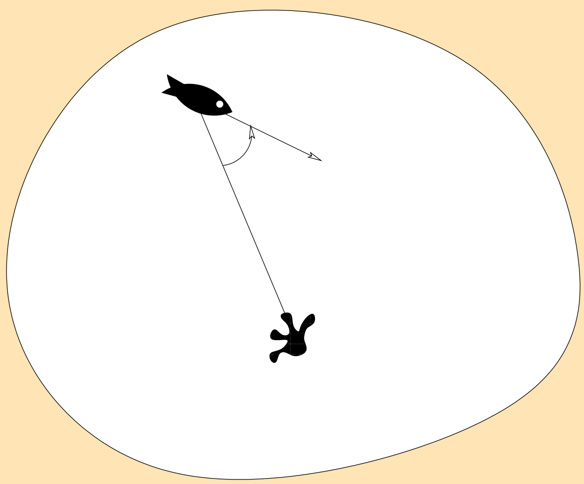
# *Zen and the Art of Bayesian Analysis*

## *C. Pozrikidis*



VERSION 26.01

CHESTER & BENNINGTON

# Zen and the Art of Bayesian Analysis

C. Pozrikidis

Chester & Bennington

2025

Zen and the Art of Bayesian Analysis



Published by Chester & Bennington



Pozrikidis, C.
Zen and the Art of Bayesian Analysis/ C. Pozrikidis – Version 26.01

# *Contents*

# *Preface*

You walk into an expensive restaurant and ask a customer to reveal their degree of general happiness on a scale from zero to ten. Bayes' theorem will tell you how to estimate the annual income of that person, given some information on income against happiness non-specific to the person. Zen will ask you not to be presumptuous about the person and wish them well, no matter how well off they are.

You want to rediscover yourself and you are thinking about taking a road trip. Bayes' theorem will tell you whether the trip will help, given some information non-specific to your current frame of mind. Zen will tell you that you can rediscover yourself and escape the unpleasant, the mundane, and the tedious in a wheelchair.

You are a frog in a pond and you sense in your toes that your friend, Funny Fins, is approaching. Bayes' theorem will tell you how far Funny Fins is and if she is moving toward you, given some generic analysis of fluid flow pertinent to swimming. Zen will tell you that there is always time for Bayesian analysis, but not enough time to spend with a dear friend, especially when the friend's name is groundhog Bilbo.

To apply Bayes' theorem, elementary concepts from statistics and mathematics are required. To understand the consequences of the theorem, an open mind full, creativity, and willingness to think instead of accept is required. The former is explained and the latter is encouraged in this book.

One mental block in applying Bayes' theorem is that forward and backward reasoning must be simultaneously exercised. For example, we may consider a tomato plant that thrived and ask: *how much was it watered?* or we may water a tomato plant by a certain amount and ask: *will it wither or thrive?* Once this mental block has been overcome, the rest is easy. Deductions and inferences based on Bayesian analysis range from useful, to thought provoking, to profound.

My two main goals in this book are to (*a*) introduce Bayes' theorem from a rigorous yet informal standpoint, and (*b*) discuss methods of Bayesian analysis in a broad range of applications and diverse settings.

All necessary concepts are defined and introduced for a self-contained discourse; elementary background information from combinatorics is provided in an appendix.

This book is addressed to a diverse audience including teachers, professors, students, and anyone is interested in learning the essence of Bayesian analysis. Familiarity with high-school level mathematics is only required in most early sections, while college-level mathematics is required in more advanced sections. The reader may select the desired level of mathematical comfort and skip sections that appear too mathematical, without compromising the understanding of subsequent material. Several Matlab[1] codes performing computations, simulations, and visualization are listed for illustration.

Narratives in the form of commentary and short stories are interspersed in the book. The selection of the narratives is guided by a prime directive that is consistent with the Bayesian approach: inner goodness amounts to developing an internal prism that projects the rays of fairness onto the subconsciousness, while rejecting the harmful components. Bayes' theorem allows us to anticipate, estimate, and then reject the harmful components, preferably in foresight and hopefully in hindsight. All short stories included in this books are real.

By reading this book, you will learn what you already know. You approach a problem, situation, suggestion, concept with some idea; you get some data, input, measurements, observations or insights to test your idea; and then you get closer to the truth and revise your initial idea. Perhaps more important, by reading this book you will affirm that nothing occurs in vacuum, and that actions have short-term, long-term, predictable, and surprising consequences. You never really know what kind of trauma people carry inside them behind a smile.

C. Pozrikidis
2025

---

[1]Matlab® is a proprietary computing environment for numerical computation and data visualization. Matlab and Simulink are registered trademarks of The MathWorks, Inc. For product information, please contact: The MathWorks, Inc., 3 Apple Hill Drive, Natick, MA 01760-2098, USA, Tel: 508-647-7000, Fax: 508-647-7001, E-mail: info@mathworks.com, Web: www.mathworks.com

*It's not impossible that the Hummer that just cut me off is maybe being driven by a father whose little child is hurt or sick in the seat next to him, and he's trying to get this kid to the hospital, and he's in a bigger, more legitimate hurry than I am: it is actually I who am in HIS way.*

This is water

David Foster Wallace

# *Notation*

| | |
|---|---|
| $\Pi_{(\circ)}$ | Unconditional probability of event $\circ$<br>Prior probability |
| $\Pi_{(\circ\,\vert\,\oplus)}$ | Conditional probability of event $\circ$, given event $\oplus$<br>Posterior probability |
| $\mathcal{L}_{\mathbf{x}}$ | Likelihood function for data $\mathbf{x}$ |
| | |
| $\mathcal{B}^n_m(\theta)$ | Binomial distribution with Bernoulli probability $\theta$ |
| $\mathcal{C}^n_m$ | Binomial coefficient |
| $\mathfrak{Be}_{\alpha,\beta}(\theta)$ | Beta distribution with Bernoulli probability $\theta$ |
| $\mathcal{N}_{\overline{\theta},\sigma}(\theta)$ | Normal (Gaussian) distribution<br>with expected value $\overline{\theta}$ and standard deviation $\sigma$ |
| | |
| | A Boolean variable can take the value of $1$<br>standing for true, or $0$, standing for false |
| | |
| $q$ | Boolean variable for *quality* |
| $p$ | Boolean variable for *pass* |
| | |
| $\equiv$ | Equal by definition, defined as |
| $=$ | Equal sign |
| $\times$ | Multiplication sign |
| $\cdot$ | (dot) Multiplication sign |
| $\cap$ | Intersection symbol for the overlap of two events |
| $\cup$ | Union symbol |
| $\bar{\circ}$ | Complement of $\circ$ |
| $[a, b]$ | A closed interval; a variable $x$ can take<br>any value $a \leq x \leq b$ |
| $(a, b)$ | An open interval; a variable $x$ can take<br>any value $a < x < b$ |
| $\int$ | Integration symbol |
| $\sum$ | Summation symbol |
| $\prod$ | Product symbol |

# Chapter 1

# Coping with uncertainty

We have invented the concept of probability to quantify certainty and cope with uncertainty regarding a tangible or intangible entity that involves an element of chance.

Examples include *an observation, a tentative concept, a notion, a measurement, an assessment, an outcome, an event, an occurrence, a shaken belief, a conjecture, an opinion, an alleged natural law, a theory, a hypothesis, or the result of an election.*

The notion of joint probability provides us with a framework for handling multiple events, whereas the notion of conditional probability provides us with a framework for handling events that occur in a specified context.

Fundamental probability concepts and a suitable mathematical framework are discussed in this chapter. Bayes' rule is introduced in Chapter 2 and applications of the rule are discussed in the remainder of this book.

## 1.1 The concept of probability

The probability of an entity that admits a probabilistic interpretation, denoted by $\circ$, is a variable, denoted by

$$\Pi_{(\circ)}. \tag{1.1.1}$$

For example, if $\circ$ is the assessment that a cake is sweet, then $\Pi_{(\circ)}$ is the probability that a cake is sweet. If $\circ$ is the assertion that a certain integer is even, then $\Pi_{(\circ)}$ is the probability that the integer is even.

More generally, the entity $\circ$ can be represented by a mathematical statement or described in narrative form. For example, $\circ$ can stand for the hypothesis that *our universe is part of a multiverse*.

By stipulation, the probability can take any value between and including zero and unity,

$$0 \le \Pi_{(\circ)} \le 1. \tag{1.1.2}$$

The limiting values, $\Pi_{(\circ)} = 0, 1$, indicate absolute certainty, where the value of zero indicates undoubtful negation and the value of one indicates absolute affirmation. Intermediate values indicate various degrees of doubt or uncertainty.

Sometimes, a probabilistic entity is also described as stochastic. The word *stochastic* derives from the Greek word $\sigma\tau\acute{o}\chi o\varsigma$ which means *target*. The underlying implication is that an effort misses a target and spreads around the centerpoint in some random fashion.

### 1.1.1 Probabilistic against deterministic

The probabilistic description should be contrasted with the deterministic description underlying a strict and rigid natural, universal, spiritual, religious, or other law.

An event or outcome is deterministic when it follows from an absolute truth or else it arises without a doubt under certain conditions. For example, if one insults a seminar speaker, one will surely be given a chair at the far end of the dinner table. A deterministic outcome never misses a physical, abstract, or conceptual target.

The probabilistic/deterministic distinction is analogous to the relative/absolute moral distinction where a continuous spectrum of gray truths rather than a single monochromatic truth are allowed. The moral spectrum is a prime source of internal conflict.

### 1.1.2 Negative probability?

The first axiom of probability theory stipulates that negative probability values and values higher than unity are allowed only as a figure of speech.

However, quasi-probability distributions in quantum mechanics and elsewhere allow for negative probabilities in an appropriate context.

### *1.1.3 Rain*

A farmer may look over her corn fields, then look at the sky with her expert eyes and whisper to her horses "*there is a 10% chance that it will rain tomorrow*".

Her prediction can be rephrased as "*on a scale from 0 to 10, the chance of rain is 1*", and expressed as

$$\Pi_{(\texttt{It will rain tomorrow})} = 0.10. \tag{1.1.3}$$

These equivalent statements convey a near certainty that it will not rain tomorrow (low probability for rain), though they allow for a small chance that it may rain.

The complementary probability is conveyed by the statement "*there is a 90% chance that it will not rain tomorrow*", which can be rephrased as "*on a scale from 0 to 10, the chance of no rain is 9*", and expressed as

$$\Pi_{(\texttt{It will not rain tomorrow})} = 0.90. \tag{1.1.4}$$

Since the events "*it will rain*" and "*it will not rain*" are complementary and mutually exclusive, the two aforementioned probabilities *must* add to unity,

$$\Pi_{(\texttt{It will rain tomorrow})} + \Pi_{(\texttt{It will not rain tomorrow})} = 1 \tag{1.1.5}$$

according to the second axiom of probability theory.

### *1.1.4 Interpretation*

Any time the word *probability* is encountered or employed, an appropriate interpretation should be sought through a systematic mental search for the most suitable context.

One should feel free to interpret probability in any way that *feels right* according to our life experiences, insights, and intuition.

Sometimes, probability should be understood strictly quantitatively as *fraction*, *percentage*, or *frequency* of repeatable events.

Other times, probability should be understood in less quantitative terms as *possibility*, *hunch*, or *preponderance of the evidence.*

Yet other times, probability should be understood in the framework of groundhog days, years, or lifetimes.

### 1.1.5 Parking

A person may say: "*there is a 10% chance we will find a parking spot this morning.*"

Another person may rephrase: "*on a scale from 0 to 10, the chance of finding a parking spot is 1.0.*"

A professional statistician may awkwardly rephrase: "*the probability of finding a parking spot this morning is 0.1.*"

These equivalent statements could be interpreted in one of the following, and perhaps other, ways:

(*a*) One out of ten parking spots will be empty.
(*b*) One out of ten drivers finds a parking spot.
(*c*) Only one out of ten times I found a parking spot.

The parking example supports our assertion that the most sensible interpretation of probability should be chosen by exercising common sense, experience, and intuition.

### 1.1.6 Groundhog days

In a brilliant movie entitled *Groundhog day*, fictitious character Phil Connors wakes up every morning in Punxsutawney, Pennsylvania, as though he fell asleep on the same previous day. Different events occur during each repeated day, and the spell is removed only when Phil becomes a genuinely kind person.

Important lessons can be learned from Phil's struggles and eventual success in breaking the cycle. The movie script teaches us that if one stops criticizing and manipulating others, one will feel much better about themselves, and thus be accepted much better by others, thereby replacing a vicious with a virtuous circle.

This possibility is supported with Einstein's astute observation: "*insanity is doing the same thing over and over and expecting a different outcome*." In fact, in Chapter 3 we will see that Einstein's observation is a corollary by Bayes' rule derived in Chapter 2.

### *1.1.7 Groundhog years*

One may think with anticipation: "*if I go to the lunch room to get some coffee, I may run into my favorite colleague and exchange our latest thoughts on the meaning of Leaving Las Vegas*." The sense of probability conveyed by this statement is best interpreted in the framework of groundhog days. If you say: "*I will probably never finish reading The Great Gatsby*", you are most certainly expressing a disappointment in the context of groundhog lifetimes.

Human nature has a propensity for familiarity and comfort found in regularity and repetition: things are and work as they should. The toothbrush should be in the medicine cabinet where we left it when we went to sleep, and an apple should always fall straight down under a tree. It is unsettling when things change suddenly beyond our control.

### *1.1.8 Meaning by comparison*

The probability of an event or anything else that admits a probabilistic description acquires practical meaning when compared meaningfully to the probabilities of alternative events. For example, the probability of having an English muffin for breakfast can be compared to that of having bacon and eggs. It makes little sense to compare the probability of having an English muffin to the price of crude oil.

### *1.1.9 Interpretation of probability in Bayesian analysis*

We will see that, in the context of Bayesian analysis, probability is often and best interpreted as *plausibility*, *degree of assurance*, *belief*, *assumption*, *theory*, or *intuitive likelihood.*

For example, we may consider the probability that the universal constants determining the physical laws were chosen by God or a higher force, or the probability that a loved one will call on the phone today. Similarly, we may consider the probability that a certain person will help us at a time of need.

The Bayesian interpretation of probability is less strict than that of the classical probability defined with regard to a repetition of events, called the frequenistic interpretation. The relaxed context is both forgiving and empowering.

More important, the Bayesian framework allows us to revise a prior rigorous or intuitive probability based on a single piece of data or a sequence of data.

It is interesting that the notion of probability in quantum mechanics appears to be inconsistent with both the frequenistic and the Bayesian interpretation and admits its own abstract mathematical interpretation. Some secrets of the universe are hard to unravel.

***Exercises***

**1.1.1** A retail store carries shirts in small, medium, large, and extra large size. You are told that the probability of finding a large size is small. Discuss the meaning of this statement.

**1.1.2** Two woodchuck siblings want to know the probability of day or night outside their hole as they wake up on the first day of spring. Discuss the best interpretation of this probability.

**1.1.3** What is the probability that the average person on the street is familiar with the movie *Groundhog day?* Discuss the meaning of this probability and the essence of the screenplay.

## 1.2 The sample space

A defined collection of possible mutually exclusive outcomes, events, or occurrences defines a sample space. Common sense suggests that the probabilities of these events must add precisely to unity $(1.0)$. An implicit stipulation is that one of these events has to occur.

For example, in the case of two mutually exclusive outcomes regarding snow, we require that

$$\Pi_{(\texttt{It will snow})} + \Pi_{(\texttt{It will not snow})} = 1. \tag{1.2.1}$$

If you keep tea, coffee, and peppermint in your pantry, then the probabilities of having one of them for breakfast add up to unity,

$$\Pi_{(\texttt{Will have tea})} + \Pi_{(\texttt{Will have coffee})} + \Pi_{(\texttt{Will have peppermint})} = 1. \tag{1.2.2}$$

An implied assumption is that you will have a beverage. The beverage collection could be extended or diminished.

Probabilities of outcomes, events, or occurrences that belong to the same sample space could be compared, added or subtracted, as appropriate.

Probabilities of events that belong to different sample spaces could also be added, subtracted, compared, and multiplied under the notion of the *joint probability* discussed in Section 1.3.

### 1.2.1 Implied assumptions

Implied assumptions endow a sample space with a framework or context. Such assumptions are sometimes taken for granted in comparing alternative events. Implied assumptions are best conveyed by the notion of conditional probability discussed in Section 1.9.

It is important to keep in mind that misunderstandings regarding underlying context may lead to false deductions and inaccurate interpretations leading to arguments and quarrels.

### 1.2.2 Complement of an event

Consider an event, denoted by $\circ$. The complement of this event is the union of all other events in the same sample space, denoted by $\bar{\circ}$, such that

$$\Pi_{(\circ)} + \Pi_{(\bar{\circ})} = 1. \tag{1.2.3}$$

By definition, the probability of an event and the probability of the complement of the event add up to unity. The reason is that an event and its complement are mutually exclusive.

For example, the complement of "*I will have tea*" is "*I will not have tea*", which amounts to having coffee, another available beverage, or nothing at all.

If the sample space contains all integers, and an event is that a number is odd, the complement of this event is that the number is even.

### 1.2.3 The moon rotates around the earth

The complement of "*everyone agreed that the moon rotates about the earth*" is "*at least one person did not agree that the moon rotates around the earth*."

Note that the complement is *not* that one person did not agree that the moon rotates around the earth, since two, three, or a higher number of persons may have disagreed.

### 1.2.4 Venn diagrams

An illuminating visual description of a sample space and its constituents is provided by a Venn diagram.

In this diagram, the sample space is the interior of a rectangle or another container, and an event is represented by a blob or another regular or irregular shape inside the container, as shown in Figure 1.2.1. The complement of an event is the region outside the blob inside the container.

In a *quantitative Venn diagram*, the container is a unit square or rectangle, and the area occupied by a blob, representing a certain event, is equal to the probability of that event.

### 1.2.5 Mutually exclusive events

A Venn diagram showing two mutually exclusive events is illustrated in Figure 1.2.1(*a*).

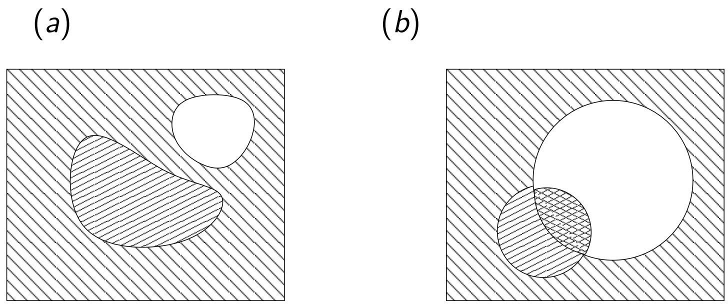



FIGURE 1.2.1 In a Venn diagram, the sample space is the interior of a rectangle or another container, while an event is represented by a blob inside the container. Illustration of (*a*) two mutually exclusive events and (*b*) two non-mutually exclusive events.

With reference to a radio, the first event can be "*the radio is tuned in the AM band*" and the second event can be "*the radio is tuned in the FM band.*" The region outside both blobs is the event that "*the radio is turned off.*" The three regions constitute a sample space defined by three mutually exclusive states.

### 1.2.6 Non-mutually exclusive events

Two events do not have to be mutually exclusive. A Venn diagram illustrating two non-mutually exclusive events is shown in Figure 1.2.1(*b*). The lenticular overlap of the two disks is the overlap of both events.

For example, the statements "*the ignition coils are bad*" and "*the transmission is bad*" are not mutually exclusive. One can be true, the second can be true, or both can be true if a car is a true lemon.

The lenticular overlap in Figure 1.2.1(*b*) is the joint event that the ignition coils are bad *and* the transmission is bad. The area outside the blobs is the happy event that "*neither the ignition coils are bad nor the transmission is bad.*"

### 1.2.7 Four regions

The four regions shown in Figure 1.2.1(*b*) represent four members of a joint-probability sample space defined by the two non-mutually exclusive events and their complements, as discussed in Section 1.4.

### *Exercises*

**1.2.1** What is the complement of the event "*the dog ate my homework?*"

**1.2.2** Define a sample space of your choice.

## 1.3 Arbitrary number of random events

As a generalization, we consider $N$ mutually exclusive entities or events defining a sample space, denoted as

$$\circ_1, \quad \circ_2, \quad \ldots, \quad \circ_N. \tag{1.3.1}$$

The sum of the associated probabilities is unity,

$$\Pi_{(\circ_1)} + \Pi_{(\circ_2)} + \cdots + \Pi_{(\circ_N)} = 1. \tag{1.3.2}$$

Using the summation symbol, we write

$$\sum_{i=1}^{N} \Pi_{(\circ_i)} = 1. \tag{1.3.3}$$

In fact, the number of events, $N$, does not have to be finite, that is, it can be infinite as the result of mathematical idealization.

### 1.3.1 Maximum entropy principle

For convenience, we denote $p_i \equiv \Pi_{(\circ_i)}$, where

$$\sum_{i=1}^{N} p_i = 1. \tag{1.3.4}$$

The information entropy is the expected value of the negative of the logarithm of each probability,

$$s \equiv -\sum_{i=1}^{N} p_i \times \ln p_i. \tag{1.3.5}$$

Since $p_i \le 1$ and thus $\ln p_i < 0$, the information entropy is positive and becomes zero only when $p_i = 1$ for some $i$.

Using the properties of the logarithm, we find that

$$\mathrm{e}^{-s} = \prod_{i=1}^{N} p_i^{p_i}, \tag{1.3.6}$$

where $\prod$ denotes the product.

To find a probability distribution, $p_i$, when it is not specified, we introduce the constrained information entropy

$$\widehat{s} \equiv s + \lambda \left( \sum_{i=1}^{N} p_i - 1 \right), \tag{1.3.7}$$

where $\lambda$ is an *a prior* unknown coefficient regarded as a Lagrange multiplier.

Next, we set $\partial \widehat{s}/\partial p_i = 0$ for $i = 1, \ldots, N$ and also $\partial \widehat{s}/\partial \lambda = 0$ to find the conditions for maximum entropy. Carrying out the differentiations, we obtain

$$-\ln(p_i) = 1 - \lambda, \qquad \sum_{i=1}^{N} p_i = 1. \tag{1.3.8}$$

The solution is readily found to be

$$p_i = \frac{1}{N}, \qquad \lambda = 1 - \ln N \tag{1.3.9}$$

for $i = 1, \ldots, N$. We have derived the uniform probability distribution, which is the intuitive choice in the absence of information.

Other probability distributions can be derived by incorporating further constraints on the right-hand side of (1.3.7).

### *1.3.2 Random integers*

A collection of $N$ mutually exclusive random events defining a sample space, $\circ_i$, can be labelled by a random integer, $i$, taking values inside a specified range

$$i_{\min} \le i \le i_{\max}, \tag{1.3.10}$$

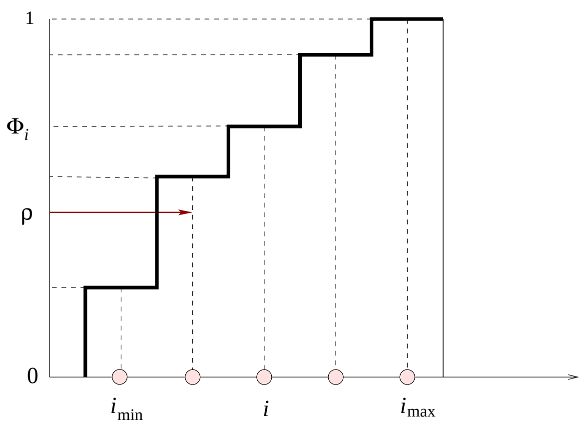



FIGURE 1.3.1 Illustration of the cumulative discrete probability distribution associated with random integers.

with corresponding probability distribution $p_i$, where $i_{\min}$ and $i_{\max}$ are specified lower and upper limits and the number of events is

$$N = i_{\max} - i_{\min} + 1. \tag{1.3.11}$$

Normalization requires that all probabilities add to unity,

$$\sum_{i=i_{\min}}^{i_{\max}} p_i = 1. \tag{1.3.12}$$

The associated cumulative discrete probability distribution is defined in terms of a sum as

$$\Phi_i \equiv \sum_{j=i_{\min}}^{i} p_j. \tag{1.3.13}$$

A graph of $\Phi_i$ against $i$ provides us with a generally uneven staircase function, as shown in Figure 1.3.1.

### *1.3.3 Generating a sample*

To generate a sequence of integers, $i$, consistent with a specified probability distribution, we introduce a random variable, $\varrho$, with uniform probability distribution in the interval $[0, 1]$, called *random deviate or variate*, In Matlab, the random deviate $\varrho$ can be generated using the internal function *rand*. The pertinent value of $i$ is then assigned as follows:

- If $\varrho \le p_{i_{\min}}$, then $i = i_{\min}$
- If $p_{j-1} < \varrho \le p_j$, then $i = j$ for $j = 2, \ldots, i_{\max}$

The mapping is illustrated with the horizontal arrow in Figure 1.3.1.

The method is implemented in the following Matlab function *dpd* that receives `imin`, `imax`, the probability distribution (`p`), and a random deviate (`r`), and returns the value of $i$ corresponding to $r$, consistent with the specified probability distribution, (dpd):

```
function i = dpd(imin,imax,p,r)

%---
% Phi: cumulative distribution
% r:    random deviate
%---

Phi = p(imin);

if(Phi >= r)
  i = imin;
  return
end

for j=imin+1:imax
 Phi = Phi + p(j);
 if(Phi >= r)
   i=j;
   return;
 end
end

%---
return
%---
```

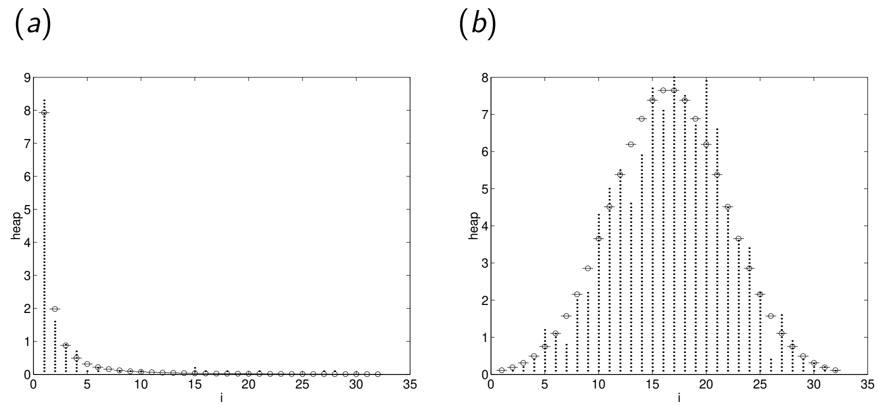



FIGURE 1.3.2 Heaps of random numbers corresponding to (*a*) the discrete probability distribution described by (1.3.14) and (*b*) the Gaussian distribution with unit standard deviation.

Heaps of random numbers described by the probability distribution

$$p_i = \frac{1}{c}\,\frac{1}{i^2}, \tag{1.3.14}$$

in the range $i = 1, \ldots, 32$ are shown in Figure 1.3.2(*a*), where

$$c = \sum_{j=1}^{32} \frac{1}{j^2} \tag{1.3.15}$$

is an appropriate normalization constant expressed by a convergent sum.

Heaps of random numbers in the range $i = 1, \ldots, 32$ conforming with the Gaussian probability density function with mean value $16.5$ and unit standard deviation are shown in Figure 1.3.2(*b*).

***Exercise***

**1.3.1** Compute and plot heaps of random numbers described by the discrete probability distribution $\Pi_i = c/i^3$ in the range $i = 1, \ldots, 32$, where $c$ is an appropriate normalization constant.

## 1.4 Joint probability

We have seen that events may occur simultaneously if they are not mutually exclusive: we can take walk in the park *and* enjoy a hot dog at the same time.

By definition, the joint probability of two events is the probability that *both* events take place at the same time or at different times according to context. The joint probability is sometimes called the conjoint probability.

### 1.4.1 Notation

The joint probability of an event, denoted as $\circ$, and another event, denoted as $\oplus$, is denoted as

$$\Pi_{(\circ \text{ and } \oplus)}. \tag{1.4.1}$$

In abbreviated notation, the *and* is replaced by a comma, so that

$$\Pi_{(\circ\,,\,\oplus)} \equiv \Pi_{(\circ \text{ and } \oplus)}, \tag{1.4.2}$$

where the equivalence sign, $\equiv$, can be read as *same as.*

### 1.4.2 Set intersection

In formal mathematics and logic, the joint probability is denoted by the set intersection symbol, $\cap$, that is,

$$\Pi(\circ \cap \oplus) \equiv \Pi_{(\circ\,,\,\oplus)}. \tag{1.4.3}$$

This notation has its origin in the set theory of mathematics where Venn diagrams are commonplace. In this book, we will use the comma notation $(,)$.

### 1.4.3 Snow and sardines

One example of a joint probability is

$$\Pi_{(\texttt{It will snow today}\,,\,\texttt{The price of sardines will rise})}. \tag{1.4.4}$$

This is the probability that it will snow today *and* the price of sardines will rise. Another example is

$$\Pi_{(\texttt{It will rain today}\,,\,\texttt{The buses will be running late in the afternoon})}. \tag{1.4.5}$$

We will see that each one of these two joint probabilities refers to a sample space defined by the two constituent sample spaces and their complements.

### 1.4.4 Irrelevance of listing order

Common sense suggests that

$$\Pi_{(\circ\,,\,\oplus)} = \Pi_{(\oplus\,,\,\circ)}, \tag{1.4.6}$$

that is, the order by which the events are listed inside the parentheses, separated by a comma, is immaterial. For example,

$$\Pi_{(\texttt{My cat is pretty}\,,\,\texttt{My dog is hillarious})} = \\ \Pi_{(\texttt{My dog is hillarious}\,,\,\texttt{My cat is pretty})}. \tag{1.4.7}$$

The irrelevance of the listing order is the cornerstone of Bayes' rule discussed in Chapter 2.

### 1.4.5 Joint-probability sample space

The joint probability refers to a composite sample space with four constituents represented by the pairs

$$(\circ\,,\,\oplus), \qquad (\circ\,,\,\overline{\oplus}), \qquad (\overline{\circ}\,,\,\oplus), \qquad (\overline{\circ}\,,\,\overline{\oplus}), \tag{1.4.8}$$

where we recall that an overline denotes the complement. The four pairs of events underwriting these probabilities are mutually exclusive.

With regard to the Venn diagram shown in Figure 1.3.1(*b*), the four combinations shown in (1.4.8) are represented by the four shaded regions. Because one shaded region does not appear in the Venn diagram shown in Figure 1.3.1(*a*), the corresponding probability is zero.

Since the four combinations exhaust all possibilities, the four probabilities must add to unity,

$$\Pi_{(\circ\,,\,\oplus)} + \Pi_{(\circ\,,\,\overline{\oplus})} + \Pi_{(\overline{\circ}\,,\,\oplus)} + \Pi_{(\overline{\circ}\,,\,\overline{\oplus})} = 1. \tag{1.4.9}$$

If three of these probabilities are known, the fourth one can be computed from this equation.

### 1.4.6 First day of spring

For example, we may consider the following events and their complements:

$\circ$: day $\overline{\circ}$: night $\oplus$: warm $\overline{\oplus}$: cold

A woodchuck may guess that it is a warm day, a cold day, a warm night, or a cold night outside his hole on the first day of spring as he emerges to taste some wild onions after a long slumber.

### 1.4.7 Cartesian or tensor product

To describe the sample space of a joint probability concisely, we say that it is a tensor product or Cartesian product of two constituent sample spaces, denoted by the symbol $\otimes$. With regard to the sample space shown in (1.4.8), we write

$$\text{sample space} = (\circ, \overline{\circ}) \otimes (\oplus, \overline{\oplus}) \tag{1.4.10}$$

to obtain four combinations. The corresponding probabilities can be arranged in a $2 \times 2$ matrix

$$\begin{bmatrix} \Pi_{(\circ\,,\,\oplus)} & \Pi_{(\circ\,,\,\overline{\oplus})} \\ \Pi_{(\overline{\circ}\,,\,\oplus)} & \Pi_{(\overline{\circ}\,,\,\overline{\oplus})} \end{bmatrix}. \tag{1.4.11}$$

The four elements of this matrix add to unity.

The events and their complements could be resolved further into partitions, yielding a matrix with higher dimensions.

### 1.4.8 Law of total probability

Since an event, $\oplus$, and it complement, $\overline{\oplus}$, are mutually exclusive, it must be that

$$\Pi_{(\circ)} = \Pi_{(\circ\,,\,\oplus)} + \Pi_{(\circ\,,\,\overline{\oplus})}. \tag{1.4.12}$$

This equation expresses the simplest version of the law of total probability. The right-hand side is the sum of the elements in the first row of the matrix shown in (1.4.11). For example,

$$\Pi_{(\texttt{eggs})} = \Pi_{(\texttt{eggs with bacon})} + \Pi_{(\texttt{eggs without bacon})}. \tag{1.4.13}$$

Similarly, we write

$$\Pi_{(\overline{\circ})} = \Pi_{(\overline{\circ}\,,\,\oplus)} + \Pi_{(\overline{\circ}\,,\,\overline{\oplus})}, \tag{1.4.14}$$

and also

$$\Pi_{(\oplus)} = \Pi_{(\oplus\,,\,\circ)} + \Pi_{(\oplus\,,\,\overline{\circ})} \tag{1.4.15}$$

and

$$\Pi_{(\overline{\oplus})} = \Pi_{(\overline{\oplus}\,,\,\circ)} + \Pi_{(\overline{\oplus}\,,\,\overline{\circ})}. \tag{1.4.16}$$

Identifying the event $\oplus$ with the event $\circ$, we obtain

$$\Pi_{(\circ\,,\,\circ)} = \Pi_{(\circ)}, \qquad \Pi_{(\overline{\circ}\,,\,\overline{\circ})} = \Pi_{(\overline{\circ})}, \qquad \Pi_{(\circ\,,\,\overline{\circ})} = 0. \tag{1.4.17}$$

The probability of being hot and hot at the same time is equal to the probability of being hot. The probability of being hot and cold at the same time is zero. When $\oplus = \circ$, the matrix shown in (1.4.11) is diagonal.

### *1.4.9 Marginal probabilities*

In the context of the sums shown on the right-hand sides of (1.4.12) through (1.4.16), the pure probabilities $\Pi_{(\circ)}$ and $\Pi_{(\oplus)}$ are called *marginal probabilities.*

The origin of this terminology can be traced to accounting where sums are added and recorded at the margins of a spreadsheet. Recall that the joint probabilities and the marginal probabilities belong to different sample spaces.

### *1.4.10 Inequalities*

Expansions (1.4.12) and (1.4.12) imply that a joint probability is no greater than any marginal probability,

$$\Pi_{(\circ\,,\,\oplus)} \leq \Pi_{(\circ)}, \qquad \Pi_{(\circ\,,\,\oplus)} \leq \Pi_{(\oplus)}. \tag{1.4.18}$$

Note that the probabilities on the left and right side of these inequalities belong to different sample spaces.

For example, the probability that we may run out of coffee *and* sugar is less than or equal to the probability that we may run out of coffee or sugar alone.

### *1.4.11 Law of total probability*

Let the following $N$ mutually exclusive events define a sample space,

$$\circ_1, \quad \circ_2, \quad \ldots, \quad \circ_N, \tag{1.4.19}$$

that is, one of the events *has* to occur, so that the sum of the associated probabilities is unity,

$$\sum_{i=1}^{N} \Pi_{(\circ_i)} = 1, \tag{1.4.20}$$

as discussed in Section 1.2. The probability of another arbitrary event, $\circ$, is given by the sum of $N$ joint probabilities according to the law of total probability

$$\Pi_{(\circ)} = \sum_{i=1}^{N} \Pi_{(\circ\,,\,\circ_i)}. \tag{1.4.21}$$

The law of total probability finds important applications in Bayesian analysis, as discussed in Chapter 2.

We may identity $\circ$ with $\circ_j$, where $j = 1, \ldots, N$, set $\Pi_{(\circ_j\,,\,\circ_i)} = 0$ for $j \neq i$, and recover $\Pi_{(\circ_i\,,\,\circ_i)} = \Pi_{(\circ_i)}$.

### 1.4.12 Continuous range of events

The $N$ events, $\circ_i$, can be labeled by a set of parameter values $\theta_i$ that vary monotonically from the first event, $i = 1$, to the last event, $i = N$, by equal intervals, $\Delta\theta$. The value $\theta_i$ implies event $\circ_i$, and *vice versa*.

The law of total probability expressed by (1.4.21) may then be written as

$$\Pi_{(\circ)} = \Delta\theta \sum_{i=1}^{N} \varphi(\circ\,,\ \theta_i), \tag{1.4.22}$$

where

$$\varphi(\circ\,,\ \theta_i) \equiv \frac{1}{\Delta\theta}\,\Pi_{(\circ\,,\,\circ_i)}. \tag{1.4.23}$$

When $N$ is large, the right-hand side of (1.4.22) can be approximated with an integral, yielding

$$\Pi_{(\circ)} = \int \varphi(\circ\,,\ \theta)\,\mathrm{d}\theta, \tag{1.4.24}$$

where $\varphi(\circ, \theta)$ is a joint probability distribution density function (pdf). The integration is performed over an appropriate integration domain.

### 1.4.13 Pairs of events

A joint probability sample space consists of $M$ pairs of events, $(\circ_i, \oplus_i)$, whose probabilities add to unity,

$$\sum_{i=1}^{M} \Pi_{(\circ_i, \oplus_i)} = 1. \tag{1.4.25}$$

Events with zero probability could be excluded from the sum. A set of pairs of interest to a car dealer is: $(\circ_1, \oplus_1) =$ (suv, red), $(\circ_2, \oplus_2) =$ (suv, green), $(\circ_3, \oplus_3) =$ (suv, blue), $(\circ_4, \oplus_4) =$ (truck, red) , $(\circ_5, \oplus_5) =$ (truck, yellow).

### *Exercises*

**1.4.1** Define and discuss the sample space of a joint probability of your choice.

**1.4.2** Define and discuss the sample space of the joint probability that a person speaks softly and carries a metaphorical big stick. Evaluate the pertinent probabilities based on your life experiences.

## 1.5 Mutual exclusivity

The joint probability of two mutually exclusive events, denoted by $\circ$ and $\oplus$, is precisely zero,

$$\Pi_{(\circ\,,\,\oplus)} = 0. \tag{1.5.1}$$

Conversely, if the joint probability of two events is zero, the events are mutually exclusive.

For example, with regard to beer, we recall the irrelevance of the listing order and write

$$\Pi_{(\texttt{Great taste}\,,\,\texttt{Less filling})} = \Pi_{(\texttt{Less filling}\,,\,\texttt{Great taste})}. \tag{1.5.2}$$

An experienced beer drinker will insist that the two attributes underwriting this joint probability are not necessarily mutually exclusive, that is, the joint probability is not necessarily zero. Others will disagree.

### 1.5.1 Lunch

Suppose that you invite to lunch a business associate or someone you just met, you don't know much about their eating preferences, and you are afraid to ask. In contemplating you choice of restaurant, you are thinking that:

(*a*) The probability of her being a vegetarian ($\circ$) and having a steak for dinner ($\oplus$) is zero.

(*b*) The probability of her being a vegetarian ($\circ$) and having something else for dinner ($\overline{\oplus}$) can be nonzero.

(*c*) The probability of her being a non-vegetarian ($\overline{\circ}$) and having a steak for dinner ($\oplus$) can be nonzero.

(*d*) The probability of her being a non-vegetarian ($\overline{\circ}$) and having something else for dinner ($\overline{\oplus}$) can be nonzero.

With reference to equation (1.4.9), repeated below for convenience,

$$\Pi_{(\circ\,,\,\oplus)} + \Pi_{(\circ\,,\,\overline{\oplus})} + \Pi_{(\overline{\circ}\,,\,\oplus)} + \Pi_{(\overline{\circ}\,,\,\overline{\oplus})} = 1, \tag{1.5.3}$$

only the first probability on the left-hand side is zero.

### 1.5.2 Steak and eggplant

At the end of a dinner reception that offered a choice between steak and eggplant, a wedding planner asked ten guests four questions:

(*a*) How many are vegetarian and had the steak?
(the answer was $0$.)

(*b*) How many are vegetarian and had the broccoli?
(the answer was $3$.)

(*c*) How many are non-vegetarian and had the steak?
(the answer was $5$.)

(*d*) How many are non-vegetarian and had the eggplant?
(the answer was $2$.)

The wedding planner recorded these answers, divided them by $10$, and regarded the results as joint probabilities for use in ordering food in the next wedding event.

### 1.5.3 Lunch on a cruise boat

Can the wedding data also be used to get an idea of the eating preferences of a group of people on the second deck of a cruise boat? That is, can data from one context be transferred to another context? We will see that this transfer of contexts is a valid point of criticism in the practical implementation of Bayesian analysis.

### 1.5.4 Eyes of the beholder

Mutual exclusivity may lie in the eyes of the beholder: if you reclassify beef as fruit, then the statements "*I ate beef*" and "*I ate a piece of fruit*" are not mutually exclusive. In fact, the concept of mutual exclusivity raises interesting questions in scientific, philosophical, sociological, and spiritual contexts, and is related to the quality of authenticity and the defaulted state of hypocrisy.

Misbehavior, mistakes, wrongdoings, and other ill-advised actions can often be traced to a misclassification of mutually exclusive events as not necessarily so, inadvertently, for the purpose of circumventing or appeasing one's conscience, or deliberately in order to achieve a certain goal. Often, the misclassification is driven by circumstances, peer pressure, and faulty norms.

### *1.5.5 Spiritual and worldly?*

One may ask: *could a spiritual person also be worldly and vice versa*? Are the states of spirituality and social unawareness manifested by the pursuit of wealth, fame, and distinctions mutually exclusive so that their joint probability is zero? The answers depend on who you ask, the person's personal circumstances, internal moral compass and perspectives in life, as well as on the prevailing religious, philosophical, or societal norms.

In some cultures and environments, it is not acceptable for someone to seek recognition or make statements of superiority above and beyond the norm; silent approval and implied respect suffice. A monk in Mount Athos who thinks he is closer to God than an ordinary person commits a serious sin.

#### *Exercise*

**1.5.1** Discuss a pair of states or events that are falsely classified as mutually independent, and a pair of states or events that are falsely classified as non-mutually-independent.

## *1.6 Independence*

Even in the face of randomness, the search for reason and consequence is a human instinct and a spontaneous reflex. In real life, things not always happen for a reason, the just-world fallacy is a cognitive bias, and the notion of karma is a myth. People get away with immoral acts and cruelty all the time. People hurt others for no reason all the time.

Independence describes the absence of correlation and our inability to identify a forward or backward cause-and-effect relation between two events.

There is good reason to believe that the two events involved in the following joint probability are independent:

$$\Pi_{(\texttt{It is Monday}\,,\,\texttt{We expect sunshine})}. \tag{1.6.1}$$

It is reasonable to expect that the two events involved in the following joint probability are *not* independent:

$$\Pi_{(\texttt{It is Sunday}\,,\,\texttt{The bakery is closed,})} \tag{1.6.2}$$

assuming that the bakery is closed on most Sundays.

### 1.6.1 Separability of the joint probability

The joint probability of two independent events is easy to calculate. If two events, $\oplus$ and $\circ$, are independent, and only then, the joint probability is the product of the individual probabilities,

$$\Pi_{(\circ\,,\,\oplus)} = \Pi_{(\circ)} \times \Pi_{(\oplus)}. \tag{1.6.3}$$

We recall that the joint probability refers to a sample space that is the tensor product of the spaces of the individual events and their complements.

Expression (1.6.3) states that, if two events are independent, the joint probability is *separable* with regard to multiplication into individual probabilities, and the joint probability can be factorized.

### 1.6.2 Sunshine

Given the independence of the following two events, we may write

$$\Pi_{(\texttt{It is Wednesday}\,,\,\texttt{We expect sunshine})} = \frac{1}{7} \times \Pi_{(\texttt{We expect sunshine})}, \tag{1.6.4}$$

assuming that today can be any day of the week.

### *1.6.3 Independence of complements*

If an event $\oplus$ is independent of another event $\circ$, the complement $\overline{\oplus}$ is also independent of the complement $\bar{\circ}$, and *vice versa*. In fact, either of $\circ$ and $\bar{\circ}$ is independent of either of $\oplus$ and $\overline{\oplus}$.

If (1.6.3) is true, then all four joint probabilities of the four doublets displayed in (1.4.8) constituting a complete sample space, repeated below for convenience, are separable:

$$(\circ\,,\ \oplus), \qquad (\circ\,,\ \overline{\oplus}), \qquad (\bar{\circ}\,,\ \oplus), \qquad (\bar{\circ}\,,\ \overline{\oplus}). \tag{1.6.5}$$

### *1.6.4 Hidden dependencies*

Sometimes, we will be surprised. Are the two events in the joint probability

$$\Pi_{(\texttt{It will snow in Texas}\ ,\ \texttt{A butterfly is flapping its wings in Alabama})} \tag{1.6.6}$$

independent? Nonlinear dynamical systems theory regarding unpredictability will argue that they may not be independent.

The reason is that weather prediction is an ill-posed mathematical problem where infinitesimal differences in the initial or boundary conditions may have an overwhelming effect on the long-time solution. Stated differently, small perturbations inherent in any system will grow exponentially with a positive Lyapunov exponent.

Other differential equations describing the evolution of mathematical, physical, financial, psychological, and other abstract systems may also end up as ill-posed problems, underscoring the fragility of the underlying processes. A seemingly minor childhood experience may have a huge impact on the psychology of an adult.

### *1.6.5 Unravelling hidden dependencies*

Hidden dependencies are sometimes hard to identify and accept, including dependencies mediated by undiscovered interactions and those based on the notion of karma. Profound theories in science, astronomy, psychology, sociology, and other fields have unravelled hidden or even mysterious dependencies in a variety of contexts.

Hidden dependencies are sometimes framed as intuition, premonition, or psychic ability, and studied under the auspices of the paranormal.

### 1.6.6 Actions have consequences

If we accept that all creatures great and small and all human beings are connected in some way in an all encompassing universe, if we believe that every action evokes a reaction and invites consequences, then we will conclude that no two events are truly independent, even if the events do not pertain to us as individuals but to others.

Words and actions *always* have individual, community-wide, world-wide, or universe-wide consequences manifested on short or long time scales. Nothing ever occurs in vacuum in the space-time domain. Negative consequences will occur when we fabricate our own reality and moral code to serve our wants and needs instead of accepting the true reality and moral code encoded in the human DNA. Exceptions arise.

In a supernatural scenario that verges on science fiction, an external observer keeps track of a collective behavioral index of planet Earth, where each person contributes positive or negative points by his/her actions or inactions.

### 1.6.7 Independence and mutual exclusivity

There is an interesting connection between independence and mutual exclusivity: *two mutually exclusive events cannot be independent.*

For example, if $\circ$ is day and $\oplus$ is night, the joint probability $\Pi_{(\circ\,,\,\oplus)}$ is clearly zero. Independence requires that $\Pi_{(\circ)} \times \Pi_{(\oplus)} = 0$, which cannot be true, unless it is always day or always night in the planetary system where we reside.

Conversely, *two independent events cannot be mutually exclusive.* Mutual exclusivity implies that one event knows something about the other so that it does not occur when the other takes place: a groundhog is nowhere to be found when a fox is around, and an unpleasant neighbor never sees you.

***Exercises***

**1.6.1** Discuss two independent events of your choice.

**1.6.2** Confirm that (1.4.9) is satisfied for two independent events and their complements by factorizing the joint probabilities.

## 1.7 NEXOR probability

The probability of an event, denoted by $\oplus$, or another event, denoted by $\circ$, or both events, is denoted by

$$\Pi_{(\circ\ \mathrm{OR}\ \oplus)} = \Pi_{(\oplus\ \mathrm{OR}\ \circ)}. \tag{1.7.1}$$

Note that the OR is non-exclusive (NEX), that is, it does not exclude the occurrence of both events. By contrast, the exclusive or (XOR) excludes the occurrence of both events.

One example of a NEXOR probability is

$$\Pi_{(\texttt{He will have coffee}\ \mathrm{OR}\ \texttt{She will have tea})}. \tag{1.7.2}$$

This probability accounts for the possibility that he may have coffee, she may have tea, or he may have coffee and she may have tea.

### 1.7.1 Union symbol

In formal mathematics and logic, the NEXOR probability is expressed by the union symbol, $\cup$, so that

$$\Pi(\circ \cup \oplus) \equiv \Pi_{(\circ\ \mathrm{OR}\ \oplus)}. \tag{1.7.3}$$

This notation has its roots in the set theory of mathematics. In this book, we will use the $\mathrm{OR}$ notation.

### 1.7.2 Venn diagram

Two events, $\circ$ and $\oplus$, are represented by two white disks inside a square container in Figure 1.7.1. Let $\bar{\circ}$ be the complement of $\circ$ in their sample space, and $\overline{\oplus}$ be the complement of $\oplus$ in their sample space. The

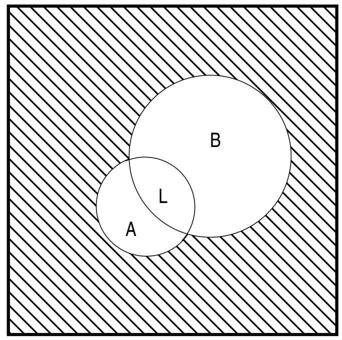



FIGURE 1.7.1 Two events, $\circ$ and $\oplus$, are represented by white disks, $A$ and $B$. The event $\circ\mathrm{OR}\oplus$ is represented by the white area consisting of the union of the disks. The joint probability, $(\circ\,\mathrm{OR}\,\oplus)$, is represented by the lenticular area, $L$.

lenticular white area labeled $L$ in Figure 1.7.1 represents the joint event $(\circ, \oplus)$, the white eclipsed disk shape labeled $A$ represents $(\circ\,,\,\overline{\oplus})$, and the second white eclipsed disk shape labeled $B$ represents $(\overline{\circ}\,,\,\oplus)$.

The XOR event $(\circ\mathrm{OR}\oplus)$ is represented by the white area consisting of the union of $A$, $B$, and $L$. The complement of this event is the shaded shown in Figure 1.7.1.

### 1.7.3 Relation to the joint probability

Now referring to Figure 1.7.1, we find that

$$\Pi_{(\circ\,\mathrm{OR}\,\oplus)} = \Pi_{(\circ\,,\,\oplus)} + \Pi_{(\circ\,,\,\overline{\oplus})} + \Pi_{(\overline{\circ}\,,\,\oplus)}. \tag{1.7.4}$$

Note that this expression is consistent with the transposition of the listing order, as shown in (1.7.1). Rearranging the right-hand side of (1.7.4), we obtain

$$\begin{aligned} \Pi_{(\circ\,\mathrm{OR}\,\oplus)} &= \big(\Pi_{(\circ\,,\,\oplus)} + \Pi_{(\circ\,,\,\overline{\oplus})}\big) \\ &\quad + \big(\Pi_{(\oplus\,,\,\circ)} + \Pi_{(\oplus\,,\,\overline{\circ})}\big) - \Pi_{(\oplus\,,\,\circ)}. \end{aligned} \tag{1.7.5}$$

The sum of the two terms in the first pair of parentheses on the right-hand side is $\Pi_{(\circ)}$, and the sum of the two terms in the second pair of parentheses is $\Pi_{(\oplus)}$. Consequently,

$$\Pi_{(\circ\,\mathrm{OR}\,\oplus)} = \Pi_{(\circ)} + \Pi_{(\oplus)} - \Pi_{(\circ\,,\,\oplus)}. \tag{1.7.6}$$

The role of the last term on the right-hand side of (1.7.6) is to take away one of the two $\Pi_{(\oplus,\circ)}$ involved in the sum of $\Pi_{(\oplus)}$ and $\Pi_{(\circ)}$. Rearranging (1.7.6), we obtain

$$\Pi_{(\circ\,\mathrm{OR}\,\oplus)} + \Pi_{(\circ\,,\,\oplus)} = \Pi_{(\circ)} + \Pi_{(\oplus)}. \tag{1.7.7}$$

The NEXOR probability can be employed as a venue for extracting the joint probability according to (1.7.6).

### *1.7.4 Complement*

Now let $(\overline{\circ\,\mathrm{OR}\,\oplus})$ be the complement of the event $(\circ\mathrm{OR}\oplus)$ represented by the hatched area in Figure 1.7.1. We find that

$$\Pi_{(\overline{\circ\,\mathrm{OR}\,\oplus})} = \Pi_{(\overline{\oplus})} - (\Pi_{(\circ)} - \Pi_{(\circ\,,\,\oplus)}) \tag{1.7.8}$$

and

$$\Pi_{(\overline{\circ\,\mathrm{OR}\,\oplus})} = \Pi_{(\overline{\circ})} - (\Pi_{(\oplus)} - \Pi_{(\oplus\,,\,\overline{\circ})}). \tag{1.7.9}$$

Using either of these expressions and (1.7.6), we may confirm that

$$\Pi_{(\overline{\circ\,\mathrm{OR}\,\oplus})} + \Pi_{(\circ\,\mathrm{OR}\,\oplus)} = 1, \tag{1.7.10}$$

as required.

### *1.7.5 Mutually exclusive events*

In the case of two mutually exclusive events, $\Pi_{(\circ\,,\,\oplus)} = 0$ and thus

$$\Pi_{(\circ\,\mathrm{OR}\,\oplus)} = \Pi_{(\circ)} + \Pi_{(\oplus)}. \tag{1.7.11}$$

The straightforward generalization of this equation to an arbitrary number of $N$ mutually exclusive events, $\circ_i$ constitutes the third axiom of probability theory expressed by

$$\Pi_{(\circ_1\,\mathrm{OR}\,\circ_1\,\mathrm{OR}\,\cdots\,\mathrm{OR}\,\circ_N)} = \Pi_{(\circ_1)} + \Pi_{(\circ_2)} + \cdots + \Pi_{(\circ_N)}. \tag{1.7.12}$$

### *1.7.6 Sample space*

Equation (1.7.6) demonstrates that the NEXOR probability refers to the sample space of the two constituent and joint events. Because the NEXOR probabilities

$$\Pi_{(\circ\,\mathrm{OR}\,\oplus)}, \qquad \Pi_{(\circ\,\mathrm{OR}\,\overline{\oplus})}, \qquad \Pi_{(\overline{\circ}\,\mathrm{OR}\,\oplus)}, \qquad \Pi_{(\overline{\circ}\,\mathrm{OR}\,\overline{\oplus})}, \tag{1.7.13}$$

do not refer to mutually exclusive events, they do not have to add to unity. In fact, using (1.7.6), we find that

$$\Pi_{(\circ\,\mathrm{OR}\,\oplus)} + \Pi_{(\circ\,\mathrm{OR}\,\overline{\oplus})} + \Pi_{(\overline{\circ}\,\mathrm{OR}\,\oplus)} + \Pi_{(\overline{\circ}\,\mathrm{OR}\,\overline{\oplus})} = 3. \tag{1.7.14}$$

Working similarly, we find that

$$\Pi_{(\circ\,\mathrm{OR}\,\oplus)} + \Pi_{(\circ\,\mathrm{OR}\,\overline{\oplus})} = \Pi_{(\circ)} + 1. \tag{1.7.15}$$

### *1.7.7 Expansion into joint probabilities*

Let the following $N$ mutually exclusive events define a sample space,

$$\circ_1, \quad \circ_2, \quad \ldots, \quad \circ_N, \tag{1.7.16}$$

that is, one of the events *has* to occur, so that the sum of the associated probabilities is unity,

$$\sum_{i=1}^{N} \Pi_{(\circ_i)} = 1, \tag{1.7.17}$$

as discussed in Section 1.2. Now let $\circ$ be an arbitrary unrelated event. Using (1.7.6), we find that

$$\Pi_{(\circ\,\mathrm{OR}\,\circ_i)} = \Pi_{(\circ)} + \Pi_{(\circ_i)} - \Pi_{(\circ\,,\,\circ_i)}. \tag{1.7.18}$$

Using the law of total probability, we find that

$$\sum_{i=1}^{N} \Pi_{(\circ\,\mathrm{OR}\,\circ_i)} = (N-1)\,\Pi_{(\circ)} + 1. \tag{1.7.19}$$

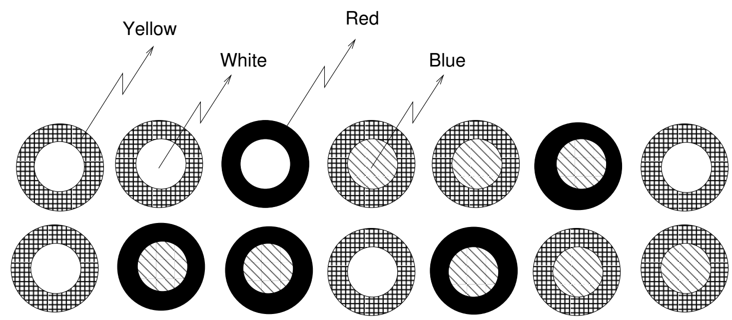



FIGURE 1.8.1 Color schemes of funky trucks with red or yellow exterior and white or blue interior found in the parking lot of *MotoMobil, LLC*.

For $N = 1$, we obtain $\Pi_{(\circ\,\mathrm{OR}\,\circ_i)} = 1$, which is true in light of the stipulation that $\Pi_{(\circ_1)} = 1$.

***Exercises***

**1.7.1** Demonstrate by example that the events $(\circ\,\mathrm{OR}\,\oplus)$ and $(\overline{\circ}\,\mathrm{OR}\,\overline{\oplus})$ are not mutually exclusive.

**1.7.2** What is the counterpart of (1.7.6) for the exclusive $\mathrm{OR}$ (XOR)?

**1.7.3** Write the counterpart of (1.7.6) for three events.

## *1.8 MotoMobil, LLC*

Reputable auto dealer *MotoMobil, LLC* sells funky trucks in four color schemes: *red or yellow exterior* and *white or blue interior*, as shown in Figure 1.8.1. Marketing research has shown that these color combinations are the fastest sellers.

At the end of the day on October 6, 1985, the sales manager recorded the sales figures in the matrix shown in Table 1.8.1. The

| | *Blue interior* | *White interior* | $\to$ Total |
|---|---|---|---|
| *Yellow exterior* | 4 | 5 | 9 |
| *Red exterior* | 2 | 3 | 5 |
| Total $\downarrow$ | 6 | 8 | 14 |

TABLE 1.8.1 Fourteen trucks with four combinations of exterior/interior colors were sold by *MotoMobil, LLC* on October 6, 1989.

total number of trucks sold on that day was 14. Trucks in all color combinations were sold on that day.

### *1.8.1 Marginal probabilities*

Based on these data, the sales manager computed the following probabilities interpreted as fractions of units sold,

$$\Pi_{(\texttt{Yellow exterior})} = \frac{9}{14}, \qquad \Pi_{(\texttt{Red exterior})} = \frac{5}{14} \tag{1.8.1}$$

and

$$\Pi_{(\texttt{Blue interior})} = \frac{6}{14}, \qquad \Pi_{(\texttt{White interior})} = \frac{8}{14}. \tag{1.8.2}$$

Note that the two probabilities shown in (1.8.1) add to unity and the two probabilities shown in (1.8.2) also add to unity. Because these probabilities are at the margins of Table 1.8.1, they are recognized as marginal probabilities.

### *1.8.2 Joint probabilities*

The exterior-interior color combinations define a $2 \times 2$ sample space of joint probabilities as Cartesian product,

$$\text{exterior color} \otimes \text{interior color} \tag{1.8.3}$$

Using the data shown in Table 1.8.1, the sales manager computed the joint probabilities

$$\begin{aligned}\Pi_{(\texttt{Yellow exterior}\,,\,\texttt{Blue interior})} &= \tfrac{4}{14},\\ \Pi_{(\texttt{Yellow exterior}\,,\,\texttt{White interior})} &= \tfrac{5}{14},\\ \Pi_{(\texttt{Red exterior}\,,\,\texttt{Blue interior})} &= \tfrac{2}{14},\\ \Pi_{(\texttt{Red exterior}\,,\,\texttt{White interior})} &= \tfrac{3}{14}.\end{aligned} \tag{1.8.4}$$

Note that these four probabilities add to unity, as required.

### 1.8.3 NEXOR probabilities

Using equation (1.7.6), repeated below for convenience,

$$\Pi_{(\oplus\ \mathrm{OR}\ \circ)} = \Pi_{(\oplus)} + \Pi_{(\circ)} - \Pi_{(\oplus\,,\,\circ)}, \tag{1.8.5}$$

the sales manager computed the non-exclusive (NEXOR) probabilities

$$\Pi_{(\texttt{Yellow ext}\ \mathrm{OR}\ \texttt{Blue int})} = \tfrac{9}{14} + \tfrac{6}{14} - \tfrac{4}{14} = \tfrac{11}{14}, \tag{1.8.6}$$

where $11 = 4 + 5 + 2$,

$$\Pi_{(\texttt{Yellow ext}\ \mathrm{OR}\ \texttt{White int})} = \tfrac{9}{14} + \tfrac{8}{14} - \tfrac{5}{14} = \tfrac{12}{14}, \tag{1.8.7}$$

where $12 = 4 + 5 + 3$,

$$\Pi_{(\texttt{Red ext}\ \mathrm{OR}\ \texttt{Blue int})} = \tfrac{5}{14} + \tfrac{6}{14} - \tfrac{2}{14} = \tfrac{9}{14}, \tag{1.8.8}$$

where $9 = 4 + 2 + 3$, and

$$\Pi_{(\texttt{Red ext}\ \mathrm{OR}\ \texttt{White int})} = \tfrac{5}{14} + \tfrac{8}{14} - \tfrac{3}{14} = \tfrac{10}{14}, \tag{1.8.9}$$

where $10 = 5 + 3 + 2$. We confirm that these four probabilities add to $3.0$.

### 1.8.4 No color preferences

In the event that the customers chose exterior and interior colors independently and with no particular bias (by flipping a fair coin), it would have to be that

$$\Pi_{(\texttt{Yellow exterior}\,,\,\texttt{Blue interior})} = \Pi_{(\texttt{Yellow exterior})} \times \Pi_{(\texttt{Blue interior})}. \tag{1.8.10}$$

| | *Blue* | *White* | → Total |
|---|---|---|---|
| *Yellow* | 3.8571 | 5.1429 | 9 |
| *Red* | 2.1429 | 2.8591 | 5 |
| Total ↓ | 6 | 8 | 14 |

TABLE 1.8.2 Sales matrix in the event that the choice of interior and exterior colors were uncorrelated, that is, car buyers chose color combinations on a whim.

Using the data, the sales manager computed

$$\Pi_{(\texttt{Yellow ext}\,,\,\texttt{Blue int})} = \frac{9}{14} \times \frac{6}{14} = 0.2755, \tag{1.8.11}$$

which differs only somewhat from the actual value $4/14 = 0.2857$ based on the sales record.

In fact, in the absence of color scheme preference, the sales matrix would appear as shown in Table 1.8.2. For example, with reference to the first entry,

$$3.8571 = 14 \times \Pi_{(\texttt{Yellow exterior}\,,\,\texttt{Blue interior})} = \frac{9\times 6}{14}. \tag{1.8.12}$$

Notwistanding the non-integer values, a result a mathematical abstraction, the matrix shown in Table 1.8.2 is nearly identical to that of the actual sales shown in Table 1.8.1. The marginal probabilities shown in the last column and last last row of Table 1.8.2 are the same as those shown in Table 1.8.1.

### *1.8.5 Uncoordinated color choices*

The sales manager explained to the receptionist that, given that the exterior and interior color preferences are nearly uncoordinated, if $k$ trucks were sold on one day, $n$ of these trucks had yellow exterior, and $m$ of these trucks had blue interior, then the sales matrix will have the form shown in Table 1.8.3.

| | *Blue* | *White* | $\rightarrow$ Total |
|---|---|---|---|
| *Yellow* | $\frac{n\,m}{k}$ | $\frac{n\,(k-m)}{k}$ | $n$ |
| *Red* | $\frac{(k-n)\,m}{k}$ | $\frac{(k-n)(k-m)}{k}$ | $k-n$ |
| Total $\downarrow$ | $m$ | $k-m$ | $k$ |

TABLE 1.8.3 Sales matrix for $k$ truck sales with uncorrelated color scheme preferences; $n$ of these trucks have yellow exterior, and $m$ trucks have blue interior.

The sum of the four records inside this matrix is equal to the number of units sold, $k$. Pair of marginal sums also add to $k$.

Using the sales matrix shown in Table 1.8.3, the receptionist derives a formula for the joint probability

$$\Pi_{(\texttt{Yellow exterior}\,,\,\texttt{Blue interior})} = \frac{nm}{k^2}, \tag{1.8.13}$$

and three other similar joint probabilities that arise by diving the entries of Table 1.8.3 by $k$. The sum of the four records inside the joint probability matrix is equal to unity.

### *1.8.6 Fruit flies and plastics*

In her employee application to *MotoMobil, LLC* a few years back, the sales manager stated that she has a college degree in kinesiology with a minor in applied statistics. In the *Notes* section of her application, she stated that she would rather sell funky trucks than study the effect of exercise on the agility of fruit flies. The branch manager appreciated her honesty and smiled.

The receptionist herself has a college degree in mass media communications with a minor in finance. In the *Notes* section of her application, she stated that she would rather help sell funky trucks than analyze the effect of quantitative easing on the price of plastics. The sales manager appreciated her honesty and smiled.

### 1.8.7 Revise your goals

Absolutely no regrets from the sales manager or the receptionist, you have to do what makes you happy in life, not necessarily what you set out to do at an earlier less informative stage of your life.

We will see that the Bayesian framework helps us assess our priorities and revise our goals taking into consideration youthful expectations and our own or others' life experiences.

#### Exercises

**1.8.1** Present the counterparts of Tables 1.8.1 and Tables 1.8.2 in the event of three exterior color choices, yellow, red, and mystic white.

**1.8.2** Let $a$ and $b$ be the entries in the first row of Table 1.8.1 or 1.8.2, and let $c$ and $d$ be the entries in the second row. In both cases,

$$a+b=9, \quad c+d=5, \quad a+d=6, \quad b+d=8. \tag{1.8.14}$$

Demonstrate that these equations do not define uniquely $a$–$d$ and explain the underlying mathematical reason.

## 1.9 Interpolation

The daily total sales of trucks, $k$, yellow trucks, $n$, and trucks with blue interior, $m$, will be changing from day to day, subject to the obvious restrictions $n \le k$, and $m \le k$.

The sales manager of *MotoMobil, LLC* observed that the triplet $(k, n, m)$ evolves systematically (non-randomly) from the beginning to the end of a month. This systematic change could be a reflection of the customers' general mood and balance in their saving accounts after receiving a paycheck or paying their mortgage, rent, and other monthly bills.

### 1.9.1 Interpolation from data taken on the 5th and 23nd of the month

Suppose that a sales record is taken on the fifth day of the month, $d = 5$, and then on the twenty-third day of the month, $d = 23$, where

$d$ stands for *day*. Linear interpolation can be performed to estimate the sales record in any given intermediate day using two methods.

Method A

Linear interpolation amounts to setting

$$\left(\frac{n}{k}\right)_d = \frac{23-d}{23-5} \times \left(\frac{n}{k}\right)_{d=5} + \frac{d-5}{23-5} \times \left(\frac{n}{k}\right)_{d=23} \tag{1.9.1}$$

and

$$\left(\frac{m}{k}\right)_d = \frac{23-d}{23-5} \times \left(\frac{m}{k}\right)_{d=5} + \frac{d-5}{23-5} \times \left(\frac{m}{k}\right)_{d=23} \tag{1.9.2}$$

for $d = 5, \ldots, 23$. We may compute

$$\Pi_{(\texttt{Yellow exterior}\,,\,\texttt{Blue interior})} = \left(\frac{n}{k}\right)_d \times \left(\frac{m}{k}\right)_d \tag{1.9.3}$$

and

$$\Pi_{(\texttt{Yellow exterior}\,,\,\texttt{White interior})} = \left(\frac{n}{k}\right)_d \times \left(1 - \left(\frac{m}{k}\right)_d\right) \tag{1.9.4}$$

for any given day, as well as three additional joint probabilities.

Method B

Alternatively, we employ linear interpolation to write

$$\Pi_{(\texttt{Yellow exterior}\,,\,\texttt{Blue interior})} = \frac{23-d}{23-5} \times \left(\frac{nm}{k^2}\right)_{d=5} + \frac{d-5}{23-5} \times \left(\frac{nm}{k^2}\right)_{d=23}. \tag{1.9.5}$$

Similar expressions can be written for the other three joint probabilities of the corresponding sample space.

The sum of the four interpolated joint probabilities is equal to unity on any given day, as required, for both methods A and B. The four joint probabilities (truck fractions) computed using both methods are reasonably close, as shown in Figure 1.9.1.

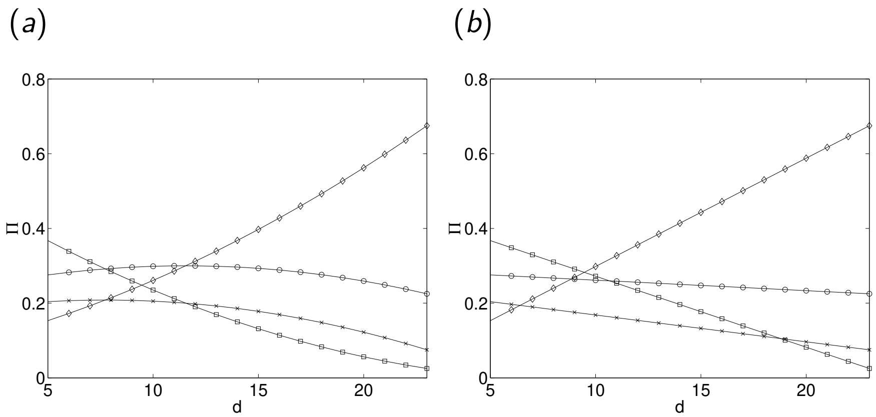



FIGURE 1.9.1 Joint daily probabilities of yellow-blue (•), yellow-white (□), red-blue (⋄), and red-white (×) combinations for $k_5 = 14$, $n_5 = 9$, $m_5 = 6$, and $k_{23} = 20$, $n_{23} = 18$, $m_{23} = 15$, computed by interpolation using $(a)$ method A or $(b)$ method B.

### 1.9.2 *Regional sales meeting*

The sales manager presented her tabulation scheme and observations of customer color preferences to other regional sales managers at a luncheon. At the end of her talk, she was presented with an appreciation plaque and a courtesy check for one thousand dollars.

The sales manager thought she will use the money to replace her washer and dryer with a red stackable pair, and then buy a nice artichoke-colored spring jacket that she had noticed on a rainy day in a local vintage clothing shop.

### 1.9.3 *Zen hoarder*

Late in the afternoon, the sales manager headed home in a yellow truck with blue interior and *Dealer* license plates across familiar farmland. The sun was setting over the hills and she felt good about herself, her customers, and her prospects of becoming a partner in *MotoMobil, LLC.* Her honesty, integrity, fairness, and good disposition toward others was paying off in a way that was consistent with her expectations. Things seemed to make sense in the world.

A customer who buys a truck from the sales manager discovers a *thank-you* note in the trunk, a bag of sea-salted roasted almonds in the glove compartment, and a copy of a cartoon entitled *Zen hoarder* in the spare-tire compartment, meant to release stress at a time of distress. The cartoon depicts a virtually empty room with two small potted plants.

As the sales manager was entering the suburbs, she noticed a familiar grocery store sign on her right. Acting on an impulse, she took the exit, cashed her check, and loaded a shopping cart with basic food staples and some nice treats. She then drove home and left twenty-one bags of groceries in her neighbors yard, making sure she did not make eye contact with the dog.

### 1.9.4 No washer or drier

The day before, the neighbor was given notice that he had lost his job at the assembly line where he worked building eighteen-wheelers for the past twenty-seven years, due to a temporary plant shutdown following an acquisition by an international conglomerate. Everyone, including the plant manager, was stunned that this had happened, especially since the assembly line produced high-quality trucks.

The red washer and dryer and the nice artichoke-colored spring jacket must wait for another year; no big deal in the greater scheme of things.

#### Exercise

**1.9.1** Confirm that formulas (1.9.1) and (1.9.2) are consistent with the data at $d = 5$ and $d = 23$.

## 1.10 Conditional probability

A student may mention to her roommate on a Sunday night: "*if we go to Corleone's for dinner, I will probably have one of the vegetarian* *dishes.*" This is a conditional probability, denoted by

$$\Pi_{(\circ\,|\,\oplus)}, \tag{1.10.1}$$

where:

• The circular symbol ($\circ$) stands for an *outcome or event* (have one of the vegetarian dishes for dinner)

• The crossed circle ($\oplus$) denotes a specified *condition* or set of conditions (go to Panetone's)

The event and the condition are separated by a vertical bar.

The vertical bar can be pronounced as "*given that*", which is definitely awkward in terms of grammar and style, but nevertheless absolutely clear. In Unix programming, the vertical bar is a *pipe.*

### 1.10.1 Cause and effect

Any conditional probability varies in the range $[0, 1]$, where the extreme values $0$ and $1$ indicate absolute certainty, given a condition. Negative values and values higher than unity are allowed only as a figure of speech.

For example, the following conditional probability expresses absolute certainty:

$$\Pi_{(\texttt{The engine will overheat}\,|\,\texttt{The radiator is empty of coolant})} = 1. \tag{1.10.2}$$

A vehicle service adviser will say: "*if the radiator is empty of coolant, the engine will certainly overheat.*" If the adviser is also a statistician, she will say: "*given that the radiator is empty of coolant, the engine will overheat*" (awkward.)

By contrast, the following conditional probability may have any value in the range $[0, 1]$:

$$\Pi_{(\texttt{An ignition coil went bad}\,|\,\texttt{The check engine light is on})}. \tag{1.10.3}$$

The reason is that, if the *check engine* light is on, an ignition coil may be bad, an oxygen sensor may be bad, or another engine component may be malfunctioning.

This example illustrates that a conditional probability carries only partial information on cause and effect.

### 1.10.2 Relation to the joint probability

Common sense suggests that the joint probability, $\Pi_{(\circ\,,\,\oplus)}$, is related to the conditional probability by

$$\Pi_{(\circ\,,\,\oplus)} = \Pi_{(\circ\,|\,\oplus)} \times \Pi_{(\oplus)}. \tag{1.10.4}$$

For example,

$$\Pi_{(\texttt{Out of butter and bread})} = \Pi_{(\texttt{Out of butter}\,|\,\texttt{Out of bread})} \times \Pi_{(\texttt{Out of bread})}. \tag{1.10.5}$$

If the probability of being out of bread is zero, then the probability of being out of butter *and* bread is also zero.

Equation (1.10.4) can be rearranged into

$$\Pi_{(\circ\,|\,\oplus)} = \frac{\Pi_{(\circ\,,\,\oplus)}}{\Pi_{(\oplus)}}, \tag{1.10.6}$$

assuming that $\Pi_{(\oplus)}$ is not zero. We see that, if the events $\circ$ and $\oplus$ are mutually exclusive, $\Pi_{(\circ\,,\,\oplus)} = 0$, the conditional probability is also zero.

In fact, in formal statistics, equation (1.10.6) is regarded as the definition of the conditional probability in terms the joint probability, $\Pi_{(\circ\,,\,\oplus)}$. The ratio provides us with a restricted number of combinations divided by all possible combinations.

Equation (1.10.6) demonstrates that the concept of conditional probability requires the introduction of the joint-probability sample space.

### 1.10.3 Tiny over tiny

The probabilities $\Pi_{(\circ\,,\,\oplus)}$ and $\Pi_{(\oplus)}$ could be tiny, but their ratio may yield a non-tiny conditional probability. For example,

$$\Pi_{(\texttt{Out of butter}\,|\,\texttt{Out of bread})} = \frac{\Pi_{(\texttt{Out of butter and bread})}}{\Pi_{(\texttt{Out of bread})}}. \tag{1.10.7}$$

If $\Pi_{(\texttt{Out of butter and bread})} = 0.0001$ and $\Pi_{(\texttt{Out of bread})} = 0.0002$, then the probability that we are out of butter, given that we are out of bread, is $\frac{1}{2}$. This will happen if we always buy butter in proportion to bread.

### *1.10.4 Non-reversibility*

The condition and the event may *not* be reversed in a conditional probability. For example, the probability

$$\Pi_{(\texttt{Kept the doctor away}\,|\,\texttt{Ate an apple a day})} \tag{1.10.8}$$

is not generally the same as the probability

$$\Pi_{(\texttt{Ate an apple a day}\,|\,\texttt{Kept the doctor away})}. \tag{1.10.9}$$

The second probability could be less than unity since the doctor may have been kept away because of other professional engagements. In fact, these two probabilities are related by Bayes' rule introduced in Chapter 2.

In another example, the probability

$$\Pi_{((\texttt{It will rain today}\,|\,\texttt{It rained yesterday})} \tag{1.10.10}$$

is not generally the same as the probability

$$\Pi_{(\texttt{It rained yesterday}\,|\,\texttt{It will rain today})}, \tag{1.10.11}$$

though they will be the same if the unconditional probabilities of raining yesterday and raining today are the same.

### *1.10.5 It is what it is*

Since an event cannot occur until it occurs,

$$\Pi_{(\circ\,|\,\circ)} = 1, \tag{1.10.12}$$

which is mathematical expression of the profound colloquialism "*it is what it is*", often heard on sports radio stations. An inadvisable-passive aggressive instance of this colloquialism is the statement "*it will get done when it gets done*".

### *1.10.6 Independence*

If the two events $\circ$ and $\oplus$ are independent, the corresponding joint probability can be factorized,

$$\Pi_{(\circ\,,\,\oplus)} = \Pi_{(\circ)} \times \Pi_{(\oplus)}, \tag{1.10.13}$$

as discussed in Section 1.5. Consequently,

$$\Pi_{(\circ\,|\,\oplus)} = \Pi_{(\circ)}. \tag{1.10.14}$$

For example,

$$\Pi_{(\texttt{The car overheats}\,|\,\texttt{I am wearing red socks})} = \Pi_{(\texttt{The car overheats})}. \tag{1.10.15}$$

However, if the engine receives heat from radiant red socks causing it to overheat, the equality does not apply.

### *1.10.7 Quantitative Venn diagram*

A quantitative Venn diagram is shown in Figure 1.10.1($a$) where the outer container enclosing the sample space is a unit rectangle. An event $\circ$ is represented by the disk whose area is $\Pi_{(\circ)}$. Another event $\oplus$ is represented by the inner rectangle whose area is $\Pi_{(\oplus)}$.

The four patterns represent the four constituents of the joint probability sample space, $(\circ, \oplus)$, $(\circ, \overline{\oplus})$, $(\bar{\circ}, \oplus)$, and $(\bar{\circ}, \overline{\oplus})$. The joint probability $\Pi_{(\circ,\oplus)}$ is the area of the disk-rectangle overlap.

### *1.10.8 Losing love is like a window in your heart (Paul Simon)*

The condition $\oplus$ in the conditional probability $\Pi_{(\circ\,|\,\oplus)}$ illuminates only part of the sample space. The remainder of the sample space is not visible, as shown in Figure 1.10.1($b$). Consequently, only part of the whole disk representing $\circ$ can be seen. If the events $\circ$ and $\oplus$ are independent, the condition $\oplus$ does not illuminate any portion of $\circ$.

The conditional probability $\Pi_{(\circ\,|\,\oplus)}$ is the ratio of the visible area of the disk, $\Pi_{(\circ\,,\,\oplus)}$, to the area enclosed by the inner rectangle, $\Pi_{(\oplus)}$.

