

# 1

## Role of probability theory in science

### 1.1 Scientific inference

This book is primarily concerned with the philosophy and practice of inferring the laws of nature from experimental data and prior information. The role of inference in the larger framework of the scientific method is illustrated in Figure 1.1.

In this simple model, the scientific method is depicted as a loop which is entered through initial observations of nature, followed by the construction of testable hypotheses or theories as to the working of nature, which give rise to the prediction of other properties to be tested by further experimentation or observation. The new data lead to the refinement of our current theories, and/or development of new theories, and the process continues.

The role of deductive inference<sup>1</sup> in this process, especially with regard to deriving the testable predictions of a theory, has long been recognized. Of course, any theory makes certain assumptions about nature which are assumed to be true and these assumptions form the axioms of the deductive inference process. The terms deductive inference and deductive reasoning are considered equivalent in this book. For example, Einstein's Special Theory of Relativity rests on two important assumptions; namely, that the vacuum speed of light is a constant in all inertial reference frames and that the laws of nature have the same form in all inertial frames.

Unfortunately, experimental tests of theoretical predictions do not provide simple yes or no answers. Our state of knowledge is always incomplete, there are always more experiments that could be done and the measurements are limited in their accuracy. Statistical inference is the process of inferring the truth of our theories of nature on the basis of the incomplete information. In science we often make progress by starting with simple models. Usually nature is more complicated and we learn in what direction to modify our theories from the differences between the model predictions and the measurements. It is much like peeling off layers of an onion. At any stage in this iterative process, the still hidden layers give rise to differences from the model predictions which guide the next step.

<sup>1</sup> Reasoning from one proposition to another using the strong syllogisms of logic (see Section 2.2.4).

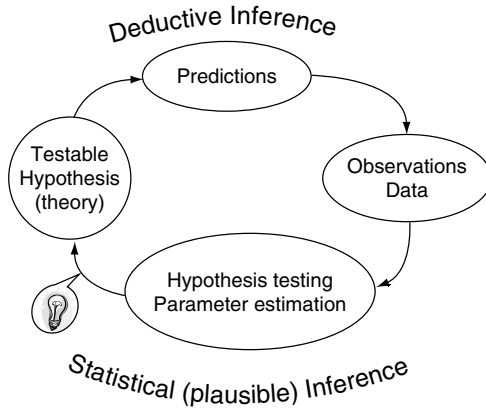


Figure 1.1 The scientific method.

### 1.2 Inference requires a probability theory

In science, the available information is always incomplete so our knowledge of nature is necessarily probabilistic. Two different approaches based on different definitions of probability will be considered. In conventional statistics, the probability of an event is identified with the long-run relative frequency of occurrence of the event. This is commonly referred to as the “frequentist” view. In this approach, probabilities are restricted to a discussion of random variables, quantities that can meaningfully vary throughout a series of repeated experiments. Two examples are:

1. A measured quantity which contains random errors.
2. Time intervals between successive radioactive decays.

The role of random variables in frequentist statistics is detailed in Section 5.2.

In recent years, a new perception of probability has arisen in recognition that the mathematical rules of probability are not merely rules for calculating frequencies of random variables. They are now recognized as uniquely valid principles of logic for conducting inference about any proposition or hypothesis of interest. This more powerful viewpoint, “Probability Theory as Logic,” or Bayesian probability theory, is playing an increasingly important role in physics and astronomy. The Bayesian approach allows us to directly compute the probability of any particular theory or particular value of a model parameter, issues that the conventional statistical approach can attack only indirectly through the use of a random variable statistic. In this book, I adopt the approach which exposes probability theory as an extended theory of logic following the lead of E. T. Jaynes in his book,<sup>2</sup> *Probability Theory* –

<sup>2</sup> The book was finally submitted for publication four years after his death, through the efforts of his former student G. Larry Bretthorst.

