault. How to become a Bayesian in eight easy steps: An annotated reading list. *Psychonomic Bulletin & Review*, 25:219–234, 2018a.

A. Etz, J. M. Haaf, J. N. Rouder, and J. Vandekerckhove. Bayesian inference and testing any hypothesis you can specify. *Advances in Methods and Practices in Psychological Science*, 1:281–295, 2018b.

B. Eva. Principles of indifference. *The Journal of Philosophy*, 116:390–411, 2019.

M. Evans. *Measuring Statistical Evidence Using Relative Belief*. CRC Press, Boca Raton, FL, 2015.

I. W. Evett. Bayesian inference and forensic science: Problems and perspectives. *The Statistician*, 36:99–105, 1987.

I. W. Evett. Implementing bayesian methods in forensic science, 1991.

I. W. Evett, G. Jackson, J. A. Lambert, and S. McCrossan. The impact of the principles of evidence interpretation on the structure and content of statements. *Science & Justice*, 40:233–239, 2000.

P. Eymard and J.-P. Lafon. *The Number* π. American Mathematical Society, Providence, Rhode Island, 2004.

R. E. Fancher. *Pioneers of Psychology*. W. W. Norton & Company, New York, 3 edition, 1996.

J. Feldman. Bayesian inference and "truth": A comment on Hoffman, Singh, and Prakash. *Psychonomic Bulletin & Review*, 22:1523–1525, 2015.

O. E. Fellows and S. F. Milliken. *Buffon*. Twayn, New York, 1972.

P. K. Feyerabend. *Against Method: Outline of an Anarchistic Theory of Knowledge*. Verso, London, 3 edition, 1993.

R. Feynman. *The Character of Physical Law*. Penguin Books, London, 1965/1992.

R. Feynman. *The Pleasure of Finding Things Out*. Perseus Books, Cambridge, MA, 1999.

L. N. G. Filon, G. U. Yule, H. Westergaard, M. Greenwood, and K. Pearson. *Speeches Delivered at a Dinner Held in University College, London in Honour of Professor Karl Pearson 23 April 1934*. Privately Printed at Cambridge University Press, Cambridge, 1934. URL `https://archive.org/details/filon-et-al-1934-speeches-delivered-at-a-dinner`.

R. A. Fisher. The nature of probability. *The Centennial Review of Arts & Science*, 2:261–274, 1958.

E. Fong and C. C. Holmes. On the marginal likelihood and cross-validation. *Biometrika*, 107:489–496, 2020.

M. Forster and E. Sober. How to tell when simpler, more unified, or less *Ad Hoc* theories will provide more accurate predictions. *British Journal for the Philosophy of Science*, 45:1–35, 1994.

M. R. Forster. Predictive accuracy as an achievable goal of science. *Philosophy of Science*, 69:S124–S134, 2002.

B. Franklin, M.-J. Majault, J.-B. Le Roy, C. L. Sallin, J. S. Bailly, J. d'Arcet, G. de Bory, J.-I. Guillotin, and A. Lavoisier. *Rapport Des Commissaires Chargés par le Roi de L'examen Magnétisme Animal*. L'Imprimerie Royale, Paris, 1784.

B. Franklin, M.-J. Majault, J.-B. Le Roy, C. L. Sallin, J. S. Bailly, J. d'Arcet, G. de Bory, J.-I. Guillotin, and A. Lavoisier. Report of the commissioners charged by the king with the examination of animal magnetism (translated from the 1784 French original). *International Journal of Clinical and Experimental Hypnosis*, 50:332–363, 2002.

J. Franklin. *The Science of Conjecture: Evidence and Probability Before Pascal (2nd ed.)*. Johns Hopkins University Press, Baltimore, 2015.

M. C. Galavotti. *A Philosophical Introduction to Probability*. CSLI Publications, Stanford, 2005.

M. C. Galavotti. Pragmatism and the birth of subjective probability. *European Journal of Pragmatism and American Philosophy*, 11:1509, 2019.

Galileo. *Dialogues Concerning Two New Sciences.* (H. Crew, & A. de Salvio, Trans.). The Macmillian Company, New York, 1638/1914.

F. Galton. *Natural Inheritance*. Macmillan, London, 1889.

D. Gamerman and H. F. Lopes. *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. Chapman & Hall/CRC, Boca Raton, FL, 2006.

A. Gandy and J. A. Scott. Unit testing for MCMC and other Monte Carlo methods. *ArXiv Preprint*, 2021. URL `https://arxiv.org/abs/2001.06465`.

M. Gardner. Mathematical games: Problems involving questions of probability and ambiguity. *Scientific American*, 201:174–182, 1959a.

M. Gardner. Mathematical games: How three modern mathematicians disproved a celebrated conjecture of Leonhard Euler. *Scientific American*, 201:181–188, 1959b.

M. Gardner. *The Scientific American Book of Mathematical Puzzles & Diversions*. Simon and Schuster, New York, 1961.

M. Gardner. On the fabric of inductive logic, and some probability paradoxes. *Scientific American*, 234:119–122, 1976.

M. Gardner. *Time Travel And Other Mathematical Bewilderments*. W. H. Freeman and Company, New York, 1988.

A. Gelman. Bayes, Jeffreys, prior distributions and the philosophy of statistics. *Statistical Science*, 24:176–178, 2009.

A. Gelman. Induction and deduction in Bayesian data analysis. *Rationality, Markets and Morals*, 2:67–78, 2011.

A. Gelman and J. Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, Cambridge, 2007.

A. Gelman and H. Stern. The difference between "significant" and "not significant" is not itself statistically significant. *The American Statistician*, 60:328–331, 2006.

A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian Data Analysis (3rd ed.)*. Chapman & Hall/CRC, Boca Raton (FL), 2014a.

A. Gelman, J. Hwang, and A. Vehtari. Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, 24: 997–1016, 2014b.

S. J. Gershman. How to never be wrong. *Psychonomic Bulletin & Review*, 26:13–28, 2019.

J. Geweke. Getting it right: Joint distribution tests of posterior simulators. *Journal of the American Statistical Association*, 99:799–804, 2004.

G. Gigerenzer, Z. Swijtink, T. Porter, L. Daston, J. Beatty, and L. Krüger. *The Empire of Chance*. Cambridge University Press, Cambridge, 1989.

G. Gigerenzer, J. Multmeier, A. Föhring, and O. Wegwarth. Do children have Bayesian intuitions? *Journal of Experimental Psychology: General*, 150: 1041–1070, 2021.

J. Glucker. Probabile, veri simile, and related terms. In J. G. F. Powell, editor, *Cicero the Philosopher*, pages 115–143. Clarendon Press, Oxford, 1995.

T. Gneiting and E. A. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102:359–378, 2007.

J. I. Gold and M. N. Shadlen. Banburismus and the brain: Decoding the relationship between sensory stimuli, decisions, and reward. *Neuron*, 36: 299–308, 2002.

B. Goldacre. *Bad Pharma: How Drug Companies Mislead Doctors and Harm Patients*. Fourth Estate, London, 2012.

I. J. Good. *Probability and the Weighing of Evidence*. Charles Griffin, London, 1950.

I. J. Good. The paradox of confirmation. *The British Journal for the Philosophy of Science*, 11:145–149, 1960.

I. J. Good. The paradox of confirmation (II). *The British Journal for the Philosophy of Science*, 12:63–64, 1961.

I. J. Good. The white shoe is a red herring. *The British Journal for the Philosophy of Science*, 17:322, 1967.

I. J. Good. Explicativity, corroboration, and the relative odds of hypotheses. *Synthese*, 30:39–73, 1975.

I. J. Good. Some logic and history of hypothesis testing. In J. C. Pitt, editor, *Philosophical Foundations of Economics*, pages 149–174. D. Reidel Publishing Company, Dordrecht–Holland, 1981.

I. J. Good. *Good Thinking: The Foundations of Probability and Its Applications*. University of Minnesota Press, Minneapolis, 1983.

I. J. Good. C205. Monotonic properties of the moments of a Bayes factor and the relationship to measures of divergence. *Journal of Statistical Computation and Simulation*, 19:320–325, 1984.

I. J. Good. Weight of evidence: A brief survey. In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, editors, *Bayesian Statistics 2*, pages 249–269. Elsevier, New York, 1985.

I. J. Good. C252. A minor comment concerning Hempel's paradox of confirmation. *Journal of Statistical Computation and Simulation*, 24: 320–321, 1986.

I. J. Good. C318. the paradox of confirmation and stoogian observations. *Journal of Statistical Computation and Simulation*, 31:121, 1989.

I. J. Good. C421. Turing's little theorem is not really paradoxical. *Journal of Statistical Computation and Simulation*, 49:242–244, 1994.