Conversely, the condition $\circ$ in the conditional probability $\Pi_{(\oplus\,|\,\circ)}$ illuminates only part of the sample space. The remainder of the sample

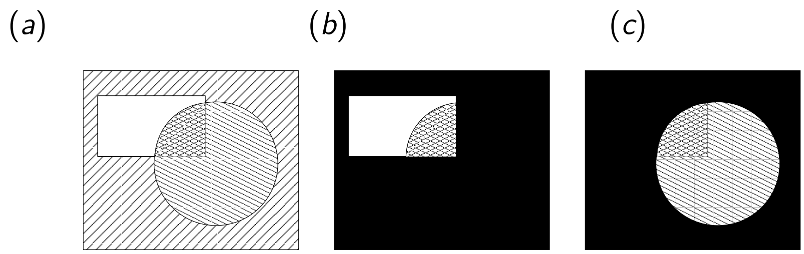



FIGURE 1.10.1 (*a*) Venn diagram of two events represented by the disk and the square in the joint-probability sample space. (*b*) The condition $\oplus$ in the conditional probability $\Pi_{(\circ\,|\,\oplus)}$ illuminates only part of the Venn diagram. (*c*) The condition $\circ$ in the conditional probability $\Pi_{(\oplus\,|\,\circ)}$ illuminates only part of the Venn diagram.

space is not visible, as shown in Figure 1.10.1(*c*). Consequently, only part of the inner rectangle representing $\oplus$ can be seen. If the events $\circ$ and $\oplus$ are independent, the condition $\circ$ does not illuminate any portion of $\oplus$.

The conditional probability $\Pi_{(\oplus\,|\,\circ)}$ is the ratio of the visible area of the inner rectangle, $\Pi_{(\oplus\,,\,\circ)}$, to the area of the disk, $\Pi_{(\circ)}$.

### *1.10.9 Relation to unconditional probabilities*

The interpretation of the condition $\oplus$ as a spot light suggests that the conditional probability $\Pi_{(\circ\,|\,\oplus)}$ can be less than, equal to, or greater than the unconditional probability, $\Pi_{(\circ)}$. For example,

$$\Pi_{(\texttt{Has blue eyes}\,|\,\texttt{Mother has blue eyes})} > \Pi_{(\texttt{Has blue eyes})} \qquad (1.10.16)$$

on account of heredity.

Using the law of total probability, we find that

$$\Pi_{(\circ\,|\,\oplus)} = \frac{\Pi_{(\circ\,,\,\oplus)}}{\Pi_{(\oplus)}} = \frac{\Pi_{(\circ)} - \Pi_{(\circ\,,\,\overline{\oplus})}}{\Pi_{(\oplus)}} \leq \frac{\Pi_{(\circ)}}{\Pi_{(\oplus)}}. \qquad (1.10.17)$$

The equality applies when $\Pi_{(\circ\,,\,\overline{\oplus})} = 0$

### *1.10.10 Complement of an event*

The illustrations in Figure 1.10.1 confirm that the conditional probability of an event and its complement add to unity,

$$\Pi_{(\circ\,|\,\oplus)} + \Pi_{(\overline{\circ}\,|\,\oplus)} = 1, \tag{1.10.18}$$

where we recall that an overbar denotes the complement. The presence of a condition does not alter this fundamental probability law. Further conditions can be attached.

### *1.10.11 Law of total probability*

The law of total probability expressed by (1.4.12), stating that $\Pi_{(\circ)} = \Pi_{(\circ\,,\,\oplus)} + \Pi_{(\circ\,,\,\overline{\oplus})}$, takes the form

$$\Pi_{(\circ)} = \Pi_{(\circ\,|\,\oplus)} \times \Pi_{(\oplus)} + \Pi_{(\circ\,|\,\overline{\oplus})} \times \Pi_{(\overline{\oplus})}. \tag{1.10.19}$$

Setting $\Pi_{(\overline{\oplus})} = 1 - \Pi_{(\oplus)}$, we obtain

$$\Pi_{(\circ)} = \Pi_{(\circ\,|\,\oplus)} \times \Pi_{(\oplus)} + \Pi_{(\circ\,|\,\overline{\oplus})} \times (1 - \Pi_{(\oplus)}). \tag{1.10.20}$$

Rearranging, we obtain

$$\Pi_{(\circ\,|\,\overline{\oplus})} = \frac{\Pi_{(\circ)} - \Pi_{(\circ\,|\,\oplus)} \times \Pi_{(\oplus)}}{1 - \Pi_{(\oplus)}}. \tag{1.10.21}$$

The denominator is always less than or equal to the numerator. A similar expression can be written for $\Pi_{(\circ\,|\,\oplus)}$.

### *1.10.12 NEXOR probability*

Equation (1.7.6) involving the NEXOR probability is generalized into

$$\Pi_{(\circ\,\mathrm{OR}\,\oplus\,|\,\square)} = \Pi_{(\circ\,|\,\square)} + \Pi_{(\oplus\,|\,\square)} - \Pi_{(\circ\,,\,\oplus\,|\,\square)}. \tag{1.10.22}$$

Additional conditions can be attached.

### *1.10.13 MotoMobil LLC*

Consider the matrix of daily truck sales shown in Table 1.7.1, repeated for convenience in Table 1.10.1. In this case, probability is interpreted as fraction of units sold.

| | *Blue interior* | *White interior* | → Total |
|---|---|---|---|
| *Yellow exterior* | 4 | 5 | 9 |
| *Red exterior* | 2 | 3 | 5 |
| Total ↓ | 6 | 8 | 14 |

TABLE 1.10.1 Fourteen trucks with four combinations of exterior/interior colors were sold by *MotoMobil, LLC* on October 6, 1989.

Two sets of conditional probabilities can be calculated,

$$\begin{aligned} &\Pi_{(\texttt{Blue interior}\,|\,\texttt{Yellow exterior})} = \frac{4}{9}, \\ &\Pi_{(\texttt{White interior}\,|\,\texttt{Yellow exterior})} = \frac{5}{9}, \\ &\Pi_{(\texttt{Blue interior}\,|\,\texttt{Red exterior})} = \frac{2}{5}, \\ &\Pi_{(\texttt{White interior}\,|\,\texttt{Red exterior})} = \frac{3}{5} \end{aligned} \qquad (1.10.23)$$

and

$$\begin{aligned} &\Pi_{(\texttt{Yellow exterior}\,|\,\texttt{Blue interior})} = \frac{4}{6}, \\ &\Pi_{(\texttt{Red exterior}\,|\,\texttt{Blue interior})} = \frac{2}{6}, \\ &\Pi_{(\texttt{Yellow exterior}\,|\,\texttt{White interior})} = \frac{5}{8}, \\ &\Pi_{(\texttt{Red exterior}\,|\,\texttt{White interior})} = \frac{3}{8}, \end{aligned} \qquad (1.10.24)$$

Note that the two probabilities in each row add to unity. The inequality (1.10.17) is satisfied. Referring to Table 1.10.1, we confirm that

$$\begin{aligned} &\Pi_{(\texttt{Yellow exterior}\,,\,\texttt{Blue interior})} \\ &\qquad = \Pi_{(\texttt{Yellow exterior}\,|\,\texttt{Blue interior})} \times \Pi_{(\texttt{Blue interior})} \\ &\qquad = \frac{4}{6} \times \frac{6}{14} = \frac{4}{14} \end{aligned} \qquad (1.10.25)$$

and their counterparts corresponding to different color combinations. Note that the two conditional probabilities

$$\Pi_{(\texttt{Blue interior} \,|\, \texttt{Yellow exterior})},$$
$$\Pi_{(\texttt{Yellow exterior} \,|\, \texttt{Blue interior})} \tag{1.10.26}$$

are not the same.

***Exercises***

**1.10.1** State in plain English (or any language of your choice) the narrative expressed by a conditional probability of your choice.

**1.10.2** Discuss a situation with a tiny-over-tiny occurrence.

## 1.11 Properties of the conditional probability

Because of its importance in Bayesian analysis, the concept of conditional probability warrants further consideration.

### 1.11.1 NEXOR condition

The conditional probability

$$\Pi_{(\circ \,|\, (\circ \,\mathrm{OR}\, \oplus))} \tag{1.11.1}$$

may have any value in the range $[0, 1]$. For example, a truck driver may assess that

$$\Pi_{(\texttt{Tire 1 is flat} \,|\, \texttt{Tire 1 is flat} \,\mathrm{OR}\, \texttt{tire 3 is flat})} = \frac{1}{2}. \tag{1.11.2}$$

We recall that "*tire 1 or tire 3 is flat*" accounts for the possibility that tire 1 is flat, tire 3 is flat, or both tires are flat.

Common sense suggests that

$$\Pi_{(\circ \,|\, (\circ \,\mathrm{OR}\, \oplus))} \geq \Pi_{\circ}, \qquad \Pi_{(\oplus \,|\, (\circ \,\mathrm{OR}\, \oplus))} \geq \Pi_{\oplus}. \tag{1.11.3}$$

In this context, $\Pi_{\circ}$ and $\Pi_{\oplus}$ are interpreted as contrived conditional probabilities, given nothing.

### 1.11.2 Law of total probability

Let the following $N$ mutually exclusive events define a sample space,

$$\circ_1, \quad \circ_2, \quad \ldots, \quad \circ_N, \tag{1.11.4}$$

that is, one of the events *has* to occur, so that the sum of the associated probabilities is unity. Normalization requires that

$$\sum_{i=1}^{N} \Pi_{(\circ_i \,|\, \oplus)} = 1 \tag{1.11.5}$$

for any event or condition, $\oplus$.

The probability of another arbitrary event, denoted by $\circ$, is given by the sum of $N$ joint probabilities,

$$\Pi_{(\circ)} = \sum_{i=1}^{N} \Pi_{(\circ\,,\,\circ_i)}, \tag{1.11.6}$$

as shown in (1.4.21). The joint probabilities in the sum can be expressed in terms of conditional probabilities using the familiar formula

$$\Pi_{(\circ\,,\,\circ_i)} = \Pi_{(\circ\,|\,\circ_i)} \times \Pi_{(\circ_i)}, \tag{1.11.7}$$

yielding an important expansion,

$$\Pi_{(\circ)} = \sum_{i=1}^{N} \Pi_{(\circ\,|\,\circ_i)} \times \Pi_{(\circ_i)}. \tag{1.11.8}$$

The expression in terms of the sum is a common step in Bayesian analysis where the individual probabilities, $\Pi_{(\circ_i)}$, are known or otherwise assumed, as discussed in Chapter 2.

If all conditional probabilities $\Pi_{(\circ\,|\,\circ_i)}$ are equal to $1.0$, then $\Pi_{(\circ)}$ will also be $1.0$. If all conditional probabilities are equal to $0$, then $\Pi_{(\circ)}$ will also be $0$.

### *1.11.3 Continuous range of events*

The $N$ events $\circ_i$ can be labeled by a parameter $\theta_i$ that varies monotonically from the first event, $i = 1$, to the last event, $i = N$, by equal intervals, $\Delta\theta$. The value $\theta_i$ implies event $\circ_i$ and *vice versa*.

Expansion (1.11.8) may then be written as

$$\Pi_{(\circ)} = \Delta\theta \sum_{i=1}^{N} \Pi_{(\circ\,|\,\theta_i)}\, \phi(\theta_i), \tag{1.11.9}$$

where

$$\phi(\theta_i) \equiv \frac{1}{\Delta\theta}\, \Pi_{(\circ_i)}. \tag{1.11.10}$$

When $N$ is large, the right-hand side of (1.11.9) can be approximated with an integral, yielding

$$\Pi_{(\circ)} = \int \Pi_{(\circ\,|\,\theta)}\, \phi(\theta)\, \mathrm{d}\theta \tag{1.11.11}$$

where the probability density function $\phi(\theta)$ is defined by interpolation in terms of the discrete values $\phi(\theta_i)$ shown in (1.11.10), subject to the normalization condition

$$\int \phi(\theta)\, \mathrm{d}\theta = 1. \tag{1.11.12}$$

By definition,

$$\Pi_{(\circ)} = \int \phi(\circ\,,\ \theta)\, \mathrm{d}\theta, \tag{1.11.13}$$

where the conditional probability density function

$$\phi(\circ\,,\ \theta) \equiv \Pi_{(\circ\,|\,\theta)}\, \phi(\theta) \tag{1.11.14}$$

is defined likewise by interpolation in terms of the discrete values

$$\phi(\circ\,,\ \theta_i) \equiv \frac{1}{\Delta\theta}\, \Pi_{(\circ\,,\,\circ_i)} \times \Pi(\circ_i). \tag{1.11.15}$$

The integrations over $\theta$ are performed over an appropriate integration domain.

### 1.11.4 Tomatoes

The event $\circ$ can be that a tomato plant survives the night, $\theta$ can be the overnight temperature, $\phi(\theta)$ can be the probability density function of the overnight temperature, and $\Pi_{(\circ\,|\,\theta)}$ is the probability that the tomato plant survives at temperature $\theta$. If the temperature is expected to vary with equal probability between $\theta_{\rm min}$ and $\theta_{\rm max}$, then

$$\phi(\theta) = \frac{1}{\theta_{\rm max} - \theta_{\rm max}} \tag{1.11.16}$$

is a flat distribution.

### 1.11.5 Attached conditions

Everything can be subject to a universal pervasive condition, such as *subject to availability of funds*. This pervasive condition can be attached to any conditional or unconditional probability. The condition of a condition is a joint condition,

$$\Pi_{(\circ\,|\,(\oplus,\square))}, \tag{1.11.17}$$

where $\square$ is a pervasive condition. An example is *a student will receive a fellowship* (event $\circ$), if she has a high enough Grade Point Average (event $\oplus$), subject to availability of funds (event $\square$).

Now we may write

$$\Pi_{(\circ\,|\,(\oplus,\square))} = \frac{\Pi_{(\circ,(\oplus,\square))}}{\Pi_{(\oplus,\square)}}. \tag{1.11.18}$$

Setting

$$\Pi_{(\circ,(\oplus,\square))} = \Pi_{(\circ,\oplus,\square)}, \qquad \Pi_{(\oplus,\square)} = \Pi_{(\oplus\,|\,\square)}\,\Pi_{(\square)} \tag{1.11.19}$$

respectively, in the numerator and denominator, we obtain

$$\Pi_{(\circ\,|\,((\oplus,\square))} = \frac{\Pi_{(\circ,\oplus,\square)}}{\Pi_{(\oplus\,|\,\square)}\,\Pi_{(\square)}}. \tag{1.11.20}$$

Generalizing by mathematical induction, we obtain

$$\Pi_{(\bigcirc\,|\,((A,\ldots,Z))} = \frac{\Pi_{(\bigcirc,A,B,\ldots,X,Y,Z)}}{\Pi_{(A\,|\,(B,\cdots,Z))}\cdots\Pi_{(X\,|\,(Y,Z))}\Pi_{(Y\,|\,Z)}\,\Pi_{(Z)}}, \qquad (1.11.21)$$

subject to straightforward notation.

### 1.11.6 Implied conditions

It can be argued that implied conditions should always be attached to probabilities to define the appropriate context beyond doubt and prevent the occurrence of paradoxes and oxymora. While this is strictly true, it does make for cumbersome notation. When faced with an oxymoron or contradiction, we should look for a missed implied condition defining context.

Brilliant crime novels conclude with a twist due to a carefully concealed implied condition, the most common condition being that the main character, the maid or an incidental person is a sociopath. After reading enough novels by a certain author, you start recognizing implied conditions that the author favors in his or her discourse.

### 1.11.7 Law of total probability

We may introduce an arbitrary event, $\oplus$, and a complete set of $N$ mutually exclusive events, $\circ_i$, and write the law of total probability

$$\Pi_{(\bigcirc\,|\,\oplus)} = \sum_{i=1}^{N} \Pi_{((\bigcirc\,,\,\circ_i)\,|\,\oplus)} = \frac{1}{\Pi_{(\oplus)}} \sum_{i=1}^{N} \Pi_{((\bigcirc\,,\,\circ_i)\,,\,\oplus)}, \qquad (1.11.22)$$

which can be written as

$$\Pi_{(\bigcirc\,|\,\oplus)} = \frac{1}{\Pi_{(\oplus)}} \sum_{i=1}^{N} \Pi_{(\oplus\,|\,(\bigcirc\,,\,\circ_i))}\,\Pi_{(\bigcirc\,,\,\circ_i)} \qquad (1.11.23)$$

or

$$\Pi_{(\bigcirc\,|\,\oplus)} = \frac{\Pi_{(\bigcirc)}}{\Pi_{(\oplus)}} \sum_{i=1}^{N} \Pi_{(\oplus\,|\,(\bigcirc\,,\,\circ_i))}\,\Pi_{(\circ_i\,|\,\bigcirc)}. \qquad (1.11.24)$$

The sum on the right-hand side of (1.11.23) is the conditional probability

$$\Pi_{(\oplus\,|\,\circ)} = \sum_{i=1}^{N} \Pi_{(\oplus\,|\,(\circ_i\,,\circ))}\, \Pi_{(\circ_i\,|\,\circ)}. \tag{1.11.25}$$

The preceding four equations provide us with a glimpse of how joint and conditional probabilities can be combined in different ways, leading us to the Bayes rule formalized in Chapter 2.

### *1.11.8 Continuous range of events*

Working as previously in this section for a continuous range of events, we derive the counterpart of equation (1.11.24),

$$\Pi_{(\circ\,|\,\oplus)} = \frac{\Pi_{(\circ)}}{\Pi_{(\oplus)}} \int \Pi_{(\oplus\,|\,(\circ\,,\,\theta))}\, \phi(\theta\,|\,\circ)\, \mathrm{d}\theta, \tag{1.11.26}$$

where $\phi(\theta\,|\,\circ)$ is a probability density function defined by interpolation in terms of discrete values,

$$\phi(\theta_i\,|\,\circ) \equiv \frac{1}{\Delta\theta}\, \Pi_{(\theta_i\,|\,\circ)}, \tag{1.11.27}$$

and the integration is performed over an appropriate integration domain. Based on equation (1.11.26), we conclude that

$$\Pi_{(\oplus\,|\,\circ)} = \int \Pi_{(\oplus\,|\,(\circ\,,\,\theta))}\, \phi(\theta\,|\,\circ)\, \mathrm{d}\theta. \tag{1.11.28}$$

***Exercises***

**1.11.1** You roll two fair dice that land randomly any one of their six faces. Compute the conditional probability

$$\Pi_{(\texttt{First dice shows 2}\,|\,\texttt{First dice shows 2 OR second dice shows 4})}. \tag{1.11.29}$$

**1.11.2** Discuss a situation where $\Pi_{(\circ\,|\,\oplus)} > \Pi_{(\circ)}$ and another situation where $\Pi_{(\circ\,|\,\oplus)} < \Pi_{(\circ)}$.

## 1.12 Backpacks and DNA

At the Thursday faculty meeting, the Head of the Department of Kinesiology asked $m$ faculty members to join her in a weekend hiking trip up the hill behind their college town. A retreat will be held at the top of the hill around a camp fire where important announcements will be made, the excellent state of the department will be affirmed, and other vital departmental matters will be discussed.

### 1.12.1 Limited selection of colors

The faculty members have not exercised in years and secretly run to the store to buy designer backpacks, comfortable sleeping bags, and other supplies. Some entered the store wearing Groucho Marx glasses and mustaches so that they could not be recognized.

The store carries backpacks in $n$ colors. In their rush to prepare for the retreat, faculty members grab backpacks with random colors. Professor S. T. Retch picked up a green backpack and wants to know the probability that no one else shows up with this color,

$$\Pi_{(\texttt{Nobody else has green}\,|\,\texttt{Retch has green})}. \tag{1.12.1}$$

Being statistically challenged, the professor asks his friend, a professor of applied statistics, to provide insights.

### 1.12.2 Expert opinion

The professor of statistics explains that the probability that another faculty member chooses randomly and without discrimination any color other than green is

$$\frac{n-1}{n}. \tag{1.12.2}$$

For example, when $n = 2$ (red or green), the probability of choosing red is $1/2$. The joint probability that two other faculty members choose randomly and independently any color other than green is

$$\frac{n-1}{n} \times \frac{n-1}{n}. \tag{1.12.3}$$

The joint probability that all other $m-1$ faculty members choose randomly and independently any color other than green is

$$\Pi_{(\texttt{Nobody else has green}\,|\,\texttt{Retch has green})} = \Big(\frac{n-1}{n}\Big)^{m-1}. \qquad (1.12.4)$$

If the store carries only green backpacks, $n = 1$, then this probability is zero. If the store carries a large selection of colors, $n$ is high and this probability is nearly unity.

Consequently, the probability that at least one other faculty member chose green is given by the complementary probability

$$\Pi_{(\texttt{At least one other person has green}\,|\,\texttt{Retch has green})} = 1 - \Big(\frac{n-1}{n}\Big)^{m-1}. \qquad (1.12.5)$$

For $n = 3$ colors and $m = 10$ faculty members, this probability is $0.974$. It appears that professor S. T. Retch cannot make a fashion statement and becomes so distressed that his cat develops an ulcer.

### *1.12.3 The importance of midges*

The retreat turned out to be a disaster, essentially because of intense annoyance caused by midges. At the next meeting of the department Heads with the Dean, the Head of kinesiology offered to her colleagues an important piece of advice: *do not underestimate the importance of literal, figurative, or metaphorical midges.*

### *1.12.4 DNA matching*

A DNA database consists of $n$ unique DNA sequences (backpack colors). Building the database was terminated when further DNA sequencing generated an entry that already existed in the database due to imperfect detection. Since half the DNA of a human is the same as that of a banana, some people were falsely identified as bananas.

Consider a population of $p$ persons, where $p > n$. The DNA of a random person in this population matches randomly a DNA sequence in the database. The DNA of a person identified as Mr. Foo matches one such sequence (green backpack).

By repeating the preceding analysis for the delusional retreat, we find that the conditional probability that at least one other person matches Mr. Foo's DNA is given by

$$\Pi_{(\texttt{At least one other person matches} \,|\, \texttt{Mr. Foo matches})} = 1 - \Big(\frac{n-1}{n}\Big)^{m-1}. \tag{1.12.6}$$

For $n = 50,000$ DNA sequences in the database and a population of $m = 200,000$ people, we find the conditional probability

$$1 - \Big(\frac{50,000-1}{50,000}\Big)^{200,000-1} = 0.9817, \tag{1.12.7}$$

which is convincingly high. Mr. Foo is not necessarily guilty of any crime.

The important and somewhat unexpected conclusion is that, even though the fraction on the left-hand side of (1.12.7) is nearly unity, the power of this fraction to a large number is much less than unity.

### *Exercises*

**1.12.1** Explain why (*a*)

$$\Pi_{(\texttt{Everyone else has green} \,|\, \texttt{Retch has green})} = \frac{1}{n^{m-1}}, \tag{1.12.8}$$

and (*b*)

$$\Pi_{(\texttt{Everyone has green})} = \frac{1}{n^{m}}. \tag{1.12.9}$$

**1.12.2** Discuss technical reasons as to why the DNAs of two different people could be falsely regarded as being the same.

**1.12.3** Prepare and discuss a graph of the conditional probability shown in (1.12.6) for $n = 100,000$ DNA sequences and varying populations $m$.

# Chapter 2

# Bayes' rule

Bayes' rule relates two conditional probabilities to two unconditional probabilities involving two events. The unconditional probabilities can be chosen or assumed, one of the conditional probabilities involving data can also be chosen or assumed, and the second conditional probability constitutes an educated estimate.

The application of Bayes' rule to make predictions, draw conclusions, and infer probabilities and parameters values is best explained by case studies and examples drawn from various settings and disciplines. Scientific, engineering, forensic, financial, philosophical, cognitive, sociological, behavioral, metaphysical, and other applications and accompanying interpretations of Bayes' rule have been developed.

Bayes' rule provides us with a venue for making educated guesses, making the most out of our experiences, and easing our anxiety regarding the randomness and unpredictability of the real world. The overall deduction process is described as Bayesian analysis.

## 2.1 Reciprocity and Bayes' rule

Equation (1.10.4) states that the joint probability of two events, $\circ$ and $\oplus$, can be expressed in terms of a conditional probability as

$$\Pi_{(\circ\,,\,\oplus)} = \Pi_{(\circ\,|\,\oplus)} \times \Pi_{(\oplus)}. \tag{2.1.1}$$

For example,

$$\Pi_{(\texttt{hungry and thirsty})} = \Pi_{(\texttt{hungry}\,|\,\texttt{thirsty})} \times \Pi_{(\texttt{thirsty})}. \tag{2.1.2}$$

The first probability on the right-hand side is the conditional probability of being hungry when we are thirsty.

### 2.1.1 Reversal of order

Reversing the role of the two events, we obtain

$$\Pi_{(\oplus\,,\,\circ)} = \Pi_{(\oplus\,|\,\circ)} \times \Pi_{(\circ)}. \tag{2.1.3}$$

For example,

$$\Pi_{(\texttt{thirsty and hungry})} = \Pi_{(\texttt{thirsty}\,|\,\texttt{hungry})} \times \Pi_{(\texttt{hungry})}. \tag{2.1.4}$$

The first probability on the right-hand side is the conditional probability of being thirsty given that we are hungry .

We recall that the two conditional probabilities,

$$\Pi_{(\texttt{hungry}\,|\,\texttt{thirsty})} \qquad \Pi_{(\texttt{thirsty}\,|\,\texttt{hungry})} \tag{2.1.5}$$

are generally different.

### 2.1.2 Getting from A to B in two ways

Combining (2.1.1) and (2.1.3), we obtain a reciprocal relation involving two unconditional and two conditional probabilities,

$$\Pi_{(\circ\,|\,\oplus)} \times \Pi_{(\oplus)} = \Pi_{(\oplus\,|\,\circ)} \times \Pi_{(\circ)}, \tag{2.1.6}$$

which can be rearranged into Bayes' rule,

$$\Pi_{(\circ\,|\,\oplus)} = \frac{\Pi_{(\oplus\,|\,\circ)}}{\Pi_{(\oplus)}} \times \Pi_{(\circ)}. \tag{2.1.7}$$

The simplicity of this rule is in sharp contrast with its profound significance in a broad range of applications.

For example, using Bayes' rule we write

$$\Pi_{(\texttt{Power outage}\,|\,\texttt{Storm})} = \frac{\Pi_{(\texttt{Storm}\,|\,\texttt{Power outage})}}{\Pi_{(\texttt{Storm})}} \times \Pi_{(\texttt{Power outage})}. \tag{2.1.8}$$

Since the conditional probability is less than unity by convention, the right-hand side of (2.1.7) must be no greater than unity, which is true since

$$\Pi_{(\oplus\,|\,\circ)} \equiv \frac{\Pi_{(\oplus\,,\,\circ)}}{\Pi_{(\circ)}} \leq \frac{\Pi_{(\oplus)}}{\Pi_{(\circ)}}. \tag{2.1.9}$$

Estimates or suggestions that do not conform with this constraint should be dismissed as invalid.

### 2.1.3 Rule of relative decrease

Rearranging (2.1.7), we obtain an alternative version of Bayes' equation,

$$\frac{\Pi_{(\circ)} - \Pi_{(\circ\,|\,\oplus)}}{\Pi_{(\circ)}} = \frac{\Pi_{(\oplus)} - \Pi_{(\oplus\,|\,\circ)}}{\Pi_{(\oplus)}}. \tag{2.1.10}$$

The differences in the numerator express the decrease in probability due to an imposed condition. The fractions express the *relative* or fractional change in probability due to an imposed condition. Bayes rule states that the two relative changes are the same.

### 2.1.4 An event and its complement

Bayes' rule (2.1.7) for an event, $\circ$, and its complement, $\bar{\circ}$, read

$$\Pi_{(\circ\,|\,\oplus)} = \frac{\Pi_{(\oplus\,|\,\circ)}}{\Pi_{(\oplus)}} \times \Pi_{(\circ)}, \tag{2.1.11}$$

and

$$\Pi_{(\bar{\circ}\,|\,\oplus)} = \frac{\Pi_{(\oplus\,|\,\bar{\circ})}}{\Pi_{(\oplus)}} \times \Pi_{(\bar{\circ})}. \tag{2.1.12}$$

By the second axiom of probability theory,

$$\Pi_{(\circ)} + \Pi_{(\bar{\circ})} = 1 \tag{2.1.13}$$

and

$$\Pi_{(\circ\,|\,\oplus)} + \Pi_{(\bar{\circ}\,|\,\oplus)} = 1. \tag{2.1.14}$$

Since by the law of total probability

$$\Pi_{(\oplus)} = \Pi_{(\oplus\,|\,\circ)} \times \Pi_{(\circ)} + \Pi_{(\oplus\,|\,\bar{\circ})} \times \Pi_{(\bar{\circ})}, \tag{2.1.15}$$

Bayes rule provides us with the equations

$$\Pi_{(\circ\,|\,\oplus)} = \frac{\Pi_{(\oplus\,|\,\circ)}}{\Pi_{(\oplus\,|\,\circ)} \times \Pi_{(\circ)} + \Pi_{(\oplus\,|\,\bar{\circ})} \times (1 - \Pi_{(\circ)})} \times \Pi_{(\circ)} \qquad (2.1.16)$$

and

$$\Pi_{(\bar{\circ}\,|\,\oplus)} = \frac{\Pi_{(\oplus\,|\,\bar{\circ})}}{\Pi_{(\oplus\,|\,\circ)} \times \Pi_{(\circ)} + \Pi_{(\oplus\,|\,\bar{\circ})} \times (1 - \Pi_{(\circ)})} \times \Pi_{(\bar{\circ})}, \qquad (2.1.17)$$

where $\Pi_{(\bar{\circ})} = 1 - \Pi_{(\circ)}$.

### 2.1.5 Which form to use

Equation (2.1.7) is used when $\Pi_{(\oplus)}$ is available, whereas equation (2.1.16) is used when $\Pi_{(\oplus\,|\,\bar{\circ})}$ is available. In both cases, $\Pi_{(\oplus\,|\,\circ)}$ is known, specified, measured, or otherwise assumed.

To deduce $\Pi_{(\oplus\,|\,\bar{\circ})}$ from $\Pi_{(\oplus)}$, we substitute into (2.1.15) the expression $\Pi_{(\bar{\circ})} = 1 - \Pi_{(\circ)}$, and obtain

$$\Pi_{(\oplus)} = \Pi_{(\oplus\,|\,\circ)} \times \Pi_{(\circ)} + \Pi_{(\oplus\,|\,\bar{\circ})} \times (1 - \Pi_{(\circ)}). \qquad (2.1.18)$$

Rearranging, find that

$$\Pi_{(\oplus\,|\,\bar{\circ})} = \frac{\Pi_{(\oplus)} - \Pi_{(\oplus\,|\,\circ)} \times \Pi_{(\circ)}}{1 - \Pi_{(\circ)}}. \qquad (2.1.19)$$

The denominator is always less than or equal to the numerator. A similar expression can be written for $\Pi_{(\circ\,|\,\bar{\oplus})}$.

### 2.1.6 Zero and unity conditional probabilities

Assume that $\Pi_{(\bar{\circ}\,|\,\oplus)} = 1$, in which case $\Pi_{(\circ\,|\,\oplus)} = 0$. Equation (2.1.6) requires that $\Pi_{(\oplus\,|\,\circ)} = 0$, unless $\Pi_{(\circ)} = 0$. We conclude that

$$\Pi_{(\bar{\circ}\,|\,\oplus)} = 1 \rightarrow \qquad \Pi_{(\circ\,|\,\oplus)} = 0, \quad \Pi_{(\oplus\,|\,\circ)} = 0, \quad \Pi_{(\bar{\oplus}\,|\,\circ)} = 1, \qquad (2.1.20)$$

where the arrow stands for *implies*. Correspondingly,

$$\Pi_{(\circ\,|\,\oplus)} = 1 \rightarrow \qquad \Pi_{(\bar{\circ}\,|\,\oplus)} = 0, \quad \Pi_{(\oplus\,|\,\bar{\circ})} = 0, \quad \Pi_{(\bar{\oplus}\,|\,\bar{\circ})} = 1. \qquad (2.1.21)$$

For example, since $\Pi_{(\texttt{light}\,|\,\texttt{day})} = 1$ by definition, $\Pi_{(\texttt{dark}\,|\,\texttt{day})} = 0$, $\Pi_{(\texttt{day}\,|\,\texttt{dark})} = 0$, and $\Pi_{(\texttt{night}\,|\,\texttt{dark})} = 1$.

### *2.1.7 Expansion over a sample space*

Let the following $N$ mutually exclusive events define a sample space,

$$\circ_1, \quad \circ_2, \quad \ldots, \quad \circ_N, \tag{2.1.22}$$

that is, one of these events *has* to occur so that the sum of the associated probabilities is unity,

$$\sum_{i=1}^{N} \Pi_{(\circ_i)} = 1. \tag{2.1.23}$$

The probability of another arbitrary event, denoted by $\oplus$, is given by the sum of $N$ joint probabilities that can be expressed in terms of conditional probabilities,

$$\Pi_{(\oplus)} = \sum_{j=1}^{N} \Pi_{(\oplus\,,\,\circ_j)} = \sum_{j=1}^{N} \Pi_{(\oplus\,|\,\circ_j)} \times \Pi_{(\circ_j)}. \tag{2.1.24}$$

Applying Bayes' rule (2.1.7) with the constituent event $\circ_i$ in place of the generic event $\circ$, where $\circ_i$ is chosen arbitrarily from its sample space, and using (2.1.24), we obtain

$$\Pi_{(\circ_i\,|\,\oplus)} = \frac{\Pi_{(\oplus\,|\,\circ_i)}}{\sum_{j=1}^{N} \Pi_{(\oplus\,|\,\circ_j)} \times \Pi_{(\circ_j)}} \times \Pi_{(\circ_i)} \tag{2.1.25}$$

for $i = 1, \ldots, N$, which is the formal mathematical representation of Bayes' rule over a sample space.

When $N = 2$, we obtain the results discussed earlier in this section for an event and its complement, $\circ_1 = \circ$ and $\circ_2 = \bar{\circ}$.

Bayes' formula (2.1.25) for a set of mutually exclusive events drawn from a sample space is a starting point in Bayesian analysis for assessing cause and effect with multiples outcomes and events, where the event $\oplus$ typically represents the data.

*Exercises*

**2.1.1** Confirm Bayes' equation (2.1.6) with reference to the truck color schemes discussed in Section 1.4.

**2.1.2** Confirm that Bayes' equation (2.1.6) is satisfied for two independent events, $\circ$ and $\oplus$.

**2.1.3** Discuss the application of (2.1.25) when $\oplus$ is $\circ_k$ for $k = 1, \ldots, N$.

## 2.2 Data and events

In its most general interpretation, Bayes' equation derived in Section 2.1 is expressed by the following template equation:

$$\Pi_{(\texttt{event}\,|\,\texttt{data})} = \frac{\Pi_{(\texttt{data}\,|\,\texttt{event})}}{\Pi_{(\texttt{data}\,|\,\texttt{any event})}} \times \Pi_{(\texttt{event})}, \tag{2.2.1}$$

where:

• *event* may have a host of real or conceptual interpretations, including *hypothesis*, *suggestion*, *belief*, theory, or something else.

• *data* can be replaced by *information*, *circumstances*, *news*, *evidence*, *observation*, or something else.

### 2.2.1 Prior and posterior probabilities

In essence, the conceptual Bayes' rule relates the probability of an event, shown at the far right of (2.2.1),

$$\Pi_{(\texttt{event})}, \tag{2.2.2}$$

called the *prior probability*, to the probability of the same event deduced from data or information, or else occurring under certain conditions, shown at the far left of (2.2.1),

$$\Pi_{(\texttt{event}\,|\,\texttt{data})}, \tag{2.2.3}$$

called the *posterior probability*. We see that the posterior probability is, in fact, a conditional probability.

### 2.2.2 *Marginal probability*

The denominator of the fraction on the right-hand side of (2.2.1), expressed by the conditional probability

$$\Pi_{(\mathtt{data}\,|\,\mathtt{any\ event})} = \Pi_{(\mathtt{data})}, \tag{2.2.4}$$

is a marginal probability defined as the probability that the data arises under any conditions, not necessarily conditions related to a particular event of interest or a suggested hypothesis. Sometimes the marginal probability is called the *evidence*, although this terminology is uncomfortably reminiscent of forensics.

### 2.2.3 *Likelihood*

The numerator of the fraction on the right-hand side of equation (2.2.1), expressed by the conditional probability

$$\Pi_{(\mathtt{data}\,|\,\mathtt{event})}, \tag{2.2.5}$$

is the *likelihood*. The likelihood is the probability that the measured data are obtained from all possible data that could have been obtained, given an event. In some cases, the likelihood expresses the probability that the data or information procured are trustworthy, given a hypothesis or event.

### 2.2.4 *Correction factor*

The fraction on the right-hand side of (2.2.1) can be interpreted as a correction factor. In extreme cases, the correction factor would be zero. In the case or irrelevant or meaningless data, the correction factor would be unity.

### 2.2.5 *Measurable v. unmeasurable*

Bayes' rule is especially useful when the posterior probability $\Pi_{(\mathtt{event}\,|\,\mathtt{data})}$ is practically unmeasurable, whereas the likelihood $\Pi_{(\mathtt{data}\,|\,\mathtt{event})}$ is practically measurable or otherwise available. Sometimes the latter is borrowed with some trepidation or even malicious intent from different contexts.

For example, the conditional probability

$$\Pi_{(\texttt{Entered Heaven}\,|\,\texttt{Had a near-death experience})} \tag{2.2.6}$$

can be assessed by interviewing those who had near-death experiences, whereas the conditional probability

$$\Pi_{(\texttt{Had a near-death experience}\,|\,\texttt{Entered Heaven})} \tag{2.2.7}$$

is unmeasurable as it requires crossing the barricades of Heaven, only to return (Jackson Browne.)

### 2.2.6 Mathematical and conceptual models and propositions

In scientific applications, an event can be a tentative physical theory or mathematical model. Examples are Einstein's special theory of relativity, the proposition that the speed of light is a universal constant relating by equivalence mass to energy and *vice versa*, and the erroneous proposition that the sun rotates around the earth and not *vice versa*. Competing and complementary theories are often encountered.

Experiments can be designed and data can be collected to assess the probability that a theory or model is correct or reliable compared to alternatives. The advantage of Bayesian analysis lies in its ability to (*a*) suggest an experiment that generates compelling evidence or data and (*b*) help revise the probability that a model is correct or reliable in the face of data.

#### *Exercise*

**2.2.1** Describe an event and a set of data amenable to Bayesian analysis.

## 2.3 SSHL

Hearing loss may occur for two reasons: (*a*) mechanical damage of the middle ear causing *conductive loss* as in eardrum rupture, ossicular disruption, or otosclerosis, (*b*) nerve damage in the inner ear involving cochlear hair cells and associated neural pathways, leading to *sensorineural loss.* The latter is typically attributed to aging.

More than one hundred causes of SSHL *sudden* sensorineural hearing loss (SSHL) are known, including noise exposure and ototoxic medication. Since about $10\%$ of the population in the United states is currently on medication XMED, and since hearing loss is stated as a side effect of this medication, there is reason to be concerned that taking this medication will lead to SSHL.

### 2.3.1 Bayes rule

We are interested in the conditional probability $\Pi_{(\texttt{SSHL}\,|\,\texttt{Medication})}$ and make the following event identification:

$\circ$ : A person develops SSHL
$\oplus$ : A person takes medication XMED

Using Bayes' equation (2.1.7), we find that

$$\Pi_{(\texttt{SSHL}\,|\,\texttt{Medication})} = \frac{\Pi_{(\texttt{Medication}\,|\,\texttt{SSHL})}}{\Pi_{(\texttt{Medication})}} \times \Pi_{(\texttt{SSHL})}, \tag{2.3.1}$$

where $\Pi_{(\texttt{Medication})} = 0.10$.

Approximately $4,000$ new cases of *sudden* sensorineural hearing loss (SSHL) are diagnosed in the United States every year. For a population of $350$ million, we compute

$$\Pi_{(\texttt{SSHL})} = \frac{4,000}{350,000,000} = \frac{40}{35} \times 10^{-5} = 0.000014\cdots, \tag{2.3.2}$$

which is small.

Substituting these data into the Bayes formula, we find that

$$\Pi_{(\texttt{SSHL}\,|\,\texttt{Medication})} = 0.00014 \times \Pi_{(\texttt{Medication}\,|\,\texttt{SSHL})}. \tag{2.3.3}$$

Even if everyone who develops SSHL is on medication, which means that $\Pi_{(\texttt{Medication}\,|\,\texttt{SSHL})} = 1$, the probability of developing SSHL while taking the medication is extremely low.

### 2.3.2 At the ENT

Assume that a person who takes medication visits an ear-nose-throat (ENT) specialist who treats SSHL with the administration of steroids.

By talking to other patients in the waiting room, the patient discovers that every patient who developed SSHL was on XMED medication.

A much erroneous deduction would be that medication certainly leads to SSHL. However, we recall the deductions stated in (2.1.21),

$$\Pi_{(\circ\,|\,\oplus)} = 1 \rightarrow \qquad \Pi_{(\overline{\circ}\,|\,\oplus)} = 0, \quad \Pi_{(\oplus\,|\,\overline{\circ})} = 0, \quad \Pi_{(\overline{\oplus}\,|\,\overline{\circ})} = 1. \tag{2.3.4}$$

In fact, the deductions in (2.1.21) specify that

$$\Pi_{(\texttt{Medication}\,|\,\texttt{SSHL})} = 1 \quad \rightarrow \quad \Pi_{(\texttt{No medication}\,|\,\texttt{SSHL})} = 0, \qquad \Pi_{(\texttt{SSHL}\,|\,\texttt{No medication})} = 0, \quad \Pi_{(\texttt{No SSHL}\,|\,\texttt{No medication})} = 1, \tag{2.3.5}$$

which do not imply that $\Pi_{(\texttt{SSHL}\,|\,\texttt{Medication})} = 1$. Consistent with intuition, given that every patient who developed SSHL was on antidepressant medication, if a person does not take medication he/she will *not* develop SSHL.

***Exercise***

**2.3.1** Collect data and estimate the probability of developing lung cancer due to smoking.

## 2.4 Spam and the return of Dr. Who

We recall that Bayes' rule relates two conditional and two unconditional probabilities,

$$\Pi_{(\circ\,|\,\oplus)} = \frac{\Pi_{(\oplus\,|\,\circ)}}{\Pi_{(\oplus)}} \times \Pi_{(\circ)}, \tag{2.4.1}$$

where $\circ$ and $\oplus$ are two events, and proceed to make deductions.

We have mentioned that, often in practice, the conditional probability on the right-hand side, $\Pi_{(\oplus\,|\,\circ)}$, is *measurable,* whereas the conditional probability on the left-hand side, $\Pi_{(\circ\,|\,\oplus)}$ is *practically unmeasurable.* The hard-to-measure conditional probability on the left-hand side is desired.

### *2.4.1 Spam*

An email contains the words "*Dear respected*." To assess the probability that the email is spam, we make the following event identification:

$\circ$ : An email is spam
$\oplus$ : The email contains the words "*Dear respected*"

Referring to Bayes' equation (2.4.1), we gather the following three pieces of information:

1. A large fraction of emails are spam, $\Pi_{(\circ)} = 0.80$ (sad but true.)

2. A significant fraction of all emails, spam or legitimate, contain the words *Dear respected*, $\Pi_{(\oplus)} = 0.35$.

3. A significant fraction of spam emails contain the words *Dear respected*, $\Pi_{(\oplus\,|\,\circ)} = 0.40$.

Note that the second and third probabilities can be readily assessed. Also note that

$$\Pi_{(\oplus\,|\,\circ)} = 0.40 < \frac{\Pi_{(\oplus)}}{\Pi_{(\circ)}} = \frac{0.35}{0.80} = 0.4375, \tag{2.4.2}$$

as required by (1.10.17).

The complement of *the email is spam* ($\circ$) is *the email is not spam* ($\bar{\circ}$), and thus legitimate. Using the law of total probability expressed by (2.1.19), repeated below for convenience,

$$\Pi_{(\oplus\,|\,\bar{\circ})} = \frac{\Pi_{(\oplus)} - \Pi_{(\oplus\,|\,\circ)} \times \Pi_{(\circ)}}{1 - \Pi_{(\circ)}}, \tag{2.4.3}$$

we find that the probability that a legitimate email contains the words *Dear respected* is

$$\Pi_{(\oplus\,|\,\bar{\circ})} = \frac{0.35 - 0.40 \times 0.80}{1 - 0.80} = 0.15. \tag{2.4.4}$$

Now using Bayes' rule (2.4.1), we find that the probability that an email that contains the words *Dear respected* is spam is

$$\Pi_{(\circ\,|\,\oplus)} = \frac{0.40}{0.35} \times 0.80 = 0.914. \tag{2.4.5}$$

Consequently, there is a $91.4\%$ chance that the email is spam and should be redirected to the spam folder.

Also using Bayes' equation for $\bar{\circ}$,

$$\Pi_{(\bar{\circ}\,|\,\oplus)} = \frac{\Pi_{(\oplus\,|\,\bar{\circ})}}{\Pi_{(\oplus)}} \times \Pi_{(\bar{\circ})}, \tag{2.4.6}$$

we find that

$$\Pi_{(\bar{\circ}\,|\,\oplus)} = \frac{0.15}{0.35} \times (1 - 0.8) = 0.086, \tag{2.4.7}$$

which is expected in light of the requirement $\Pi_{(\circ\,|\,\oplus)} + \Pi_{(\bar{\circ}\,|\,\oplus)} = 1$. Consequently, there is only a $8.6\%$ chance that the email is legitimate and should not be redirected to the spam folder.

### 2.4.2 The return of Dr. Who

*Dr. Who* is a brilliant British science-fiction television series produced by the BBC, chronicling the adventures of a brilliant and compassionate Doctor. Consider the following event identification:

| | | |
|---|---|---|
| $\circ$ | : | A person is familiar with the special theory of relativity |
| $\oplus$ | : | A person watches reruns of *Dr. Who* |

We are provided with three pieces of information:

1. Only a small fraction of the general population are familiar with Einstein's general theory of relativity, $\Pi_{(\circ)} = 0.0023$.

2. Only a small fraction of the general population watch reruns of *Dr. Who*, $\Pi_{(\oplus)} = 0.0092$.

3. A significant fraction of those who are familiar with the special theory of relativity watch reruns of *Dr. Who*, $\Pi_{(\oplus\,|\,\circ)} = 0.72$.

The conditional probability $\Pi_{(\oplus\,|\,\circ)}$ can be assessed by polling those who are familiar with Einstein's general theory of relativity (they can be found on Internet physics forums.)

Using the law of total probability expressed by (2.1.19), repeated below for convenience,

$$\Pi_{(\oplus\,|\,\bar{\circ})} = \frac{\Pi_{(\oplus)} - \Pi_{(\oplus\,|\,\circ)} \times \Pi_{(\circ)}}{1 - \Pi_{(\circ)}}, \tag{2.4.8}$$

we find that the probability that a person who is not familiar with the theory of relativity watches reruns of *Dr. Who* is

$$\Pi_{(\oplus\,|\,\bar{\circ})} = \frac{0.0092 - 0.72 \times 0.0023}{1 - 0.0023} = 0.0076. \tag{2.4.9}$$

Using Bayes' rule (2.4.1), we find that the probability that a person who watches reruns of *Dr. Who* is familiar with the special theory of relativity is

$$\Pi_{(\circ\,|\,\oplus)} = \frac{0.72}{0.0092} \times 0.0023 = 0.18. \tag{2.4.10}$$

Stated differently, $18\%$ of those who watch reruns of *Dr. Who* are familiar with Einstein's general theory of relativity. By contrast, only $0.23\%$ of the general population are familiar with Einstein's special theory of relativity. We conclude that everyone should be watching reruns of *Dr. Who*.

### *Exercises*

**2.4.1** A person exhibits charm and charisma. In your estimation, what is the probability that these are genuine attributes meant to cheer up, entertain, and comfort others? (*Hint:* the answer is related to the *halo effect* discussed in Wikipedia.)

**2.4.2** Estimate the probability that a person who watches your favorite reality television show is familiar with Planck's constant.

## 2.5 Trial by jury

Prosecutors and defense attorneys make arguments to influence the subjective estimates of conditional probabilities of the individual jurors. In a trial by jury, all parties are interested in the probability of the following event and its complement:

| | | |
|---|---|---|
| $\circ$ | : | The defendant did it |
| $\bar{\circ}$ | : | Anybody else could have done it |

An implied assumption is that a crime was committed. If the committal of the crime is called into question, all probabilities become conditional.

Bayes' rule (2.1.7) for the guilt and its complement read

$$\Pi_{(\circ\,|\,\oplus)} = \frac{\Pi_{(\oplus\,|\,\circ)}}{\Pi_{(\oplus)}} \times \Pi_{(\circ)} \tag{2.5.1}$$

and

$$\Pi_{(\bar{\circ}\,|\,\oplus)} = \frac{\Pi_{(\oplus\,|\,\bar{\circ})}}{\Pi_{(\oplus)}} \times \Pi_{(\bar{\circ})}, \tag{2.5.2}$$

where $\oplus$ stands for partial or total evidence. Since someone must have done it in the presence or absence of evidence,

$$\Pi_{(\circ\,|\,\oplus)} + \Pi_{(\bar{\circ}\,|\,\oplus)} = 1. \tag{2.5.3}$$

If we know $\Pi_{(\circ\,|\,\oplus)}$, we can calculate $\Pi_{(\bar{\circ}\,|\,\oplus)}$ and vice versa.

To estimate the likelihood expressed by the conditional probability $\Pi_{(\oplus\,|\,\circ)}$, we ask the question: *If the defendant is guilty, what is the probability of observing the evidence?*

To estimate the likelihood expressed by the conditional probability $\Pi_{(\oplus\,|\,\bar{\circ})}$, we ask the question: *If someone else is guilty, what is the probability of observing the evidence?*

### 2.5.1 Odds ratio

The jurors and the judge are interested in the odds ratio defined as

$$\mathcal{O} \equiv \frac{\Pi_{(\circ\,|\,\oplus)}}{\Pi_{(\bar{\circ}\,|\,\oplus)}} = \frac{\Pi_{(\circ\,|\,\oplus)}}{1 - \Pi_{(\circ\,|\,\oplus)}}, \tag{2.5.4}$$

taking any zero or positive value. According to Bayes' rule,

$$\mathcal{O} = \frac{\Pi_{(\oplus\,|\,\circ)}}{\Pi_{(\oplus\,|\,\bar{\circ})}} \times \mathcal{O}_{\text{prior}}, \tag{2.5.5}$$

where

$$\mathcal{O}_{\text{prior}} \equiv \frac{\Pi_{(\circ)}}{\Pi_{(\bar{\circ})}} = \frac{\Pi_{(\circ)}}{1 - \Pi_{(\circ)}} \tag{2.5.6}$$

is the odds ratio before evidence taking any zero or positive value.

### *2.5.2 Bayes factor*

The fraction on the right-hand side of (2.5.5) is sometimes called the Bayes factor

$$\mathcal{B} \equiv \frac{\Pi_{(\oplus\,|\,\circ)}}{\Pi_{(\oplus\,|\,\bar{\circ})}}. \tag{2.5.7}$$

By definition then,

$$\mathcal{O} = \mathcal{B}\,\mathcal{O}_{\text{prior}}. \tag{2.5.8}$$

The odds ratio after evidence can be computed from the odds ratio before evidence and an assumed or speculative value for $\mathcal{B}$.

From (2.5.4), we find that

$$\Pi_{(\circ\,|\,\oplus)} = \frac{\mathcal{O}}{1+\mathcal{O}} = \frac{\mathcal{B}\mathcal{O}_{\text{prior}}}{1+\mathcal{B}\mathcal{O}_{\text{prior}}} = \frac{1}{1+1/(\mathcal{B}\mathcal{O}_{\text{prior}})}, \tag{2.5.9}$$

which is zero only if $\mathcal{B} = 0$ or $\mathcal{O}_{\text{prior}} = 0$.

### *2.5.3 Presumption of innocence*

Although any accused person should be presumed innocent, the odds ratio before evidence, $\mathcal{O}_{\text{prior}}$, cannot be zero, otherwise $\mathcal{O}$ will remain zero in spite of any evidence. This observation explains why trials of certain crimes are held at different jurisdictions.

Presumption of innocence should be interpreted to mean that, in a qualified population of $m$ persons, everyone has an equal probability of being guilty, yielding the prior odds ratio

$$\mathcal{O}_{\text{prior}} = \frac{1}{m-1}. \tag{2.5.10}$$

In a city of one million people, the prior guilt ratio is $10^{-6}$. If the crime is stealing egg rolls, the population is restricted to those who enjoy egg rolls. In practice, the prior guilt ratio is reset by the prosecution's and defense's initial arguments.

### 2.5.4 Not guilty and innocent

A person who is not guilty is innocent. In a trial by jury, a prosecutor needs to prove that a person is guilty beyond reasonable doubt. However, the defense attorney does not necessarily need to prove that the person is innocent. Only when all evidence points to a person's guilt, wrongly or rightly, the defense attorney needs to prove innocence by solving the crime.

This truism is the basic premise of mystery novels and television shows involving brilliant defense attorneys who discover the real perpetrator, such as the attorney portrayed by Raymond Barr.

#### Exercise

**2.5.1** Suppose that the first ten prime numbers were stolen from the Universe. What is a qualified set of perpetrators?

## 2.6 Probable cause

Bayes' rule provides us with insights on cause and effect, with the understanding that an effect may be due to different causes. The merits of different hypotheses can be quantified. As an example, we consider the following events:

$\circ$ : A person visits a car dealership
$\oplus$ : A person develops a headache

We are provided with three pieces of information:

1. A small fraction of the population visit a car dealership each day, $\Pi_{(\circ)} = 0.001$.

2. A good fraction of the general population suffer from headaches due to various causes, $\Pi_{(\oplus)} = 0.40$.

3. A significant fraction of those who visit a car dealership develop a headache, $\Pi_{(\oplus\,|\,\circ)} = 0.20$. Note that this conditional probability is measurable; you can imagine how.

   Using (2.1.19), we find that the probability of a person who did not visit a car dealership develops a headache is

$$\Pi_{(\oplus\,|\,\overline{\circ})} = \frac{0.40 - 0.20 \times 0.001}{1 - 0.001} \simeq 0.40 \cdots . \tag{2.6.1}$$

   In this case, $\Pi_{(\circ\,|\,\overline{\oplus})} \simeq \Pi_{(\circ)}$.

Using Bayes formula (2.1.7), we find that the daily probability that a person who develops a headache has just visited a car dealership is

$$\Pi_{(\circ\,|\,\oplus)} = \frac{0.20}{0.40} \times 0.001 = 0.0005. \tag{2.6.2}$$

Stated differently, $0.05\%$ of those who have a headache have just visited a car dealership; equivalently, $5$ out of $10,000$ people who have a headache have just visited a car dealership.

### *2.6.1 Another headache*

A headache may be caused by a different reason. Consider the following event identification:

$\circ$ : A person spends quality time with their dentist
$\oplus$ : A person develops a headache

We are provided with three pieces of information:

1. A good fraction of the population visit a dentist any given day, $\Pi_{(\circ)} = 0.003$.

2. A large fraction of the general population suffer from headaches due to various causes, $\Pi_{(\oplus)} = 0.40$.

3. A large fraction of those who visit a dentist develop a headache $\Pi_{(\oplus\,|\,\circ)} = 0.80$. Note that this conditional probability is measurable; you can imagine how.

   Using (2.1.19), we find that the probability of a person who did not visit a dentist develops a headache is

$$\Pi_{(\oplus\,|\,\overline{\circ})} = \frac{0.40 - 0.80 \times 0.003}{1 - 0.003} \simeq 0.40. \tag{2.6.3}$$

   In this case, $\Pi_{(\circ\,|\,\overline{\oplus})} \simeq \Pi_{(\circ)}$.

Using the Bayes formula (2.1.7), we find that the probability that a person who develops a headache has just visited their dentist is

$$\Pi_{(\circ\,|\,\oplus)} = \frac{0.80}{0.40} \times 0.003 = 0.006. \tag{2.6.4}$$

Stated differently, $0.6\%$ of those who have a headache have just visited a dentist; equivalently, $6$ out of $1,000$ people who have a headache have just visited a dentist.

### 2.6.2 Probable cause

The checkout cashier at the grocery store is telling us that she is having a splitting headache. Is this because she just visited a car dealership, or is it because she just visited her dentist? Bayesian analysis will tell us that it is because of the dentist, with probability $0.6\%$ v. $0.05\%$. In this context, Bayesian analysis was performed for the purpose of hypothesis testing.

#### *Exercises*

**2.6.1** A person exhibits charm and charisma. In your estimation, what is the probability that these are genuine attributes meant to cheer up, entertain, and comfort others? (*Hint:* the answer is related to the *halo effect* discussed in Wikipedia.)

**2.6.2** Compute the Bayes factor of visiting a car dealership v. visiting a dentist as the ratio of the corresponding conditional probabilities, $\Pi_{(\oplus\,|\,\circ)}$.

## 2.7 Accident at an intersection

The probability of an automobile accident at a particular intersection increases when it gets dark. Applying Bayes' equation (2.1.17), we write

$$\Pi_{(\texttt{accident}\,|\,\texttt{driving after dark})} = \frac{\Pi_{(\texttt{driving after dark}\,|\,\texttt{accident})}}{\Pi_{(\texttt{driving after dark}\,|\,\texttt{accident or no accident})}} \times \Pi_{(\texttt{accident})}. \quad (2.7.1)$$

The left-hand side is the probability of an accident occurring after dark. The marginal probability in the denominator on the right-hand side is given by

$$\Pi_{(\texttt{driving after dark}\,|\,\texttt{accident or no accident})} = \Pi_{(\texttt{driving after dark}\,|\,\texttt{accident})} \times \Pi_{(\texttt{accident})} + \Pi_{(\texttt{driving after dark}\,|\,\texttt{no accident})} \times \Pi_{(\texttt{no accident})}. \quad (2.7.2)$$

Two conditional probabilities appear on the right-hand side: the probability that a vehicle is involved in an accident after dark, and the probability that a vehicle not involved in an accident passes through the intersection unscathed after dark.

### 2.7.1 Data

The following pieces of information pertinent to the right-hand side of (2.7.1) is available:

1. $n$ automobiles pass through the intersection each week.
2. $80\%$ of the traffic occurs during daylight; correspondingly, $20\%$ of the traffic occurs after dark.
3. $n_a$ accidents occur each week, where $n_a \leq n$.
4. $n_d$ accidents occur after dark each week, where $n_d \leq n_a \leq n$.

The first two pieces of information are retrieved from automated records of a pressure-sensitive cable sensor stretched across the intersection. The third and fourth pieces of information are retrieved from police records.

Using these data, we compute the unconditional probabilities

$$\Pi_{(\texttt{accident})} = \frac{n_a}{n}, \qquad \Pi_{(\texttt{no accident})} = \frac{n - n_a}{n}, \tag{2.7.3}$$

and the conditional probabilities

$$\begin{aligned} \Pi_{(\texttt{driving after dark}\,|\,\texttt{accident})} &= \frac{n_d}{n_a}, \\ \Pi_{(\texttt{driving after dark}\,|\,\texttt{no accident})} &= 0.20. \end{aligned} \tag{2.7.4}$$

All necessary pieces of information are available.

### 2.7.2 Prediction

Substituting these expressions into Bayes' formula, we find that

$$\Pi_{(\texttt{accident}\,|\,\texttt{driving after dark})} = \frac{\dfrac{n_d}{n_a}}{\dfrac{n_d}{n_a} \times \dfrac{n_a}{n} + 0.20 \times \dfrac{n - n_a}{n}} \times \frac{n_a}{n}. \tag{2.7.5}$$

Simplifying, we find that the conditional probability of an accident after dark is given by

$$\Pi_{(\texttt{accident}\,|\,\texttt{driving after dark})} = \frac{n_d}{n_d + 0.20 \times (n - n_a)}. \tag{2.7.6}$$

If $n_a = n$ (every car has an accident), the probability is $1$, as expected. If $n_d = 0$ (no accidents occur after dark), the probability is $0$, as expected.

### 2.7.3 Bypassing the absurd

Formula (2.7.6) was derived based on measurable quantities. If we had to estimate the probability of having an accident after dark directly,

$$\Pi_{(\texttt{accident}\,|\,\texttt{after dark})}, \tag{2.7.7}$$

we would have to drive through the intersection at night several hundred times and count how may times we had an accident, which is unwise.

*Exercises*

**2.7.1** In your estimation, what is the probability that a person who was robbed regards a crowbar as an instrument of crime?

**2.7.2** You are waiting at an intersection for a school bus to pass. Estimate the probability that the school bus is followed by a car.

## 2.8 Defective light switch

Mr. Smith bought a three-way light switch from the local Big-Box store. The pertinent shelf contained two boxes from two different manufacturers, *Iswitch* and *Onoff*. Mr. Smith chose a box randomly and returned home to find out that the switch is defective. This is certainly aggravating, every trip to the store costs him three dollars for gas.

### 2.8.1 Iswitch or Onoff?

Mr. Smith googles the manufacturers and finds that $12\%$ of the switches made by *Iswitch* are defective and $19\%$ of the switches made by *Onoff* are defective. Mr. Smith is interested in the probabilities that the switch he bought was made by one or the other manufacturer,

$$\Pi_1 \equiv \Pi_{(\texttt{Made by Iswitch}\,|\,\texttt{Switch is defective})}, \tag{2.8.1}$$

and

$$\Pi_2 \equiv \Pi_{(\texttt{Made by Onoff}\,|\,\texttt{Switch is defective})}. \tag{2.8.2}$$

In the absence of other suppliers, the sum of these two conditional probabilities is unity. Mr. Smith uses Bayes' equation to write

$$\Pi_1 = \frac{\Pi_{(\texttt{Switch is defective}\,|\,\texttt{Made by Iswitch})}}{\Pi_{(\texttt{Switch is defective})}} \times \Pi_{(\texttt{Made by Iswitch})}. \tag{2.8.3}$$

The marginal probability in the denominator is given by

$$\begin{aligned}
&\Pi_{(\texttt{Switch is defective})} \\
&\quad = \Pi_{(\texttt{Switch is defective}\,|\,\texttt{Made by Iswitch})} \times \Pi_{(\texttt{Made by Iswitch})} \\
&\quad + \Pi_{(\texttt{Switch is defective}\,|\,\texttt{Made by Onoff})} \times \Pi_{(\texttt{Made by Onoff})}.
\end{aligned} \tag{2.8.4}$$

We expect that this marginal probability will lie between $0.12$ and $0.19$.

### 2.8.2 Bottom shelf

It suddenly occurred to Mr. Smith that *Iswitch* boxes are usually placed at the bottom shelf, while *Onoff* boxes are placed at the top shelf of the store. Since Mr. Smith usually picks up items from the bottom shelf, he sets

$$\Pi_{(\texttt{Made by Iswitch})} = \tfrac{3}{4} \qquad \Pi_{(\texttt{Made by Onoff})} = \tfrac{1}{4}. \tag{2.8.5}$$

Consequently,

$$\Pi_{(\texttt{Switch is defective})} = 0.12 \times \tfrac{3}{4} + 0.19 \times \tfrac{1}{4} = 0.1375, \tag{2.8.6}$$

and thus

$$\Pi_{(\texttt{Made by Iswitch}\,|\,\texttt{Switch is defective})} = \frac{0.12}{0.1375} \times \tfrac{3}{4} = 0.655. \tag{2.8.7}$$

Even though *Iswitch* is more reliable, the defective switch was probably made by them.

#### *Exercise*

**2.8.1** Mr. Smith examined all the switches who bought from the Big-Box store in the last year and observed that $8$ out of $10$ were made by *Onoff*. How does this observation affect his deduction?

## 2.9 The way we eat

In a business convention at an impressive hotel at Las Vegas, Nevada, a dinner of the $23$ most successful and influential local entrepreneurs was held. At the end of the dinner, young dishwasher Jeremiah mentioned to his young friend Meifeng that $15$ of the entrepreneurs grew up in poor families.

### 2.9.1 Kalamata olives

Jeremiah explained that $15$ of these entrepreneurs arranged neatly what they could not eat on the side of their plates, as though it were saved to be eaten later in the day or on the next day. The other $8$ entrepreneurs picked on the food and ate selectively only what they deemed most worthy. Some of them left precious Kalamata olives and delicious little balls of goat cheese on their plates. In fact, Jeremiah is a graduate student specializing in forensic psychology.

### 2.9.2 Appreciate food

It so happens that Meifeng is also a graduate student specializing in applied Bayesian statistics. Meifeng agrees that there is a correlation between growing up poor and appreciating food on a plate as a grown up. The pertinent conditional probability is

$$\Pi_1 \equiv \Pi_{(\texttt{Person grew up poor}\,|\,\texttt{Appreciate food})}. \tag{2.9.1}$$

Meifeng is thinking that there is a poor correlation between being wealthy and being classy.

It is fair to speculate that

$$\Pi_{(\texttt{Appreciate food}\,|\,\texttt{Person grew up poor})} = 0.99 \tag{2.9.2}$$

and

$$\Pi_{(\texttt{Appreciate food}\,|\,\texttt{Person did not grow up poor})} = 0.40, \tag{2.9.3}$$

which are measurable conditional probabilities. Based on current demographics, we set

$$\begin{aligned}\Pi_{(\texttt{Person grew up poor})} &= 0.30,\\ \Pi_{(\texttt{Person did not grow up poor})} &= 0.70.\end{aligned} \tag{2.9.4}$$

These unconditional probabilities add to unity, as required. Using Bayes' formula, Meifeng finds that

$$\Pi_1 = \frac{0.99}{0.99 \times 0.30 + 0.40 \times 0.70} \times 0.30 = 0.5147. \tag{2.9.5}$$

This calculation suggests that only $0.5147 \times 15 \simeq 8$ of the $15$ entrepreneurs who appreciated their food grew up in poor families. This estimate is close to that made based on the national average of a person that grew up poor, $23 \times 0.3 \simeq 7$.

### 2.9.3 Messed up food

Referring to (2.9.2) and (2.9.3), Meifeng finds that

$$\Pi_{(\texttt{Mess up food}\,|\,\texttt{Person grew up poor})} = 0.01 \tag{2.9.6}$$

and

$$\Pi_{(\texttt{Mess up food}\,|\,\texttt{Person did not grow up poor})} = 0.60, \tag{2.9.7}$$

which are measurable conditional probabilities. Using Bayes' formula, Meifeng finds that the probability

$$\Pi_2 \equiv \Pi_{(\texttt{Person grew up poor}\,|\,\texttt{Mess up food})} \tag{2.9.8}$$

is given by

$$\Pi_2 = \frac{0.01}{0.01 \times 0.30 + 0.60 \times 0.70} \times 0.30 = 0.0071. \tag{2.9.9}$$

This calculation suggests that only $0.0071 \times 8 \simeq 1$ of the $8$ entrepreneurs who messed up the food on their plates grew up in poor families. This estimate is far from that made based on the national average of a person that grew up poor, $23 \times 0.30 \simeq 7$.

Altogether, $8 + 1 = 9$ of the $23$ entrepreneurs grew up poor. Jeremiah and Meifeng googled the $23$ entrepreneurs who attended the dinner and confirmed that nine of them grew up in impoverished villages at the high desert.

### 2.9.4 Small things are the giveaway

The dominant clues are sometimes needles in a haystack. Looking for them is a matter of astute observation and raw intelligence assisted by experience. In the case of Meifeng and Jeremiah, the needles were the Kalamata olives. One can think of a variety of dominant clues for testing the sincerity and loyalty of a public person, friend, coworker, or acquaintance.

#### *Exercise*

**2.9.1** Discuss a dominant clue for testing the honesty of your financial adviser or stock broker (assuming that you have one.)

## 2.10 Walnut and almond cookies

Bo and Rebo pick up two boxes of walnut and almond cookies from the neighborhood bakery shop run by a young couple. The first box contains $w_1$ walnut and $a_1$ almond cookies, and the second box contains $w_2$ walnut and $a_2$ almond cookies. Bo and Rebo agree that they will enjoy the cookies in Bo's box and save the cookies in Rebo's box for their parents.

On the way home, Bo opens his box and offers to Rebo a randomly selected cookie. It is clear that

$$\Pi_{(\text{Bo gave Rebo a walnut cookie})} = \frac{w_1}{w_1 + a_1} \qquad (2.10.1)$$

and

$$\Pi_{(\text{Bo gave Rebo an almond cookie})} = \frac{a_1}{w_1 + a_1}. \qquad (2.10.2)$$

Rebo decides to save the cookie until they get home so that she can enjoy it with a glass of milk, and puts it into her box.

### 2.10.1 At home

At home, Bo and Rebo are greeted by their dad who is trimming the hedge of their garden. Rebo opens her box and offers him a randomly selected cookie. Dad remarks with appreciation that this is one of his favorite walnut cookies.

Given that dad was offered a walnut cookie, what is the probability that Bo had offered Rebo a walnut cookie? Formally, this expressed by the conditional probability

$$\Pi_1 \equiv \Pi_{(\text{Bo gave Rebo a walnut cookie} \,|\, \text{Rebo gave dad a walnut cookie})}. \qquad (2.10.3)$$

We know by now that Bayes' rule will allow us to compute this probability in terms of two reverse conditional probabilities

$$\Pi_2 \equiv \Pi_{(\text{Rebo gave dad a walnut cookie} \,|\, \text{Bo gave Rebo a walnut cookie})} \qquad (2.10.4)$$

and

$$\Pi_3 \equiv \Pi_{(\text{Rebo gave dad a walnut cookie} \,|\, \text{Bo gave Rebo an almond cookie})}, \qquad (2.10.5)$$

where

$$\Pi_2 = \frac{w_2 + 1}{w_2 + a_2 + 1}, \qquad \Pi_3 = \frac{w_2}{w_2 + a_2 + 1}. \tag{2.10.6}$$

Note that these conditional probabilities do not add to unity.

Using Bayes' rule, we obtain a formula for the desired probability,

$$\Pi_1 = \frac{\dfrac{w_2+1}{w_2+a_2+1}}{\dfrac{w_2+1}{w_2+a_2+1} \times \dfrac{w_1}{w_1+a_1} + \dfrac{w_2}{w_2+a_2+1} \times \dfrac{a_1}{w_1+a_1}} \times \frac{w_1}{w_1+a_1}. \tag{2.10.7}$$

Simplifying, we obtain

$$\Pi_1 = \frac{(w_2+1) \times w_1}{(w_2+1) \times w_1 + w_2 \times a_1}. \tag{2.10.8}$$

For $w_2 = 0$, we find that this probability is precisely unity, as expected. For $w_1 = 0$, we find that this probability is precisely zero, as expected. For $w_1 = 8$, $a_1 = 8$, $w_2 = 4$, and $a_2 = 12$, the probability is $0.556$.

***Exercise***

**2.10.1** Repeat the analysis in the event that Bo offered Rebo two randomly selected cookies.

## 2.11 Gardening on Labor Day

A sample space consisting of four events, $\circ_1$, $\circ_2$, $\circ_3$, and $\circ_4$, regarding the work shift of a sales associate is displayed in Table 2.11.1 along with the associated probabilities. Note that the probabilities listed in the last column add to unity, as required. The last probability in this table, $\Pi_{(\circ_4)} = 0.05$, allows for the possibility of being called off duty due to low customer presence or quitting the job due to an unsavory supervisor.

| *Event* | *Description* | *Probability* | *Numerical values* |
|---|---|---|---|
| $\circ_1$ | Working the morning shift | $\Pi_{(\circ_1)}$ | 0.50 |
| $\circ_2$ | Working the evening shift | $\Pi_{(\circ_2)}$ | 0.25 |
| $\circ_3$ | Working the graveyard shift | $\Pi_{(\circ_3)}$ | 0.20 |
| $\circ_4$ | Taking the day off | $\Pi_{(\circ_4)}$ | 0.05 |

TABLE 2.11.1 An example of four events ($N = 4$) constituting a sample space.

| *Conditional* | *Implied action* | *Value* |
|---|---|---|
| $\Pi_{(\oplus \mid \circ_1)}$ | Gardening in the afternoon | 0.2 |
| $\Pi_{(\oplus \mid \circ_2)}$ | Gardening in the morning | 0.3 |
| $\Pi_{(\oplus \mid \circ_3)}$ | Gardening in the morning or afternoon | 0.7 |
| $\Pi_{(\oplus \mid \circ_4)}$ | Gardening any time | 0.9 |

TABLE 2.11.2 Assigned values for the conditional probabilities, with reference to the events shown in Table 2.11.1.

Now consider an unrelated event denoted by:

$$\oplus \quad : \quad \text{Gardening}$$

The conditional probabilities listed in Table 2.11.2 appear reasonable. For example, $\Pi_{(\oplus \mid \circ_1)}$ is the probability of gardening, given that a person worked in the morning shift. Note that the conditional probabilities shown in Table 2.11.2 do not have to add to unity. In fact, each one of them is free to vary independently in the range $[0, 1]$.

Conversely, $\Pi_{(\circ_i \mid \oplus)}$ is the conditional probability that a person worked a selected shift or took a day off, given that the person spent some time gardening on that day ($\oplus$).

### *2.11.1 Not gardening*

The complement of gardening is:

$\overline{\oplus}$ : Not gardening

By definition,

$$\Pi_{(\oplus)} + \Pi_{(\overline{\oplus})} = 1. \tag{2.11.1}$$

The events represented by $\oplus$ (gardening) and $\oplus$ (not gardening) constitute a two-member sample space.

### *2.11.2 Matrix of conditional probabilities*

The two sample spaces of working or gardening can be accommodated in the $2 \times 4$ table of conditional probabilities shown in Table 2.11.3(*a*).

The two numbers in each column must add to unity, as indicated by the vertical arrows near the bottom. The reason is that the probability of gardening or not, given that one has worked any shift or took the day off, is unity.

Plugging in the numbers given in Table 2.11.2, we obtain the evaluated matrix shown in Table 2.11.3(*b*). Note that there is no reason that the four numbers in each row must add to unity. More generally,

$$\sum_i \Pi_{(\oplus, \circ_i)} \neq 1 \tag{2.11.2}$$

for any set of events, $\circ_i$, defining a sample space.

### *2.11.3 Matrix of joint probabilities*

The associated matrix of joint probabilities is shown in Table 2.11.4(*a*). In terms of the conditional probabilities, the joint probabilities shown in this table are given by

$$\Pi_{(\oplus\,,\,\circ_i)} = \Pi_{(\oplus\,|\,\circ_i)} \times \Pi_{(\circ_i)}, \qquad \Pi_{(\overline{\oplus}\,,\,\circ_i)} = \Pi_{(\oplus\,|\,\circ_i)} \times \Pi_{(\circ_i)} \tag{2.11.3}$$

for $i = 1, \ldots, 4$. Thus, the first column of the table of joint probabilities arises by multiplying the first column of the table of conditional

(*a*)

| | $\circ_1$ | $\circ_2$ | $\circ_3$ | $\circ_4$ |
|---|---|---|---|---|
| $\oplus$ | $\Pi_{(\oplus\,\vert\,\circ_1)}$ | $\Pi_{(\oplus\,\vert\,\circ_2)}$ | $\Pi_{(\oplus\,\vert\,\circ_3)}$ | $\Pi_{(\oplus\,\vert\,\circ_4)}$ |
| $\overline{\oplus}$ | $\Pi_{(\overline{\oplus}\,\vert\,\circ_1)}$ | $\Pi_{(\overline{\oplus}\,\vert\,\circ_2)}$ | $\Pi_{(\overline{\oplus}\,\vert\,\circ_3)}$ | $\Pi_{(\overline{\oplus}\,\vert\,\circ_4)}$ |
| | $\downarrow$ | $\downarrow$ | $\downarrow$ | $\downarrow$ |
| | 1.0 | 1.0 | 1.0 | 1.0 |

(*b*)

| | $\circ_1$ | $\circ_2$ | $\circ_3$ | $\circ_4$ |
|---|---|---|---|---|
| $\oplus$ | 0.2 | 0.3 | 0.7 | 0.9 |
| $\overline{\oplus}$ | 0.8 | 0.7 | 0.3 | 0.1 |
| | $\downarrow$ | $\downarrow$ | $\downarrow$ | $\downarrow$ |
| | 1.0 | 1.0 | 1.0 | 1.0 |

TABLE 2.11.3 (*a*) Matrix of conditional probabilities and (*b*) evaluated matrix of conditional probabilities.

probabilities shown in Table 2.11.3(*a*) by $\Pi_{(\circ_1)}$; similar multiplications are carried out for the other columns.

Plugging in the numbers, we obtain the evaluated matrix shown in Table 2.11.4(*b*). To compute the marginal probabilities shown in the last column, we have added the corresponding probabilities in each row. To compute the marginal probabilities shown in the last row, we have added the corresponding probabilities in each column. By construction, the sum of the marginal probabilities in the last column is unity and the sum of the marginal probabilities in the last row is also unity.

### *2.11.4 Gardening on Labor Day*

Big-Box Store Sales Associate Maria does not remember which shift she worked on Labor Day last year and assumes the following prior

(*a*)

| | $\circ_1$ | $\circ_2$ | $\circ_3$ | $\circ_4$ | | Marginal |
|---|---|---|---|---|---|---|
| $\oplus$ | $\Pi_{(\oplus\,,\circ_1)}$ | $\Pi_{(\oplus\,,\circ_2)}$ | $\Pi_{(\oplus\,,\circ_3)}$ | $\Pi_{(\oplus\,,\circ_4)}$ | $\rightarrow$ | $\Pi_{(\oplus)}$ |
| $\overline{\oplus}$ | $\Pi_{(\overline{\oplus}\,,\circ_1)}$ | $\Pi_{(\overline{\oplus}\,,\circ_2)}$ | $\Pi_{(\overline{\oplus}\,,\circ_3)}$ | $\Pi_{(\overline{\oplus}\,,\circ_4)}$ | $\rightarrow$ | $\Pi_{(\overline{\oplus})}$ |
| | $\downarrow$ | $\downarrow$ | $\downarrow$ | $\downarrow$ | | $\downarrow$ |
| Marginal | $\Pi_{(\circ_1)}$ | $\Pi_{(\circ_2)}$ | $\Pi_{(\circ_3)}$ | $\Pi_{(\circ_4)}$ | $\rightarrow$ | 1.0 |

(*b*)

| | $\circ_1$ | $\circ_2$ | $\circ_3$ | $\circ_4$ | | Marginal |
|---|---|---|---|---|---|---|
| $\oplus$ | 0.100 | 0.075 | 0.140 | 0.045 | $\rightarrow$ | $\Pi_{(\oplus)} = 0.360$ |
| $\overline{\oplus}$ | 0.400 | 0.175 | 0.060 | 0.005 | $\rightarrow$ | $\Pi_{(\overline{\oplus})} = 0.640$ |
| | $\downarrow$ | $\downarrow$ | $\downarrow$ | $\downarrow$ | | $\downarrow$ |
| Marginal | 0.50 | 0.25 | 0.20 | 0.05 | $\rightarrow$ | 1.0 |

TABLE 2.11.4 (*a*) Matrix of joint probabilities and (*b*) evaluated matrix of joint probabilities.

probabilities shown in Table 2.4.1,

$$\Pi_{(\circ_1)} = 0.50, \qquad \Pi_{(\circ_2)} = 0.25,$$
$$\Pi_{(\circ_3)} = 0.20, \qquad \Pi_{(\circ_4)} = 0.05. \tag{2.11.4}$$

However, Maria does remember gardening and enjoying the sunshine on that day.