Table 1.1 *Frequentist and Bayesian approaches to probability.*

Approach	Probability definition
FREQUENTIST STATISTICAL INFERENCE	$p(A)$ = long-run relative frequency with which $A$ occurs in identical repeats of an experiment. “ $A$ ” restricted to propositions about random variables.
BAYESIAN INFERENCE	$p(A B)$ = a real number measure of the plausibility of a proposition/hypothesis $A$ , given (conditional on) the truth of the information represented by proposition $B$ . “ $A$ ” can be any logical proposition, <i>not</i> restricted to propositions about random variables.

*The Logic of Science* (Jaynes, 2003). The two approaches employ different definitions of probability which must be carefully understood to avoid confusion.

The two different approaches to statistical inference are outlined in Table 1.1 together with their underlying definition of probability. In this book, we will be primarily concerned with the Bayesian approach. However, since much of the current scientific culture is based on “frequentist” statistical inference, some background in this approach is useful.

The frequentist definition contains the term “identical repeats.” Of course the repeated experiments can never be identical in all respects. The Bayesian definition of probability involves the rather vague sounding term “plausibility,” which must be given a precise meaning (see Chapter 2) for the theory to provide quantitative results. In Bayesian inference, a probability distribution is an encoding of our uncertainty about some model parameter or set of competing theories, based on our current state of information. The approach taken to achieve an operational definition of probability, together with consistent rules for manipulating probabilities, is discussed in the next section and details are given in Chapter 2.

In this book, we will adopt the plausibility definition<sup>3</sup> of probability given in Table 1.1 and follow the approach pioneered by E. T. Jaynes that provides for a unified picture of both deductive and inductive logic. In addition, Jaynes brought

<sup>3</sup> Even within the Bayesian statistical literature, other definitions of probability exist. An alternative definition commonly employed is the following: “probability is a measure of the degree of belief that any well-defined proposition (an event) will turn out to be true.” The events are still random variables, but the term is generalized so it can refer to the distribution of results from repeated measurements, or, to possible values of a physical parameter, depending on the circumstances. The concept of a coherent bet (e.g., D’Agostini, 1999) is often used to define the value of probability in an operational way. In practice, the final conditional posteriors are the same as those obtained from the extended logic approach adopted in this book.

great clarity to the debate on objectivity and subjectivity with the statement, “the only thing objectivity requires of a scientific approach is that experimenters with the same state of knowledge reach the same conclusion.” More on this later.

### ***1.2.1 The two rules for manipulating probabilities***

It is now routine to build or program a computer to execute deductive logic. The goal of Bayesian probability theory as employed in this book is to provide an extension of logic to handle situations where we have incomplete information so we may arrive at the relative probabilities of competing hypotheses for a given state of information.

Cox and Jaynes showed that the desired extension can be arrived at uniquely from three “desiderata” which will be introduced in Section 2.5.1. They are called “desiderata” rather than axioms because they do not assert that anything is “true,” but only state desirable goals of a theory of plausible inference.

The operations for manipulating probabilities that follow from the desiderata are the sum and product rules. Together with the Bayesian definition of probability, they provide the desired extension to logic to handle the common situation of incomplete information. We will simply state these rules here and leave their derivation together with a precise operational definition of probability to the next chapter.

$$\text{Sum Rule: } p(A|B) + p(\bar{A}|B) = 1 \quad (1.1)$$

$$\begin{aligned} \text{Product Rule: } p(A, B|C) &= p(A|C)p(B|A, C) \\ &= p(B|C)p(A|B, C), \end{aligned} \quad (1.2)$$

where the symbol  $A$  stands for a proposition which asserts that something is true. The symbol  $B$  is a proposition asserting that something else is true, and similarly,  $C$  stands for another proposition. Two symbols separated by a comma represent a compound proposition which asserts that both propositions are true. Thus  $A, B$  indicates that both propositions  $A$  and  $B$  are true and  $p(A, B|C)$  is commonly referred to as the joint probability. Any proposition to the right of the vertical bar  $|$  is assumed to be true. Thus when we write  $p(A|B)$ , we mean the probability of the truth of proposition  $A$ , given (conditional on) the truth of the information represented by proposition  $B$ .

#### **Examples of propositions:**

$A \equiv$  “The newly discovered radio astronomy object is a galaxy.”