S. N. Goodman. Toward evidence-based medical statistics. 2: The Bayes factor. *Annals of Internal Medicine*, 130:1005–1013, 1999.

P. Gorroochurn. Errors of probability in historical context. *The American Statistician*, 65:246–254, 2011.

P. Gorroochurn and B. Levin. On two historical aspects of Buffon's needle problem. *Electronic Journal for History of Probability and Statistics*, 9, 2013.

M. A. Goss–Sampson. *Bayesian Inference in JASP: A Guide for Students*. 2020. URL `https://jasp-stats.org/jasp-materials/`.

N. T. Gridgeman. Geometric probability and the number π. *Scripta Mathematica*, 25:183–195, 1960.

Q. F. Gronau and E.-J. Wagenmakers. Bayesian evidence accumulation in experimental mathematics: A case study of four irrational numbers. *Experimental Mathematics*, 27:277–286, 2018.

Q. F. Gronau and E.-J. Wagenmakers. Rejoinder: More limitations of Bayesian leave-one-out cross-validation. *Computational Brain & Behavior*, 2:35–47, 2019.

Q. F. Gronau, A. Sarafoglou, D. Matzke, A. Ly, U. Boehm, M. Marsman, D. S. Leslie, J. J. Forster, E.-J. Wagenmakers, and H. Steingroever. A tutorial on bridge sampling. *Journal of Mathematical Psychology*, 81:80–97, 2017.

Q. F. Gronau, K. N. A. Raj, and E.-J. Wagenmakers. Informed Bayesian inference for the A/B test. *Journal of Statistical Software*, 100:1–39, 2021.

P. Grünwald and T. van Ommen. Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it. *Bayesian Analysis*, 12:1069–1103, 2017.

X. Gu, H. Hoijtink, J. Mulder, and Y. Rosseel. Bain: A program for Bayesian testing of order constrained hypotheses in structural equation models. *Journal of Statistical Computation and Simulation*, 89:1526–1553, 2019.

J. M. Haaf, A. Ly, and E.-J. Wagenmakers. Retire significance, but still test hypotheses. *Nature*, 567:461, 2019.

I. Hacking. *The Taming of Chance*. Cambridge University Press, Cambridge, 1990.

A. Hald. *A History of Probability and Statistics and Their Applications Before 1750*. John Wiley & Sons, New York, 1990.

J. B. S. Haldane. On being the right size. *Harper's Magazine*, 152:424–427, 1926.

J. B. S. Haldane. A note on inverse probability. *Mathematical Proceedings of the Cambridge Philosophical Society*, 28:55–61, 1932.

A. Hall. On an experimental determination of π. *Messenger of Mathematics*, 2:113–114, 1872.

T. Hastie, A. Montanari, S. Rosset, and R. J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *Annals of Statistics*, 50:949–986, 2022.

D. W. Heck. A caveat on the Savage–Dickey density ratio: The case of computing Bayes factors for regression parameters. *British Journal of Mathematical and Statistical Psychology*, 72:316–333, 2019.

L. Held and M. Ott. How the maximal evidence of p–values against point null hypotheses depends on sample size. *The American Statistician*, 70:335–341, 2016.

L. Henderson, N. D. Goodman, J. B. Tenenbaum, and J. F. Woodward. The structure and dynamics of scientific theories: A hierarchical Bayesian perspective. *Philosophy of Science*, 77:172–200, 2010.

R. Hertwig and G. Gigerenzer. The 'conjunction fallacy' revisited: How intelligent inferences look like reasoning errors. *Journal of Behavioral Decision Making*, 12:275–305, 1999.

J. D. Hey, T. M. Neugebauer, and C. M. Pasca. George-Louis Leclerc de Buffon's 'Essays on moral arithmetic'. In A. Ockenfels and A. Sadrieh, editors, *A Collection of Essays in Honor of Reinhard Selten*, pages 245–282. Springer, Berlin, 2010.

R. Hill. Reflections on the cot death cases. *Significance*, 2:13–15, 2005.

M. Hinne, Q. F. Gronau, D. van den Bergh, and E.-J. Wagenmakers. A conceptual introduction to Bayesian model averaging. *Advances in Methods and Practices in Psychological Science*, 3:200–215, 2020.

H. Hoijtink, J. Mulder, C. van Lissa, and X. Gu. A tutorial on testing hypotheses using the Bayes factor. *Psychological Methods*, 24:539–556, 2019.

J. Horgan. *The End Of Science: Facing The Limits Of Knowledge In The Twilight Of The Scientific Age*. Basic Books, New York, 2015.

S. Hossenfelder, editor. *Existential Physics: A Scientist's Guide to Life's Biggest Questions*. Viking, 2022.

D. Howie. *Interpreting Probability: Controversies and Developments in the Early Twentieth Century*. Cambridge University Press, Cambridge, 2002.

C. Howson. Must the logical probability of laws be zero? *The British Journal for the Philosophy of Science*, 24:153–163, 1973.

C. Howson. Popper, prior probabilities, and inductive inference. *The British Journal for the Philosophy of Science*, 38:207–224, 1987.

C. Howson. On the consistency of Jeffreys' simplicity postulate and its role in Bayesian inference. *Philosophical Quarterly*, 38:68–83, 1988.

C. Howson and P. Urbach. *Scientific Reasoning: The Bayesian Approach (3rd. ed.)*. Open Court, Chicago, 2006.

T. E. Hudson. *Bayesian Data Analysis for the Behavioral and Neural Sciences*. Cambridge University Press, Cambridge, 2021.

D. Hume. *A Treatise of Human Nature*. 1739.

V. S. Huzurbazar. On the certainty of an inductive inference. *Mathematical Proceedings of the Cambridge Philosophical Society*, 51:761–762, 1955.

G. R. Iversen, W. H. Longcor, F. Mosteller, J. P. Gilbert, and C. Youtz. Bias and runs in dice throwing and recording: A few million throws. *Psychometrika*, 36:1–19, 1971.

E. T. Jaynes. *Probability Theory: The Logic of Science*. Cambridge University Press, Cambridge, 2003.

T. Jefferson. *Notes on the State of Virginia*. John Stockdale, London, 2 edition, 1787.

W. H. Jefferys and J. O. Berger. Ockham's razor and Bayesian analysis. *American Scientist*, 80:64–72, 1992.

H. Jeffreys. On the early history of the solar system. *Monthly Notices of the Royal Astronomical Society*, 78:424–441, 1918.

H. Jeffreys. The rigidity of the earth's central core. *Geophysical Journal International*, 1:371–383, 1926.

H. Jeffreys. *Scientific Inference*. Cambridge University Press, Cambridge, UK, 1 edition, 1931.

H. Jeffreys. Probability, statistics, and the theory of errors. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 140:523–535, 1933a.

H. Jeffreys. On the prior probability in the theory of sampling. *Proceedings of the Cambridge Philosophical Society*, 29:83–87, 1933b.

H. Jeffreys. Some tests of significance, treated by the theory of probability. *Proceedings of the Cambridge Philosophy Society*, 31:203–222, 1935a.

H. Jeffreys. Comment on "The logic of inductive inference" by Ronald A. Fisher. *Journal of the Royal Statistical Society*, 98:70–71, 1935b.

H. Jeffreys. On some criticisms of the theory of probability. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 22: 337–359, 1936a.

H. Jeffreys. Further significance tests. *Mathematical Proceedings of the Cambridge Philosophical Society*, 32:416–445, 1936b.

H. Jeffreys. The unconscious significance of numbers. *International Journal of Psychoanalysis*, 17:217–223, 1936c.

H. Jeffreys. The problem of inference. *Mind*, 45:324–333, 1936d.

H. Jeffreys. Scientific method, causality, and reality. *Proceedings of the Aristotelian Society*, 37:61–70, 1937a.

H. Jeffreys. *Scientific Inference*. Cambridge University Press, Cambridge, UK, 1 edition, 1937b.

H. Jeffreys. The tests for sampling differences and contingency. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 162:479–495, 1937c.

H. Jeffreys. Significance tests when several degrees of freedom arise simultaneously. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 165:161–198, 1938a.

H. Jeffreys. The comparison of series of measures on different hypotheses concerning the standard errors. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 167:367–384, 1938b.