Using formula (2.1.25) and the conditional probabilities shown in Table 2.11.2, Maria finds that

$$\Pi_{(\circ_i\,|\,\oplus)} = \frac{\Pi_{(\oplus\,|\,\circ_i)}}{0.2 \times 0.50 + 0.3 \times 0.25 + 0.7 \times 0.20 + 0.9 \times 0.05} \times \Pi_{(\circ_i)}, \tag{2.11.5}$$

which amounts to

$$\Pi_{(\circ_i\,|\,\oplus)} = \frac{\Pi_{(\oplus\,|\,\circ_i)}}{0.36} \times \Pi_{(\circ_i)}, \tag{2.11.6}$$

where $\Pi_{(\oplus)} = 0.36$ is the probability of gardening on any given day. Performing the calculations, Maria finds that

$$\Pi_{(\circ_1\,|\,\oplus)} = 0.277, \qquad \Pi_{(\circ_2\,|\,\oplus)} = 0.208,$$
$$\Pi_{(\circ_3\,|\,\oplus)} = 0.390, \qquad \Pi_{(\circ_4\,|\,\oplus)} = 0.125. \tag{2.11.7}$$

Thanks to the recollection of gardening, there is only $27.7\%$ chance that Maria worked the morning shift ($\circ_1$) last year, which is much lower than she initially thought. It is likely that Maria worked the graveyard shift ($\circ_3$).

### *2.11.5 Supervisor X or Y*

Maria's daughter Hannah pointed out that the priors listed in (2.11.4) should be chosen according to whether supervisor X or Y did the shift scheduling. Hannah remarked that, if supervisor X did the scheduling, then

$$\Pi_{(\circ_1)} = 0.70, \qquad \Pi_{(\circ_2)} = 0.25,$$
$$\Pi_{(\circ_3)} = 0.04, \qquad \Pi_{(\circ_4)} = 0.01. \tag{2.11.8}$$

If supervisor Y did the scheduling, then

$$\Pi_{(\circ_1)} = 0.30, \qquad \Pi_{(\circ_2)} = 0.50,$$
$$\Pi_{(\circ_3)} = 0.10, \qquad \Pi_{(\circ_4)} = 0.10. \tag{2.11.9}$$

In fact, these are conditional probabilities with respect to the choice of supervisor. This observation underlines the notion that no set of probabilities are truly unconditional. If you look hard enough, you can always find ifs, buts, fine prints, and pivot points.

If the probability that supervisor X did the scheduling is $\Pi_{(\mathrm{D})} = q$, and the probability that supervisor Y did the scheduling is $\Pi_{(\mathrm{M})} = 1-q$, then

$$\Pi_{(\circ_i)} = \Pi_{(\circ_i\,|\,\mathrm{A})} \times q + \Pi_{(\circ_i\,|\,\mathrm{B})} \times (1-q) \tag{2.11.10}$$

for $i = 1, 2, 3, 4$, where the parameter $q$ varies between $0$ and $1$. This expression can be substituted into the denominator on the right-hand side of (2.1.25).

### 2.11.6 Python

In fact, Hannah and her sisters are third-year students in computer science proficient in programming. Although Hannah's favorite computer language is *python*, for expedience, she wrote the following Matlab code named *rosa* to generate graphs of the posterior probabilities $\Pi_{(\bigcirc_i \,|\, \oplus)}$ against $q$:

```
%===
% compute and plot probabilities
% using Bayes' formula
%===

%---
% data
%---

p1A = 0.70; p2A = 0.25; p3A = 0.04; p4A = 0.01;
p1B = 0.30; p2B = 0.50; p3B = 0.10; p4B = 0.10;

%---
% prepare
%---

Nq = 64;
Dq = 1.0/Nq;

%---
% scan
%---

for i=1:Nq+1
 q = (i-1)*Dq;
 qc = 1.0-q;
 prior1 = p1A*q+p1B*qc;
 prior2 = p2A*q+p2B*qc;
 prior3 = p3A*q+p3B*qc;
 prior4 = p4A*q+p4B*qc;
 den = 0.2*prior1+0.3*prior2+0.7*prior3+0.9*prior4;
 post1(i) = 0.2*prior1/den;
```

```
 post2(i) = 0.3*prior2/den;
 post3(i) = 0.7*prior3/den;
 post4(i) = 0.9*prior4/den;
 qvect(i) = q;
end

%---
% plot
%---

plot(qvect,post1,'k')
plot(qvect,post2,'r--')
plot(qvect,post3,'b.-')
plot(qvect,post4,'k-.')
plot([0.5 0.5],[0,0.6],'k:')
plot([0 1.0],[0.5,0.5],'k:')
xlabel('q','fontsize',15)
ylabel('\Pi_o','fontsize',15)
```

The graphics display generated by the code is shown in Figure 2.11.1. For $q = \frac{1}{2}$, the probability $\Pi_{(\bigcirc_2)}$ is slightly higher than $\Pi_{(\bigcirc_1)}$, which suggests that Maria worked the evening shift.

***Exercises***

**2.11.1** Present the counterpart of Table 2.11.4(*b*) when the first entry of Table 2.11.2 is changed from $0.2$ to $0.1$.

**2.11.2** Repeat the Bayesian analysis with three supervisors.

## *2.12 At the DMV*

Ms. Horváth lived frugally for many years while working as a pharmacy technician for sixty hours a week. She never enjoyed an avocado toast or a fancy coffee drink prepared by a barista, she ordered items only from the dollar menu and a large hot or iced coffee for lunch, she used a no-contract flip-phone and Internet hot-spot device from Walmart,

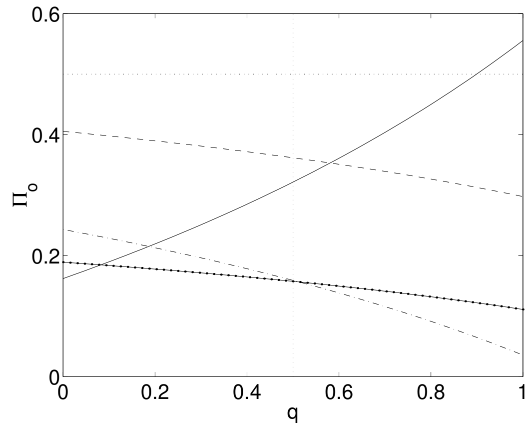



FIGURE 2.11.1 Dependence of the posterior probabilities $\Pi_{(\circ_i\,|\,\oplus)}$ on the supervisor's probability, $q$. The solid line is for $\circ_1$, the red dashed line is for $\circ_2$, the blue connected dots is for $\circ_3$, and the black dot-dashed line is for $\circ_4$.

and she only watched free television on the air-transmitted channels she could receive (no cable or satellite.)

Finally, she was able to save enough money to buy a precious yellow/blue truck from *MotoMobil, LLC* paying, as she always does, the full amount in cash. Ms. Horváth noticed that, as the years go by, the engine horsepower of cars and trucks get smaller but the electronic devices proliferate. Ms. Horváth once noticed a truck equipped with a coffee maker in the central console.

### 2.12.1 Date of sale

Ms. Horváth is now at the Department of Motor Vehicles (DMV) to register her vehicle but does not remember the exact date of purchase, though she does remember that it was either the fifth or the sixth of September when trucks are offered at a discounted prices in anticipation of the following year's models.

To find out the exact date of sale, Ms. Horváth calls from her flip-phone *MotoMobil, LLC*. Unfortunately, the receptionist informs her that

| | *Blue* | *White* | → Total |
|---|---|---|---|
| *Yellow* | $\frac{n\,m}{k}$ | $\frac{n\,(k-m)}{k}$ | $n$ |
| *Red* | $\frac{(k-n)\,m}{k}$ | $\frac{(k-n)(k-m)}{k}$ | $k-n$ |
| Total ↓ | $m$ | $k-m$ | $k$ |

TABLE 2.12.1 Sales matrix for $k$ truck sales with uncorrelated color scheme preferences; $n$ of these trucks have yellow exterior, and $m$ trucks have blue interior.

she only has access to the general sales records at this time of the day. The receptionist explains that, if $k$ trucks were sold, $n$ of these trucks had yellow exterior, and $m$ of these trucks had blue interior, then the sales matrix will appear as shown in Table 1.8.3, repeated in Table 2.12.1 for convenience.

The receptionist is kind enough to provide Ms. Horváth with the data

$$k = 14, \qquad n = 9, \qquad m = 6 \tag{2.12.1}$$

for the fifth of the month, and with the data

$$k = 20, \qquad n = 18, \qquad m = 15 \tag{2.12.2}$$

for the sixth of the month. The receptionist asks Ms. Horváth to keep these data confidential so that other dealers do not gain an advantage and start acting smart.

### *2.12.2 Recollections from Biostatistics*

Ms. Horváth recalls the principles of Bayesian analysis from the biostatistics class she took at college, taught by her favorite professor whose last name could not remember but is sure it started with a P and ended with an s. The professor always had a sad expression in his eyes.

To carry out the Bayesian inference, Ms. Horváth defines two events:

$\circ_5$ : Purchased the truck on the fifth day of the month
$\circ_6$ : Purchased the truck on the sixth day of the month

and the event:

$\oplus$ : Purchased a yellow/blue truck

She then applies Bayes' equation (2.1.25) to obtain

$$\Pi_{(\circ_i \,|\, \oplus)} = \frac{\Pi_{(\oplus \,|\, \circ_i)}}{\Pi_{(\oplus \,|\, \circ_5)} \times \Pi(\circ_5) + \Pi_{(\oplus \,|\, \circ_6)} \times \Pi(\circ_6)} \times \Pi_{(\circ_i)} \quad (2.12.3)$$

for $i = 5, 6$, where

$$\Pi_{(\circ_5)} = q, \qquad \Pi_{(\circ_6)} = 1 - q \quad (2.12.4)$$

are the priors, and $q$ is a parameter in the range $[0, 1]$ expressing Ms. Horváth's vague recollection.

Using the sales table provided to her by the receptionist, Ms. Horváth finds that

$$\Pi_{(\circ_5 \,|\, \oplus)} = \frac{\left(\frac{nm}{k^2}\right)_5}{\left(\frac{nm}{k^2}\right)_5 \times q + \left(\frac{nm}{k^2}\right)_6 \times (1-q)} \times q \quad (2.12.5)$$

for the fifth of the month, and

$$\Pi_{(\circ_6 \,|\, \oplus)} = \frac{\left(\frac{nm}{k^2}\right)_6}{\left(\frac{nm}{k^2}\right)_5 \times q + \left(\frac{nm}{k^2}\right)_6 \times (1-q)} \times (1-q) \quad (2.12.6)$$

for the sixth of the month.

Ms. Horváth pulls out her Unix laptop and codes these equations into a following Matlab program namer *horvath*, listed below:

```
%===
% plot probabilities using Bayes' formula
%===
```

```
%---
% data
%---

k5 = 14; n5 =  9; m5 = 6;
k6 = 20; n6 = 18; m6 = 15;
%---
% prepare
%---

A5 = n5*m5/k5^2; A6 = n6*m6/k6^2;

Nq = 64;    % number of point points
Dq = 1/Nq;  % step in q

%---
% scan
%---

for i=1:Nq+1
 q = (i-1)*Dq;
 qvect(i) = q;
 prob5(i) = A5*q/(A5*q+A6*(1-q));
 prob6(i) = 1.0-prob5(i);
end

hold on
plot(qvect,prob5,'k')
plot(qvect,prob6,'k--')
plot([0.5 0.5],[0,1],'k')
plot([0 1.0],[0.5,0.5],'k')
xlabel('q','fontsize',15)
ylabel('\Pi_{5}                 \Pi_{6}','fontsize',15)
```

Running the code generates the graph shown in Figure 2.12.1, where the solid line represents the probability that she had bought the truck on the fifth of the month and the broken line represents the probability that she bought the truck on the sixth of the month. For each $q$, the two probabilities add to unity, as required.

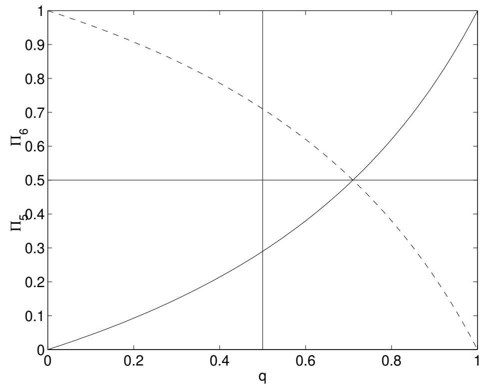



FIGURE 2.12.1 Probability of truck sale on the fifth of the month (solid line) or sixth of the month (broken line) as a function of the degree of prior recollection, $q$.

### *2.12.3 Registered*

Ms. Horváth had a vague recollection that she bought the truck on the fifth of the month and guesses $q = 0.6$. The graph shown in Figure 2.12.1 suggests that the sale occurred on the sixth of the month with probability around $\Pi_6 = 0.6$.

As Ms. Horváth proceeds to fill out the registration form, putting with some trepidation the sixth of the month as purchase date, the receptionist calls on her flip phone to confirm that the sale occurred on the sixth of the month. Ms. Horváth will enjoy her truck for many years to come.

### *2.12.4 On the sublime*

While standing in line at the Department of Motor Vehicles (DMV) to register her brand-new truck, Ms. Horváth chatted with gardener Diego who had taken time off work to register his thirty-three year-old Toyota work truck that barely passed inspection. Diego was holding tight in his fist the required fee in cash, making sure it was correct to the last penny. Diego mentioned that he was in a hurry to get back to work so that he did not miss any more hourly pay.

### *2.12.5 Act of generosity*

Diego appeared to have misunderstood Ms. Horváth's situation and thought she was distressed because she too had trouble paying the fees and also filling the gas tank. In a stunning display of generosity, Diego offered to pay gratis Ms. Horváth's fees and return to register his own truck on the next day after working overtime beyond his usual twelve-hour work day.

### *2.12.6 Vanished*

Ms. Horváth's politely and evasively declined and proceeded to complete her transaction when her number was called. When she exited the DMV, she could not locate in the parking lot a thirty-three year old Toyota work truck. When she returned to the building, she could not find Diego. Diego seems to have vanished in thin air; or was he ever there?

### *2.12.7 Trays of sushi and egg-rolls*

Ms. Horváth never forgot Diego's sublime act of generosity and made a habit of delivering six giant trays of the most expensive sushi she could find to the local Shelter in the first week of January and in the last week of May each year, so that the underprivileged could taste something nice as they plow through life, one hour at a time. In delivering the sushi, she was joined by her friend Chenguang who delivered two trays of her delicious home-made egg-rolls.

### *2.12.8 In Industrial Engineering*

Ms. Horváth's husband mentioned the incidence to a few of his colleagues in the Department of Industrial Engineering. The faculty decided with a nearly unanimously vote $21$–$1$ to add to the Bylaws a clause that cancels the Departmental holiday reception and requires instead using the funds to deliver to the local homeless shelter six giant trays of the most expensive sushi they could find on the last day of the final exams of the Fall semester each year, so that the underprivileged could taste something nice as they plow through life, one hour at a time.

Most homeless women and men are compassionate and intelligent human beings whose hopes and dreams have been shattered at some early point in their lives. Many homeless men are war vets who dream of the fight fast asleep at the traffic light (Jackson Browne.) Most people don't realize that the difference between an able bodied person and a disabled person is a bicycle ride.

#### 2.12.9 No sushi from Eaky

Professor of fluid mechanics S. N. Eaky voted against the sushi measure in the Department of Industrial Engineering. At the faculty meeting, he argued that the holiday reception fosters collegiality in an informal setting that facilitates the exchange of ideas on how to secure large research grants in response to calls for interdisciplinary proposals.

The professor discourages his graduate students from volunteering at the local hospital and requires instead that they spend all of their time at the lab. The students must comply with his requirements to ensure supportive recommendation letters upon graduation.

#### 2.12.10 Buttering the rim

On his way to the annual meeting of the American Physical Society, the professor sat next to Mr. P. L. Easant, a vice president of a reputable community bank.

The vice president is a delightful and interesting man with a broad range of interests, including the repair of small machines and the cooking of delicious meals on a dime. When he retires in a couple of years, the vice president plans to use his skills to help organic farmers fill out tax returns and repair farm equipment.

The vice president explained to the professor that, if you butter the rim of a pot, water will not boil over. Professor Eaky made a note of this observation in his fancy hand-held device as soon as he deplaned.

#### 2.12.11 Distinguished

While attending the fluid mechanics meeting and pretending to be listening to presentations, the professor wrote a grant proposal entitled "*Investigation of the effect of contact angle discontinuity on the velocity of an advancing three-phase contact angle.*" In the *Impact to Society* section, the professor argued that his research will help family-owned restaurants and underprivileged populations efficiently boil spaghetti without wasting precious boiled-over water.

The grant was awarded, as professor Eaky is well connected, influential, a member of the scientific establishment, and an expert in writing proposals that cannot be declined.

### *2.12.12 No eye contact*

After the grant was awarded, Dr. Eaky was elevated by his institution to the upper echelon of scholars, chosen to deliver a Distinguished Lecture, and subsequently elected to yet another national academy.

In the award bestowing reception, Dr. Eaky stated with euphoria that he will never retire, and vowed to continue writing grant proposals and keep accumulating honors and awards for the benefit of his institution. A few members of the audience rolled their eyes and most avoided making eye contact with him for several months. Others profusely congratulated him by all available means.

The vice president became aware of Eaky's achievements from the local newspaper and called to sincerely congratulate him in a rare display of class and generosity. When Eaky realized who was on the other end of the phone, he hung up, fearing liability, but no remorse, for his unethical actions.

### *Exercises*

**2.12.1** Perform the Bayesian analysis for the case where Ms. Horváth thought that the sale occurred on the fifth, sixth, or seventh of the month and the receptionist provided her with the pertinent data.

**2.12.2** Write a two-paragraph essay that describes the psychological profiles of Mr. Easant and Dr. Eaky.

## 2.13 Who ate the pastrami?

A set of fingerprints were found on a refrigerator door, and Carol accuses her husband, Frank, of eating the pastrami while everyone else was asleep after midnight.

In Carol's defense, this would not be the first time that Frank pulled a fast one, although he typically goes for the peppered salami. Years of lecturing on the importance of eating fruit and vegetables have gone by in vain. Frank argues that the vast majority of nutritional advice is unfounded and should be ignored.

### 2.13.1 Frank proclaims his innocence

Frank dismisses the accusation and argues that the following conditional probability of interest is significant,

$$\Pi_1 \equiv \Pi_{(\texttt{Someone else ate the pastrami}\,|\,\texttt{Fingerprint found})}, \qquad (2.13.1)$$

where *Fingerprint* is the fingerprint found on the fridge. Consequently, Frank states with indignity that the following conditional probability of vital interest is small:

$$\Pi_{(\texttt{Frank ate the pastrami}\,|\,\texttt{Fingerprint found})} = 1 - \Pi_1. \qquad (2.13.2)$$

To make his case, Frank invokes the Bayes formula to write on the chores board

$$\Pi_1 = \frac{\Pi_{(\texttt{Fingerprint found}\,|\,\texttt{Someone else ate the pastrami})}}{\Pi_{(\texttt{Fingerprint found})}} \times \Pi_{(\texttt{Someone else})}. \qquad (2.13.3)$$

Frank has a way of estimating the probabilities on the right-hand side.

### 2.13.2 Pastrami eaters

Frank did some research to find that a fraction $q$ of illegal pastrami eaters leave fingerprints on other peoples' refrigerators, that is,

$$\Pi_{(\texttt{Fingerprint found}\,|\,\texttt{Someone else ate the pastrami})} = q. \qquad (2.13.4)$$

Frank also states that he never wears gloves when he picks pastrami slices,

$$\Pi_{(\texttt{Fingerprint found}\,|\,\texttt{Frank ate the pastrami})} = 1.0. \tag{2.13.5}$$

This means that, if Frank ate the pastrami, his fingerprints would be found for sure all over the place.

Frank recalls vaguely that, on one occasion, he ate pastrami while wearing gloves as he was painting the front porch. However, this exception is put aside as a member of a different sample space. Surgeons were gloves in order to prevent infections rather than ensure that fingerprints are not left behind.

### *2.13.3 Fingerprint forensics*

Frank makes his case by arguing that $1$ out of $L$ people are detected to have the same set of fingerprints by the forensic department's outdated software. From a general population of $N$ people, a number of

$$n = \frac{N}{L} \tag{2.13.6}$$

people will be detected to have the same fingerprints, where generally $N > L$.

The prior probability that Frank ate the pastrami is the probability that he is one of these people,

$$\Pi_{(\texttt{Frank})} = \frac{1}{n} = \frac{L}{N}. \tag{2.13.7}$$

The prior probability that someone else ate the pastrami is given by Frank's complement,

$$\Pi_{(\texttt{Someone else})} = 1 - \Pi_{(\texttt{Frank})} = \frac{n-1}{n} = \frac{N-L}{N}. \tag{2.13.8}$$

Frank has collected all necessary data.

The marginal probability in the denominator on the right-hand side of (2.13.3) is given by

$$\begin{aligned} &\Pi_{(\texttt{Fingerprint found})} \\ &\quad = \Pi_{(\texttt{Fingerprint found}\,|\,\texttt{Someone else ate the pastrami})} \times \Pi_{(\texttt{Someone else})} \\ &\quad\ + \Pi_{(\texttt{Fingerprint found}\,|\,\texttt{Frank ate the pastrami})} \times \Pi_{(\texttt{Frank})}. \end{aligned} \tag{2.13.9}$$

Substituting the preceding expressions, we obtain

$$\Pi_{(\texttt{Fingerprint found})} = q \times \frac{N-L}{N} + 1.0 \times \frac{L}{N}. \tag{2.13.10}$$

All pieces of information are now available to conduct the Bayesian analysis.

### 2.13.4 Preponderance of guilt

Bayes' formula yields

$$\Pi_1 = \frac{q}{q \times \dfrac{N-L}{N} + 1.0 \times \dfrac{L}{N}} \times \frac{N-L}{N}. \tag{2.13.11}$$

Simplifying, we obtain

$$\Pi_1 = \frac{q \times (N-L)}{q \times N + (1.0-q) \times L}. \tag{2.13.12}$$

For a population of $N = sL$ people, where $s$ is a defined coefficient, we obtain

$$\Pi_1 = \frac{q \times (s-1)}{q \times s + 1.0 - q}, \tag{2.13.13}$$

We recall that $q$ is the fraction of illegal pastrami eaters that leave fingerprints on other peoples' refrigerators. We observe that:

- When $s = 1$, the probability that someone else ate the pastrami is zero. The reason is that Frank's fingerprints are unique.

- When $q = 0$, the probability that someone else ate the pastrami is also zero. The reason is that no other illegal pastrami eater leaves fingerprints.

- As $s$ tends to infinity, the pool of potential illegal eaters increases and the probability that someone else ate the pastrami tends to unity (the fraction on the right-hand side of (2.13.13) tends to unity.)

For $s = 2$ and $q = 0.5$, we find that $\Pi_1 = 1/3$, and this suggests that Frank is the culprit. Frank is planning to spend the night in the dog house.

### *2.13.5 The liver is also gone*

Carol noticed that whoever ate the pastrami, also ate the liver. Frank hates liver. The probability that Frank ate the liver is $0.033$. The general population likes liver. The probability that an intruder ate the liver is $0.90$.

The posterior probability given in (2.13.13) now provides us with a new prior estimate, denoted for convenience as

$$\varsigma = \frac{q \times (s-1)}{q \times s + 1.0 - q}. \tag{2.13.14}$$

Frank defines the conditional probability

$$\Pi_2 \equiv \Pi_{(\texttt{Someone else ate the pastrami and the liver} \mid \texttt{Fingerprint found})} \tag{2.13.15}$$

and repeats the Bayesian analysis to finds that

$$\Pi_2 = \frac{0.90}{0.90 \times \varsigma + 0.033 \times (1 - \varsigma)} \times \varsigma. \tag{2.13.16}$$

Substituting the expression for $\varsigma$, Frank obtains

$$\Pi_2 = \frac{q \times 0.90 \times (s-1)}{0.90 \times q \times s + 0.033 - q \times 0.90}. \tag{2.13.17}$$

For $s = 2$ and $q = 0.5$, we find that $\Pi_2 = 0.9319$, which suggests that Frank should be exonerated.

Had Carol not observed the missing liver, poor Frank would have spent the night in the dog house.

### *2.13.6 Dominant clues*

Carol's observation that the liver was also gone was a key piece of evidence in assessing Frank's innocence. Some clues or observations (data) are mundane in that they carry an expected amount of information according to conventional wisdom. Other clues carry an overwhelming amount of information.

### 2.13.7 Bayesian deduction

Hercules Poireau and other brilliant detectives and defense lawyers habitually and instinctively look for seemingly tertiary data that determine without bias the posterior probability of someone's guilt or innocence in a Bayesian deduction. One such evidence can be the amount of ash in an ashtray.

Hercules' final conclusion is complete only when all data, especially seemingly tertiary data, are consistent with his conclusions. Hercules enjoys explaining his deductions and train of thought in a much anticipated and craftily staged denouement.

#### Exercise

**2.13.1** Frank's daughter points out that, whoever ate the pastrami and the liver, also drank milk. What is the new probability that someone other than Frank is the culprit?

## 2.14 Rolling the dice

A person rolls two fair (unbiased) dice and reports that the sum of the two numbers facing up is $12$. What it the probability that the first dice has rolled on $6$, given that the sum is $12$?

Since the sum can be $12$ only if both dice have rolled on $6$, the relevant conditional probability is unity,

$$\Pi_{(\texttt{First dice shows 6}\,|\,\texttt{Sum is 12})} = 1, \qquad (2.14.1)$$

which means that the first dice has rolled on $6$ beyond any doubt.

### 2.14.1 Bayes' theorem

To confirm the obvious, we apply Bayes' formula (2.2.1) and find that the conditional probability is given by

$$\Pi_{(\texttt{First dice shows 6}\,|\,\texttt{Sum is 12})} = \frac{\Pi_{(\texttt{Sum is 12}\,|\,\texttt{First dice shows 6})}}{\Pi_{(\texttt{Sum is 12}\,|\,\texttt{First dice shows 1 OR 2 OR 3 OR 4 OR 5 OR 6})}} \times \Pi_{(\texttt{First dice shows 6})}, \tag{2.14.2}$$

where

$$\Pi_{(\texttt{First dice shows 6})} = \frac{1}{6} \tag{2.14.3}$$

is an unconditional probability and

$$\Pi_{(\texttt{Sum is 12}\,|\,\texttt{First dice shows 6})} = \frac{1}{6} \tag{2.14.4}$$

is the conditional probability that the sum is $12$, given that the first dice rolled on $6$; the second dice could have rolled on $1, 2, 3, 4, 5$, or $6$.

The marginal probability in the denominator is given by

$$\begin{aligned}\Pi&_{(\texttt{Sum is 12}\,|\,\texttt{First dice shows 1 OR 2 OR 3 OR 4 OR 5 OR 6})} \\ &= \Pi_{(\texttt{Sum is 12}\,|\,\texttt{First dice shows 1})} \times \frac{1}{6} + \Pi_{(\texttt{Sum is 12}\,|\,\texttt{First dice shows 2})} \times \frac{1}{6} \\ &\quad + \cdots + \Pi_{(\texttt{Sum is 12}\,|\,\texttt{First dice shows 6})} \times \frac{1}{6}.\end{aligned} \tag{2.14.5}$$

We note that

$$\Pi_{(\texttt{Sum is 12}\,|\,\texttt{First dice shows m})} = 0 \tag{2.14.6}$$

for $m = 1, \ldots, 5$, and find that

$$\Pi_{(\texttt{Sum is 12}\,|\,\texttt{First dice shows 1 OR 2 OR 3 OR 4 OR 5 OR 6})} = \frac{1}{6^2}. \tag{2.14.7}$$

Substituting these results into (2.14.2), we confirm the conditional probability shown in (2.14.1).

### 2.14.2 *Rolling twice*

If the sum of the two dice were $11$, we would have considered the probability

$$\Pi_{(\texttt{First dice shows 6}\,|\,\texttt{Sum is 11})} = \frac{\Pi_{(\texttt{Sum is 11}\,|\,\texttt{First dice shows 6})}}{\Pi_{(\texttt{Sum is 11}\,|\,\texttt{First dice shows 1 OR 2 OR 3 OR 4 OR 5 OR 6})}} \times \Pi_{(\texttt{First dice shows 6})}, \tag{2.14.8}$$

where

$$\Pi_{(\texttt{First dice shows 6})} = \frac{1}{6},$$
$$\Pi_{(\texttt{Sum is } 11\,|\,\texttt{First dice shows 6})} = \frac{1}{6}, \tag{2.14.9}$$

and

$$\Pi_{(\texttt{Sum is } 11\,|\,1 \text{ OR } 2 \text{ OR } 3 \text{ OR } 4 \text{ OR } 5 \text{ OR } 6)} = \Pi_{(\texttt{Sum is } 11\,|\,\texttt{First dice shows 1})} \times \frac{1}{6} + \Pi_{(\texttt{Sum is } 11\,|\,\texttt{First dice shows 2})} \times \frac{1}{6} + \cdots + \Pi_{(\texttt{Sum is } 11\,|\,\texttt{First dice shows 6})} \times \frac{1}{6}. \tag{2.14.10}$$

We note that

$$\Pi_{(\texttt{Sum is } 11\,|\,\texttt{First dice shows m})} = 0 \tag{2.14.11}$$

for $m = 1, \ldots, 4$, and also

$$\Pi_{(\texttt{Sum is } 11\,|\,\texttt{First dice shows 5})} = \frac{1}{6},$$
$$\Pi_{(\texttt{Sum is } 11\,|\,\texttt{First dice shows 6})} = \frac{1}{6}. \tag{2.14.12}$$

Making substitutions, we find that

$$\Pi_{(\texttt{Sum is } 11\,|\,\texttt{First dice shows } 1 \text{ OR } 2 \text{ OR } 3 \text{ OR } 4 \text{ OR } 5 \text{ OR } 6)} = 2 \times \frac{1}{6^2}, \tag{2.14.13}$$

yielding

$$\Pi_{(\texttt{First dice shows 6}\,|\,\texttt{Sum is } 11)} = \frac{1}{2}. \tag{2.14.14}$$

Working in a similar fashion, we find that

$$\Pi_{(\texttt{First dice shows 5}\,|\,\texttt{Sum is } 11)} = \frac{1}{2} \tag{2.14.15}$$

and

$$\Pi_{(\texttt{First dice shows } m\,|\,\texttt{Sum is } 11)} = 0 \tag{2.14.16}$$

for $m = 1, 2, 3, 4$.

***Exercise***

**2.14.1** After rolling two dice, we are informed that the sum of the two outcomes is $10$. The probability that the first dice has rolled on $1, 2$, or $3$ is zero. What is the probability that the first dice has rolled on $4, 5$, or $6$?

## *2.15 Multiple events*

Consider the joint probability of three events, denoted by $\circ$, $\oplus$, and $\boxdot$,

$$\Pi_{(\circ\,,\,\oplus\,,\,\boxdot)}. \tag{2.15.1}$$

The order by which the three events are listed is immaterial.

### *2.15.1 Sample space*

The three events may belong to the same sample space or different sample spaces. Either way, the joint probability is a member of a composite joint-probability Cartesian sample space consisting of $2^3 = 8$ joint probabilities,

$$\begin{array}{llll} \Pi_{(\circ\,,\,\oplus\,,\,\boxdot)}, & \Pi_{(\circ\,,\,\overline{\oplus}\,,\,\boxdot)}, & \Pi_{(\overline{\circ}\,,\,\oplus\,,\,\boxdot)}, & \Pi_{(\overline{\circ}\,,\,\overline{\oplus}\,,\,\boxdot)}, \\ \Pi_{(\circ\,,\,\oplus\,,\,\overline{\boxdot})}, & \Pi_{(\circ\,,\,\overline{\oplus}\,,\,\overline{\boxdot})}, & \Pi_{(\overline{\circ}\,,\,\oplus\,,\,\overline{\boxdot})}, & \Pi_{(\overline{\circ}\,,\,\overline{\oplus}\,,\,\overline{\boxdot})}, \end{array} \tag{2.15.2}$$

where a line over an event denotes the complement. Since these combinations exhaust all possibilities, these eight joint probabilities must add to unity.

We can think of the event combinations as a partition of a cube mediated by a double tensor or Cartesian product,

$$(\circ, \overline{\circ}) \otimes (\oplus, \overline{\oplus}) \otimes (\boxdot, \overline{\boxdot}). \tag{2.15.3}$$

### *2.15.2 Factorization into conditional probabilities*

By definition of the conditional probability, we may write

$$\begin{aligned} \Pi_{(\circ\,,\,\oplus\,,\,\boxdot)} &= \Pi_{(\circ\,|\,(\oplus\,,\,\boxdot))} \times \Pi_{(\oplus\,,\,\boxdot)} \\ &= \Pi_{(\oplus\,|\,(\boxdot,\,\circ))} \times \Pi_{(\boxdot\,,\,\circ)} = \Pi_{(\boxdot\,|\,(\circ\,,\,\oplus))} \times \Pi_{(\circ\,,\,\oplus)}. \end{aligned} \tag{2.15.4}$$

The pairwise joint probabilities can be expressed in terms of conditional probabilities, yielding

$$\begin{aligned} \Pi_{(\circ\,,\,\oplus\,,\,\square)} &= \Pi_{(\circ\,|\,(\oplus\,,\,\square))} \times \Pi_{(\oplus\,|\,\square)} \times \Pi_{(\square)} \\ &= \Pi_{(\oplus\,|\,(\square\,,\,\circ))} \times \Pi_{(\square\,|\,\circ)} \times \Pi_{(\circ)} \\ &= \Pi_{(\square\,|\,(\circ\,,\,\oplus))} \times \Pi_{(\circ\,|\,\oplus)} \times \Pi_{(\oplus)}. \end{aligned} \tag{2.15.5}$$

Alternatively, we may write

$$\begin{aligned} \Pi_{(\circ\,,\,\oplus\,,\,\square)} &= \Pi_{((\circ\,,\,\oplus)\,|\,\square)} \times \Pi_{(\square)} \\ &= \Pi_{((\oplus\,,\,\square)\,|\,\circ)} \times \Pi_{(\circ)} = \Pi_{((\square\,,\,\circ)\,|\,\oplus)} \times \Pi_{(\oplus)}. \end{aligned} \tag{2.15.6}$$

Using the preceding relations, we can derive expressions for conditional probabilities of interest in terms of others.

For example, setting the first expressions in (2.15.5) and (2.15.6) equal, we obtain

$$\Pi_{((\circ\,,\,\oplus)\,|\,\square)} = \Pi_{(\circ\,|\,(\oplus\,,\,\square))} \times \Pi_{(\oplus\,|\,\square)}, \tag{2.15.7}$$

where $\square$ can be regarded as an attached condition, as discussed in Section 1.6.

### *2.15.3 Bayes rules*

Setting the last expression in (2.15.4) equal to the first expression, in (2.15.6), we obtain

$$\Pi_{(\square\,|\,(\circ\,,\,\oplus))} \times \Pi_{(\circ\,,\,\oplus)} = \Pi_{((\circ\,,\,\oplus)\,|\,\square)} \times \Pi_{(\square)}, \tag{2.15.8}$$

which can be rearranged into a Bayes rule,

$$\Pi_{(\square\,|\,(\circ\,,\,\oplus))} = \frac{\Pi_{((\circ\,,\,\oplus)\,|\,\square)}}{\Pi_{(\circ\,,\,\oplus)}} \times \Pi_{(\square)}. \tag{2.15.9}$$

Setting the second expression in (2.15.4) equal to the third expression, we obtain another Bayes rule,

$$\Pi_{(\circ\,|\,(\oplus,\square))} \times \Pi_{(\oplus\,,\,\square)} = \Pi_{(\oplus\,|\,(\circ,\square))} \times \Pi_{(\circ\,,\,\square)}. \tag{2.15.10}$$

Rearranging, we obtain a high-order Bayes' rule

$$\Pi_{(\circ\,|\,(\oplus\,,\,\square))} = \frac{\Pi_{(\oplus\,|\,(\circ\,,\,\square))}}{\Pi_{(\oplus\,,\,\square)}} \times \Pi_{(\circ\,,\,\square)}. \tag{2.15.11}$$

Writing

$$\Pi_{(\circ\,,\,\square)} = \Pi_{(\circ\,|\,\square)} \times \Pi_{\square}, \qquad \Pi_{(\oplus\,,\,\square)} = \Pi_{(\oplus\,|\,\square)} \times \Pi_{\square}, \tag{2.15.12}$$

we obtain

$$\Pi_{(\circ\,|\,(\oplus\,,\,\square))} = \frac{\Pi_{(\oplus\,|\,(\circ,\square))}}{\Pi_{(\oplus\,|\,\square)}} \times \Pi_{(\circ\,|\,\square)}, \tag{2.15.13}$$

which simply attaches the condition $\square$ to the regular Bayes' rule.

The high-order Bayes rules are useful when the conditional probabilities are available or can be calculated on the right-hand side.

### *Exercise*

2.15.1 Construct the sample space of the joint probability of three specific events of your choice.

# Chapter 3

# Bayesian deductions

In Chapter 2, we introduced Bayes' rule and discussed applications of Bayesian analysis in several settings and contexts. In this chapter, we continue the discussion of applications in further contexts and more advanced frameworks.

Certain consequences of Bayes' rule will turn out to surprising, to the extent that the basic premise may appear to be deceptively wrong or else we get something out of nothing. We will conclude this chapters by discussing how decisions can be made on right and wrong.

## 3.1 The Monte Hall problem

Wikipedia describes the Monte Hall question submitted by a reader as follows:

"*Suppose you're on a game show, and you're given the choice of three doors. Behind one door is a car; behind the others, goats.*

*You pick a door, say #1, and the host, who knows what's behind the doors, opens another door, say #3, which has a goat.*

*He then says to you, "Do you want to pick door #2?" Is it to your advantage to switch your choice?*"

### 3.1.1 The wrong answer

On first impulse, one may answer that it does not matter whether you stay with the chosen door or switch to the other closed door. The (false) reasoning is that there is an equal probability for the prize to be behind either closed door. However, much to many's surprise this answer is wrong.

### *3.1.2 The right answer*

Before the host opens a door, the prior unconditional probability that the car is behind the door that you picked is $1/3$ and the prior probability that the car is behind the other two closed doors is $2/3$.

If the car is not behind the second door, the probability that it is behind the third door remains equal to $2/3$. Likewise, if the car is not behind the third door, the probability that it is behind the second door remains equal to $2/3$.

We conclude that it is to the player's advantage to switch the initial choice to double his chance of winning. A mathematical proof in a generalized framework will be carried out in this section based on Bayesian analysis.

An important lesson to be learnt from this stunningly simple reasoning is that the initial player's choice of door and response of the host, whatever these may be, do not affect the location of the prize.

It does not matter how many company CEOs or university chancellors order their subordinates to develop a unified field theory in six months, otherwise cost-of-living salary adjustment will not be approved and tenure will be denied, if such a theory requires an unknown framework that hinges on what today is classified as paranormal. Two hundred years ago the idea of space–time continuum and warped space would be considered instances of the paranormal.

### *3.1.3 Bayesian analysis*

To carry out the Bayesian analysis, we introduce the player sample space consisting of three events:

$\mathrm{P}_1$: Player chooses Door 1 with probability $\alpha$
$\mathrm{P}_2$: Player chooses Door 2 with probability $\beta$
$\mathrm{P}_3$: Player chooses Door 3 with probability $\gamma$

where

$$\alpha + \beta + \gamma = 1. \tag{3.1.1}$$

Next, we introduce the host sample space consisting of three different events:

$H_1$: Host opens Door 1
$H_2$: Host opens Door 2
$H_3$: Host opens Door 3

The joint events $(P_1,\ H_1)$, $(P_2,\ H_2)$, and $(P_3,\ H_3)$, are mutually exclusive.

The car can be behind door 1 with probability $\Pi_{(\texttt{car behind 1})}$, behind door 2 with probability $\Pi_{(\texttt{car behind 2})}$, or behind door 3 with probability $\Pi_{(\texttt{car behind 3})}$. For each car location behind door $\#n$, we introduce the joint probabilities, $\Pi^{\mathtt{n}}_{(P_i\,,\,H_j)}$, for $i, j = 1, 2, 3$, whose event doublets define a joint sample space. The nine probabilities can be arranged in three joint probability matrices,

$$\mathbf{\Pi}^{\mathtt{n}} \equiv \begin{bmatrix} \Pi^{\mathtt{n}}_{(P_1\,,\,H_1)} & \Pi^{\mathtt{n}}_{(P_1\,,\,H_2)} & \Pi^{\mathtt{n}}_{(P_1\,,\,H_3)} \\ \Pi^{\mathtt{n}}_{(P_2\,,\,H_1)} & \Pi^{\mathtt{n}}_{(P_2\,,\,H_2)} & \Pi^{\mathtt{n}}_{(P_2\,,\,H_3)} \\ \Pi^{\mathtt{n}}_{(P_3\,,\,H_1)} & \Pi^{\mathtt{n}}_{(P_3\,,\,H_2)} & \Pi^{\mathtt{n}}_{(P_3\,,\,H_3)} \end{bmatrix} \tag{3.1.2}$$

for $\mathtt{n} = 1, 2, 3$. The sum of all elements of this matrix is equal to $1.0$ for any car location. A key observation is that the values of these joint probabilities depend on the car location according to the rules of the game.

Since the host never opens the door that the player has chosen initially, the diagonal elements of this matrix are zero for any $\mathtt{n}$,

$$\Pi^{\mathtt{n}}_{(P_1\,,\,H_1)} = 0, \quad \Pi^{\mathtt{n}}_{(P_2\,,\,H_2)} = 0, \quad \Pi^{\mathtt{n}}_{(P_3\,,\,H_3)} = 0 \tag{3.1.3}$$

for any car location.

Since the host never opens the door leading to a car,

$$\Pi^{\mathtt{n}}_{(P_i\,,\,H_n)} = 0 \tag{3.1.4}$$

for $i = 1, 2, 3$ and any $\mathtt{n}$. Consequently, the nth column of $\mathbf{\Pi}^{\mathtt{n}}$ is zero.

### 3.1.4 Probability trees

To evaluate the joint probabilities, we introduce the probability trees shown in Table 3.1.1.

For the sake of generality, we have assumed that, after the player's initial door selection, the host opens one of the other two available doors with probabilities $q$, $r$, and $s$, respectively, for $\mathtt{n} = 1, 2, 3$, and the other available door with probabilities $1 - q$, $1 - r$, and $1 - s$.

Based on the expressions shown in Table 3.1.1, we find that

$$\mathbf{\Pi}^1 \equiv \begin{bmatrix} 0 & \alpha\, q & \alpha\,(1-q) \\ 0 & 0 & \beta \\ 0 & \gamma & 0 \end{bmatrix} \tag{3.1.5}$$

when the car is behind door 1,

$$\mathbf{\Pi}^2 \equiv \begin{bmatrix} 0 & 0 & \alpha \\ \beta\,(1-r) & 0 & \beta\, r \\ \gamma & 0 & 0 \end{bmatrix} \tag{3.1.6}$$

when the car is behind door 2, and

$$\mathbf{\Pi}^3 \equiv \begin{bmatrix} 0 & \alpha & 0 \\ \beta & 0 & 0 \\ \gamma\, s & \gamma\,(1-s) & 0 \end{bmatrix} \tag{3.1.7}$$

when the car is behind door 3. The diagonal elements of each matrix are zero.

### 3.1.5 Bayes formula

Bayes' formula provides us with the expressions

$$\Pi_{(\mathtt{car\ behind\ n}\,|\,\mathrm{P}_i,\mathrm{H}_j)} = \frac{\Pi^{\mathtt{n}}_{(\mathrm{P}_i,\mathrm{H}_j)}}{\Pi^{\mathtt{any\ n}}_{(\mathrm{P}_i,\mathrm{H}_j)}} \times \Pi_{(\mathtt{car\ behind\ n})} \tag{3.1.8}$$

for $i, j, n = 1, 2, 3$. The breakthrough afforded by the Bayesian approach is that the conditional probabilities in the numerator on the right-hand side,

$$\Pi^{\mathtt{n}}_{(\mathrm{P}_i,\mathrm{H}_j)} \tag{3.1.9}$$

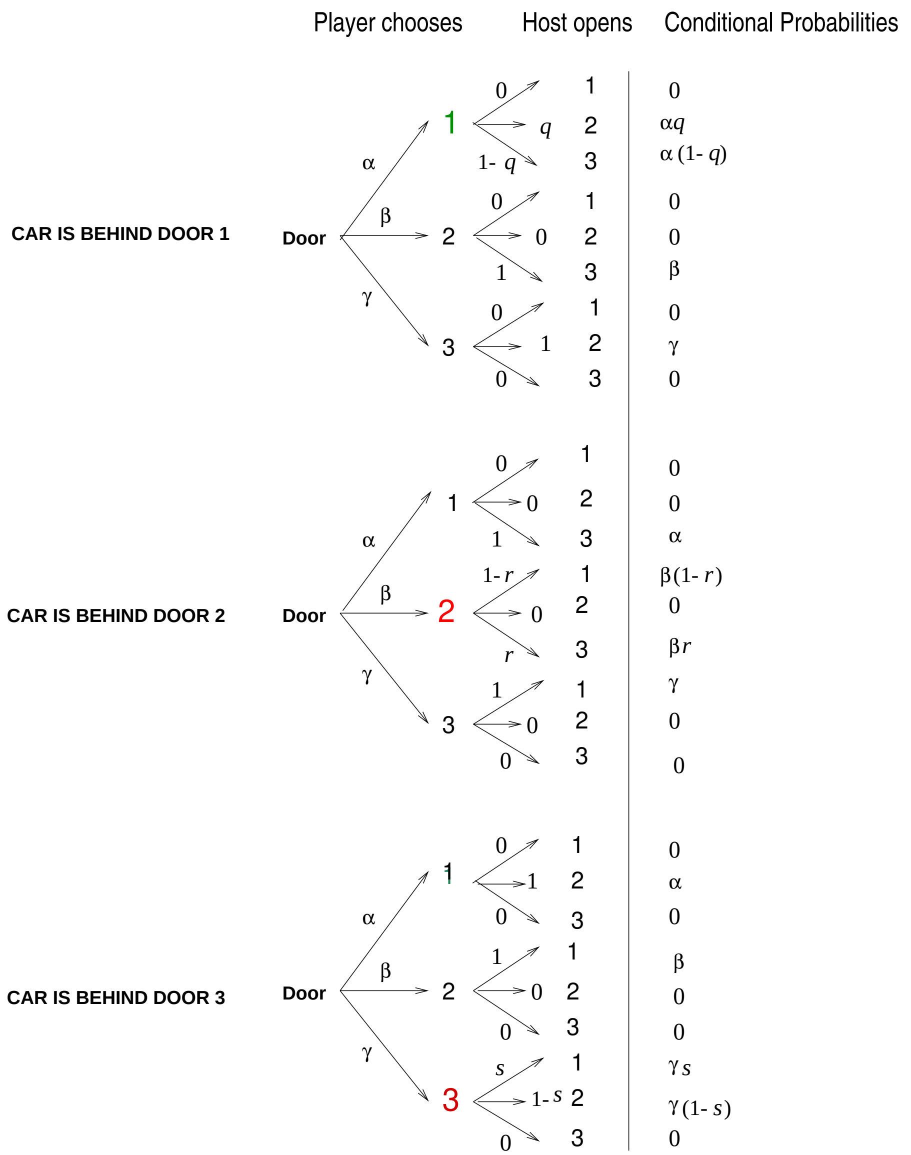



TABLE 3.1.1 Monte Hall probability tree diagrams for three car locations. The player chooses initially door #1 with probability $\alpha$, door #2 with probability $\beta$, and door #3 with probability $\gamma$, where $\alpha + \beta + \gamma = 1$.

are available in (3.1.5), (3.1.6), and (3.1.7). All parameters involved in these probabilities are determined by the game designers, not by the player.

### *3.1.6 Case of P1-H2*

In the event that the player chooses the first door and the host opens the second door, we consider the probabilities

$$\Pi_{(\texttt{car behind 1}\,|\,\mathrm{P}_1,\mathrm{H}_2)} = \frac{\alpha\, q}{\Pi^{\texttt{any n}}_{(\mathrm{P}_1,\mathrm{H}_2)}} \times \Pi_{(\texttt{car behind 1})} \tag{3.1.10}$$

and

$$\Pi_{(\texttt{car behind 3}\,|\,\mathrm{P}_1,\mathrm{H}_2)} = \frac{\alpha}{\Pi^{\texttt{any n}}_{(\mathrm{P}_1,\mathrm{H}_2)}} \times \Pi_{(\texttt{car behind 3})}, \tag{3.1.11}$$

where

$$\Pi^{\texttt{any n}}_{(\mathrm{P}_1,\mathrm{H}_2)} = \alpha\, q \times \Pi_{(\texttt{car behind 1})} + \alpha \times \Pi_{(\texttt{car behind 3})}. \tag{3.1.12}$$

The ratio between these posterior probabilities is

$$\frac{\Pi_{(\texttt{car behind 3}\,|\,\mathrm{P}_1,\mathrm{H}_2)}}{\Pi_{(\texttt{car behind 1}\,|\,\mathrm{P}_1,\mathrm{H}_2)}} = \frac{1}{q} \times \frac{\Pi_{(\texttt{car behind 3})}}{\Pi_{(\texttt{car behind 1})}}. \tag{3.1.13}$$

When

$$\frac{\Pi_{(\texttt{car behind 3})}}{\Pi_{(\texttt{car behind 1})}} > q, \tag{3.1.14}$$

the fraction on the left-hand side of (3.1.13) is higher than unity and consequentially

$$\Pi_{(\texttt{car behind 3}\,|\,\mathrm{P}_1,\mathrm{H}_2)} > \Pi_{(\texttt{car behind 1}\,|\,\mathrm{P}_1,\mathrm{H}_2)}. \tag{3.1.15}$$

The player should abandon door #1 that she initially chose and switch to door #3. The recommendation should be reversed when

$$\frac{\Pi_{(\texttt{car behind 3})}}{\Pi_{(\texttt{car behind 1})}} < q. \tag{3.1.16}$$

When this condition is met, the player should stick with #1 that she initially chose.

### 3.1.7 Case of P1-H3

In the event that the player chooses the first door and the host opens the third door, we consider the probabilities

$$\Pi_{(\texttt{car behind 1}\,|\,\mathrm{P}_1,\mathrm{H}_3)} = \frac{\alpha\,(1-q)}{\Pi^{\texttt{any n}}_{(\mathrm{P}_1,\mathrm{H}_3)}} \times \Pi_{(\texttt{car behind 1})} \tag{3.1.17}$$

and

$$\Pi_{(\texttt{car behind 2}\,|\,\mathrm{P}_1,\mathrm{H}_3)} = \frac{\alpha}{\Pi^{\texttt{any n}}_{(\mathrm{P}_1,\mathrm{H}_3)}} \times \Pi_{(\texttt{car behind 2})}, \tag{3.1.18}$$

where

$$\Pi^{\texttt{any n}}_{(\mathrm{P}_1,\mathrm{H}_3)} = \alpha\,(1-q) \times \Pi_{(\texttt{car behind 1})} + \alpha \times \Pi_{(\texttt{car behind 2})}. \tag{3.1.19}$$

The ratio between these posterior probabilities is

$$\frac{\Pi_{(\texttt{car behind 2}\,|\,\mathrm{P}_1,\mathrm{H}_3)}}{\Pi_{(\texttt{car behind 1}\,|\,\mathrm{P}_1,\mathrm{H}_3)}} = \frac{1}{1-q} \times \frac{\Pi_{(\texttt{car behind 2})}}{\Pi_{(\texttt{car behind 1})}}. \tag{3.1.20}$$

When

$$\frac{\Pi_{(\texttt{car behind 2})}}{\Pi_{(\texttt{car behind 1})}} > 1-q, \tag{3.1.21}$$

the fraction on the left-hand side of (3.1.20) is higher than unity and consequentially

$$\Pi_{(\texttt{car behind 2}\,|\,\mathrm{P}_1,\mathrm{H}_3)} > \Pi_{(\texttt{car behind 1}\,|\,\mathrm{P}_1,\mathrm{H}_3)}. \tag{3.1.22}$$

The player should abandon door #1 that she has initially chose and switch to door #2.

The recommendation should be reversed when

$$\frac{\Pi_{(\texttt{car behind 2})}}{\Pi_{(\texttt{car behind 1})}} < 1-q. \tag{3.1.23}$$

When this condition is met, the player should stick with #1 that she initially chose.

### 3.1.8 Case of P2-H3

Results for other cases follow from those just derived with straightforward changes in notation and cyclic permutation of $q, r, s$.

In case of $\mathrm{P2}-\mathrm{H3}$, the player should abandon door #2 and switch to door $\#1$ when

$$\frac{\Pi_{(\texttt{car behind 1})}}{\Pi_{(\texttt{car behind 2})}} > r, \tag{3.1.24}$$

or stick with door #2 otherwise.

### 3.1.9 Case of P2-H1

In case of $\mathrm{P2}-\mathrm{H1}$, the player should abandon door #2 and switch to door $\#3$ when

$$\frac{\Pi_{(\texttt{car behind 3})}}{\Pi_{(\texttt{car behind 2})}} > 1 - r, \tag{3.1.25}$$

or stick with door #2 otherwise.

### 3.1.10 Case of P3-H1

In case of $\mathrm{P3}-\mathrm{H1}$, the player should abandon door #3 and switch to door $\#2$ when

$$\frac{\Pi_{(\texttt{car behind 2})}}{\Pi_{(\texttt{car behind 3})}} > s, \tag{3.1.26}$$

or stick with door #3 otherwise.

### 3.1.11 Case of P3-H2

In case of $\mathrm{P3}-\mathrm{H2}$, the player should abandon door #3 and switch to door $\#1$ when

$$\frac{\Pi_{(\texttt{car behind 1})}}{\Pi_{(\texttt{car behind 3})}} > 1 - s, \tag{3.1.27}$$

or stick with door #3 otherwise.

### *3.1.12 Absence of bias*

In the absence of any kind of bias or preference,

$$\Pi_{(\texttt{car behind 1})} = \Pi_{(\texttt{car behind 2})} = \Pi_{(\texttt{car behind 3})} = \tfrac{1}{3} \qquad (3.1.28)$$

and

$$q = r = s = \tfrac{1}{2}. \qquad (3.1.29)$$

In all cases, the player should abandon her initial choice and switch to the only other available closed door.

### *3.1.13 Know your friends and your frenemies*

In a game of strategy, you make a move (choose a door) and a person who is not acting in your best interest (host) responds with another move (open a door).

To get ahead, it is wise to reverse your move (choose another door), and thus compensate for the other person's calculated advantage. However, if the other person acts in your best interest, you should stay with your initial choice.

A practical piece of advice is that you should always submit an annual employee evaluation report to an adversarial boss or an unhinged university Dean at the last possible minute to deprive them from the advantage of time and force them to make an erratic move.

#### *Exercise*

**3.1.1** Play the Monte Hall game with your favorite delivery person, where you are the player and the delivery person is the host. Report the outcome for a large number of trials.

## 3.2 Toasters

A fraction of toasters manufactured by *Flying Toasters, LLC* are defective for various reasons. This means that, if $N$ units are manufactured over a certain period of time, $d \times N$ units will have wire soldering issues and $(1-d) \times N$ units will work fine, where $d$ is the fraction of defective units.

If $d = 0$, the product is outstanding, completely free of defects, and worthy of five stars on Internet customer reviews. If $d = 1$, every unit is problematic; the company should hold an equipment liquidation sale and file for bankruptcy.

For example, if $873$ toasters were sold over a period of one week, and if $3$ of these toasters were defective, then

$$d = \frac{3}{873} \simeq 0.00344. \tag{3.2.1}$$

This is a reasonable rate of failure. We may say that the probability of a toaster being defective is $0.344\%$, which is low.

### 3.2.1 Quality index

It is helpful to introduce a *Boolean variable* serving as a binary quality index, $q$, taking the value of $0$ or $1$, where:

- $q = 0$ signifies a defective toaster
- $q = 1$ signifies a good toaster

### 3.2.2 Prior probabilities

By definition, the *unconditional* (prior) probability that a randomly chosen toaster is defective is

$$\Pi_{(q=0)} = d. \tag{3.2.2}$$

The corresponding unconditional (prior) probability that a randomly chosen toaster makes toast is

$$\Pi_{(q=1)} = 1 - \Pi_{(q=0)} = 1 - d. \tag{3.2.3}$$

In the absence of information, these probabilities are *a priori* unknown.

### 3.2.3 Quality control

*Flying Toasters, LLC* runs a product inspection station to implement quality control. A Boolean variable, $p$, is introduced to quantify the results of an inspection, taking the value $0$ or $1$, as follows:

- $p = 0$ signifies that a toaster failed inspection
- $p = 1$ signifies that a toaster passed inspection

A rejection sticker with a sad face is attached to a toaster that failed inspection.

However, the testing procedure is not full-proof and a quality assessment can be erroneous. On occasion, a good toaster will be mistaken for a bad toaster (false negative) and a bad toaster will be mislabeled as a good toaster (false positive).

### 3.2.4 Sensitivity

Inspection of $N_{\text{bad}}$ toasters *that are known to be defective* $(q = 0)$ resulted in $\lambda \times N_{\text{bad}}$ rejections (true positive for defect) and $(1 - \lambda) \times N_{\text{bad}}$ approvals (false negative for defect), where $0 \le \lambda \le 1$ is a positive factor called the *sensitivity*, varying in the range $[0, 1]$. By definition,

$$\lambda = \frac{\texttt{True Positives}}{\texttt{True Positives} + \texttt{False Negatives}}. \tag{3.2.4}$$

The sensitivity is also called the true positive rate (TPR).

We say that the inspection station is $100 \times \lambda\%$ sensitive. For example, if $\lambda = 0.97$, the inspection station is $97\%$ sensitive. Unless everyone at the inspection station is asleep, $\lambda$ will be close to $1$.

We note that $\Pi_{(p=0\,|\,q=0)}$ is the probability that a bad toaster failed inspection and was deemed defective, whereas $\Pi_{(p=1\,|\,q=0)}$ is the probability that a bad toaster passed inspection and was deemed functional, and write

$$\begin{aligned} \Pi_{(p=0\,|\,q=0)} &= \lambda, \\ \Pi_{(p=1\,|\,q=0)} &= 1 - \Pi_{(t=0\,|\,q=0)} = 1 - \lambda. \end{aligned} \tag{3.2.5}$$

The sensitivity is

$$\lambda = \frac{\Pi_{(p=0\,|\,q=0)}}{\Pi_{(p=0\,|\,q=0)} + \Pi_{(p=1\,|\,q=0)}}. \tag{3.2.6}$$

### 3.2.5 *Specificity*

Inspection of $N_{\text{good}}$ toasters *that are known to work* $(q = 1)$ has resulted in $\mu \times N_{\text{good}}$ approvals (true negative for defect) and $(1-\mu) \times N_{\text{good}}$ rejections (false positive for defect), where $\mu$ is a positive factor called the *specificity*, varying in the range $[0, 1]$. By definition,

$$\mu = \frac{\texttt{True Negatives}}{\texttt{True Negatives} + \texttt{False Positives}}. \tag{3.2.7}$$

The specificity is also called the true negative rate (TNR).

We say that the inspection station is $100 \times \mu\%$ specific. For example, if $\mu = 0.95$, the inspection station is $95\%$ specific. Since the inspection station is reliable, more or less, $\mu$ will be close to $1.0$.

Since $\Pi_{(p=1\,|\,q=1)}$ is the probability that a good toaster $(q = 1)$ has passed inspection $(p = 1)$, and $\Pi_{(p=0\,|\,q=1)}$ is the probability that a good toaster $(q = 1)$ has failed inspection $(p = 0)$,

$$\begin{aligned} &\Pi_{(p=1\,|\,q=1)} = \mu, \\ &\Pi_{(p=0\,|\,q=1)} = 1 - \Pi_{(p=1\,|\,q=1)} = 1 - \mu. \end{aligned} \tag{3.2.8}$$

The specificity is given by

$$\mu = \frac{\Pi_{(p=1\,|\,q=1)}}{\Pi_{(p=1\,|\,q=1)} + \Pi_{(p=0\,|\,q=1)}}. \tag{3.2.9}$$

### 3.2.6 *Conditional probability matrix*

The four conditional probabilities introduced can be accommodated in a table encapsulated in a $2 \times 2$ matrix,

$$\begin{bmatrix} \Pi_{(p=0\,|\,q=0)} & \Pi_{(p=1\,|\,q=0)} \\ \Pi_{(p=0\,|\,q=1)} & \Pi_{(p=1\,|\,q=1)} \end{bmatrix} = \begin{bmatrix} \lambda & 1-\lambda \\ 1-\mu & \mu \end{bmatrix}. \tag{3.2.10}$$

The sum of the elements in each row is $1.0$. Ideally, this matrix should have $1$ on the diagonals and $0$ on the off-diagonals.

### *3.2.7 Assessing $\lambda$ and $\mu$*

The sensitivity, $\lambda$, and specificity, $\mu$, can be readily assessed. A quality control specialist would send to the inspection station a batch of known bad toasters and another batch of known good toasters, and consider what they have to say.

If $\lambda = 1$ and $\mu = 1$, the testing station is flawless in identifying good or bad products. If $\lambda < 1$ or $\mu < 1$, the testing station is imperfect. If $\lambda = 0$ and $\mu = 0$, the testing station is totally unreliable.

It is possible that $\lambda = \mu$. In the language of information theory, the production line is a *binary symmetric channel.*

### *3.2.8 Probability tree*

In the cafeteria, product quality engineer Bujar drew on a napkin the probability tree shown in Table 3.2.1. By unraveling this tree guided by Bayes' theorem, Bujar was able to assess the probability that *a toaster that has tested bad is truly defective* and the probability that *a toaster that has tested good works fine*, as is now explained.

### *3.2.9 A lemon is not necessarily a lemon*

An arbitrary toaster has been examined at the testing station, was deemed defective, and was given the label $p = 0$ and a sticker with a sad face. However, because of the imperfect performance of the testing station, the toaster is not necessarily a lemon.

We are interested in the probability that a rejected toaster ($p = 0$) is truly defective ($q = 0$), denoted by

$$\Pi_{(q=0\,|\,p=0)}. \tag{3.2.11}$$

Using Bayes' equation (2.2.1), we find that

$$\Pi_{(q=0\,|\,p=0)} = \frac{\Pi_{(p=0\,|\,q=0)}}{\Pi_{(p=0\,|\,q=0 \text{ OR } 1)}} \times \Pi_{(q=0)}. \tag{3.2.12}$$

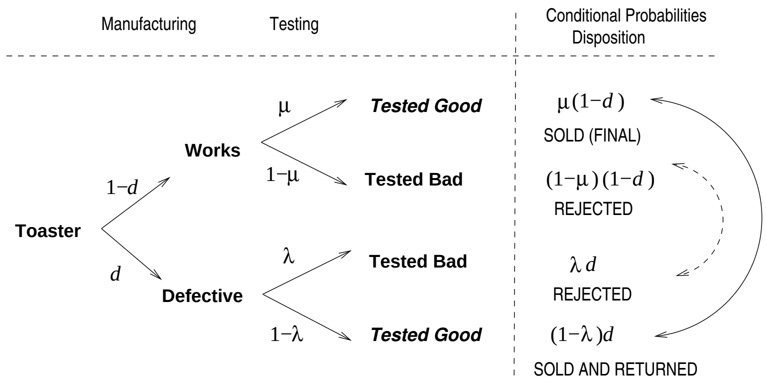



TABLE 3.2.1 Toaster probability tree diagram involving three fractions, $d$, $\lambda$, and $\mu$. The four conditional probabilities shown in the right column add to unity.

The denominator of the fraction on the right-hand side is the marginal probability that a toaster has been rejected at the testing station irrespective of whether the toaster is well made or defective,

$$\Pi_{(p=0\,|\,q=0\text{ OR }1)} = \Pi_{(p=0\,|\,q=0)} \times \Pi_{(q=0)} + \Pi_{(p=0|q=1)} \times \Pi_{(q=1)}. \tag{3.2.13}$$

Making substitutions, we find that

$$\Pi_{(p=0\,|\,q=0\text{ OR }1)} = \lambda \times d + (1-\mu) \times (1-d), \tag{3.2.14}$$

which is the sum of the two innermost conditional probabilities connected with a pointed arc drawn with the broken line in Table 3.2.1.

Consequently,

$$\Pi_{(q=0\,|\,p=0)} = \frac{\lambda \times d}{\lambda \times d + (1-\mu) \times (1-d)}, \tag{3.2.15}$$

which can be rearranged into

$$\Pi_{(q=0\,|\,p=0)} = \frac{1}{1 + \dfrac{1-\mu}{d} \times \dfrac{1-d}{\lambda}}. \tag{3.2.16}$$

This formula provides us with the probability that a rejected toaster is truly defective. Two special cases are of interest:

- When $\mu = 1$, we find that $\Pi_{(q=0\,|\,p=0)} = 1$, independent of $\lambda$ and $d$, as expected. The reason is that the inspection station identifies every single good toaster. If a toaster is well made, it will surely be detected at the inspection station.

- When $d = 1$, we also find that $\Pi_{(q=0\,|\,p=0)} = 1$ independent of $\lambda$ and $\mu$, as expected. The reason is that all toasters are bad and no good toaster (there is none) can be mislabeled as bad.

### *3.2.10 873 toasters*

As a numerical example, we assume that $3$ out of $873$ toasters are true lemons, that is,

$$d = \frac{3}{873} \simeq 0.00344, \tag{3.2.17}$$

as shown in (3.2.1). A fraction $\lambda = 0.97$ of inspections of defective toasters are successful, and a fraction $\mu = 0.95$ of inspections of well-made toasters are successful. Numerical evaluation of formula (3.2.15) shows that

$$\Pi_{(q=0\,|\,p=0)} = \frac{1}{1 + \dfrac{1-0.95}{0.00344} \times \dfrac{1-0.00344}{0.97}} = 0.0627, \tag{3.2.18}$$

and thus

$$\Pi_{(q=1\,|\,p=0)} = 1 - \Pi_{(q=0\,|\,p=0)} = 0.9373, \tag{3.2.19}$$

which is unexpectedly high. Even though the toaster has been deemed defective $(p = 0)$, there is a nearly $94\%$ chance that the toaster is good $(q = 1)$.

These numerical results are surprising to the extent that we may become suspicious that the theory is misleading or flawed and the whole premise of Bayesian analysis amounts to snake oil. In fact, the theory is absolutely correct for a good reason.

### *3.2.11 To measure a weak signal, we need a high-precision instrument*

Consider the second fraction in the denominator on the right-hand side of (3.2.16),

$$\frac{1-d}{\lambda}. \tag{3.2.20}$$

Since the numerator and denominator are close to unity, the fraction is also close to unity.

Now consider the first fraction in the denominator on the right-hand side of (3.2.16),

$$\frac{1-\mu}{d}. \tag{3.2.21}$$

The numerator and denominator are both small. In our numerical example, this fraction is

$$\frac{1-0.95}{0.00344} = \frac{5}{0.344} = 14.535, \tag{3.2.22}$$

which makes the right-hand side of (3.2.16) small. We want

$$d \gg 1-\mu \tag{3.2.23}$$

or

$$\mu \gg 1-d, \tag{3.2.24}$$

so that the fraction is small, that is, *we want the inspection station to be more reliable than the product.* Any reasonable person will agree that this is a fair stipulation.

One corollary is that to measure a weak signal, we need a high-precision instrument. A second corollary is that it is unreasonable to spend a lot of time analyzing or trying to understand the behavior of a person or institution that is known to be erratic.

In Figure 3.2.1, the conditional probability $\Pi_{(q=0\,|\,p=0)}$ is plotted against $d$ and $\mu$ for $\lambda = \mu$. When $d$ is small, $\mu$ must be near $1$ to yield $\Pi_{(q=0\,|\,p=0)} \simeq 1$.

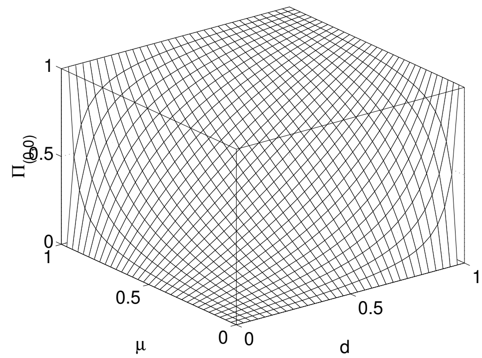



FIGURE 3.2.1 Conditional probability $\Pi_{(0,0)} \equiv \Pi_{(q=0\,|\,p=0)}$ plotted against the fraction of defective toasters, $d$, and correct positives fraction, $\mu$, for $\lambda = \mu$.

### *3.2.12 A good toaster is a good toaster*

A toaster has been tested and deemed to be functional ($p = 1$). Given the imperfect record of the testing station, we are interested in the probability that the toaster is truly functional ($q = 1$),

$$\Pi_{(q=1\,|\,p=1)}. \tag{3.2.25}$$

Using Bayes' equation (2.2.1), we find that

$$\Pi_{(q=1\,|\,p=1)} = \frac{\Pi_{(p=1\,|\,q=1)}}{\Pi_{(p=1\,|\,q=0\text{ OR }1)}} \times \Pi_{(q=1)}. \tag{3.2.26}$$

The denominator of the fraction on the right-hand side is the marginal probability that a toaster has been approved irrespective of whether it is well made or a lemon, given by

$$\Pi_{(p=1\,|\,q=0\text{ OR }1)} = \Pi_{(p=1\,|\,q=0)} \times \Pi_{(q=0)} + \Pi_{(p=1\,|\,q=1)} \times \Pi_{(q=1)}. \tag{3.2.27}$$

Making substitutions, we find that

$$\Pi_{(p=1\,|\,q=0\text{ OR }1)} = (1-\lambda) \times d + \mu \times (1-d), \tag{3.2.28}$$

which is the sum of the two outermost conditional probabilities connected by a pointed arc in the last column of Figure 3.2.1.

Using expressions (3.2.14) and (3.2.28), we confirm that the two marginal probabilities add up to unity, as required,

$$\Pi_{(p=0\,|\,q=0\text{ OR }1)} + \Pi_{(p=1\,|\,q=0\text{ OR }1)} = \lambda\, d + (1-\mu)(1-d) + (1-\lambda)\, d + \mu\,(1-d) = 1. \tag{3.2.29}$$

Each marginal probability involves two cases, and the sum involves four cases which exhaust the available sample space.

Consequently,

$$\Pi_{(q=1\,|\,p=1)} = \frac{\mu \times (1-d)}{(1-\lambda)\times d + \mu \times (1-d)}, \tag{3.2.30}$$

which can be rearranged into

$$\Pi_{(q=1\,|\,p=1)} = \frac{1}{1 + \dfrac{d}{1-d} \times \dfrac{1-\lambda}{\mu}}. \tag{3.2.31}$$

Both fractions in the denominator on the right-hand side are small. When $\lambda = 1$, we find that $\Pi_{(q=1\,|\,p=1)} = 1$, in agreement with intuition. The reason is that a poorly made toaster is always flagged.

### 3.2.13 873 toasters

As a numerical example, we substitute into (3.2.30) the values $d = 3/873$, $\lambda = 0.97$, and $\mu = 0.95$, and find that

$$\Pi_{(q=1\,|\,p=1)} = 0.999891, \tag{3.2.32}$$

and thus

$$\Pi_{(q=0\,|\,p=1)} = 1 - \Pi_{(q=1\,|\,p=1)} = 0.00001, \tag{3.2.33}$$

which is reassuring. A toaster that passed inspection is a good toaster within only a shadow of doubt.

### *3.2.14 Conditional probability matrix*

The four posterior probabilities can be accommodated in a table encapsulated in a $2 \times 2$ matrix,

$$\begin{bmatrix} \Pi_{(q=0\,|\,p=0)} & \Pi_{(q=0\,|\,p=1)} \\ \Pi_{(q=1\,|\,p=0)} & \Pi_{(q=1\,|\,p=1)} \end{bmatrix}. \tag{3.2.34}$$

where

$$\Pi_{(q=0\,|\,p=1)} = 1 - \Pi_{(q=1\,|\,p=1)} \tag{3.2.35}$$

and

$$\Pi_{(q=1\,|\,p=0)} = 1 - \Pi_{(q=0\,|\,p=0)}. \tag{3.2.36}$$

The sum of the elements in each column of this matrix is $1.0$. If the testing station is flawless, this matrix is the identity matrix, that is, it has ones on the diagonal and zeros on the off-diagonal positions.

### *3.2.15 Estimating the fraction of defective toasters*

The probability $\Pi_{(q=0\,|\,p=0)}$ can be measured by collecting a group of rejected toasters ($p = 0$) and counting how many of them do not work ($q = 0$) by attempting to make a toast. Expression (3.2.15) may then be inverted to estimate the fraction of truly defective toasters,

$$d = \frac{(1-\mu) \times \Pi_{(q=0\,|\,p=0)}}{\lambda + (1-\mu-\lambda) \times \Pi_{(q=0\,|\,p=0)}}. \tag{3.2.37}$$

We observe that, if $\Pi_{(q=0\,|\,p=0)} = 0$, all toasters are deemed functional, and thus $d = 0$; whereas if $\Pi_{q=0,p=0} = 1$, all toasters are deemed defective, and thus $d = 1$.

The probability $\Pi_{(q=1\,|\,p=1)}$ can be measured by collecting a group of approved toasters ($p = 1$) and counting how many of them work by attempting to make a toast ($q = 1$). Expression (3.2.30) may then be inverted to infer the fraction of truly defective toasters,

$$d = \mu \times \frac{1 - \Pi_{(q=1\,|\,p=1)}}{\mu + (1-\lambda-\mu) \times \Pi_{(q=1\,|\,p=1)}}. \tag{3.2.38}$$

This expression arises from (3.2.37) by replacing (*a*) $d$ with $1-d$, (*b*) $\Pi_{(q=0\,|\,p=0)}$ with $\Pi_{(q=1\,|\,p=1)}$, (*c*) $\mu$ with $\lambda$, and (*d*) $\lambda$ with $\mu$.

If we ask the testing station to provide us with the fraction of defective toasters, they will suggest the value

$$d_{\text{reported}} = \Pi_{(p=0\,|\,q=0 \text{ OR } 1)} = \lambda \times d + (1-\mu) \times (1-d). \quad (3.2.39)$$

Simplifying, we obtain

$$d_{\text{reported}} = 1 - \mu - d \times (1 - \lambda - \mu), \quad (3.2.40)$$

which is equal to $d$ only if

$$d = \frac{1-\mu}{2-\lambda-\mu}. \quad (3.2.41)$$

For an arbitrary value of $d$, this is true only when $\lambda = 1$ and $\mu = 1$. When $\lambda = \mu$, this is true when $d = \frac{1}{2}$.

### *3.2.16 Into the dumpster*

The disposition of manufactured toasters is indicated in the last column of Figure 3.2.1. A good toaster may be sold to a satisfied customer or falsely rejected and tossed into a dumpster if erroneously deemed defective. A defective toaster may be tossed into the dumpster if correctly deemed bad, or returned by an unhappy customer and tossed into the dumpster.

The fraction of manufactured toasters that escape the dumpster is $\mu\,(1-d)$, which is equal to unity only if $d = 0$ and $\mu = 1$.

***Exercise***

**3.2.1** Reproduce the graph shown in Figure 3.2.2 for $\lambda = \frac{1}{10}\,\mu$ and discuss the differences with the case $\lambda = \mu$.

## 3.3 Hope

Hope has been working as a customer service representative at *Flying Toasters, LLC* for several years. In her private life, Hope is a recluse, living with her beloved cat and two hamsters.

Hope realized early on that it is impossible to ridicule a cat, and contemplated the reasons as to why. The cat taught Hope a few things: (*a*) be apprehensive of those who you don't know well, (*b*) trust anyone only after an extended period of time, and (*c*) meow relentlessly until a desired goal has been achieved.

### 3.3.1 Kind and unkind

Of the $N_{\text{unkind}}$ unkind people (defective toasters) Hope has met in the past twenty years, she had a feeling that $\lambda \times N_{\text{unkind}}$ of them were selfish and unkind (true positive for unkind), and $(1-\lambda) \times N_{\text{unkind}}$ of them of them were kind (false negative for unkind).

Of the $N_{\text{kind}}$ kind people (good toasters) Hope has met in the past twenty years, she had a feeling that $\mu \times N_{\text{kind}}$ of them were kind (true negative for unkind) and $(1-\mu) \times N_{\text{kind}}$ were unkind (false positive for unkind).

By recounting all the people (toasters in a store) that she has met in the past five years, Hope made the rough estimates $\lambda = 0.90$ and $\mu = 0.60$. It seems that Hope has good judgment and the ability to identify the kind and the unkind.

### 3.3.2 Feeling blue

On days that Hope feels blue and frustrated with aggressive customers, she thinks that only $20\%$ of all people are kind and the rest of them are indifferent or unkind. The fraction of unkind people is rather large, $d = 0.80$.

Now referring to Bayes equation (3.2.31), repeated here for easy reference,

$$\Pi_{(q=1\,|\,p=1)} = \frac{\mu \times (1-d)}{(1-\lambda) \times d + \mu \times (1-d)}, \tag{3.3.1}$$

we substitute the values $d = 0.80$, $\lambda = 0.90$, and $\mu = 0.60$, and find that

$$\Pi_{(q=1\,|\,p=1)} = 0.60. \tag{3.3.2}$$

This result shows that, if Hope has assessed that a particular person is kind ($p = 1$), the probability that the person is truly kind ($q = 1$) is only $60\%$.

***Exercise***

**3.3.1** On days that Hope feels optimistic and cheerful, she thinks that $50\%$ of all people are kind and the rest of them are either indifferent or unkind. What is the probability that a person that Hope assessed to be unkind ($p = 0$) is truly unkind ($q = 0$)?

## 3.4 Deer tick bites

A small percentage of the population in the United States, approximately $1\%$, suffer from lyme disease due to deer tick bites. This percentage corresponds to population fraction $d = 0.01$, where $d$ stands for disease. Fortunately, the disease can be treated or prevented at an early stage with a single high dosage or a prolonged course of antibiotics.

A person decides to get tested by way of a blood sample. A phlebotomist is visited and a blood sample is taken and sent to a testing laboratory. In medical diagnostics, *positive testing* is bad news, whereas *negative testing* is good news.

### 3.4.1 Sensitive and specific

The testing laboratory is not totally reliable for reasons that are beyond their control. To carry out the Bayesian analysis, we introduce a Boolean variable to indicate the presence of disease, $q$, such that the value $q = 0$ indicates an unhealthy sample (defective toaster) and the value $q = 1$ indicates an healthy sample (functioning toaster).

Moreover, we introduce another Boolean variable for passing the test, $p$, where:

• The value $p = 0$ indicates that the result of a test is positive because the sample contains a pathogen (did not pass the test).

• The value $p = 1$ indicates that the result of a test is negative because the sample does not contain a pathogen (passed the test).

A fraction $\lambda = 0.97$ of unhealthy samples ($q = 0$) are correctly tested positive ($p = 0$) and are classified as *true positive* for disease. The rest of the samples are incorrectly tested negative ($p = 1$) and are classified as *false negative* for disease. We say that the test is $\lambda \times 100\%$ sensitive.

A fraction $\mu = 0.95$ of healthy samples ($q = 1$) are correctly tested negative ($p = 1$) and are correctly classified as *true negative* for disease. The rest of the samples are incorrectly tested positive ($p = 0$) and are classified as *false positive* for disease. We say that the test is $\mu \times 100\%$ specific.

By repeating the analysis of Section 3.2 for toasters, we derive equation (3.2.15), repeated below for convenience,

$$\Pi_{(q=0\,|\,p=0)} = \frac{\lambda \times d}{\lambda \times d + (1-\mu) \times (1-d)}. \tag{3.4.1}$$

This formula provides us with the probability that a person whose test came back positive ($p = 0$) carries the disease ($q = 0$) and should seek treatment.

Plugging in the numbers, we find that

$$\Pi_{(q=0\,|\,p=0)} = \frac{0.97 \times 0.01}{0.97 \times 0.01 + (1-0.95) \times (1-0.01)} = 0.164, \tag{3.4.2}$$

which indicates that there is only a $16.4\%$ chance that a person whose test came back positive carries the disease. Although this is higher than the generic probability of $1.0\%$, it is still small. The tested person most likely does not carry the disease.

