3

Introduction to Bayesian methods

3.1 Bayesian philosophy

The popularity of Bayesian methods has constantly increased since the 1950s, and is now at its peaks, with Bayesian models used in virtually every research area, from social science (Jackman, 2009) to medicine and public health (Berry and Stangl, 1999), from finance (Rachev *et al.*, 2008) to ecology (McCarthy, 2007) to health economics (Baio, 2012), and econometrics (Gómez-Rubio *et al.*, 2014). Nevertheless, the history of Bayesian thinking dates back to the eighteen century; in this section we are going to follow the steps of the great scientists who have invented and developed the theory which is the core of this book, but for an extensive and complete history of Bayesian theory, we encourage the reader to refer to the following publications: Howie (2002), Fienberg (2006), and Bertsch Mcgrayne (2011).

3.1.1 Thomas Bayes and Simon Pierre Laplace

All started with the work of two men: the reverend Thomas Bayes and the scientist Simon Pierre Laplace. Thomas Bayes, born in 1701 in Hertfordshire (England), was a presbyterian minister and mathematician. He started thinking about probability as a way to explain cause–effect relationships in a mathematical way. He was inspired by the work of the Scottish philosopher David Hume who in 1748 published an essay where he attacked some of the pillars of Christianity based on traditional belief and affirmed that we can only base our knowledge on our experience. Without knowing it, he was introducing a cornerstone of Bayesian theory, that experience plays an essential role in informing about cause–effect mechanism. He did not

Spatial and Spatio-temporal Bayesian Models with R-INLA, First Edition.
Marta Blangiardo and Michela Cameletti.

provide any mathematical detail, but concluded that "we can only find probable causes from probable effects" (Bertsch Mcgrayne, 2011).

Bayes' interest laid on demonstrating that knowing an effect it was possible to obtain the probability of the causes that might have generated it, essentially giving the foundations for the so-called inverse probability theory which is embraced by the Bayesian philosophy. In "An Essay Towards Solving a Problem in the Doctrine of Chances" (Bayes, 1763), published posthumously by Richard Price in 1763, he wrote: *Given the number of times in which an unknown event has happened and failed: required the chance that the probability of its happening in a single trial lies somewhere between any two degrees of probability that can be named.* This sentence might seem a little obscure, but can be rephrased using the simple following example: an event which occurs n times (e.g., tossing a coin and getting head) is governed by a probability of occurrence π, which is characterized by a suitable probability distribution; in other words, Bayes proposes a direct elicitation of the cause which rules the occurrence of the effect, that is the probability of the occurrence for the particular event.

Independently from Bayes, Simon Pierre Laplace, a French scientist, presented in 1774 the fundamental principle of Bayesianism and inverse probability theory, simply stating that the probability of a cause is proportional to the probability of an event (given the cause) and providing with the first version of what we call today Bayes theorem (Laplace, 1774; Stigler, 1986). In his "Mémoire sur la probabilité des causes par les évènemens" he presented his principle as follows: *If an event can be produced by a number n of different causes, then the probabilities of these causes given the event are to each other as the probabilities of the event given the causes, and the probability of the existence of each of these is equal to the probability of the event given the cause, divided by the sum of all the probabilities of the event given each of these causes.* He first assumed that the probability of the causes must be equal and worked on applications of his principle, first on gambling, which was the only field at that time where probability was used (even though the notion of inverse probability was not). Later he wanted to test how this worked with large amount of data. To do so, demography was the natural field where to look; parishes had a large collection of data from births to marriages and deaths. He concentrated on births by gender and concluded that based on the evidence the probability that a baby boy was born was slightly higher than a baby girl. He started with equal probabilities for the two events (being born as a boy or as a girl) and then applying the inverse probability theory he included the evidence from the data to update his probabilities, using the very same idea that is at the basis of Bayesian inference and which will be introduced in Section 3.4.1.

Laplace worked extensively on demography and in 1781 he estimated the size of the French population, using data from parishes and also additional information about the size of the population available for the eastern regions of the country; in the same year he also generalized his inverse probability principle, allowing for the causes to have different probabilities and proposing the final form of Bayes theorem (which will be introduced in Section 3.3). Subjectivism is a natural characteristic of inverse probability theory: different people will have a different view of what cause

is more probable for a particular event, based on their experience; in case of no information, the individual is allowed to consider all the causes as equally probable.

After Laplace died in 1827, the concept of probability used to quantify uncertainty due to unknown was despised for almost a century and so was subjectivism. Francis Ysidro Edgeworth and Karl Pearson returned on the idea of inverse probability at the end of the nineteenth/early twentieth century (Edgeworth, 1887; Pearson, 1920) and after that the subjectivism became the central topic in the work of Bruno de Finetti.