H. Jeffreys. Maximum likelihood, inverse probability and the method of moments. *Annals of Eugenics*, 8:146–151, 1938c.

H. Jeffreys. The nature of mathematics. *Philosophy of Science*, 5:434–451, 1938d.

H. Jeffreys. Science, logic and philosophy. *Nature*, 141:716–719, 1938e.

H. Jeffreys. *Theory of Probability*. Oxford University Press, Oxford, UK, 1 edition, 1939.

H. Jeffreys. Epistemology and modern physics. *Philosophical Magazine, Series 7*, 32:177–205, 1941.

H. Jeffreys. Does a contradiction entail every proposition? *Mind*, 51: 90–91, 1942.

H. Jeffreys. *Theory of Probability*. Oxford University Press, Oxford, UK, 2 edition, 1948.

H. Jeffreys. Bertrand russell on probability. *Mind: A Quarterly Review of Psychology and Philosophy*, 59:313–319, 1950.

H. Jeffreys. The origin of the solar system. *Proceedings of the Royal Society Series A*, 214:281–291, 1952.

H. Jeffreys. The present position in probability theory. *The British Journal for the Philosophy of Science*, 5:275–289, 1955.

H. Jeffreys. probability theory in astronomy. *Monthly Notices of the Royal Astronomical Society*, 117:347–355, 1957.

H. Jeffreys. *Theory of Probability*. Oxford University Press, Oxford, UK, 3 edition, 1961.

H. Jeffreys. Review of "the foundations of statistical inference". *Technometrics*, 3:407–410, 1963.

H. Jeffreys. *Scientific Inference*. Cambridge University Press, Cambridge, UK, 3 edition, 1973.

H. Jeffreys. Fisher and inverse probability. *International Statistical Review*, 42:1–3, 1974.

H. Jeffreys. Probability theory in geophysics. *Journal of the Institute of Mathematics and its Applications*, 19:87–96, 1977.

H. Jeffreys. Some general points in probability theory. In A. Zellner, editor, *Bayesian Analysis in Econometrics and Statistics: Essays in Honor of Harold Jeffreys*, pages 451–453. North-Holland Publishing Company, Amsterdam, The Netherlands, 1980.

H. Jeffreys and B. Swirles, editors. *Collected Papers of Sir Harold Jeffreys on Geophysics and Other Sciences. Volume 6: Mathematics, Probability and Miscellaneous Other Sciences*. Gordon and Breach Science Publishers, London, 1977.

A. Jern, K.-M. K. Chang, and C. Kemp. Belief polarization is not always irrational. *Psychological Review*, 121:206–224, 2014.

H. W. Jevons and H. S. Jevons. William stanley jevons. *Econometrica*, 2: 225–237, 1934.

W. S. Jevons. On the mechanical performance of logical inference. *Philosophical Transactions*, 160:497–521, 1870a.

W. S. Jevons. On the mechanical performance of logical inference. *Proceedings of the Royal Society of London*, 18:166–169, 1870b.

W. S. Jevons. The power of numerical discrimination. *Nature*, 3:281–282, 1871.

W. S. Jevons. *The Principles of Science: A Treatise on Logic and Scientific Method*. MacMillan, London, 1874/1913.

E. O. John. Instrumentalism in science: Comments and criticisms. *Global Journal of Humanities*, 8:1–9, 2018.

P. E. B. Jourdain. Sir George Darwin: A biographical sketch. *The Open Court*, 27:193–201, 1913.

J. M. Joyce. A non–pragmatic vindication of probabilism. *Philosophy of Science*, 65:575–603, 1998.

P. Juola. Authorship attribution. *Foundations and Trends® in Information Retrieval*, 1:233–334, 2006.

M. L. Kalish, T. L. Griffiths, and S. Lewandowsky. Iterated learning: Intergenerational knowledge transmission reveals inductive biases. *Psychonomic Bulletin & Review*, 14:288–294, 2007.

G. Karabatsos. A menu–driven software package of Bayesian nonparametric (and parametric) mixed models for regression analysis and density estimation. *Behavior Research Methods*, 49:335–362, 2017.

R. E. Kass and A. E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90:773–795, 1995.

Z. Kekecs, B. Palfi, B. Szaszi, P. Szecsi, M. Zrubka, M. Kovacs, B. E. Bakos, D. Cousineau, P. Tressoldi, K. Schmidt, M. Grassi, T. R. Evans, Y. Yamada, J. K. Miller, H. Liu, F. Yonemitsu, D. Dubrov, J. P. Röer, M. Becker, R. Schnepper, A. Ariga, P. Arriaga, R. Oliveira, N. Põldver, K. Kreegipuu, B. Hall, S. Wiechert, B. Verschuere, K. Girán, and B. Aczel. Raising the value of research studies in psychological science by increasing the credibility of research reports: The Transparent Psi Project. *Royal Society Open Science*, 10:191375, 2023.

J. B. Keller. The probability of heads. *The American Mathematical Monthly*, 93:191–197, 1986.

A. W. Kemp and C. D. Kemp. Weldon's dice data revisited. *The American Statistician*, 45:216–222, 1991.

M. G. Kendall and P. A. P. Moran. *Geometrical Probability*. Charles Griffin & Company Limited, London, 1963.

J. M. Keynes. *A Treatise on Probability*. Macmillan & Co, London, 1921.

J. M. Keynes. William Stanley Jevons 1835–1882: A centenary allocation on his life and work as economist and statistician. *Journal of the Royal Statistical Society*, 99:516–555, 1936.

R. Khalifa. *Quran – The Final Testament: Authorized English Version of the Original*. 2010. URL `https://www.quranalone.com/media/quran-english.pdf`.

B. J. K. Kleijn and A. W. van der Vaart. Misspecification in infinite–dimensional Bayesian statistics. *Annals of Statistics*, 34:837–877, 2006.

W. E. Knowles Middleton. Archimedes, Kircher, Buffon, and the burning–mirrors. *Isis*, 52:533–543, 1961.

D. E. Knuth. *The Art of Computer Programming. Volume I: Fundamental Algorithms*. Addison–Wesley, Reading, MA, 2 edition, 1973.

E. Kreyszig. Archimedes and the invention of burning mirrors: An investigation of work by Buffon. In J. M. Rassias, editor, *Geometry, Analysis and Mechanics*, pages 139–147. World Scientific Publishing Co., Singapore, 1994.

J. K. Kruschke. *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan*. Academic Press/Elsevier, Amsterdam, 2 edition, 2015.

M. Kumar. *Quantum: Einstein, Bohr and the Great Debate about the Nature of Reality*. Icon Books, London, 2009.

J. Kunert, A. Montag, and S. Pöhlmann. The quincunx: History and mathematics. *Statistical Papers*, 42:143–169, 2001.

W. Kurt. *Bayesian Statistics the Fun Way*. No Starch Press, San Francisco, 2019.

T. O. Kvålseth. Fechner's psychophysical law as a special case of Stevens' three–parameter power law. *Perceptual and Motor Skills*, 75:1205–1206, 1992.

H. E. Kyburg Jr. and H. E. Smokler, editors. *Studies in Subjective Probability*. John Wiley & Sons, New York, 1964.

Z. Labby. Weldon's dice, automated. *CHANCE*, 22:6–13, 2009.

H. Lagerlund. The assimilation of Aristotelian and Arabic logic up to the later thirteenth century. In D. M. Gabbay and J. Woods, editors, *Handbook of the History of Logic. Volume 2: Mediaeval and Renaissance Logic*, pages 281–346. Elsevier, 2008.

B. Lambert. *A Student's Guide to Bayesian Statistics*. SAGE, London, 2018.

P.-S. Laplace. Memoir on the probability of the causes of events. *Statistical Science*, 1:364–378, 1774/1986.

P.-S. Laplace. *Théorie Analytique des Probabilités*. Courcier, Paris, 1812.

P.-S. Laplace. *A Philosophical Essay on Probabilities (Translated by F. W. Truscott and F. L. Emory from the 6th French edition, 1840; first edition 1814)*. Chapman & Hall, London, 1814/1902.

P.-S. Laplace. *Philosophical Essay on Probabilities (Translated by Andrew I. Dale from the 5th French edition, 1825; first edition 1814: Essai Philosophique sur les Probabilités, Paris: Courcier)*. Springer-Verlag, New York, 1814/1995.

M. Lazzarini. Un'applicazione del calcolo dell probilità – alla ricerca sperimentale di un valore approsimato di π. *Periodico di Matematica per l'Insegnamento Secondario*, 17:140–143, 1901.

M. D. Lee. Bayesian methods in cognitive modeling. In J. T. Wixted and E.-J. Wagenmakers, editors, *Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience (4th ed.): Volume 4: Methodology*, pages 37–84. Wiley, New York, 2018.

M. D. Lee and W. Vanpaemel. Determining informative priors for cognitive models. *Psychonomic Bulletin & Review*, 25:114–127, 2018.

M. D. Lee and E.-J. Wagenmakers. *Bayesian Cognitive Modeling: A Practical Course*. Cambridge University Press, 2013.

I. Lehmann. p'. *Publications du Bureau Central Séismologique International Serie A: Travaux Scientifiques*, 14:87–115, 1936.