The underlying reason is that the fraction of successful detection of the laboratory, $1-\mu$, is lower than the fraction of unhealthy individuals, $d$, as discussed in Section 3.4.2.

*Exercise*

**3.4.1** Repeat the Bayesian analysis for $d = 0.10$ and discuss the results.

## 3.5 Gambler's fallacy

A person flips a coin whose sides are labeled $\mathrm{A}$ or $\mathrm{B}$. In the context of probability, each coin flip is called a *Bernoulli trial.*

The coin may be rigged so that the probability that it shows face $\mathrm{A}$ is $p$ and the probability that it shows face $\mathrm{B}$ is $1-p$, where $p$ is the *Bernoulli probability* taking values in the range $[0, 1]$. This means that, if the coin is flipped $N$ times, approximately $pN$ times will show face $\mathrm{A}$ and approximately $(1-p)N$ times will show face $\mathrm{B}$ if $N$ is sufficiently high. In the case of a fair coin, $p = \frac{1}{2}$.

Assume that the coin has shown face $\mathrm{A}$ in the first two trials. What is the probability that the coin will also show face $\mathrm{A}$ in the third trial? This probability is exactly equal to $p$, as it if the first two trials had not taken place. Those who argue otherwise fall prey to the gambler's fallacy.

### 3.5.1 Gambler's fallacy formally dismissed

To demonstrate the fallacy in no uncertain terms, we apply (2.15.9) to obtain an expression for the conditional probability of interest,

$$\Pi_{(\mathrm{A}_3\,|\,(\mathrm{A}_1,\mathrm{A}_2))} = \frac{\Pi_{((\mathrm{A}_1,\mathrm{A}_2)\,|\,\mathrm{A}_3)}}{\Pi_{((\mathrm{A}_1,\mathrm{A}_2)\,|\,(\mathrm{A}_3\ \mathrm{OR}\ \mathrm{B}_3))}} \times \Pi_{(\mathrm{A}_3)} \tag{3.5.1}$$

where $\Pi_{(\mathrm{A}_3\,|\,(\mathrm{A}_1,\mathrm{A}_2))}$ on the left-hand side is the conditional probability that the third outcome is $\mathrm{A}$, given that first and second outcomes are also $\mathrm{A}$, subject to the following definitions:

- $\Pi_{(A_3)}$ at the far end of the right-hand side is the unconditional probability that the third outcome is $A$,

$$\Pi_{(A_3)} = p. \tag{3.5.2}$$

- $\Pi_{((A_1,A_2)\,|\,A_3)}$ in the numerator of the fraction on the right-hand side is the conditional probability that the first and second outcomes are $A$, given that the third outcome is also $A$,

$$\Pi_{((A_1,A_2)\,|\,A_3)} = p \times p = p^2. \tag{3.5.3}$$

- The denominator of the fraction on the right-hand side is the probability that the first and second outcomes are $A$, given that the third outcome is $A$ or $B$, given by

$$\Pi_{((A_1,A_2)\,|\,(A_3\,\mathrm{OR}\,B_3))} = p^2 \times p + p^2 \times (1-p) = p^2. \tag{3.5.4}$$

Making substitutions, we find that

$$\Pi_{(A_3\,|\,A_1,A_2)} = p, \tag{3.5.5}$$

which proves our assertion beyond doubt.

### *3.5.2 Of manholes and drunk driving*

Gambler's fallacy is teaching us that, if we walk blindfolded down a street with uncovered manholes, and if we were lucky enough not to fall through the first two manholes, nothing guarantees that we will not fall through the third manhole. If we committed ten moving violations and were not caught, we do not have a higher chance of not being caught the next time.

More important, if we did not get caught drunk driving forty times, we should not think that we will not be caught driving drunk the forty-first time. In fact, the probability of being caught at any time is the equal to ratio of the number of daily DUI offense citations to the number of daily drivers, which has nothing to do with the number of our own previous infractions.

### 3.5.3 Correlated events

Gambler's fallacy applies to situations where the repeated trials are uncorrelated, that is, they occur as though previous ones have not taken place. For example, we you pull two socks from a chest drawer in the dark, the probability of finding a matching pair remains the same each time we make a draw. The fallacy fails if you keep one of the two unmatched socks and toss the second one back into the drawer.

#### *Exercise*

**2.5.1** Discuss further the example of matching socks with two pairs of socks in the drawer.

## 3.6 Idig

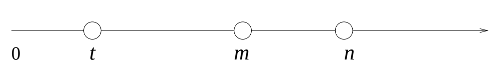
Agricultural tractor manufacturer *Idig* produced a successful model of tractors for just one year in 1965, and was then acquired and absorbed by a larger manufacturer with a better name recognition. Fortunately, all employee jobs were spared.

A tractor sales, service, and repair shop in rural Ohio has serviced $t$ tractors and recorded the serial numbers in the invoices. Master mechanic Jeremiah noticed that the maximum serial number was $m$ and wants to estimate the number of tractors that *Idig* has manufactured, denoted by $n$.

Common sense suggests that

$$m \geq t, \quad n \geq t, \quad n \geq m, \tag{3.6.1}$$

or $n \geq \max(t, m)$, in the following order:




In the context of probability, Jeremiah wants to estimate the conditional probability

$$\Pi_{(n\,|\,(m,t))}, \tag{3.6.2}$$

that is, the probability that $n$ tractors were manufactured given the maximum serial number, $m$, and number of units serviced, $t$. The estimation can be done using Bayes' rule for multiple events.

You can visit a parking lot, count the number of Škodas that you see, $t$, and record their serial numbers, $m$. The following analysis will help you estimate the number of Škodas manufactured.

### 3.6.1 Bayes rule

According to Bayes' formula (2.15.13) involving three events, repeated below for convenience,

$$\Pi_{(\circ \,|\, (\oplus\,,\,\square))} = \frac{\Pi_{(\oplus \,|\, (\circ,\square))}}{\Pi_{(\oplus \,|\, \square)}} \times \Pi_{(\circ \,|\, \square)}, \tag{3.6.3}$$

the desired probability is given by

$$\Pi_{(n \,|\, (m,t))} = \frac{\Pi_{(m \,|\, (n,t))}}{\Pi_{(m \,|\, t)}} \times \Pi_{(n \,|\, t)}, \tag{3.6.4}$$

where $m$ and $t$ are provided in Jeremiah's records. The data are encapsulated in the doublet $(m, t)$.

### 3.6.2 The prior

The conditional probability that $n$ tractors have been manufactured given that $t$ tractors were serviced, $\Pi_{(n \,|\, t)}$, plays the role of a prior.

To provide a reasonable estimate for this probability, we can impose a maximum on the number of units that could have possibly be manufactured taking into consideration time and material constraints, $n_{\max} \geq t$, and assume a uniform probability distribution,

$$\Pi_{(n \,|\, t)} = \frac{1}{n_{\max} - t + 1} \tag{3.6.5}$$

for $n = t, \ldots, n_{\max}$. One may confirm that

$$\sum_{n=t}^{n_{\max}} \Pi_{(n \,|\, t)} = 1, \tag{3.6.6}$$

as required. After all is said and done, we can take the theoretical limit $n_{\max} \to \infty$.

### 3.6.3 Likelihood

The conditional probability that a certain maximum serial number $m$ was observed if $t \le m$ units were serviced and $n \ge m$ units were manufactured, $\Pi_{(m\,|\,n,t)}$, plays the role of the likelihood.

If only one tractor was serviced, $t = 1$, the serial number observed can be any integer in the range $1, \ldots, n$, with equal probability,

$$\Pi_{(m\,|\,(n,1))} = \frac{1}{n} \tag{3.6.7}$$

for $m = 1, \ldots, n$. To assess the likelihood for a higher number of units serviced, $t > 1$, we resort to combinatorics.

### 3.6.4 Likelihood from the binomial coefficient

Assume that two tractors were serviced, $t = 2$. The number of possible unordered pairs of serial numbers is equal to the binomial coefficient, $\mathcal{C}_2^n$, as discussed in Appendix A. The probability of each pair is $1/\mathcal{C}_2^n$, and the number of pairs involving a particular serial number, $k$, is $k-1$.

Identifying $k$ with the maximum serial number observed, $m$, we obtain

$$\Pi_{(m\,|\,(n,2))} = \frac{m-1}{\mathcal{C}_2^n} \tag{3.6.8}$$

for $m = 1, \ldots, n$, provided that $m \ge 2$.

Now we assume that an arbitrary number of $t$ tractors have been serviced. The number of possible unordered $t$-tuples of serial numbers is equal to the binomial coefficient $\mathcal{C}_t^n$. The probability of each $s$-tuple is $1/\mathcal{C}_t^n$, and the number of $t$-tuples involving a particular serial number, $k$, is $\mathcal{C}_{t-1}^{k-1}$.

Identifying $k$ with the maximum serial number observed, $m$, we obtain

$$\Pi_{(m\,|\,(n,t))} = \frac{\mathcal{C}_{t-1}^{m-1}}{\mathcal{C}_t^n} \tag{3.6.9}$$

for $m = 1, \ldots, n$, provided that $m \geq t$.

We have derived an expression for the likelihood in terms of the binomial coefficient.

### *3.6.5 Marginal probability*

The marginal probability in the denominator on the right-hand side of Bayes's formula (3.6.4) is given by

$$\Pi_{(m\,|\,t)} = \sum_{n=m}^{n_{\max}} \Pi_{(m\,|\,(n,t))} \times \Pi_{(n\,|\,t)} \qquad (3.6.10)$$

for $m \geq t$. Note that summation begins at $n = m$. Substituting expressions (3.6.5) and (3.6.9), we find that

$$\Pi_{(m\,|\,t)} = \sum_{n=m}^{n_{\max}} \Big(\frac{\mathcal{C}_{t-1}^{m-1}}{\mathcal{C}_t^n} \times \frac{1}{n_{\max} - t + 1}\Big) \qquad (3.6.11)$$

for $m \geq t$.

The following Matlab code named *idig* performs the computations:

```
nmax = 200;  % specified maximum number of units
t = 5;  % number of tractors serviced
m = 40; % max serial number observed
margie = 0.0;

for n=m:nmax
   post(n) = binomial_coeff(m-1,t-1) ...
            /binomial_coeff(n,t)/(nmax-t+1);
   margie = margie + post(n);
end

for n=m:nmax
  post(n) = post(n)/margie;
  nplot(n) = n;
end

figure(1)
```

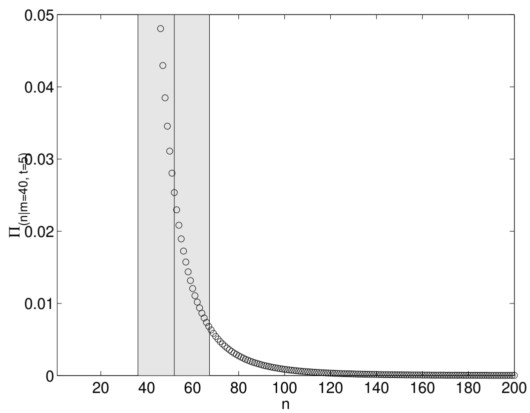



FIGURE 3.6.1 Probability distribution of tractors manufactured by *Idig* for $t = 5$ units serviced and $m = 40$ maximum serial number observed. The shaded region is centered at the expected value and extends over one standard deviation on either side.

```
plot(nplot(m:nmax),post(m:nmax),'ko')
xlabel('n','fontsize',15)
ylabel('\Pi_{(n | m=40, t=5)}','fontsize',15)
```

The code calls the function *binomial_coeff* listed in Appendix A to compute the binomial coefficient. The graphics display generated by the code is shown in Figure 3.6.1.

A detailed calculation shows that, in the limit as $n_{\max}$ tends to infinity, the expected value of $n$ is

$$\overline{n} \equiv \sum_n n \, \Pi_{(n \,|\, m,t)} = \frac{(t-1)(m-1)}{t-2}, \tag{3.6.12}$$

and the variance is

$$s^2 \equiv \sum_n (n-\overline{n})^2 \, \Pi_{(n \,|\, m,t)} = \frac{(t-1)(m-1)(m-t+1)}{(t-3)(t-2)^2} \tag{3.6.13}$$

(https://en.wikipedia.org/wiki/German_tank_problem), where $s$ is the standard deviation. The number of tractors produced by *Idig* is expected to fall inside the shaded area in Figure 3.6.1.

***Exercise***

**2.6.1** Jeremiah calls his friend Jonas at another repair shop and is provided with the data $t = 3$ and $m = 60$. Revise the Bayesian analysis to include these data.

## 3.7 Hierarchical ordering

Consider the product decomposition of the joint probability of three arbitrary events, denoted by $\circ$, $\oplus$, and $\square$, as shown in equation (2.15.5), repeated below for convenience,

$$\Pi_{(\circ\,,\,\oplus\,,\,\square)} = \Pi_{(\circ\,|\,(\oplus\,,\,\square))} \times \Pi_{(\oplus\,|\,\square)} \times \Pi_{(\square)}. \tag{3.7.1}$$

It is possible that the order of the events can be arranged so that

$$\Pi_{(\circ\,|\,(\oplus\,,\,\square))} \simeq \Pi_{(\circ\,|\,\oplus)}, \tag{3.7.2}$$

and thus

$$\Pi_{(\circ\,,\,\oplus\,,\,\square)} \simeq \Pi_{(\circ\,|\,\oplus)} \times \Pi_{(\oplus\,|\,\square)} \times \Pi_{(\square)}, \tag{3.7.3}$$

which is reminiscent of the chain rule. The approximation (3.7.2) assumes that the joint probability of $\circ$ is affected primarily by $\oplus$ and to a lesser degree by $\square$.

### 3.7.1 Red Corvettes

For example, if red Corvettes are in demand, then the probability that a Corvette is sold ($\circ$) depends primarily on it being red ($\oplus$), and to a lesser degree on whether a cigarette lighter is available in the central console ($\square$).

### 3.7.2 Reverse engineering a recipe

Celebrity cooking judge Bowtie tastes a spoonful of hot salsa and wants to assess of probability that habanero peppers ($\circ$), crushed red pepper ($\oplus$), and cilantro ($\square$) were used in the ingredients.

The good judge knows by experience that recipes that call for crushed red pepper almost always call for habanero peppers, no matter how much cilantro was also used, if any, and sets

$$\Pi_{(\circ \,|\, (\oplus \,,\, \square))} \simeq \Pi_{(\circ \,|\, \oplus)} = 0.8. \tag{3.7.4}$$

Nearly half of all recipes that call for cilantro also call for crushed red pepper,

$$\Pi_{(\oplus \,|\, \square)} = 0.5. \tag{3.7.5}$$

About $90\%$ of all recipes call for cilantro,

$$\Pi_{(\square)} = 0.90. \tag{3.7.6}$$

The good judge computes

$$\Pi_{(\circ \,,\, \oplus \,,\, \square)} \simeq 0.8 \times 0.5 \times 0.9 = 0.36. \tag{3.7.7}$$

It seems unlikely that the recipe uses all three ingredients.

### *3.7.3 Growing older*

The judge reveals his prediction to the cook who is extremely impressed by the judge's educated guess. In fact, the judge's guess is the result of life-time experiences. The judge likes to say tongue-in-cheek that experiences are a poor substitute for ideals. I agree.

Unfortunately or fortunately, accumulated experience and wisdom are linear functions of biological time elapsed, possibly exhibiting a maximum at a certain stage. This is a main reason as to why some people do not mind growing older but look forward to gaining further wisdom in spite of declining hearing and eyesight. I am not one of these people.

### *3.7.4 Multitude of events*

In the case of an arbitrary number of $N$ events, hierarchical ordering amounts to setting

$$\Pi_{(\circ_1 \,,\, \dots \,,\, \circ_N)} \simeq \Pi_{(\circ_1 \,|\, \circ_2)} \times \Pi_{(\circ_2 \,|\, \circ_3)} \times \cdots \times \Pi_{(\circ_{N-1} \,|\, \circ_N)} \times \Pi_{(\circ_N)}. \tag{3.7.8}$$

The wise ordering of the events is or paramount importance. In practice, this approximation is tolerated when the joint probability is regarded as an educated prior.

### *3.7.5 The best academics make the best administrators*

Hierarchical approximations are not always possible. The usefulness of a hierarchical approximation, when appropriate, hinges on the observation that the conditional probabilities $\Pi_{(\circ_i \,|\, \circ_{i+1})}$ can be borrowed from different, though reasonably similar, settings.

Corroborative support is provided by the well-known but carefully concealed fact that the best scholars make the best college administrators. In the context of statistics,

$$\Pi_{(\text{good administrator}\,|\,(\text{good academic}\,,\,\text{smooth talker}))} \simeq \Pi_{(\text{good administrator}\,|\,\text{good academic})}. \tag{3.7.9}$$

Being a smooth talker and touting one's horn are truly irrelevant in substance. An essential quality is the ability to grasp and evaluate subjects, situations, and circumstances on a short time scale and in a broad range of contexts.

### *3.7.6 Be careful who you appoint*

These and other similar realizations are a key to the appointment of effective leaders in corporations, communities, and institutions. History has shown that a single person alone can inflict a great deal of damage to a corporation, community, or institution.

#### *Exercise*

3.7.1 Discuss an example where the hierarchical approximation is appropriate.

## *3.8 Heavy Equipment, Inc.*

The main copy machine of *Heavy Equipment, Inc.* has become unreliable and needs to be repaired. Roughly two out of three photocopies roll out good, but the third copy rolls out blurry, streaky, or blotchy. The machine's rate of success is $66\%$, signified by a parameter ,

$$\mu = 0.66. \tag{3.8.1}$$

If $\mu = 1$, every copy would come out good, and if $\mu = 0$, every copy would come out bad.

### *3.8.1 The boss and the receptionist*

The boss gets exasperated and asks that a technician be called to repair the machine. The receptionist makes the call. The repair company promises to send a technician the very same evening after hours. The dispatcher provides assurances that the repair will improve the machine's rate of success to $90\%$, signified by the parameter

$$\lambda = 0.90. \tag{3.8.2}$$

This means that nine out of ten copies will be good.

### *3.8.2 At 7:00 am*

The boss shows up bright and early in the morning and wonders whether the technician had shown up and the machine had been fixed. In the absence of information, she is thinking there is a fifty-fifty chance that the technician had showed up

$$s = 0.5, \tag{3.8.3}$$

where $s = 1.0$ signifies with absolute certainty that the technician showed up and $s = 0$ signifies with absolute certainty that the technician did not show up. The pertinent probability tree is shown in Figure 3.8.1.

### *3.8.3 $r$ and $q$*

It is convenient to introduce a Boolean variable, $r$, so that $r = 1$ if the repair has been performed and $r = 0$ is a repair has not been performed. Moreover, we introduce another Boolean variable, $q$, so that $q = 1$ if a copy comes out good and $q = 0$ if a copy comes out blurry.

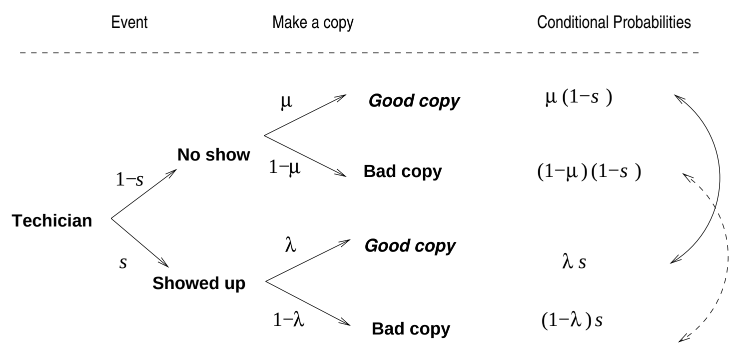



FIGURE 3.8.1 Photocopy probability tree involving three fractions, $s$, $\mu$, and $\lambda$. The conditional probabilities shown in the right column add to unity.

### *3.8.4 First photocopy*

Let's make a copy and see what happens. The boss pushes the button, and the copy is good. According to Bayes' theorem, given that the copy is good, $q = 1$, the probability that the repair has been performed, $r = 1$, is given by the conditional probability

$$\Pi_{(r=1\,|\,q=1)} = \frac{\Pi_{(q=1\,|\,r=1)}}{\Pi_{(q=1)}} \times \Pi_{(r=1)}, \tag{3.8.4}$$

where we recall that $\Pi_{(r=1)} = s$. The denominator of the fraction on the right-hand side is the marginal probability that a copy is good, irrespective of whether the technician has shown up, given by

$$\Pi_{(q=1)} = \Pi_{(q=1\,|\,r=0)} \times \Pi_{(r=0)} + \Pi_{(q=1\,|\,r=1)} \times \Pi_{(r=1)}. \tag{3.8.5}$$

Making substitutions, we find the expression

$$\Pi_{(q=1)} = \mu\,(1-s) + \lambda\,s. \tag{3.8.6}$$

Consequently,

$$\Pi_{(r=1\,|\,q=1)} = \frac{\lambda \times s}{\mu \times (1-s) + \lambda \times s}. \tag{3.8.7}$$

Substituting into equation (3.8.7) the values $s = 0.5$, $\mu = 0.66$, and $\lambda = 0.90$, the boss obtains

$$\Pi_{(r=1\,|\,q=1)} = 0.577, \tag{3.8.8}$$

which is encouraging. The boss revises her initial estimate from $s = 0.5$ to

$$s = 0.577. \tag{3.8.9}$$

Based on this estimate, the boss tends to believe that a repair has been performed.

### 3.8.5 Second photocopy

The boss needs further confirmation. She pushes the button once again and a second good copy comes out. Substituting into equation (3.8.7) the values $s = 0.577$, $\mu = 0.66$, and $\lambda = 0.90$, she obtains

$$\Pi_{(r=1\,|\,q=1)} = 0.650. \tag{3.8.10}$$

The boss happily revises her current estimate to

$$s = 0.650. \tag{3.8.11}$$

The boss is almost certain that the repair has been performed.

### 3.8.6 Third photocopy

The boss decides to push the button for a third time. Unfortunately, a bad copy with a mocking circular smudge at the top of the page comes out. The boss regrets that she pushed the button for the third time.

According to Bayes' theorem, given that the copy is bad, $q = 0$, the probability that the repair has been performed, $r = 1$, is given by the conditional probability

$$\Pi_{(r=1\,|\,q=0)} = \frac{(1-\lambda)\times s}{(1-\mu)\times(1-s) + (1-\lambda)\times s}. \tag{3.8.12}$$

Substituting into this equation $s = 0.650$, $\mu = 0.66$, and $\lambda = 0.90$, we obtain

$$\Pi_{(r=1\,|\,q=0)} = 0.354. \tag{3.8.13}$$

It is nearly certain that the technician has not shown up. The boss heads for the coffee machine in a state of visible distress.

### *3.8.7 Bayesian analysis*

The boss has examined the probability of two states (copier broken or copier fixed) guided by the data (quality of a photocopy.) As data came in, she revised her conclusions. This viewpoint is the cornerstone of the Bayesian approach.

### *3.8.8 Alejandro*

The boss is a good boss. She is patient, straightforward, ethical, truthful, respectful of others, never yells, always gives credit where credit is due, and speaks softly but carries a big stick. Someone once remarked that most fictional heros of novels and television shows share these attributes.

In one staff meeting, engineer Alejandro spent an entire twenty minutes complaining about the steep helical pitch of some replacement carriage bolts, as he usually does about one thing or another. The boss graciously let him ramble while working on her mental to-do list.

In the minutes of the staff meeting, the boss wrote: *Engineer Alejandro does not approve of the new carriage bolts.* When the minutes were handed out in the next staff meeting, everyone laughed, including Alejandro who is a good-natured yet tightly wound colleague.

### *3.8.9 A second set of copies*

Engineer Gevorg sees the boss in a state of distress and wants to help. Hoping for good luck, he is headed for the machine and makes three copies.

Unfortunately, the first copy comes out bad, while the second and third copies come out good. Starting with $s = 0.5$ and using (3.8.12) for the first copy and (3.8.7) for the second and third copies, Gevorg computes the $s$ sequence

$$0.500, \quad 0.227, \quad 0.286, \quad 0.354. \tag{3.8.14}$$

The final number is consistent with that of the boss. It seems that it does not matter whether the bad copy comes out first or last.

### 3.8.10 A third set of copies

Gevorg is not ready to give up and pushes the bottom three more times. The first copy comes out good, the second copy comes out bad, and the third copy comes out good. Starting with $s = 0.5$ and using (3.8.12) for the first and third copies and (3.8.7) for the second copy, Gevorg computes the $s$ sequence

$$0.500, \quad 0.577, \quad 0.286, \quad 0.354. \tag{3.8.15}$$

The final number is still the same, it does not matter if the bad copy comes out first or last.

Gevorg is intrigued that the sequence of good and bad copies is immaterial. The probability of show, $s$, is the same, so long as there are two good copies and one bad copy. Gevorg gets in a pensive mood and heads for the water cooler.

### 3.8.11 Back-of-the-envelop calculations

Gevorg sits down in the lunch room and considers the boss's good-good-bad sequence. He expresses the third estimate of $s$ (denoted by $s_3$) in terms of the second estimate of $s$ (denoted by $s_2$), and the second estimate in terms of the first estimate (denoted by $s_1$), using the familiar equations

$$s_3 = \frac{(1-\lambda) \times s_2}{(1-\mu) \times (1-s_2) + (1-\lambda) \times s_2}, \tag{3.8.16}$$

$$s_2 = \frac{\lambda s_1}{\mu \times (1-s_1) + \lambda \times s_1}, \tag{3.8.17}$$

$$s_1 = \frac{\lambda s_0}{\mu \times (1-s_0) + \lambda \times s_0}, \tag{3.8.18}$$

where $s_0$ is the initial guess before any photocopies have been made.

Back substitution amounts to substituting the expression for $s_1$ given in (3.8.18) into (3.8.17), and then substituting the resulting expression into (3.8.16). After simplifications, we obtain

$$s_3 = \frac{\lambda \times (1-\lambda) \times s_1}{\mu \times (1-\mu) \times (1-s_1) + \lambda \times (1-\lambda) \times s_1}, \tag{3.8.19}$$

and then

$$s_3 = \frac{\lambda^2 \times (1-\lambda) \times s_0}{\mu^2 \times (1-\mu) \times (1-s_0) + \lambda^2 \times (1-\lambda) \times s_0}. \tag{3.8.20}$$

Substituting into (3.8.20) $s_0 = 0.5$, $\mu = 0.66$, and $\lambda = 0.90$, we obtain the familiar value $s_3 = 0.354$. Gevorg briefs the boss on his findings.

### *3.8.12 At lunch*

The boss and Gevorg get together for brown-bag lunch involving tuna and egg sandwiches, and discover by example that, if $n$ photocopies were made, and if $m$ of these photocopies were good and the remaining $n-m$ photocopies where bad, then the probability that the technician has shown up is given by the formula

$$s_m^n = \frac{\lambda^m \times (1-\lambda)^{n-m}}{\mu^m \times (1-\mu)^{n-m} \times (1-s_0) + \lambda^m \times (1-\lambda)^{n-m} \times s_0} \times s_0, \tag{3.8.21}$$

where $s_0$ is the initial guess. For $n = 3$ and $m = 2$, we recover the formula for $s_3$ shown in (3.8.20).

In fact, this formula can be confirmed by mathematical induction. Using Bayes' theorem, we find that

$$s_{m+1}^{n+1} = \frac{\lambda}{\mu \times (1-s_m^n) + \lambda \times s_m^n} \times s_m^n \tag{3.8.22}$$

and

$$s_m^{n+1} = \frac{(1-\lambda)}{(1-\mu) \times (1-s_m^n) + (1-\lambda) \times s_m^n} \times s_m^n, \tag{3.8.23}$$

which are consistent with (3.8.21).

The order by which the sequence of photocopies turned out good or bad is confirmed to be immaterial. Gevorg's formula will be derived formally in the next few sections following a brief summary of combinatorics.

### 3.8.13 Goofies

Early in the afternoon, the copy machine technician shows up with a red toolbox and a pink lunch box decorated with Goofies. He profusely apologizes that he could not show up the previous evening due to a flat tire. Because the point of puncture was on the wall of the tire, the tire could not be repaired and had to be replaced.

The service station advisor tried to convince him that all four tires must be replaced, but the technician knew better than that and was well aware that the shop was run by con artists. The technician gets to work and the copy machine is repaired in no time.

#### *Exercise*

**3.8.1** Compute the counterpart of the quadruplet shown in (3.8.15) with initial estimate $s = 0.80$ and discuss the results.

## 3.9 The binomial distribution

The boss of *Heavy Equipment, Inc* is intrigued by formula (3.8.21), repeated below for convenience,

$$s_m^n = \frac{\lambda^m \times (1-\lambda)^{n-m}}{\mu^m (1-\mu)^{n-m} \times (1-s_0) + \lambda^m (1-\lambda)^{n-m} \times s_0} \times s_0. \tag{3.9.1}$$

We recall that $s_0$ is the prior probability that the technician showed up, $1 - s_0$ is the prior probability that the technician did not show up, $s_m^n$ is the posterior probability that the technician showed up and $1 - s_m^n$ is the posterior probability that the technician did not show up.

### 3.9.1 Bayes equation

The boss recalls that the receptionist holds a Master's degree in applied mathematics and a Master's degree in English literature, and turns to her for insights. The receptionist points out that formula (3.9.1) is an instance of Bayes' equation written in the generic form

$$\Pi_{(\text{event}\,|\,\text{data})} = \frac{\Pi_{(\text{data}\,|\,\text{event})}}{\Pi_{(\text{data}\,|\,\text{any event})}} \times \Pi_{(\text{event})}, \tag{3.9.2}$$

subject to the following substitutions:

event $\rightarrow$ the technician showed up
any event $\rightarrow$ the technician showed up or did not show up
data $\rightarrow$ of $n$ photocopies made, $m$ were good

By definition, $\Pi_{(\text{event})} = s_0$ on the right-hand side is the prior (unconditional) probability. The conditional probabilities in the numerator and denominator on the right-hand side are computed using single-copy (Bernoulli) probabilities, $\lambda$ or $\mu$. Subject to these substitutions, Bayes' formula becomes

$$\begin{aligned} &\Pi_{(\texttt{technician showed up}\,|\,m\ \texttt{out of}\ n\ \texttt{copies are good})} \\ &\qquad = \frac{\Pi_1}{\Pi_2} \times \Pi_{(\texttt{the technical showed up})}, \end{aligned} \tag{3.9.3}$$

where

$$\begin{aligned} \Pi_1 &= \Pi_{(m\ \texttt{out of}\ n\ \texttt{copies are good}\,|\,\texttt{the technician showed up})}, \\ \Pi_2 &= \Pi_{(m\ \texttt{out of}\ n\ \texttt{copies are good}\,|\,\texttt{the technician showed up or not})}. \end{aligned} \tag{3.9.4}$$

The ratio $\Pi_1/\Pi_2$ is the Bayes factor.

We may define

$$\Pi_0 = \Pi_{(m\ \texttt{out of}\ n\ \texttt{copies are good}\,|\,\texttt{the technician did not show up})}, \tag{3.9.5}$$

and note $\Pi_2$ is not necessarily equal to $\Pi_0 + \Pi_1$.

### 3.9.2 The technician showed up

The probability $\Pi_1$ plays the role of the likelihood. The receptionist argues that, if the technician had shown up, the copy machine would have been repaired and the conditional probability in the numerator on the right-hand side would be given by the binomial distribution,

$$\Pi_1 = \mathcal{B}_m^n(\lambda) \equiv \mathcal{C}_m^n \, \lambda^m \, (1-\lambda)^{n-m}, \tag{3.9.6}$$

where $\lambda$ is the Bernoulli probability that a copy is good after repair and $\mathcal{C}_m^n$ is the binomial coefficient discussed in Appendix A.

The binomial coefficient accounts for the possibility that a good copy comes out first, second, third or at any stage, and sums all possible permutations of $m$-good and $(n-m)$-bad copies in a total of $n$ copies. These are then multiplied by the corresponding single copy (Bernoulli) probability, $\lambda$ or $1-\lambda$, as shown in (3.9.6).

We observe that $\mathcal{B}_n^n(\lambda) = \lambda^n$ for $m = n$ and $\mathcal{B}_0^n(\lambda) = (1-\lambda)^n$ for $m = 0$, expressing the probabilities of all good or bad photocopies. As expected, $\mathcal{B}_n^n(1) = 1$ and $\mathcal{B}_m^n(1) = 0$ for $m = 0, 1, \ldots, n-1$.

### *3.9.3 Partition of unity*

In a photocopying session of $n$ copies, the set of possible outcomes is: not one good copy $(m = 0)$, only one good copy $(m = 1)$, ..., all copies are good $(m = n)$. The corresponding probabilities must sum to unity,

$$\sum_{m=0}^{n} \mathcal{B}_m^n(\lambda) = 1. \tag{3.9.7}$$

In fact, this constraint is an identity stemming from the partition of unity according to equation (A.27), Appendix A,

$$\sum_{m=0}^{n} \mathcal{B}_m^n(\lambda) = \big(\, \lambda + (1-\lambda) \,\big)^n = 1^n = 1 \tag{3.9.8}$$

for any arbitrary $\lambda$.

### *3.9.4 The technician did not show up*

If the technician had not shown up, the machine would not have been repaired and the conditional probability $\Pi_0$ would be given by the binomial distribution:

$$\Pi_0 = \mathcal{B}_m^n(\mu) \equiv \mathcal{C}_m^n \, \mu^m \, (1-\mu)^{n-m}. \tag{3.9.9}$$

This estimate is the same as that shown in (3.9.6), except that $\lambda$ has been replaced by $\mu$. We recall that $\mu$ is the probability that a copy turns out to be good before repair.

### *3.9.5 Marginal probability*

The denominator on the right-hand side of Bayes' equation (3.9.3) is the marginal probability

$$\Pi_2 = \Pi_1 \times s + \Pi_0 \times (1 - s), \tag{3.9.10}$$

given by

$$\Pi_2 = \mathcal{B}_m^n(\lambda) \times s + \mathcal{B}_m^n(\mu) \times (1 - s), \tag{3.9.11}$$

where $s$ is the probability that the technician has shown up. Substituting into equation (3.9.3) the preceding expressions, we find that

$$\Pi_{(\texttt{the technician showed up}\,|\,m\ \texttt{out of}\ n\ \text{good copies})} = \frac{\mathcal{B}_m^n(\lambda)}{\mathcal{B}_m^n(\lambda) \times s + \mathcal{B}_m^n(\mu) \times (1 - s)} \times s. \tag{3.9.12}$$

Simplifying by deleting the binomial coefficient from the numerator and denominator of the fraction on the right-hand side, we derive precisely formula (3.8.21).

### *3.9.6 Probability tree for a photocopying session*

The probability tree for a photocopying session where $n$ copies are made is shown in Table 3.9.1. The tree has two main branches reflecting the two states of the machine, broken (technician no show) or fixed (technician show).

In revising the prior probability, $s$, only two of the four conditional probabilities listed in the last column of Table 3.9.1 are employed; appropriate pairs are connected by pointed arrows. Using this tree bypasses the calculations of a revised technician show-up fraction, $s$, after each exploratory copy has been made in a photocopying session, good or bad.

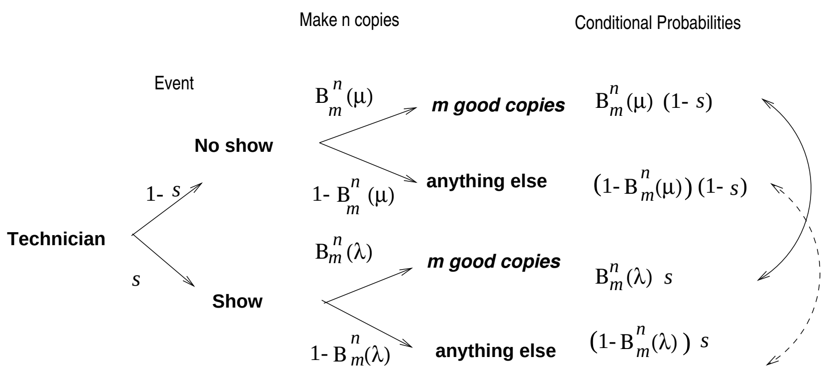



TABLE 3.9.1 Photocopying session probability tree involving three fractions, $s$, $\mu$, and $\lambda$, and two parameters, $n$ and $m$. The conditional probabilities shown in the right column add to unity.

### 3.9.7 *Numerical evaluation of the binomial distribution*

The following Matlab function computes the binomial distribution $\mathcal{B}_m^n(\theta)$ for specified $n$, $m$, and $\theta$:

```
function Bnm = binomial_distro(m,n,theta)

  kmax = m;
  if(n-m < kmax) kmax = n-m; end
% kmax = min(m,n-m);  % alternative

  combo = 1.0;
  for k=1:kmax
   combo = combo*(n-k+1)/k;
  end

  Bnm = combo* theta^m *(1.0-theta)^(n-m);

return
```

The first two lines identity the maximum of $m$ and $n-m$ involved in the computation of the binomial coefficient. Matlab and other applications encapsulate an internal function named *min* that computes the minimum of two integers.

### *3.9.8 Expected value and variance*

Given a Bernoulli probability, $\theta$, the expected value of $m$ for fixed $n$ is

$$\overline{m} \equiv \sum_{m=0}^{n} m\,\mathcal{B}_m^n(\theta) = n\,\theta, \tag{3.9.13}$$

and the associated variance is

$$\sigma^2 \equiv \sum_{m=0}^{n} (m - \overline{m})^2\,\mathcal{B}_m^n(\theta) = n\,\theta(1-\theta), \tag{3.9.14}$$

where $\sigma$ is the standard deviation. As expected, the variance is zero when $\theta = 0$ or $1$, and reaches a maximum at the mid-point, $\theta = \frac{1}{2}$.

### *3.9.9 Maximum likelihood estimate (MLE)*

Engineer Sashka of *Heavy Equipment, Inc.* is interested in the physical interpretation of the binomial distribution. Let $\theta$ be a generic Bernoulli probability that an arbitrary photocopy comes out good, and correspondingly $1 - \theta$ be the generic Bernoulli probability that an arbitrary photocopy comes out bad.

In a photocopying session, $n$ copies are made, and $m$ of these copies are good, where $m \leq n$. Sashka is thinking that, given a pair of $n$ and $m$, the corresponding Bernoulli probability $\theta$ can be computed by maximizing the binomial distribution with respect to $\theta$,

$$\mathcal{B}_m^n(\theta) \equiv \mathcal{C}_m^n\,\theta^m\,(1-\theta)^{n-m}. \tag{3.9.15}$$

This is done by setting the derivative of the binomial distribution with respect to $\theta$ to zero to obtain an algebraic equation for $\theta$.

In the formal language of statistics, the solution is the *maximum likelihood estimate* (MLE). From this viewpoint, the binomial distribution appears as a likelihood function.

### *3.9.10 Mary Boas*

Sashka pulls out the paperback international edition of her favorite calculus book (Boas, M. L. (2005) *Mathematical Methods in the Physical Sciences*, Third Edition, Wiley) and computes the derivative

$$\frac{\mathrm{d}\mathcal{B}_m^n(\theta)}{\mathrm{d}\theta} = \mathcal{C}_m^n \left( m\,\theta^{m-1}\,(1-\theta)^{n-m} - (n-m)\,\theta^{m-1}\,(1-\theta)^{n-m-1} \right), \tag{3.9.16}$$

which can be rearranged into the expression

$$\frac{\mathrm{d}\mathcal{B}_m^n(\theta)}{\mathrm{d}\theta} = \mathcal{B}_m^n(\theta)\,\frac{m-n\theta}{\theta\,(1-\theta)}. \tag{3.9.17}$$

Using the differentiation rule for the logarithm, Sashka obtains

$$\frac{\mathrm{d}\,\ln\mathcal{B}_m^n(\theta)}{\mathrm{d}\theta} = \frac{m-n\theta}{\theta\,(1-\theta)}. \tag{3.9.18}$$

Setting this derivative to zero, Sashka finds that

$$\theta_{\max} = \frac{m}{n}. \tag{3.9.19}$$

Sashka is intrigued that the logarithm of the likelihood function appears naturally in his calculations.

A graph of $\mathcal{B}_m^n(\theta)$ against $\theta$ for $n = 10$ and $m = 8$ is shown in Figure 3.9.1. The results confirm that the maximum is achieved when $\theta = m/n$, indicated by the broken vertical line. Similar results are obtained for other combinations of $n$ and $m$.

### *3.9.11 Behavioral pattern*

Sashka is thinking that, if $\theta_{\max} = m/n$ is precisely equal to $\lambda$, then the technician has showed up beyond any doubt.

If $\theta_{\max}$ is precisely equal to $\mu$, the technician has not showed up beyond any doubt.

If $\theta_{\max}$ is closer to $\lambda$ than $\mu$, it is fair to deduce that the technician has shown up.

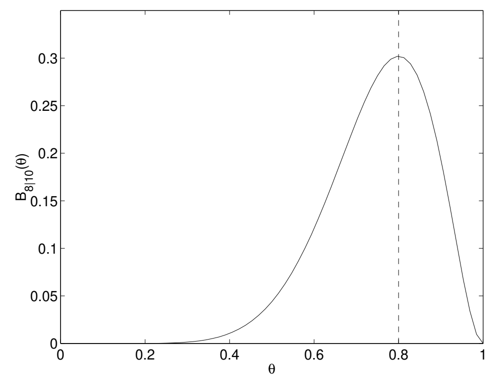



FIGURE 3.9.1 Graph of the binomial distribution $\mathcal{B}_{m|n}(\theta)$ against the Bernoulli probability, $\theta$, for $n = 10$ and $m = 8$. The maximum is achieved when $\theta = m/n$, indicated by the broken vertical line. Note that the distribution is zero in the extreme cases $\theta = 0$ or $1$.

If $\theta_{\max}$ is closer to $\mu$ than $\lambda$, it is fair to deduce that the technician has not shown up.

The last two deductions carry some uncertainty and are made with some trepidation. The notion of confidence intervals comes to mind.

However, this approach does not take into consideration that the technician has shown up in the evening on every previous occasion, and therefore does not incorporate a known behavioral pattern.

By contrast, the Bayesian analysis has provided us with an unambiguous estimate for the probability that the technician has showed up based on available data and a known behavioral pattern mediated by the prior. Other factors, such as inclement weather on the day of repair, can be taken into account.

***Exercises***

**3.9.1** Derive formula (3.9.13) for the expected value of $m$.

**3.9.2** Based on Figure 3.9.1, would you say that the technician showed up if $\mu = 0.6$ and $\lambda = 1.0$?

## 3.10 Three cousins

Photocopier repair company, *Spinning Wheels, LLC* is a family-owned business that employs as dispatched technicians three of the owner's cousins. The cousins have different levels of experience and varying degrees of mechanical skill.

After a machine has been repaired by the $i$th cousin, the machine's fraction of good copies is $\theta_i$ for $i = 1, 2, 3$, where $\theta_i$ is a quality index varying in the range $[0, 1]$. In the broader context of probability and statistics, $\theta_i$ is a Bernoulli probability.

### 3.10.1 Baja

A customer who calls for help is not sure of which technician will be dispatched. To account for this uncertainty, the boss assigns to each technician a probability, $p_i$ for $i = 1, 2, 3$, subject to the compulsory constraint that

$$p_1 + p_2 + p_3 = 1. \tag{3.10.1}$$

In the absence of insights, it is fair to assume that $p_i = \frac{1}{3}$ for $i = 1, 2, 3$, that is, any cousin could be dispatched with equal probability. If one cousin is known to take frequent holidays in Baja, then his probability would be lower.

After the copy machine has been repaired, the boss makes $n$ copies and counts $m$ good copies and $n - m$ bad copies. The boss wants to infer the posterior probability of which technician was dispatched. The pertinent technician probability tree is shown in Table 3.10.1.

### 3.10.2 Bayes equation

The boss introduces the prior probability $p_i$ and applies Bayes' rule to

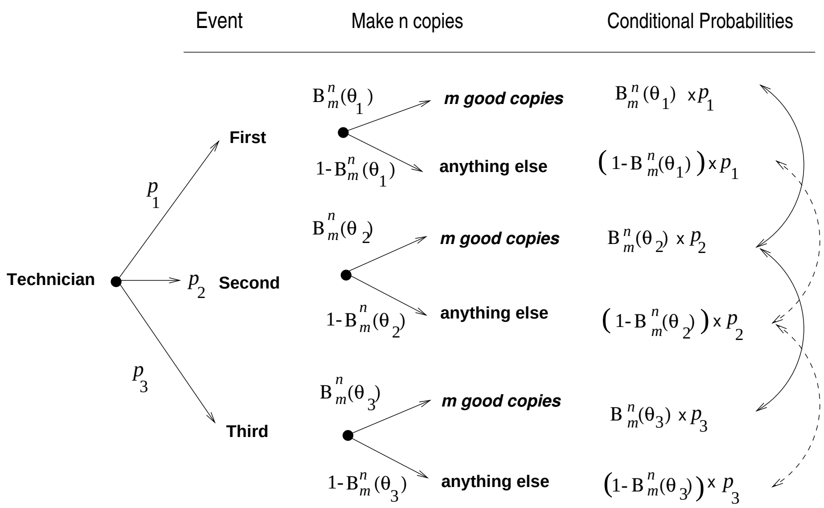



TABLE 3.10.1 Three-technician probability tree involving probabilities of dispatch, $p_1$, $p_2$, and $p_3$, and corresponding repair effectiveness fractions, $\theta_1$, $\theta_2$, and $\theta_3$ (Bernoulli probabilities). The conditional probabilities shown in the right column add to unity. As more technicians are added, the tree becomes longer.

write the equation

$$(p_i)_{\text{posterior}} = \frac{\mathcal{B}^n_m(\theta_i)}{\mathcal{B}^n_m(\theta_1) \times p_1 + \mathcal{B}^n_m(\theta_2) \times p_2 + \mathcal{B}^n_m(\theta_3) \times p_3} \times p_i \tag{3.10.2}$$

for $i = 1, 2, 3$, where the binomial distribution, $\mathcal{B}_{m|n}(\theta)$, plays the role of of likelihood function. The denominator of the fraction on the right-hand side of (3.10.2) is the sum of the conditional probabilities connected by the solid arcs in the probability tree shown in Table 3.10.1.

In other applications, the likelihood function will be given by some other distribution defined with respect to a measurement or quantifiable observation, as discussed in Chapter 4.

### 3.10.3 Any number of cousins

The boss explains her calculations to engineer Gevorg and to the receptionist over fish and chips snacks at the local pub, compliments of the boss, while Tom Petty sings *like a refugee* from inside the juke box.

The receptionist suggests that, if $n_t$ technicians were available, the posterior probability of the $i$th technician would be given by the generalized formula

$$(p_i)_{\text{posterior}} = \frac{\mathcal{B}_m^n(\theta_i)}{\sum_{j=1}^{n_t} \mathcal{B}_m^n(\theta_j) \times p_j} \times p_i \tag{3.10.3}$$

for $i = 1, \ldots, n_t$. The probability distribution, $p_i$, can be revised (updated) each time a set of $n$ copies have been made and the number of good copies, $m$, have been counted.

### 3.10.4 Performance ranking

Without loss of generality, each technician can be assigned a unique integer identification number, $i$, such that technician #1 is the poorest performer and technician $\#n_t$ is the best performer. This means that

$$0 \le \theta_1 < \theta_2 < \cdots < \theta_{n_t-1} < \theta_{n_t} \le 1. \tag{3.10.4}$$

An example for $n_t = 32$ technicians is shown in Figure 3.10.1. For the purpose of illustration, the distribution shown in Figure 3.10.1 was generated using the formula

$$\theta_i = \frac{1}{2^{(n_t-i)/(n_t-1)}} \left( 1 - \sin\left(\frac{i-1}{n_t-1}\,\pi\right) \right) \tag{3.10.5}$$

for $i = 1, \ldots, n_t$. In this case, the work of the first technician is marginal, $\theta_1 = 0.5$, whereas the work of the best technician is outstanding, $\theta_{32} = 0.99$.

### 3.10.5 On the seventh floor

Word has spread out on the seventh floor that the boss, Gevorg, and the receptionist are engaged in some kind of parameter estimation concerning the copy machine technician using a Bayesian approach, and

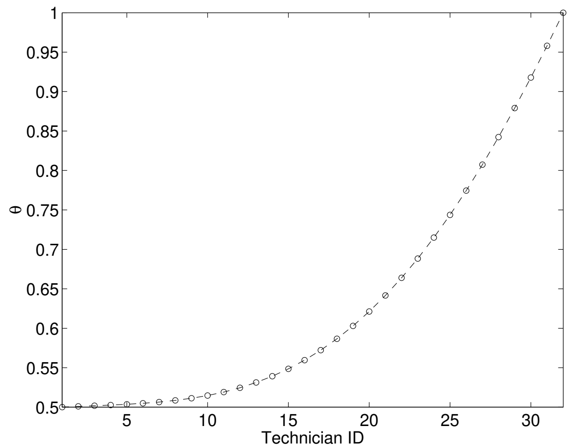



FIGURE 3.10.1 Distribution of the technician performance index, $\theta_i$, in terms of the technician ID, $i$.

coworkers gather around the water cooler to discuss. The term *statistical inference* is mentioned by someone who had the chance to loop up the topic on a mobile devices. A usually mundane workday is turning into a metaphysical experience.

It appears that mathematics and other abstract sciences open a doorway to a hidden world that is free of anxiety, earthly demands, and the futile pursuit of rewards in physical life.

### 3.10.6 *Lining up to make photocopies*

After some deliberations, the boss and her coworkers decide to conduct the following experiment:

1. Make a guess for the *initial prior* probabilities, $p_i$ for $i = 1, \ldots, n_t$. In the absence of information, these probabilities are all the same, $p_i = 1/n_t$.
2. Ask the first volunteer make $n_1$ copies and count the number of good copies, $m_1$.
3. Use equation (3.10.3) with likelihood function $\mathcal{B}_{m_1}^{n_1}(\theta_i)$ to compute the revised probabilities $p_i$ for $i = 1, \ldots, n_t$.

4. Replace the old with the revised $p_i$ for $i = 1, \ldots, n_t$.
5. Have the second volunteer make $n_2$ copies and count $m_2$.
6. Use equation (3.10.3) with likelihood function $\mathcal{B}_{m_2}^{n_2}(\theta_i)$ to compute the revised probabilities $p_i$ for $i = 1, \ldots, n_t$.
7. Repeat to the last volunteer.

At the end of the experiment, inspect and draw conclusions from the final posterior probability distribution.

### *3.10.7 Devil's advocate*

A well known devil's advocate, poses a question: *why break up the task into different photocopying sessions instead of performing one large session incorporating all photocopying sessions?* Everyone scratches their heads and are hard pressed to find an easy answer. In fact, the answer is a key to Bayesian analysis.

### *3.10.8 Compound likelihood function*

To address the devil advocate's comment, the receptionist points out that, each time a set of photocopies are made, the set of current prior probabilities, $p_i$, are revised by multiplication with the current likelihood function, as shown in (3.10.3). Multiplication is followed by normalization implemented by division to ensure that the posterior probabilities sum to unity on the right-hand side of (3.10.3).

Chain threading these actions, we find that the posterior probabilities after $k$ volunteers have made photocopies is given by

$$(p_i)_{\text{posterior}} = \frac{1}{\varphi}\,\Lambda(\theta_i) \times (p_i)_{\text{initial prior}} \tag{3.10.6}$$

for $i = 1, \ldots, n_t$, where

$$\Lambda(\theta_i) \equiv \mathcal{B}_{m_k}^{n_k}(\theta_i) \times \cdots \times \mathcal{B}_{m_1}^{n_1}(\theta_i) \tag{3.10.7}$$

is a compound likelihood function and

$$\varphi = \sum_{j=1}^{n_t} \Lambda(\theta_j) \times (p_j)_{\text{initial prior}} \tag{3.10.8}$$

is a normalization constant ensuring that the posterior probabilities add to unity at any stage.

Substituting the expression for the binomial distribution and simplifying, we find that the likelihood function of the accumulated sessions can be replaced by the cumulative binomial distribution

$$\Lambda(\theta_i) \to \mathcal{B}_{m_k+\cdots+m_1}^{n_k+\cdots+n_1}(\theta_i) \tag{3.10.9}$$

for $i = 1, \ldots, n_t$, involving cumulative photocopies. We conclude that the photocopying sessions can be consolidated into a single session. Conversely, a single session can be broken up into multiple sessions.

### *3.10.9 Numerical simulation*

The physical experiment conceived by the boss and her coworkers can be replaced by numerical simulation. Unfortunately, because none of the engineers of *Heavy Equipment, Inc.* took a computer programming class at college, all calculations must be done by hand. The engineers agreed that the engineering curricula at the Universities they attended were significantly watered down and decided to submit grievances at the next fund-raising alumni events.

### *3.10.10 Quality of education*

Engineer Gevorg suggested that the reason for the weakening of the curricula is that the Deans of Colleges of Engineering place emphasis on favorable student ratings and instructional gimmicks instead of emphasizing quality. Quality of education is an attribute that cannot be quantified and therefore cannot be embedded in PowerPoint presentations to potential donors. How do you convince someone that the performance of a hand-tightened bolt is superior to that of a pneumatically-tightened bolt?

### *3.10.11 Code*

The experiment conceived by the boss and her coworkers is implemented in the following Matlab code named *photocopies*:

```
%---
% data
```

```
%---

nt = 32; % number of technicians
ns = 5;  % number of photocopying sessions

% number of copies per session:

n = [10; 14; 12; 19; 22];

% number of good copies:

m = [09; 07; 11; 15; 19];

%---
% bad copies per session
%---

for i=1:ns
 l(i) = n(i)-m(i);
end

%---
% assign Bernoulli probability
% per technician (typical)
%---

fc = 1.0/2^(1/(nt-1));

for i=1:nt
 fcc = 1-0.2*sin((i-1)*pi/(nt-1));
 theta(i) = fcc*fc^(1*(nt-i));
end

%---
% initial prior probabilities (uniform)
%---

for i=1:nt
 p(i) = 1/nt;
```

```
end

%===
% sequential testing over sessions
%===

for i=1:ns
  sum = 0.0;
  for j=1:nt
    L(j) = binomial_distro(m(i),n(i),theta(j));
    sum = sum + L(j)*p(j);
  end
  for j=1:nt
   pnew(j) = L(j)*p(j)/sum;
  end
  p = pnew;
end
```

The code calls the function *binomial_distro* listed in Section 3.10 to evaluate the binomial distribution.

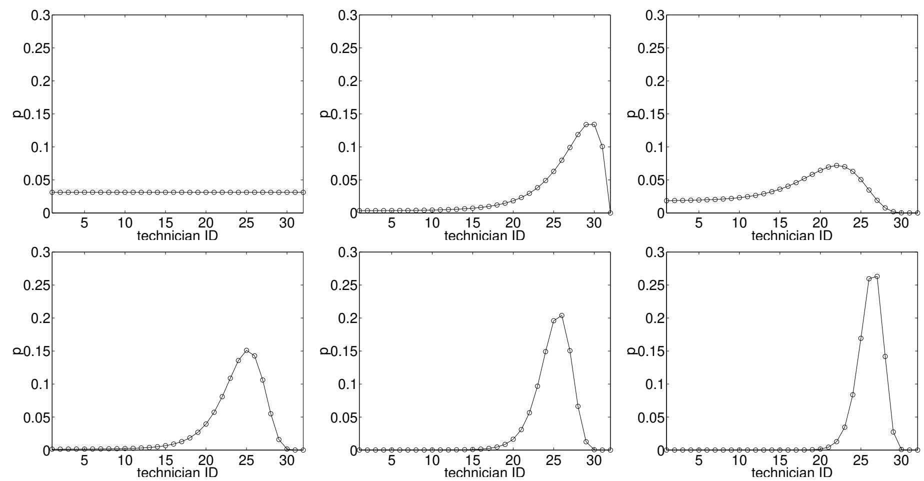
### *3.10.12 Uniform initial prior*

Results of computations with $n_t = 32$ technicians and a uniform prior are shown in Figure 3.10.2. The number of copies and the number of good copies per session are defined in two arrays in the code.

In Figure 3.10.2(*a*), the probability $p_i$ is plotted against the technician ID, $i$. In Figure 3.10.2(*b*), the probability $p_i$ is plotted against the Bernoulli probability, $\theta_i$, to preserve anonymity. Note that the data are spaced evenly in Figure 3.10.2(*a*) but unevenly in Figure 3.10.2(*b*). The results suggests that one of the best technicians was dispatched.

The cumulative binomial distribution concentric with the larger circular symbols arising from the simulation precisely coincides with the distributions shown in Figure 3.10.2(*b*). The dashed vertical lines mark the corresponding location of the maximum. The maximum occurs at the ratio $m_1/n_1$ in the second panel and at the ratio $(m_1+m_2)/(n_1+n_2)$ in the third panel of Figure 3.10.2(*b*); similar expressions for the maxima apply to subsequent panels.

(*a*)




(*b*)

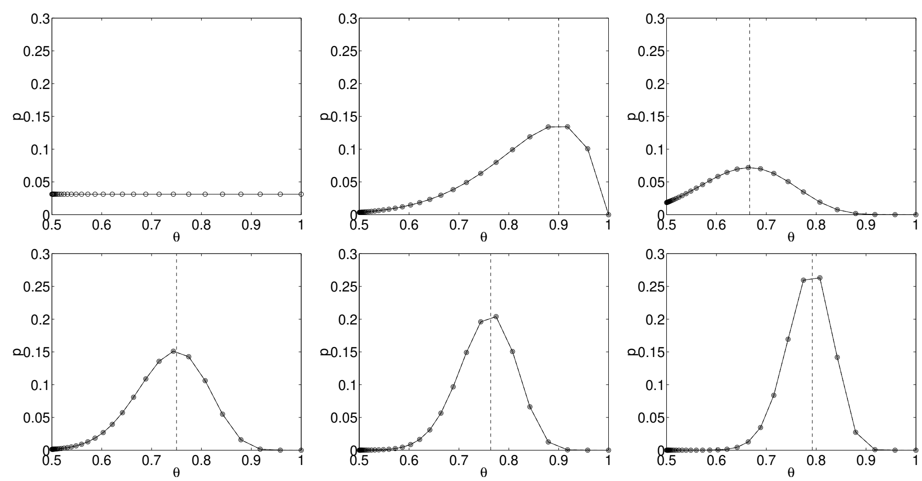



FIGURE 3.10.2 Evolving distribution of the probability of (*a*) the technician ID and (*b*) anonymous repair performance index, $\theta$, after each photocopying session for a uniform initial prior distribution and $n_t = 32$ technicians.

### *3.10.13 Non-uniform initial prior*

Results of computations with $n_t = 32$ technicians and a non-uniform prior are shown in Figure 3.10.3. The small circular symbols connected by a broken line in Figure 3.10.3($b$) represent the cumulative binomial distribution, which is different from that arising from the Bayesian analysis with a nonuniform initial prior. The effect of the initial prior on the evolving probability distribution becomes decreasingly important as more photocopying sessions are carried out.

### *3.10.14 Experimentation*

If a new photocopy session data comes in in the midst of an update, we should not start afresh using the new data immediately. Instead, we should use the current data for a complete update, and then repeat the process for the new data for another complete update. If we do not do this, the current data will have only a partial effect on the conclusions.

#### *Exercises*

**3.10.1** What is the name of Tom Petty's song with lyrics *She was born in an Indiana town?*

**3.10.2** Reproduce the graphs shown in Figure 3.10.3 for a different prior of your choice.

## *3.11 Broad features*

We have discussed the basic principles and applications of Bayesian analysis in several contexts. Filtered through the prism of Robert M. Pirsig's ideas discussed in his brilliant essay *Zen and the art of motorcycle maintenance*, Bayes' rule provides us with guidelines for making decisions, acknowledging and correcting mistakes, and defining behavioral and spiritual paths.

The brilliance of Pirsig's writing is that it cannot be classified in a genre. One spends much more time thinking about what was said in a paragraph than reading the lines. Regarding morality, Pirsig's writes:

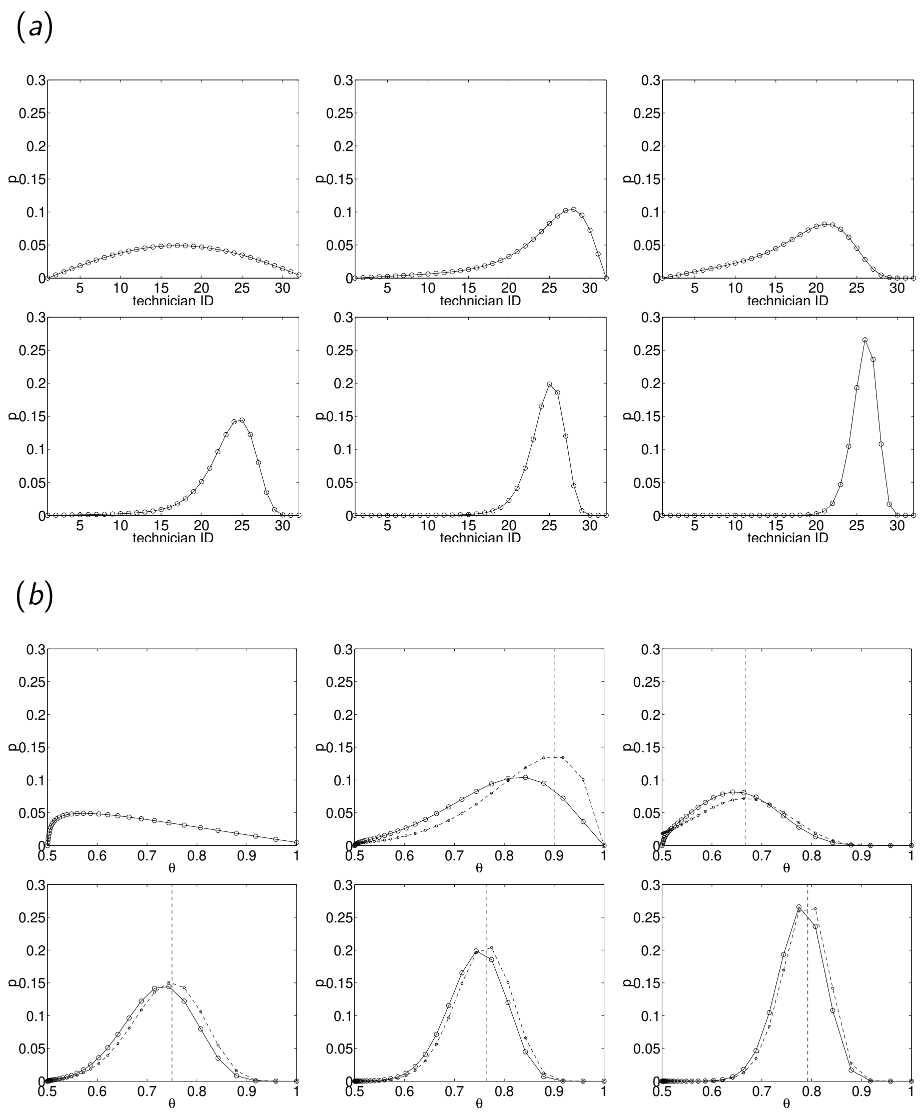



FIGURE 3.10.3 Evolving distribution of the probability of (*a*) the technician ID and (*b*) anonymous repair performance index, $\theta$, after each photocopying session for a non-uniform initial prior distribution and $n_t = 32$ technicians.

```
And what is good, Phaedrus,
And what is not good --
Need we ask anyone to tell us these things?
```

Phaedrus is Pirsig's previous self acting as an interlocutor, which can be anyone's inner self.

Pirsig poses a seemingly mundane question: *How much one should maintain one's own motorcycle?* The answer requires a great deal of thought.

Now we can look back to highlight salient features of the Bayesian approach, and indicate associations with Zen-like awareness and cognitive distortions.

### *3.11.1 The meaning of probability in Bayesian analysis*

In Section 1.1, we stated that the concept of probability can be interpreted in a multitude of ways, including fraction or frequency of events in real time or groundhog lifetime. In the context of Bayesian analysis, these interpretations can be retained or replaced by a factual or intuitive, tangible or undefined sense of plausibility.

A sense of plausibility is encapsulated in statements such as "*I expect that she will become a successful dentist*" and "*I expect that this will be an improved tractor model*". The Bayesian interpretation places the notion of probability squarely within the context of epistemology.

### *3.11.2 Under-determination*

Overdetermination describes the availability of redundant data (evidence) regarding the veracity of an event or the truthfulness of a hypothesis. In mathematics, a system of equations is overdetermined when it encapsulates more equations than unknowns.

In the Bayesian framework, because we work with probabilities that are always less than unity, an event of interest cannot be or become overdetermined in spite of overwhelming evidence: beliefs and ideas are always underdetermined, originating from an initial state that allows for even the slightest amount of doubt. This is another way of saying that there should always be some doubt regarding any fact, opinion, or assessment.

Doubt and self-doubt are signs of an intelligent mind. Strange though it may seem, Bayesian underdetermination is thoughtful, tolerant, and forgiving. If everyone votes that the earth rotates around the moon, the Bayesian approach will allow for the possibility that they may all be wrong.

### 3.11.3 Conservatism bias

Conservatism bias describes a human cognitive condition whereby a person under-revises a belief with respect to that dictated by Bayes' rule in the face of new evidence.

A large percentage of people and groups fall prey to this bias in that they believe that their individual or collective behavior and modus operandi does not fully reveal their intentions, motivations, and prejudices. The colloquialism "*Fooled me once shame on you, fooled me twice shame on me*" can be regarded as a Bayesian corollary.

### 3.11.4 Murphy's law

Captain Edward A. Murphy was the director of a NASA laboratory at the Edwards United States Air Force Base in 1949. When an important experiment was miswired, Captain Murphy remarked: "*If there are two or more ways of doing something, and one of them can lead to catastrophe, then someone will do it.*"

Captain Murphy's comment was generalized to become the universal Murphy's law, one version of which reads: "*Given enough time, if something can go wrong, it will go wrong.*" Consequently, imperfections will be amplified and malfunctioning or dysfunctional communities and institutions will inevitably reach a point of breakdown.

In fact, Murphy's law can be regarded as a consequence of Bayes law. One should bear in mind that Murphy's law is not the result of some kind of malevolence on behalf of the cosmos but a reflection of the complexity of the real world.

### *3.11.5 Experience-based assessments*

Bayes' rule provides us with a venue for testing a hypothesis motivated by experiences, recollections, hunches, impulses, or data gathered. Moreover, Bayes' rule provides us with a starting point for weighing the wisdom of explicit or implicit decisions, predictions, or expectations.

One concern is that experiences, hunches, impulses, and data may have been gained or collected in a limited, misleading, or inappropriate context. For example, a person who has never been poor can hardly discuss poverty; a person who has never written a well-cited research paper can hardly manage an educational institution; a person with a poor understanding of world history cannot process the long-term significance of current events.

Our consciousness reconstructs recollections by interpolating through selected snapshots. The interpolation itself is based on experience-driven expected actions that may vary widely across witnesses of the same event. Honest eye witnesses sometimes provide different accounts of the same event.

### *3.11.6 Subjectivity*

In sociological and some scientific contexts, the conditional probabilities involved in Bayes' rule can be subjective. For example, we hear that $80\%$ of all dentists believe that fluoride supplement is good for the teeth. If you ask different people to suggest a value for the conditional probability

$$\Pi_{(\texttt{Feeling well}\,|\,\texttt{Feeling good})}, \tag{3.11.1}$$

they will give you different answers. Those who accept the dominance of mind over body will suggest a high estimate, and those who only accept the traditional medicinal approach will suggest a low estimate. A group of people believe that disease can be cured by holistic methods and prayer. The subjectivity is especially important in trial by jury, as discussed in Section 2.3.

Exaggerated and wrongly framed arguments that appeal to fear, prejudice, and other instincts can be made to influence the subjective estimates of conditional probabilities of others. People with mental pathology make arguments and take actions to influence the frame of mind of unsuspecting victims and innocent bystandards.

### *3.11.7 Proliferation of theories and white noise*

As the number of scientists grows along with the number of opinion outlets, so does the number of hypotheses, theories, and suggestions. The test the veracity or validity of any of these according to Bayes' rule, reliable hard data are necessary. Hard data should be distinguished from generic data in the way that intelligent discourse should be distinguished from gossip and small talk.

Generating hard data requires painstaking work, persistent thinking, and substantial investment in time, effort, and brain power. The casual Internet surfer realizes with extreme discomfort that the true and the precious are dispersed in white noise that cannot be suppressed but is archived instantaneously to be available in perpetuity. The more noise is removed, the more noise is introduced by a relentless supply. It is now nearly impossible to discover a simple or absolute truth.

### *3.11.8 Bayesian filtering*

The proliferation of inaccurate information, false arguments, spam, and irreproducible data underscores the importance of Bayesian filtering well before inaccurate information, false arguments, spam, and wrong data has been archived. Unfortunately, because of the complexity of modern communication channels, this is a daunting task.

### *3.11.9 Interpretation of dreams*

It is both interesting and underappreciated that Bayesian analysis provides us with a framework for unifying disparate subjects and identifying underlying causes.

One example is Sigmund Freud's interpretation of dreams. A long time ago, the idea that someone's personality is determined nearly exclusively by early-life experiences would seem absurd. The usefulness and implications of Bayesian analysis have not been fully explored, especially in cognitive and behavioral sciences.

### *3.11.10 Peter's principle*

The world of scientific computing is divided into uploaders and downloaders. The rest of the world is divided into givers and takers. The world of statistics is fractured into frequenists and Bayesianists.

Bayesian analysis has been criticized by the frequenistic establishment for being fundamentally flawed in interpretation, application, and implementation. The exaggerated wording is a clue that the criticism is largely unfounded, with some reservations and exceptions, as discussed in later sections.

Although it is true that the mindless or manipulative application of Bayesian analysis can lead to erroneous and misleading conclusions, this is the case for any other method of analysis. The Bayesian approach is particularly susceptible to the generalized version of Peter's principle: *stretch it until it breaks, and then some*. The classic Peter's principle prescribes that managers rise to their level of (in)competence. However, the broadband criticism should not be directed to the machinery or method, but rather to the operator.

### *3.11.11 Three stages of acceptance*

The inability of an establishment to fully accept a new framework is consistent with the well known resistance of the scientific community to new concepts, methods, and ideas. The three stages of resistance are as follows:

*A new concept is wrong; the concept is correct but is not important; the concept is correct and important, but we knew it all along.*

True scholars know by experience that it is much easier to publish a mediocre research paper, article, or book than a ground-breaking manuscript. A fair number of scholars give up or become marginalized, defeated by a powerful academic establishment.

### *Exercises*

**2.11.1** Describe an experiment that can be implemented to assess the proposition that the earth is not flat.

**3.11.2** Discuss a real-life application of Peter's principle.

**3.11.3** Describe a set of circumstances where the three stages of acceptance apply.

# Chapter 4

# Probability density functions (pdf)

The notion of uncertainty quantified in terms of probability was introduced in Chapter 1 and discussed in the first three chapters with reference to a sample space consisting of a collection of discrete events. Bayes' equation was derived in Chapter 2 and applications were elaborated in Chapter 3 for a variety of setting.

To handle sample spaces containing data, variables, and events that vary in a continuous range, now we introduce the notion of a probability density function (pdf), not to be confused with the portable document format (pdf). Pertinent Bayes rules are derived in this chapter and illustrative applications are discussed in the remainder of this book

## 4.1 Probability density function (pdf)

Assume that a random variable, $\theta$, takes positive, negative, or zero values between two specified limits, $a$ and $b$, as shown in Figure 4.1.1, For example, $\theta$ can be the temperature of a star, the time we receive a phone call, or the amount of rainfall on any given day.

The associated probability density function (pdf), $\phi(\theta)$, is defined such that the probability that $\theta$ falls inside an arbitrary interval, $[\theta_-, \theta_+]$, that is included in $[a, b]$, is equal to the area under the graph of the pdf between two vertical lines drawn at $\theta = \theta_-$ and $\theta_+$,

$$\Pi(\theta_-, \theta_+) = \int_{\theta_-}^{\theta_+} \phi(\theta) \, \mathrm{d}\theta. \tag{4.1.1}$$

An essential stipulation is that $\phi(\theta) \geq 0$ for any $\theta$ in $[a, b]$; a pdf cannot

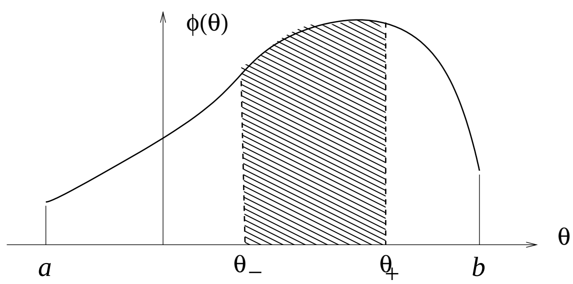



FIGURE 4.1.1 Illustration of the probability density function (pdf) of one random variable, $\theta$, defined between two specified limits, $a$ and $b$,

be negative. In a generalized context, $\phi(\theta)$ can be defined in terms of the Dirac delta function expressing a spike.

By construction, the area under the entire curve above the $\theta$ axis in Figure 4.1.1 is equal to unity,

$$\Pi(a, b) = \int_a^b \phi(\theta)\,\mathrm{d}\theta = 1. \tag{4.1.2}$$

This constraint ensures that the probability that $\theta$ takes some value, any value, between $a$ and $b$ is unity.

### *4.1.1 Cumulative probability distribution function*

The cumulative probability distribution function (cpdf) is defined as

$$\Phi(\theta) = \Pi(a, \theta) = \int_a^\theta \phi(\theta')\,\mathrm{d}\theta', \tag{4.1.3}$$

where $\theta'$ is an integration variable and $\Phi(b) = 1$ due to normalization. The probability that $\theta$ falls inside an infinitesimal interval $\mathrm{d}\theta$ containing or adjacent to $\theta$ is

$$\mathrm{d}\Phi(\theta) = \phi(\theta)\,\mathrm{d}\theta. \tag{4.1.4}$$

Consequently,

$$\phi(\theta) = \frac{\mathrm{d}\Phi}{\mathrm{d}\theta}. \tag{4.1.5}$$

### *4.1.2 Integrable singularities*

A pdf, $\phi(\theta)$, is allowed to exhibit an *integrable singularity*, that is, the graph of $\phi(\theta)$ is allowed to rise or plunge to infinity at one point or a collection of points, provided that the area under the curve is finite (non-infinite.)

### *4.1.3 Expected value and variance*

The expected value of $\theta$ is

$$\overline{\theta} \equiv \int_a^b \theta\, \phi(\theta)\, \mathrm{d}\theta, \tag{4.1.6}$$

and the associated variance is

$$\sigma^2 \equiv \int_a^b (\theta - \overline{\theta})^2\, \phi(\theta)\, \mathrm{d}\theta, \tag{4.1.7}$$

where $\sigma$ is the standard deviation.

### *4.1.4 Mapped pdf*

Assume that a random variable, $\theta$, defined in an interval, $[a, b]$, is related to another random variable, $\eta$, by a monotonic function

$$\eta = f(\theta). \tag{4.1.8}$$

In agricultural applications, $\theta$ can be the amount of rainfall and $\eta$ can be the crop yield.

By way of this association, the variable $\eta$ is also a random variable whose pdf is denoted as $\widehat{\phi}(\eta)$. The associated cumulative pdf is

$$\widehat{\Phi}(\eta) = \Pi(a_\eta, \eta) = \pm \int_{a_\eta}^{\eta} \widehat{\phi}(\eta')\, \mathrm{d}\eta', \tag{4.1.9}$$

where $a_\eta = f(a)$. The plus sign is chosen when $f'(\theta) > 0$ so that $\eta > a_\eta$, and the minus sign is chosen when $f'(\theta) < 0$ so that $\eta < a_\eta$, where a prime denotes the first derivative with respect to $\theta$.

Now we require that

$$\mathrm{d}\Phi = \mathrm{d}\widehat{\Phi} \tag{4.1.10}$$

or

$$\phi(\theta)\,\mathrm{d}\theta = \pm\widehat{\phi}(\eta)\,\mathrm{d}\eta, \tag{4.1.11}$$

for corresponding differentials $\mathrm{d}\theta$ and $\mathrm{d}\eta$.

The two pdfs are related by the equation

$$\frac{\phi(\theta)}{\widehat{\phi}(\eta)} = \pm f'(\theta) = \pm\frac{\mathrm{d}f}{\mathrm{d}\theta}. \tag{4.1.12}$$

The ratio of the two corresponding probabilities is non-negative.

### *4.1.5 Random variables with a specified pdf*

In applications, we require a sequence of random values, $\theta$, that conform with a specified probability density function $\phi(\theta)$, inside a specified interval, $[a, b]$.

To generate this sequence, we introduce a random variable, $\varrho$, with uniform probability density function taking values in the range $[0, 1]$. We then map the interval of interest, $[a, b]$, to the standard interval, $[0, 1]$, such that

$$\phi(\theta)\,\mathrm{d}\theta = \mathrm{d}\varrho. \tag{4.1.13}$$

Next, we introduce the cumulative distribution

$$\Phi(\theta) \equiv \int_a^\theta \phi(\theta')\,\mathrm{d}\theta'. \tag{4.1.14}$$

Integrating both sides of equation (4.1.13) we derive the mapping function

$$\Phi(\theta) = \varrho. \tag{4.1.15}$$

Matlab incorporates an internal function named *rand* that generates random numbers $\varrho$, called random deviates or variates. A $\theta$ sequence arises from a $\varrho$ sequence by solving equation (4.1.15) for $\theta$, as discussed in Section 4.3.

### 4.1.6 Discretizing a pdf

Consider a random variable, $\theta$, defined in an interval, $[a, b]$, according to a probability distribution function, $\phi(\theta)$. It is useful to divide the domain of definition of $\theta$ into an arbitrary number of $N$ intervals with size

$$\Delta\theta = \frac{b-a}{N}, \tag{4.1.16}$$

and introduce an underlying provisional discrete probability distribution (dpd) defined in terms of the cumulative distribution as

$$\Delta\Phi_i = \phi(\theta_i)\,\Delta\theta \tag{4.1.17}$$

for $i = 1, \ldots, N$, where

$$\theta_i = a + (i - \tfrac{1}{2})\,\Delta\theta \tag{4.1.18}$$

is the mid-point of the $i$th interval. The associated discrete probability distribution is given by

$$p_i = \frac{\Delta\Phi_i}{\sum_{j=1}^{N} \Delta\Phi_j} = \frac{\phi(\theta_i)}{\sum_{j=1}^{N} \phi(\theta_j)}. \tag{4.1.19}$$

The denominator in the right-hand side is a normalization constant ensuring that

$$\sum_{i=1}^{N} p_i = 1. \tag{4.1.20}$$

Consistent with the definition of the pdf, $p_i$ is the probability that $\theta$ falls inside the $i$th division with some small error that depends on the size of the division. The method described in Section 1.3 can be applied to generate a sample.

The discrete probability distribution can be manipulated in some desired fashion, and the results can be transferred back to the probability density function by taking the limit as $N$ tends to infinity and transitioning from summation to integration.

***Exercise***

**4.1.1** Discuss a physical application where a random, $\eta$, variable is related to another random variable, $\theta$.

## 4.2 Maximum entropy principle

The information entropy associated with a pdf, $\phi(\theta)$, is defined as the expected value of the negative of the logarithm of the pdf,

$$s \equiv -\int_a^b \phi(\theta) \ln \phi(\theta)\, \mathrm{d}\theta. \tag{4.2.1}$$

Since $\phi \le 1$ and thus $\ln \phi < 0$, the information entropy is positive.

The principle of maximum entropy stipulates that the function $\phi(\theta)$ should maximize the constrained entropy functional

$$\widehat{s} = s + \lambda \big( \int_a^b \phi(\theta)\, \mathrm{d}\theta - 1 \big), \tag{4.2.2}$$

where $\lambda$ is a Lagrange multiplier.