3.1.2 Bruno de Finetti and colleagues

Bruno de Finetti, an Italian actuary (1906–1985) worked all his life on the subjective concept of probability. He wrote 17 papers by the 1930 (according to Lindley, 1986), but the most famous are two: one published in 1931 where he put the basis for his subjective probability theory and the other "Theory of Probability," written in Italian in 1970 and translated in English (de Finetti, 1974) where he wrote: *My thesis, paradoxically, and a little provocatively, but nonetheless genuinely, is simply this: Probability does not exist. The abandonment of superstitious beliefs about the existence of the Phlogiston, the Cosmic Ether, Absolute Space and Time, ... or Fairies and Witches was an essential step along the road to scientific thinking. Probability, too, if regarded as something endowed with some kind of objective existence, is no less a misleading misconception, an illusory attempt to exteriorize or materialize our true probabilistic beliefs.* What he means by this statement is that the concept of objective probability is nonexistent and each individual is entitled to their own *degree of belief* in the causes which govern the occurrence of events.

Independently from de Finetti, Frank Ramsey, an English mathematician and philosopher, arrived to the same conclusions and in his "Studies in Subjective Probability" (Ramsey, 1931), he introduced the idea of subjective belief comparing it to the lowest odds, which a person will accept in a bet. On the other hand, Harold Jeffrey, an English mathematician, while maintaining a Bayesian framework, focused his theory on the concept of objective probability, which assumes ignorance on all the probabilities of occurrence for the events. In his "Theory of Probability" (Jeffrey, 1939), he insisted on the importance of learning from experience.

During the Second World War, many scientists contributed to the development of Bayesian methods, mostly in an applied perspective: amongst the others, Alan Turing, an English mathematician and computer scientist, who is mostly known for the invention of the modern computers, pioneered sequential data analysis using a Bayesian framework.

3.1.3 After the Second World War

It is after the Second World War that the so-called neo-Bayesianism revival started (Fienberg, 2006), focusing mainly on decision theory: the father of this philosophy is Leonard Savage, an American mathematician and statistician, who in his book "The Foundations of Statistics", published in 1954 (Savage, 1954), synthesized

the subjectivism theory of de Finetti and Ramsey, introducing the basis for the maximization of subjective expected utility. At the same time, universities like Chicago (where Savage worked) and Harvard experienced a large affluence of Bayesian scholars. Dennis Lindley, a British statistician, was visiting Chicago University when Savage's book was published and was strongly influenced by his subjective view of probability. In a paper in 1957, he discussed how the Frequentist and Bayesian approach can be in disagreement regarding hypothesis testing in certain situations, something which is now known as Lindley's paradox (Lindley, 1957). His work was always centered on the concept of subjective uncertainty and in his latest book "Understanding uncertainty", he wrote *Uncertainty is a personal matter; it is not the uncertainty but your uncertainty* (Lindley, 2006).

Such great advances in Bayesian theory did not correspond to the same advances in empirical applications of the theories due to the computational issues arising when dealing with large datasets. It is finally in 1962 that we see the first use of Bayesian framework for solving a practical problem: John Tukey, an American statistician, and colleagues proposed a Bayesian methodology to predict the results of the 1962 US election which were presented at the NBC television during that night. They used various data on previous elections and expert opinion from leading social scientists and included a hierarchical structure (in his paper on Bayesianism, Fienberg (2006) refers that Tukey spoke extensively about borrowing strength, the main feature of hierarchical models, which will be central in Chapter 5).

3.1.4 The 1990s and beyond

The advent of computational algorithms to perform Bayesian inference and modeling started in the 1990s with the development of Gibbs sampling, and then generalized to Markov chain Monte Carlo (MCMC) methods (Gelfand and Smith, 1990; Casella and George, 1992). The availability of such computational methods allowed statisticians to perform Bayesian computation on large datasets and to specify complex models. In 1989, the project "Bayesian inference Using Gibbs Sampling" developed a user friendly software (BUGS, Spiegelhalter *et al.*, 1995) to perform Gibbs sampling simulations, later extended to include different MCMC methods, such as Metropolis–Hastings and slice sampling. Toward the end of the 1990s, the first version of WinBUGS (BUGS operating under Windows, Spiegelhalter *et al.*, 1999) was launched and in less than a decade has made it possible for applied researchers like social scientists, epidemiologists, geographers, etc., to perform Bayesian modeling on large datasets.

A more recent alternative to MCMC, based on integrated nested Laplace approximations (INLA) was developed by Rue *et al.* (2009). This method ensures faster computation than MCMC for a particular class of models, the so-called latent Gaussian models. The methodology behind INLA will be extensively treated in Chapter 4 and Chapters 5–8 will provide several examples of how INLA can be used in real data problems.