I. Lehmann. Seismology in the days of old. *Eos*, 68:33–35, 1987.

M. Li and D. B. Dunson. Comparing and weighting imperfect models using D-probabilities. *Journal of the American Statistical Association*, 115: 1349–1360, 2020.

D. V. Lindley. *Introduction to Probability & Statistics from a Bayesian Viewpoint. Part 2. Inference*. Cambridge University Press, Cambridge, 1965.

D. V. Lindley. *Bayesian Statistics, a Review*. SIAM, Philadelphia (PA), 1972.

D. V. Lindley. Bayesian statistics. In W. L. Harper and C. A. Hooker, editors, *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science, Vol. II*, pages 353–363. D. Reidel Publishing Company, Dordrecht–Holland, 1976.

D. V. Lindley. *Making Decisions*. Wiley, London, 2 edition, 1985.

D. V. Lindley. The analysis of experimental data: The appreciation of tea and wine. *Teaching Statistics*, 15:22–25, 1993.

D. V. Lindley. Some comments on Bayes factors. *Journal of Statistical Planning and Inference*, 61:181–189, 1997.

D. V. Lindley. The philosophy of statistics. *The Statistician*, 49:293–337, 2000.

D. V. Lindley. That wretched prior. *Significance*, 1:85–87, 2004.

D. V. Lindley. *Understanding Uncertainty*. Wiley, Hoboken, 2006.

J. Lukasiewicz. Zur Geschichte der Aussagenlogik. *Erkenntnis*, 5:111–131, 1935.

D. Lunn, C. Jackson, N. Best, A. Thomas, and D. Spiegelhalter. *The BUGS Book: A Practical Introduction to Bayesian Analysis*. Chapman & Hall/CRC, Boca Raton (FL), 2012.

A. Ly and E.-J. Wagenmakers. Bayes factors for peri–null hypotheses. *TEST*, 31:1121–1142, 2022.

A. Ly, M. Marsman, A. J. Verhagen, R. P. P. P. Grasman, and E.-J. Wagenmakers. A tutorial on Fisher information. *Journal of Mathematical Psychology*, 80:40–55, 2017.

A. Ly, A. Etz, M. Marsman, and E.-J. Wagenmakers. Replication Bayes factors from evidence updating. *Behavior Research Methods*, 51:2498–2508, 2019.

A. Ly, A. Stefan, J. van Doorn, F. Dablander, D. van den Bergh, A. Sarafoglou, Š. Kucharský, K. Derks, Q. F. Gronau, A. Raj, U. Boehm, E.-J. van Kesteren, M. Hinne, D. Matzke, M. Marsman, and E.-J. Wagenmakers. The Bayesian methodology of Sir Harold Jeffreys as a practical alternative to the p–value hypothesis test. *Computational Brain & Behavior*, 3:153–161, 2020.

A. Ly, D. van den Bergh, F. Bartoš, and E.-J. Wagenmakers. Bayesian inference with JASP. *The ISBA Bulletin*, 28:7–15, 2021.

W. J. Ma, K. P. Kording, and D. Goldreich. *Bayesian Models of Perception and Action: An Introduction*. MIT Press, Cambridge, MA, in press. URL `https://www.cns.nyu.edu/malab/bayesianbook.html`.

H. Maas. *William Stanley Jevons and the Making of Modern Economics*. Cambridge University Press, New York, 2005.

D. J. C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, Cambridge, 2003.

D. M. MacKay. Psychophysics of perceived intensity: A theoretical basis for Fechner's and Stevens' laws. *Science*, 139:1213–1216, 1963.

J. Maddox. False calculation of π by experiment. *Nature*, 370:323, 1994.

M. Maier, F. Bartoš, and E.-J. Wagenmakers. Robust Bayesian meta–analysis: Addressing publication bias with model–averaging. *Psychological Methods*, 28:107–122, 2023.

E. B. Mallon and N. R. Franks. Ants estimate area using Buffon's needle. *Proceedings of the Royal Society B: Biological Sciences*, 267:765–770, 2000.

N. Mantel. An extension of the Buffon needle problem. *The Annals of Mathematical Statistics*, 24:674–677, 1953.

J.-M. Marin and C. P. Robert. *Bayesian Core: A Practical Approach to Computational Bayesian Statistics*. Springer, New York, 2007.

S. Marks and G. Smith. The two–child paradox reborn? *CHANCE*, 24: 54–59, 2011.

E. Martin. *The Calculating Machines: Their History And Development.* (P. Kidwell and M. Williams, Trans.). The MIT Press; Tomash Publishers, London; Los Angeles/San Francisco, 1925/1992.

R. McElreath. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. Chapman & Hall/CRC Press, Boca Raton (FL), 2016.

J. McFadden. Razor sharp: The role of Occam's razor in science. *Annals of the New York Academy of Sciences*, 1530:8–17, 2023.

S. B. McGrayne. *The Theory That Would Not Die: How Bayes' Rule Cracked The Enigma Code, Hunted Down Russian Submarines, And Emerged Triumphant From Two Centuries Of Controversy*. Yale University Press, Yale, 2011.

E. C. Merkle and Y. Rosseel. blavaan: Bayesian structural equation models via parameter expansion. *Journal of Statistical Software*, 85, 2018.

D. E. Meyer and R. W. Schvaneveldt. Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, 90:227–234, 1971.

A. A. Mills and R. Clift. Reflections on the 'Burning mirrors of Archimedes'. With a consideration of the geometry and intensity of sunlight reflected from plane mirrors. *European Journal of Physics*, 13: 268–279, 1992.

C. Misak. *Frank Ramsey: A Sheer Excess of Powers*. Oxford University Press, Oxford, 2020.

T. J. Mitchell and J. J. Beauchamp. Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83:1023–1032, 1988.

T. Miyake. Scientific inference and the earth's interior: Dorothy Wrinch and Harold Jeffreys at Cambridge. In F. Stadler, editor, *Integrated History and Philosophy of Science, Vol. 20*, pages 81–91. Springer, Cambridge, 2017.

R. D. Morey and J. N. Rouder. Bayes factor approaches for testing interval null hypotheses. *Psychological Methods*, 16:406–419, 2011.

R. D. Morey, J. N. Rouder, A. J. Verhagen, and E.-J. Wagenmakers. Why hypothesis tests are essential for psychological science: A comment on Cumming. *Psychological Science*, 25:1289–1290, 2014.

R. D. Morey, R. Hoekstra, J. N. Rouder, M. D. Lee, and E.-J. Wagenmakers. The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review*, 23:103–123, 2016a.

R. D. Morey, J. W. Romeijn, and J. N. Rouder. The philosophy of Bayes factors and the quantification of statistical evidence. *Journal of Mathematical Psychology*, 72:6–18, 2016b.

Richard D. Morey and Jeffrey N. Rouder. BayesFactor 0.9.12-4.2. Comprehensive R Archive Network, 2018. URL `http://cran.r-project.org/web/packages/BayesFactor/index.html`.

B. Mosselmans. *William Stanley Jevons and the Cutting Edge of Economics*. Routledge, London, 2007.

F. Mosteller and D. L. Wallace. Inference in an authorship problem. *Journal of the American Statistical Association*, 58:275–309, 1963.

F. Mosteller and D. L. Wallace. *Applied Bayesian and Classical Inference: The Case of The Federalist Papers*. Springer, New York, 2 edition, 1984.

F. Mosteller and C. Youtz. Quantifying probabilistic expressions. *Statistical Science*, 5:2–12, 1990.

M. Mulder, E.-J. Wagenmakers, R. Ratcliff, W. Boekel, and B. U. Forstmann. Bias in the brain: A diffusion model analysis of prior probability and potential payoff. *Journal of Neuroscience*, 32:2335–2343, 2012.

I. J. Myung. The importance of complexity in model selection. *Journal of Mathematical Psychology*, 44:190–204, 2000.

I. J. Myung. Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology*, 47:90–100, 2003.

I. J. Myung and M. A. Pitt. Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review*, 4:79–95, 1997.

D. J. Navarro, D. R. Foxcroft, and T. J. Faulkenberry, editors. *Learning Statistics With JASP: A Tutorial for Psychology Students and Other Beginners*. 2019.

E. I. Newman. A method of estimating the total length of root in a sample. *Journal of Applied Ecology*, 3:139–145, 1966.

I. Newton. *The Mathematical Principles of Natural Philosophy*. (A. Motte, Trans.). Daniel Adee, New York, 1726/1846.

R. S. Nickerson. Ambiguities and unstated assumptions in probabilistic reasoning. *Psychological Bulletin*, 120:410–433, 1996.

S. Nieuwenhuis, B. U. Forstmann, and E.-J. Wagenmakers. Erroneous analyses of interactions in neuroscience: A problem of significance. *Nature Neuroscience*, 14:1105–1107, 2011.