$B \equiv$  “The measured redshift of the object is  $0.150 \pm 0.005$ .”

$A \equiv$  “Theory  $X$  is correct.”

$\bar{A} \equiv$  “Theory  $X$  is not correct.”

$A \equiv$  “The frequency of the signal is between  $f$  and  $f + df$ .”

We will have much more to say about propositions in the next chapter.

Cambridge University Press

052184150X - Bayesian Logical Data Analysis for the Physical Sciences: A Comparative Approach with Mathematica™ Support

P. C. Gregory

Excerpt

[More information](#)

## 1.3 Usual form of Bayes' theorem

5

*Bayes' theorem* follows directly from the product rule (a rearrangement of the two right sides of the equation):

$$p(A|B, C) = \frac{p(A|C)p(B|A, C)}{p(B|C)}. \quad (1.3)$$

Another version of the sum rule can be derived (see Equation (2.23)) from the product and sum rules above:

$$\text{Extended Sum Rule: } p(A + B|C) = p(A|C) + p(B|C) - p(A, B|C), \quad (1.4)$$

where  $A + B \equiv$  proposition  $A$  is true or  $B$  is true or both are true. If propositions  $A$  and  $B$  are mutually exclusive – only one can be true – then Equation (1.4) becomes

$$p(A + B|C) = p(A|C) + p(B|C). \quad (1.5)$$

## 1.3 Usual form of Bayes' theorem

$$p(H_i|D, I) = \frac{p(H_i|I)p(D|H_i, I)}{p(D|I)}, \quad (1.6)$$

where  $H_i \equiv$  proposition asserting the truth of a hypothesis of interest

$I \equiv$  proposition representing our prior information

$D \equiv$  proposition representing data

$p(D|H_i, I)$  = probability of obtaining data  $D$ , if  $H_i$  and  $I$  are true  
(also called the likelihood function  $\mathcal{L}(H_i)$ )

$p(H_i|I)$  = prior probability of hypothesis

$p(H_i|D, I)$  = posterior probability of  $H_i$

$$p(D|I) = \sum_i p(H_i|I)p(D|H_i, I)$$

(normalization factor which ensures  $\sum_i p(H_i|D, I) = 1$ ).

## 1.3.1 Discrete hypothesis space

In Bayesian inference, we are interested in assigning probabilities to a set of competing hypotheses perhaps concerning some aspect of nature that we are studying. This set of competing hypotheses is called the *hypothesis space*. For example, a problem of current interest to astronomers is whether the expansion of the universe is accelerating or decelerating. In this case, we would be dealing with a discrete hypothesis

Cambridge University Press

052184150X - Bayesian Logical Data Analysis for the Physical Sciences: A Comparative Approach with Mathematica™ Support

P. C. Gregory

Excerpt

[More information](#)

6

*Role of probability theory in science*

space<sup>4</sup> consisting of  $H_1$  ( $\equiv$  accelerating) and  $H_2$  ( $\equiv$  decelerating). For a discrete hypothesis space,  $p(H_i|D, I)$  is called a *probability distribution*. Our posterior probabilities for  $H_1$  and  $H_2$  satisfy the condition that

$$\sum_{i=1}^2 p(H_i|D, I) = 1. \quad (1.7)$$

### 1.3.2 Continuous hypothesis space

In another type of problem we might be dealing with a hypothesis space that is continuous. This can be considered as the limiting case of an arbitrarily large number of discrete propositions.<sup>5</sup> For example, we have strong evidence from the measured velocities and distances of galaxies that we live in an expanding universe. Astronomers are continually seeking to refine the value of Hubble's constant,  $H_0$ , which relates the recession velocity of a galaxy to its distance. Estimating  $H_0$  is called a parameter estimation problem and in this case, our hypothesis space of interest is continuous. In this case, the proposition  $H_0$  asserts that the true value of Hubble's constant is in the interval  $h$  to  $h + dh$ . The truth of the proposition can be represented by  $p(H_0|D, I)dH$ , where  $p(H_0|D, I)$  is a *probability density function* (PDF). The probability density function is defined by

$$p(H_0|D, I) = \lim_{\delta h \rightarrow 0} \frac{p(h \leq H_0 < h + \delta h|D, I)}{\delta h}. \quad (1.8)$$

#### Box 1.1 Note about notation

The term "PDF" is also a common abbreviation for *probability distribution function*, which can pertain to discrete or continuous sets of probabilities. This term is particularly useful when dealing with a mixture of discrete and continuous parameters.