Having looked back at the essential steps taken by Bayesianism and how it has affirmed itself during more than three centuries, in the rest of the chapter we are

going to present the fundamental principles of Bayesian methods, starting with the concept of conditional and inverse probability. We will then focus on the three components of Bayesian inference: likelihood, prior, and posterior, and how the posterior can be obtained through Bayes theorem. The families of conjugate models will follow, particular types of models which are easy to treat analytically, but that are limited in their flexibility. We then conclude the chapter with a section presenting different choices for prior distribution.

3.2 Basic probability elements

3.2.1 What is an event?

Given a random experiment, Ω is the sample space, that is the collection of all its possible outcomes. For example, if we roll a six faces dice once, the sample space will be $\Omega = \{1, 2, 3, 4, 5, 6\}$ or if we toss a coin twice it will be $\Omega =$ {Tail–Tail, Tail–Head, Head–Tail, Head–Head}. An event can be defined as the outcome(s) from the experiment that we are interested in. Events always belong to Ω. For instance an event A could be "Tossing a coin twice will give me two tails", which can also be written as $A =$ {Tail–Tail}. For each event A there is a *complement* event A^C which can only happen if A does not occur. For instance, A^C will be "Tossing a coin twice will NOT give me two tails", which can also be written as $A^C =$ {Head–Head, Tail–Head, Head–Tail}.

Considering two simple events A and B, the *intersection* event, which can be written as $A \cap B$, occurs when both happen at the same time, while the *union* event, which can be written as $A \cup B$, occurs when at least one of the two events happens. Because the entire sample space is the collection of all the possible events (and therefore one of them will occur), Ω is said to be the *certain (or sure)* event. Conversely, an empty set $\emptyset$ (i.e., a set containing no events) is called the *impossible* event. If $A \cap B = \emptyset$ the two events are called *incompatible* or *mutually exclusive*. The union of each event with its complement gives the certain event, while the intersection of each event with its complement gives the empty set. For instance going back to the coin example, $A =$ {Tail–Tail}, $A^C =$ {Head–Head, Tail–Head, Head–Tail}, we can write the union and the intersection between these two events as follows:

$$A \cup A^C = \{\text{Tail–Tail, Tail–Head, Head–Tail, Head–Head}\} = \Omega$$
$$A \cap A^C = \{\} = \emptyset.$$

3.2.2 Probability of events

Having defined sample space and events in an experiment, the next step is that of assigning these a probability measure, which represents the chance each event will occur. Such probability can vary between 0 and 1, representing, at its extremes, complete certainty that the event is occurring (probability = 1) or is not occurring

(probability = 0); anything between these two values provides the degree of uncertainty on the event being true.

There are different formulations of probability (see Dawid, 1991 for an exhaustive description); as this book focuses on the Bayesian framework, we will introduce here the subjective approach, which is its foundation. Nevertheless, we will also briefly present the classical or frequentist approach, commonly used in introductory statistical courses, and generally used as a benchmark.

Classical probability

In the frequentist approach the concept of probability is the limit of a long-run relative frequency. This means that for an event A the uncertainty on its occurrence is calculated as the ratio of the number of times the event occurred to the number of trials. For instance, applying this formulation to the dice example described above, considering $A = \{2\}$ and assuming that the dice is fair, if we roll it for a large M number of times, A will occur approximately $M/6$ of the times; thus, its probability (frequency) will be $Pr(A) = 1/6$. From this definition it follows that the probability is objective, it is a characteristic of objects (e.g., of the dice) and cannot differ for different subjects.

In this framework, a probability must satisfy the three Kolmogorov axioms:

(I) for an event A in Ω, the probability $Pr(A) \geq 0$;

(II) the probability for the certain event Ω is equal to 1;

(III) the probability of the union of n incompatible events is equal to the sum of the probabilities of each of them $Pr(\cup_{i=1}^{n} A_i) = \sum_{i=1}^{n} Pr(A_i)$. This is also called axiom of total probabilities, which we will see in Section 3.3 is linked to Bayes theorem.