I. Niiniluoto. *Truthlikeness*. D. Reidel, Dordrecht, 1987.

I. Niiniluoto. Truthlikeness: Old and new debates. *Synthese*, 197:1581–1599, 2020.

R. Nobles and D. Schiff. Misleading statistics within criminal trials: The Sally Clark case. *Significance*, 2:17–19, 2005.

A. Nordgaard, R. Ansell, W. Drotz, and L. Jaeger. Scale of conclusions for the value of evidence. *Law, Probability and Risk*, 11:1–24, 2012.

T. Norsen and S. Nelson. Yet another snapshot of foundational attitudes toward quantum mechanics. 2013. URL `https://arxiv.org/pdf/1306.4646v2.pdf`.

A. O'Hagan. Fractional Bayes factors for model comparison. *Journal of the Royal Statistical Society B*, 57:99–138, 1995.

A. O'Hagan. Dicing with the unknown. *Significance*, 1:132–133, 2004.

A. O'Hagan and J. Forster. *Kendall's Advanced Theory of Statistics Vol. 2B: Bayesian Inference (2nd ed.)*. Arnold, London, 2004.

A. M. Overstall and R. King. conting: An R package for Bayesian analysis of complete and incomplete contingency tables. *Journal of Statistical Software*, 58:1–27, 2014.

B. Pascal. *Traité du Triangle Arithmétique, Avec Quelques Autres Petits Traitez sur la Mesme Matière*. Guillaume Desprez, Paris, 1665.

R. Patai and J. Patai. *The Myth of the Jewish Race (revised edition)*. Wayne State University Press, Detroit, 1989.

S. Pawel, F. Aust, L. Held, and E.-J. Wagenmakers. Normalized power priors always discount historical data. *Stat*, 12:e591, 2023.

S. Pawel, F. Aust, L. Held, and E.-J. Wagenmakers. Power priors for replication studies. *TEST*, in press.

S. Pawel, A. Ly, and E.-J. Wagenmakers. Evidential calibration of confidence intervals. *The American Statistician*, in press.

K. Pearson. *The Grammar of Science*. J. M. Dent & Sons, London, 1892/1937.

K. Pearson. X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine Series 5*, 50:157–175, 1900.

K. Pearson. *National Life From the Standpoint of Science*. Adam & Charles Black, London, 1901.

K. Pearson. Walter Frank Raphael Weldon. 1860–1906. *Biometrika*, 5: 1–52, 1906.

K. Pearson. *The Life, Letters and Labours of Francis Galton. Volume I: Birth 1822 to Marriage 1853*. Cambridge University Press, Cambridge, 1914.

K. Pearson. *The Life, Letters and Labours of Francis Galton. Volume II: Researches Of Middle Life*. Cambridge University Press, Cambridge, 1924.

K. Pearson. *The Life, Letters and Labours of Francis Galton. Volume IIIa: Correlation, Personal Identification and Eugenics*. Cambridge University Press, Cambridge, 1930a.

K. Pearson. *The Life, Letters and Labours of Francis Galton. Volume IIIb: Characterization, Especially by Letters*. Cambridge University Press, Cambridge, 1930b.

K. Pearson and M. Moul. The problem of alien immigration into Great Britain, illustrated by an examination of Russian and Polish Jewish children: Part II. *Annals of Eugenics*, 1:56–127, 1925.

S. Peart. *The Economics of W. S. Jevons*. Routledge, London, 1996.

C. S. Peirce. The probability of induction. *Popular Science Monthly*, 12: 705–718, 1878.

J. Pek and T. Van Zandt. Frequentist and Bayesian approaches to data analysis: Evaluation and estimation. *Psychology Learning & Teaching*, 19: 21–35, 2020.

J. M. Pérez and J. O. Berger. Expected–posterior prior distributions for model selection. *Biometrika*, 89:491–511, 2002.

W. Perks. Some observations on inverse probability including a new indifference rule. *Journal of the Institute of Actuaries*, 73:285–334, 1947.

M. D. Perlman and M. J. Wichura. Sharpening Buffon's needle. *The American Statistician*, 29:157–163, 1975.

J. M. Pfadt, D. van den Bergh, K. Sijtsma, M. Moshagen, and E.-J. Wagenmakers. Bayesian estimation of single–test reliability coefficients. *Multivariate Behavioral Research*, 57:620–641, 2022.

S. T. Piantadosi. One parameter is always enough. *AIP Advances*, 8: 095118, 2018.

D. Pisa, R. Alonso, A. Rábano, I. Rodal, and L. Carrasco. Different brain regions are infected with fungi in Alzheimer's disease. *Scientific Reports*, 5:15015, 2015.

M. A. Pitt and I. J. Myung. When a good fit can be bad. *Trends in Cognitive Sciences*, 6:421–425, 2002.

M. Plummer. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In K. Hornik, F. Leisch, and A. Zeileis, editors, *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*. Vienna, Austria, 2003.

H. Poincaré. *Calcul des Probabilités*. Gauthier-Villars et Fils, Paris, 1896.

H. Poincaré. *The Foundations of Science* (G. B. Halsted, Trans.). The Science Press, New York, 1913.

G. Pólya. *Mathematics and Plausible Reasoning: Vol. I. Induction and Analogy in Mathematics*. Princeton University Press, Princeton, NJ, 1954a.

G. Pólya. *Mathematics and Plausible Reasoning: Vol. II. Patterns of Plausible Inference*. Princeton University Press, Princeton, NJ, 1954b.

G. Pólya. *How to Solve It*. Princeton University Press, Princeton, NJ, 2 edition, 1957.

K. R. Popper. What is dialectic? *Mind*, 49:403–426, 1940.

K. R. Popper. Are contradictions embracing? *Mind*, 52:47–50, 1943.

K. R. Popper. *The Logic of Scientific Discovery*. Harper Torchbooks, New York, 1959.

K. R. Popper. *Conjectures And Refutations: The Growth of Scientific Knowledge*. Routledge and Kegan Paul, London, 4 edition, 1972.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2004. URL `http://www.R-project.org`. ISBN 3–900051–00–3.

H. Raiffa and R. Schlaifer. *Applied Statistical Decision Theory*. Harvard Business School, Boston, 1961.

J. F. Ramaley. Buffon's noodle problem. *The American Mathematical Monthly*, 76:916–918, 1969.

F. P. Ramsey. Truth and probability. In R. B. Braithwaite, editor, *The Foundations of Mathematics and Other Logical Essays*, pages 156–198. Kegan Paul, London, 1926.

R. Ratcliff. A theory of memory retrieval. *Psychological Review*, 85: 59–108, 1978.

R. Ratcliff. A diffusion model account of response time and accuracy in a brightness discrimination task: Fitting real data and failing to fit fake but plausible data. *Psychonomic Bulletin & Review*, 9:278–291, 2002.

R. Ratcliff and J. N. Rouder. Modeling response times for two–choice decisions. *Psychological Science*, 9:347–356, 1998.

J. Rissanen. A universal prior for integers and estimation by minimum description length. *Annals of Statistics*, 11:416–431, 1983.

J. Rissanen. Complexity of models. In W. H. Zurek, editor, *Complexity, Entropy and the Physics of Information*, pages 117–125. CRC Press, Boca Raton, FL, 1990.

C. P. Robert. *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation (2nd ed.)*. Springer, New York, 2007.

C. P. Robert. On the Lindley–Jeffreys paradox. *Philosophy of Science*, 81: 216–232, 2014.

C. P. Robert, N. Chopin, and J. Rousseau. Harold Jeffreys's Theory of Probability revisited. *Statistical Science*, 24:141–172, 2009.

S. Roberts and H. Pashler. How persuasive is a good fit? A comment on theory testing in psychology. *Psychological Review*, 107:358–367, 2000.

R. M. Robertson. Jevons and his precursors. *Econometrica*, 19:229–249, 1951.

J. Roger. *Buffon: A Life in Natural History (Translated by S. L. Bonnefoi)*. Cornell University Press, Ithaca, 1997.

J. W. Romeijn and R. van de Schoot. A philosopher's view on Bayesian evaluation of informative hypotheses. In H. Hoijtink, I. Klugkist, and P. A. Boelen, editors, *Bayesian Evaluation of Informative Hypotheses*, pages 181–207. Springer Verlag, New York, 2008.

L. Rosa, E. Rosa, L. Sarner, and S. Barrett. A close look at therapeutic touch. *The Journal of the American Medical Association*, 279:1005–1010, 1998.

J. Rosenberg, M. Kubsch, E.-J. Wagenmakers, and M. Dogucu. Making sense of uncertainty in the science classroom: A Bayesian approach. *Science & Education*, 31:1239–1262, 2022.