We will use the same symbol,  $p(\dots)$ , for probabilities and PDFs; the nature of the argument will identify which use is intended. To arrive at a final numerical answer for the probability or PDF of interest, we eventually need to convert the terms in Bayes' theorem into algebraic expressions, but these expressions can become very complicated in appearance. It is useful to delay this step until the last possible moment.

<sup>4</sup> Of course, nothing guarantees that future information will not indicate that the correct hypothesis is outside the current working hypothesis space. With this new information, we might be interested in an expanded hypothesis space.

<sup>5</sup> In Jaynes (2003), there is a clear warning that difficulties can arise if we are not careful in carrying out this limiting procedure explicitly. This is often the underlying cause of so-called paradoxes of probability theory.

Cambridge University Press

052184150X - Bayesian Logical Data Analysis for the Physical Sciences: A Comparative Approach with Mathematica™ Support

P. C. Gregory

Excerpt

[More information](#)

## 1.3 Usual form of Bayes' theorem

7

Let  $W$  be a proposition asserting that the numerical value of  $H_0$  lies in the range  $a$  to  $b$ . Then

$$p(W|D, I) = \int_a^b p(H_0|D, I) dH_0. \quad (1.9)$$

In the continuum limit, the normalization condition of Equation (1.7) becomes

$$\int_{\Delta H} p(H|D, I) dH = 1, \quad (1.10)$$

where  $\Delta H$  designates the range of integration corresponding to the hypothesis space of interest.

We can also talk about a *joint probability distribution*,  $p(X, Y|D, I)$ , in which both  $X$  and  $Y$  are continuous, or, one is continuous and the other is discrete. If both are continuous, then  $p(X, Y|D, I)$  is interpreted to mean

$$p(X, Y|D, I) = \lim_{\delta x, \delta y \rightarrow 0} \frac{p(x \leq X < x + \delta x, y \leq Y < y + \delta y|D, I)}{\delta x \delta y}. \quad (1.11)$$

In a well-posed problem, the prior information defines our hypothesis space, the means for computing  $p(H_i|I)$ , and the likelihood function given some data  $D$ .

## 1.3.3 Bayes' theorem – model of the learning process

Bayes' theorem provides a model for inductive inference or the learning process. In the parameter estimation problem of the previous section,  $H_0$  is a continuous hypothesis space. Hubble's constant has some definite value, but because of our limited state of knowledge, we cannot be too precise about what that value is. In all Bayesian inference problems, we proceed in the same way. We start by encoding our prior state of knowledge into a prior probability distribution,  $p(H_0|I)$  (in this case a density distribution). We will see a very simple example of how to do this in Section 1.4.1, and many more examples in subsequent chapters. If our prior information is very vague then  $p(H_0|I)$  will be very broad, spanning a wide range of possible values of the parameter.

It is important to realize that a Bayesian PDF is a **measure of our state of knowledge** (i.e., ignorance) of the value of the parameter. The actual value of the parameter is not distributed over this range; it has some definite value. This can sometimes be a serious point of confusion, because, in frequentist statistics, the argument of a probability is a random variable, a quantity that can meaningfully take on different values, and these values correspond to possible outcomes of experiments.