To find the solution, we approximate the integrals using the midpoint integration rule with $N$ divisions, obtaining

$$\widehat{s} = -\Delta\theta \sum_{i=1}^{N} \phi(\theta_i) \ln \phi(\theta_i) + \lambda \big( \Delta\theta \sum_{i=1}^{N} \phi(\theta_i) - 1 \big), \tag{4.2.3}$$

where $\Delta\theta = (b-a)/N$ is the size of each division and

$$\theta_i = a + (i - \tfrac{1}{2})\, \Delta\theta \tag{4.2.4}$$

is the mid-point. We find that the entropy becomes maximum when

$$\phi(\theta_i) = \frac{b-a}{N} \tag{4.2.5}$$

for $i = 1, \ldots, N$, and $\lambda = 1 - \ln N$, thereby deriving a uniform pdf.

In case of a specified expected value, $\overline{\theta}$, the following constraint entropy functional is employed:

$$\widehat{s} = s + \lambda \Big( \int_a^b \phi(\theta)\, \mathrm{d}\theta - 1 \Big) + \mu \Big( \int_a^b \theta\, \phi(\theta)\, \mathrm{d}\theta - \overline{\theta} \Big), \tag{4.2.6}$$

where $\lambda$ and $\mu$ are two Lagrange multipliers. When $a = 0$ and $b = \infty$, the pdf that maximizes the entropy functional shown in (**??**) is the exponential distribution

$$\phi(\theta) = \frac{1}{\overline{\overline{\theta}}} \exp \Big( - \frac{\theta}{\overline{\overline{\theta}}} \Big). \tag{4.2.7}$$

In case of a specified expected value, $\overline{\theta}$, and variance, $\sigma^2$, the following constraint entropy functional is employed:

$$\begin{aligned} \widehat{s} = s + \lambda \Big( \int_a^b \phi(\theta)\, \mathrm{d}\theta - 1 \Big) + \mu \Big( \int_a^b \theta\, \phi(\theta)\, \mathrm{d}\theta - \overline{\theta} \Big) \\ + \nu \Big( \int_a^b (\theta - \overline{\theta})^2 \phi(\theta)\, \mathrm{d}\theta - \sigma^2 \Big), \end{aligned} \tag{4.2.8}$$

where $\lambda$, $\mu$, and $\nu$ are three Lagrange multipliers. When $a = -\infty$ and $b = \infty$, the pdf that maximizes the entropy functional shown in (4.2.8) is the Gaussian (normal) distribution

$$\phi(\theta) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp \Big( - \tfrac{1}{2} \frac{(\theta - \overline{\theta})^2}{\sigma^2} \Big), \tag{4.2.9}$$

where $\sigma$ is the standard deviation.

***Exercise***

**4.2.1** Derive the pdf that maximizes the entropy functional shown in (4.2.6) .

## 4.3 Miscellaneous pdfs

Having introduced the notion of the probability density function (pdf), we proceed to consider specific distributions.

### 4.3.1 Flat distribution

Assume that a pdf, $\phi(\theta)$, has the uniform value $1/(b-a)$ in an interval of interest, $[a, b]$,

$$\phi(\theta) = \frac{1}{b-a}. \tag{4.3.1}$$

The expected value is

$$\overline{\theta} \equiv \int_a^b \theta\, \phi(\theta)\, \mathrm{d}\theta = \tfrac{1}{2}\,(a+b), \tag{4.3.2}$$

and the variance is

$$s^2 \equiv \int_a^b (\theta - \overline{\theta})^2\, \phi(\theta)\, \mathrm{d}\theta = \tfrac{1}{12}\,(b-a)^2, \tag{4.3.3}$$

where $s$ is the standard deviation. The associated cumulative distribution computed from (4.1.14) is

$$\Phi(\theta) = \frac{\theta - a}{b-a}. \tag{4.3.4}$$

Using equation (4.1.15), we set

$$\frac{\theta - a}{b-a} = \varrho \tag{4.3.5}$$

and obtain the linear mapping function

$$\theta = a + (b-a)\,\varrho, \tag{4.3.6}$$

which shows that $\theta$ is merely a rescaled random deviate.

### 4.3.2 Linear distribution

In the case of a linear distribution with slope $\beta$ defined in an interval of interest, $[a, b]$,

$$\phi(\theta) = \beta\,(\theta - \frac{a+b}{2}) + \frac{1}{b-a}. \tag{4.3.7}$$

The mandatory normalization condition

$$\int_a^b \phi(\theta)\,\mathrm{d}\theta = 1 \tag{4.3.8}$$

is satisfied for any slope, $\beta$. The associated cumulative distribution is

$$\Phi(\theta) = \int_a^\theta \big(\,\beta\,(u - \frac{a+b}{2}) + \frac{1}{b-a}\,\big)\,\mathrm{d}u. \tag{4.3.9}$$

Performing the integration, we obtain

$$\Phi(\theta) = \big(\beta\,\frac{\theta-b}{2} + \frac{1}{b-a}\big)\,(\theta - a), \tag{4.3.10}$$

which is a quadratic function of $\theta$.

Now using equation (4.1.15), we set

$$\big(\beta\,\frac{\theta-b}{2} + \frac{1}{b-a}\big)\,(\theta - a) = \varrho. \tag{4.3.11}$$

Rearranging, we obtain

$$\tfrac{1}{2}\,\beta\,(\theta-a)^2 + \Big(-\tfrac{1}{2}\,\beta\,(b-a) + \frac{1}{b-a}\Big)\,(\theta - a) = \varrho, \tag{4.3.12}$$

which is a quadratic equation in $\theta - a$. Using the quadratic formula, we obtain the solution

$$\theta = a + \tfrac{1}{2}\,L - \frac{1}{\beta L} + \Big(\big(\tfrac{1}{2}\,L - \frac{1}{\beta L}\big)^2 + 2\,\frac{\varrho}{\beta}\Big)^{1/2}, \tag{4.3.13}$$

where $L = b - a$.

The following Matlab function named *randlin* receives a random deviate and returns $\theta$:

```
function theta = randlin (a,b,alpha,r)

%===
% map a uniform deviate (r) to the
% interval [a, b] with a linear pdf
% with slope beta
%===

L = b-a;

if(abs(beta)<0.0000001)
 theta = a + L*r;
else
 tmp = 0.5*L-1.0/(beta*L);
 theta = a + tmp + sqrt(tmp^2+2*r/beta);
end

%---
% done
%---

return
```

A straightforward generalization provides us with a function that returns an array of values $\theta$ consistent with the linear pdf.

### 4.3.3 Exponential distribution

Consider the exponential probability density function defined over the entire $\theta$ axis, for $-\infty < \theta < \infty$,

$$\phi(\theta) = \tfrac{1}{2}\,\mu\,\exp\big(-\mu\,|\theta|\big), \tag{4.3.14}$$

where $\mu$ is a positive constant. Using equation (4.1.15), we set

$$\tfrac{1}{2}\,\mu\int_{-\infty}^{\theta}\exp\big(-\mu\,|u|\big)\,\mathrm{d}u = \varrho. \tag{4.3.15}$$

Performing the integration, we obtain

$$\tfrac{1}{2}\,\exp(\mu\,\theta) = \varrho \tag{4.3.16}$$

for $\theta < 0$, and

$$1 - \tfrac{1}{2}\exp(-\mu\,\theta) = \varrho \tag{4.3.17}$$

for $\theta > 0$. Rearranging, we obtain

$$\theta = \begin{cases} \dfrac{1}{\mu}\ln(2\varrho) & \text{for} \quad 0 < \varrho < \tfrac{1}{2}, \\ -\dfrac{1}{\mu}\ln[2(1-\varrho)] & \text{for} \quad \tfrac{1}{2} < \varrho < 1. \end{cases} \tag{4.3.18}$$

A $\theta$ sequence arises from a $\varrho$ sequence by way of these expressions.

### 4.3.4 Gaussian distribution

The Gaussian distribution, also known as the normal distribution, with mean value $\overline{\theta}$ and standard deviation $\sigma$, is given by

$$\phi(\theta) = \frac{1}{\sqrt{2\pi}\,\sigma}\exp\Big[-\tfrac{1}{2}\Big(\frac{\theta-\overline{\theta}}{\sigma}\Big)^2\Big], \tag{4.3.19}$$

where the random variable $\theta$ ranges over the entire real axis.

The general procedure described previously in this section results in an intractable algebraic equation, $\Phi(\theta) = \varrho$, involving the error function. The algebraic equation is also impractical to invert by numerical methods. Matlab includes an internal function *randn* that generates random numbers with zero mean, $\overline{\theta} = 0$, and unit standard deviation, $\sigma = 1$.

### 4.3.5 Gaussian sampling by way of the central-limit theorem

As an alternative, a random sequence for $\theta$ can be obtained by exploiting an implication of the central-limit theorem.

First, we select a group of random numbers with a uniform probability distribution in the interval $[0, 1]$, $\varrho_i^{(1)}$ for $i = 1, \ldots, \nu$, and compute the number

$$\theta_1 = \overline{\theta} + \sigma \sum_{i=1}^{\nu} (\,\varrho_i^{(1)} - \tfrac{1}{2}\,). \tag{4.3.20}$$

Second, we select another group of numbers, $\varrho_i^{(2)}$ for $i = 1, 2, \ldots, \nu$, and compute the corresponding number $\theta_2$.

The selection and calculation are repeated many times.

As $\nu$ becomes increasingly large, the random variables, $\theta_i$, obey a Gaussian distribution with mean value $\overline{\theta}$ and standard deviation $\sigma$. In practice, setting $\nu = 12$ yields a satisfactory Gaussian-like behavior.

***Exercise***

**4.3.1** Derive the expression for the variance shown in (4.3.3).

## *4.4 Continuous data*

Consider a continuous random variable, $x$, associated with an event, $\circ$. For example, the event $\circ$ can be receiving a phone call, and $x$ can be the time the call is made.

Now we express the differential probability that $x$ lies between $x$ and $x + \mathrm{d}x$ as

$$\mathrm{d}\Phi_{(x\,|\,\circ)} = \phi_{(x\,|\,\circ)}\,\mathrm{d}x, \tag{4.4.1}$$

where

- $\phi_{(x\,|\,\circ)}$ is a conditional probability density function (pdf)
- $\Phi_{(x\,|\,\circ)}$ is the associated cumulative probability density function (cpdf)

The domain of definition of $x$ can be extended by setting $\phi_{(x\,|\,\circ)}$ to zero beyond the boundaries of the native range (support) where $\phi_{(x\,|\,\circ)}$ is non-zero.

Since the event $\circ$ is associated with some value of $x$, the differential probabilities must add to unity over an appropriate range of $x$,

$$\int \phi_{(x\,|\,\circ)}\,\mathrm{d}x = 1. \tag{4.4.2}$$

### *4.4.1 Bayes theorem*

Bayes equation requires that

$$\Pi_{(\circ\,|\,x)} = \frac{\mathrm{d}\Phi_{(x\,|\,\circ)}}{\mathrm{d}\Phi_{(x)}} \times \Pi_{(\circ)}. \tag{4.4.3}$$

Using (4.4.1) and simplifying, we obtain

$$\Pi_{(\circ\,|\,x)} = \frac{\phi_{(x\,|\,\circ)}}{\phi_{(x)}} \times \Pi_{(\circ)}, \tag{4.4.4}$$

where $\phi_{(x)}$ is marginal pdf.

### *4.4.2 An event and its complement*

Now we consider the complement of the event, $\bar{\circ}$, which along with $\circ$ defines a complete sample space. The corresponding Bayes equation is

$$\Pi_{(\bar{\circ}\,|\,x)} = \frac{\phi_{(x\,|\,\bar{\circ})}}{\phi_{(x)}} \times \Pi_{(\bar{\circ})}, \tag{4.4.5}$$

where the marginal pdf is given by

$$\phi_{(x)} = \phi_{(x\,|\,\circ)}\,\Pi_{(\circ)} + \phi_{(x\,|\,\bar{\circ})}\,\Pi_{(\bar{\circ})}, \tag{4.4.6}$$

subject to the constraint $\Pi_{(\circ)} + \Pi_{(\bar{\circ})} = 1$. The mandatory normalization conditions

$$\int \phi_{(x\,|\,\circ)}\,\mathrm{d}x = 1, \qquad \int \phi_{(x\,|\,\bar{\circ})}\,\mathrm{d}x = 1, \qquad \int \phi_{(x)}\,\mathrm{d}x = 1, \tag{4.4.7}$$

are satisfied.

### *4.4.3 Likelihoods*

The pdfs $\phi_{(x\,|\,\circ)}$ and $\phi_{(x\,|\,\bar{\circ})}$ are the likelihoods of $x$ pertaining to the event and its complement,

$$\mathcal{L}_x(\circ) \equiv \phi_{(x,\circ)}, \qquad \mathcal{L}_x(\bar{\circ}) \equiv \phi_{(x,\bar{\circ})}. \tag{4.4.8}$$

Bayes equations (4.4.4) and (4.4.5) become

$$\Pi_{(\circ\,|\,x)} = \frac{\mathcal{L}_x(\circ)}{\mathcal{L}_x} \times \Pi_{(\circ)}, \tag{4.4.9}$$

and

$$\Pi_{(\bar{\circ}\,|\,x)} = \frac{\mathcal{L}_x(\bar{\circ})}{\mathcal{L}_x} \times \Pi_{(\bar{\circ})}, \tag{4.4.10}$$

where

$$\mathcal{L}_x = \mathcal{L}_x(\circ)\,\Pi_{(\circ)} + \mathcal{L}_x(\bar{\circ})\Pi_{(\bar{\circ})} \tag{4.4.11}$$

is a marginal likelihood.

### *4.4.4 Tomatoes*

Denote by $x$ the daily amount of water that a farmer delivers to a tomato plant according to a pdf, $\phi(x)$. The event that a tomato plant thrives is denoted by $\circ$, and the event that the tomato plant withers is denoted by $\bar{\circ}$.

The likelihood $\phi_{(x\,|\,\circ)}$ is the pdf of the amount of water delivered to a plant that thrived, exhibiting a peak at the recommended amount, as shown in Figure 4.4.1. The likelihood $\phi_{(x\,|\,\bar{\circ})}$ is the pdf of the amount of water delivered to a plant that withered, exhibiting a peak at low values of $x$ and another peak at high values of $x$, as shown in Figure 4.4.1. The amount of water is restricted in the range between $0$ and $2$ liters.

Equations (4.4.4) and (4.4.5) provide us with the probabilities that a plant that was watered by an $x$ amount will survive or wither this year. The prior $\Pi_{(\circ)}$ could be the fraction of plants that survived last year, and the prior $\Pi_{(\bar{\circ})}$ could be the fraction of plants that withered last year.

Referring to Figure 4.4.1 and following the vertical broken line drawn at $1$ lt, we read

$$\mathcal{L}_{1\,lt}(\circ) = 3.5, \qquad \mathcal{L}_{1\,lt}(\bar{\circ}) = 0.5. \tag{4.4.12}$$

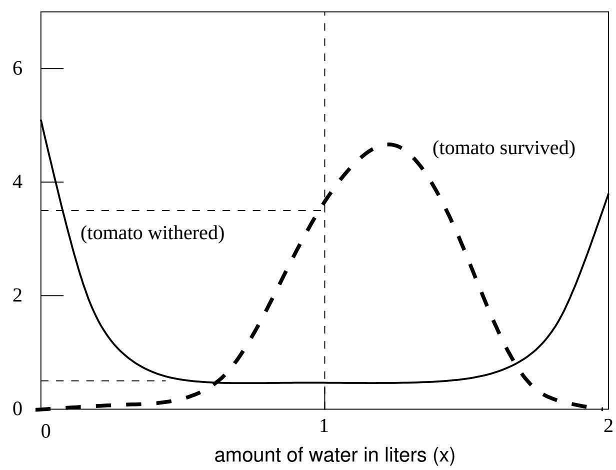



FIGURE 4.4.1 Probability density functions (pdf) of the daily amount of water in liters ($x$) delivered to a tomato plant that survived plant or withered, playing the role of likelihood functions. The area under each curve is equal to unity.

Bayes rule prescribes that

$$\Pi_{(\circ\,|\,1\,lt)} = \frac{\mathcal{L}_{1\,lt}(\circ)}{\mathcal{L}_{1\,lt}} \times \Pi_{(\circ)} \tag{4.4.13}$$

and

$$\Pi_{(\bar{\circ}\,|\,1\,lt)} = \frac{\mathcal{L}_{1\,lt}(\bar{\circ})}{\mathcal{L}_{1\,lt}} \times \Pi_{(\bar{\circ})}, \tag{4.4.14}$$

where

$$\mathcal{L}_{1\,lt} = \mathcal{L}_{1\,lt}(\circ) \times \Pi_{(\circ)} + \mathcal{L}_{1\,lt}(\bar{\circ}) \times \Pi_{(\bar{\circ})}. \tag{4.4.15}$$

For $\Pi_{(\circ)} = 0.8$ and $\Pi_{(\bar{\circ})} = 0.2$, we find that $\mathcal{L}_{1lt} = 2.9$, and then

$$\Pi_{(\circ\,|\,1\,lt)} = 0.966, \qquad \Pi_{(\bar{\circ}\,|\,1\,lt)} = 0.034. \tag{4.4.16}$$

The tomato plant has an excellent chance of surviving at the watering volume of $1\ lt$ this year.

### 4.4.5 Dosage of chemotherapy

The tomato example can be reluctantly rephrased in terms of an ill person who receives a certain amount of radiation or medication. Dosage of chemotherapy can be difficult to decide: If the dose is too low, it will be ineffective against a tumor, whereas, at excessive doses, the toxicity (side-effects) will be intolerable to the person receiving it (from https://en.wikipedia.org/wiki/Chemotherapy#Dosage).

### 4.4.6 Odds ratio

Dividing side by side equations (4.4.9) and (4.4.10), we obtain

$$\mathcal{O} = \mathcal{B}\mathcal{O}_{\rm prior}, \tag{4.4.17}$$

where

$$\mathcal{O} \equiv \frac{\Pi_{(\circ\,|\,x)}}{\Pi_{(\bar{\circ}\,|\,x)}} = \frac{\Pi_{(\circ\,|\,x)}}{1-\Pi_{(\circ\,|\,x)}} \tag{4.4.18}$$

is the odds ratio,

$$\mathcal{B} \equiv \frac{\mathcal{L}_x(\circ)}{\mathcal{L}_x(\bar{\circ})} \tag{4.4.19}$$

is the Bayes factor, and

$$\mathcal{O}_{\rm prior} \equiv \frac{\Pi_{(\circ)}}{\Pi_{(\bar{\circ})}} = \frac{\Pi_{(\circ)}}{1-\Pi_{(\circ)}} \tag{4.4.20}$$

is the prior odds ratio. From the second equality in (4.4.18), we find that

$$\Pi_{(\circ\,|\,x)} = \frac{\mathcal{O}}{1+\mathcal{O}} = \frac{\mathcal{B}\mathcal{O}_{\rm prior}}{1+\mathcal{B}\mathcal{O}_{\rm prior}} = \frac{1}{1+1/(\mathcal{B}\mathcal{O}_{\rm prior})}, \tag{4.4.21}$$

which is zero only when $\mathcal{B} = 0$ or $\mathcal{O}_{\rm prior} = 0$.

### 4.4.7 Logistic regression

Without loss of generality, we may set

$$\mathcal{O}_{\rm prior} = \exp(\beta_0), \tag{4.4.22}$$

where $\beta_0 = \ln \mathcal{O}_{\text{prior}}$ is a positive parameter. Next, we introduce the empirical approximation

$$\mathcal{B} = \exp(-\alpha_1(x - x_0)), \tag{4.4.23}$$

where $\alpha_1$ and $x_0$ are two constants. Combining the last two expressions, we find that

$$\mathcal{B}\mathcal{O}_{\text{prior}} = \exp(\beta_0 - \alpha_1(x - x_0)). \tag{4.4.24}$$

Substituting this approximation into (4.4.21), we obtain the logistic regression formula

$$\Pi_{(\bigcirc)\,|\,x)} = \frac{1}{1 + \exp(-\alpha_1 x - \alpha_0)}, \tag{4.4.25}$$

where

$$\alpha_0 = \beta_0 - \alpha_1 x_0. \tag{4.4.26}$$

The empiricism underlying the approximation (4.4.23) is reason enough to raise eyebrows concerning the purity of this Bayesian approach.

***Exercises***

**4.4.1** Discuss Bayes equation for an item that sold or did not sell as a function of its price.

**4.4.2** Rephrase the tomato example in terms of a tasty or non-tasty cake, where $\theta$ is the amount of sugar added to the ingredients.

## *4.5 Phone calls*

Let the $N$ mutually exclusive events $\circ_1, \circ_2, \ldots, \circ_N$ define a sample space, that is, one of these events *has* to occur so that the sum of the associated probabilities is unity,

$$\sum_{i=1}^{N} \Pi_{(\circ_i)} = 1. \tag{4.5.1}$$

The counterpart of Bayes equation (4.4.4) is

$$\Pi_{(\circ_i\,|\,x)} = \frac{\mathcal{L}_x(\circ_i)}{\mathcal{L}_x} \times \Pi_{(\circ_i)} \tag{4.5.2}$$

for $i = 1, \ldots, N$, where

$$\mathcal{L}_x = \sum_{j=1}^{N} \mathcal{L}_x(\circ_j)\, \Pi_{(\circ_j)} \tag{4.5.3}$$

is a marginal likelihood. We see that

$$\sum_{i=1}^{N} \Pi_{(\circ_i\,|\,x)} = 1, \tag{4.5.4}$$

as required.

Now we assume that the pdf $\mathcal{L}_x(\circ_1)$ describes the probability that a telemarketer calls at time $x$, which is expected to peak at dinner time, as shown in Figure 4.5.1. The pdf $\mathcal{L}_x(\circ_2)$ describes the probability that a collection agency calls at time $x$, which is expected to be non-zero during business hours, as shown in Figure 4.5.1. The pdf $\mathcal{L}_x(\circ_3)$ describes the probability of another person calling at time $x$, which is expected to be significant before midnight, as shown in Figure 4.5.1.

Equation (4.5.2) with $N = 3$ provides us with the probability that a call received at time $x$ is from a telemarketer, a collection agency, or someone else. The prior probabilities $\Pi_{(\circ_i)}$ can be identified with the fraction of $i$th phone calls received over a twenty-four hour period on a previous day, for $i = 1, 2, 3$. Absent this information, the uninformative probabilities $\Pi_{(\circ_i)} = 1/3$ for $i = 1, 2, 3$ can be used.

### 4.5.1 A phone call at noon

For example, if we receive a phone call at noon, we refer to Figure 4.5.1 and read

$$\mathcal{L}_{12:00}(\circ_1) = 1.5, \quad \mathcal{L}_{12:00}(\circ_2) = 4.0, \quad \mathcal{L}_{12:00}(\circ_3) = 2.0. \tag{4.5.5}$$

Bayes rule requires that

$$\Pi_{(\circ_i\,|\,12:00)} = \frac{\mathcal{L}_{12:00}(\circ_i)}{\mathcal{L}_{12:00}} \times \Pi_{(\circ_i)} \tag{4.5.6}$$

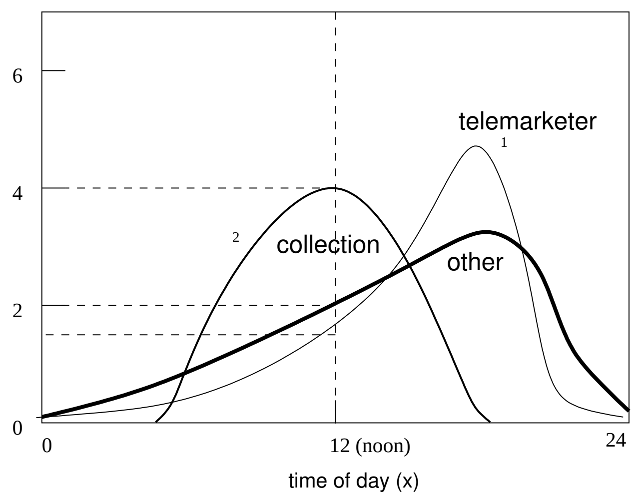



FIGURE 4.5.1 Probability density functions (pdf) of time of phone calls, playing the role of likelihood functions for telemarketers, collection agencies, or other. The area under each curve is equal to unity.

for $i = 1, 2, 3$, where

$$\mathcal{L}_{12:00} = \mathcal{L}_{12:00}(\circ_1) \times \Pi_{(\circ_1)} + \mathcal{L}_{12:00}(\circ_2) \times \Pi_{(\circ_2)} + \mathcal{L}_{12:00}(\circ_3) \times \Pi_{(\circ_3)} \tag{4.5.7}$$

is the likelihood that the phone rings at noon. For prior probabilities $\Pi_{(\circ_1)} = \frac{1}{3}$, $\Pi_{(\circ_2)} = \frac{1}{3}$, and $\Pi_{(\circ_3)} = \frac{1}{3}$, we find that $\mathcal{L}_{12:00} = 2.5$, and then

$$\Pi_{(\circ_1 \,|\, 12:00)} = 0.200, \qquad \Pi_{(\circ_2 \,|\, 12:00)} = 0.533, \qquad \Pi_{(\circ_3 \,|\, 12:00)} = 0.267. \tag{4.5.8}$$

Regrettably, it appears that the phone call is from a collection agency.

On the other hand, if no debt is owed, we set $\Pi_{(\circ_1)} = \frac{1}{2}$, $\Pi_{(\circ_2)} = 0$, and $\Pi_{(\circ_3)} = \frac{1}{2}$, and find that $\mathcal{L}_{12:00} = 1.75$ and then

$$\Pi_{(\circ_1 \,|\, 12:00)} = 0.429, \qquad \Pi_{(\circ_2 \,|\, 12:00)} = 0.000, \qquad \Pi_{(\circ_3 \,|\, 12:00)} = 0.571. \tag{4.5.9}$$

Fortunately, it appears that the phone call is from a charming Massey–Ferguson dealer.

***Exercise***

**4.5.1** In the phone call example, add another potential caller of your choice ($N = 4$) and discuss the corresponding likelihood function.

## 4.6 Riding Harleys

Two retired professors are sitting near a pond next to a country road on a beautiful Sunday morning. There is a biker club in the town nearby, and the bikers own exclusively Sportsters and Tourings. The bikers met for breakfast early in the morning and are now riding down the country road past the professors. The professors are fascinated by the bikes and regret their prior complacency and conservative lifestyle.

### 4.6.1 Waiting for God

In their conversations about the essence of a true scholar, the professors agreed that the standard academic publishing scheme has a single idea at the apex, four refereed journal papers at the first level, sixteen conference abstracts at the second level, and sixty-four or more seminar presentations at the third level.

True scholars understand that one paper per idea is plenty. Brilliant artists generate a masterpiece, and then move on to generate another masterpiece in due time. True thinkers remain silent for a long periods of time and speak only when they have something important to say.

### 4.6.2 Acoustic signatures

Over the summer, the professors saved part of their pensions to buy a decibel meter so that they can study the acoustic signatures of the machines. They found that the Sportsters are generally louder than the Tourings, but the sound level at a certain distance also depends on the speed of the bike, the personality of the driver, the presence of the passengers, and the desire to impress the passengers.

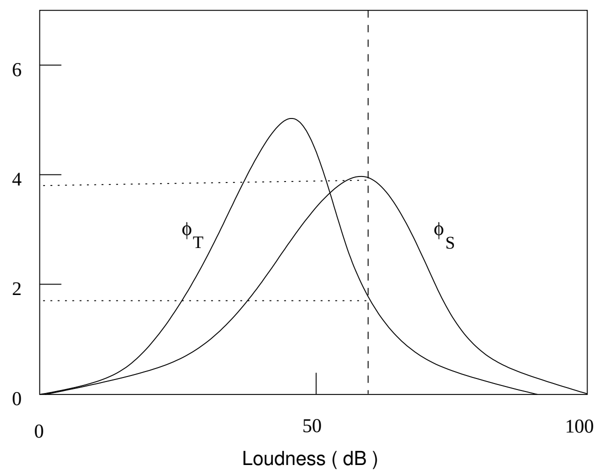



FIGURE 4.6.1 Probability density functions (pdf) of sound loudness measured in dB, playing the role of likelihood functions for Sportsters and Tourings. The area under each curve is equal to unity.

The professors measured the sound from a large number of Sportsters and a large number of Tourings and generated the corresponding loudness probability density functions, $\phi_{\rm T}(\ell)$ and $\phi_{\rm S}(\ell)$, where $\ell$ stands for loudness, as shown in Figure 4.6.1. Normalization requires that

$$\int_0^\infty \phi_{\rm T}(\ell)\,{\rm d}\ell = 1, \qquad \int_0^\infty \phi_{\rm S}(\ell)\,{\rm d}\ell = 1, \tag{4.6.1}$$

which means that area under each curve in the figure is equal to unity.

The professors did visit the bike dealer fifty miles away in a timid attempt to familiarize themselves with the equipment, and found by asking tangential questions that $60\%$ of the bikers ride Tourings and $40\%$ of the bikers ride Sportsters. The bike dealer and her tattoos couldn't have been nicer.

### *4.6.3 Breaking free*

As the professors are sitting near the water watching the tadpoles and contemplating the random nature of events, they hear a bike approaching at a distance. Is it a Sportster or a Touring?

The professors agree on an impulse that if it is a Sportster, they will each buy a Sportster on the next day. If it is a Touring, they will both buy a Touring on the next day. In either case, they will cover their fragile interior with leather and enjoy the comradeship of the biker club.

With hands trembling from excitement, the professors pull out the decibel meter and read $60$ dB, which is rather loud. Then they apply Bayes' equation

$$\Pi_{(\circ\,|\,\ell)} = \frac{\phi_{(\ell\,|\,\circ)}}{\phi(\ell)} \times \Pi_{(\circ)} \tag{4.6.2}$$

to write

$$\Pi_{(\texttt{Touring}\,|\,60\,dB)} = \frac{\phi_{\rm T}(60\;db)}{\phi(60\;db)} \times \Pi_{(\texttt{Touring})}, \tag{4.6.3}$$

where

$$\phi(60\;db) = \phi_{\rm T}(60\;db) \times \Pi_{(\texttt{Touring})} + \phi_{\rm S}(60\;db) \times \Pi_{(\texttt{Sportster})} \tag{4.6.4}$$

is a marginal probability. Based on information from the dealer,

$$\Pi_{(\texttt{Touring})} = 0.6, \qquad \Pi_{(\texttt{Sportster})} = 0.4. \tag{4.6.5}$$

Based on the data plotted in Figure 4.6.1,

$$\phi_{\rm T}(60\;db) = 1.7 \qquad \phi_{\rm S}(60\;db) = 3.8. \tag{4.6.6}$$

Substituting these numbers into Bayes equation, the professors find that

$$\Pi_{(\texttt{Touring}\,|\,60\,dB)} = \frac{1.7}{1.7 \times 0.6 + 3.8 \times 0.4} \times 0.6 = 0.4016, \tag{4.6.7}$$

and thus,

$$\Pi_{(\texttt{Sportster}\,|\,60\,dB)} = 1 - \Pi_{(\texttt{Touring}\,|\,60\,dB)} = 0.5984. \tag{4.6.8}$$

The odds ratio is

$$\mathcal{O} \equiv \frac{\Pi_{(\texttt{Sportster}\,|\,60\,dB)}}{\Pi_{(\texttt{Touring}\,|\,60\,dB)}} = \frac{\phi_{\rm S}(60\;db)}{\phi_{\rm T}(60\;db)} \times \frac{\Pi_{(\texttt{Sportster})}}{\Pi_{(\texttt{Touring})}}, \tag{4.6.9}$$

yielding

$$\mathcal{O} = \frac{3.8}{1.7} \times \frac{0.4}{0.6} = 1.490, \tag{4.6.10}$$

which indicates that the bike is a Sportster. Indeed, as the bike approaches, it clearly appears to be a Sportster.

### 4.6.4 Sleeping on cots

On the very next day, at the crack of dawn on a crisp Monday morning, the professors visited the dealership and purchased a sparkling white and a bright red Sportster. Both professors slept on cots in their garage next to their bikes for several days. Sometimes, a shinny new bike is all you need to stop the bleeding of old wounds, at least for a while.

#### Exercise

**4.6.1** Which Harley would the professors have bought if the decibel meter read $45$?

## 4.7 The $h$ index

The feeling of happiness is hard to describe, though the state of happiness can be detected by several diagnostics: empathy for others, a positive attitude in life, and, most important, the urge to work hard, cook nice meals, laugh at life's absurdities, smile at Rodney Dangerfield's jokes, and enjoy George Carlin's insights: "*Inside every cynical person, there is a disappointed idealist.*"

### 4.7.1 TCS

The Tax Collection Service has kindly asked taxpayers to indicate on their tax returns their degree of happiness, on a scale from $0$ to $10$. The Service then divided the taxpayer income into nine brackets, as shown in Table 4.7.1. For each bracket, $\mathrm{X}$, the Service analyzed the data to generate a probability density function (pdf) pertinent to the bracket, $\phi_{\mathrm{X}}(h)$, with regard to a scaled happiness index, $h$, that varies between $0$ and $1$.

| Bracket | Income in US Dollars |
|---|---|
| A | $0 - 20,000$ |
| B | $20,000 - 40,000$ |
| C | $40,000 - 60,000$ |
| D | $60,000 - 80,000$ |
| E | $80,000 - 100,000$ |
| F | $100,000 - 150,000$ |
| G | $150,000 - 300,000$ |
| H | $300,000 - 600,000$ |
| I | $600,000 - \infty$ |

TABLE 4.7.1 Income brackets per taxpayer identified by the income Tax Collection Service for the purpose of assessing a happiness index, $h$, ranging from zero for extremely unhappy to unity for extremely happy.

### 4.7.2 Happiness below poverty level

The pdf of the lowest bracket, $\phi_{\rm A}(h)$, is shown in Figure 4.7.1(*a*). The area under this curve is equal to unity.

$$\int_0^\infty \phi_{\rm A}(h)\,{\rm d}h = 1. \tag{4.7.1}$$

We observe that taxpayers far below the poverty level are mostly unhappy due to uncertainties on whether they and their loved ones will be able to afford their next meal.

Some of these taxpayers are reasonably happy for reasons that have absolutely nothing to do with monetary assets. Double-blind studies have shown that believing in God or another High Authority, and having a high regard for spirituality contribute to a person's happiness a great deal.

### 4.7.3 Happiness of the rich

The pdf of the highest bracket, $\phi_{\rm I}(h)$, is shown in Figure 4.7.1(*b*). The

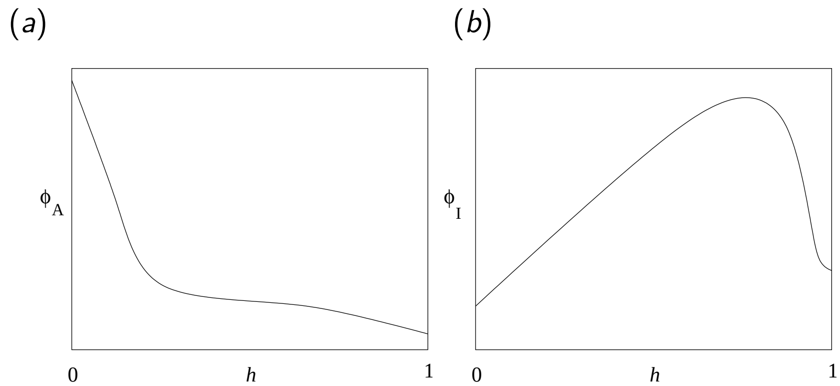



FIGURE 4.7.1 Probability density function (pdf) of a happiness index, $h$,, for (*a*) taxpayers far below the poverty line, $\phi_{\mathrm{A}}$, and (*b*) taxpayers that are filthy rich, $\phi_{\mathrm{I}}$.

area under this curve is equal to unity,

$$\int_0^\infty \psi_{\mathrm{I}}(h)\,\mathrm{d}h = 1. \qquad (4.7.2)$$

We observe that extremely wealthy taxpayers are mostly happy due to an elevated self-esteem, a sense of accomplishment, and the relentless flattery and affirmation by others.

Some of these taxpayers are unhappy for reasons that have absolutely nothing to do with monetary assets. Double-blind studies funded by public agencies have shown that glowing in the dark and overcooking the bacon causes distress.

### *4.7.4 Money can't buy happiness*

The Service observed that the pdfs for brackets E–I are essentially the same. They disclosed this interesting finding to the media, and the discovery spawned a chain of mind-numbing news releases entitled "Money can't buy happiness." In most of these news releases, the word "happiness" was misspelled or else replaced by the spell-checker by "hapless".

Having at our disposal the nine pdfs, $\phi_{\mathrm{X}}(h)$, we want to predict the income bracket of a taxpayer who informs us of their degree of happiness, where $\mathrm{X}$ denotes the income bracket. The prediction is done in terms of the conditional probability,

$$\Pi_{(\mathrm{X}\,|\,h)}. \tag{4.7.3}$$

Bayes' equation requires that

$$\Pi_{(\mathrm{X}\,|\,h)} = \frac{\phi_{\mathrm{X}}(h)}{\phi(h)} \times \Pi_{(\mathrm{X})}, \tag{4.7.4}$$

where

$$\phi(h) = \phi_{\mathrm{A}}(h)\,\Pi_{(\mathrm{A})} + \cdots + \phi_{\mathrm{I}}(h)\,\Pi_{(\mathrm{I})} \tag{4.7.5}$$

is a marginal probability. In the language of Bayesian statistics, the pdfs $\phi_{\mathrm{X}}(h)$ provided by the Service are likelihood functions.

### *4.7.5 La Jolla, San Diego, California*

A patron of a pricey restaurant in La Jolla, San Diego, California, discloses to the waiter her $h$ index. It is natural to assume that the patron falls into the highest brackets by setting $\Pi_{(\mathrm{X})} = 0$ for $\mathrm{X} = \mathrm{A}{-}\mathrm{F}$, and then

$$\Pi_{(\mathrm{G})} = 0.25, \qquad \Pi_{(\mathrm{H})} = 0.50, \qquad \Pi_{(\mathrm{I})} = 0.25. \tag{4.7.6}$$

Substituting these speculative estimates into (4.7.4), we obtain a revised probability distribution.

### *4.7.6 Being presumptuous*

We have fallen victims of an insidious cognitive bias. A patron of an expensive restaurant could be a person with modest income who saved her money over the last year to celebrate her spouse's retirement. Or the patron could have decided to spend all of her savings to treat a friend who was just diagnosed with a serious and possibly life-threatening illness to a nice meal.

### *4.7.7 Another $h$ index*

Professors and institutions derive a great deal of pleasure and esteem by counting the number of times their publications have been cited in the literature.

An accepted measure of success is the $h$ index, defined as the number of publications that have been cited a least $h$ times. A professor or researcher with a low $h$ index suffers from professional frustration and low self-esteem manifested by erratic behavior and the failing of graduate students supervised by their colleagues.

#### *Exercise*

**4.7.1** Discuss whether the happiness $h$ index is related to the citation index of those who publish.

## *4.8 Continuous events*

The $N$ mutually exclusive events $\circ_i$ covering a sample space introduced in Section 4.4 can be labeled by a parameter $\theta_i$ that varies monotonically from the first to the last event. Bayes equation may then be stated as

$$\Pi_{(\theta_i\,|\,x)} = \frac{\mathcal{L}_x(\theta_i)}{\sum_{j=1}^{N} \mathcal{L}_x(\theta_j) \times \Pi_{(\theta_j)}} \times \Pi_{(\theta_i)} \qquad (4.8.1)$$

for $i = 1, \ldots, N$, where $x$ is the continuously varying data, $\mathcal{L}_x(\theta_j)$ are the likelihoods satisfying

$$\int_{\mathcal{I}_i} \mathcal{L}_x(\theta_i)\,\mathrm{d}x = 1, \qquad (4.8.2)$$

and the integration is performed over an appropriate integration domain.

### *4.8.1 From sums to integrals*

In the event that $\theta_i$ are separated by equal intervals, $\Delta\theta$, that may generally depend on $x$, we set

$$\Pi_{(\theta_i)} = \phi(\theta_i)\,\Delta\theta \tag{4.8.3}$$

and

$$\Pi_{(\theta_i\,|\,x)} = \phi(\theta_i\,|\,x)\,\Delta\theta, \tag{4.8.4}$$

where $\phi(\theta_i)$ and $\phi(\theta_i\,|\,x)$ are discrete representation of probability density functions, and obtain

$$\phi(\theta_i\,|\,x) = \frac{\mathcal{L}_x(\theta_i)}{\Delta\theta\ \sum_{j=1}^{N} \mathcal{L}_x(\theta_j) \times \phi(\theta_j)} \times \phi(\theta_i). \tag{4.8.5}$$

When $N$ is large, the denominator on the right-hand side of (4.8.5) can be replaced with an integral to obtain Bayes rule for a pdf with continuous data and a continuous spectrum of events parametrized by $\theta$,

$$\phi(\theta\,|\,x) = \frac{\mathcal{L}_x(\theta)}{\displaystyle\int_{\Theta(x)} \mathcal{L}_x(\theta') \times \phi(\theta')\,\mathrm{d}\theta'} \times \phi(\theta), \tag{4.8.6}$$

where $\theta'$ is an integration variable and $\Theta(x)$ is the integration domain. In Section 6.1, we will derive (4.8.6) in terms of the joint probability density function for $\theta$ and $x$.

### 4.8.2 *Everyone is calling*

For example, a virtually infinite collection of people who may give you a phone call can be labeled by $\theta$. The pdf that a person gives you a call at time $x$ is the likelihood $\mathcal{L}_x(\theta)$. Equation (4.8.6) provides us with the probability that a phone call received at time $x$ originates from a person labeled $\theta$, subject to a prior distribution, $\phi(\theta)$.

In another example, $\theta$ can be the distance from a cat and $x$ can be the intensity of the cat's meow heard the cat's owner, as discussed in Section 5.8, Further interpretations of equation (4.8.6) will be discussed in the remainder of this chapter and in Chapter 6.

***Exercise***

**4.8.1** Discuss an interpretation of equation (4.8.6) of your choice.

## 4.9 Spinning Wheels, LLC

*Spinning Wheels, LLC* is a copier repair company in rural Ohio with outstanding reputation and consistent five-star reviews. They always return calls within the hour, they treat their customers with respect, and they have a wicked sense of humor that everyone appreciates.

After a copier has been repaired, they leave a few small chocolates and a fortune cookie on the intake tray. One fortune cookie once read "*He who throws mud at others is losing ground.*" Profound or awkward?

### 4.9.1 Repair performance index

To attract further business, *Spinning Wheels, LLC* published a brochure that discusses a repair performance index, $\theta$, varying in the interval $[0, 1]$. The brochure explains that the performance index, $\theta$, is a stochastic (random) variable due to uncertainties in the quality of replacement copier parts, especially brittle parts involving plastic and metallic pieces.

In the old days, metallic replacement parts were made of hard carbon steel; nowdays they are made mostly of low-carbon steel and wear twice as fast.

### 4.9.2 Probability density function (pdf)

The brochure includes a graph of the probability density function (pdf), $\phi(\theta)$, of the performance index, $\theta$, as shown in Figure 4.9.1. The graph shown in this figure is similar to the graphs shown in Figures 3.10.2(*b*) and 3.10.3(*b*). For the particular pdf shown in Figure 4.9.1, the probability that $\theta$ lies in the range $(0.80, 1.0)$ is roughly $0.5$, which is remarkable considering that almost everyone's unfavorable experiences with copy machines.

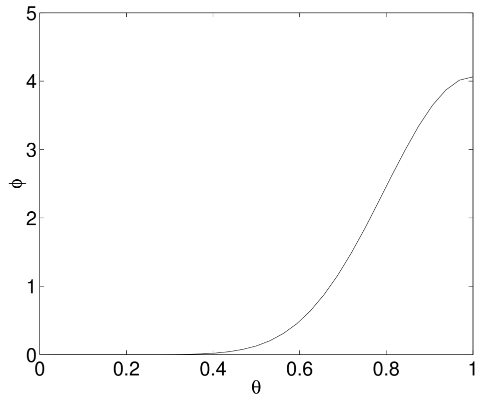



FIGURE 4.9.1 Probability density function (pdf) of a repair performance index, $\theta$, advertised by Spinning Wheels, LLC.

In fact, the probability density function (pdf) shown in Figure 4.9.1 was generated by the following analytical formula:

$$\phi(\theta) = \frac{2 \cdot 4 \cdot 6 \cdot 8 \cdot 10}{1 \cdot 3 \cdot 5 \cdot 7 \cdot 9} \sin^{10}(\tfrac{1}{2}\,\theta\pi) \tag{4.9.1}$$

for $0 \le \theta \le 1$, where a dot denotes multiplication. One may verify readily with the help of tables of definite integrals that this distribution satisfies the constraint (4.1.2) (e.g., Gradshteyn, I. S. & Ryzhik, I. M. (1980) *Table of Integrals, Series, and Products.* Academic Press, p. 369).

To simplify the notation, the product in the numerator on the fraction on the right-hand side could have been written as $10!!$, and the denominator would have been written as $9!!$, where the double exclamation mark has obvious meanings for even or odd integers.

### *4.9.3 Discretization of the pdf and back*

The receptionist of *Heavy Equipment, Inc.* receives a copy of the prospectus and is interested in the graph shown in Figure 4.9.1. To make practical sense of the graph shown in Figure 4.9.1, she divides the $\theta$ abscissa into $n_{\rm div}$ intervals of equal length

$$\Delta\theta = \frac{1}{n_{\rm div}}. \tag{4.9.2}$$

The mid-points of the divisions are located at

$$\theta_i = (i - \tfrac{1}{2}) \times \Delta\theta \tag{4.9.3}$$

for $i = 1, \ldots, n_{\rm div}$. She then assigns to the mid-point of each division a probability

$$\Pi_i = \phi(\theta_i) \times \Delta\theta \tag{4.9.4}$$

for $i = 1, \ldots, n_{\rm div}$, constituting a discrete probability distribution. Consistent with the definition of the pdf, $\Pi_i$ is the probability that $\theta$ falls inside the $i$th division with some small error that depends on the size of the divisions, $\Delta\theta$.

The sum of all probabilities, $\Pi_i$, approximates an integral by way of the *mid-point integration rule*, that is,

$$\sum_{i=1}^{n_{\rm div}} \Pi_i = \Delta\theta \sum_{i=1}^{n_{\rm div}} \phi(\theta_i) \simeq \int_0^1 \phi(\theta)\, {\rm d}\,\theta = 1. \tag{4.9.5}$$

By introducing divisions, the receptionist is able to work with a finite set of discrete probabilities, $\Pi_i$ for $i = 1, \ldots, n_{\rm div}$.

### *4.9.4 The binomial distribution*

The receptionist interprets the repair performance index $\theta$ as the single-copy quality Bernoulli probability. When $\theta = 1$, all copies are good after repair; when $\theta = 0$, all copies are blurry or wrinkled after repair. After a copy machine has been repaired, the probability that $m$ out of $n$ copies are good is given by the binomial distribution, $\mathcal{B}_m^n(\theta)$.

By repeating the analysis of Section 3.10, the receptionist of Heavy Equipment Inc. finds that the generalized formula (3.10.3) takes the form

$$(\Pi_i)_{\rm posterior} = \frac{\mathcal{B}_m^n(\theta_i)}{\sum_{j=1}^{n_{\rm div}} \mathcal{B}_m^n(\theta_j)\, \Pi_j} \times \Pi_i \tag{4.9.6}$$

for $i = 1, \ldots, n_{\text{div}}$. Expressing $\Pi_i$ in terms of $\phi_i \equiv \phi(\theta_i)$ by way of (4.9.4), the receptionist finds that

$$(\phi_i)_{\text{posterior}} = \frac{\mathcal{B}_m^n(\theta_i)}{\Delta\theta \sum_{j=1}^{n_{\text{div}}} \mathcal{B}_m^n(\theta_j)\,\phi_j} \times \phi_i \tag{4.9.7}$$

for $i = 1, \ldots, n_{\text{div}}$.

### *4.9.5 Sum to integral*

The denominator on the right-hand side of (4.9.7) is the mid-point rule approximation to an integral, yielding the pointwise revisions

$$(\phi_i)_{\text{posterior}} = \frac{\mathcal{B}_m^n(\theta_i)}{\int_0^1 \mathcal{B}_m^n(\theta')\,\phi(\theta')\,\mathrm{d}\theta'} \times \phi_i, \tag{4.9.8}$$

where $\theta'$ is an integration variable. This expression leads us to an instance of Bayes' formula for a pdf,

$$\phi(\theta)_{\text{posterior}} = \frac{\mathcal{B}_m^n(\theta)}{\mathcal{L}_m^n} \times \phi(\theta), \tag{4.9.9}$$

where

$$\mathcal{L}_m^n \equiv \int_0^1 \mathcal{B}_m^n(\theta')\,\phi(\theta')\,\mathrm{d}\theta' \tag{4.9.10}$$

is a likelihood function and $\theta$ can have any value in its domain of definition, $[0, 1]$. The discrete data are encapsulated in the doublet $(m, n)$.

### *4.9.6 Numerical simulation*

The receptionist is interested in assessing whether the pdf shown in Figure 4.9.1 is consistent with her own repair records. For this purpose, she goes through her files and reads the number of good copies, $m$, from a total number of copies made, $n$, each time the machine has been repaired in the past five year.

To perform the Bayesian analysis, the receptionist uses equation (4.9.7) to obtain a thread of posterior pdfs, using as initial prior pdf the one shown in Figure 4.9.1, described by (4.9.1). The computations are implemented in the following Matlab code named *spinning*:

```
nr = 5;  % number of repairs

n = [14; 10; 12; 19; 22];
m = [13; 10; 11; 17; 20];

ndiv = 2*32;  % number of divisions

%---
% bad copies per session
%---

for i=1:nr   % run over repairs
 l(i) = n(i)-m(i);
end

%---
% discretization
%---

Dtheta = 1.0/ndiv;

for i=1:ndiv
  theta(i) = (i-0.5)*Dtheta;
end

%---
% initial prior
%---

 fc = 1*3*5*7*9/(2*4*6*8*10);

 for i=1:ndiv
  phi(i) = 1.0/fc*sin(theta(i)*0.5*pi)^10;
 end

%===
% sequential testing over repairs
```

```
% (photocopying sessions)
%===

for r=1:nr  % run over repairs

  accum = 0.0;

  for j=1:ndiv % run over divisions
    L(j) = binomial(m(i),n(i),theta(j));
    accum = accum + L(j)*phi(j);
  end

  accum = accum*Dtheta;

  for i=1:ndiv
   phinew(i) = L(i)*phi(i)/accum;
  end

  phi = phinew;

end
```

The results of the simulation are shown in Figure 4.9.2($a$).

### *4.9.7 MAP*

The *maximum a posteriori estimate* (MAP) is defined as the value of $\theta$ where the posterior probability density function reaches a maximum. The small circular symbols connected with dashed lines represent the cumulative binomial distribution. In this case, the MAP differs only slightly from the maximum likelihood estimate corresponding to the maximum of the cumulative binomial distribution, indicated by the vertical dashed lines.

### *4.9.8 A different initial prior*

The receptionist also performs simulations with a a flat prior distribution, as shown in Figure 4.9.2($b$). The results are qualitatively similar to those presented in Figure 4.9.2($a$). The receptionist plans to email these graphs to *Spinning Wheels, LLC* for their information.

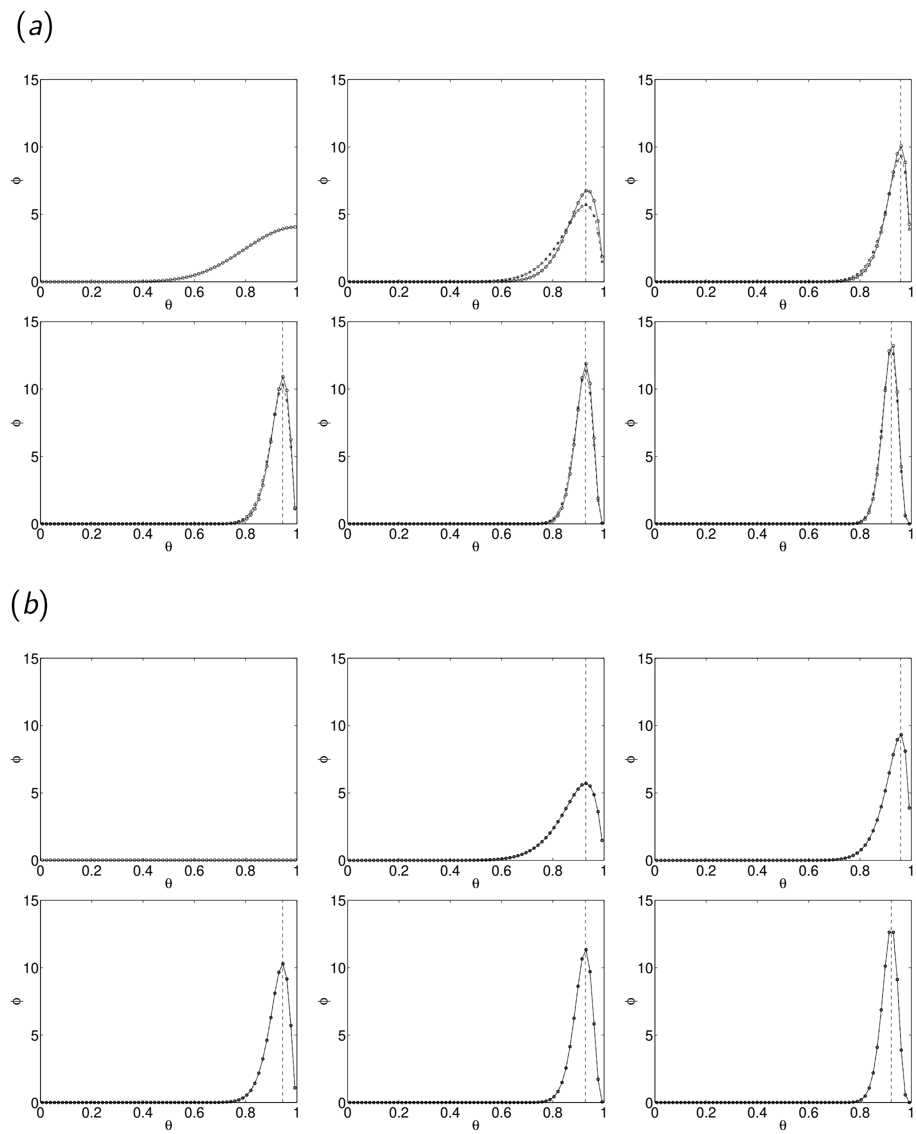



FIGURE 4.9.2 Evolving distribution of the probability density function (pdf) of the photocopy machine repair effectiveness index, $\theta$, for two different prior distributions. The small squares connected by dashed lines represent the cumulative binomial distribution.

***Exercise***

**4.9.1** Reproduce the graphs shown in Figure 4.9.2 for a prior pdf of your choice and discuss the results.

## *4.10 Bayes theorem for pdfs*

Previously in this chapter, we have seen that events and data can occur in four combinations:

- discrete events with discrete data
- discrete events with continuous data
- continuous events with discrete data
- continuous events with continuous data

Discrete probability distributions (dpd) and probability density functions (pdf) are employed for events and data likelihoods, as appropriate.

### *4.10.1 Discrete events*

Bayes rule for the probabilities of $N$ discrete events with a discrete or continuous data array, $\mathbf{x}$, reads

$$\Pi(\circ_i \,|\, \mathbf{x}) = \frac{\mathcal{L}_{\mathbf{x}}(\circ_i)}{\mathcal{L}_{\mathbf{x}}} \times \Pi_{(\circ_i)} \tag{4.10.1}$$

for $i = 1, \ldots, N$, where the events, $\circ_i$, cover a complete sample space and

$$\mathcal{L}_{\mathbf{x}} \equiv \sum_{j=1}^{N} \mathcal{L}_{\mathbf{x}}(\circ_j) \times \Pi_{(\circ_j)} \tag{4.10.2}$$

is $\mathcal{L}_{\mathbf{x}}$ is the marginal likelihood.

Bayes equation (4.10.1) relates the posterior discrete probability distribution (dpd), denoted by $\Pi(\circ_i \,|\, \mathbf{x})$, to the prior discrete probability distribution, denoted by $\Pi_{(\circ_i)}$, in terms of a set of suitable likelihood functions, $\mathcal{L}_{\mathbf{x}}(\circ_i)$, sometimes also called partition functions.

Summing equation (4.10.1) with respect to $i$, we confirm the mandatory normalization condition

$$\sum_{i=1}^{N} \Pi(\circ_i \,|\, \mathbf{x}) = 1, \tag{4.10.3}$$

thanks to the denominator of the fraction on the right-hand side of (4.10.1).

### *4.10.2 Likelihood function*

The likelihood function, $\mathcal{L}_{\mathbf{x}}(\circ_i)$, is the probability that the observed data, $\mathbf{x}$, arise from all possible data that could present themselves for the $i$th event, $\circ_i$, as discussed in Section 4.5. In the case of discrete data, $\mathcal{L}_{\mathbf{x}}(\circ_i)$ is a discrete probability distribution (dpd) with respect to $\mathbf{x}$. In the case of continuous data, $\mathcal{L}_{\mathbf{x}}(\circ_i)$ is a probability density function (pdf) with respect to $\mathbf{x}$.

### *4.10.3 Continuous events*

Bayes' formula (4.9.10) is an instance of Bayes' equation for the probability density function (pdf) of a range of continuous events parametrized by a variable, $\theta$, for a discrete or continuous data array, $\mathbf{x}$,

$$\phi(\theta \,|\, \mathbf{x}) = \frac{\mathcal{L}_{\mathbf{x}}(\theta)}{\varphi_{\mathbf{x}}} \times \phi(\theta), \tag{4.10.4}$$

where

$$\varphi_{\mathbf{x}} \equiv \int_a^b \mathcal{L}_{\mathbf{x}}(\theta')\, \phi(\theta')\, \mathrm{d}\theta', \tag{4.10.5}$$

the random variable $\theta$ is defined in the interval $[a, b]$, and $\theta'$ in the denominator is an auxiliary integration variable.

Bayes equation (4.10.4) relates the posterior probability density function, $\phi(\theta \,|\, \mathbf{x})$, to the prior probability density function, $\phi(\theta)$, in terms of a suitable likelihood function, $\mathcal{L}_{\mathbf{x}}(\theta)$, sometimes also called a partition function.

Integrating both sides of equation (4.10.4) with respect to $\theta$ from $a$ to $b$, we confirm the mandatory normalization condition

$$\int_a^b \phi(\theta \,|\, \mathbf{x})\, \mathrm{d}\theta = 1, \tag{4.10.6}$$

thanks to the denominator of the fraction on the right-hand side of (4.10.4).

It is important to bear in mind that the definite integral in the denominator of Bayes' equation (4.10.4) is merely a number that depends on $\mathbf{x}$, whereas the numerator is a function of $\theta$ for given data encapsulated in the array $\mathbf{x}$.

### 4.10.4 Likelihood function

The likelihood function, $\mathcal{L}_{\mathbf{x}}(\theta)$, is the probability that the data, $\mathbf{x}$, arise from any possible data that could be obtained for a certain value of $\theta$, as discussed in Section 4.9. In the case of discrete data, $\mathcal{L}_{\mathbf{x}}(\theta)$ is a discrete probability distribution (dpd) with respect to $\mathbf{x}$. In the case of continuous data, $\mathcal{L}_{\mathbf{x}}(\theta)$ is a probability density function (pdf) with respect to $\mathbf{x}$.

### 4.10.5 Revision by multiplication

Bayes equation (4.10.4) for continuous events updates the prior pdf, $\phi(\theta)$, by multiplying it with the likelihood function. The product is then divided by the value of the definite integral in the denominator. To signify this action, we write

$$\phi(\theta)_{\text{posterior}} \sim \mathcal{L}_{\mathbf{x}}(\theta) \times \phi(\theta), \tag{4.10.7}$$

where the symbol $\sim$ indicates functional proportionality. The product of two functions is another function computed by pointwise multiplication.

Revision by multiplication is the simplest method of updating a function. More involved methods employed in applied mathematics and scientific computing involve convolutions.

### *4.10.6 MAP*

The maximum *a posteriori* estimate (MAP) is the value of $\theta$ where $\phi(\theta)_{\text{posterior}}$ reaches a maximum. When the prior, $\phi(\theta)$, is a flat (uniform) function, the MAP coincides with the maximum of the likelihood function, which is the maximum likelihood estimate, (MLE).

The MAP provides us with a single-estimate prediction, which appears to defy the basic premise of Bayesian statistics as a framework for inclusiveness by allowing all possibilities to survive. In the framework of Bayesian analysis, probability density functions replace pointwise estimates accompanied by confidence intervals.

### *4.10.7 Sequential updating*

Updating a prior pdf after a first data set becomes available, $\mathbf{x}_1$, and then after another data set becomes available, $\mathbf{x}_2$, we obtain

$$\phi(\theta \,|\, \mathbf{x}_1, \mathbf{x}_2) = \frac{\mathcal{L}_{\mathbf{x}_1}(\theta) \times \mathcal{L}_{\mathbf{x}_2}(\theta)}{\varphi_{\mathbf{x}_1,\mathbf{x}_2}} \times \phi(\theta), \tag{4.10.8}$$

where

$$\varphi_{\mathbf{x}_1,\mathbf{x}_2} = \int_a^b \mathcal{L}_{\mathbf{x}_1}(\theta')\, \mathcal{L}_{\mathbf{x}_2}(\theta')\, \phi(\theta')\, \mathrm{d}\theta'. \tag{4.10.9}$$

The order by which the two data sets arrive is immaterial. The prior pdf on the right-hand side, $\phi(\theta)$, is truly the initial prior. The pdf on the left-hand side, $\phi(\theta \,|\, \mathbf{x}_1, \mathbf{x}_2)$, is the second posterior.

Generalization to further data sets is straightforward. The sequence by which the data sets arrive is immaterial.

### *4.10.8 Allegation of snake oil*

Cynics have argued that statistical analysis can be manipulated to yield a desired result with little regard to truth or reality. In Bayesian analysis, a prior pdf is multiplied by a likelihood function to yield a posterior pdf, and then normalized to ensure that the integral of the pdf is unity, as required. When all is said and done, Bayesian inference involves multiplying two functions to get another function. Three questions could be asked:

(*a*) Where does the prior come from when not rigorously derived or intuitively defined?

(*b*) Where does the likelihood function come from when not obvious?

(*c*) How are the data come and are they reliable?

Sometimes the answers to these questions are unclear or unsatisfactory, especially in sociological, financial, and psychological settings.

### *4.10.9 Conjugate pairs*

In Chapter 5, we will discuss conjugate likelihood function-prior pairs. The temptation to use such pairs with no regard to physical or conceptual relevance is encouraged by the availability of these pairs in pull-down menus of Bayesian statistics software. When the boundaries of a rigorous method of analysis are crossed by falling prey to Peter's principle, the method is classified as art at best or snake oil at worst.

#### *Exercise*

**4.10.1** Discuss an example of a continuous stochastic parameter $\theta$ and state the limits of variation.

## *4.11 The likelihood function*

Of all the data that we could have measured pertaining to an event $\circ$, we obtained those encapsulated in a data array $\mathbf{x}$. The likelihood function, $\mathcal{L}_{\mathbf{x}}(\circ)$ for discrete events or $\mathcal{L}_{\mathbf{x}}(\theta)$ for continuous events parametrized by $\theta$, is the probability of the measured data, $\mathbf{x}$, in the sample space of all possible data. The data themselves may take discrete values over a finite or infinite set, or vary continuously over a specified range.

### *4.11.1 Discrete data*

Assuming that an admissible discrete data point, $\mathbf{x}_k$, arises with probability

$$\Pi_{(\mathbf{x}_k)}(\circ) \qquad \text{or} \qquad \Pi_{(\mathbf{x}_k)}(\theta), \tag{4.11.1}$$

respectively, for discrete or continuous events, we set

$$\mathcal{L}_{\mathbf{x}_k}(\circ) = \Pi_{(\mathbf{x}_k)}(\circ) \qquad \text{or} \qquad \mathcal{L}_{\mathbf{x}_k}(\theta) = \Pi_{(\mathbf{x}_k)}(\theta) \tag{4.11.2}$$

for $k = 1, \ldots, n_{data}$, where $n_{data}$ is the number of available data. An infinite number of data is acceptable.

Normalization requires that

$$\sum_{k=1}^{n_{data}} \Pi_{(\mathbf{x}_k)}(\circ) = 1 \qquad \text{or} \qquad \sum_{k=1}^{n_{data}} \Pi_{(\mathbf{x}_k)}(\theta) = 1 \tag{4.11.3}$$

for any $\circ$ or $\theta$. By definition of the likelihood function then,

$$\sum_{k=1}^{n_{data}} \mathcal{L}_{\mathbf{x}_k}(\circ) = 1 \qquad \text{or} \qquad \sum_{k=1}^{n_{data}} \mathcal{L}_{\mathbf{x}_k}(\theta) = 1, \tag{4.11.4}$$

for any $\circ$ or $\theta$.

### *4.11.2 Continuously varying data*

If an admissible continuous data point, $\mathbf{x}$, arises with probability density function $\phi_{\mathbf{x}}(\theta)$, we set

$$\mathcal{L}_{\mathbf{x}}(\circ) = \phi_{\mathbf{x}}(\circ) \qquad \text{or} \qquad \mathcal{L}_{\mathbf{x}}(\theta) = \phi_{\mathbf{x}}(\theta). \tag{4.11.5}$$

Normalization requires that

$$\int \mathcal{L}_{\mathbf{x}}(\circ)\, \mathrm{d}\mathbf{x} = 1 \qquad \text{or} \qquad \int \mathcal{L}_{\mathbf{x}}(\theta)\, \mathrm{d}\mathbf{x} = 1 \tag{4.11.6}$$

for any $\circ$ or $\theta$ over an appropriate integration domain.

### *4.11.3 Binomial likelihood*

In a process with $m$ successes in $n$ trials and associated $n-m$ failures, we obtain discrete data expressed by the doublet $\mathbf{x}_m = (m, n)$ for $m = 0, \ldots, n$. In the case of independent Bernoulli trials, the likelihood function is given by the binomial distribution,

$$\mathcal{L}_{\mathbf{x}}(\circ) = \mathcal{B}_m^n(\lambda) \qquad \text{or} \qquad \mathcal{L}_{\mathbf{x}}(\theta) = \mathcal{B}_m^n(\theta), \tag{4.11.7}$$

where $\lambda$ is a process parameter associated with event $\circ_i$. Normalization requires that

$$\sum_{m=0}^{n} \mathcal{B}_m^n(\lambda) = 1 \qquad \text{or} \qquad \sum_{m=0}^{n} \mathcal{B}_m^n(\theta) = 1 \tag{4.11.8}$$

for any $\lambda$ or $\theta$. An implicit assumption is that the trials are uncorrelated; a necessary condition is the absence of bias. Modifications of the binomial distribution can be employed in the case of non-Bernoulli trials.

### *4.11.4 Correlated events*

Previously in this chapter, we introduced the binomial distribution as a likelihood function with reference to the Bernoulli probability of a photocopy in *Heavy Equipment, Inc.* The binomial distribution expresses the probability that, in a session of $n$ events, such as photocopies, $m$ outcomes $\mathrm{A}$, such as a photocopy is good, and $n-m$ outcomes $\mathrm{B}$, such as a photocopy is bad, will be observed, where $m \leq n$, $\theta$ is the Bernoulli probability of outcome $\mathrm{A}$, and $1-\theta$ is the Bernoulli probability of outcome $\mathrm{B}$.

However, if a bad photocopy leaves a smudge of ink on the up-take roller, the next photocopy will be affected and the events will no longer be uncorrelated. If it rains one day, it is more likely that it will also rain on the next day.

The likelihood function of correlated events is hard to describe due to the general lack of precise information on duration, memory, cause and effect. In practice, we employ non-binomial likelihood functions whose derivation or adoption typically but not always requires a leap of faith. For example, the normal distribution is justified by the central limit theorem for events that arise as the result a large number of random decisions, influences, or pathways.

### *4.11.5 Zero of a likelihood function*

If the likelihood function, $\mathcal{L}_{\mathbf{x}}(\theta)$, is zero at some critical value of $\theta$, the posterior pdf will also be and remain zero at that value, independent of the prior pdf. The reason is that the data cannot possibly appear at the critical value of $\theta$.

### *4.11.6 Dirac delta function*

If $\mathcal{L}_{\mathbf{x}}(\theta)$ is the Dirac delta function forced at a certain value, $\theta_0$, the posterior pdf will also be the Dirac delta function forced at that value. This means that $\theta_0$ is the only possible value that the parameter $\theta$ can take, independent of any evidence. Consequently, the Bayesian updating is deterministic. A similar behavior occurs when $\mathcal{L}_{\mathbf{x}}(\theta)$ is the sum of modulated Dirac delta functions forced at different values $\theta$.

#### *Exercise*

**4.11.1** Discuss an example where the notion of Bernoulli probability fails due to institutional, physical, or cognitive memory effects.

## *4.12 The prior pdf*

The prior probability distribution, $\phi(\theta)$, is assigned to convey knowledge, experience, wisdom, insights, and psychic ability into the physical or conceptual process parametrized by the variable $\theta$.

### *4.12.1 Zeros of the prior*

If the prior, is zero at a certain value of $\theta$, it will remain zero at that value after it has been updated to yield the posterior according to Bayes' equation (4.10.4). This property can be exploited to minimize the difference between the prior and the posterior, and thus enforce a known behavior.

For example, the pdf of the temperature of a cup of coffee, $\theta$, vanishes at the zero of the absolute temperature where the Universe freezes irrespective of the quality of the coffee prepared.

The probability that an arbitrary supervisor will write only good things in an annual employee evaluation is also zero. An arbitrary employee will receive some constructive criticism that will serve as a basis for dismissal at times of financial hardship.

### 4.12.2 Faith-based approach

If the prior pdf, $\phi(\theta)$, is the Dirac delta function in one dimension representing an impulse at some point, the posterior $\phi(\theta|\mathbf{x})$ will also be the same delta function, independent of the likelihood function and the data. This means that no amount of data or evidence will deter a Bayesian operator from making a pre-determined assessment.

The faith-based approach is related to the ostrich effect, the status quo bias, the confirmation bias, and other related cognitive biases (https://en.wikipedia.org /wiki/List_of_cognitive_biases). Stubbornness is compelling evidence of a faith-based approach or attitude at best, and cognitive dissonance at worst. While a rational person will change their opinion, assessment, or conclusions as data become available, an irrational, stubborn, brainwashed, or biased person will remain unmoved.

Potential jurors entering a courtroom with delta functions in their opinions or beliefs are rightly excused by prosecutors or defense attorneys.

### 4.12.3 Uniform prior pdf

In the absence of preconceived notions, bias, experiences, or clues, we adopt a uniform prior pdf,

$$\phi(\theta) = \frac{1}{b-a}, \tag{4.12.1}$$

and obtain the posterior pdf

$$\phi(\theta\,|\,\mathbf{x})_{\text{from uniform}} = \frac{\mathcal{L}_{\mathbf{x}}(\theta)}{\displaystyle\int_a^b \mathcal{L}_{\mathbf{x}}(\theta')\,\mathrm{d}\theta'}, \tag{4.12.2}$$

which is simply the scaled likelihood function. Further data will generate a posterior pdf that generally differs from the likelihood function. Potential jurors entering a courtroom with uniform priors in their opinions or beliefs are accepted as unbiased participants.

### 4.12.4 Inexperience and naivete'

A uniform prior may convey a sense of inexperience or naiveté. Sometimes, a uniform prior is graciously called non-informative. For example, a competent and honest car mechanic has a good idea of what needs to be fixed when engine cylinder #6 misfires, and does not assume that the cause may lie in any component malfunction. By contrast, a dishonest mechanic will suggest replacing all spark plugs, all spark plug wires, and all ignition coils along with the associated electronic modules. The problem will surely be fixed, albeit at an exorbitant and unnecessary cost.

#### Exercise

**4.12.1** Discuss an example where an uninformative prior is appropriate.

## 4.13 Numerical integration and sampling

The computation of the integral in the denominator on the right-hand side of Bayes' equation (4.10.4),

$$\varphi_{\mathbf{x}} \equiv \int_a^b \mathcal{L}_{\mathbf{x}}(\theta')\, \phi(\theta')\, \mathrm{d}\theta', \tag{4.13.1}$$

is an important task. We recall that $\mathcal{L}_{\mathbf{x}}(\theta)$ is the likelihood function, the array $\mathbf{x}$ encapsulates the data, and $\phi(\theta)$ is an available prior. Only in a small number of simple cases we will be able to compute the integral by analytical methods.

Of equal importance is the computation of the posterior expected value

$$\overline{\theta} \equiv \int_a^b \theta\, \phi(\theta\,|\,\mathbf{x})\, \mathrm{d}\theta, \tag{4.13.2}$$

and posterior variance,

$$\sigma^2 \equiv \int_a^b (\theta - \overline{\theta})^2\, \phi(\theta\,|\,\mathbf{x})\, \mathrm{d}\theta \tag{4.13.3}$$

where $\phi(\theta \,|\, \mathbf{x})$ is the posterior pdf.

Fortunately, a variety of numerical methods are available for computing these integrals (e.g., Pozrikidis, C (2008) *Numerical Computation in Science and Engineering*, Second Edition, Oxford University Press.)

### *4.13.1 Mid-point rule*

Consider the definite integral of a function, $f(\theta)$, between two specified limits, $a$ and $b$,

$$\mathcal{J} \equiv \int_a^b f(\theta)\, \mathrm{d}\theta. \qquad (4.13.4)$$

The simplest version of the mid-point rule is implemented by dividing the integration domain, $[a, b]$, into $N$ intervals of equal length, $\Delta\theta = (b - a)/N$, and approximating the graph of the function with a flat function inside each interval to obtain

$$\mathcal{J} \simeq \Delta\theta \sum_{i=1}^{N} f(\theta_i), \qquad (4.13.5)$$

where

$$\theta_i = a + (i - \tfrac{1}{2})\, \Delta\theta \qquad (4.13.6)$$

for $i = 1, \ldots, N$ are the interval mid-points. When the integral is finite and the integrand is sufficiently regular, the accuracy of the computation improves as a higher number of intervals are employed.

### *4.13.2 Gaussian quadrature*

The theory of numerical integration reveals that an integral is best approximated by an appropriate Gauss integration quadrature.

The methodology amounts to replacing the integral with a weighted sum of integrand values at a chosen number of $N_Q$ base points, $\theta_i$ for $i = 1, \ldots, N_Q$,

$$\mathcal{J} \simeq c_Q \sum_{i=1}^{N_Q} f(\theta_i)\, w_i, \qquad (4.13.7)$$

where $c_Q$ is an appropriate coefficient and $w_i$ are integration weights. The base points are distributed in some optimal fashion in the integration domain of interest, $[a, b]$. Having chosen the number of base points at will, $N_Q$, we compute their precise location and associated weights guided by the theory of orthogonal polynomials. Tables of base points and corresponding weights and associated codes are broadly available.

Different quadratures have been developed for cases where the lower and upper integration limits, $a$ and $b$, define a finite, semi-infinite, or infinite integration domain. Quadratures for functions that exhibit integrable singularities are also available. The accuracy of the computation improves dramatically as a higher number of base points are employed.

### 4.13.3 Nystrom's method

Approximating the integral in the denominator of Bayes' equation with a quadrature, and evaluating the equation at the integration base points, we obtain

$$\phi(\theta_j \,|\, \mathbf{x}) = \frac{\mathcal{L}_{\mathbf{x}}(\theta_j)}{c_Q \sum_{i=1}^{N_Q} \mathcal{L}_{\mathbf{x}}(\theta_i)\, \phi(\theta_i)\, w_i} \times \phi(\theta_j) \tag{4.13.8}$$

for $j = 1, \ldots, N_Q$. The posterior expected value and variance may also be computed in terms of the integration quadrature as

$$\overline{\theta} \simeq c_Q \sum_{i=1}^{N_Q} \theta_i\, \phi(\theta_i \,|\, \mathbf{x})\, w_i, \tag{4.13.9}$$

and

$$\sigma^2 \simeq c_Q \sum_{i=1}^{N_Q} (\theta_i - \overline{\theta})^2\, \phi(\theta_i \,|\, \mathbf{x})\, w_i. \tag{4.13.10}$$

Only a complete set of base points and the associated weights appear in the last three equations.

### 4.13.4 Gauss–Legendre quadrature

When both integration limits, $a$ and $b$, are finite, that is, they are not infinite, we apply the Gauss–Legendre quadrature. The base points are given by

$$\theta_i = \tfrac{1}{2}\,(a+b) + \tfrac{1}{2}\,(b-a)\,z_i, \tag{4.13.11}$$

where $z_i$ are the known zeros of the $N_Q$-degree Legendre polynomial. The coefficient $c_Q$ in front of the quadrature sum is equal to half the length of the integration domain,

$$c_Q = \tfrac{1}{2}\,(b-a). \tag{4.13.12}$$

The weights $w_i$ are available from tables. The Gauss–Legendre quadrature is used routinely in scientific computing.

### *4.13.5 Monte Carlo integration*

Consider the definite integral of the product of a sufficiently regular function, $\psi(\theta)$, and a probability density function (pdf), $\phi(\theta)$, defined in an interval of interest, $[a, b]$,

$$\mathcal{J} \equiv \int_a^b \psi(\theta)\,\phi(\theta)\,\mathrm{d}\theta. \tag{4.13.13}$$

Using the *law of large numbers*, we find that the integral can be approximated with the mean value of the function $\psi(\theta)$ at $N$ random points, $\theta_i$, sampled according to the pdf $\phi(\theta)$,

$$\mathcal{J} \simeq \frac{1}{N}\sum_{i=1}^{N} \psi(\theta_i). \tag{4.13.14}$$

The accuracy improves as a higher number of points are employed.

In Monte Carlo integration,

$$\varphi_{\mathbf{x}} \simeq \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}_{\mathbf{x}}(\theta_i), \qquad \overline{\theta} \simeq \frac{1}{N}\sum_{i=1}^{N} \theta_i,$$

$$\sigma^2 \simeq \frac{1}{N}\sum_{i=1}^{N} (\theta_i - \overline{\theta})^2. \tag{4.13.15}$$

Internal software functions that generate random numbers with a specified pdf are available.

To compute the integral of an arbitrary suitable function, $Q(\theta)$,

$$\mathcal{J} \equiv \int_a^b Q(\theta)\,\mathrm{d}\theta, \tag{4.13.16}$$

we apply Monte Carlo integration with the function $\psi(\theta) = Q(\theta)/\phi(\theta)$.

### 4.13.6 Monte Carlo implementation

The Monte Carlo algorithm involves the following steps:

1. Generate a sequence of random numbers with a flat pdf in the interval $[0, 1]$ (uniform random deviates), $\varrho_i$ for $i = 1, \ldots N$, using, for example, the Matlab function *rand*.

2. Find the corresponding sequence, $\theta_i$, by solving the equation

$$\int_a^{\theta_i} \phi(u)\,\mathrm{d}u = \varrho_i \tag{4.13.17}$$

   or the equation

$$\int_a^{\theta_i} \phi(u)\,du = 1 - \varrho_i, \tag{4.13.18}$$

   whichever is appropriate, where $u$ is a dummy integration variable, as shown in (4.1.15).

3. Approximate the integral using (4.13.14).

For example, Matlab encapsulates an internal function, *randn* that generates random data obeying the normal (Gaussian) distribution.

The procedure described in Step 2 is called inversion. An alternative is the acceptance–rejection method. In practice, both methods encounter significant limitations.