R. D. Rosenkrantz. Simplicity. In W. L. Harper and C. A. Hooker, editors, *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science: Volume I Foundations and Philosophy of Epistemic Applications of Probability Theory*, pages 167–196. Reidel, Dordrecht, 1976.

R. D. Rosenkrantz. *Inference, Method and Decision*. Reidel, Dordrecht, 1977.

R. D. Rosenkrantz. Measuring truthlikeness. *Synthese*, 45:463–487, 1980.

R. D. Rosenkrantz. Does the philosophy of induction rest on a mistake? *The Journal of Philosophy*, 79:78–97, 1982.

R. D. Rosenkrantz. Why Glymour is a Bayesian. In J. Earman, editor, *Testing Scientific Theories*, pages 69–98. University of Minnesota Press, Minneapolis, 1983.

R. L. Rosnow and R. Rosenthal. Statistical procedures and the justification of knowledge in psychological science. *American Psychologist*, 44: 1276–1284, 1989.

J. N. Rouder and J. Lu. An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin & Review*, 12:573–604, 2005.

J. N. Rouder and R. D. Morey. Teaching Bayes' theorem: Strength of evidence as predictive accuracy. *The American Statistician*, 73:186–190, 2019.

J. N. Rouder, J. Lu, P. L. Speckman, D. Sun, and Y. Jiang. A hierarchical model for estimating response time distributions. *Psychonomic Bulletin & Review*, 12:195–223, 2005.

J. N. Rouder, J. Lu, R. D. Morey, D. Sun, and P. L. Speckman. A hierarchical process dissociation model. *Journal of Experimental Psychology: General*, 137:370–389, 2008.

J. N. Rouder, R. D. Morey, A. J. Verhagen, J. M. Province, and E.-J. Wagenmakers. Is there a free lunch in inference? *Topics in Cognitive Science*, 8:520–547, 2016a.

J. N. Rouder, R. D. Morey, and E.-J. Wagenmakers. The interplay between subjectivity, statistical practice, and psychological science. *Collabra*, 2:1–12, 2016b.

J. N. Rouder, J. M. Haaf, and F. Aust. From theories to models to predictions: A Bayesian model comparison approach. *Communication Monographs*, 85:41–56, 2018.

R. Royall. *Statistical Evidence: A Likelihood Paradigm*. Chapman & Hall, London, 1997.

B. Russell. Is there a God? In B. Russell, editor, *Last Philosophical Testament 1943–68*, pages 542–548. Routledge, London, 1952/1997.

B. Russell. *Autobiography*. Routledge, New York, 1975/2009.

C. Sagan. *The Demon-Haunted World: Science as a Candle in the Dark*. Ballantine Books, New York, 1995.

F. J. Samaniego. *A Comparison of the Bayesian and Frequentist Approaches to Estimation*. Springer, New York, 2010.

F. J. Samaniego. An ode to Dennis Lindley. In A. O'Hagan, editor, *A Book For Dennis*, pages 125–129. Blurb, 2013.

A. N. Sanborn and T. T. Hills. The frequentist implications of optional stopping on Bayesian hypothesis tests. *Psychonomic Bulletin & Review*, 21: 283–300, 2014.

L. A. Santaló. *Integral Geometry and Geometric Probability*. Addison–Wesley, Reading, MA, 1976.

A. Sarafoglou, F. Aust, M. Marsman, E.-J. Wagenmakers, and J. M. Haaf. multibridge: An R package to evaluate informed hypotheses in binomial and multinomial models. *Behavior Research Methods*, in press.

M. Schabas. *A World Ruled by Number: William Stanley Jevons and the Rise of Mathematical Economics*. Princeton University Press, Princeton, 1990.

D. J. Schad, B. Nicenboim, P.-C. Bürkner, M. Betancourt, and S. Vasishth. Workflow techniques for the robust use of Bayes factors. *Psychological Methods*, in press.

M. Schlosshauer, J. Kofler, and A. Zeilinger. A snapshot of foundational attitudes toward quantum mechanics. 2013. URL `https://arxiv.org/pdf/1301.1069v1.pdf`.

H. Schmidli, S. Gsteiger, S. Roychoudhury, A. O'Hagan, D. Spiegelhalter, and B. Neuenschwander. Robust meta–analytic–predictive priors in clinical trials with historical control information. *Biometrics*, 70:1023–1032, 2014.

A. Schopenhauer. *Die Beiden Grundprobleme der Ethik*. Johann Christian Hermannsche Buchhandlung (F. E. Suchsland), Frankfurt am Main, 1841.

A. Schopenhauer. *The Two Fundamental Problems of Ethics*. Cambridge University Press, Cambridge, 2009.

L. L. Schroeder. Buffon's needle problem: An exciting application of many mathematical concepts. *The Mathematics Teacher*, 67:183–186, 1974.

J. N. Schupbach. *Bayesianism and Scientific Reasoning*. Cambridge University Press, Cambridge, 2022.

J. Scott. 2. On the burning mirrors of Archimedes, and on the concentration of light produced by reflectors. *Proceedings of the Royal Society of Edinburgh*, 6:232–235, 1869.

J. G. Scott and J. O. Berger. An exploration of aspects of Bayesian multiple testing. *Journal of Statistical Planning and Inference*, 136:2144–2162, 2006.

J. G. Scott and J. O. Berger. Bayes and empirical–Bayes multiplicity adjustment in the variable–selection problem. *Annals of Statistics*, 38: 2587–2619, 2010.

M. Senechal. *I Died for Beauty: Dorothy Wrinch and the Cultures of Science*. Oxford University Press, New York, 2012.

S. Senn. *Statistical Issues in Drug Development*. John Wiley & Sons, Chichester, 2 edition, 2007.

O. B. Sheynin. H. Poincaré's work on probability. *Archive for History of Exact Sciences*, 42:137–171, 1991.

R. M. Shiffrin and S. H. Chandramouli. Model selection, data distributions, and reproducibility. In H. Atmanspacher and S. Maasen, editors, *Reproducibility: Principles, Problems, Practices, and Prospects*, pages 115–140. John Wiley & Sons, Hoboken (NJ), 2016.

R. M. Shiffrin, M. D. Lee, W. Kim, and E.-J. Wagenmakers. A survey of model evaluation approaches with a tutorial on hierarchical Bayesian methods. *Cognitive Science*, 32:1248–1284, 2008.

R. M. Shiffrin, S. H. Chandramouli, and P. D. Grünwald. Bayes factors, relations to minimum description length, and overlapping model classes. *Journal of Mathematical Psychology*, 72:56–77, 2016.

A. Shimony. Scientific inference. In R. C. Colodny, editor, *The Nature and Function of Scientific Theories: Essays in Contemporary Science and Philosophy*, pages 79–172. The University of Pittsburgh Press, Pittsburgh, 1970.

N. Silver. *The Signal and the Noise: The Art and Science of Prediction*. Allen Lane, London, 2012.

E. Siniksaran. Throwing Buffon's needle with Mathematica. *The Mathematica Journal*, 11:71–90, 2008.

S. Sivasundaram and K. H. Nielsen. Surveying the attitudes of physicists concerning foundational issues of quantum mechanics. 2016. URL `https://arxiv.org/pdf/1612.00676.pdf`.

A. F. M. Smith. Discussion of O'Hagan. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57:120–122, 1995.

C. A. B. Smith. Personal probability and statistical analysis. *Journal of the Royal Statistical Society A*, 128:469–499, 1965.

R. Smith? *Mathematical Modelling of Zombies*. University of Ottawa Press, Canada, 2014.

E. Sober. Instrumentalism, parsimony, and the Akaike framework. *Philosophy of Science*, 69:S112–S123, 2002.

E. Sober. *Ockham's Razors: A User's Manual*. Cambridge University Press, Cambridge, 2015.

H. Solomon. *Geometric Probability*. SIAM, Philadelphia, 1978.

R. J. Solomonoff. A formal theory of inductive inference. Part I. *Information and Control*, 7:1–22, 1964a.

R. J. Solomonoff. A formal theory of inductive inference. Part II. *Information and Control*, 7:224–254, 1964b.

C. Sommer. Another survey of foundational attitudes towards quantum mechanics. 2013. URL `https://arxiv.org/pdf/1303.2719.pdf`.

H. Spall. Sir Harold Jeffreys: An interview. *Earthquake Information Bulletin*, 12:48–53, 1980.

D. J. Spiegelhalter, K. R. Abrams, and J. P. Myles. *Bayesian Approaches to Clinical Trials and Health–Care Evaluation*. John Wiley & Sons, Chichester, UK, 2004.

J. Sprenger and S. Hartmann. *Bayesian Philosophy of Science*. Oxford University Press, Oxford, 2019.

Stan Development Team. RStan: The R interface to Stan, 2020. URL `http://mc-stan.org/`. R package version 2.21.2.