We then acquire some new data,  $D_1$ . Bayes' theorem provides a means for combining what the data have to say about the parameter, through the likelihood function, with our prior, to arrive at a posterior probability density,  $p(H_0|D_1, I)$ , for the parameter.

$$p(H_0|D_1, I) \propto p(H_0|I_0)p(D_1|H_0, I). \quad (1.12)$$

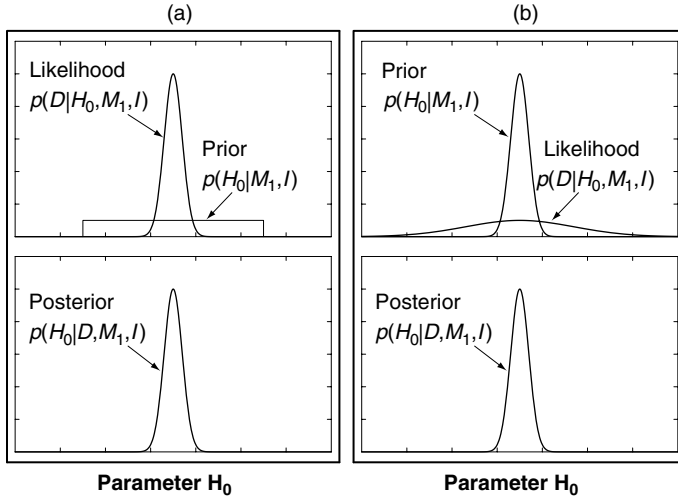


Figure 1.2 Bayes' theorem provides a model of the inductive learning process. The posterior PDF (lower graphs) is proportional to the product of the prior PDF and the likelihood function (upper graphs). This figure illustrates two extreme cases: (a) the prior much broader than likelihood, and (b) likelihood much broader than prior.

Two extreme cases are shown in Figure 1.2. In the first, panel (a), the prior is much broader than the likelihood. In this case, the posterior PDF is determined entirely by the new data. In the second extreme, panel (b), the new data are much less selective than our prior information and hence the posterior is essentially the prior.

Now suppose we acquire more data represented by proposition  $D_2$ . We can again apply Bayes' theorem to compute a posterior that reflects our new state of knowledge about the parameter. This time our new prior,  $I'$ , is the posterior derived from  $D_1, I$ , i.e.,  $I' = D_1, I$ . The new posterior is given by

$$p(H_0|D_2, I') \propto p(H_0|I')p(D_2|H_0, I'). \quad (1.13)$$

### 1.3.4 Example of the use of Bayes' theorem

Here we analyze a simple model comparison problem using Bayes' theorem. We start by stating our prior information,  $I$ , and the new data,  $D$ .

$I$  stands for:

- Model  $M_1$  predicts a star's distance,  $d_1 = 100$  light years (ly).
- Model  $M_2$  predicts a star's distance,  $d_2 = 200$  ly.
- The uncertainty,  $e$ , in distance measurements is described by a Gaussian distribution of the form



Cambridge University Press

052184150X - Bayesian Logical Data Analysis for the Physical Sciences: A Comparative Approach with Mathematica™ Support

P. C. Gregory

Excerpt

[More information](#)

## 1.3 Usual form of Bayes' theorem

9

$$p(e|I) = \frac{1}{\sqrt{2\pi} \sigma} \exp\left(-\frac{e^2}{2\sigma^2}\right),$$

where  $\sigma = 40$  ly.

d) There is no current basis for preferring  $M_1$  over  $M_2$  so we set  $p(M_1|I) = p(M_2|I) = 0.5$ .

$D \equiv$  "The measured distance  $d = 120$  ly."

The prior information tells us that the hypothesis space of interest consists of models (hypotheses)  $M_1$  and  $M_2$ . We proceed by writing down Bayes' theorem for each hypothesis, e.g.,

$$p(M_1|D, I) = \frac{p(M_1|I)p(D|M_1, I)}{p(D|I)}; \quad (1.14)$$

$$p(M_2|D, I) = \frac{p(M_2|I)p(D|M_2, I)}{p(D|I)}. \quad (1.15)$$

Since we are interested in comparing the two models, we will compute the *odds ratio*, equal to the ratio of the posterior probabilities of the two models. We will abbreviate the odds ratio of model  $M_1$  to model  $M_2$  by the symbol  $O_{12}$ .