### 4.13.7 Numerical example

To illustrate the Monte Carlo method, we consider the integral

$$\mathcal{J} \equiv \int_0^\infty \theta^{2m} \mathrm{e}^{-\alpha\theta^2}\,\mathrm{d}\theta, \tag{4.13.19}$$

where $m$ is a specified integer and $\alpha$ is a specified positive parameter. The exact value of the integral is known to be

$$\mathcal{J} = \frac{1 \cdot 3 \cdots (2m-1)}{2\,(2\alpha)^m} \sqrt{\frac{\pi}{\alpha}} \tag{4.13.20}$$

(e.g., Gradshteyn, I. S. & Ryzhik, I. M. (1980) *Table of Integrals, Series, and Products.* Academic Press, p. 337, § 3.461.2.) Introducing the transformation $\theta = x/\sqrt{2\alpha}$, we obtain

$$\mathcal{J} = \frac{\sqrt{2\pi}}{(2\alpha)^{m+1/2}} \int_0^\infty x^{2m} \Big( \frac{1}{\sqrt{2\pi}}\, \mathrm{e}^{-\frac{1}{2}x^2} \Big)\, \mathrm{d}x. \tag{4.13.21}$$

The function enclosed by the tall parentheses inside the integral is the normal (Gaussian) distribution with zero mean and unit standard deviation. Consequently, we may approximate

$$\mathcal{J} \simeq \tfrac{1}{2} \frac{\sqrt{2\pi}}{(2\alpha)^{m+1/2}} \frac{1}{N} \sum_{i=1}^{N} x_i^{2m}, \tag{4.13.22}$$

where the coefficient $\frac{1}{2}$ accounts for the infinite domain of definition of the Gaussian distribution and the points $x_i$ are sampled from the Gaussian distribution.

The method is implemented in the following Matlab code named *monte_carlo* in terms of the internal Matlab function *randn*:

```
m = 4;
alpha = 0.3;
N = 2^20;
monte_carlo ...
= sqrt(pi/2)/ (2*alpha)^(m+0.5) * mean(randn(1,N).^(2*m))
exact = 1*3*5*7/2/(2*alpha)^m * sqrt(pi/alpha)
```

A Monte Carlo integration session with $N = 2^{20} = 1,048,576$ points yields the value $1,287.0$; another session with the same number of points yields the value $1,341.5$. Both values are reasonably close to the exact value, $1,310.9$. However, we note that a high number of sample points, $N$, are employed.

***Exercise***

**4.13.1** Compute by Monte Carlo integration the integral (4.13.19) for $\alpha = 0.2$ and $m = 3$.

## 4.14 Revision by projection

Consider the Bayes equation (4.10.4) for a probability density function (pdf), repeated below for convenience,

$$\phi(\theta \,|\, \mathbf{x}) = \frac{\mathcal{L}_{\mathbf{x}}(\theta)}{\varphi_{\mathbf{x}}} \times \phi(\theta), \tag{4.14.1}$$

where $\mathbf{x}$ is the data,

$$\varphi_{\mathbf{x}} \equiv \int_a^b \mathcal{L}_{\mathbf{x}}(\theta')\, \phi_{\mathbf{x}}(\theta')\, \mathrm{d}\theta' \tag{4.14.2}$$

is a marginal likelihood, and $\theta'$ is an integration variable.

Now apply this equation at an array of points, $\theta_i$ for $i = 1, \ldots, N$, that lie inside the domain of definition of $\theta$, to obtain $N$ equations,

$$\phi(\theta_i \,|\, \mathbf{x}) = \frac{\mathcal{L}_{\mathbf{x}}(\theta_i)}{\varphi_{\mathbf{x}}} \times \phi(\theta_i), \tag{4.14.3}$$

for $i = 1, \ldots, N$.

It is convenient to introduce two $N$-dimensional vectors containing the prior and the posterior,

$$\boldsymbol{\phi} \equiv \big(\, \phi(\theta_1), \ldots, \phi(\theta_N)\, \big) \tag{4.14.4}$$

and

$$\boldsymbol{\phi}_{\mathbf{x}} \equiv \big(\, \phi(\theta_1 \,|\, \mathbf{x}), \ldots, \phi(\theta_N \,|\, \mathbf{x})\, \big). \tag{4.14.5}$$

The $N$ equations in (4.14.3) can be compiled into the compact form

$$\boldsymbol{\phi}_{\mathbf{x}} = \mathbf{P}_{\mathbf{x}} \cdot \boldsymbol{\phi}, \tag{4.14.6}$$

where $\mathbf{P_x}$ is a diagonal matrix with components

$$(P_\mathbf{x})_{ij} = \begin{cases} \mathcal{L}_\mathbf{x}(\theta_i)/\varphi_\mathbf{x} & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases} \tag{4.14.7}$$

In the standard language of applied mathematics, $\mathbf{P_x}$ is a projection matrix; conceptually, the prior is projected onto the posterior.

The updating difference under data, $\mathbf{x}$, defined as $\Delta\boldsymbol{\phi} \equiv \boldsymbol{\phi}_\mathbf{x} - \boldsymbol{\phi}$, is given by

$$\Delta\boldsymbol{\phi} = \mathbf{B_x} \cdot \boldsymbol{\phi}, \tag{4.14.8}$$

where

$$\mathbf{B_x} \equiv \mathbf{P_x} - \mathbf{I} \tag{4.14.9}$$

is a diagonal matrix and $\mathbf{I}$ is the identity matrix.

Further updating under new data provide us with the formula

$$\boldsymbol{\phi}_{\mathbf{x}_1,\mathbf{x}_2,\ldots} = \mathbf{P}_{\mathbf{x}_1} \cdot \mathbf{P}_{\mathbf{x}_2} \cdots \boldsymbol{\phi}, \tag{4.14.10}$$

where $\boldsymbol{\phi}$ is the initial prior and a dot denotes matrix multiplication. If all data are the same, the projection matrices are raised to increasingly high powers, and the largest diagonal element dominates.

### *4.14.1 Evolution equation*

The updating difference equation underlies a nonlinear differential equation in virtual time, $t$,

$$\frac{\partial\,\phi(\theta)}{\partial t} = \Big( \frac{1}{\varphi(\mathbf{x}(t))}\,\mathcal{L}_{\mathbf{x}(t)}(\theta) - 1 \Big) \times \phi(\theta), \tag{4.14.11}$$

where the data $\mathbf{x}(t)$ are implicit functions of time, and

$$\varphi(\mathbf{x}(t)) \equiv \int_a^b \mathcal{L}_{\mathbf{x}(t)}(\theta')\,\phi(\theta')\,\mathrm{d}\theta'. \tag{4.14.12}$$

The explicit forward-time discretization of this differential equation with a unit time interval $\Delta t$ yields (4.14.10). The evolution equation for the pdf is accompanied by an evolution equation for the data,

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = \mathbf{F}(t), \tag{4.14.13}$$

where $\mathbf{F}(t)$ is a given or measured function of time.

### *4.14.2 Update by diffusion*

To screen out unwanted statistical or numerical noise, we may replace the diagonal matrix $\mathbf{B}_\mathbf{x}$ with a tridiagonal matrix $\boldsymbol{\mathcal{B}}_\mathbf{x}$ with elements

$$(\mathcal{B}_\mathbf{x})_{ij} = \begin{cases} (1-2\epsilon)\,(c_\mathbf{x}\,\mathcal{L}_\mathbf{x}(\theta_i) - 1) & \text{if } i = j, \\ \epsilon\,(c_\mathbf{x}\,\mathcal{L}_\mathbf{x}(\theta_i) - 1) & \text{if } i = j \pm 1, \\ 0 & \text{otherwise}, \end{cases} \tag{4.14.14}$$

where $c_\mathbf{x} = 1/\varphi_\mathbf{x}$ and $\epsilon$ is an arbitrary positive parameter. When $\epsilon = 0$, we recover $\boldsymbol{\mathcal{B}}_\mathbf{x} = \mathbf{B}_\mathbf{x}$.

Correspondingly, we replace the diagonal projection matrix $\mathbf{P}_\mathbf{x}$ with a tridiagonal matrix $\boldsymbol{\mathcal{P}}_\mathbf{x}$ with elements

$$(\mathcal{P}_\mathbf{x})_{ij} = \begin{cases} c_\mathbf{x}\,\mathcal{L}_\mathbf{x}(\theta_i) + 2\epsilon\,(1 - c_\mathbf{x}\,\mathcal{L}_\mathbf{x}(\theta_i)) & \text{if } i = j, \\ -\epsilon\,(1 - c_\mathbf{x}\,\mathcal{L}_\mathbf{x}(\theta_i)) & \text{if } i = j \pm 1, \\ 0 & \text{otherwise}. \end{cases} \tag{4.14.15}$$

These forms implement a diffusion process that smooths profiles and screens out artificial oscillations.

#### *Exercise*

**4.14.1** Discuss the behavior of $\phi_{\mathbf{x}_1,\mathbf{x}_2,\ldots}$ for repeated doublets of data.

# Chapter 5

# Analysis with pdfs

In Chapter 4, we introduced Bayes' rule for the probability density function (pdf) of a parameter, $\theta$, that determines a continuous range of random events in terms of data encapsulated in an array, $\mathbf{x}$. In this chapter, we proceed to discuss illustrative applications from a variety of settings. Different likelihood functions and prior pdfs will be used with strong or weak justification according to context.

In Chapter 6, the formulation will be extended to frameworks involving random occurrences and events determined by two or a higher number of random parameters described by joint probability distributions.

## 5.1 Binomial likelihood

Assume that the number of acquittals ($m$) in a number of trials ($n$) of a defense attorney are encapsulated in her performance record expressed by the doublet $\mathbf{x} \equiv (m, n)$. The associated likelihood function is given by the binomial distribution

$$\mathcal{L}_{\mathbf{x}}(\theta) = \mathcal{B}_m^n(\theta) \equiv \mathcal{C}_m^n\, \theta^m\, (1-\theta)^{n-m}, \tag{5.1.1}$$

where $\mathcal{C}_m^n$ is the binomial coefficient defined and discussed in Appendix A, and the parameter $\theta$ value is the probability that a single trial will result in acquittal (Bernoulli probability). A person who is interested in hiring the attorney will want to know the attorney's $\theta$ curve in terms of a probability density function (pdf).

Other interpretations of $n$, $m$, and $\theta$ were discussed in Chapter 4 with regard to photocopying sessions, and further interpretations will be discussed in this chapter.

### 5.1.1 Bayes formula

In the Bayesian framework, $\theta$ is an unknown parameter described by a probability density function, $\phi(\theta)$. Bayes formula with the binomial distribution as likelihood reads

$$\phi(\theta|\mathbf{x}) = \frac{\mathcal{B}_m^n(\theta)}{\varphi_m^n} \times \phi(\theta), \tag{5.1.2}$$

where

$$\varphi_m^n \equiv \int_0^1 \mathcal{B}_m^n(\theta')\,\phi(\theta')\,\mathrm{d}\theta' \tag{5.1.3}$$

and $\theta'$ is an integration variable. The integral appears formidable, and it usually is.

### 5.1.2 Sequential updating

Updating the pdf after a pair of $\mathbf{x}_1 \equiv (m_1, n_1)$ data has been received, and then again after another pair of $\mathbf{x}_2 \equiv (m_2, n_2)$ data has been received, we obtain

$$\phi(\theta|\mathbf{x}_1, \mathbf{x}_2) = \frac{\mathcal{B}_{m_1}^{n_1}(\theta)\,\mathcal{B}_{m_2}^{n_2}(\theta)}{\varphi_{m_1,m_2}^{n_1,n_2}} \times \phi(\theta), \tag{5.1.4}$$

where

$$\varphi_{m_1,m_2}^{n_1,n_2} \equiv \int_0^1 \mathcal{B}_{m_1}^{n_1}(\theta')\,\mathcal{B}_{m_2}^{n_2}(\theta')\,\phi(\theta')\,\mathrm{d}\theta'. \tag{5.1.5}$$

Substituting the expression for the binomial distribution and simplifying, we obtain

$$\phi(\theta|\mathbf{x}_1, \mathbf{x}_2) = \frac{\mathcal{B}_{m_1+m_2}^{n_1+n_2}(\theta)}{\varphi_{m_1,m_2}^{n_1,n_2}} \times \phi(\theta), \tag{5.1.6}$$

where

$$\varphi^{n_1,n_2}_{m_1,m_2} \equiv \int_0^1 \mathcal{B}^{n_1+n_2}_{m_1+m_2}(\theta')\,\phi(\theta')\,\mathrm{d}\theta'. \tag{5.1.7}$$

This formula admits a direct generalization to an arbitrary number of data sets.

### 5.1.3 Uniform prior pdf

In the case of a uniform prior distribution, $\phi(\theta) = 1$, we obtain a binomial posterior distribution

$$\phi(\theta|\mathbf{x})_{\text{from uniform}} = \frac{\mathcal{B}^n_m(\theta)}{\displaystyle\int_0^1 \mathcal{B}^n_m(\theta')\,\mathrm{d}\theta'}. \tag{5.1.8}$$

Substituting the explicit form of the binomial distribution and simplifying by eliminating the combinatorial coefficient, we obtain

$$\phi(\theta|\mathbf{x})_{\text{from uniform}} = \frac{\theta^m(1-\theta)^{n-m}}{\displaystyle\int_0^1 \theta'^{\,m}(1-\theta')^{n-m}\,\mathrm{d}\theta'}. \tag{5.1.9}$$

We recall that the integral in the denominator is merely a scalar normalization factor.

### 5.1.4 The Beta and Gamma functions

To compute the normalization factor in the denominator on the right-hand side of (5.1.9), we consider the Beta function

$$\mathrm{B}(\alpha,\beta) \equiv \int_0^1 \varrho^{\alpha-1}(1-\varrho)^{\beta-1}\,\mathrm{d}\varrho, \tag{5.1.10}$$

which is related to the Gamma function by $\Gamma$, by

$$\mathrm{B}(\alpha,\beta) = \frac{\Gamma(\alpha)\,\Gamma(\beta)}{\Gamma(\alpha+\beta)}, \tag{5.1.11}$$

(e.g., Gradshteyn, I. S. & Ryzhik, I. M. (1980) *Table of Integrals, Series, and Products.* Academic Press, p. 285, § 3.196.2).

Setting in (5.1.10)

$$\varrho = \theta', \qquad \alpha = m + 1, \qquad \beta = n - m + 1, \tag{5.1.12}$$

and noting that

$$\Gamma(r+1) = r\,! \equiv 1 \cdot 2 \cdots r \tag{5.1.13}$$

for any integer, $r$, we obtain the integration formula

$$\int_0^1 \theta'^m (1-\theta')^{n-m} \, \mathrm{d}\theta' = \frac{m\,!\,(n-m)\,!}{(n+1)\,!}, \tag{5.1.14}$$

where the exclamation mark denotes the factorial. Substituting this expression into (5.1.9), we obtain

$$\phi(\theta|\mathbf{x})_{\text{from uniform}} = \frac{(n+1)\,!}{m\,!\,(n-m)\,!}\,\theta^m (1-\theta)^{n-m}, \tag{5.1.15}$$

which can be restated as

$$\phi(\theta\,|\,\mathbf{x})_{\text{from uniform}} = (n+1)\,\mathcal{B}_m^n(\theta). \tag{5.1.16}$$

Since the definite integral of the posterior pdf with respect to $\theta$ from $0$ to $1$ must be unity, it must be that

$$\int_0^1 \mathcal{B}_m^n(\theta)\,\mathrm{d}\theta = \frac{1}{n+1}, \tag{5.1.17}$$

independent of $m$ and $n$.

### 5.1.5 Uniform and beta distribution priors

When the prior, $\phi(\theta)$, is the uniform distribution, $\phi(\theta) = 1$, or beta distribution $\phi(\theta) = \mathfrak{Be}_{\alpha,\beta}(\theta)$ discussed in Section 5.3, the posterior remains a binomial or beta distribution. Under these circumstances, the maximum of the posterior pdf occurs at the maximum of the likelihood function, which is the binomial distribution, for aggregated sets of data, $\mathbf{x}_1 \equiv (m_1, n_1)$, $\mathbf{x}_2 \equiv (m_2, n_2)$, ....

***Exercise***

**5.1.1** Confirm the integral in (5.1.17) for $\mathcal{B}_n^n(\theta)$ and $\mathcal{B}_0^n(\theta)$.

## *5.2 Pizza, medical records, and strawberries*

A person flips a coin whose sides are labeled $\mathrm{A}$ or $\mathrm{B}$. Each coin flip is a *Bernoulli trial.* The coin could be rigged so that the probability that it lands with face $\mathrm{A}$ up is $\theta$ and the probability that it lands with face $\mathrm{B}$ up is $1-\theta$, where $\theta$ is a *Bernoulli probability* ranging in the interval $[0,1]$. In the case of a fair coin, $\theta=\frac{1}{2}$.

The Bernoulli probability, $\theta$, can be regarded as an unknown parameter described by a certain probability density function (pdf), $\phi(\theta)$. In the absence of insights on the coin fairness, we may assume that the coin has been rigged with equal probability in every possible way, and adopt a uniform prior pdf, $\phi(\theta)=1$.

A uniform or another pdf can be updated after a session of $n$ coin flips using equation (5.1.2), involving the binomial distribution as the likelihood function, where $m$ is the number of observed faces $\mathrm{A}$ and $n-m$ is the number of observed faces $\mathrm{B}$. After a number of sessions, the pdf will give us a good idea about the coin fairness.

### *5.2.1 Pizza or hamburger*

Eighteen-wheeler driver Anna flips a coin as she approaches Highway Exit 15 to decide whether she will have a hamburger or a slice of pizza for lunch. She keeps in her pocket two coins: an unfair coin with $\theta=0.1$, and another unfair coin with $\theta=0.9$.

Anna fabricated these unfair coins by coating layers of epoxy on the first or second face with a blue or gray color. An alternative would have been to bent the coins with pliers toward the head or tail face. When Anna feels like pizza, she flips the first coin; when she feels like a hamburger, she flips the second coin.

### *5.2.2 Medical records*

Medical records were collected from health centers, hospitals, and individual practices across the nation and have been analyzed to describe the fraction of different populations that suffer from heart disease, $\theta$, in terms of a probability density function, $\phi(\theta)$.

Some populations have high $\theta$ due to healthy eating, vigorous exercise, clean air and water, and good genes Other populations have low $\theta$ due to compromised genes, heavy smoking, polluted air and dirty water. Conversely, $\theta$ can be regarded as the Bernoulli probability that a person will develop heart disease.

In one particular township with population $n = 945$, a registered nurse (RN) found that $m = 123$ residents developed heart disease. Assuming that these incidents of heart disease are unrelated (non-hereditary), she adopts the binomial likelihood function and applies formula (5.1.2) to obtain the posterior pdf specific to the township from the prior pdf corresponding to the entire nation.

### *5.2.3 Strawberries*

A lovely farmhouse is sitting pretty at the edge of a farm next to a country road in rural Ohio. A horizontal sliding window facing the porch has been opened partially to an unknown location, exposing an opening of length $\theta \times w$ and a transparent glass surface of length $(1 - \theta) \times w$, where $w$ is the width of the sliding sash and $\theta$ is an openness parameter, as shown in Figure 5.2.1.

### *5.2.4 Bilbo*

A few woodchucks are having fun in the spring by throwing $n$ strawberries at the window with their backs turned against the window. The session is orchestrated by the biggest woodchuck, named Bilbo, who borrowed the strawberries from his favorite field.

Between giggles, Bilbo counts that $m$ out of the $n$ strawberries have gone through the open area of the sliding window sash, and the remaining $n - m$ strawberries have landed and splashed on the sliding sash.

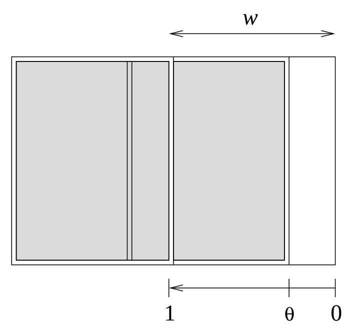



FIGURE 5.2.1 Illustration of a horizontal sliding window that has been opened partially to a location determined by a parameter, $\theta$. When $\theta = 0$ the window is shut.

The strawberries that stuck to the glass are eaten by birds. The strawberries that make it through the opening land on a strawberry pie that the farmer deliberately put on a table behind the window to entertain the woodchucks and get some benefit from their silly acts.

Based on this information, Bilbo wants to estimate the fraction of the open window, $\theta$, in terms of a probability density function (pdf.) If the probability of each strawberry landing on the pie is proportional to the open fraction of the window, $\theta$, and if the woodchucks do not adjust their throwing technique during a session based on dynamic feedback, then the corresponding likelihood function is given by the binomial distribution

$$\mathcal{L}_{\mathbf{x}}(\theta) = \mathcal{B}^n_m(\theta), \tag{5.2.1}$$

where $\mathbf{x} = (n, m)$ is the data. Bayes' theorem can be used to obtain a posterior pdf from an assumed prior pdf, $\phi(\theta)$. If the window is assumed to be open at any position, the prior pdf is uniform.

### *5.2.5 Smart farmer*

The farmer hosting Bilbo and his friends is watching with much amusement the woodchucks. The farmer observes that, for some unclear reason, the woodchucks are more successful in landing strawberries on the pie when the window is open in the mid-way position.

To account for this observation, the farmer is thinking of using an empirical likelihood function which is a modification of the binomial distribution,

$$\mathcal{L}_{\mathbf{x}}(\theta) = \theta\,(1-\theta)\,\mathcal{B}^n_m(\theta). \tag{5.2.2}$$

The farmer cannot help but think that the subjectiveness of this modification based on his own observations can be manipulated to reach any desired conclusions, true or perceived.

Fudging analysis, data, or both, by subjective modifications may appear like an act of deception or desperation. The farmer begins to understand why Bayesian analysis can be viewed as snake oil by critics.

***Exercises***

**5.2.1** Discuss a situation of your choice where the notion of a Bernoulli trial is appropriate.

**5.2.2** Suggest a modification of the binomial distribution in the event that the woodchucks are more successful in landing strawberries on the pie when the window is one-third open.

## 5.3 The beta distribution

The beta distribution is defined as

$$\mathfrak{Be}_{\alpha,\beta}(\theta) \equiv \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)} \times \theta^{\alpha-1}(1-\theta)^{\beta-1} \tag{5.3.1}$$

for any real and positive parameters $\alpha$ and $\beta$ and any value of $\theta$ in the interval $[0, 1]$, When $\alpha = 1$ and $\beta = 1$, the beta distribution reduces to the uniform distribution.

When $\alpha < 1$, $\beta < 1$, or both, a benign integrable singularity occurs at $\theta = 0$, $1$, or both. Such singularity is allowed in probability density functions.

The definite integral in (5.1.11) suggests that the probability density function (pdf)

$$\phi(\theta) = \mathfrak{Be}_{\alpha,\beta}(\theta), \tag{5.3.2}$$

defined for any value of $\theta$ in the interval $[0, 1]$, is acceptable in that it satisfies the obligatory normalization condition

$$\int_0^1 \mathfrak{Be}_{\alpha,\beta}(\theta)\, \mathrm{d}\theta = 1, \tag{5.3.3}$$

where $\theta$ is regarded as a Bernoulli probability.

The beta distribution, $\mathfrak{Be}_{\alpha,\beta}(\theta)$ should be confused neither with the Beta function, $\mathrm{B}(\theta)$, which is defined in terms of the Gamma function, nor with the binomial distribution,

$$\mathcal{B}^n_m(\theta) \equiv \mathcal{C}^n_m\, \theta^m\, (1-\theta)^{n-m}, \tag{5.3.4}$$

where $n$ is an integer, $m = 0, \ldots n$ is another integer, and $\mathcal{C}^n_m$ is the binomial coefficient. The beta distribution is a generalization of the binomial distribution arising for $\alpha = m + 1$ and $\beta = n - m + 1$.

### *5.3.1 Thumbs up and thumbs down*

In practice, the parameter $\theta$ is interpreted as a measure of quality or success, where $\theta = 1$ conveys supreme quality and unqualified success, and $\theta = 0$ conveys poor quality and total failure. $\mathfrak{Be}_{1,1}(\theta) = 1$.

More generally, the parameter $\alpha$ measures the number of thumbs up and the parameter $\beta$ measures the number of thumbs down obtained by literal, figurative, or metaphorical polling. Thumbs can be extended at any position between up and down to indicate ambivalence. The total number of whole or partial thumbs is $\alpha + \beta$.

### *5.3.2 Expected value and variance*

The expected value of a random parameter $\theta$ that obeys the beta distribution is found to be

$$\overline{\theta} \equiv \int_0^1 \theta\, \mathfrak{Be}_{\alpha,\beta}(\theta)\, \mathrm{d}\theta = \frac{\alpha}{\alpha+\beta}. \tag{5.3.5}$$

When $\alpha = \beta$, the expected value is $1/2$, independent of $\alpha$. We observe that the expected value is simply the number of thumbs up divided by the sum of thumbs up and thumbs down. The associated variance is

$$\sigma^2 \equiv \int_0^1 (\theta - \overline{\theta})^2 \, \mathfrak{Be}_{\alpha,\beta}(\theta) \, \mathrm{d}\theta = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}. \tag{5.3.6}$$

When $\alpha = 1$ and $\beta = 1$, we find that $\overline{\theta} = \frac{1}{12}$ and $\sigma^2 = \frac{1}{12}$.

Conversely, the last two equations can be used to recover the parameters $\alpha$ and $\beta$ in terms of $\overline{\theta}$ and $\sigma^2$, in a process known as data fitting.

### 5.3.3 Cumulative pdf

The cumulative pfd associated with the beta distribution is defined as

$$\Phi_{\alpha,\beta}(\theta) \equiv \int_0^\theta \mathfrak{Be}_{\alpha,\beta}(\theta') \, \mathrm{d}\theta', \tag{5.3.7}$$

where $\theta'$ is an integration variable. Explicitly,

$$\Phi_{\alpha,\beta}(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)} \int_0^\theta \theta'^{\alpha-1} (1-\theta')^{\beta-1} \, \mathrm{d}\theta'. \tag{5.3.8}$$

The last integral on the right-hand side is the incomplete Gamma function.

### 5.3.4 Bayesian updating with binomial likelihood

Assume that a prior pdf is represented by the beta distribution and the likelihood function is the binomial distribution

$$\mathcal{L}_{\mathbf{x}}(\theta) = \mathcal{B}_m^n(\theta). \tag{5.3.9}$$

The posterior pdf is given by

$$\phi(\theta \,|\, \mathbf{x})_{\text{from beta}} = \frac{\mathcal{B}_m^n(\theta)}{\displaystyle\int_0^1 \mathcal{B}_m^n(\theta') \, \mathfrak{Be}_{\alpha,\beta}(\theta') \, \mathrm{d}\theta'} \times \mathfrak{Be}_{\alpha,\beta}(\theta), \tag{5.3.10}$$

where the data $\mathbf{x}$ encapsulates the doublet $(n, m)$, and $\theta'$ is an integration variable. Cursory inspection reveals that the posterior pdf is another beta distribution, given by

$$\phi(\theta \,|\, \mathbf{x})_{\text{from beta}} = \mathfrak{Be}_{\alpha+m,\beta+n-m}(\theta). \tag{5.3.11}$$

where $m$ is typically interpreted as thumbs up and $n - m$ is interpreted as thumbs down. This feature underscores the convenience of likelihood function–prior conjugacy in Bayesian statistics. Other conjugate pairs are available.

### *5.3.5 Malpractice*

The convenience of using conjugate pairs can be so overpowering and enticing that such pairs are often used indiscriminantly with no regard to physical or conceptual relevance. The temptation to use conjugate pairs in the absence of sound reasoning is encouraged further by the availability of such pairs in pull-down menus of Bayesian statistics software. One must be careful not to formulate the opinion that this malpractice is a fundamental flaw of Bayesian analysis.

#### *Exercise*

**5.3.1** A professor senses that five students like her class, two students dislike her class, and two students sleep through the class. What is the pdf of the quality of the class perceived by the students quantified by a parameter $\theta$?

## *5.4 Sea-salted roasted almonds*

Author P. H. Aedrus woke up at 3:30 am every single morning for five straight years to work for a few hours on a book entitled "*What Happened to Klappa Clue? Adventures of an Inquisitive Tree Frog.*" The narrative is based on a brilliant framework that combines fiction with sociological, theological, philosophical, and epistemological commentary.

### 5.4.1 No holidays for Aedrus

Author Aedrus spent entire weekends working on the manuscript and was oblivious to all holidays. When he thought the manuscript was good enough to be read by others, he submitted his work to twenty publishers.

### 5.4.2 Submission and rejection

Five of the twenty publishers responded with ten reviews altogether. Three reviews were positive, three reviews were sitting on the fence, and four reviews were unfavorable. All publishers praised the manuscript but stated that extensive revisions are needed to address the issues raised by the reviewers. After the initial shock of rejection verging on depression that lasted for several months, Author Aedrus got back on his feet and was able to revisit his work.

### 5.4.3 At the Post Office

A mathematician colleague at the Post Office who was also doubling as a grossly underpaid adjunct professor at the local community college mentioned that the probability density function (pdf) of a variable, $\theta$, that indicates whether the manuscript will be accepted ($\theta = 1$) or declined ($\theta = 0$) is described by the beta distribution with parameters

$$\begin{aligned} \alpha &= 3 + 0.5 + 0.5 + 0.5 = 4.5, \\ \beta &= 4 + 0.5 + 0.5 + 0.5 = 5.5. \end{aligned} \tag{5.4.1}$$

A graph of this prior distribution is represented by the solid line in Figure 5.4.1. The maximum is achieved at the expected value

$$\overline{\theta} = \frac{4.5}{4.5 + 5.5} = 0.45, \tag{5.4.2}$$

which is somewhat less than $0.5$, consistent with the publishers' response.

### 5.4.4 Revision

On the advice and encouragement of his colleague, author Aedrus spent two more years essentially rewriting the book on the same schedule.

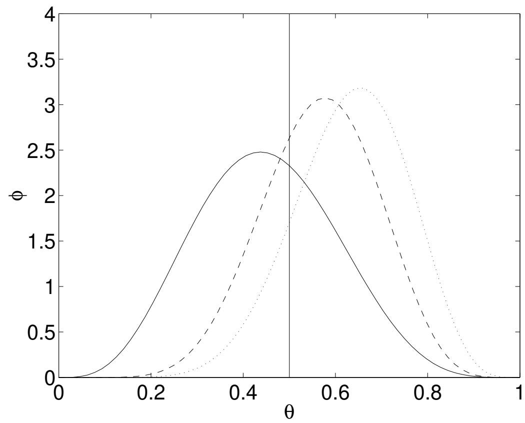



FIGURE 5.4.1 Graph of the prior beta distribution, $\mathfrak{Be}_{4.5,5.5}(\theta)$, drawn with the solid line, and the posterior beta distribution, $\mathfrak{Be}_{8.5,6.5}(\theta)$ drawn with the red broken dashed line. The posterior beta distribution, $\mathfrak{Be}_{9.5,5.5}(\theta)$ is drawn with the dotted line.

This time he took time off from work at the Post Office and gave up his precious sea-salted roasted almonds to compensate for the lost income. However, he could not bring himself to compromise on the quality of coffee he prepared each morning and sharply at noon.

When author Aedrus thought he was ready, he asked five people whose opinion he trusted to read the revision. Four of the readers thought the revised work was brilliant. The fifth reader suggested that a large portion of the commentary needs to be deleted or grounded on historical or philosophical perspectives.

Author Aedrus wants to get an idea of the publishers' response before resubmitting the manuscript. Using Bayes' equation (5.3.11), author Aedrus finds that the posterior pdf is given by the new beta distribution

$$\phi\big(\theta \,|\, (4,1),(4.5,5.5)\big) = \mathfrak{Be}_{4+4.5,1+5.5}(\theta) = \mathfrak{Be}_{8.5,6.5}(\theta). \quad (5.4.3)$$

A graph of this posterior distribution is shown with the dashed line in

Figure 5.4.1. The maximum is achieved roughly at $\theta = 0.6$,

$$\overline{\theta} = \frac{8.5}{8.5 + 6.5} = 0.566 \cdots , \tag{5.4.4}$$

which shifts the probability toward acceptance. If all five readers of the revised work were positive, the posterior pdf would be given by $\mathfrak{Be}_{9.5,5.5}(\theta)$, which is represented by the dotted line in Figure 5.4.1.

Based on these estimates, author Aedrus decided to submit the book with significant trepidation. The daily trip to his mail box was a fearful act followed by relief when a publisher response did not arrive.

### *5.4.5 Final act*

A publisher accepted and published the revised manuscript and the book turned out to be a huge success. However, readers' opinions were bimodal.

Most readers found the writing inspiring and the ideas interesting and compelling. Some readers even wrote in Internet reviews that the insights presented in the book had a profound effect on their perspectives in life. Other readers found that the book was utterly uninteresting and placed it at the level of pulp fiction.

Notwistanding the critical reviews, author Aedrus earned sufficient royalties to guarantee a life-time supply of his favorite sea-salted roasted almonds that had to give up for two years.

### *5.4.6 Sensitive*

Author Aedrus mulled in his mind the question as to why anyone would feel compelled to publicly berate an author, especially since buying or reading a book is not mandatory. His neighbor at the upstairs apartment, professor of religious studies S. E. N. Sitive, suggested that the reviewing and blogging business has become a profitable industry capitalizing on the primitive instinct to criticize what we do not understand and diminish what lies above us.

### *5.4.7 The last temptation of the professor of Religious Studies*

Professor S. E. N. Sitive stopped believing in God when she learned that her mother passed away suddenly at her homeland, having spent her entire life in poverty raising four children as a single mom. The professor slowly recovered her faith, but only partially, in later years.

The professor kept her personal struggle private and transparent to students, administrators, and colleagues. She only confided to the custodian, a war veteran who helped her go through the crisis. The veteran himself suffered from excruciating sciatica and strong recurrent bouts of PTSD and survivor's guilt.

### 5.4.8 Retraining camp

While she was going through this personal struggle, a student emailed Professor S. E. N. Sitive with carbon copy (cc) to the Dean and the Provost, stating with intense emotion that the professor should be ashamed of herself for not making the class more engaging and for not interjecting in her lectures experiences from her homeland, personal stories, and entertaining anecdotes.

Having adopted a business model for the university they were hired to manage at exorbitant salaries, and being eager to satisfy their customers, the Dean and the Provost mandated that the Professor attend a series teaching effectiveness seminars in a retraining camp of their choice as a condition for her contract's renewal. The demand was dropped after the Professors' Union intervened.

#### Exercises

**5.4.1** Compute the variance of the prior and posterior distributions shown in Figure 5.4.1.

**5.4.2** Should Professor S. E. N. Sitive have submitted to the humiliating actions of the administration?

## 5.5 Internet book sales

Internet book seller *Clwyd* carries thousands of titles and allows verified book buyers to post reviews and rate books on a scale from one to five stars. A review is considered favorable if it is accompanied by four or five stars, and unfavorable otherwise.

### 5.5.1 Number of reviews

Some books receive no reviews, other books receive many reviews. The department of analytics has described the probability that a book receives $r$ reviews by a probability power-law distribution,

$$\pi_0 = \frac{\gamma}{c} \qquad \pi_r = \frac{1}{c}\frac{1}{r^s} \tag{5.5.1}$$

for $r \geq 1$, where $\gamma$ is a adjustable coefficient expressing the percentage of unreviewed books, $s$ is an adjustable exponent, and $c$ is a normalization factor given by

$$c = \gamma + \sum_{r=1}^{\infty} \frac{1}{r^s}, \tag{5.5.2}$$

so that

$$\sum_{r=0}^{\infty} \pi_r = 1, \tag{5.5.3}$$

as required for any discrete probability distribution.

### 5.5.2 Riemann's zeta function

In fact, the sum on the right-hand side of (5.5.2) is Riemann's zeta function defined as

$$\zeta_s \equiv \sum_{n=1}^{\infty} \frac{1}{r^s} \tag{5.5.4}$$

for any exponent $s > 1$. The sum diverges when $s \leq 1$ and the zeta function is not defined.

A graph of the zeta function confirming that the sum diverges as the exponent $s$ tends to $1$ is shown in Figure 5.5.1. Special values of the zeta function are

$$\zeta_1 = \infty, \qquad \zeta_2 = \tfrac{1}{6}\,\pi^2, \qquad \zeta_4 = \tfrac{1}{90}\,\pi^4. \tag{5.5.5}$$

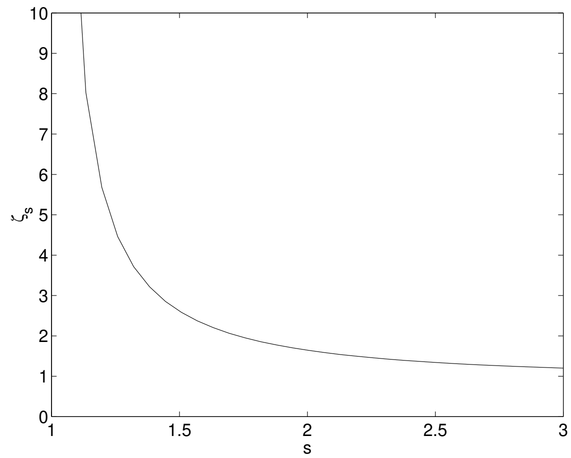



FIGURE 5.5.1 Graph of the Riemann zeta function, $\zeta_s$, defined as an infinite sum of inverse integer powers.

As $s$ tends to infinity, the first term in the sum dominates and the zeta function tends to unity.

### 5.5.3 Favorable and unfavorable reviews

The department of analytics found that the percentage of favorable reviews (four or five stars) for a certain book depends on the number of reviews written for the book. The higher the number of reviews, the higher the percentage of positive reviews. This interesting observation has been discussed in the context of group psychology and meta-critics.

The dependence of the percentage of favorable reviews on the number of reviews is mediated by a function, $f(r)$, which is defined such that the fraction of favorable reviews is

$$\alpha_r = f(r) \tag{5.5.6}$$

and the fraction of unfavorable reviews is

$$\beta_r = 1 - f(r) \tag{5.5.7}$$

for $r > 0$. An implicit assumption is that $0 \leq f(r) \leq 1$ for any number of reviews written, $r$. Since $\alpha_r + \beta_r = 1$, the number of positive and negative reviews is equal to the number of total reviews.

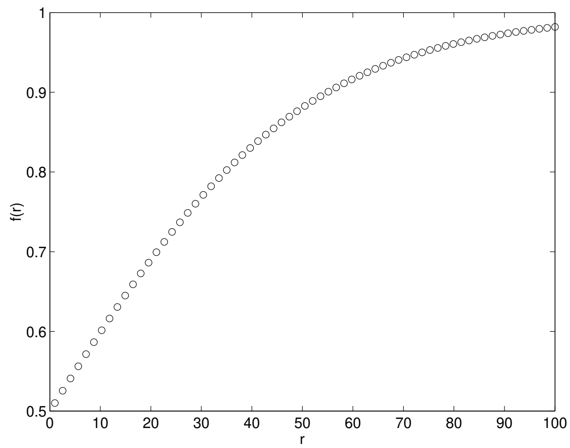



FIGURE 5.5.2 Graph of the fraction of good reviews, $f(r)$, describing the overall favorability of books with $r$ reviews for behavioral turning point $k = 50$.

The department of analytics fitted the function $f(r)$ with a formula involving the hyperbolic tangent function, $\tanh$,

$$f(r) = \xi + (1 - \xi) \tanh \frac{r}{k}, \tag{5.5.8}$$

where $\xi$ and $k$ are two adjustable parameters. A graph of this function for $\xi = \frac{1}{2}$ and $k = 50$ is shown in Figure 5.5.2. We confirm that the percentage of favorable reviews increases steadily with the number of reviews written, $r$. In this context, the parameter $k$ can be interpreted as a behavioral turning point.

### *5.5.4 Quality index*

The nominal quality of a randomly selected book can be quantified by an index, $\theta$, defined such that $\theta = 1$ describes high quality (Man Booker prize) and $\theta = 0$ describes low quality (pulp fiction.) The marketing department interprets $\theta$ as an index of commercial success, intellectual triumph, or both. The favorability index, $\theta$, obeys a probability density function given by

$$\phi(\theta) = \pi_0\,\phi_0(\theta) + \sum_{r=1}^{\infty} \pi_r \times \mathfrak{Be}_{\alpha(r),\beta(r)}(\theta), \tag{5.5.9}$$

where $\phi_0(\theta)$ is an assumed favorability distribution of books that received no reviews. In the absence of any insights, we set $\phi_0(\theta) = 1$. This choice hinges on the assumption that a book that has received no reviews could be either extremely interesting ($\theta = 1$) or extremely uninteresting ($\theta = 0$).

In a staff meeting, someone pointed out that a book that reveals the true meaning of life ($\theta = 1$) would receive no reviews, as readers would immediately abandon their computers upon reading the book and head for the mountains. A perfunctory book would also receive no reviews.

It is straightforward to verify that the pdf given in (5.5.9) satisfies the mandatory normalization condition

$$\int_0^1 \phi(\theta)\,\mathrm{d}\theta = 1. \tag{5.5.10}$$

Making substitutions into (5.5.9), we obtain

$$\phi(\theta) = \frac{1}{\gamma + \zeta_s}\Big(\gamma\,\phi_0(\theta) + \sum_{r=1}^{\infty} \frac{1}{r^s}\,\frac{\Gamma(\alpha_r + \beta_r)}{\Gamma(\alpha_r)\,\Gamma(\beta_r)}\,\theta^{\alpha_r - 1}(1-\theta)^{\beta_r - 1}\Big). \tag{5.5.11}$$

The precise form of this pdf depends on four adjustable parameters: $s$, $\gamma$, $\xi$, and $k$, where the last two parameters are involved in $\alpha_r$ and $\beta_r$.

A graph of this pdf for $s = 2$, $\gamma = 0.40$, $\xi = \frac{1}{2}$, and $k = 50$ is shown with the solid line in Figure 5.5.3. Even though the pdf diverges to infinity at $\theta = 0$ and $1$, the singularity is integrable, yielding a finite cumulative distribution,

$$\Phi(\theta) \equiv \int_0^{\theta} \phi(\theta')\,\mathrm{d}\theta' \tag{5.5.12}$$

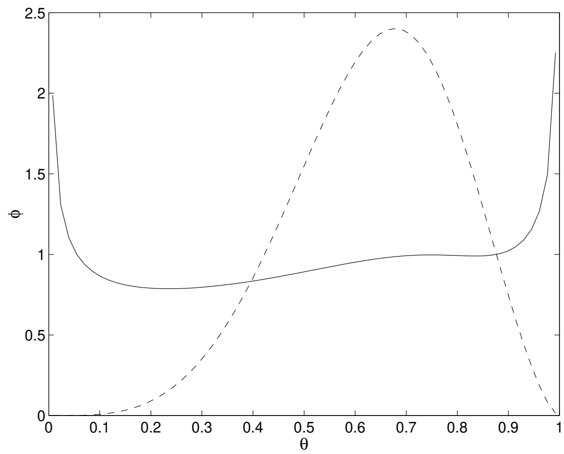



FIGURE 5.5.3 Graph of the pdf of a book impact index, $\theta$ for $s = 2$, $\gamma = 0.40$, and $k = 50$ (solid line). The posterior pdf is shown with the broken line.

and a finite complement, $1 - \Phi(\theta)$.

### 5.5.5 A new book is published

A new book received $6$ reviews, including $4$ favorable reviews and $2$ unfavorable reviews. Using Bayes's theorem with the binomial likelihood function $\mathcal{B}_4^6(\theta)$ and the pdf shown in (5.5.11) as a prior, we find that the posterior pdf of the quality index is

$$\phi(\theta|(4,5)) = \frac{1}{c}\,\theta^4 \times (1-\theta)^2 \times \phi(\theta), \tag{5.5.13}$$

where $c$ is a normalization factor. Note that this pdf is *not* just another beta distribution. A graph of this pdf is shown with the broken line in Figure 5.5.3. The results suggest that this is a good book.

Even if a has book received only positive reviews, there would still be a chance that the quality of the book is less than perfect. One reason is that all reviews may have been written by the author's cousins. Car dealers and others are known to post self-serving reviews.

***Exercises***

**3.5.1** Confirm that the normalization condition (5.5.10) is satisfied.

**3.5.2** Compute the posterior pdf subject to a non-informative (uniform) prior and compare it with that shown in Figure 5.5.3.

## *5.6 Cosmic particles through a window*

We revisit the partially open sliding window discussed in Section 5.3 involving woodchuck Bilbo, and assume that $n$ cosmic particles hit the window randomly over a certain period of time, $\tau$, according to the Poisson probability distribution,

$$\mathcal{P}_n(\lambda) = \frac{\lambda^n}{n\,!}\, e^{-\lambda} \tag{5.6.1}$$

for $n = 0, 1, \ldots,$ where $\lambda = \alpha\tau$ is a dimensionless parameter and $\alpha$ is a positive rate constant. A graph of the Poisson distribution for $\lambda = 6.5$ is shown in Figure 5.6.1.

To validate the Poisson distribution, we recall the Maclaurin series of the exponential function,

$$\mathrm{e}^{\lambda} \equiv \sum_{n=0}^{\infty} \frac{\lambda^n}{n\,!}, \tag{5.6.2}$$

which is, in fact, the definition of the exponential function. Summing all probabilities, we find that

$$\sum_{n=0}^{\infty} \mathcal{P}_n(\lambda) = e^{-\lambda} \sum_{n=0}^{\infty} \frac{\lambda^n}{n\,!} = e^{-\lambda} \times e^{\lambda}, \tag{5.6.3}$$

and thus

$$\sum_{n=0}^{\infty} \mathcal{P}_n(\lambda) = 1, \tag{5.6.4}$$

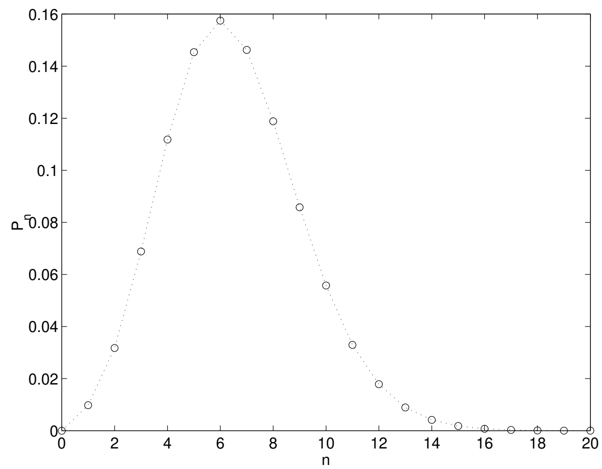



FIGURE 5.6.1 Graph of the Poisson distribution, $\mathcal{P}_n(\lambda)$, for $\lambda = 6.5$. Note that the maximum is achieved at the expected value $n = \lambda$.

as required for normalization.

The expected value of $n$ is equal to $\lambda$,

$$\overline{n} \equiv \sum_{0}^{\infty} n\,\mathcal{P}_n(\lambda) = \lambda, \tag{5.6.5}$$

and the standard deviation, $s$, is equal to $\lambda$,

$$s^2 \equiv \sum_{0}^{\infty} (n - \overline{n})^2\,\mathcal{P}_n(\lambda) = \lambda^2. \tag{5.6.6}$$

To confirm these expressions, we note that the distribution represented by the dotted lines in Figure 5.6.1 peaks at $n = 6.5$. The non-integer value may appear nonsensical in light of our assumption of an integer $n$. However, extending physical to abstract entities is the hallmark of mathematics.

It can be shown that the Poisson distribution arises from the binomial distribution in a certain limit.

### 5.6.1 Cosmic blockage

A fraction of the cosmic particles that hit the window land on the window glass and are reflected backward, while the remaining particles enter the room. The probability that $m$ out of $n$ particles are allowed through the open portion of the window during a period of time of interest is given by the likelihood function

$$\mathcal{L}_m(\theta, \lambda) = \sum_{n=m}^{\infty} \mathcal{B}_m^n(\theta)\, \mathcal{P}_n(\lambda). \tag{5.6.7}$$

We recall that $\theta$ is the fraction of the opened window sash, whereas $\lambda$ is a dimensionless parameter involving a rate constant.

We note that, in the mathematical expression for the likelihood function, the Poisson distribution plays the role of an unconditional probability, whereas the binomial distribution plays the role of a conditional probability.

### 5.6.2 Consolidation

Substituting the expression for the Poisson probability distribution into the expression for the likelihood function, we obtain

$$\mathcal{L}_m(\theta, \lambda) = e^{-\lambda} \sum_{n=m}^{\infty} \frac{\lambda^n}{n\,!}\, \mathcal{B}_m^n(\theta). \tag{5.6.8}$$

Substituting the definition of the binomial distribution, we obtain

$$\mathcal{L}_m(\theta, \lambda) = \frac{1}{m\,!}\, e^{-\lambda}\, \theta^m \sum_{n=m}^{\infty} \frac{\lambda^n}{(n-m)\,!}\, (1-\theta)^{n-m}, \tag{5.6.9}$$

which can be rearranged into

$$\mathcal{L}_m(\theta, \lambda) = \frac{1}{m\,!}\, e^{-\lambda}\, (\lambda\theta)^m \sum_{q=0}^{\infty} \frac{1}{q\,!} \big( \lambda\, (1-\theta)\, \big)^q, \tag{5.6.10}$$

where we have defined for convenience $q \equiv n - m$. The sum on the right-hand side is the exponential $\mathrm{e}^{\lambda(1-\theta)}$, yielding the likelihood function

$$\mathcal{L}_m(\theta, \lambda) = \frac{1}{m!}\, (\lambda\theta)^m\, e^{-\lambda\theta} \tag{5.6.11}$$

for $m = 0, 1, \ldots,$ which is precisely the Poisson distribution with expected value and standard deviation both equal to $\lambda\theta$,

$$\mathcal{L}_m(\theta, \lambda) = \mathcal{P}_m(\lambda\theta). \tag{5.6.12}$$

The mandatory normalization condition,

$$\sum_{m=0}^{\infty} \mathcal{L}_m(\theta, \lambda) = 1, \tag{5.6.13}$$

is satisfied.

### 5.6.3 Known $\lambda$

Assume that the rate constant, $\lambda$, is known. Bayes' equation takes the specific form

$$\phi(\theta|m) = \frac{\mathcal{L}_m(\theta, \lambda)}{\displaystyle\int_0^1 \mathcal{L}_m(\theta', \lambda)\, \phi(\theta')\, \mathrm{d}\theta'}\, \phi(\theta), \tag{5.6.14}$$

where the denominator of the fraction on the right-hand side is a marginal probability. Substituting the expression for the likelihood function from (5.6.11) and simplifying, we obtain

$$\phi(\theta|m) = \frac{\theta^m\, e^{-\lambda\theta}}{\displaystyle\int_0^1 \theta'^m\, e^{-\lambda\theta'}\, \phi(\theta')\, \mathrm{d}\theta'}\, \phi(\theta). \tag{5.6.15}$$

In the absence of any information, we may assume that the window has been opened with equal probability to any location, set $\phi(\theta) = 1$, and revise the probability based on a measured $m$ for given $\lambda$.

### 5.6.4 Mail delivery to a dentist

The postal delivery officer does what it takes to protect himself against angry dogs. When delivering mail to the home of retired dentist Gregory R. Umpy, owner of two protective pitbulls, he tosses the mail from a safe distance through a kitchen window to avoid the dogs. Some pieces of mail enter the kitchen, while the remaining pieces bounce back into a moat that the dentist built to prevent solicitors and past unhappy patients from knocking on his door.

The dentist is supposed to receive on average of $6.5$ pieces of mail each day, most of them coupons for dental cleaning supplies, gutter installations, and free second pizzas. The daily number of mail pieces that the dentist receives, $m$, is described by the Poisson distribution with $\lambda = 6.5$.

### *5.6.5 No mail on Tuesday*

Last Tuesday, the dentist received no mail, $m = 0$. Was this because there was no mail, or was it because the window was closed and all mail fell into the moat?

Applying (5.6.15) for $m = 0$ and a uniform prior pdf, $\phi(\theta) = 1$, meaning that the window could have been opened at any position. Dentist Umpy obtains the posterior pdf

$$\phi(\theta \,|\, m = 0) = \frac{e^{-\lambda\theta}}{\displaystyle\int_0^1 e^{-\lambda\theta'}\, \mathrm{d}\theta'} = \frac{\lambda}{\mathrm{e}^{\lambda} - 1}\, \mathrm{e}^{\lambda(1-\theta)}, \tag{5.6.16}$$

for $0 \le \theta \le 1$, which reveals an exponential decay of the probability that the window was partially open.

A graph of this function is shown in Figure 5.6.1. The probability that the window was shut, $\theta = 0$, and thus the mail ended up in the moat, is high. The probability that the window was entirely open, $\theta = 1$, and no mail was delivered is essentially null.

#### *Exercises*

**5.6.1** Explain how the Poisson distribution arises from the binomial distribution in a certain limit.

**5.6.2** Plot and discuss the counterpart of the graph shown in Figure 5.6.1 for $m = 2$.

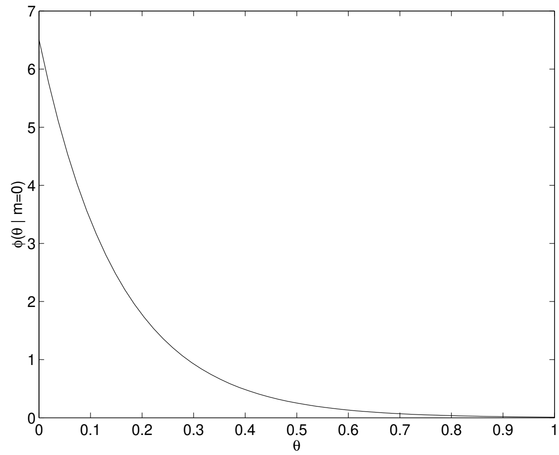



FIGURE 5.6.1 Graph of the probability density function of the dentist's open window fraction.

## 5.7 *Normal after normal*

A popular likelihood function is described by the normal distribution with variance $\sigma^2$, centered at an expected value $\theta = x$. We write

$$\mathcal{L}_{x,\sigma}(\theta) = \mathcal{N}_{x,\sigma}(\theta) \tag{5.7.1}$$

for $-\infty < \theta < \infty$, where $\theta$ is a parameter of interest and

$$\mathcal{N}_{x,\sigma}(\theta) \equiv \frac{1}{\sqrt{2\pi}\,\sigma}\,\exp(-\tfrac{1}{2}\,\frac{(\theta - x)^2}{\sigma^2}) \tag{5.7.2}$$

is the Gaussian distribution. A graph of the Gaussian distribution shifted with respect to the mean and scaled by the standard deviation, $\sigma$, is shown in Figure 5.7.1. The standard deviation is the square root of the variance.

We observe that $\mathcal{N}_{x,\sigma}(\theta)$ drops essentially to zero when $\theta$ is a only few distances $\sigma$ away from the expected value, $x$. This fast decay is due to the negative square of the power in the argument of the exponential. As the standard deviation $\sigma$ tends to zero, $\mathcal{N}_{x,\sigma}(\theta)$ spikes

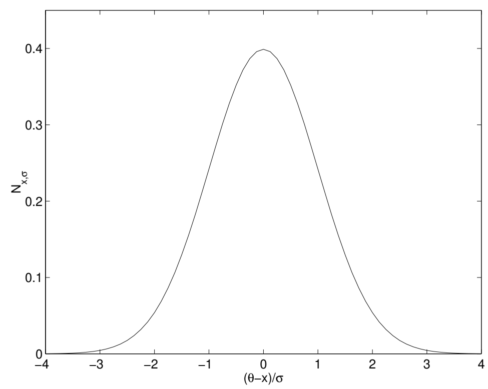



FIGURE 5.7.1 Graph of the Gaussian distribution with respect to the independent variable, $\theta$, shifted with respect to the expected value, $x$, and scaled with respect to the standard deviation, $\sigma$. The area under the curve is equal to unity.

at the expected values, $x$, and gradually reduces to the Dirac delta function representing an impulse.

The normal distribution satisfies the identities

$$\int_{-\infty}^{\infty} \mathcal{N}_{x,\sigma}(\theta)\,\mathrm{d}\theta = 1, \qquad \int_{-\infty}^{\infty} \mathcal{N}_{x,\sigma}(x)\,\mathrm{d}x = 1, \tag{5.7.3}$$

expressing mandatory normalization conditions.

### 5.7.1 Bayes' equation

Bayes' equation for a pdf, $\phi(\theta)$, with the normal distribution as likelihood function with a given $\sigma$ and measured $x$ takes the form

$$\phi(\theta)_{\text{posterior}} = \frac{\mathcal{N}_{x,\sigma}(\theta)}{\displaystyle\int_{-\infty}^{\infty} \mathcal{N}_{x,\sigma}(\theta')\,\phi(\theta')\,\mathrm{d}\theta'}\,\phi(\theta), \tag{5.7.4}$$

where $\theta'$ is an integration variable. A prior pdf can be assumed, $\phi(\theta)$, and then revised according to values of $x$ and $\sigma$ suggested by data.

### 5.7.2 Normal after normal

When the prior pdf is uniform, $\phi(\theta) = 1$, the posterior pdf is normal and remains normal as more doublets of data, $x$ and $\sigma$, become available. To show this, we assume that the prior distribution is a normal distribution,

$$\phi(\theta) = \mathcal{N}_{\overline{\theta},\tau}(\theta), \tag{5.7.5}$$

where $\overline{\theta}$ is the expected value of $\theta$ corresponding to the maximum of the Gaussian distribution, and $\tau$ is the associated standard deviation.

Applying the Bayes equation (5.7.4), we find that the posterior distribution is given by the scaled product of two normal distributions,

$$\phi(\theta)_{\text{posterior}} = \frac{\mathcal{N}_{x,\sigma}(\theta) \times \mathcal{N}_{\overline{\theta},\tau}(\theta)}{\displaystyle\int_{-\infty}^{\infty} \mathcal{N}_{x,\sigma}(\theta')\, \mathcal{N}_{\overline{\theta'},\tau}(\theta)\, \mathrm{d}\theta'}, \tag{5.7.6}$$

where $\theta'$ is an integration variable. In fact, the product of two normal distributions is another normal distribution. To show this, we substitute the expression for the normal distributions and obtain

$$\phi(\theta)_{\text{posterior}} = \frac{1}{\displaystyle\int_{-\infty}^{\infty} F(\theta')\, \mathrm{d}\theta'}\, F(\theta), \tag{5.7.7}$$

where, by definition,

$$F(\theta) \equiv \exp\big( -\tfrac{1}{2}\, \frac{(\theta - x)^2}{\sigma^2} - \tfrac{1}{2}\, \frac{(\theta - \overline{\theta})^2}{\tau^2} \big). \tag{5.7.8}$$

An important observation is that $F(\theta)$ can be expressed as

$$F(\theta) = c\, \exp\big( -\tfrac{1}{2}\, \frac{(\theta - \overline{\theta}_{\text{posterior}})^2}{\tau^2_{\text{posterior}}} \big), \tag{5.7.9}$$

where $c$ is some constant,

$$\frac{1}{\tau^2_{\text{post}}} = \frac{1}{\tau^2} + \frac{1}{\sigma^2}, \qquad \frac{\overline{\theta}_{\text{post}}}{\tau^2_{\text{post}}} = \frac{\overline{\theta}}{\tau^2} + \frac{x}{\sigma^2}, \tag{5.7.10}$$

and *post* stands for posterior. We conclude that

$$\phi(\theta)_{\text{posterior}} = \mathcal{N}_{\overline{\theta}_{\text{post}},\tau_{\text{post}}}(\theta), \tag{5.7.11}$$

which is a revised normal distribution.

The normal posterior can be revised further taking into consideration data fitted to another normal distribution.

### 5.7.3 Chainsaws

A marketing firm has discovered that the industry standard of the pdf of the net daily sales of chainsaws, denoted by $\theta$, is described by a normal distribution, $\mathcal{N}_{x,\sigma}(\theta)$. Negative values of $\theta$ or $x$ correspond to number of sales offset by the number of returns.

The daily number of units sold is affected by a large number of independent random factors, including the occurrence of storms with damaging winds and the general consumers' spending mood.

The daily number of units returned is also affected by a large number of independent random factors, including the manufacturing quality of plastic carburetor parts.

The normal distribution is justified as a consequence of the central limit theorem for the compounded effect of a large number of random parameters.

### 5.7.4 Sharp teeth

Last year's net daily sales of chainsaw manufacturer *Sharp Teeth, LLC* were described by a normal prior pdf, $\phi(\theta)$, with a measured expected value and variance. A prediction for this year's sales can be made using (5.7.11). A naive approach would have been to assume that this year's sales will be the same as last years' sales.

#### *Exercise*

**5.7.1** Discuss the similarities between the chainsaw problem and the medical records problem discussed in Section 37.

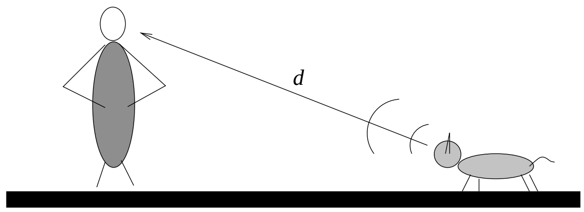



FIGURE 5.8.1 Professor's Ritewell cat Byte is meowing at a distance $d$ from her in the dark.

## 5.8 How far is the cat?

Professor of English Literature U. Ritewell stepped out to her backyard when she heard her cat Byte meow. She can't tell how far Byte is in the dark and wants to estimate the distance, $d$, as shown in Figure 5.8.1. The professor is afraid that she may accidentally step on Byte's bushy tail.

### 5.8.1 Deterministic meow

Under ideal conditions, the strength of the meow heard by the Professor has a unique and well-defined value, $s_{\text{ideal}}$, that depends on distance, $d$, alone. We can set

$$s_{\text{ideal}}(d) = s_0\, \psi(d), \tag{5.8.1}$$

where $s_0$ is the loudest possible meow heard by a nervous grasshopper sitting under the cat and pretending not to be there, and $s$ stands for sound. The function

$$\psi(d) = \frac{a^2}{a^2 + d^2} \tag{5.8.2}$$

expresses an acoustic attenuation law, where $a$ is the cat's height. Note that, if Byte's height is $a = 1$ ft, then the sound intensity falls off nearly by a factor or $100$ at a distance of $d = 10$ ft.

### *5.8.2 Ideal likelihood*

The ideal meow intensity can be described by a probability density function (pdf) for $s$, parametrized by $d$, involving the Dirac delta function, $\delta(w)$, representing an infinite spike. This pdf is expressed by the following likelihood function for $d$,

$$\mathcal{L}_s^{\text{ideal}}(d) = \delta(\, s - s_{\text{ideal}}(d)\,). \tag{5.8.3}$$

In the present ideal state, and in any other ideal state, the delta function expresses a deterministic response in the absence of uncertainty.

### *5.8.3 Mediating field*

Before we proceed, we make an important observation: *The likelihood function for the variable of interest, $d$, is determined by a field pertaining to $d$ according to an appropriate law mediated by the function $\psi(d)$ embedded in $s_{\text{ideal}}(d)$.*

This realization is a key to performing Bayesian analysis in a broad range of physical and engineering applications.

### *5.8.4 Random meow*

The actual meow intensity heard by the Professor is less than the ideal meow $s_{\text{ideal}}$ for several reasons: is the Professor wearing her hearing aids? is the Professor's tinnitus loud this evening? is it windy? is the cat hoarse from meowing relentlessly to be let outside all day?

Because of these and other uncertainties, the meow intensity heard by the professor is a random variable with probability density function (pdf) described by $\mathcal{L}_s(d) = 0$ for $s > s_{\text{ideal}}(d)$ and

$$\mathcal{L}_s(d) = \frac{1}{s_{\text{ideal}}(d)}\,\mathrm{H}(\eta) \tag{5.8.4}$$

for $s \le s_{\text{ideal}}(d)$, where

$$\eta = \frac{s}{s_{\text{ideal}}(d)} \tag{5.8.5}$$

is a scaled intensity and $\mathrm{H}(\eta)$ is a universal function defined in the interval $0 \le \eta \le 1$. Because

$$\int_0^{s_{\rm ideal}(d)} \mathcal{L}_s(d)\,{\rm d}s = \frac{1}{s_{\rm ideal}} \int_0^{s_{\rm ideal}} \mathrm{H}(\eta)\,{\rm d}s = 1, \qquad (5.8.6)$$

we must have that

$$\int_0^1 \mathrm{H}(\eta)\,{\rm d}\eta = 1. \qquad (5.8.7)$$

The stipulation $\mathcal{L}_s(d) = 0$ for $s > s_{\rm ideal}(d)$ ensures that the Professor cannot hear a meow louder than the ideal meow for a certain distance, $d$. The likelihood function $\mathcal{L}_s(d)$ incorporates random imperfections of both the receiver (professor) and the source (Byte).

Using expression (5.8.1), we find that

$$\eta = \frac{s}{s_0}\,\frac{1}{\psi(d)}. \qquad (5.8.8)$$

Consequently,

$$\mathcal{L}_s(d) = \frac{1}{s_0}\,\frac{1}{\psi(d)}\,\mathrm{H}(\frac{s}{s_0}\,\frac{1}{\psi(d)}) \qquad (5.8.9)$$

for $s \le s_{\rm ideal}(d)$ and

$$\mathcal{L}_s(d) = 0 \qquad (5.8.10)$$

for $s > s_{\rm ideal}(d)$. Substituting the acoustic law (5.8.2) for $\psi(d)$, we find that

$$\mathcal{L}_s(d) = \frac{1}{s_0}\,\frac{a^2+d^2}{a^2}\,\mathrm{H}(\frac{s}{s_0}\,\frac{a^2+d^2}{a^2}) \qquad (5.8.11)$$

for $s \le s_{\rm ideal}(d)$ or

$$\mathcal{L}_s(d) = 0 \qquad (5.8.12)$$

for $s > s_{\rm ideal}(d)$.

For a specified scaled loudness $s/s_0 < 1$, the maximum value of $d$ beyond which $\mathcal{L}_s$ is zero, denoted by $d_{\rm max}$, arises when $\psi(d_{\rm max}) = s/s_0$, yielding

$$d_{\rm max} = a\sqrt{\frac{s_0}{s} - 1}. \tag{5.8.13}$$

This formula arises from the inversion of the acoustic law (5.8.2).

### *5.8.5 The* H *function*

To assess the function $\mathrm{H}(\eta)$, the professor positions the cat at a certain distance, $d$, asks the cat to meow repeatedly, and makes a histogram of the cat meow intensity heard. To minimize Byte's annoyance, the Professor performs this calibration only at one distance, $d$, and assumes that the scaled response applies for any distance.

The professor decides to fit her measurements to a family of functions parametrized by an integer, $k$,

$$\mathrm{H}(\eta) = \mathcal{A}_k(\eta), \tag{5.8.14}$$

where

$$\mathcal{A}_k(\eta) = c_k \sin^{2k}(\tfrac{1}{2}\,\eta\pi) \tag{5.8.15}$$

for $0 \le \eta \le 1$, and

$$c_k = \frac{(2k)\,!\,!}{(2k-1)\,!\,!} \equiv \frac{2\cdot 4\cdots(2k)}{1\cdot 3\cdots(2k-1)}. \tag{5.8.16}$$

The professor was careful to verify with the help of tables of definite integrals that this distribution satisfies the constraint (5.8.7) (e.g., Gradshteyn, I. S. & Ryzhik, I. M. (1980) *Table of Integrals, Series, and Products.* Academic Press, p. 369, § 3.621.3).

Graphs of the function $A_k(\eta)$ for several values of the half-exponent, $k$, is shown in Figure 5.8.2(*a*). The area under each curve is equal to unity. As the exponent $k$ increases, $A_k(\eta)$ tends to twice a Dirac delta function forced at $\eta = 1$.

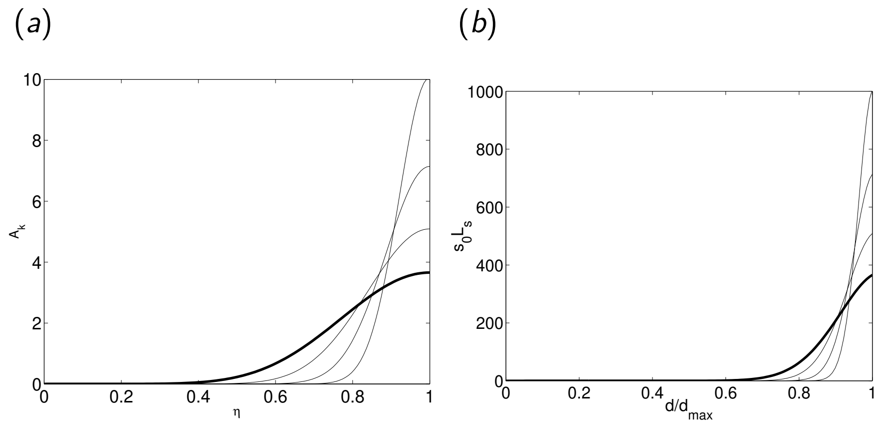



FIGURE 5.8.2 Graphs of the (*a*) dimensionless function $\mathcal{A}_k(\eta)$ determining the probability density function (pdf) of the meow heard by the professor and (*b*) likelihood function, $\mathcal{L}_s$ scaled by $1/s_0$ for a meow heard by the Professor with intensity $s = 0.01 \times s_0$ at a certain distance, for $k = 4$ (heavy line), $8, 16$, and $32$.

### 5.8.6 The likelihood function

The likelihood function is given by

$$\mathcal{L}_s(d) = \frac{1}{s_0} \frac{a^2 + d^2}{a^2} \mathcal{A}_k(\frac{s}{s_0} \frac{a^2 + d^2}{a^2}) \tag{5.8.17}$$

for $s \le s_{\text{ideal}}(d)$ or

$$\mathcal{L}_s(d) = 0 \tag{5.8.18}$$

for $s > s_{\text{ideal}}(d)$.

Graphs of the likelihood function for $a = 1$ ft and acoustic signal intensity $s = 0.01 \times s_0$ are shown in Figure 5.8.2(*b*) for several values of the half-exponent, $k$. As $k$ increases, $\mathcal{L}_k(\eta)$ tends to twice the Dirac delta function forced at $d = d_{\max}$.

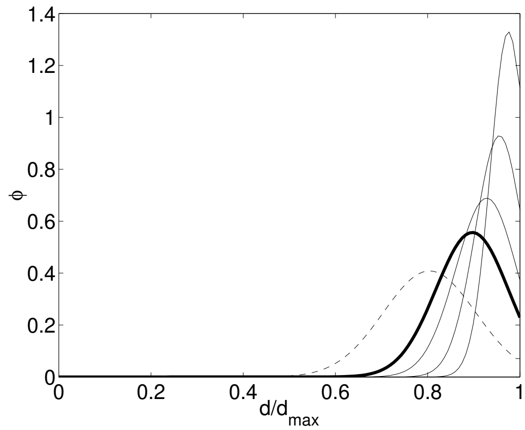



FIGURE 5.8.3 Prior (broken line) and posterior (solid line) pdf of Byte's distance for meow loudness $s = 0.01 \times s_0$ and exponent $k = 4$ (heavy red line), $8, 16$, and $32$.

### *5.8.7 Bayesian analysis*

The professor uses Bayes's theorem to obtain the posterior pdf

$$\phi(d \mid s) = \frac{\mathcal{L}_s(d)}{\displaystyle\int_0^\infty \mathcal{L}_s(d')\,\phi(d')\,\mathrm{d}d'} \times \phi(d). \tag{5.8.19}$$

Since Bytes likes to hang around the cat mint, which is $b = 8$ feet away, the professor uses a normal prior with standard deviation equal to $a$ to account for Byte's natural curiosity,

$$\phi(d) = \frac{1}{\sqrt{2\pi}\,a}\,\exp(-\tfrac{1}{2}\,\frac{(d-b)^2}{a^2}). \tag{5.8.20}$$

The prior and posterior pdfs are shown in Figure 5.8.3 for several values of $k$. Professor Ritewell's calibration is best fitted for $k = 4$. For this value of $k$, Byte is probably a little farther away than Dr. Ritewell originally thought.

### 5.8.8 Final act

The essence of the Bayesian analysis is that the meow intensity heard by the Professor, $s$, is attributed to any cat distance, $d$, up to a maximum, through a chosen or measured likelihood function. The professor calls Byte and the cat runs to her for a vigorous petting and scratching session followed by a delicious treat.

### 5.8.9 Misguided neighbor

Professor Ritewell's neighbor is an honest and generous but misguided man. Over the years, the neighbor has been making comments on how fortunate tenured professors are in that they work so little (eight months a year or less), and yet they are rewarded by so much. In fact, the neighbor is repeating opinions expressed in television and radio talk shows by people who have not climbed the steep mountain of creative activity.

When the neighbor makes such comments, professor Ritewell looks away and smiles evasively. The professor knows by experience that true scholars take an oath of lifetime poverty, make an implied promise of daily creativity, agree to perform on demand in front of tough audiences and harsh critics, and endure a deluge of absurd administrative emails with a perfunctory "take care" greeting above the signature line (administrative spam.)

M. G., Piety wrote that "*Academics are not really the plague that they are increasingly represented as being, but there is, lamentably, a sizable contingent that gives the rest of us a bad name.*"

#### Exercise

**5.8.1** Explain why it does not make sense to use the Gaussian distribution as a likelihood function for Byte's meow intensity.

## 5.9 Find your roots

In mathematics, a function of one independent variable is a virtual

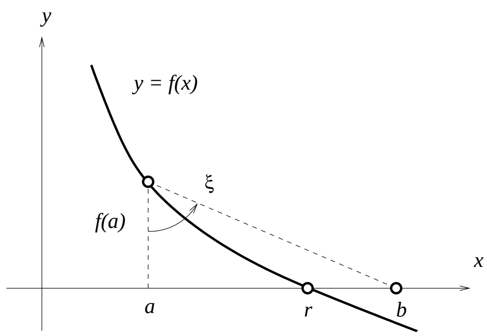



FIGURE 5.9.1 Graph of a function $f(x)$, whose root, $r$ is desired. The root is located at the intersection of the graph of the function with the $x$ axis.

machine that receives a number, $x$, and generates another number, $f(x)$. The number generation can be implemented by a mathematical formula, such as $f(x) = \exp(x)$, or else arises from a simple or involved numerical computation, physical measurement, or quantitative deduction.

In quantum chemistry and other advanced computational science disciplines, generating a single function value may require years of computational time even on the fastest available supercomputer.

### *5.9.1 Root*

We are often interested in finding a root of a function, $f(x)$, denoted by $r$, also called a zero, such that

$$f(r) = 0. \tag{5.9.1}$$

The root is located at the intersection of the graph of the function with the $x$ axis, as shown in Figure 5.9.1.

### *5.9.2 Stochasticity*

A seemingly inefficient method of locating the root involves guessing a value, $r = a$, as shown in Figure 5.9.1, and computing $f(a)$. If $f(a) = 0$, we are done.

Otherwise, we may introduce the angle $\xi$ subtended between the vertical line and a straight line that is meant to approximate the graph of the function, $f(x)$, as shown in Figure 5.9.1, where $-\frac{1}{2}\pi \le \xi \le \frac{1}{2}\pi$. By definition,

$$\tan\xi = \frac{b-a}{f(a)}, \tag{5.9.2}$$

where $b$ is an approximation to the root. For some specific yet unknown value of $\xi$, $b$ will be the root.

Since are not certain about the value of $\xi$ that leads us to the root, it appears sensible to describe it by a probability density function (pdf), denoted by $\psi(\xi)$.

Correspondingly, for given $a$ and $f(a)$, the estimate $b$ is described by a related pdf, $\phi(b)$. The two pdfs are related by the equation

$$\frac{\phi(b)}{\psi(\xi)} = \left|\frac{\mathrm{d}\xi}{\mathrm{d}b}\right|, \tag{5.9.3}$$

where the vertical bars enclose the absolute value, as discussed in Section 5.3. Differentiating both sides of (5.9.2) with respect to $b$, we obtain

$$\frac{1}{\cos^2\xi}\frac{\mathrm{d}\xi}{\mathrm{d}b} = \frac{1}{f(a)}. \tag{5.9.4}$$

Noting that

$$\frac{1}{\cos^2\xi} = 1 + \tan^2\xi = 1 + \frac{(b-a)^2}{f^2(a)} = \frac{f^2(a)+(b-a)^2}{f^2(a)}, \tag{5.9.5}$$

we obtain

$$\frac{\mathrm{d}\xi}{\mathrm{d}b} = \frac{f(a)}{f^2(a)+(b-a)^2}. \tag{5.9.6}$$

Finally, we refer to (5.9.3) and find that

$$\phi(b) = \frac{|f(a)|}{f^2(a)+(b-a)^2}\,\psi(\xi). \tag{5.9.7}$$

### *5.9.3 Bayesian analysis*

The pdf shown in (5.9.7) for $\phi(b)$ plays the role of a likelihood function in a Bayesian analysis for $b$, derived by geometrical reasoning alone,

$$\mathcal{L}_{a,f(a)}(b) = \frac{|f(a)|}{f^2(a) + (b-a)^2}\,\psi(\xi), \tag{5.9.8}$$

where the doublet

$$\mathbf{x} \equiv (a, f(a)) \tag{5.9.9}$$

is the data. To validate this likelihood function, we compute

$$\int_{-\infty}^{\infty} \mathcal{L}_{a,f(a)}(b)\,\mathrm{d}a = \int_{-\infty}^{\infty} \frac{|f(a)|}{f^2(a) + (b-a)^2}\,\psi(\xi)\,\mathrm{d}a \tag{5.9.10}$$

or

$$\int_{-\infty}^{\infty} \mathcal{L}_{a,f(a)}(b)\,\mathrm{d}a = \int_{-\pi/2}^{\pi/2} \psi(\xi)\,\mathrm{d}\xi = 1, \tag{5.9.11}$$

as required. However, to derive this value we have regarded $f(a)$ as independent of $a$, which is an approximation.

Bayes equation for becomes

$$\phi(b\,|\,a, f(a)) = \frac{\mathcal{L}_{a,f(a)}(b)}{\varphi_{a,f(a)}} \times \phi(b), \tag{5.9.12}$$

where

$$\varphi_{a,f(a)} \equiv \int_{-\infty}^{\infty} \mathcal{L}_{a,f(a)}(b) \times \phi(b)\,\mathrm{d}b \tag{5.9.13}$$

is a normalization factor.

### *5.9.4 Uniform pdf for $\psi(\xi)$*

In the event of a uniform pdf $\psi(\xi) = 1/\pi$ for $-\frac{1}{2}\pi \leq \xi \leq \frac{1}{2}\pi$, we obtain the Bayes equation

$$\phi(b\,|\,a, f(a)) = \frac{1}{\pi}\,\frac{1}{\varphi_{a,f(a)}}\,\frac{|f(a)|}{f(a)^2 + (b-a)^2} \times \phi(b). \tag{5.9.14}$$

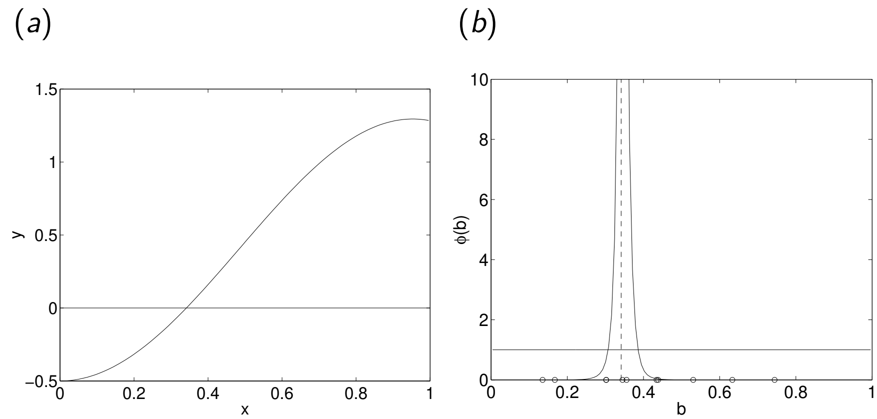



FIGURE 5.9.2 (*a*) Graph of the function $f(x) = 2\sin^2(\frac{1}{2}\pi x) - 0.5\cosh^3(\frac{1}{2}x)$ and (*b*) evolving pdf of the root location starting with a flat distribution in the interval $[0, 1]$.