A. M. Stefan, Q. F. Gronau, F. D. Schönbrodt, and E.-J. Wagenmakers. A tutorial on Bayes Factor Design Analysis using an informed prior. *Behavior Research Methods*, 51:1042–1058, 2019.

A. M. Stefan, N. J. Evans, and E.-J. Wagenmakers. Practical challenges and methodological flexibility in prior elicitation. *Psychological Methods*, 27:177–197, 2022.

S. S. Stevens. To honor Fechner and repeal his law. *Science*, 133:80–86, 1961.

S. S. Stevens. *Psychophysics: Introduction to its perceptual, neural, and social prospects*. Wiley, New York, 1975.

I. Stewart. *In Pursuit of the Unknown: 17 Equations That Changed the World*. Basic Books, New York, 2012.

I. Stewart. *Do Dice Play God? The Mathematics of Uncertainty*. Basic Books, New York, 2019.

S. M. Stigler. *The History of Statistics: The Measurement of Uncertainty Before 1900*. Harvard University Press, Cambridge, MA, 1986a.

S. M. Stigler. Laplace's 1774 memoir on inverse probability. *Statistical Science*, 1:359–378, 1986b.

S. M. Stigler. *Statistics on the Table: The History of Statistical Concepts and Methods*. Harvard University Press, Cambridge, MA, 1999.

J. V. Stone. *Bayes' Rule with R: A Tutorial Introduction to Bayesian Analysis*. Sebtel Press, 2016.

J. Strachey and A. Strachey. *Bloomsbury/Freud: The Letters of James and Alix Strachey 1924–1925.* (P. Meisel and W. Kendrick, Eds.). Basic Books, New York, 1986.

S. Subramanian. *A Dominant Character: How J. B. S. Haldane Transformed Genetics, Became a Communist, and Risked his Neck for Science*. W. W. Norton & Company, New York, 2019.

C. Sutton. 'nullius in verba' and 'nihil in verbis': Public understanding of the role of language in science. *The British Journal for the History of Science*, 27:55–64, 1994.

R. G. Swinburne. The paradoxes of confirmation: A survey. *American Philosophical Quarterly*, 8:318–330, 1971.

B. Swirles. Reminiscences and discoveries: Harold Jeffreys from 1891 to 1940. *Notes and Records: The Royal Society of the History of Science*, 46: 301–308, 1992.

W. Świątkowski and A. Carrier. There is nothing magical about Bayesian statistics: An introduction to epistemic probabilities in data analysis for psychology starters. *Basic and Applied Social Psychology*, 42:387–412, 2020.

G. 't Hooft. *The Cellular Automaton Interpretation of Quantum Mechanics*. Springer Open, Cham, 2016.

S. Talts, M. Betancourt, D. Simpson, A. Vehtari, and A. Gelman. Validating Bayesian inference algorithms with simulation–based calibration. *ArXiv Preprint*, 2018. URL `https://arxiv.org/abs/1804.06788`.

D. G. Taylor. *Games, Gambling, and Probability: An Introduction to Mathematics*. CRC Press, Boca Raton, FL, 2 edition, 2021.

M. Tegmark. The interpretation of quantum mechanics: Many worlds or many words? 1997. URL `https://arxiv.org/pdf/quant-ph/9709032.pdf`.

M. Tegmark and J. A. Wheeler. 100 years of quantum mysteries. *Scientific American*, 284:68–75, 2001.

A. R. Templeton. Coherent and incoherent inference in phylogeography and human evolution. *Proceedings of the National Academy of Sciences*, 107: 6376–6381, 2010a.

A. R. Templeton. Reply to Berger et al.: Improving ABC. *Proceedings of the National Academy of Sciences*, 107:E158, 2010b.

M. Theil. The role of translations of verbal into numerical probability expressions in risk management: A meta–analysis. *Journal of Risk Research*, 5:177–186, 2002.

S. P. Thompson. *Calculus Made Easy*. MacMillan and Co., London, 1910.

W. M. Thorburn. The myth of Occam's razor. *Mind*, 27:345–353, 1918.

S. Thrun, W. Burgard, and D. Fox. *Probabilistic Robotics*. The MIT Press, Cambridge, MA, 2005.

I. Todhunter. *A History of the Mathematical Theory of Probability From the Time of Pascal to That of Laplace*. Cambridge, MacMillan and Co., 1865.

G. F. Townsend. *Three Hundred Æsop's Fables, Literally Translated from the Greek*. George Routledge and Sons, London, 1887.

J. T. Townsend. The mind–body equation revisited. In C. Cheng, editor, *Philosophical Aspects of the Mind–Body Problem*, pages 200–218. Honolulu University Press, Honolulu, USA, 1975.

G. Tredoux. *Comrade Haldane is Too Busy to go on Holiday: The Genius Who Spied for Stalin*. Encounter Books, New York, 2018.

A. M. Turing. The applications of probability to cryptography. *UK National Archives, HW* $25/37$, 1941/2012.

F. Tuyl. A method to handle zero counts in the multinomial model. *The American Statistician*, 73:151–158, 2019.

F. Tuyl, R. Gerlach, and K. Mengersen. On the certainty of an inductive inference: The binomial case. *Statistical Science*, in press.

A. Tversky. Intransitivity of preferences. *Psychological Review*, 76:31–48, 1969.

A. Tversky and D. Kahneman. Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, 90: 293–315, 1983.

D. van den Bergh, J. M. Haaf, A. Ly, J. N. Rouder, and E.-J. Wagenmakers. A cautionary note on estimating effect size. *Advances in Methods and Practices in Psychological Science*, 4:1–8, 2021.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, Cambridge, 1998.

N. N. N. van Dongen, E.-J. Wagenmakers, and J. Sprenger. A Bayesian perspective on severity: Risky predictions and specific hypotheses. *Psychonomic Bulletin & Review*, 30:516–533, 2023.

J. van Doorn, D. Matzke, and E.-J. Wagenmakers. An in–class demonstration of Bayesian inference. *Psychology Learning and Teaching*, 19: 36–45, 2020.

J. van Doorn, D. van den Bergh, U. Böhm, F. Dablander, K. Derks, T. Draws, A. Etz, N. J. Evans, Q. F. Gronau, M. Hinne, Š. Kucharský, A. Ly, M. Marsman, D. Matzke, A. R. Komarlu Narendra Gupta, A. Sarafoglou, A. Stefan, J. G. Voelkel, and E.-J. Wagenmakers. The JASP guidelines for conducting and reporting a Bayesian analysis. *Psychonomic Bulletin & Review*, 28:813–826, 2021.

J. B. van Doorn. A brief introduction to Bayesian inference: From tea to beer. 2023. URL `https://johnnydoorn.github.io/BayesBookQuarto/`.

B. C. Van Fraassen. *Laws and Symmetry*. Clarendon Press, Oxford, 1989.

H. van Maanen. Het stokje van Lazzarini. *Skepter*, 31:8–12, 2018.

J. Vandekerckhove, D. Matzke, and E.-J. Wagenmakers. Model comparison and the principle of parsimony. In J. Busemeyer, J. Townsend, Z. J. Wang, and A. Eidels, editors, *Oxford Handbook of Computational and Mathematical Psychology*, pages 300–319. Oxford University Press, 2015.

J. Vandekerckhove, J. N. Rouder, and J. K. Kruschke. Editorial: Bayesian methods for advancing psychological science. *Psychonomic Bulletin & Review*, 25:1–4, 2018.

W. Vanpaemel. Measuring model complexity with the prior predictive. *Advances in Neural Information Processing Systems*, 22:1919–1927, 2009.

W. Vanpaemel. Prior sensitivity in theory testing: An apologia for the Bayes factor. *Journal of Mathematical Psychology*, 54:491–498, 2010.

W. Vanpaemel. Strong theory testing using the prior predictive and the data prior. *Psychological Review*, 127:136–145, 2020.

D. Veen, D. Stoel, N. Schalken, K. Mulder, and R. van de Schoot. Using the data agreement criterion to rank experts' beliefs. *Entropy*, 20:592, 2018.

S. Velasco, F. L. Román, A. González, and J. A. White. Statistical estimation of some irrational numbers using an extension of Buffon's needle experiment. *International Journal of Mathematical Education in Science and Technology*, 37:735–740, 2006.

I. Verdinelli and L. Wasserman. Computing Bayes factors using a generalization of the Savage–Dickey density ratio. *Journal of the American Statistical Association*, 90:614–618, 1995.

A. J. Verhagen and E.-J. Wagenmakers. Bayesian tests to quantify the result of a replication attempt. *Journal of Experimental Psychology: General*, 143:1457–1475, 2014.