$$O_{12} = \frac{p(M_1|D, I)}{p(M_2|D, I)} = \frac{p(M_1|I)p(D|M_1, I)}{p(M_2|I)p(D|M_2, I)} = \frac{p(D|M_1, I)}{p(D|M_2, I)}. \quad (1.16)$$

The two prior probabilities cancel because they are equal and so does  $p(D|I)$  since it is common to both models. To evaluate the likelihood  $p(D|M_1, I)$ , we note that in this case, we are assuming  $M_1$  is true. That being the case, the only reason the measured  $d$  can differ from the prediction  $d_1$  is because of measurement uncertainties,  $e$ . We can thus write  $d = d_1 + e$  or  $e = d - d_1$ . Since  $d_1$  is determined by the model, it is certain, and so the probability,<sup>6</sup>  $p(D|M_1, I)$ , of obtaining the measured distance is equal to the probability of the error. Thus we can write

$$\begin{aligned} p(D|M_1, I) &= \frac{1}{\sqrt{2\pi} \sigma} \exp\left(-\frac{(d-d_1)^2}{2\sigma^2}\right) \\ &= \frac{1}{\sqrt{2\pi} 40} \exp\left(-\frac{(120-100)^2}{2(40)^2}\right) = 0.00880. \end{aligned} \quad (1.17)$$

Similarly we can write for model  $M_2$

<sup>6</sup> See Section 4.8 for a more detailed treatment of this point.

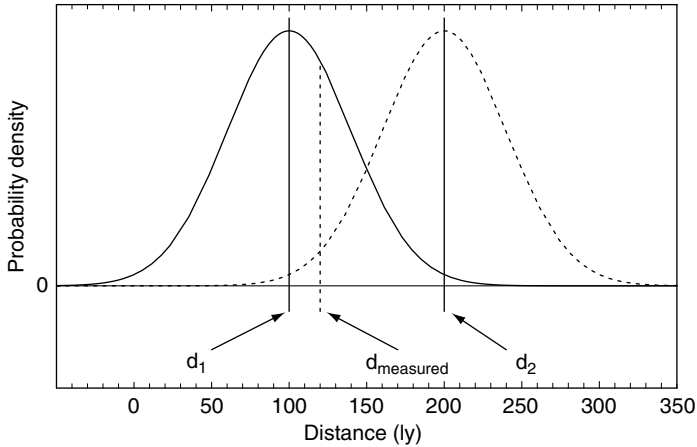


Figure 1.3 Graphical depiction of the evaluation of the likelihood functions,  $p(D|M_1, I)$  and  $p(D|M_2, I)$ .

$$\begin{aligned}
 p(D|M_2, I) &= \frac{1}{\sqrt{2\pi} \sigma} \exp\left(-\frac{(d-d_2)^2}{2\sigma^2}\right) \\
 &= \frac{1}{\sqrt{2\pi} 40} \exp\left(-\frac{(120-200)^2}{2(40)^2}\right) = 0.00135.
 \end{aligned}
 \tag{1.18}$$

The evaluation of Equations (1.17) and (1.18) is depicted graphically in Figure 1.3. The relative likelihood of the two models is proportional to the heights of the two Gaussian probability distributions at the location of the measured distance. Substituting into Equation (1.16), we obtain an odds ratio of 6.52 in favor of model  $M_1$ .

#### 1.4 Probability and frequency

In Bayesian terminology, a *probability* is a representation of our state of knowledge of the real world. A *frequency* is a factual property of the real world that we measure or estimate.<sup>7</sup> One of the great strengths of Bayesian inference is the ability to incorporate relevant prior information in the analysis. As a consequence, some critics have discounted the approach on the grounds that the conclusions are subjective and there has been considerable confusion on that subject. We certainly expect that when scientists from different laboratories come together at an international meeting, their state of knowledge about any particular topic will differ, and as such, they may have arrived at different conclusions. It is important to recognize

<sup>7</sup> For example, consider a sample of 400 people attending a conference. Each person sampled has many characteristics or attributes including sex and eye color. Suppose 56 are found to be female. Based on this sample, the frequency of occurrence of the attribute female is  $56/400 \equiv 14\%$ .