A prior pdf $\phi(b)$ must be assumed.

As an example, we consider the function

$$f(x) = 2\,\sin^2(\frac{1}{2}\pi x) - \frac{1}{2}\,\cosh^3(\frac{1}{2}\,x) \tag{5.9.15}$$

displayed in Figure 5.9.2(*a*). A root in the interval $[0, 1]$ is known to be $r = 0.3414469$, to shown accuracy.

In the numerical method, random trial values of $a$ are selected by a standard random generator, indicated by the circular symbols on the horizontal axis of Figure 5.9.2(*b*). The pdf evolves by the repeated application of Bayes equation (5.9.14).

The initial flat prior and the evolved pdf after $10$ trials are plotted in Figure 5.9.2(*b*). After ten trials, the pdf exhibits a definite peak at the location of the root, $r$, indicated by the broken vertical line.

### *5.9.5 Extensions and improvements*

The method can be extended such that the straight dashed line in Figure 5.9.1 is replaced by a parabola whose slope angle and curvature at the trial point $a$ are independent stochastic variables. Moreover, the method could be extended to systems of two or more nonlinear equations.

*Exercise*

**5.9.1** Reproduce the counterpart of Figure 5.9.2 for a linear function of your choice.

# Chapter 6

# Joint pdfs

Bayes formula for the probability density function (pdf) of a physical or conceptual system that depends on one parameter, $\theta$, such as the Bernoulli probability of a coin flip or an unsmudged photocopy, can be generalized to events that depend on two or more stochastic parameters.

In applications, the set of these parameters may describe a multi-component event or hypothesis, or a multivariate model. The notion and basic properties of a joint probability density function (jpdf) and the corresponding Bayes equation are discussed in this chapter with illustrative applications.

## 6.1 Joint pdf in two dimensions

Consider two continuous stochastic (random) variables, $\xi_1$ and $\xi_2$, taking values inside a specified domain $\mathcal{D}$ in the $\xi_1\xi_2$ plane, as shown in Figure 6.1.1.

For example, the random variable $\xi_1$ can be the temperature, and the random variable $\xi_2$ can be the humidity of the atmosphere. In another example, $\xi_1$ can be the amount of fuel left in the tank of a spacecraft, and $\xi_2$ can be the distance to the nearest planet.

The joint probability density function (pdf) of the doublet $\boldsymbol{\xi} \equiv (\xi_1, \xi_2)$ is denoted as $\phi(\xi_1, \xi_2)$ or $\phi(\boldsymbol{\xi})$, with the understanding that $\phi(\boldsymbol{\xi})$ is non-negative inside $\mathcal{D}$ and uniformly zero outside $\mathcal{D}$, as shown in Figure 6.1.1.

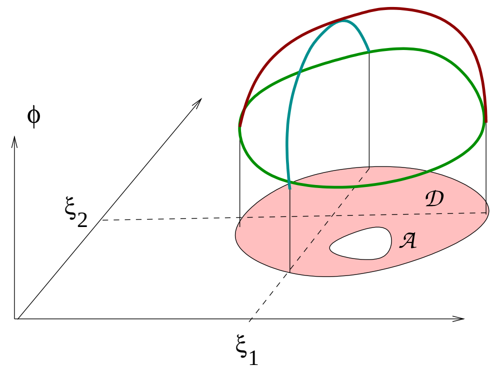



FIGURE 6.1.1 Surface graph of a joint probability density function of two random variables, $\xi_1$ and $\xi_2$, taking values inside a certain domain, $\mathcal{D}$.

### 6.1.1 Definition

By definition, the probability that $\boldsymbol{\xi}$ takes values inside an arbitrary area $\mathcal{A}$ that lies entirely inside $\mathcal{D}$, as shown in Figure 6.1.1, is given by the areal integral

$$\Pi_{\mathcal{A}} = \iint_{\mathcal{A}} \phi(\boldsymbol{\xi}) \, \mathrm{d}A(\boldsymbol{\xi}), \tag{6.1.1}$$

where $\mathrm{d}A$ is a differential area in the $\boldsymbol{\xi}$ plane. Physically, the integral is the volume between the surface graph of the joint pdf, the $\xi_1\xi_2$ plane, and a vertical cylindrical surface tracing the boundary of $\mathcal{A}$.

The integration in the $\xi_1\xi_2$ plane can be performed in Cartesian, polar, or any other convenient curvilinear coordinates. In Cartesian coordinates, $\mathrm{d}A = \mathrm{d}\xi_1 \, \mathrm{d}\xi_2$. In plane polar coordinates, $\mathrm{d}A = r \, \mathrm{d}r \, \mathrm{d}\theta$.

The probability that $\boldsymbol{\xi}$ falls inside a small rectangular area $\mathrm{d}\mathcal{A}$ with differential sides $\mathrm{d}\xi_1$ and $\mathrm{d}\xi_2$ is given by

$$\Pi_{\mathrm{d}\mathcal{A}} = \phi(\boldsymbol{\xi}) \, \mathrm{d}\xi_1 \, \mathrm{d}\xi_2. \tag{6.1.2}$$

Similar expressions can be written for differential finite or infinitesimal areas with different shapes.

### *6.1.2 Normalization*

Normalization imposes the constraint

$$\Pi_{\mathcal{D}} = \iint_{\mathcal{D}} \phi(\boldsymbol{\xi})\, \mathrm{d}A(\boldsymbol{\xi}) = 1, \tag{6.1.3}$$

which stipulates that the doublet $\boldsymbol{\xi}$ must take some value, any value, inside $\mathcal{D}$. The integral on the left-hand side is the volume confined by the surface graph of the joint pdf $\phi$, the $\xi_1\xi_2$ plane, and a vertical cylindrical surface tracing the boundary of $\mathcal{D}$, as shown in Figure 6.1.1.

Performing the integration in Cartesian coordinates, we find that

$$\iint_{\mathcal{D}} \phi(\xi_1, \xi_2)\, \mathrm{d}\xi_1\, \mathrm{d}\xi_2 = 1, \tag{6.1.4}$$

for suitable integration limits. In the literature, it is routinely assumed that the limits of integration cover the entire $\boldsymbol{\xi}$ plane, with the understanding that $\phi(\xi_1, \xi_2)$ is uniformly zero outside $\mathcal{D}$.

### *6.1.3 Marginal probabilities*

The joint probability density function can be used to compute the first marginal probability density function

$$\phi_1(\xi_1) = \int \phi(\xi_1, \xi_2)\, \mathrm{d}\xi_2 \tag{6.1.5}$$

and the first marginal probability density function

$$\phi_2(\xi_2) = \int \phi(\xi_1, \xi_2)\, \mathrm{d}\xi_1, \tag{6.1.6}$$

for suitable integration limits.

The marginal probability $\phi_1(\xi_1)$ is the area in the vertical plane normal to the $\xi_1$ axis under the graph of the joint pdf, as shown in Figure 6.1.1. Likewise, the marginal probability $\phi_2(\xi_2)$ is the area in the vertical plane normal to the $\xi_2$ axis under the graph of the joint pdf, as shown in Figure 6.1.1.

Since by definition

$$\int \phi_1(\xi_1)\, \mathrm{d}\xi_1 = 1, \qquad \int \phi_2(\xi_2)\, \mathrm{d}\xi_2 = 1, \tag{6.1.7}$$

the normalization constraint (6.1.4) is satisfied.

### 6.1.4 Expected values and variances

The expected values of $\xi_1$ and $\xi_2$ are given by

$$\overline{\xi}_1 \equiv \iint_{\mathcal{D}} \xi_1\, \phi(\xi_1, \xi_2)\, \mathrm{d}\xi_1\, \mathrm{d}\xi_2 = \int \xi_1\, \phi_1(\xi_1)\, \mathrm{d}\xi_1 \tag{6.1.8}$$

and

$$\overline{\xi}_2 \equiv \iint_{\mathcal{D}} \xi_2\, \phi(\xi_1, \xi_2)\, \mathrm{d}\xi_1\, \mathrm{d}\xi_2 = \int \xi_2\, \phi_2(\xi_1)\, \mathrm{d}\xi_2. \tag{6.1.9}$$

The corresponding variances are defined by similar integrals,

$$\sigma_1 \equiv \iint_{\mathcal{D}} (\xi_1 - \overline{\xi}_1)^2\, \phi(\xi_1, \xi_2)\, \mathrm{d}\xi_1\, \mathrm{d}\xi_2 \tag{6.1.10}$$

and

$$\sigma_2 \equiv \iint_{\mathcal{D}} (\xi_2 - \overline{\xi}_2)^2\, \phi(\xi_1, \xi_2)\, \mathrm{d}\xi_1\, \mathrm{d}\xi_2. \tag{6.1.11}$$

The expected value of an arbitrary random function, $f(\xi_1, \xi_2)$, is given by

$$\overline{f} \equiv \iint_{\mathcal{D}} f(\xi_1, \xi_2)\, \phi(\xi_1, \xi_2)\, \mathrm{d}\xi_1\, \mathrm{d}\xi_2. \tag{6.1.12}$$

For example, $\xi_1$ can be the temperature, $\xi_2$ can be the humidity, and $f(\xi_1, \xi_2)$ can be the crop yield.

### *6.1.5 Conditional probabilities*

The conditional probabilities associated with a joint probability are defined in terms of the marginal probabilities by $\phi(\xi_1)$ and $\phi(\xi_2)$ by

$$\phi(\xi_1 \,|\, \xi_2) = \frac{\phi(\xi_1\,,\,\xi_2)}{\phi_2(\xi_2)}, \qquad \phi(\xi_2 \,|\, \xi_1) = \frac{\phi(\xi_2\,,\,\xi_1)}{\phi_1(\xi_1)}. \tag{6.1.13}$$

The graphs of the conditional probabilities are scaled intersections of two vertical planes and the graph of the joint pdf, as shown in Figure 6.1.1.

Integrating the first equation in (6.1.13), we find that

$$\int \phi(\xi_1 \,|\, \xi_2)\, \mathrm{d}\xi_1 = \frac{1}{\phi_2(\xi_2)} \int \phi(\xi_1\,,\,\xi_2)\, \mathrm{d}\xi_1 = 1, \tag{6.1.14}$$

and similarly

$$\int \phi(\xi_2 \,|\, \xi_1)\, \mathrm{d}\xi_2 = 1. \tag{6.1.15}$$

The marginal probabilities are given by

$$\phi_1(\xi_1) = \int \phi(\xi_1 \,|\, \xi_2)\, \phi_2(\xi_2)\, \mathrm{d}\xi_2 \tag{6.1.16}$$

and

$$\phi_2(\xi_2) = \int \phi(\xi_2 \,|\, \xi_1)\, \phi_1(\xi_1)\, \mathrm{d}\xi_1. \tag{6.1.17}$$

### *6.1.6 Bayes rule*

It follows from the definition of conditional probabilities that

$$\phi(\xi_1, \xi_2) = \phi(\xi_1 \,|\, \xi_2) \times \phi_2(\xi_2) = \phi(\xi_2 \,|\, \xi_1) \times \phi_1(\xi_1). \tag{6.1.18}$$

Rearranging, we derive the Bayes equation

$$\phi(\xi_1 \,|\, \xi_2) = \frac{\phi(\xi_2 \,|\, \xi_1)}{\phi_2(\xi_2)} \times \phi_1(\xi_1). \tag{6.1.19}$$

Substituting the definition of the marginal probability in the denominator, we obtain

$$\phi(\xi_1 \,|\, \xi_2) = \frac{\phi(\xi_2 \,|\, \xi_1)}{\displaystyle\int \phi(\xi_2 \,|\, \xi_1')\, \phi_1(\xi_1')\, \mathrm{d}\xi_1'} \times \phi_1(\xi_1). \tag{6.1.20}$$

Integrating both sides with respect to $\xi_1$, we find that

$$\int \phi(\xi_1 \,|\, \xi_2)\, \mathrm{d}\xi_1 = 1, \tag{6.1.21}$$

as required.

In fact, equation (6.1.20) is the same as equation (4.8.6) applicable to continuous events parametrized by $\theta$ and continuous data parametrized by $x$, repeated below for convenience,

$$\phi(\theta \,|\, x) = \frac{\mathcal{L}_x(\theta)}{\displaystyle\int_{\Theta(x)} \mathcal{L}_x(\theta') \times \phi(\theta')\, \mathrm{d}\theta'} \times \phi(\theta), \tag{6.1.22}$$

subject to the substitutions

$$\xi_1 \to \theta, \quad \xi_2 \to x, \quad \phi(\xi_2 \,|\, \xi_1) \to \mathcal{L}_x(\theta), \quad \phi_1(\xi_1) \to \phi(\theta). \tag{6.1.23}$$

This comparison indicates that, in practice, $\xi_1$ plays the role of an event parameter and $\xi_2$ plays the role of continuous data.

In the feline application discussed in Section 6.8, $\xi_1$ is the distance from a cat named Byte and $\xi_2$ is the intensity of Byte's meow.

### *6.1.7 Independence*

If two individual causes or events parametrized by $\xi_1$ and $\xi_2$ are independent, the joint pdf is separable,

$$\phi(\xi_1, \xi_2) = \phi_1(\xi_1) \times \phi_2(\xi_2), \tag{6.1.24}$$

where $\phi_1(\xi_1)$ and $\phi_2(\xi_2)$ are the marginal probabilities. Using the definition of the marginal probabilities, we confirm that

$$\begin{aligned}\int \phi(\xi_1, \xi_2)\, \mathrm{d}\xi_2 &= \int \phi_1(\xi_1)\, \phi_2(\xi_2)\, \mathrm{d}\xi_2 \\ &= \phi_1(\xi_1) \int \phi_2(\xi_2)\, \mathrm{d}\xi_2 = \phi_1(\xi_1),\end{aligned} \tag{6.1.25}$$

and similarly

$$\begin{aligned}\int \phi(\xi_1, \xi_2)\, \mathrm{d}\xi_1 &= \int \phi_1(\xi_1)\, \phi_2(\xi_2)\, \mathrm{d}\xi_1 \\ &= \phi_2(\xi_2) \int \phi_1(\xi_1)\, \mathrm{d}\xi_1 = \phi_2(\xi_2).\end{aligned} \tag{6.1.26}$$

### *6.1.8 Discrete distributions*

If the pair $(\xi_1, \xi_2)$ is only allowed to take $n$ discrete values $\boldsymbol{\xi}^{(i)}$ for $i = 1, \ldots, n$, then

$$\phi(\xi_1, \xi_2) = \sum_{i=1}^{n} \pi_i\, \delta_2(\mathbf{x} - \boldsymbol{\xi}^{(i)}), \tag{6.1.27}$$

where $\delta_2$ is the Dirac delta function in two dimensions and the pointwise probabilities $\pi_i$ must add to unity,

$$\sum_{i=1}^{n} \pi_i = 1. \tag{6.1.28}$$

The pdf is zero everywhere and spikes to infinity at the isolated points $\boldsymbol{\xi}^{(i)}$ for $i = 1, \ldots, n$.

The expected value of an arbitrary random function, $f(\xi_1, \xi_2)$, is

$$\overline{f} = \sum_{i=1}^{n} \pi_i f(\xi_1^{(i)}, \xi_2^{(i)}). \tag{6.1.29}$$

***Exercise***

**6.1.1** Express the normalization condition (6.1.3) in plane polar coordinates in the $\xi_1\xi_2$ plane.

## 6.2 Bilinear joint pdf

Assume that the domain of definition, $\mathcal{D}$, is a rectangle in the $\xi_1\xi_2$ plane confined inside $a_1 \le \xi_1 \le b_1$ and $a_2 \le \xi_2 \le b_2$. The joint pdf can be approximated with a bilinear function,

$$\phi(\xi_1, \xi_2) = c_0 + c_1\,\xi_1 + c_2\,\xi_2, \tag{6.2.1}$$

where $c_0, c_1, c_2$ are three constants. Normalization requires that

$$\int_{a_2}^{b_2} \int_{a_1}^{b_1} \big(c_0 + c_1\,\xi_1 + c_2\,\xi_2\big)\,\mathrm{d}\xi_1\,\mathrm{d}\xi_2 = 1. \tag{6.2.2}$$

Performing the integration and rearranging, we obtain

$$c_0 + \tfrac{1}{2}\,(a_1 + b_1)\,c_1 + \tfrac{1}{2}\,(a_2 + b_2)\,c_2 = \frac{1}{(b_1 - a_1)(b_2 - a_2)}. \tag{6.2.3}$$

Any two of the three constants, $c_0$, $c_1$, $c_2$, can be assigned independently, and the third constant must be computed to satisfy this mandatory normalization condition.

### 6.2.1 Two known points

If the joint pdf is known at two points in the parameter plane,

$$\boldsymbol{\xi}^{(1)} = (\xi_1^{(1)}, \xi_2^{(1)}), \qquad \boldsymbol{\xi}^{(2)} = (\xi_1^{(2)}, \xi_2^{(2)}), \tag{6.2.4}$$

then

$$\phi(\xi_1^{(1)}, \xi_2^{(1)}) = v_1, \qquad \phi(\xi_1^{(2)}, \xi_2^{(2)}) = v_2, \tag{6.2.5}$$

where $v_1$ and $v_2$ are given values. Substituting the linear approximation, we obtain two equations,

$$c_0 + c_1\,\xi_1^{(1)} + c_2\,\xi_2^{(1)} = v_1, \qquad c_0 + c_1\,\xi_1^{(2)} + c_2\,\xi_2^{(2)} = v_2. \tag{6.2.6}$$

The system of three linear equations consisting of (6.2.6) and (6.2.3) can be solved to obtain $c_0$, $c_1$, and $c_2$. In matrix notation,

$$\begin{bmatrix} 1 & \xi_1^{(1)} & \xi_2^{(1)} \\ 1 & \xi_1^{(2)} & \xi_2^{(2)} \\ 1 & \frac{1}{2}(a_1+b_1) & \frac{1}{2}(a_2+b_2) \end{bmatrix} \cdot \begin{bmatrix} c_0 \\ c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} v_1 \\ v_2 \\ \dfrac{1}{(b_1-a_1)(b_2-a_2)} \end{bmatrix}. \tag{6.2.7}$$

Data points where the matrix on the left-hand side is singular should be avoided.

### 6.2.2 *Marginal and conditional probabilities*

The marginal probabilities are given by

$$\phi_1(\xi_1) = \int_{a_2}^{b_2} (c_0 + c_1\,\xi_1 + c_2\,\xi_2)\,\mathrm{d}\xi_2 \tag{6.2.8}$$

and

$$\phi_2(\xi_2) = \int_{a_1}^{b_1} (c_0 + c_1\,\xi_1 + c_2\,\xi_2)\,\mathrm{d}\xi_1. \tag{6.2.9}$$

Performing the integrations, we obtain

$$\phi_1(\xi_1) = \mu_1\xi_1 + \eta_1, \qquad \phi_2(\xi_2) = \mu_2\,\xi_2 + \eta_2, \tag{6.2.10}$$

where

$$\begin{aligned} \mu_1 &= c_1\,(b_2-a_2), \quad \eta_1 = (b_2-a_2)\left(c_0 + \tfrac{1}{2}\,(a_2+b_2)\,c_2\right) \\ \mu_2 &= c_2\,(b_1-a_1), \quad \eta_2 = (b_1-a_1)\left(c_0 + \tfrac{1}{2}\,(a_1+b_1)\,c_1\right). \end{aligned} \tag{6.2.11}$$

The conditional probabilities are given by

$$\phi(\xi_1\,|\,\xi_2) = \frac{\phi(\xi_1,\xi_2)}{\phi_2(\xi_2)} = \frac{c_0 + c_1\xi_1 + c_2\xi_2}{\mu_2\xi_2 + \eta_2} \tag{6.2.12}$$

and

$$\phi(\xi_2 \,|\, \xi_1) = \frac{\phi(\xi_1, \xi_2)}{\phi_1(\xi_1)} = \frac{c_0 + c_1 \xi_1 + c_2 \xi_2}{\mu_1 \xi_1 + \eta_1}. \qquad (6.2.13)$$

***Exercise***

**6.2.1** Discuss whether the stochastic parameters $\theta_1$ and $\theta_2$ in (6.2.1) could be independent.

## *6.3 Joint pdf in arbitrary dimensions*

Suppose that a random outcome is determined by an arbitrary number of $\nu$ random parameters encapsulated in an array,

$$\boldsymbol{\xi} = (\xi_1, \xi_2, \ldots, \xi_\nu). \qquad (6.3.1)$$

The $\nu$ individual scalar parameters, $\xi_1$, $\xi_2$, ..., $\xi_\nu$ are defined inside a specified volume, $\mathcal{D}$, in $\nu$-dimensional space according to the prevailing physical or conceptual context.

For example, $\mathcal{D}$ can be a circular disk in two dimensions, a sphere in three dimensions, or a hypersphere in higher dimensions. If each parameter varies in the interval $[0, 1]$, independent of the values of all other parameters, then the domain of definition, $\mathcal{D}$, is a square, a cube, or a hypercube.

### *6.3.1 Definition*

The associated joint probability density function (pdf) is denoted by

$$\phi(\,\xi_1, \xi_2, \ldots, \xi_\nu\,) \qquad (6.3.2)$$

or $\phi(\boldsymbol{\xi})$. By definition, the probability that $\boldsymbol{\xi}$ takes values inside a volume $\mathcal{S}$ that resides inside $\mathcal{D}$ is given by the $\nu$-dimensional integral

$$\int_{\mathcal{S}} \phi(\boldsymbol{\xi}) \, \mathrm{d}V(\boldsymbol{\xi}), \qquad (6.3.3)$$

where $\mathrm{d}V(\boldsymbol{\xi})$ is a differential volume in the $\boldsymbol{\xi}$ space consisting of the $\xi_1$, $\xi_2$, ..., $\xi_\nu$ axes.

### *6.3.2 Normalization*

Normalization imposes the constraint

$$\int_{\mathcal{D}} \phi(\boldsymbol{\xi})\,\mathrm{d}V(\boldsymbol{\xi}) = 1, \tag{6.3.4}$$

which requires that the $\nu$-tuple $\boldsymbol{\xi}$ must take some value, any value, inside the volume $\mathcal{D}$.

### *6.3.3 Marginal and conditional probabilities*

Marginal and conditional probabilities are defined as in the case of two parameters discussed in Section 6.1 for $\nu = 2$. Invoking the definition of the conditional probability, we find that

$$\phi(\,\xi_1, \ldots, \xi_\nu\,) = \phi(\,\xi_\nu\,|\,\xi_1, \ldots, \xi_{\nu-1}\,) \times \phi(\xi_1, \ldots, \xi_{\nu-1}\,), \tag{6.3.5}$$

where

$$\phi(\xi_1, \ldots, \xi_{\nu-1}) = \int \phi(\,\xi_1, \ldots, \xi_\nu\,)\,\mathrm{d}\xi_\nu \tag{6.3.6}$$

is a joint marginal probability. In the discussion of Section 6.3.1 for $\nu = 2$, this marginal probability was denoted as $\phi_1(\xi_1)$. The conditional probability $\phi(\,\xi_\nu\,|\,\xi_1, \ldots, \xi_{\nu-1}\,)$ shown in (6.3.5) satisfies the normalization condition

$$\int \phi(\,\xi_\nu\,|\,\xi_1, \ldots, \xi_{\nu-1}\,)\,\mathrm{d}\xi_\nu = 1. \tag{6.3.7}$$

Repeating the process, we find that

$$\begin{aligned}\phi(\,\xi_1, \ldots, \xi_\nu\,) = {} & \phi(\,\xi_\nu\,|\,\xi_1, \ldots, \xi_{\nu-1}\,) \\ & \times \phi(\,\xi_{\nu-1}\,|\,\xi_1, \ldots, \xi_{\nu-2}\,) \times \phi(\xi_1, \ldots, \xi_{\nu-2}\,).\end{aligned} \tag{6.3.8}$$

where

$$\phi(\xi_1,\ldots,\xi_{\nu-2}) = \iint \phi(\xi_1,\ldots,\xi_\nu)\,\mathrm{d}\xi_{\nu-1}\,\mathrm{d}\xi_\nu \tag{6.3.9}$$

is a marginal joint probability. The second conditional probability shown on the right-hand side of (6.3.8) satisfies the normalization conditions

$$\int \phi(\,\xi_{\nu-1}\,|\,\xi_1,\ldots,\xi_{\nu-2}\,)\,\mathrm{d}\xi_{\nu-1} = 1 \tag{6.3.10}$$

Repeating the process further, we obtain the ultimate chain rule

$$\phi(\,\xi_1,\xi_2,\ldots,\xi_\nu\,) = \phi(\xi_1) \times \prod_{m=1}^{\nu-1} \phi(\,\xi_{m+1}\,|\,\xi_1,\ldots,\xi_m\,), \tag{6.3.11}$$

where $\prod$ denotes the product. The first conditional probability in the product for $m=1$ is $\phi(\,\xi_2\,|\,\xi_1)$, the penultimate conditional probability for $m=\nu-2$ is $\phi(\,\xi_{\nu-1}\,|\,\xi_1,\ldots,\xi_{\nu-2}\,)$, and the last conditional probability for $m=\nu-1$ is $\phi(\,\xi_\nu\,|\,\xi_1,\ldots,\xi_{\nu-1}\,)$, as shown in (6.3.5).

### *6.3.4 Bayes rule*

An alternative to (6.3.5) is

$$\phi(\,\xi_1,\xi_2,\ldots,\xi_\nu\,) = \phi(\,\xi_1,\ldots,\xi_{\nu-1}\,|\,\xi_\nu\,) \times \phi(\xi_\nu\,), \tag{6.3.12}$$

where

$$\phi(\xi_\nu\,) = \int\cdots\int \phi(\,\xi_1,\xi_2,\ldots,\xi_\nu\,)\,\mathrm{d}\xi_1\cdots\mathrm{d}\xi_{\nu-1} \tag{6.3.13}$$

is a marginal joint probability. In the discussion of Section 6.3.1 for $\nu=2$, this marginal probability was denoted as $\phi_2(\xi_2)$. The conditional probability $\phi(\,\xi_1,\ldots,\xi_{\nu-1}\,|\,\xi_\nu\,)$ satisfies the normalization condition

$$\int\cdots\int \phi(\,\xi_1,\ldots,\xi_{\nu-1}\,|\,\xi_\nu\,)\,\mathrm{d}\xi_1\cdots\mathrm{d}\xi_{\nu-1} = 1. \tag{6.3.14}$$

Combining (6.3.5) with (6.3.12) we obtain a Bayes rule expressed by

$$\phi(\,\xi_1, \ldots, \xi_{\nu-1} \,|\, \xi_\nu\,) = \frac{\phi(\,\xi_\nu \,|\, \xi_1, \ldots, \xi_{\nu-1}\,)}{\phi(\xi_\nu\,)} \times \phi(\xi_1, \ldots, \xi_{\nu-1}\,). \tag{6.3.15}$$

In applications, $\xi_\nu$ constitutes the data, $\phi(\,\xi_1, \ldots, \xi_{\nu-1} \,|\, \xi_\nu\,)$ on the left-hand side is a posterior joint probability, and $\phi(\xi_1, \ldots, \xi_{\nu-1}\,)$ on the right-hand side is a prior joint probability. The numerator on the right-hand side is the likelihood of the data.

For $\nu = 2$, we recover the Bayes equation shown in (6.1.19) repeated below,

$$\phi(\xi_1 \,|\, \xi_2) = \frac{\phi(\xi_2 \,|\, \xi_1)}{\phi_2(\xi_2)} \times \phi_1(\xi_1), \tag{6.3.16}$$

subject to straightforward changes in notation.

Rearranging (6.3.15), we obtain an alternative form of Bayes rule in terms of relative decreases,

$$\frac{\phi(\xi_1, \ldots, \xi_{\nu-1}\,) - \phi(\,\xi_1, \ldots, \xi_{\nu-1} \,|\, \xi_\nu\,)}{\phi(\xi_1, \ldots, \xi_{\nu-1}\,)} = \frac{\phi(\xi_\nu\,) - \phi(\,\xi_\nu \,|\, \xi_1, \ldots, \xi_{\nu-1}\,)}{\phi(\xi_\nu\,)}. \tag{6.3.17}$$

### *6.3.5 Generalized Bayes rule*

Consider two non-overlapping blocks of $\boldsymbol{\xi}$, $\boldsymbol{\xi}_I$ and $\boldsymbol{\xi}_{II}$, whose union reconstructs $\boldsymbol{\xi}$. The preceding derivation can be generalized to

$$\phi(\,\boldsymbol{\xi}_I \,|\, \boldsymbol{\xi}_{II}) = \frac{\phi(\,\boldsymbol{\xi}_{II} \,|\, \boldsymbol{\xi}_I)}{\phi(\boldsymbol{\xi}_{II})} \times \phi(\boldsymbol{\xi}_I), \tag{6.3.18}$$

where

$$\phi(\boldsymbol{\xi}_{II}) = \int \phi(\,\boldsymbol{\xi}_{II} \,|\, \boldsymbol{\xi}_I')\, \phi(\boldsymbol{\xi}_I')\, \mathrm{d}V(\boldsymbol{\xi}_I') \tag{6.3.19}$$

is a marginal probability.

Rearranging (6.3.18), we obtain an alternative form of Bayes rule in terms of relative decreases,

$$\frac{\phi(\boldsymbol{\xi}_I) - \phi(\,\boldsymbol{\xi}_I \,|\, \boldsymbol{\xi}_{II})}{\phi(\boldsymbol{\xi}_I)} = \frac{\phi(\boldsymbol{\xi}_{II}) - \phi(\,\boldsymbol{\xi}_{II} \,|\, \boldsymbol{\xi}_I)}{\phi(\boldsymbol{\xi}_{II})}. \tag{6.3.20}$$

### 6.3.6 Hierarchical approximation

With reference to (6.3.11), the hierarchical approximation amounts to setting

$$\phi(\,\xi_1, \xi_2, \ldots, \xi_\nu\,) \simeq \phi(\xi_1) \times \phi(\,\xi_2 \,|\, \xi_1) \cdots \times \phi(\,\xi_\nu \,|\, \xi_{\nu-1}), \tag{6.3.21}$$

involving the nearest neighbor conditional probabilities. In consolidated notation,

$$\phi(\,\xi_1, \xi_2, \ldots, \xi_\nu\,) \simeq \phi(\xi_1) \times \prod_{m=1}^{\nu-1} \phi(\,\xi_{m+1} \,|\, \xi_m\,). \tag{6.3.22}$$

Hierarchical ordering amounts to arranging the individual scalar variables, $\xi_i$, so as to minimize the error involved in the hierarchical approximation. Some examples were discussed in Section 3.8.

#### Exercise

**6.3.1** Express the normalization condition (6.3.4) in spherical polar coordinates in the $\xi_1\xi_2\xi_3$ space.

## 6.4 Bayes equation for a joint pdf

Suppose that a random outcome is determined by an arbitrary number of $n$ random parameters encapsulated in an array,

$$\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_n). \tag{6.4.1}$$

Bayes' equation for the corresponding joint pdf takes the form

$$\phi(\boldsymbol{\theta}\,|\,\mathbf{x}) = \frac{\mathcal{L}_{\mathbf{x}}(\boldsymbol{\theta})}{\int_{\mathcal{D}} \mathcal{L}_{\mathbf{x}}(\boldsymbol{\theta}')\,\phi(\boldsymbol{\theta}')\,\mathrm{d}V(\boldsymbol{\theta}')} \times \phi(\boldsymbol{\theta}), \qquad (6.4.2)$$

where the array $\mathbf{x}$ encapsulates observed or measured, discrete or continuously varying data, $\mathrm{d}V$ is a differential volume in the $\boldsymbol{\theta}$ space, and $\boldsymbol{\theta}'$ is a multicomponent integration variable, subject to the following definitions:

- $\mathcal{L}_{\mathbf{x}}(\boldsymbol{\theta})$ is a likelihood function for the data involving $\boldsymbol{\theta}$.
- $\phi(\boldsymbol{\theta})$ is a prior joint probability density function.
- $\phi(\boldsymbol{\theta}\,|\,\mathbf{x})$ the posterior joint probability density function.
- The denominator of the fraction on the right-hand side of (6.4.2) is a marginal probability or partition function.

Equation (6.4.2) follows directly from (6.3.18). subject to the substitutions:

$$\boldsymbol{\xi}_I \to \boldsymbol{\theta}, \qquad \boldsymbol{\xi}_{II} \to \mathbf{x}, \qquad \phi(\,\boldsymbol{\xi}_{II}\,|\,\boldsymbol{\xi}_I) \to \mathcal{L}_{\mathbf{x}}(\boldsymbol{\theta}). \qquad (6.4.3)$$

Normalization requires that

$$\int \mathcal{L}_{\mathbf{x}}(\boldsymbol{\theta})\,\mathrm{d}V(\mathbf{x}) = 1. \qquad (6.4.4)$$

In the case of discrete data, the integral is replaced by a sum.

Bayes equation (6.4.2) provides us with the posterior (revised) joint probability density function of $\boldsymbol{\theta}$, $\phi(\boldsymbol{\theta}\,|\,\mathbf{x})$, for observed $\mathbf{x}$ and assumed prior $\phi(\boldsymbol{\theta})$. The size of the data array, $\mathbf{x}$, is unrelated to that of $\boldsymbol{\theta}$. For example, $\mathbf{x}$, can be a single number, independent of $n$.

### *6.4.1 Two parameters*

In the case of two parameters, $n = 2$, we obtain Bayes' equation

$$\phi(\theta_1, \theta_2|\mathbf{x}) = \frac{\mathcal{L}_{\mathbf{x}}(\theta_1, \theta_2)}{\varphi(\mathbf{x})} \times \phi(\theta_1, \theta_2), \qquad (6.4.5)$$

where

$$\varphi(\mathbf{x}) = \iint_{\mathcal{D}} \mathcal{L}_{\mathbf{x}}(\theta'_1, \theta'_2)\, \phi(\theta'_1, \theta'_2)\, \mathrm{d}\theta'_1\, \mathrm{d}\theta'_2. \tag{6.4.6}$$

### 6.4.2 Posterior from uniform

Assume that the integration volume in the $\boldsymbol{\theta}$ space is equal to $V_{\boldsymbol{\theta}}$. In the absence of bias, we set $\phi(\boldsymbol{\theta}) = 1/V_{\boldsymbol{\theta}}$ and obtain

$$\phi(\boldsymbol{\theta}|\mathbf{x})_{\text{from uniform}} = \frac{\mathcal{L}_{\mathbf{x}}(\boldsymbol{\theta})}{\displaystyle\int_{\mathcal{D}} \mathcal{L}_{\mathbf{x}}(\boldsymbol{\theta}')\, \mathrm{d}V(\boldsymbol{\theta}')}, \tag{6.4.7}$$

which shows that the posterior pdf is the scaled likelihood function determined by the data.

### 6.4.3 Posterior from a spike

If $\phi(\boldsymbol{\theta})$ is a delta function in $n$-dimensional space, $\phi(\boldsymbol{\theta}|\mathbf{x})$ will also be a delta function. This means that no amount of evidence will deter us from making a deterministic assessment.

### 6.4.4 Monte Carlo integration

With regard to the integral defined in (6.4.6), now we consider the definite integral of the product of a sufficiently regular function of two random variables, $\psi(\theta_1, \theta_2)$, and a joint probability density function (pdf) in two dimensions, $\phi(\theta_1, \theta_2)$, over in a certain domain $\mathcal{D}$ in the $\theta_1\theta_2$ plane,

$$\mathcal{J} \equiv \iint_{\mathcal{D}} \psi(\theta_1, \theta_2)\, \phi(\theta_1, \theta_2)\, \mathrm{d}\theta_1\, \mathrm{d}\theta_2. \tag{6.4.8}$$

Using the *law of large numbers*, we find that the integral can be approximated with the mean value of the function $\psi(\theta_1, \theta_2)$ at $N$ random doublets

$$\boldsymbol{\theta}^{(i)} = (\theta_1^{(i)}, \theta_2^{(i)}), \tag{6.4.9}$$

sampled according to the joint pdf, $\phi(\theta_1, \theta_2)$, that is,

$$\mathcal{J} \simeq \frac{1}{N} \sum_{i=1}^{N} \psi(\boldsymbol{\theta}^{(i)}).\mathcal{J} \simeq \frac{1}{N} \sum_{i=1}^{N} \psi(\boldsymbol{\theta}^{(i)}). \tag{6.4.10}$$

The accuracy improves as a higher number of sampling doublets are employed. The random doublets can be obtained using the Gibbs sampling method.

### *6.4.5 Gibbs sampling in two dimensions*

In the Gibbs sampling method, a starting doublet is chosen, $\boldsymbol{\theta}^{(0)}$, and a chain is constructed as follows:

1. Sample $\theta_1^{(1)}$ from the conditional pdf $\phi(\theta_1 \,|\, \theta_2^{(0)})$.
2. Sample $\theta_2^{(1)}$ from the conditional pdf $\phi(\theta_2 \,|\, \theta_1^{(1)})$.
3. Set $\boldsymbol{\theta}^{(1)} = (\theta_1^{(1)}, \theta_2^{(1)})$.
4. Repeat.

As an example, we assume that the domain of definition, $\mathcal{D}$, is a rectangle in the $\xi_1\xi_2$ plane confined inside $a_1 \le \xi_1 \le b_1$ and $a_2 \le \xi_2 \le b_2$, as discussed in Section 6.2. The joint pdf can be approximated with a bilinear function,

$$\phi(\xi_1, \xi_2) = c_0 + c_1\,\xi_1 + c_2\,\xi_2, \tag{6.4.11}$$

where $c_0, c_1, c_2$ are three constants. An illustration of this joint pdf is shown in Figure 6.4.1(*a*). for $a_1 = 0$, $b_1 = 0.5$, $a_2 = 0$, $b_2 = 1.0$, $c_1 = 1.0$, and $c_2 = 2.0$. Gibbs sampling is implemented in the following Matlab code:

```
c0 = 1.0/((b1-a1)*(b2-a2)) - 0.5*(a1+b1)*c1 - 0.5*(a2+b2)*c2;
mu1 = c1*(b2-a2);
mu2 = c2*(b1-a1);

%---
% initialize
%---
```

(*a*) (*b*)

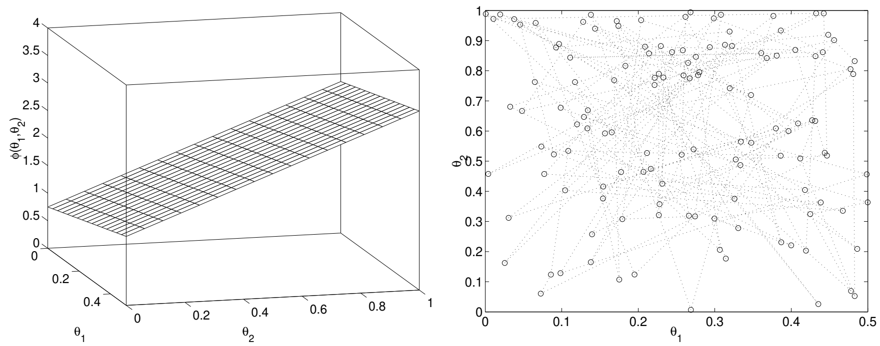


FIGURE 6.4.1 (*a*) plot of a linear joint pdf and (*b*) an associated Gibbs sampling set. for a bilinear joint pdf with $c_1 = 1.0$ and $c_2 = 2.0$.

```
theta1 = 1.0;
theta2 = 1.0;
theta1 = a1+(b1-a1)*rand;
theta2 = a2+(b2-a2)*rand;

for i=1:N
 alpha = c1/(mu2*theta2+eta2);
 theta1 = randlin(a1,b1,alpha,rand);
 alpha = c2/(mu1*theta1+eta1);
 theta2 = randlin(a2,b2,alpha,rand);
% now use sampling point: theta1, theta2
end
```

The function *randlin* generating random data with a linear pdf is listed in Section 4.5. A graph of a Gibbs sampling set is shown in Figure 6.4.1(*b*).

If an event of interest is influenced *independently* by the parameters, $\theta_1$ and $\theta_2$, then

$$\phi(\theta_1 \,|\, \theta_2) = \phi_1(\theta_1), \qquad \phi(\theta_2 \,|\, \theta_1) = \phi_2(\theta_2). \qquad (6.4.12)$$

In that case, the method amounts to generating independent random samples for $\theta_1$ and $\theta_2$ with corresponding marginal pdfs.

### *6.4.6 MCMC*

The most efficient and sophisticated method of carrying out the Bayesian analysis is encapsulated in the Markov Chain Monte Carlo sampling algorithm (MCMC), as discussed in serious books and monographs of Bayesian statistics.

#### *Exercise*

**6.4.1** Express the integral in (6.4.6) in plane polar coordinates in the $\theta_1'\theta_2'$ plane.

## *6.5 Open window with a shade*

Consider the woodchucks throwing strawberries at a partially open sliding window, as discussed in Section 5.4. For reasons that will become apparent in hindsight, we rename the open-window fraction $\theta$ as $\theta_1$ and equip the window with a vertical shade that can be pulled down to an unknown location, leaving a $\theta_2 h$ opening and an obstructed area of length $(1-\theta_2)h$ behind the sliding sash, where $h$ is the window height. as shown in Figure 6.5.1.

It is appropriate to introduce the random vectorial parameter

$$\boldsymbol{\theta} = (\,\theta_1, \theta_2\,) \tag{6.5.1}$$

pertinent to the window. The doublet $[\theta_1, \theta_2]$ takes values in a parametric square, $[0,1] \times [0,1]$.

### *6.5.1 Woodchuck Bilbo*

The woodchucks have picked up a fresh batch of $n$ strawberries and toss them against the window. Woodchuck Bilbo is having fun by observing that $m$ out of $n$ strawberries go through the open area of the window, while the remaining $n-m$ strawberries splash on the shade or sliding sash.

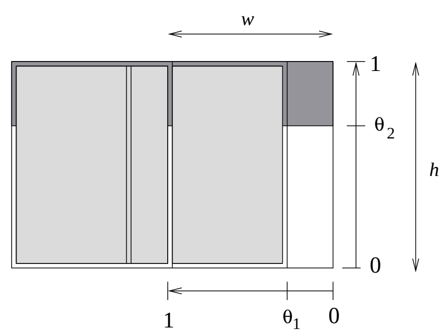



FIGURE 6.5.1 Illustration of a horizontal sliding window that has been opened partially to a location determined by the parameter $\theta_1$ and a shade that has been pulled down to a location determined by the parameter $\theta_2$.

### 6.5.2 Likelihood function

Our objective is to estimate the probability density function of the doublet, $(\theta_1, \theta_2)$, from a strawberry throwing session using Bayes' rule. The likelihood function is the conditional probability that $m$ out of $n$ strawberries enter the room for given values of $\theta_1$ and $\theta_2$, given by

$$\mathcal{L}_m(\boldsymbol{\theta}) = \sum_{k=m}^{n} \mathcal{B}_k^n(\theta_1)\, \mathcal{B}_m^k(\theta_2). \tag{6.5.2}$$

The first binomial term, $\mathcal{B}_k^n(\theta_1)$, represents the probability that $k \leq n$ out of $n$ strawberries have gone through the window opening. The second binomial term, $\mathcal{B}_m^k(\theta_1)$, represents the probability that $m \leq k$ out of these $k$ strawberries have escaped the shade.

If the window is completely open, $\theta_1 = 1$, we find that $\mathcal{B}_k^n(1) = 0$ for any $k$, except that $\mathcal{B}_n^n(1) = 1$; consequently, $\mathcal{L}_m(\boldsymbol{\theta}) = \mathcal{B}_m^n(\theta_2)$.

### 6.5.3 Some math

Substituting the expression for the binomial distribution, we obtain

$$\mathcal{L}_m(\boldsymbol{\theta}) = \frac{n!}{m!}\,\theta_2^m \sum_{k=m}^{n} \frac{1}{(n-k)!\,(k-m)!}\,\theta_1^k(1-\theta_1)^{n-k}(1-\theta_2)^{k-m}. \tag{6.5.3}$$

Setting $p = k - m$ and thus $k = p + m$, and defining $\ell = n - m$, we obtain

$$\mathcal{L}_m(\boldsymbol{\theta}) = \frac{n!}{m!\,\ell!}\,(\theta_1\theta_2)^m \sum_{p=0}^{\ell} \binom{\ell}{p} \big(\theta_1\,(1-\theta_2)\,\big)^p(1-\theta_1)^{\ell-p}. \tag{6.5.4}$$

Now we referring to the expansion (A.27), Appendix A, repeated below for convenience,

$$(a+b)^n = \sum_{m=0}^{n} \binom{n}{m}\; a^{n-m}\, b^m, \tag{6.5.5}$$

we set $a = \theta_1(1-\theta_2)$ and $b = 1 - \theta_1$, to find that the sum in (6.5.4) is equal to $(1-\theta_1\theta_2)^\ell$. Consequently, the likelihood function is given by the binomial distribution

$$\mathcal{L}_m(\boldsymbol{\theta}) = \mathcal{B}_m^n(\theta_1\theta_2). \tag{6.5.6}$$

A graph of this likelihood function for $n = 10$ and $m = 5$ is shown in Figure 6.5.1.

### 6.5.4 Stating the obvious

In retrospect, since the fraction of the open area of the window-shade system is $\theta_1\theta_2$, the final results shown in (6.5.6) is expected. The dependence of the likelihood function on the product $\theta_1\theta_2$ suggests that the Bayesian analysis will provide us with the pdf of the product $\theta_1\theta_2$ but cannot differentiate between pairs $(\theta_1, \theta_2)$ that have the same product. Based on this observation, it makes sense to chose the prior pdf that is a function of the product $\theta_1\theta_2$. The uniform pdf is a reasonable choice.

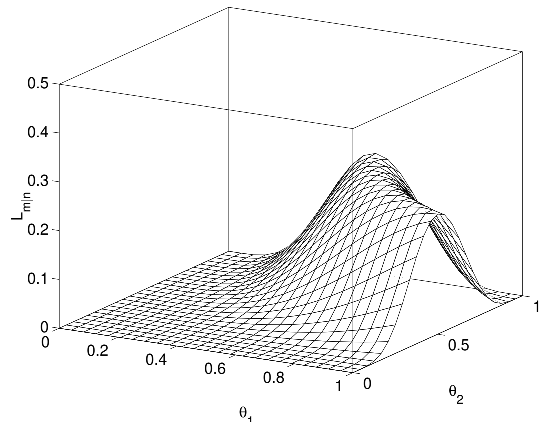



FIGURE 6.5.1 Graph of the likelihood function for $n = 10$ strawberries thrown by woodchucks against a window with a shade with $m = 5$ successful landings.

***Exercise***

**6.5.1** The shade is made by a screen that allows the small strawberries to enter the room. What is the corresponding Bayes equation for the pdf $\phi(\theta_1, \theta_2)$?

## *6.6 Habanero peppers and crushed pepper*

Deans of colleges and universities love research and creative activity, but not just any research and creative activity. They love *sponsored* research and creative activity. This means that they urge their faculty to obtain grants preceding their intellectual pursuits and irrespective of probability of success, if any.

### *6.6.1 Cook and Look*

Distinguished Professors I. Cook and U. Look have secured a grant to study the amount of water consumption after eating $10$ fl oz (fluid onces) of hot salsa as a function of the amount of chopped habanero peppers and crushed red pepper in the ingredients.

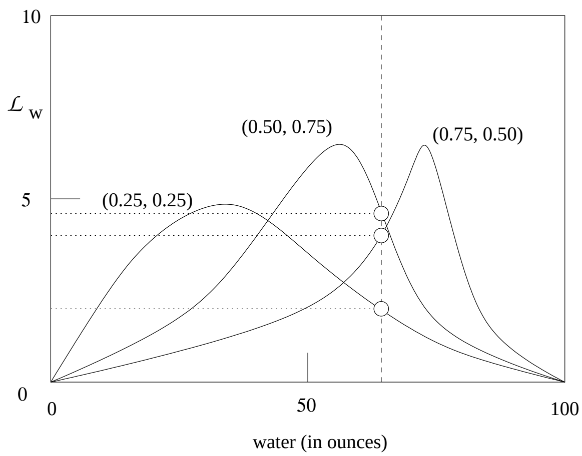



FIGURE 6.6.1 Probability density function of water consumption after eating $10$ fl oz of hot salsa, as a function of the amount of chopped chile peppers, $\theta_1$, and crushed red pepper, $\theta_2$, added to $3$ cups of salsa, shown for three pairs, $(\theta_1, \theta_2)$.

The results of this worthy study were published in the form of probability density functions,

$$\mathcal{L}_w(\theta_1, \theta_2), \tag{6.6.1}$$

where $w$ is the volume of water consumed measured in fl oz, $\theta_1$ is the amount of chopped habanero peppers measured in cups, and $\theta_2$ is the amount of crushed red pepper also measured in cups added to $3$ cups of salsa, as shown in Figure 6.6.1. One cup is equal to $8$ fl oz.

For example, for $\frac{1}{4}$ cup of chopped habanero peppers and $\frac{1}{4}$ cup of crushed red pepper, the expected amount of water consumption after consuming $10$ fl oz ($1.25$ cups) oz of salsa is roughly $35$ fl oz ($4.375$ cups). Normalization requires that

$$\int_0^\infty \mathcal{L}_w(\theta_1, \theta_2)\,\mathrm{d}w = 1 \tag{6.6.2}$$

for any pair, $(\theta_1, \theta_2)$, that is, the area under each curve in Figure 6.6.1 is unity.

The distinguished professors published three such curves for three combinations of $\theta_1$ and $\theta_2$ in a prestigious journal of food science with open access at taxpayers' expense, as shown in Figure 6.6.1. Dividing the amount of the grant award with the number of curves, $3$, we find that each curve is worth about $\$100$ K.

A key observation is that a certain amount of water is consumed for different combinations of $\theta_1$ and $\theta_2$, albeit with different probabilities. Thus, the amount of water consumed does not uniquely define $\theta_1$ and $\theta_2$, though it provides us with possible pairs.

### 6.6.2 Reverse engineering the recipe

Entrepreneur O. B. Late wants to open a sports bar that will serve hot salsa and blue tortilla chips along with pink lemonade. Instead of developing his own salsa recipe, the entrepreneur buys a number of different salsa jars from the local grocery store, tastes them all, and chooses the one that he likes the most. Since he is not sure how much chopped habanero peppers and crushed red pepper was used, he wants to reverse engineer the recipe.

Because he suffers from stress-induced ulcer, O. B. Late asks his partner, P. Rolate, if she could eat an entire ten-ounce and report how much water she had to consume to quench her mouth to a tolerable state. The partner agreed and reported that she consumed $65$ ounces of water.

### 6.6.3 The prior

O. B. Late reckons that about half a cup of chopped habaneros peppers and one quarter of a cup of crushed red pepper were probably used in the recipe, and adopts the prior pdf

$$\phi(\theta_1, \theta_2) = 72\, \theta_1 (1 - \theta_1)\, \theta_2\, (1 - \theta_2)^2 \tag{6.6.3}$$

for $0 \le \theta_1, \theta_2 \le 1$. Since this prior pdf is the product of a function of $\theta_1$ and other function of $\theta_2$, adding the two ingredients is done with an independent frame of mind. A three-dimensional plot and a contour plot of this prior pdf is shown in Figure 6.6.2. The mandatory

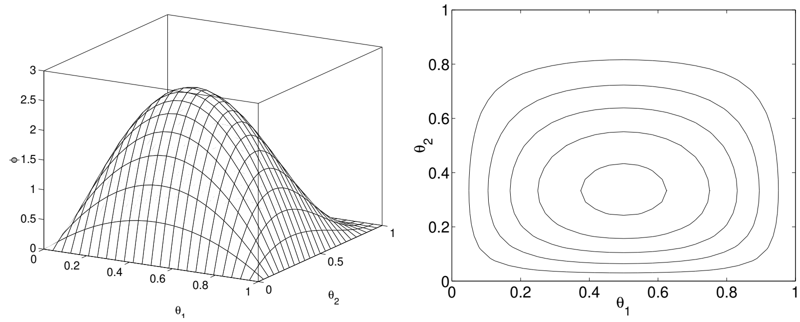



FIGURE 6.6.2 Three-dimensional and contour plot of the prior probability density function for two salsa ingredients.

normalization condition,

$$\int_0^1 \int_0^1 \phi(\theta_1, \theta_2)\, \mathrm{d}\theta_1\, \mathrm{d}\theta_2 = 1, \tag{6.6.4}$$

is satisfied. The prior is zero along the four edges of the $(\theta_1, \theta_2)$ square, which conveys the belief that no chopped habaneros peppers or crushed red pepper was used alone, and no way one cup of each was added to the hot salsa.

### *6.6.4 Bayes' equation*

Partner Late uses Bayes' equation to define the posterior pdf

$$\phi(\theta_1, \theta_2 | w = 65) = \frac{\mathcal{L}_{w=65}(\theta_1, \theta_2)}{\varphi_{(w=65)}} \times \phi(\theta_1, \theta_2), \tag{6.6.5}$$

where

$$\varphi_{(w=65)} = \int_0^1 \int_0^1 \mathcal{L}_{w=65}(\theta_1', \theta_2')\, \phi(\theta_1', \theta_2')\, \mathrm{d}\theta_1'\, \mathrm{d}\theta_2' \tag{6.6.6}$$

is a marginal probability and $\theta_1'$, $\theta_2'$ are integration variables.

The likelihood function, $\mathcal{L}_{w=65}(\theta_1, \theta_2)$, can be evaluated at particular pairs $(\theta_1, \theta_2)$ as the intersection of the vertical dashed line drawn at $w = 65$ and each of the three curves, indicated by the circles, in

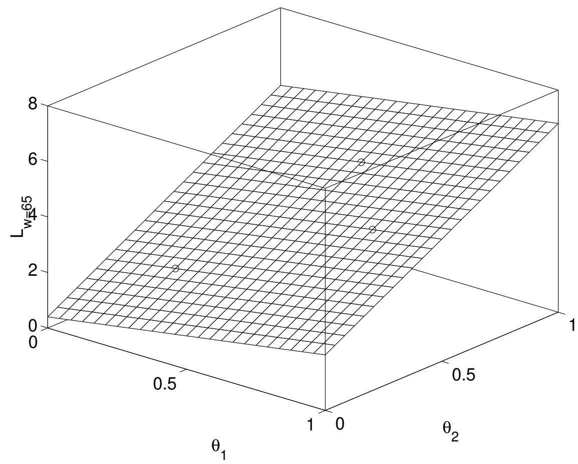



FIGURE 6.6.3 Likelihood function for two hot-salsa ingredients corresponding to 65 ounces of water consumed after each ten ounces of hot salsa.

Figure 6.6.1, compliments of the distinguished professors. Only three data points are available, described by the $(\theta_1, \theta_2, \mathcal{L}_{w=65})$ triplets

$$(0.25, 0.25, 2.0), \quad (0.50, 0.75, 4.8), \quad (0.75, 0.50, 4.0). \tag{6.6.7}$$

The graph of the likelihood function of interest can be approximated with a plane interpolating through these three data. A back-of-the-envelop calculation yields the linear function

$$\mathcal{L}_{w=65}(\theta_1, \theta_2) = 1.6\,\theta_1 + 4.8\,\theta_2 + 0.4, \tag{6.6.8}$$

as shown in Figure 6.6.3. The three data are indicated by the circular symbols.

According to Bayes' formula, the posterior pdf is

$$\phi(\theta_1, \theta_2 \,|\, w = 65) \tag{6.6.9}$$
$$= \frac{1.6\,\theta_1 + 4.8\,\theta_2 + 0.4}{\varphi_{(w=65)}} \times 72\,\theta_1\,(1-\theta_1)\,\theta_2\,(1-\theta_2)^2,$$

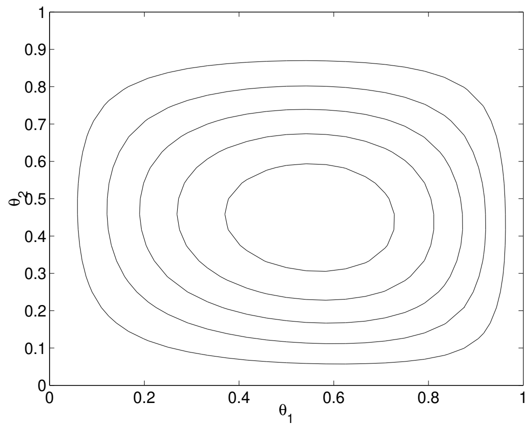



FIGURE 6.6.4 Contour plot of the posterior probability density function for two salsa ingredients.

where

$$\varphi(w{=}65) = 72 \int_0^1 \int_0^1 (1.6\,\theta_1' + 4.8\,\theta_2' + 0.4)\,\theta_1'\,(1-\theta_1') \times \theta_2'\,(1-\theta_2')^2\,\mathrm{d}\theta_1'\,\mathrm{d}\theta_2' \tag{6.6.10}$$

to ensure proper normalization. A graph of the posterior pdf is shown in Figure 6.6.4 as a contour plot. The results suggest that the maximum is achieved when $0.55$ cup of chopped habaneros peppers and $0.45$ cup of crushed red pepper are added to the salsa.

### *6.6.5 Getting a second opinion*

Rolate also asks his first cousin twice-removed to eat an entire ten-ounce jar of salsa and report how much water she had to consume afterwards.

The first cousin reported that she consumed $72$ ounces of water. Since this is close enough to $65$, partner Late decides to go ahead with $0.55$ cup of chopped habaneros peppers and $0.45$ cup of crushed red pepper in each batch of $20$ ounces of hot salsa.

### 6.6.6 Final act

The hot salsa turned out to be delicious and the sports bar was a huge success. The distinguished professors heard that the results of their sponsored research were used by the entrepreneurs in a successful venture and demanded free chips and salsa for life on Friday nights when they meet to gossip about the students, complain about the university administration, and malign their colleagues. The entrepreneurs kindly agreed.

### 6.6.7 Petty and small-minded

The professors learned from server Tara the details of the calculations the entrepreneurs performed to determine their recipe, and offered criticism and objections.

Professor I. Cook argued that, if a uniform initial pdf were used, the posterior pdf would be the likelihood function shown in Figure 6.6.3 whose maximum occurs at full pepper. Professor U. Look argued that approximating the likelihood function with a linear function is nonsensical.

The entrepreneurs pretended to be interested in these concerns, while rolling their eyes and winking at each other, and then moved to the back room for a couple of drinks and a game of pool (billiards.)

#### Exercise

**6.6.1** Repeat the Bayesian analysis with a fourth curve of your choice in Figure 6.6.1 using a bilinear approximation for the likelihood function,

$$\mathcal{L}_{w=65}(\theta_1, \theta_2) = (a\,\theta_1 + b) \times (c\,\theta_2 + d), \tag{6.6.11}$$

where $a$, $b$, $c$, and $d$ are suitable coefficients.

## 6.7 The frog and the red herring

A magnificent bullfrog named Hoppy is sitting on a lily pond minding his own business, as shown in Figure 6.7.1. Suddenly, his paws sense

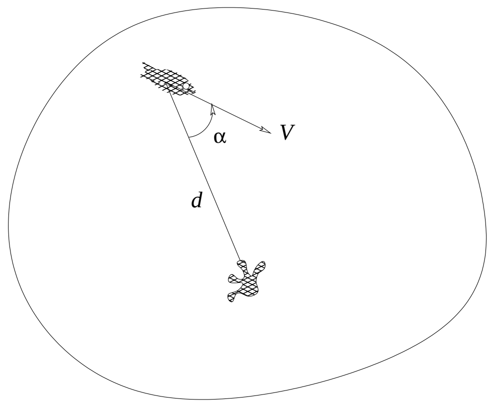



FIGURE 6.7.1 A bullfrog with yellow shades named Hoppy is sitting on a lily pond when senses that his friend Funny Fins is swimming with velocity $V$ at a distance $d$ and angle $\alpha$.

water motion by way of shear stress mediated by mechanotransduction. Surely his friend Funny Fins, a red herring, is coming to see him.

For convenience, Hoppy assumes an infinite pond. The intensity of the shear stress sensed by Hoppy's paws, $s$, depends on three parameters:

- His distance from Funny Fins, $d$, varying from $0$ to $\infty$.
- Funny Fins' direction of swimming determined by the angle $\alpha$ varying between $0$ and $2\pi$.
- Funny Fins' speed of swimming, $V$, varying between $0$ and $\infty$.

Hoppy will use Bayes' theorem to estimate the probability that Funny Fins is swimming toward him.

### *6.7.1 Likelihood function*

Because Hoppy's paws are randomly muddy, and also because some annoying tadpoles jump erratically nearby, the intensity of the shear stress, $s$, for given, $d$, $\alpha$, and $V$ varies randomly with a joint probability distribution function

$$\mathcal{L}_s(d, \alpha, V), \tag{6.7.1}$$

which plays the role of a likelihood function for sensed paw shear stress, $s$. Normalization requires that

$$\int_0^\infty \mathcal{L}_s(d, \alpha, V)\, \mathrm{d}s = 1 \tag{6.7.2}$$

for any triplet, $(d, \alpha, V)$, that is, Hoppy must experience some shear stress.

### 6.7.2 Is Funny Fins coming and how far is he?

The full joint probability density function determining Funny Fin's behavior is denoted by $\phi(d, \alpha, V)$. Since the pdf depends on three parameters, it can be visualized only in terms of slice graphs.

Hoppy may be interested in the probability density function of the angle, $\alpha$, denoted by $\phi_\alpha(\alpha)$, in the hope that it peaks at $\alpha = 0$ so that his friend is swimming toward him. Alternatively, Hoppy may be interested in the probability density function of Funny Fins' distance, $d$, denoted by $\phi_d(d)$. Finally, Hoppy may be interested in the probability density function of Funny Fins' speed of swimming, $V$, denoted by $\phi_V(V)$.

Because the pdf or interest depends on three parameters, $d$, $\alpha$, and $V$, Hoppy must consider the full joint probability, $\phi(d, \alpha, V)$, and compute

$$\phi_\alpha(\alpha) = \int_0^\infty \Big( \int_0^\infty \phi(d, \alpha, V)\, \mathrm{d}d \Big) \mathrm{d}V \tag{6.7.3}$$

for the swimming direction,

$$\phi_d(d) = \int_0^{2\pi} \Big( \int_0^\infty \phi(d, \alpha, V)\, \mathrm{d}V \Big) \mathrm{d}\alpha \tag{6.7.4}$$

for the distance, and

$$\phi_V(V) = \int_0^{2\pi} \Big( \int_0^\infty \phi(d, \alpha, V)\, \mathrm{d}d \Big) \mathrm{d}\alpha \tag{6.7.5}$$

for the velocity. Computing the double integrals is easier said that done.

### 6.7.3 Bayes equation

Hoppy uses Bayes' equation to write

$$\phi(d, \alpha, V \mid s) = \frac{\mathcal{L}_s(d, \alpha, V)}{\varphi_s} \times \phi(d, \alpha, V), \tag{6.7.6}$$

where

$$\varphi_s \equiv \int_0^\infty \int_0^{2\pi} \int_0^\infty \mathcal{L}_s(d', \alpha', V') \times \phi(d', \alpha', V') \, \mathrm{d}d' \, \mathrm{d}\alpha' \, \mathrm{d}V' \tag{6.7.7}$$

is a marginal probability.

### 6.7.4 Prior pdf

Hoppy has every reason to believe that, in the absence of information, the variables $d$, $\alpha$, and $V$ can be assigned independently in the prior joint probability distribution, and sets

$$\phi(d, \alpha, V) = \phi_d(d) \times \phi_\alpha(\alpha) \times \phi_V(V). \tag{6.7.8}$$

Moreover, Hoppy recognizes that Funny Fins usually hangs around at $d_0 = 20 \text{ ft}$ and swims with velocity $V_0 = 0,75 \text{ ft/sec}$, and sets

$$\phi_d(d) = \mathcal{N}_{d_0, \frac{1}{10} \times d_0,}(d), \qquad \phi_\alpha(\alpha) = \frac{1}{2\pi}, \\ \phi_V(V) = \mathcal{N}_{V_0, \frac{1}{10} \times V_0,}(V), \tag{6.7.9}$$

where $\mathcal{N}_{\overline{\theta}, \sigma}(\theta)$ is the normal distribution with mean $\overline{\theta}$ and standard deviation $\sigma$. In the prior, the standard deviation has been set by Hoppy arbitrarily to one tenth of the respective mean. The prior distribution of $\alpha$ is a non-informative flat.

### 6.7.5 Clean paws

If Hoppy's paws were always clean, and if the annoying tadpoles were taking a break, he would sense only one specific value of the flow-induced shear stress for each triplet $(d,\ \alpha,\ V)$, denoted by

$$s_{\text{clean}}(d, \alpha, V). \tag{6.7.10}$$

To evaluate this ideal shear stress, Hoppy introduces the water viscosity, $\mu$, assumes that the shear stress if proportional to Funny Fin's velocity, $V$, and sets

$$s_{\text{clean}} = \beta \frac{\mu V}{a} U(d, \alpha), \tag{6.7.11}$$

where $\beta$ is a dimensionless coefficient, $a$ is Funny Fin's head-to-tail distance, and $U(d, \alpha)$ is a dimensionless function of $d$ and $\alpha$.

Hoppy remembers with nostalgia his Professor of Fluid Mechanics who had mentioned in an 8:00 am lecture that the fluid flow induced by Funny Fins can be modeled in terms of a mathematical singularity called the point-source dipole. The induced streamline pattern is shown in Figure 6.7.2. Using the mathematical expression for the point-source dipole, Hoppy finds that

$$U(d, \alpha) = \frac{a^2}{d^2} \tag{6.7.12}$$

independent of $\alpha$, which is convenient. Using this expression, Hoppy finds that

$$s_{\text{clean}} = \zeta \frac{V}{V_0} \frac{a^2}{d^2}. \tag{6.7.13}$$

where $\zeta = \beta \mu V_0 / a$ is a known constant. The two fractions on the right-hand side of (6.7.13) are dimensionless.

### 6.7.6 Muddy paws

In the absence of randomness due to muddy paws and tadpoles, the likelihood function is described by Dirac's delta function, $\delta(w)$, forced at clean paws,

$$\mathcal{L}_s(d, \alpha, V) = \delta\big(s - s_{\text{clean}}(d, \alpha, V)\big). \tag{6.7.14}$$

The delta function representing a localized impulse is typical of a deterministic response.

To account for muddy paws and annoying tadpoles, Hoppy assumes that the shear stress, $s$, varies between $0$ (very muddy paws) and $s_{\text{clean}}$

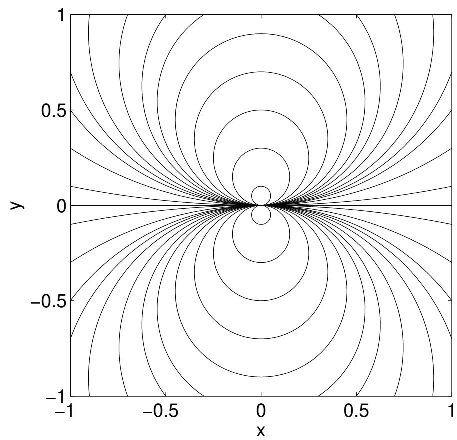



FIGURE 6.7.2 Streamline pattern induced by Funny Fin's swimming, where the $x$ axis is in the direction of swimming.

(strongest sense) according to a certain probability density function, such that

$$\mathcal{L}_s(d,\alpha,V) = \begin{cases} \dfrac{1}{s_{\mathrm{clean}}(d,\alpha,V)}\,\mathrm{H}(\eta) & \text{for } s \le s_{\mathrm{clean}}(d,\alpha,V), \\ 0 & \text{for } s > s_{\mathrm{clean}}(d,\alpha,V), \end{cases} \tag{6.7.15}$$

where

$$\eta \equiv \frac{s}{s_{\mathrm{clean}}(d,\alpha,V)} \tag{6.7.16}$$

is a scaled shear stress and $\mathrm{H}(\eta)$ is a universal function defined in the interval $0 \le \eta \le 1$, satisfying

$$\int_0^1 \mathrm{H}(\eta)\,\mathrm{d}\eta = 1. \tag{6.7.17}$$

The dependence on $\alpha$ has been retained for the sake of generality. The second case in (6.7.15) expresses the fact that Hoppy cannot feel more than the ideal shear stress for a certain combination of $d$, $\alpha$, and $V$.

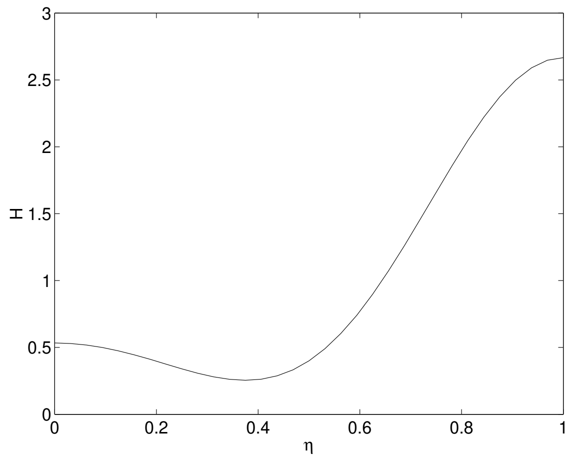



FIGURE 6.7.3 Universal probability density function (pdf) of the shear stress sensed by Hoppy's paws.

### 6.7.7 Choice of universal function

The form of the universal function $\mathrm{H}(\eta)$ depends on how often Hoppy cleans his paws and how often annoying tadpoles appear. Hoppy chooses the universal function described by

$$\mathrm{H}(\eta) = \frac{1}{c_1 + c_2} \frac{2 \times 4 \times 6}{1 \times 3 \times 5} \big( c_1 \sin^6(\tfrac{\eta}{2}\,\pi) + c_2 \sin^6(\tfrac{1-\eta}{2}\,\pi) \big) \quad (6.7.18)$$

with $c_1 = 1$ and $c_2 = 0.2$, as shown in Figure 6.7.3.

Substituting expression (6.7.13) for the clean shear stress, Hoppy finds that the likelihood function is given by

$$\mathcal{L}_s(d, V) = \begin{cases} \dfrac{1}{\zeta} \dfrac{V_0}{V} \dfrac{d^2}{a^2}\, \mathrm{H}(\dfrac{s}{\zeta} \dfrac{V_0}{V} \dfrac{d^2}{a^2}) & \text{for } s \le s_{\text{clean}}(d, \alpha, V), \\ 0 & \text{for } s > s_{\text{clean}}(d, \alpha, V). \end{cases} \quad (6.7.19)$$

A graph of this likelihood function for $s = 0.01 \times \zeta$ is shown in Figure 6.7.4(*a*).

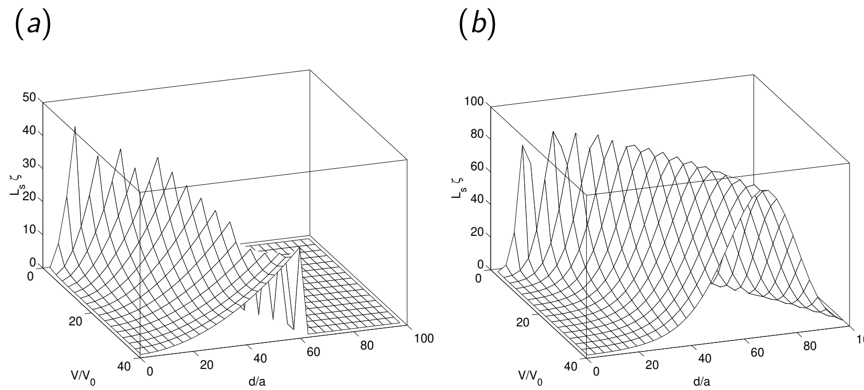



FIGURE 6.7.4 (*a*) Likelihood function associated with the function $\mathrm{H}(\eta)$ for $s = 0.01 \times \zeta$ and (*b*) corresponding likelihood function associated with the Gaussian distribution.

Bayes' equation takes the form

$$\phi(d, V \,|\, s) = \frac{\mathcal{L}_s(d, V)}{\varphi_s} \times \phi(d, V), \tag{6.7.20}$$

where

$$\varphi_s = \int_0^\infty \int_0^\infty \mathcal{L}_s(d', V')\, \phi(d', V')\, \mathrm{d}d'\, \mathrm{d}V' \tag{6.7.21}$$

is a marginal probability.

Hoppy is thinking that he can simplify things by using as likelihood function the Gaussian distribution

$$\mathcal{L}_s(d, V) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp(-\tfrac{1}{2}\,\frac{(s - s_{\mathrm{clean}})^2}{\sigma^2}) \tag{6.7.22}$$

with an arbitrarily chosen standard deviation, for example, $\sigma = \frac{1}{2}\, s_{\mathrm{clean}}$. A graph of this likelihood function for $s = 0.01 \times \zeta$ is shown in Figure 4.7.4(*b*).

Hoppy notices that $s_{\mathrm{clean}}$, and thus the likelihood function, depends on the composite variable $\xi \equiv V/d^2$ rather than on the individual variables$V$ and $d^2$, and considers introducing a probability density function that depends on the variable $\xi$. Bayes' equation becomes

$$\phi(\xi|s) = \frac{\mathcal{L}_s(\xi)}{\displaystyle\int_0^\infty \mathcal{L}_s(\xi')\,\phi(\xi')\,\mathrm{d}\xi'} \times \phi(\xi). \tag{6.7.23}$$

A Bayesian analysis based on this equation will not be able to provide information on $V$ and $d^2$ individually, but rather on their specific combination encapsulated in $\xi$.

### 6.7.8 Bffl Bilbo

As Hoppy starts pulling out his laptop to write code that computes and plots the posterior pdf, he notices that his best friend for life (bffl) Bilbo the groundhog is rolling down the hill, and hops away to greet him and ask him to recite hilarious jokes. Hoppy is thinking that there is always time for Bayesian analysis, but not enough time to spend with a dear friend, especially when the friend's name is groundhog Bilbo.

#### Exercise

**6.7.1** Show that the function $\mathrm{H}(\eta)$ given in (6.7.17) satisfies the constraint (6.7.18).

## 6.8 A spider named Mo

Consider a random variable, $\xi$, defined in a specified interval, $[a_\xi, b_\xi]$, with a specified probability density function (pdf), $\phi_\xi(\xi)$. The variable $\xi$ is related to another variable, $x$, varying in a corresponding interval, $[a_x, b_x]$, by a monotonic function

$$x = f(\xi), \tag{6.8.1}$$

where

$$a_x = f(a_\xi), \qquad b_x = f(b_\xi). \tag{6.8.2}$$

By way of this association, the variable $x$ is also a random variable with

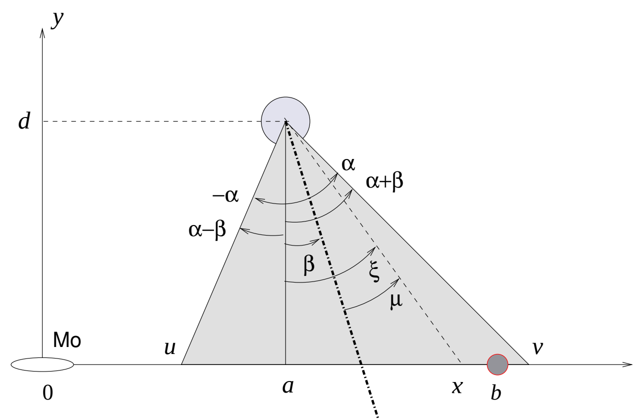



FIGURE 6.8.1 A paint sprayer with painting aperture $2\alpha$, accidentally rotated at an angle $\beta$, sprays paint onto a flat surface. The sprayer ejection mid-plane is indicated by the dashed line.

pdf $\phi_x(x)$. The two pdfs are related by

$$\pm \frac{\phi_\xi(\xi)}{\phi_x(x)} = f'(\xi) = \frac{\mathrm{d}f}{\mathrm{d}\xi}, \qquad (6.8.3)$$

where a prime denotes the first derivative, the plus sign is chosen when $f' > 0$, and the minus sign is chosen when $f' < 0$.

### *6.8.1 Sprayer head*

Now consider a paint sprayer with painting aperture $2\alpha$ spraying paint onto a flat surface, as shown in Figure 6.8.1. The sprayer head has been accidentally rotated by an angle, $\beta$. When $\beta = 0$, the spraying jet is perfectly symmetric.

### *6.8.2 An ejected droplet*

An ejected droplet may leave the sprayer head at a random angle, $\mu$, measured with respect to the sprayer head mid-plane, as shown in Figure 6.8.1. The ejection of a droplet from the sprayer head is described by an associated probability density function,

$$\phi_\mu(\mu) = \begin{cases} \Phi(\mu) & \text{for } -\alpha \le \mu \le \alpha, \\ 0 & \text{otherwise,} \end{cases} \tag{6.8.4}$$

where $\Phi(\mu)$ is a given or assumed pdf. In the case of a uniform and unclogged sprayer head, $\Phi(\mu) = 1/(2\alpha)$.

### 6.8.3 An impinging droplet

An ejected paint droplet impinges on the painted surface at an angle

$$\xi = \mu + \beta \tag{6.8.5}$$

measured with respect to the normal to the paint surface, as shown in Figure 6.8.1, in the range $-\alpha+\beta \le \xi \le \alpha+\beta$. The associated pdf is

$$\phi_\xi(\xi) = \phi_\mu(\xi-\beta) = \begin{cases} \Phi(\xi-\beta) & \text{for } -\alpha+\beta \le \xi \le \alpha+\beta, \\ 0 & \text{otherwise.} \end{cases} \tag{6.8.6}$$

The angle $\xi$ is related to the point of impingement, $x$, location of the sprayer head, $a$, and distance of the sprayer head from the surface, $d$, by the equation

$$\tan\xi = \frac{x-a}{d}, \tag{6.8.7}$$

as shown in Figure 4.8.1. Differentiating both sides with respect to $x$ and using the rules of derivative differentiation, we obtain

$$\frac{1}{\cos^2\xi}\frac{\mathrm{d}\xi}{\mathrm{d}x} = \frac{1}{d}. \tag{6.8.8}$$

Noting that

$$\frac{1}{\cos^2\xi} = 1 + \tan^2\xi = 1 + \frac{(x-a)^2}{d^2} = \frac{d^2+(x-a)^2}{d^2}, \tag{6.8.9}$$

we obtain

$$\frac{\mathrm{d}\xi}{\mathrm{d}x} = \frac{d}{d^2+(x-a)^2}. \tag{6.8.10}$$

Now we refer to equation (6.8.3), and obtain the pdf

$$\phi_x(x) = \frac{d}{d^2 + (x-a)^2} \times \phi_\xi\big(\xi(x)\big), \tag{6.8.11}$$

where $\xi$ is an implicit function of $x$, $a$, and $d$, given by

$$\xi(x) = \arctan \frac{x-a}{d} \tag{6.8.12}$$

in the range $[-\frac{1}{2}\pi,\ \frac{1}{2}\pi]$, and the inverse tangent function arises by inversion. We thus find that

$$\phi_x(x) = \frac{d}{d^2 + (x-a)^2} \times \phi_\mu\big(\xi(x) - \beta\big). \tag{6.8.13}$$

We will see that, in the Bayesian analysis, this pdf plays the role of a likelihood function *mediated by geometrical reasoning alone*.

### 6.8.4 Symmetric and unclogged

When $\beta = 0$, corresponding to a perfectly symmetric sprayer head, $\Phi(\mu) = 1/(2\alpha)$, corresponding to a uniform and unclogged sprayer head, and $\alpha = \frac{1}{2}\pi$, we obtain

$$\phi_x(x) = \frac{1}{\pi} \frac{d}{d^2 + (x-a)^2}, \tag{6.8.14}$$

which is the Cauchy distribution parametrized by the position of the sprayer, $a$, and distance from the painted surface, $d$.