J. M. Villarreal, A. J. Etz, and M. D. Lee. Evaluating the complexity and falsifiability of psychological models. *Psychological Review*, 130:853–872, 2023.

G. W. F. von Leibniz. Letter to Herman Conring. In L. E. Loemker, editor, *Philosophical Papers and Letters (2nd ed.)*, pages 186–191. University of Chicago Press, Chicago, IL, 1678/1989.

R. von Mises (Ed.). *P.S. de Laplace (1814): Philosophischer Versuch über die Wahrscheinlichkeit*. Akademische Verlagsgesellschaft M. B. R., Leipzig, 1932.

E.-J. Wagenmakers. A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, 14:779–804, 2007.

E.-J. Wagenmakers. Methodological and empirical developments for the Ratcliff diffusion model of response times and accuracy. *European Journal of Cognitive Psychology*, 21:641–671, 2009.

E.-J. Wagenmakers. *Bayesian Thinking for Toddlers*. JASP Publishing, Amsterdam, 2020.

E.-J. Wagenmakers and Q. F. Gronau. De Bayesiaanse leercyclus [the Bayesian learning cycle]. *STAtOR*, 19:8–13, 2018.

E.-J. Wagenmakers, P. Grünwald, and M. Steyvers. Accumulative prediction error and the selection of time series models. *Journal of Mathematical Psychology*, 50:149–166, 2006.

E.-J. Wagenmakers, H. J. L. van der Maas, and R. P. P. P. Grasman. An EZ–diffusion model for response time and accuracy. *Psychonomic Bulletin & Review*, 14:3–22, 2007.

E.-J. Wagenmakers, R. Ratcliff, P. Gomez, and G. McKoon. A diffusion model account of criterion shifts in the lexical decision task. *Journal of Memory and Language*, 58:140–159, 2008.

E.-J. Wagenmakers, T. Lodewyckx, H. Kuriyal, and R. Grasman. Bayesian hypothesis testing for psychologists: A tutorial on the Savage–Dickey method. *Cognitive Psychology*, 60:158–189, 2010.

E.-J. Wagenmakers, R. D. Morey, and M. D. Lee. Bayesian benefits for the pragmatic researcher. *Current Directions in Psychological Science*, 25: 169–176, 2016a.

E.-J. Wagenmakers, A. J. Verhagen, and A. Ly. How to quantify the evidence for the absence of a correlation. *Behavior Research Methods*, 48: 413–426, 2016b.

E.-J. Wagenmakers, G. Dutilh, and A. Sarafoglou. The creativity-verification cycle in psychological science: New methods to combat old idols. *Perspectives on Psychological Science*, 13:418–427, 2018a.

E.-J. Wagenmakers, Q. F. Gronau, and J. Vandekerckhove. Five Bayesian intuitions for the stopping rule principle. *Manuscript submitted for publication*, 2018b.

E.-J. Wagenmakers, J. Love, M. Marsman, T. Jamil, A. Ly, A. J. Verhagen, R. Selker, Q. F. Gronau, D. Dropmann, B. Boutin, F. Meerhoff, P. Knight, A. Raj, E.-J. van Kesteren, J. van Doorn, M. Šmíra, S. Epskamp, A. Etz, D. Matzke, T. de Jong, D. van den Bergh, A. Sarafoglou, H. Steingroever, K. Derks, J. N. Rouder, and R. D. Morey. Bayesian inference for psychology. Part II: Example applications with JASP. *Psychonomic Bulletin & Review*, 25:58–76, 2018c.

E.-J. Wagenmakers, M. Marsman, T. Jamil, A. Ly, A. J. Verhagen, J. Love, R. Selker, Q. F. Gronau, M. Šmíra, S. Epskamp, D. Matzke, J. N. Rouder, and R. D. Morey. Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review*, 25:35–57, 2018d.

E.-J. Wagenmakers, M. D. Lee, J. N. Rouder, and R. D. Morey. The principle of predictive irrelevance or why intervals should not be used for model comparison featuring a point null hypothesis. In C. W. Gruber, editor, *The Theory of Statistics in Psychology – Applications, Use and Misunderstandings*, pages 111–129. Springer, Cham, 2020.

E.-J. Wagenmakers, Q. F. Gronau, F. Dablander, and A. Etz. The support interval. *Erkenntnis*, 87:589–601, 2022.

E.-J. Wagenmakers, S. Zabell, and Q. F. Gronau. J. B. S. Haldane's rule of succession. *Statistical Science*, in press. URL `https://arxiv.org/abs/2307.09489`.

L. Wasserman. Bayesian model selection and model averaging. *Journal of Mathematical Psychology*, 44:92–107, 2000.

J. W. N. Watkins. *Science and Scepticism*. Princeton University Press, Princeton, NJ, 1985.

J. Watson. *The Double Helix*. Athenaeum, New York, 1968.

S. F. Weiss. After the fall: Political whitewashing, professional posturing, and personal refashioning in the postwar career of Otmar Freiherr von Verschuer. *Isis*, 101:722–758, 2010.

D. Wells. Are these the most beautiful? *The Mathematical Intelligencer*, 12:37–41, 1990.

R. Wetzels, J. G. W. Raaijmakers, E. Jakab, and E.-J. Wagenmakers. How to quantify support for and against the null hypothesis: A flexible WinBUGS implementation of a default Bayesian t test. *Psychonomic Bulletin & Review*, 16:752–760, 2009.

R. Wetzels, R. P. P. P. Grasman, and E.-J. Wagenmakers. An encompassing prior generalization of the Savage–Dickey density ratio test. *Computational Statistics & Data Analysis*, 54:2094–2102, 2010.

J. E. B. Wilkie and G. V. Bodenhausen. Are numbers gendered? *Journal of Experimental Psychology: General*, 141:206–210, 2012.

J. E. B. Wilkie and G. V. Bodenhausen. The numerology of gender: Gendered perceptions of even and odd numbers. *Frontiers in Psychology*, 6:810, 2015.

S. Willems, C. Albers, and I. Smeets. Variability in the interpretation of probability phrases used in Dutch news articles – a risk for miscommunication. *Journal of Science Communication*, 19:A03, 2020.

S. Willis, L. McKenna, S. McDermott, G. O'Donell, A. Barrett, B. Rasmusson, A. Nordgaard, C. Berger, M. Sjerps, J. Lucena–Molina, G. Zadora, C. Aitken, T. Lovelock, L. Lunt, C. Champod, A. Biedermann, T. Hicks, and F. Taroni. ENFSI guideline for evaluative reporting in forensic science: Strengthening the evaluation of forensic results across Europe (STEOFRAE). Technical report, 2015.

K. J. Wilson and M. Farrow. Combining judgements from correlated experts. In L. C. Dias, A. Morton, and J. Quigley, editors, *Elicitation: The Science and Art of Structuring Judgement*, pages 211–240. Springer International Publishing, Cham, 2018.

R. Wohl. Buffon and his project for a new science. *Isis*, 51:186–199, 1960.

G. R. Wood and J. M. Robertson. Buffon got it straight. *Statistics & Probability Letters*, 37:415–421, 1998.

D. Wrinch and H. Jeffreys. On some aspects of the theory of probability. *Philosophical Magazine*, 38:715–731, 1919.

D. Wrinch and H. Jeffreys. On certain fundamental principles of scientific inquiry. *Philosophical Magazine*, 42:369–390, 1921.

D. Wrinch and H. Jeffreys. On certain fundamental principles of scientific inquiry. *Philosophical Magazine*, 45:368–374, 1923.

Y. Yao, A. Vehtari, D. Simpson, and A. Gelman. Using stacking to average Bayesian predictive distributions. *Bayesian Analysis*, 13:917–1007, 2018.

K. D. S. Young and L. I. Pettit. Measuring discordancy between prior and data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58:679–689, 1996.

S. Zabell. The secret life of I. J. Good. *Statistical Science*, 38:285–302, 2023.

S. L. Zabell. W. E. Johnson's "sufficientness" postulate. *Annals of Statistics*, 10:1090–1099, 1982.

S. L. Zabell. The rule of succession. *Erkenntnis*, 31:283–321, 1989.

S. L. Zabell. Book review of "a world ruled by number: William Stanley Jevons and the rise of mathematical economics". *Journal of the History of the Behavioral Sciences*, 28:171–176, 1992.

S. L. Zabell. *Symmetry and its Discontents: Essays on the History of Inductive Probability*. Cambridge University Press, New York, 2005.

S. L. Zabell. Carnap and the logic of inductive inference. In D. M. Gabbay, S. Hartmann, and J. Woods, editors, *Handbook of the History of Logic. Volume 10: Inductive Logic*, pages 265–309. Elsevier, North-Holland, 2011.