### 6.8.5 Mo

A little spider named Mo is wandering on the painted surface in the early spring and senses that she is about to be painted red. Mo is urgently interested in assessing the position of the paint spray, determined by $a$ and $d$, and the rotation angle of the sprayer head, $\beta$, in terms of a joint probability density function, $\phi(a, d, \beta)$.

Mo observes that the first paint droplet hits the surface at $x = b$. where $u \le b \le v$, as shown in Figure 6.8.1. Given this datum, the pdf given in (6.8.13) serves as a likelihood function,

$$\mathcal{L}_b(a, d, \beta) = \frac{d}{d^2 + (b-a)^2} \times \phi_\mu\big(\xi(b) - \beta\big), \tag{6.8.15}$$

where

$$\xi(b) = \arctan \frac{b-a}{d} \tag{6.8.16}$$

according to (6.8.13). To ensure that nothing is amiss, Mo confirms that

$$\begin{aligned} \int_u^v \mathcal{L}_b(a, d, \beta)\, \mathrm{d}b &= \int_u^v \frac{d}{d^2 + (b-a)^2} \times \phi_\mu\big(\xi(b) - \beta\big)\, \mathrm{d}b \\ &= \int_{-\alpha+\beta}^{\alpha+\beta} \phi_\mu\big(\xi(b) - \beta\big)\, \mathrm{d}\xi(b) = 1, \end{aligned} \tag{6.8.17}$$

as required, where the points $u$ and $v$ are defined in Figure 6.8.1.

### *6.8.6 Bayesian analysis*

Bayes' equation for the triplet of unknown variables $\boldsymbol{\theta} \equiv (a, d, \beta)$ with scalar data $b$ takes the form

$$\phi(a, d, \beta \,|\, b) = \frac{\mathcal{L}_b(a, d, \beta)}{\varphi_b} \times \phi(a, d, \beta), \tag{6.8.18}$$

where

$$\varphi_b = \int_{-\pi}^{\pi} \Big( \int_0^\infty \big( \int_{-\infty}^\infty \mathcal{L}_b(a, d, \beta)\, \phi(a, d, \beta)\, \mathrm{d}a \big)\, \mathrm{d}d \Big)\, \mathrm{d}\beta \tag{6.8.19}$$

is a normalization factor.

### *6.8.7 Priors*

Mo is thinking that $a, d$, and $\beta$ are independent and factorizes

$$\phi(a, d, \beta) = \phi_a(a) \times \phi_d(d) \times \phi_\beta(\beta), \tag{6.8.20}$$

where $\phi_a(a)$, $\phi_d(d)$, and $\phi_\beta(\beta)$ are three independent pdfs.

Moreover, Mo is thinking that the sprayer rotation angle, $\beta$, may have any value in the range $[-\frac{1}{2}\pi, \frac{1}{2}\pi]$ and is zero outside this range so that the painter does not paint himself, and adopts the uniform prior pdf

$$\phi_\beta(\beta) = \frac{1}{\pi} \tag{6.8.21}$$

for $[-\frac{1}{2}\pi, \frac{1}{2}\pi]$. Integrating Bayes' equation (6.8.18) with respect to $\beta$, Mo finds the marginal posterior probability

$$\phi(a, d \,|\, b) = \frac{1}{\pi\varphi_b} \int_{-\pi/2}^{\pi/2} \mathcal{L}_b(a, d, \beta) \times \phi_a(a) \times \phi_d(d) \, \mathrm{d}\beta. \tag{6.8.22}$$

Substituting the likelihood function, Mo finds that

$$\phi(a, d \,|\, b) = \frac{1}{\psi_b} \frac{d}{d^2 + (b-a)^2} \int_{-\pi/2}^{\pi/2} \phi_\mu\big(\xi(b) - \beta\big)\phi_a(a) \times \phi_d(d) \, \mathrm{d}\beta, \tag{6.8.23}$$

where $\psi_b$ is a normalization factor.

In the case of a uniform sprayer head, $\phi_\mu(\mu) = 1/(2\alpha)$ and

$$\phi(a, d \,|\, b) = \frac{1}{2\alpha} \frac{1}{\psi_b} \frac{d}{d^2 + (b-a)^2} \int_{\beta_{\min}}^{\beta_{\max}} \phi_a(a) \times \phi_d(d) \, \mathrm{d}\beta, \tag{6.8.24}$$

where

$$\begin{aligned} \beta_{\min} &= \max(-\tfrac{1}{2}\,\pi, \xi(b) - \alpha), \\ \beta_{\max} &= \min(\tfrac{1}{2}\,\pi, \xi(b) + \alpha), \end{aligned} \tag{6.8.25}$$

where

$$\xi(b) = \arctan\frac{b-a}{d}. \tag{6.8.26}$$

Since the integrand in (6.8.24) is independent of $\beta$,

$$\phi(a, d \,|\, b) = \frac{1}{\chi_b} \frac{d}{d^2 + (b-a)^2} \, (\beta_{\max} - \beta_{\max}) \times \phi_a(a) \times \phi_d(d), \tag{6.8.27}$$

where $\chi_b$ is a normalization factor.

Mo sees a drop impinging the surface at a distance $b = 50\ell$, where $\ell$ is Mo's size, and adopts a uniform prior distribution for the sprayer position,

$$\phi_a(a) = \frac{1}{a_{\max}} \tag{6.8.28}$$

for $0 \le a \le a_{\max}$ where $a_{\max} = 100\ell$.

In addition, Mo adopts a uniform prior distribution for the sprayer distance from the surface,

$$\phi_d(a) = \frac{1}{d_{\max}} \tag{6.8.29}$$

for $0 \le a \le d_{\max}$, where $d_{\max} = 200\,\ell$. The posterior pdf for $b = 50\ell$ is shown in Figure 6.8.2.

After observing the location of the peak, Mo starts running as fast as he can away from the impinging drop.

***Exercise***

**6.8.1** Mo decides to use the normal prior distributions

$$\phi_a(a) = \mathcal{N}_{\overline{a},\sigma_a}(a), \qquad \phi_d(d) = \mathcal{N}_{\overline{d},\sigma_d}(d), \tag{6.8.30}$$

where $\overline{a}$ and $\overline{d}$ are assigned expected values and $\sigma_a$ and $\sigma_d$ are assigned standard deviations. Reproduce the counterpart of Figure 4.8.2 for these distributions.

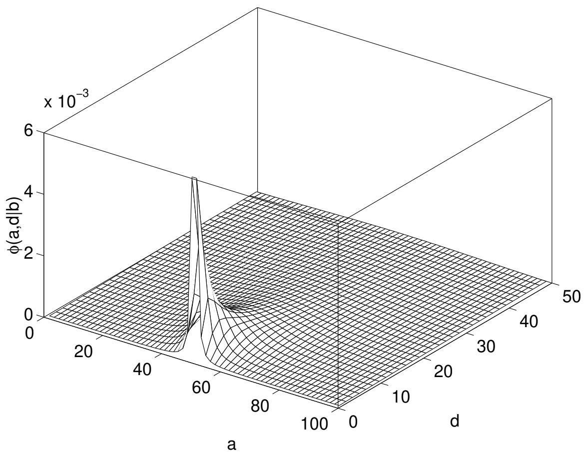



FIGURE 6.8.2 Posterior distribution of sprayer head's position in Mo's sizes, $\ell$, when a droplet impinges a surface $50 \times \ell$ away from Mo.

## 6.9 *Harry and the bone*

A forgetful golden retriever named Harry buried a bone at the backyard and is now smelling it at an imprecise distance nearby, as shown in Figure 6.9.1.

Harry may detect the same smell intensity if his nose is clean and the bone is buried at the standard depth of $2$ inches (typical bone burial), if the bone is closer but his nose is stuffed, or if the bone is farther away but hardly buried. In the case of stuffy nose, Harry's sense of smell is less than that under normal conditions.

Conversely, for a given bone location, Harry smells the bone at a variety of intensities depending on the stuffiness of his nose and the burial depth.

### *6.9.1 Likelihood function*

Harry is thinking that the probability that the bone is located at a

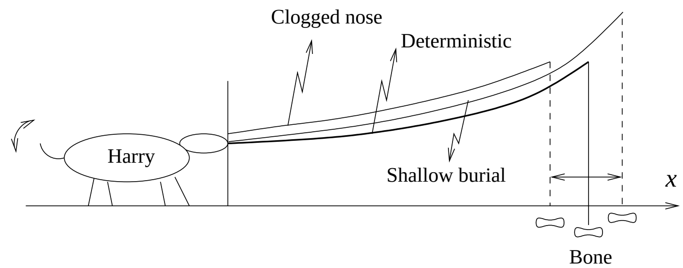



FIGURE 6.9.1 Illustration of the smell field generated by a buried bone. Harry experiences the same smell in all three cases.

certain distance, $x$, is given by the Gaussian distribution

$$\mathcal{L}_s(x,\sigma) = \frac{1}{\sqrt{2\pi}\,\sigma}\,\exp(-\tfrac{1}{2}\,\frac{\widehat{s}^2(x)}{\sigma^2}), \tag{6.9.1}$$

where $\sigma$ is an *a priori* unknown standard deviation and

$$\widehat{s} \equiv s - s_{\rm std}(x) \tag{6.9.2}$$

is the difference between Harry's smell sensation, $s$, and a standard sensation, $s_{\rm std}(x)$, discussed next. Justification for this likelihood function was given in Section 4.4 in the context of the maximum entropy principle which Harry has mastered while attending dog obedience school. Harry has no way of choosing the standard deviation and treats it as an unknown.

We observe that the likelihood function for the variable of interest, $x$, is determined by a field function involving $x$ according to an appropriate law mediated by the Gaussian distribution. As mentioned on a previous occasion, this realization is a key to performing Bayesian analysis in a broad range of physical and conceptual applications.

### *6.9.2 Standard smell*

Harry did some self-calibration to find that the standard smell is given by

$$s_{\rm std}(x) = s_0\,\mathcal{S}(\frac{x}{a}), \tag{6.9.3}$$

where $s_0$ is the smell when the bone is just under Harry's paws, $a$ is the length of Harry's tail, and the dimensionless function $\mathcal{S}(w)$ expresses the bone smell field. By definition

$$\mathcal{S}(0) = 1. \tag{6.9.4}$$

By common sense, $\mathcal{S}(x/a)$, decays to zero as $x/a$ tends to infinity.

### 6.9.3 Bayes equation

Bayes' equation takes the form

$$\phi(x, \sigma \,|\, s) = \frac{\mathcal{L}_s(x, \sigma)}{\varphi_s} \times \phi(x, \sigma), \tag{6.9.5}$$

where

$$\varphi_s = \int_0^\infty \Big( \int_{-\infty}^\infty \mathcal{L}_s(x', \sigma')\, \phi(x', \sigma')\, \mathrm{d}x' \Big)\, \mathrm{d}\sigma' \tag{6.9.6}$$

is a marginal probability.

The probability of interest in locating the bone is

$$\phi_x(x) = \int_0^\infty \phi(x, \sigma)\, \mathrm{d}\sigma. \tag{6.9.7}$$

Bayes's equation suggests that

$$\phi_x(x \,|\, s) = \frac{1}{\varphi_s} \int_0^\infty \mathcal{L}_s(x, \sigma) \times \phi(x, \sigma)\, \mathrm{d}\sigma. \tag{6.9.8}$$

A joint pdf for $x$ and $\sigma$ must be chosen as a prior.

### 6.9.4 Separable prior

Harry is thinking that the standard deviation, $\sigma$, is independent of the burial distance, $x$, and assumes a prior in the form

$$\phi(x, \sigma) = \phi_x(x) \times \phi_\sigma(\sigma), \tag{6.9.9}$$

where $\phi_x(x)$ and $\phi_\sigma(\sigma)$ are pdfs. The probability of interest in locating the bone is given by

$$\phi_x(x \,|\, s) = \phi_x(x)\, \frac{1}{\varphi_s} \int_0^\infty \mathcal{L}_s(x, \sigma)\, \phi_\sigma(\sigma)\, \mathrm{d}\sigma, \tag{6.9.10}$$

where

$$\varphi_s = \int_{-\infty}^{\infty} \Big( \int_0^{\infty} \mathcal{L}_s(x', \sigma')\, \phi_\sigma(\sigma')\, \mathrm{d}\sigma' \Big) \times \phi_x(x')\, \mathrm{d}x' \qquad (6.9.11)$$

is a marginal probability.

### 6.9.5 Uniform prior for $\sigma$

Harry may assume that the pdf for the standard deviation is uniform in a chosen interval, $[0, \Sigma]$,

$$\phi_\sigma(\sigma) = \frac{1}{\Sigma} \qquad (6.9.12)$$

and zero outside this interval for $\sigma > \Sigma$. Under this assumption, Bayes' equation (6.9.10) becomes

$$\phi_x(x \,|\, s) = \phi_x(x)\, \frac{1}{\varphi_s}\, \frac{1}{\Sigma} \int_0^{\Sigma} \mathcal{L}_s(x, \sigma)\, \mathrm{d}\sigma. \qquad (6.9.13)$$

Substituting the normal distribution for the likelihood function,

$$\mathcal{L}_s(x, \sigma) = \frac{1}{\sqrt{2\pi}\sigma}\, \exp(-\tfrac{1}{2}\, \frac{\widehat{s}^2}{\sigma^2}), \qquad (6.9.14)$$

Harry obtains

$$\phi_x(x \,|\, s) = \phi_x(x)\, \frac{1}{\varphi_s}\, \frac{1}{\Sigma} \frac{1}{\sqrt{2\pi}} \int_0^{\Sigma} \frac{1}{\sigma}\, \exp(-\tfrac{1}{2}\, \frac{\widehat{s}^2}{\sigma^2})\, \mathrm{d}\sigma. \qquad (6.9.15)$$

To compute the integral, Harry runs into the dog house and pulls out his favorite calculus book.

### 6.9.6 Computing an integral

Harry decides to introduce the auxiliary variable $v = 1/\sigma^2$ and scratches in the dirt the definition

$$\mathcal{J} \equiv \int_0^{\Sigma} \frac{1}{\sigma}\, \exp\big(-\tfrac{1}{2}\, \frac{\widehat{s}^2}{\sigma^2}\big)\, \mathrm{d}\sigma, \qquad (6.9.16)$$

which can be restated as

$$\mathcal{J} = \tfrac{1}{2} \int_{1/\Sigma^2}^{\infty} \frac{1}{v}\, \exp\big(-\tfrac{1}{2}\, \widehat{s}^2 v\big)\, \mathrm{d}v. \qquad (6.9.17)$$

Harry recognizes that

$$\mathcal{J} = \tfrac{1}{2}\,\mathrm{E}_1\big(\tfrac{1}{2}\,\frac{\widehat{s}^2}{\Sigma^2}\big), \tag{6.9.18}$$

where $\mathrm{E}_1$ is the exponential integral defined as

$$\mathrm{E}_1(z) \equiv \int_z^{\infty} \frac{\mathrm{e}^{-t}}{t}\,\mathrm{d}t. \tag{6.9.19}$$

(e.g., Abramowitz, M. & Stegun, I. A. (1972) *Handbook of Mathematical Functions.* Dover, p. 228).

Bayes equation now becomes

$$\phi_x(x\,|\,s) = \frac{1}{\chi_{\widehat{s}}}\,\mathrm{E}_1\big(\tfrac{1}{2}\,\frac{\widehat{s}^2}{\Sigma^2}\big) \times \phi(x), \tag{6.9.20}$$

where

$$\chi_{\widehat{s}} = \int_{-\infty}^{\infty} \mathrm{E}_1\big(\tfrac{1}{2}\,\frac{\widehat{s}^2}{\Sigma^2}\big) \times \phi(x)\,\mathrm{d}x \tag{6.9.21}$$

is a normalization constant. For small arguments, the exponential integral behaves like

$$\mathrm{E}_1(w) \simeq -\gamma - \ln w, \tag{6.9.22}$$

where $\gamma$ is Euler's constant. Harry is concerned that the logarithmic term will be a source of difficulty in his Bayesian analysis. Notwithstanding this difficulty, Harry sets for convenience

$$\Sigma = \beta s_0, \tag{6.9.23}$$

where $\beta$ is a numerical factor, and obtains

$$\phi_x(x\,|\,s) = \phi_x(x)\,\frac{1}{\chi_{\widehat{s}}}\,\mathrm{E}_1\big(\tfrac{1}{2}\,\frac{\zeta^2}{\beta^2}\big) \tag{6.9.24}$$

where

$$\zeta = \frac{s}{s_0} - \mathcal{S}(\frac{x}{a}) \tag{6.9.25}$$

and

$$\chi_{\widehat{s}} = \int_{-\infty}^{\infty} \mathrm{E}_1\big(\tfrac{1}{2}\,\frac{\zeta^2}{\beta^2}\big) \times \phi(x)\,\mathrm{d}x \tag{6.9.26}$$

is a normalization factor. Harry recognizes that the posterior based on the uniform pdf for $\sigma$ expressed by (6.9.12) exhibits an integrable singularity.

### *6.9.7 Another pdf for $\sigma$*

To circumvent the logarithmic term, Harry decides to switch to a different prior pdf for the standard deviation, $\sigma$, and chooses

$$\phi_\sigma(\sigma) = \tfrac{15}{2}\,\frac{1}{\Sigma^5}\,\sigma^2\,(\Sigma^2 - \sigma^2) \tag{6.9.27}$$

for $0 \le \sigma \le \Sigma$ or $\phi(\sigma) = 0$ for $\sigma > \Sigma$. Bayes equation becomes

$$\phi_x(x\,|\,s) = \phi(x)\,\frac{1}{\varphi_s}\,\frac{1}{\Sigma^5}\,\frac{15}{2\sqrt{2\pi}} \int_0^{\Sigma} \sigma\,(\Sigma^2 - \sigma^2)\,\exp(-\tfrac{1}{2}\,\frac{\widehat{s}^2}{\sigma^2})\,\mathrm{d}\sigma. \tag{6.9.28}$$

To compute the integral, Harry runs into the dog house, pulls out once again his favorite calculus book, defines the auxiliary variable $v = 1/\sigma^2$, and scratches with his front left paw in the dirt

$$\mathcal{I} \equiv \int_0^{\Sigma} \sigma\,(\Sigma^2 - \sigma^2)\,\exp\big(-\tfrac{1}{2}\,\frac{\widehat{s}^2}{\sigma^2}\big)\,\mathrm{d}\sigma, \tag{6.9.29}$$

and then

$$\mathcal{I} = \tfrac{1}{2} \int_{1/\Sigma^2}^{\infty} \frac{1}{v^2}\,(\Sigma^2 - \frac{1}{v})\,\exp\big(-\tfrac{1}{2}\,\widehat{s}^2 v\big)\,\mathrm{d}v. \tag{6.9.30}$$

Setting $v = u/s_0^2$, Harry obtains

$$\mathcal{I} = \tfrac{1}{2}\,s_0^4 \int_{1/\beta^2}^{\infty} \frac{1}{u^2}\,(\beta^2 - \frac{1}{u})\,\exp\big(-\tfrac{1}{2}\,\zeta^2 u\big)\,\mathrm{d}u, \tag{6.9.31}$$

where $\beta \equiv \Sigma/a$. The integral converges for any value of $\zeta$, including zero. Bayes equation becomes

$$\phi_x(x\,|\,s) = \phi_x(x)\,\frac{1}{\omega_s} \int_{1/\beta^2}^{\infty} \frac{1}{u^2}\,(\beta^2 - \frac{1}{u})\,\exp\big(-\tfrac{1}{2}\,\zeta^2 u\big)\,\mathrm{d}u. \tag{6.9.32}$$

The denominator,

$$\omega_s = \int_{-\infty}^{\infty} \Big( \int_{1/\beta^2}^{\infty} \frac{1}{u^2} \, (\beta^2 - \frac{1}{u}) \, \exp\big( -\tfrac{1}{2}\, \zeta^2 u \big) \, \mathrm{d}u \Big) \times \phi_x(x) \, \mathrm{d}x, \tag{6.9.33}$$

is a normalization constant.

### 6.9.8 Harry gets the bone

Harry is thinking that the bone is most likely buried $\lambda$ tail lengths away from where he stands and uses the prior pdf

$$\phi_x(x) = \mathcal{N}_{\lambda a, \mu a}, \tag{6.9.34}$$

where $\mathcal{N}$ is the normal distribution and $\mu$ is a numerical factor. Moreover, he uses the standard smell field

$$\mathcal{S}(\frac{x}{a}) = \frac{a}{|x|}. \tag{6.9.35}$$

Harry pulls out his laptop and codes the equations.

Results for $\lambda = 2$, $\mu = 2$, $s/s_0 = 0.2$, and $\beta = 0.1$ are shown in Figure 6.9.2. The graphs indicate that the bone is buried most likely only about four tail lengths away. Harry proceeds to dig out the bone and enjoy his snack.

#### Exercise

**6.9.1** Confirm that the pdf stated in (6.9.27) satisfies the mandatory normalization condition.

## 6.10 Sue and Marisol

A beautiful woodpecker whose name is Marisol is perched at the top branch of a tree, as shown in Figure 6.10.1. All of a sudden, Marisol sees a brown cat hanging around the tree. Immediately she notifies her friend, Sue, who is perched at the top branch of another tree, by chirping a particular sinusoidal frequency that encodes a warning.

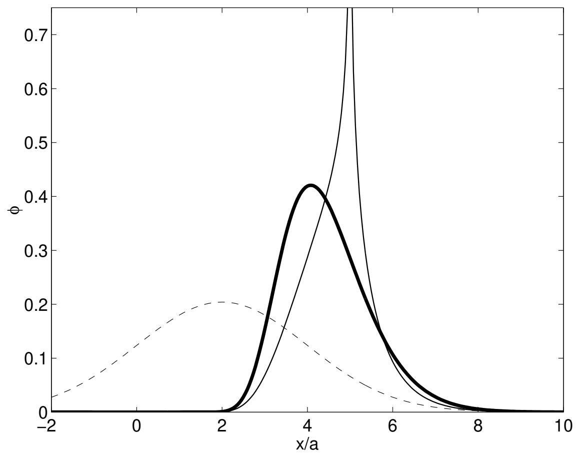



FIGURE 6.9.2 Prior (broken line) and posterior pdf computed using (6.9.12) (solid line) or (6.9.27) (heavy line) for smell intensity $s/s_0 = 0.20$.

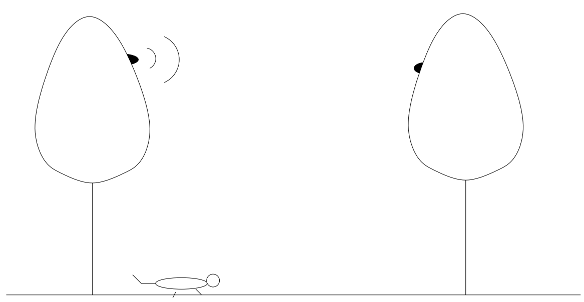


FIGURE 6.10.1 A beautiful woodpecker named Marisol, perched at the top branch of a tree, notices a brown cat is hanging around the tree, and notifies her friend, Sue, who is perched on another tree.

### *6.10.1 Acoustic signal*

The whistling is distorted by the wind, the chirping of other birds, the singing of tree frogs, the enthusiastic hymns of bullfrogs, and the roaring of remote chainsaws. Sue listens carefully and retains a time series of the acoustic signal, $s_i$ for $i = 1, \ldots, N$, encapsulated in an array

$$\mathbf{s} = (\, s_1, s_2, \ldots, \, s_N \,). \tag{6.10.1}$$

A time series is a sequence of data separated by a fixed time interval, $\Delta t$, also called the sampling time.

### 6.10.2 Sinusoidal fitting

Sue is struggling to decipher the angular frequency of the whistle, $\omega$, and fits it into a sinusoidal function,

$$f(t) = a \sin(\omega t) \tag{6.10.2}$$

for $0 \le \omega < 2\pi$, where $a$ is the amplitude. Sue interprets probability as degree of certainty and introduces a joint probability density function for $\omega$ (of prime interest) and $a$ (of low interest),

$$\phi(\omega, a), \tag{6.10.3}$$

where the amplitude, $a$, is regarded as a nuisance parameter. Recalling Bayes' theorem, Sue writes

$$\phi(\omega, a \,|\, \mathbf{s}) = \frac{\mathcal{L}_{\mathbf{s}}(\omega, a)}{\varphi_{\mathbf{s}}} \times \phi(\omega, a), \tag{6.10.4}$$

where

$$\varphi_{\mathbf{s}} = \int_0^\infty \Big( \int_0^{2\pi} \mathcal{L}_{\mathbf{s}}(\omega, a)\, \phi(\omega, a)\, \mathrm{d}\omega \Big)\, \mathrm{d}a \tag{6.10.5}$$

is a marginal probability playing the role of a normalization constant.

### 6.10.3 Likelihood function

Sue introduces the difference between the received and assumed signal at the $i$th time instant, $t_i$,

$$\widehat{s}_i \equiv s_i - f(t_i) = s_i - a\, \sin(\omega t_i), \tag{6.10.6}$$

and assumes that the likelihood function is given by the product of $N$ Gaussian distributions with a common standard deviation, $\sigma$,

$$\mathcal{L}_{\mathbf{s}}(\omega, a, \sigma) = \mathcal{N}_{\widehat{s}_1,\sigma} \times \cdots \times \mathcal{N}_{\widehat{s}_N,\sigma}, \tag{6.10.7}$$

where

$$\mathcal{N}_{\widehat{s}_i,\sigma}, = \frac{1}{\sqrt{2\pi}\,\sigma}\,\exp(-\tfrac{1}{2}\,\frac{\widehat{s}_i^2}{\sigma^2}) \tag{6.10.8}$$

for $i = 1, \ldots, N$. Justification for this choice is provided by the principle of maximum entropy discussed in Section 4.2.

Substituting into the expression for the likelihood function the $N$ Gaussian distributions, and using the properties of the exponential function, Sue obtains

$$\mathcal{L}_{\mathbf{s}}(\omega, a, \sigma) = \frac{1}{(2\pi)^{N/2}}\frac{1}{\sigma^N}\exp\big(-\tfrac{1}{2}\,N\,\frac{s_{\text{rms}}^2}{\sigma^2}\big), \tag{6.10.9}$$

where

$$s_{\text{rms}} \equiv \Big(\frac{1}{N}\sum_{i=1}^{N}\widehat{s}_i^2\Big)^{1/2} \tag{6.10.10}$$

is the root mean square (rms) deviation of the signal from the sinusoidal model.

The sum in (6.10.10) is zero only if the signal is a perfect match of the model expressed by the function $f(t)$. The maximum likelihood (ML) estimate for $\omega$ and $a$ is the value of $\omega$ and $a$ where $s_{\text{rms}}^2$ reaches a minimum (least squares minimization.)

### 6.10.4 *Extended formulation*

In the absence of insights on the standard deviation, $\sigma$, and amplitude, $a$, Sue treats them as unknowns and considers the extended Bayes equation

$$\phi(\omega, a, \sigma \,|\, \mathbf{s}) = \frac{\mathcal{L}_{\mathbf{s}}(\omega, a, \sigma)}{\psi_{\mathbf{s}}} \times \phi(\omega, a, \sigma), \tag{6.10.11}$$

where $\phi(\omega, a, \sigma)$ is a joint probability density function and

$$\psi_{\mathbf{s}} = \int_0^{\infty}\Big(\int_0^{\infty}\big(\int_0^{2\pi}\mathcal{L}_{\mathbf{s}}(\omega, a, \sigma)\,\phi(\omega, a, \sigma)\,\mathrm{d}\omega\big)\,\mathrm{d}a\Big)\,\mathrm{d}\sigma \tag{6.10.12}$$

is a normalization factor. Sue is interested in the pdf of the angular frequency,

$$\phi(\omega \,|\, \mathbf{s}) = \int_0^\infty \int_0^\infty \phi(\omega, a, \sigma \,|\, \mathbf{s}) \, \mathrm{d}a \, \mathrm{d}\sigma, \tag{6.10.13}$$

given by the marginal probability

$$\phi(\omega \,|\, \mathbf{s}) = \frac{1}{\psi_\mathbf{s}} \times \int_0^\infty \int_0^\infty \mathcal{L}_\mathbf{s}(\omega, a, \sigma) \times \phi(\omega, a, \sigma) \, \mathrm{d}a \, \mathrm{d}\sigma \tag{6.10.14}$$

according to Bayes' rule.

### 6.10.5 Priors

Sue needs to quickly adopt a distribution for the prior. In the absence of evidence to the contrary, Sue is thinking of independence and sets

$$\phi(\omega, a, \sigma) = \phi_\omega(\omega) \times \phi_a(a) \times \phi_\sigma(\sigma). \tag{6.10.15}$$

For the angular frequency, $\omega$, Sue uses the uniform distribution,

$$\phi_\omega(\omega) = \frac{1}{2\pi} \tag{6.10.16}$$

for $0 \le \omega < 2\pi$. For the amplitude, Sue uses another uniform distribution,

$$\phi_a(a) = \frac{1}{a_{\max}} \tag{6.10.17}$$

for $0 \le a < a_{\max}$ and $\phi(a) = 0$ for $a > a_{\max}$, where $A$ is the maximum amplitude that Marisol's vocal chords can generate.

### 6.10.6 Uniform prior for $\sigma$

Sue must now adopt a prior pdf for the standard deviation, $\sigma$. To make things easy in light of the ominous presence of the cat, Sue introduces another uniform distribution,

$$\phi_\sigma(\sigma) = \frac{1}{\sigma_{\max}} \tag{6.10.18}$$

for $0 \le \sigma < \sigma_{\max}$ and $\phi_\sigma(\sigma) = 0$ for $\sigma > \sigma_{\max}$, where $\sigma_{\max}$ is a specified cut-off. With this prior, Bayes' equation becomes

$$\phi(\omega \,|\, \mathbf{s}) = \frac{1}{2\pi\psi_{\mathbf{s}}} \, \frac{1}{a_{\max}\sigma_{\max}} \times \int_0^{a_{\max}} \int_0^{\sigma_{\max}} \mathcal{L}_{\mathbf{s}}(\omega, a, \sigma) \, \mathrm{d}\sigma \, \mathrm{d}a. \tag{6.10.19}$$

Substituting the expression for the likelihood function, Sue obtains

$$\phi(\omega \,|\, \mathbf{s}) = \frac{1}{\chi_{\mathbf{s}}} \int_0^{a_{\max}} \mathcal{J}(\omega, a, \sigma) \, \mathrm{d}\sigma \, \mathrm{d}a, \tag{6.10.20}$$

where

$$\mathcal{J}(\omega, a, \sigma) \equiv \int_0^{\sigma_{\max}} \frac{1}{\sigma^N} \exp\big( -\tfrac{1}{2} \, N \, \frac{s_{\mathrm{rms}}^2}{\sigma^2} \big) \, \mathrm{d}\sigma \tag{6.10.21}$$

and $\chi_{\mathbf{s}}$ is a normalization constant.

To compute the integral defining $\mathcal{J}$, Sue introduces the variable $v = 1/\sigma^2$ and obtains

$$\mathcal{J} = \tfrac{1}{2} \int_{1/\sigma_{\max}^2}^{\infty} v^{(N-3)/2} \, \exp\big( -\tfrac{1}{2} \, N \, s_{\mathrm{rms}}^2 \, v \big) \, \mathrm{d}v. \tag{6.10.22}$$

Next, Sue defines $w = v \, s_{\mathrm{rms}}^2$ and obtains

$$\mathcal{J} = \tfrac{1}{2} \frac{1}{s_{\mathrm{rms}}^{N-1}} \int_{s_{\mathrm{rms}}^2/\sigma_{\max}^2}^{\infty} w^{(N-3)/2} \, \exp\big( -\tfrac{1}{2} \, N \, w \big) \, \mathrm{d}w. \tag{6.10.23}$$

When $N \ge 2$, Sue may let $\sigma_{\max}$ tend to infinity and obtain a convergent integral from $0$ to $\infty$ on the right-hand side.

Substituting the expression for $\mathcal{J}$ given in (6.10.23) into Bayes' equation (6.10.20), Sue finds that

$$\phi(\omega \,|\, \mathbf{s}) = \frac{1}{\xi_{\mathbf{s}}} \int_0^{a_{\max}} \frac{1}{s_{\mathrm{rms}}^{N-1}} \, \mathrm{d}a, \tag{6.10.24}$$

where $\xi_{\mathbf{s}}$ is a normalization constant. We recall that the rms value is zero only if the entire signal is a perfect fit of the sinusoidal function. It is now permissible to let $a_{\max}$ tend to infinity and obtain

$$\phi(\omega \,|\, \mathbf{s}) = \frac{1}{\xi_{\mathbf{s}}} \int_0^{\infty} \frac{1}{\Big( \frac{1}{N} \, \sum_{i=1}^{N} \widehat{s}_i^{\,2} \Big)^{(N-1)/2}} \, \mathrm{d}a. \tag{6.10.25}$$

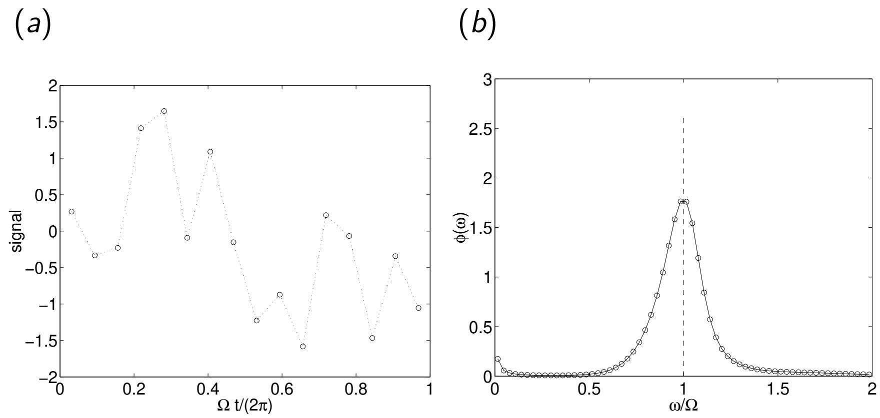



FIGURE 6.10.2 (*a*) Acoustic signal received by Sue for angular frequency $\Omega$ and (*b*) posterior frequency probability density function.

Using this equation is tantamount to performing probabilistic regression.

### *6.10.7 Garbled sinusoidal signal*

The signal received by Sue describes a sinusoidal transmission with angular frequency $\Omega$ and period $2\pi/\omega$, corrupted by significant random noise,

$$s_i = A\left( \sin(\Omega t_i) + \epsilon\,(1 - 2\,\varrho_i)\right) \tag{6.10.26}$$

for $i = 1, \ldots N$, where $A$ is the amplitude, $\epsilon$ is an arbitrary coefficient and $\varrho_i$ is a random number with uniform pdf varying in the range $[0, 1]$. A time sequence over one period is shown in Figure 6.10.2(*a*).

The posterior probability density function computed using the Bayes equation (6.10.25) is implemented in the following Matlab code:

```
%---
% parameters
%---

Om = 1.87; % arbitrary
Am = 0.90; % arbitrary
```

```
eps = 1.25; % arbitrary

N = 16; % signal size
amax = 100;
Na = 128;
No = 64;

%---
% prepare
%---

pi2 = 2.0*pi;
T = pi2/Om;
Da = amax/Na;
Dt = T/N;
Do = 2.0*Om/No;

%---
% signal
%---

for i=1:N
 time(i) = (i-0.5)*Dt;
 s(i) = sin(Om*time(i))-ramp*(rand-0.5);
 s(i) = Am*s(i);
end

 figure(1)
 plot(Om*time/pi2,s,'ko:')

%-------
% Bayes
%-------

nrm = 0.0;  % normalization of phi(omega)

%---
for io=1:No % loop over omega
%---
```

```
 om = (io-0.5)*Do;
 oplot(io) = om;

 % compute the integral over a

 smn = 0.0;

 for i=1:Na  % loop over ampl
  a = (i-0.5)*Da;
  rms = 0.0;
  for j=1:N
   rms = rms + ( s(j)-a*sin(om*time(j)))^2;
  end
  rms = sqrt(rms/N);
  smn = smn + 1.0/rms^(N-1);
 end
 phi(io) = smn*Da;
 nrm = nrm+phi(io);

end

nrm = nrm*Do;

%---
% normalize
%---

for io=1:No
 phi(io) = phi(io)/nrm;
end

%---
% plot
%---

figure(2)
plot(oplot/Om,phi,'ko-')
plot([1.0,1.0],[0,1.2*max(phi)],'k--')
```

The code invokes the Matlab internal random number generator *rand*. The posterior probability density function computed by the code is shown in Figure 6.10.2(*b*). Sue notices a pronounced peak at the uncorrupted transmitted frequency $\Omega$ and flies to a higher branch, just in case the cat gets any ideas.

#### 6.10.8 Jeffrey's prior for $\sigma$

As crazy as it may seem, Sue could have chosen Jeffrey's prior

$$\phi(\sigma) = \frac{1}{\sigma}. \tag{6.10.27}$$

The craziness arises because the integral of this distribution from $0$ to any limit is not defined. Notwistanding this difficulty, we find that

$$\phi(\omega \,|\, \mathbf{s}) = \frac{1}{\zeta_{\mathbf{s}}} \int_0^{a_{\max}} \frac{1}{s_{\rm rms}^N} \, \mathrm{d}a, \tag{6.10.28}$$

where $\zeta_{\mathbf{s}}$ is a normalization constant. The predictions are similar to those for the uniform prior.

#### Exercise

**6.10.1** Fit a signal corrupted by noise to a linear function, $f(t) = at+b$, using Sue's method.

## 6.11 Detecting a blood clot

Professor V. S. Mart has developed a medical device for sensing the onset of a blood clot attached to the exposed surface of a blood vessel, called the endothelium. His device involves an array of $N_s = 6$ sensors implanted to the endothelium, marked S1–S6, as shown in Figure 6.11.1.

#### 6.11.1 V6 engine cylinders

It is interesting that the professor subconsciously labelled the sensors in the order that his truck's manufacturer labelled the cylinders of the

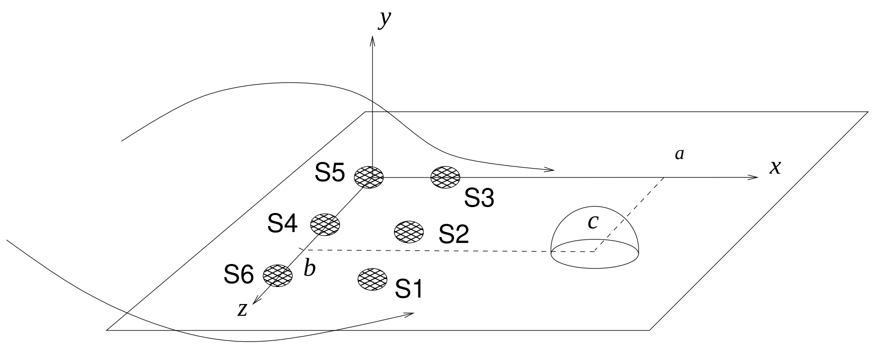



FIGURE 6.11.1 An array of six sensors are attached to the endothelium of a blood vessel to detect the development of a blood clot.

vehicle V6 engine. Cylinders S4, S5, and S6 are located at the front of the truck behind the radiator, whereas cylinders S1, S2, and S3 face the dashboard.

Replacing the spark plugs of cylinders S1, S2, and S3 requires removing the engine cover, which is a time-consuming and pricey job. A new engine cover gasket must also be installed after removing the engine cover.

### *6.11.2 Hemispherical clot model*

The idea is that an attached blood clot, modeled as a hemispherical bump of radius $c$ centered at the position $x = a$ and $y = b$, generates a perturbation flow, as shown in Figure 6.11.1. The induced shear stress is registered by the six sensors as a multi-threaded signal, denoted by

$$\odot_i(a, b, c) \tag{6.11.1}$$

for $i = 1, \ldots, 6$. The signal will be analyzed by software to infer the location and size of the clot.

The professor is working on a computational model that would allow him to detect the location and size of the clot.

### 6.11.3 Biofluiddynamics

The professor has a friend who knows a friend who specializes in scientific computing, biofluiddynamics, and blood flow. The friend has posted on his web site a computer code that calculates the shear stress at the location of the sensors for any specified triplet, $(a, b, c)$ based on the boundary-element method. The prediction of the code is denoted as

$$\oplus_i(a, b, c) \tag{6.11.2}$$

for $i = 1, \ldots, 6$. Although the code was written in the Fortran programming language to ensure ease of use and backward compatibility, it can be be easily translated into any other computer language.

### 6.11.4 Uncertainties

The professor realizes that the measured signals and calculated predictions will differ due to various assumptions, approximations, and random deviations, and considers the differences in the measured and computed values, denoted by

$$\widehat{\odot}_i(a, b, c) \equiv \odot_i - \oplus_i \tag{6.11.3}$$

for $i = 1, \ldots, 6$. To make progress, the professor regards these differences as random variables whose individual probability density functions (pdfs) are all described by the normal distribution with zero mean and variances $\sigma_i^2$,

$$\phi_i(\widehat{\odot}_i) \equiv \mathcal{N}_{0,\sigma_i}(\widehat{\odot}_i) = \frac{1}{\sqrt{2\pi}\,\sigma_i} \exp(-\tfrac{1}{2}\,\frac{\widehat{\odot}_i^{\,2}}{\sigma_i^2}) \tag{6.11.4}$$

for $i = 1, \ldots, 6$. Justification for this choice is provided by the principle of maximum entropy discussed in Section 4.5. The professor accepts the principle as an appropriate and useful framework.

### 6.11.5 Joint pdf

Assuming that the differences $\widehat{\odot}_i$ are independent, the professor considers the joint pdf

$$\phi(\widehat{\odot}_1, \ldots, \widehat{\odot}_6) = \mathcal{N}_{0,s_1}(\widehat{\odot}_1) \times \cdots \times \mathcal{N}_{0,s_6}(\widehat{\odot}_6). \tag{6.11.5}$$

Substituting the normal distributions and using the properties of the exponential function, the professor derives the joint distribution

$$\phi(\widehat{\odot}_1, \ldots, \widehat{\odot}_6) \\ = \frac{1}{(2\pi)^3} \frac{1}{\sigma_1 \times \cdots \times \sigma_6} \times \exp\left( -\tfrac{1}{2} \left( \frac{\widehat{\odot}_1^2}{\sigma_1^2} + \cdots + \frac{\widehat{\odot}_6^2}{\sigma_6^2} \right)\right). \tag{6.11.6}$$

This expression provides us with a likelihood function,

$$\mathcal{L}_{\widehat{\odot}_1,\ldots,\widehat{\odot}_6}(a, b, c, \sigma_1, \ldots, \sigma_6) = \phi(\widehat{\odot}_1, \ldots, \widehat{\odot}_6), \tag{6.11.7}$$

where the standard deviations, $\sigma_i$, are treated as unknowns.

### 6.11.6 Bayes' theorem

Now the professor introduces the joint pdf of an extended set of unknowns including the clot size and location and standard deviations of the signal,

$$\phi(a, b, c, \sigma_1, \ldots, \sigma_6). \tag{6.11.8}$$

The standard deviations are regarded as nuisance parameters playing the role of unwanted guests. According to Bayes' theorem, given six signals, the posterior pdf is given by

$$\phi(a, b, c, \sigma_1, \ldots, \sigma_6 \,|\, \widehat{\odot}_1, \ldots, \widehat{\odot}_6) \\ = \alpha\, \mathcal{L}_{\widehat{\odot}_1,\ldots,\widehat{\odot}_6}(a, b, c, \sigma_1, \ldots, \sigma_6) \times \phi(a, b, c, \sigma_1, \ldots, \sigma_6), \tag{6.11.9}$$

where $\alpha$ is a normalization constant and $\phi(a, b, c, \sigma_1, \ldots, \sigma_6)$ on the right-hand side is a prior pdf.

### 6.11.7 Pdf of interest

The joint probability of interest for detecting the location and size of the clot is given by

$$\phi(a, b, c) = \int_0^{\Sigma_1} \cdots \int_0^{\Sigma_6} \phi(a, b, c, \sigma_1, \ldots, \sigma_6)\, \mathrm{d}\sigma_1 \cdots \mathrm{d}\sigma_6 \tag{6.11.10}$$

as a prior or posterior, where $\Sigma_i$ are reasonable maximum values of the standard deviations. The six integrations serve to screen out the nuisance parameters. Bayes equation suggests that

$$\phi(a, b, c \,|\, \widehat{\odot}_1, \ldots, \widehat{\odot}_6) = \alpha \int_0^{\Sigma_1} \cdots \int_0^{\Sigma_6} \mathcal{L}_{\widehat{\odot}_1,\ldots,\widehat{\odot}_6}(a, b, c, \sigma_1, \ldots, \sigma_6) \times \phi(a, b, c, \sigma_1, \ldots, \sigma_6)\, \mathrm{d}\sigma_1 \cdots \mathrm{d}\sigma_6 \tag{6.11.11}$$

for a given sextuple of signal to prediction differences.

### 6.11.8 Choice of prior

A prior pdf must be adopted. It is reasonable to assume that the physical parameters $a, b, c$ are independent of the standard deviations in the prior, $\sigma_i$, so that

$$\phi(a, b, c, \sigma_1, \ldots, \sigma_6) = \phi(a, b, c) \times \phi(\sigma_1, \ldots, \sigma_6). \tag{6.11.12}$$

Bayes equation becomes

$$\phi(a, b, c \,|\, \widehat{\odot}_1, \ldots, \widehat{\odot}_6) = \alpha\, \phi(a, b, c) \int_0^{\Sigma_1} \cdots \int_0^{\Sigma_6} \mathcal{L}_{\widehat{\odot}_1,\ldots,\widehat{\odot}_6}(a, b, c, \sigma_1, \ldots, \sigma_6) \times \phi(\sigma_1, \ldots, \sigma_6)\, \mathrm{d}\sigma_1 \cdots \mathrm{d}\sigma_6. \tag{6.11.13}$$

Substituting the likelihood function, we obtain a long expression,

$$\phi(a, b, c \,|\, \widehat{\odot}_1, \ldots, \widehat{\odot}_6) = \alpha \times \phi(a, b, c) \times \frac{1}{(2\pi)^3} \times \int_0^{\Sigma_1} \cdots \int_0^{\Sigma_6} \frac{1}{\prod \sigma_i} \times \exp\big( -\tfrac{1}{2} \sum_{i=1}^{6} \frac{\widehat{\odot}_i^{\,2}}{\sigma_i^2} \big)\, \phi(\sigma_1, \ldots, \sigma_6)\, \mathrm{d}\sigma_1 \cdots \mathrm{d}\sigma_6, \tag{6.11.14}$$

where $\prod$ denotes the product. To make further progress, the professor must adopt a joint pdf for the six standard deviations.

### 6.11.9 Uniform prior for the standard deviations

Choosing independent uniform priors for the standard deviations, the professor sets

$$\phi(\sigma_i) = \frac{1}{\Sigma_i} \tag{6.11.15}$$

in a chosen interval, $[0, \Sigma_i]$, and obtains

$$\phi(\sigma_1, \ldots, \sigma_6) = \prod \phi(\sigma_i) = \frac{1}{\prod \Sigma_i}. \tag{6.11.16}$$

Equation (6.11.14) becomes

$$\phi(a, b, c \,|\, \widehat{\odot}_1, \ldots, \widehat{\odot}_6) = \alpha\, \phi(a, b, c) \times \frac{1}{(2\pi)^3} \frac{1}{\prod \Sigma_i}$$
$$\times \int_0^{\Sigma_1} \cdots \int_0^{\Sigma_6} \frac{1}{\prod \sigma_i} \times \exp\big( -\tfrac{1}{2} \sum_{i=1}^{6} \frac{\widehat{\odot}_i^{\,2}}{\sigma_i^2}\big)\, \mathrm{d}\sigma_1 \cdots \mathrm{d}\sigma_6. \tag{6.11.17}$$

For convenience, the professor writes

$$\phi(a, b, c \,|\, \widehat{\odot}_1, \ldots, \widehat{\odot}_6) = \alpha\, \frac{1}{(2\pi)^3} \frac{1}{\prod \Sigma_i} \times \Big( \prod \mathcal{J}_i \Big) \times \phi(a, b, c), \tag{6.11.18}$$

where

$$\mathcal{J}_i \equiv \int_0^{\Sigma_i} \frac{1}{\sigma_i}\, \exp\big( -\tfrac{1}{2} \frac{\widehat{\odot}_i^{\,2}}{\sigma_i^2}\big)\, \mathrm{d}\sigma_i. \tag{6.11.19}$$

Consulting with Harry's equation (6.9.18), the professor finds that

$$\mathcal{J}_i = \tfrac{1}{2}\, \mathrm{E}_1\big(\tfrac{1}{2} \frac{\widehat{\odot}_i^{\,2}}{\Sigma_i^2}\big), \tag{6.11.20}$$

where $\mathrm{E}_1$ is the exponential integral. Bayes equation provides us with the posterior probability

$$\phi(a, b, c \,|\, \widehat{\odot}_1, \ldots, \widehat{\odot}_6) = \alpha\, \frac{1}{(2\pi)^3} \times \frac{1}{2^6} \times \frac{1}{\prod \Sigma_i}$$
$$\times \prod \mathrm{E}_1\big(\tfrac{1}{2} \frac{\widehat{\odot}_i^{\,2}}{\Sigma_i^2}\big) \times \phi(a, b, c), \tag{6.11.21}$$

where all products are from $i = 1$ to $6$. The professor would rather not deal with the logarithmic singularity inherent with the exponential integrals.

### 6.11.10 Professional help

It turns out that the professor once met in a committee another prominent professor of Bayesian statistics. The professor explained the problem to the prominent Professor who thought that the answer is straightforward:

*Each sensor should record a time series of $N$ signals, and the root mean square (rms) value or the error should be used in the likelihood function.*

In practical terms, expression (6.11.4) should be replaced by

$$\phi_i(\widehat{\odot}_i) \equiv \mathcal{N}_{0,\sigma_i} = \frac{1}{(2\pi)^{N/2}\,\sigma_i^N}\ \exp(-\tfrac{1}{2}\,\frac{\widehat{\odot}^2_{\mathrm{rms}_i}}{\sigma_i^2}) \tag{6.11.22}$$

for $i = 1, \ldots, 6$, where

$$\widehat{\odot}_{\mathrm{rms}_i} \equiv \Big(\frac{1}{N}\sum_{j=1}^{N} \widehat{\odot}^2_{i_j}\Big)^{1/2} \tag{6.11.23}$$

is the rms deviation of the error. The sum is zero only when the signal is a perfect match of the predictions of the boundary-element code.

Bayes' equation becomes

$$\begin{aligned}\phi(a,b,c\,|\,\widehat{\odot}_{\mathrm{rms}_1},\ldots,\widehat{\odot}_{\mathrm{rms}_6}) &= \alpha \times \phi(a,b,c) \times \frac{1}{(2\pi)^{3N}} \\ \times \int_0^{\Sigma_1}\cdots\int_0^{\Sigma_6} &\frac{1}{\prod \sigma_i^N} \times \exp\big(-\tfrac{1}{2}\sum_{i=1}^{6}\frac{\widehat{\odot}^2_{\mathrm{rms}_i}}{\sigma_i^2}\big) \\ &\times\phi(\sigma_1,\ldots,\sigma_6)\,\mathrm{d}\sigma_1\cdots\mathrm{d}\sigma_6.\end{aligned} \tag{6.11.24}$$

In the event of a uniform prior for the standard deviations, the professor obtains

$$\begin{aligned}&\phi(a,b,c\,|\,\widehat{\odot}_{\mathrm{rms}_1},\ldots,\widehat{\odot}_{\mathrm{rms}_6}) \\ &\qquad = \alpha\,\frac{1}{(2\pi)^{3N}}\,\frac{1}{\prod\Sigma_i} \times \Big(\prod \mathcal{J}_i\Big) \times \phi(a,b,c),\end{aligned} \tag{6.11.25}$$

where

$$\mathcal{J}_i \equiv \int_0^{\Sigma_i} \frac{1}{\sigma_i^N} \exp\big(-\tfrac{N}{2}\,\frac{\widehat{\odot}_{\mathrm{rms}_i}^{\,2}}{\sigma_i^2}\big)\,\mathrm{d}\sigma_i. \tag{6.11.26}$$

Using Sue's equation (6.10.23), the professor finds that

$$\mathcal{J}_i = \tfrac{1}{2}\,\frac{1}{\widehat{\odot}_{\mathrm{rms}_i}^{\,N-1}} \int_{\widehat{\odot}_{\mathrm{rms}_i}^{\,2}/\Sigma_i^2}^{\infty} w_i^{(N-3)/2} \exp\big(-\tfrac{N}{2}\,w\big)\,\mathrm{d}w_i, \tag{6.11.27}$$

where

$$w_i \equiv \frac{\widehat{\odot}_{\mathrm{rms}_i}^{\,2}}{\sigma_i^2} \tag{6.11.28}$$

is an integration variable.

When $N \geq 2$, the professor may let the maximum value of the standard deviation, $\Sigma_i$, tend to infinity and obtain a convergent integral from $0$ to $\infty$ on the right-hand side. Bayes' equation then becomes

$$\phi(a,b,c\,|\,\widehat{\odot}_{\mathrm{rms}_1},\ldots,\widehat{\odot}_{\mathrm{rms}_6}) = \beta\,\Big(\prod \frac{1}{\widehat{\odot}_{\mathrm{rms}_i}^{\,N-1}}\Big) \times \phi(a,b,c), \tag{6.11.29}$$

where $\beta$ is a normalization factor. Sensible values for $a$, $b$, and $c$ arise at the maximum of the posterior pdf with respect to $a$, $b$, and $c$, for a sensible prior, $\phi(a,b,c)$.

### *6.11.11 Writing code*

A blood clot detection software code was written based on equation (6.11.29), and the compiled binary was loaded in a microchip embedded in the device. A important module is the predictive boundary-element code. encapsulated in the open-source software library FDLIB written by the author of this book. The device is currently under consideration for approval by the FDA.

### *6.11.12 Cousin Vinnie*

The professor's cousin Vinnie volunteers as a paramedic at the local hospital on Saturday evenings and Sunday mornings. On one occasion, Vinnie was explaining to a patient that the medical imaging machine uses software that employs the Galerkin finite-element method (FEM) to solve the equations of electromagnetics. The FEM is a special implementation of Ritz's method of weighted residuals based on a weak formulation of the governing equations (e.g., Pozrikidis, C. (2008) *Numerical Computation in Science and Engineering.* Second Edition, Oxford University Press).

Vinnie explained with pride to the patient that her cousin contributed to the implementation of the method in spite of annual admonitions from his Dean to "get on with the program" and submit to federal agencies collaborative applications for funding with other marginal relevant and poorly talented colleagues.

### 6.11.13 *Making a deal with God*

Before the results of the patient's medical imaging came in, the patient attempted to make a deal with God: if he survives the illness, he will write back to his favorite legislator outlining compelling reasons as to why (*a*) university administrators should be required to promote, let alone appreciate, substantive long-term research beyond the self-serving administrative measure of "research expenditures" and with the notions of quality, productivity, and accountability in mind; (*b*) food stamps should be given to anyone on an honor system (no questions asked); (*c*) affordable health care should be declared a human right, not a privilege.

Of course, deep in his heart, the patient knows that nobody has ever been able to make a deal with God.

#### *Exercise*

**6.11.1** Summarize the basic principles of the boundary-element method for viscous flow.

## *6.12 Big*

Mr. Horváth has nine apple trees, but not all apple trees are equal. Each tree produces sweet or sour apples, but the percentage varies from tree to tree according to how much sun and water each tree enjoys.

Mr. Horváth denotes the Bernoulli probability that a random apple that came from the $i$th tree is sweet as $\theta_i$ for $i = 1, \ldots, 9$, and introduces the relevant array

$$\boldsymbol{\theta} \equiv \big(\, \theta_1, \ldots, \theta_9 \,\big). \tag{6.12.1}$$

The probability that an apple that came from the $i$th tree is sour is $1 - \theta_i$. To be clear, $\theta_i$ is the fraction of sweet apples and $1 - \theta_i$ is the fraction of sour apples produced by the $i$th tree.

The probability density function of $\theta_i$ pertinent to the $i$th tree is denoted as

$$\phi_i(\theta_i) \tag{6.12.2}$$

for $i = 1, \ldots, 9$.

### *6.12.1 Beta distribution*

In the absence of data or insight, Mr. Horváth assumes that all nine pdfs are given by the same beta distribution,

$$\phi_i(\theta_i) = \mathfrak{Be}_{\alpha,\beta}(\theta_i), \tag{6.12.3}$$

where $\alpha$ and $\beta$ are assumed parameters common to all nine trees, as discussed in Section 5.1. In the absence of information, $\alpha$ and $\beta$ are described by a certain joint probability density function,

$$\phi(\alpha, \beta). \tag{6.12.4}$$

In this context, $\alpha$ and $\beta$ are regarded as hyperparameters.

Data must be collected to assess the nine pdfs shown in (6.12.2) and the joint pdf shown in (6.12.4) encapsulated in the extended pdf

$$\phi(\alpha, \beta, \boldsymbol{\theta}). \tag{6.12.5}$$

It is reasonable to assume that the performance of each tree is independent of the performance of every other tree and factorize

$$\phi(\alpha, \beta, \boldsymbol{\theta}) = \Big( \prod_{i=1}^{9} \phi_i(\theta_i) \Big) \times \phi(\alpha, \beta). \qquad (6.12.6)$$

At this point, Mr. Horváth's favorite hog Big comes into play.

### 6.12.2 Ahhh and argh

To generate data, Mr. Horváth collects apples from each tree and puts them in separate baskets labeled 1–9. Big samples apples from each basket and grunts *ahhh* if an apple is sweet or *argh* if an apple is sour.

Of $n_i$ apples collected from the $i$th tree, Big decided that $m_i$ apples are sweet and the remaining $n_i - m_i$ applies are sour. Mr. Horváth introduces the relevant $2 \times 9$ data matrix

$$\mathbf{x} \equiv \begin{bmatrix} n_1 & n_2 & \cdots & n_8 & n_9 \\ m_1 & m_2 & \cdots & m_8 & m_9 \end{bmatrix}, \qquad (6.12.7)$$

compliments of Big, puts the baskets in his truck, and drives them to his barn.

The probability that $m_i$ out of $n_i$ apples are sweet is expressed by the binomial distribution

$$\mathcal{B}_{m_i}^{n_i}(\theta_i) \equiv \binom{n_i}{m_i} \theta_i^{m_i} (1 - \theta_i)^{n_i - m_i}, \qquad (6.12.8)$$

for $i = 1, \ldots, 9$, with an unknown Bernoulli probability, $\theta_i$, described by the pdf shown in (6.12.2). The likelihood function for the union of all baskets is given by

$$\mathcal{L}_{\mathbf{x}}(\boldsymbol{\theta}) = \prod_{i=1}^{9} \mathcal{B}_{m_i}^{n_i}(\theta_i). \qquad (6.12.9)$$

An implied assumption underlying this factorization is that the performance of each tree is independent of the performance of every other tree.

### *6.12.3 Bayes equation*

Mr. Horváth applies Bayes' equation to write

$$\phi\big(\alpha, \beta, \boldsymbol{\theta}\,|\,\mathbf{x}\big) = \frac{1}{c}\,\Big(\prod_{i=1}^{9} \mathcal{B}_{m_i|n_i}(\theta_i)\Big)\Big(\prod_{i=1}^{9} \mathfrak{Be}_{\alpha,\beta}(\theta_i)\Big) \times \phi(\alpha, \beta), \tag{6.12.10}$$

where $\mathfrak{Be}_{\alpha,\beta}(\theta_i)$ is the beta distribution defined in (5.3.1) and $c$ is a normalization constant. Using the properties of the beta distribution, Mr. Horváth obtains

$$\phi\big(\alpha, \beta, \boldsymbol{\theta}\,|\,\mathbf{x}\big) = \Big(\prod_{i=1}^{9} \mathfrak{Be}_{\alpha+m_i,\beta+n_i-m_i}(\theta_i)\Big) \times \phi(\alpha, \beta). \tag{6.12.11}$$

In this context, $\alpha$ and $\beta$ are regarded as nuisance parameters.

### *6.12.4 Hierarchical formulation*

Equation (6.12.10) implements an hierarchical formulation. In this formulation, equation (6.12.11) is regarded as Bayes' equation for the pdf $\phi(\alpha, \beta)$, where the likelihood function, represented by the product on the right-hand side, is driven by all data encapsulated in $\mathbf{x}$. To see this, we write

$$\phi\big(\alpha, \beta, \boldsymbol{\theta}\,|\,\mathbf{x}\big) \sim \mathcal{L}_{\mathbf{x}}(\boldsymbol{\theta}) \times \phi(\boldsymbol{\theta}\,|\,\alpha, \beta) \times \phi(\alpha, \beta), \tag{6.12.12}$$

where

$$\mathcal{L}_{\mathbf{x}}(\boldsymbol{\theta}) = \prod_{i=1}^{9} \theta_i^{m_i} \times (1 - \theta_i)^{n_i - m_i} \tag{6.12.13}$$

is the likelihood function,

$$\phi(\boldsymbol{\theta}\,|\,\alpha, \beta) = \prod_{i=1}^{9} \mathfrak{Be}_{\alpha,\beta}(\theta_i) \tag{6.12.14}$$

is the prior, and $\phi(\alpha, \beta)$ is the hyperprior. Other likelihood functions, priors, and hyper-priors can be chosen in different applications.

In previous sections, we saw that hyperparameters are routinely defined with reference to the likelihood function. For example, the standard deviation of the Gaussian distribution can be regarded as unknown that can be estimated as part of the solution. In the present context, a hyperprior pdf is defined with reference to a prior pdf that is unrelated to the likelihood function.

***Exercise***

**6.12.1** Discuss a hierarchical formulation of your choice.

# Appendix A

# Combinatorics

Probability theory and Bayesian analysis make extensive usage of combinatorics whose main goal to provide formulas for the number of possible combinations of specified occurrences or events.

### *A.1 The factorial*

The factorial of an integer, $n$, is an integer denoted by an exclamation mark, denoted as

$$n! \equiv 1 \cdot 2 \cdots n = \prod_{k=1}^{n}, \tag{A.15}$$

subject to the convention that $0! = 1$. For large $n$, the factorial can be approximated with Stirling's formula

$$n! \simeq \sqrt{2\pi n} \left(\frac{n}{e}\right)^n. \tag{A.16}$$

The formula is remarkably accurate as it introduces an error on the order of only $2\%$ for $n$ as low as $4$.

### *A.2 Gamma function*

The factorial of an integer, $n$, can be expressed in terms of the Gamma function, $\Gamma$, as

$$n! = \Gamma(n+1). \tag{A.17}$$

The Gamma function is defined as a definite integral over the positive part of the real axis,

$$\Gamma(\xi) \equiv \int_0^{\infty} t^{\xi-1} e^{-t} \, \mathrm{d}t, \tag{A.18}$$

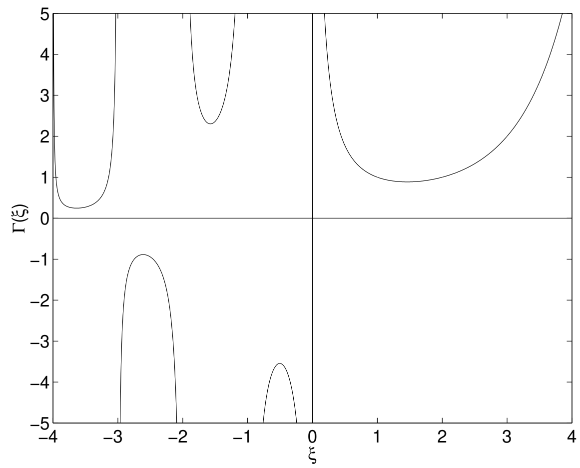



FIGURE A.1 Graph of the Gamma function. Note that $\Gamma(n+1) = n!$ for zero or any integer, $n$.

for any value, $\xi$. For our purposes, we assume that the independent parametric variable, $\xi$, is real. A graph of the Gamma function is shown in Figure 7.0.1. Note that singularities occur when $\xi$ is zero or a negative integer.

Relation (A.17) can be used to define the factorial of any real, even negative, number, $\alpha$,

$$\alpha\,! = \Gamma(\alpha + 1). \tag{A.19}$$

This extension to non-integers provides us with a bridge to fractional calculus.

### *A.3 Combinations*

In combinatorics, the factorial of an integer, $n\,!$, expresses the number of ways by which $n$ distinct objects can be arranged in an $n$-dimensional array, where each way represents a permutation.

The proof can be carried out by mathematical induction. For example, when $n = 2$ with objects labelled $\mathrm{A}$ and $\mathrm{B}$, we obtain two permutations, $\mathrm{AB}$ and $\mathrm{BA}$, which is consistent with the value $2\,! = 2$.

### *A.4 Permutations*

Dividing the factorial of an integer by selected groups of permutations, we obtain the number of constrained permutations.

The number of permutations of $n$ objects consisting of a set of $m_1$ indistinguishable objects, another set of $m_2$ indistinguishable objects, and a third set of $m_3$ indistinguishable objects, is given by

$$\frac{n\,!}{m_1\,!\, m_2\,!\, m_3\,!}, \tag{A.20}$$

where $m_1 + m_2 + m_3 = n$. A similar expression can be written for a higher number of sets consisting of indistinguishable objects.

### *A.5 The binomial coefficient*

The binomial coefficient is an integer denoted and defined as

$$\mathcal{C}_m^n \equiv \binom{n}{m} \equiv \frac{n\,!}{m\,!\,(n-m)\,!} \tag{A.21}$$

for any two integers $n$ and $m$, where $m \le n$. Substituting the definition of the factorial, we find that

$$\mathcal{C}_m^n = \frac{(n-m+1)\cdots(n-1)\,n}{m\,!} = \prod_{k=1}^{m} \frac{n-k+1}{k}, \tag{A.22}$$

which shows that $\mathcal{C}_m^n$ is an $m$th-degree polynomial in $n$.

The binomial coefficient expresses the number of possible combinations by which $m$ objects can be chosen from a set of $n$ identical objects, leaving $n-m$ objects behind, where the order by which the objects are chosen is immaterial.

For example, for $n = 3$ with objects labeled $\mathrm{A}$, $\mathrm{B}$, and $\mathrm{C}$, and $m = 2$, we have three combinations $\mathrm{AB}$, $\mathrm{AC}$, $\mathrm{BC}$, where $\mathrm{BA}$ is not included since it is considered the same as $\mathrm{AB}$.

### *A.6 The factorial symmetry*

Referring to the definition of the binomial coefficient, we find that

$$\mathcal{C}_m^n = \mathcal{C}_{n-m}^n. \tag{A.23}$$

This formula shows that the binomial coefficient is symmetric with respect to the mid-point, $m = n/2$.

### A.7 Pascal triangle

The binomial coefficients for a certain $n$ can be arranged at the rows of the Pascal triangle with $n = 0$ at the apex:

$$\begin{array}{ccccccccccc} & & & & & 1 & & & & & \\ & & & & 1 & & 1 & & & & \\ & & & 1 & & 2 & & 1 & & & \\ & & 1 & & 3 & & 3 & & 1 & & \\ & 1 & & 4 & & 6 & & 4 & & 1 & \\ 1 & & 5 & & 10 & & 10 & & 5 & & 1 \\ \dots & & \dots & & \dots & & \dots & & \dots & & \dots \end{array} \tag{A.24}$$

The second row corresponds to $n = 1$ and the third row corresponds to $n = 2$.

The elements of each row correspond to $m = 0, \dots, n$, from left to right. Each element in this triangle is the sum of the two entries immediately above, and the outermost elements are equal to unity. The symmetry property (A.23) is apparent.

### A.8 Properties

It can be shown by mathematical induction that the binomial coefficient satisfies the recursion formula

$$\mathcal{C}_m^n = \mathcal{C}_{m-1}^{n-1} + \mathcal{C}_m^{n-1}, \tag{A.25}$$

for appropriate integers, $m$ and $n$, which may also be expressed as

$$\mathcal{C}_m^n + \mathcal{C}_{m+1}^n = \mathcal{C}_{m+1}^{n+1}, \tag{A.26}$$

for appropriate integers, $m$ and $n$.

### A.9 Binomial expansion

The binomial coefficient provides us with a systematic way of computing the $n$th power of the sum of two arbitrary numbers, $a$ and $b$, in terms of the binomial expansion

$$(a+b)^n = \sum_{m=0}^{n} \mathcal{C}_m^n \, a^{n-m} \, b^m = \sum_{m=0}^{n} \binom{n}{m} \, a^{n-m} \, b^m, \tag{A.27}$$

involving $n+1$ summed terms. For $n=2$, we obtain the well-known quadratic expansion

$$(a+b)^2 = a^2 + 2\,ab + b^2. \tag{A.28}$$

The coefficients, $1, 2, 1$ can be found at the second row of the Pascal triangle discussed later in this section.

Setting $b = 1-a$, we obtain the partition-of-unity expansion

$$1 = \sum_{m=0}^{n} \mathcal{C}_m^n \, a^{n-m} \, (1-a)^m = \sum_{m=0}^{n} \binom{n}{m} \, a^{n-m} \, (1-a)^m, \tag{A.29}$$

for any $a$ and $n$.

### *A.10 Identities*

Taking the partial derivative of (A.27) with respect to $b$, we obtain

$$n\,(a+b)^{n-1} = \sum_{m=1}^{n} m\,\mathcal{C}_m^n \, a^{n-m} \, b^{m-1}. \tag{A.30}$$

For $a=1$ and $b=1$, we derive the identity

$$\frac{1}{2^{n-1}} \sum_{m=1}^{n} m\,\mathcal{C}_m^n = n. \tag{A.31}$$

Now taking the partial derivative of (A.30) with respect to $b$, we obtain

$$n\,(n-1)\,(a+b)^{n-2} = \sum_{m=2}^{n} m\,(m-1)\,\mathcal{C}_m^n \, a^{n-m} \, b^{m-2}. \tag{A.32}$$

For $a=1$ and $b=1$, we derive the identity

$$\frac{1}{2^{n-2}} \sum_{m=2}^{n} m\,(m-1)\,\mathcal{C}_m^n = n\,(n-1). \tag{A.33}$$

Rearranging, we obtain

$$\frac{1}{2^{n-2}} \sum_{m=1}^{n} m^2 \, \mathcal{C}_m^n = \frac{1}{2^{n-2}} \sum_{m=1}^{n} m \, \mathcal{C}_m^n + n^2 - n. \tag{A.34}$$

Using (A.31) to evaluate the sum on the right-hand side, we derive the identity

$$\frac{1}{2^{n-2}} \sum_{m=1}^{n} m^2 \, \mathcal{C}_m^n = n^2 + n. \tag{A.35}$$

Further partial derivatives of (A.27) with respect to $b$ can be taken to derive further identities.

### *A.11 Product differentiation*

The $n$th derivative of the product of two functions, $f$ and $g$, is given by

$$(fg)^{(n)} = \sum_{m=0}^{n} \mathcal{C}_m^n \, f^{(n-m)} \, g^{(m)}. \tag{A.36}$$

When either $f$ or $g$ is a constant, only one term in the sum survives.

### *A.12 Numerical evaluation*

In practice, the binomial coefficient can be computed from the expression

$$\mathcal{C}_m^n = \prod_{k=1}^{m} \frac{n-k+1}{k} = \prod_{k=1}^{k_{\max}} \frac{n-k+1}{k}, \tag{A.37}$$

where $k_{\max}$ is the minimum of $m$ and $n-m$.

The following Matlab function computes the binomial coefficient $\mathcal{C}_m^n$ based on formula (A.37):

```
function Cnm = binomial_coeff(n,m)

 kmax = m;
```

```
 if((n-m)<kmax) kmax = n-m; end

 Cnm = 1.0;
 for k=1:kmax
  Cnm = Cnm*(n-k+1)/k;
 end

return
```

### A.13 Fractional binomial coefficient

A fractional binomial coefficient can be defined terms of the Gamma function,

$$\mathcal{C}^{\alpha}_{\beta} \equiv \binom{\alpha}{\beta} = \frac{\Gamma(\alpha+1)}{\Gamma(\beta+1)\,\Gamma(\alpha-\beta+1)} \tag{A.38}$$

for any two real numbers, $\alpha$ and $\beta$.

When $\beta$ is an integer, $m$, the fractional binomial coefficient is given by

$$\mathcal{C}^{\alpha}_{m} = \frac{(\alpha-m+1)\cdots(\alpha-1)\,\alpha}{1\cdot 2\cdots m} = \prod_{k=1}^{m} \frac{\alpha-k+1}{j} \tag{A.39}$$

for any real positive number, $\alpha$ and any integer $m \geq 1$, subject to the convention that $\mathcal{C}^{\alpha}_{0} = 1$. Using the properties of the Gamma function, we find that

$$\mathcal{C}^{\alpha}_{m} = (-1)^m \frac{\Gamma(m-\alpha)}{\Gamma(-\alpha)\,\Gamma(m+1)}. \tag{A.40}$$

Given this definition, we may write the generalized binomial expansion to an non-integer exponents,

$$(a+b)^{\alpha} = \sum_{m=0}^{n} \mathcal{C}^{\alpha}_{m}\, a^{\alpha-m}\, b^m = \sum_{m=0}^{\alpha} \mathcal{C}^{\alpha}_{m}\, a^{\alpha-m}\, b^m, \tag{A.41}$$

involving an infinite number of summed terms. For $b = 1-a$ we obtain a partition-of-unity equation.

***Exercises***

**A.1** A random drawing will be held where $m$ out of $n$ people in an audience on the merits of time-share condos will be given tickets for a weekend trip to Las Vegas, where $m \leq n$. Explain why the probability that one particular person in the audience will receive a ticket is

$$\frac{\mathcal{C}_{m-1}^{n-1}}{\mathcal{C}_m^n} = \frac{m}{n}. \tag{A.42}$$

What is the probability that one particular person and the person sitting on his/her right receive tickets?

**A.2** Confirm that the binomial coefficient satisfies the property (A.25).

**A.3** Derive an identity by taking the third partial derivative of (A.27) with respect to $b$.

# Appendix B

# Quick reference

Bayes' rule takes a variety of forms depending on the definition of the necessary sample spaces and whether discrete or continuous events or data are considered.

### 1. A pair of events

For any two events, denoted by $\circ$ and $\oplus$, Bayes' rule prescribes that

$$\Pi_{(\circ\,|\,\oplus)} = \frac{\Pi_{(\oplus\,|\,\circ)}}{\Pi_{(\oplus)}} \times \Pi_{(\circ)},$$

where $\oplus$ is a specified event regarded as data,

$$\Pi_{(\oplus)} = \Pi_{(\oplus\,|\,\circ)} \times \Pi_{(\circ)} + \Pi_{(\oplus\,|\,\bar{\circ})} \times \Pi_{(\bar{\circ})}$$

is a marginal probability, and $\bar{\circ}$ is the complement of $\circ$ in its sample space.

### 2. Expansion over a sample space

Let the following $N$ mutually exclusive events define a sample space,

$$\circ_1, \quad \circ_2, \quad \ldots, \quad \circ_N.$$

One of the events *has* to occur, so that the associated probabilities add to unity,

$$\sum_{i=1}^{N} \Pi_{(\circ_i)} = 1.$$

Since

$$\Pi_{(\oplus)} = \sum_{j=1}^{N} \Pi_{(\oplus\,|\,\circ_j)} \times \Pi(\circ_j),$$

the corresponding Bayes rule takes the form

$$\Pi_{(\circ_i \,|\, \oplus)} = \frac{\Pi_{(\oplus \,|\, \circ_i)}}{\sum_{j=1}^{N} \Pi_{(\oplus \,|\, \circ_j)} \times \Pi(\circ_j)} \times \Pi_{(\circ_i)}$$

for $i = 1, \ldots, N$. The likelihoods of the data are defined as and

$$\mathcal{L}_{\oplus}(\circ_i) \equiv \Pi_{(\oplus \,|\, \circ_i)}.$$

In terms of the likelihoods, Bayes rule takes the form

$$\Pi_{(\circ_i \,|\, \oplus)} = \frac{\Pi_{(\oplus \,|\, \circ_i)}}{\sum_{j=1}^{N} \mathcal{L}_{\oplus}(\circ_j) \times \Pi(\circ_j)} \times \Pi_{(\circ_i)}$$

for $i = 1, \ldots, N$.

## 3. Continuous data

Let an event $\oplus$ depend on a continuous random variable, $x$. It is convenient to denote

$$\Pi_{(\circ_i)}(x) \equiv \Pi_{(\circ_i \,|\, \oplus)}$$

for $i = 1, \ldots, N$.

Now we chose a small interval, $\mathrm{d}x$, and write

$$\Pi_{(\oplus)}(x) = \psi_{\oplus}(x)\,\mathrm{d}x$$

and

$$\Pi_{(\oplus \,|\, \circ_i)}(x) = \psi_i(x)\,\mathrm{d}x,$$

where $\psi_{\oplus}(x)$ and $\psi_i(x)$ are probability density functions (pdfs). Invoking the law of total probability, we write

$$\psi_{\oplus}(x) = \sum_{j=1}^{N} \psi_j(x) \times \Pi_{(\circ_j)}.$$

The corresponding Bayes rule takes the form

$$\Pi_{(\circ_i)}(x) = \frac{\psi_i(x)}{\psi_{\oplus}(x)} \times \Pi_{(\circ_i)}$$

involving discrete probabilities and pdfs, where $x$ is regarded as data, $\mathcal{L}_x(\circ_i) = \psi_i(x)$ are conditional densities of the likelihoods of the data, and $\mathcal{L}_x(\oplus) = \psi_\oplus(x)$ is the likelihood probability density of the data.

## 4. Continuous events

Bayes' equation for continuous events parametrized by a random variable $\theta$ takes the form

$$\phi(\theta \,|\, x) = \frac{\mathcal{L}_x(\theta)}{\displaystyle\int \mathcal{L}_x(\theta') \times \phi(\theta') \, \mathrm{d}\theta'} \times \phi(\theta),$$

where $x$ is discrete or continuous data, $\mathcal{L}_x(\theta)$ is the likelihood of the data, and $\phi(\theta)$ is a probability density function.

## 5. Continuous events in high dimensions

Bayes' equation for an arbitrary number of continuous variables takes the form

$$\phi(\boldsymbol{\theta} \,|\, \mathbf{x}) = \frac{\mathcal{L}_{\mathbf{x}}(\boldsymbol{\theta})}{\int \cdots \int \mathcal{L}_{\mathbf{x}}(\boldsymbol{\theta}') \, \phi(\boldsymbol{\theta}') \, \mathrm{d}V(\boldsymbol{\theta}')} \times \phi(\boldsymbol{\theta}),$$

where the array $\mathbf{x}$ encapsulates observed or measured, discrete or continuously varying data, $\mathrm{d}V$ is a differential volume in the $\boldsymbol{\theta}$ space, and $\boldsymbol{\theta}'$ is a multicomponent integration variable.

# Index

# *Zen and the Art of Bayesian Analysis*

## *C. Pozrikidis*

Bayes theorem and its generalizations allow us to infer or deduce the probability of an event, proposition, theory or quantifiable cause, given an observed consequence or measured effect. The theorem is especially useful when a consequence or effect may be due to a multitude of causes. In this concise book, the author introduces Bayes' theorem from the ground up and discusses methods of Bayesian analysis in a broad range of applications. Familiarity with elementary mathematics is assumed in most early sections, while college-level mathematics is required in more advanced sections.

Key features include the following:

- Bayes theorem is introduced from the ground up following the introduction and interpretation of uncertainty quantified by probability.
- A broad range of real-life and scientific applications are discussed.
- The reader may select the tolerated level of mathematical comfort and skip sections that appear too mathematical without compromising the understanding of subsequent material.
- All necessary concepts are defined and introduced for a self-contained discourse; elementary background information from combinatorics is provided in an appendix.
- Matlab codes performing computations, simulations, and visualization are listed and discussed.

Chester & Bennington