From these axioms it also follows that

1. the probability of the empty set is 0. As the intersection between the certain event and the impossible event is empty ($\Omega \cap \emptyset = \emptyset$) and the union between the certain event and the impossible event is the sure event ($\Omega \cup \emptyset = \Omega$), then Ω and $\emptyset$ are incompatible and for the axiom of total probabilities it follows that $Pr(\Omega) = Pr(\Omega) + Pr(\emptyset)$ from which $Pr(\emptyset) = Pr(\Omega) - Pr(\Omega) = 0$;
2. given the event A, the probability of its contrary event A^C is $1 - Pr(A)$. As $A \cup A^C = \Omega$, then $Pr(A \cup A^C) = Pr(\Omega) = 1$; also as A and A^C are incompatible, from the axiom of total probabilities $Pr(A \cup A^C) = Pr(A) + Pr(A^C) = 1$ from which $Pr(A^C) = 1 - Pr(A)$;
3. given the two events A and B, if A is included in B ($A \subseteq B$), then B can be written as the union of the event A and of the difference between B and A ($A \cup (B - A) = A \cup (B \cap A^C)$), which are incompatible; then from the law of total probabilities it follows that $Pr(B) = Pr(A) + Pr(B - A)$, from which, $Pr(B) \geq Pr(A)$, as from the first axiom of probability $Pr(B - A) \geq 0$;

4. the probability of the union of any two events A and B, $Pr(A \cup B)$ can be written as $Pr(A) + Pr(B) - Pr(A \cap B)$. Using the Venn diagram[1] in Figure 3.1, $A \cup B$ can be seen as the union of the three following incompatible events: only A occurs (left-hand side of the Venn diagram), characterized by the probability $Pr(A - (A \cap B))$; only B occurs (right-hand side of the Venn diagram), characterized by the probability $Pr(B - (A \cap B))$; A and B occur at the same time, (central part of the Venn diagram) characterized by the probability $Pr(A \cap B)$; as these three events are incompatible, applying the axiom of total probabilities, we obtain $Pr(A - (A \cap B)) = Pr(A) - Pr(A \cap B)$ (and similarly for $Pr(B - (A \cap B))$, so that $Pr(A \cup B) = Pr(A) - Pr(A \cap B) + Pr(B) - Pr(B \cap A) + Pr(B \cap A) = Pr(A) + Pr(B) - Pr(A \cap B)$. Note that if A and B are incompatible $Pr(A \cap B) = 0$, so $Pr(A \cup B)$ becomes the sum of the probabilities of the two events, as already seen in the axiom of total probabilities.

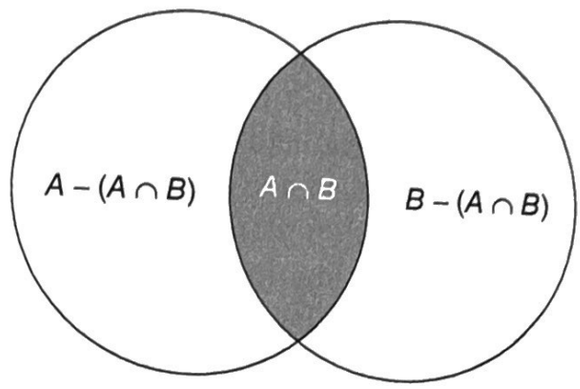

Figure 3.1 Venn diagram for two nonmutually exclusive events A and B.

There are two main concerns related to the classical (or frequentist) definition of probability: first considering the probability of an event A as a relative frequency means that we are only able to calculate it if we know the entire sample space Ω. Such condition can be satisfied in a simple example (e.g., in the dice or coin experiments previously presented), but it becomes problematic in more realistic and complex settings; for instance, if we are interested in assessing the probability of underage drinking in men living in London, or the probability of experiencing breast cancer in France, we would only be able to use the proportion if we knew the occurrence of the event of interest in the entire target population, which is in practice very unlikely. Second, this probability definition is based on the concept of repeatability, which is not necessarily a characteristic of the event of interest: for instance the events "Caesar crossed the Rubicon" or "The next US president will be a woman" do not satisfy this assumption as they can only happen once. In these cases the classical probability fails and a different approach is needed.

[1] The Venn diagram is formed by a collection of overlapping circles and is used to visualize logical relationships between classes (e.g., events).

Subjective probability

A substantially different definition is that of the subjective probability, has been introduced in Section 3.1, is the basis of Bayesian thinking. inally introduced by de Finetti (1931) in terms of betting, that is the pro the rate at which the individual is willing to bet on the occurrence of a more general terms, this probability formulation reflects a degree of belie individual has on the occurrence of an event. It is clear that the great between this and the classical approach lies on the idea that each individu tled to their own degree of belief. Thus, the probability is *subjective* (rem Section 3.1 that de Finetti claimed that probability does not exist; it shoul clear that he meant the probability in the frequentist perspective). In de definition, the only rule for a measure to be considered a probability is which he formulated in betting terms as follows: an individual will never a occurrence of an event if they are certain that they will lose.

De Finetti (1931) proved that this concept of coherence is equivalent t mogorov axioms (presented in Section 3.2.2), meaning that these rules valid in the subjective definition of probability. An important aspect is tha subjective interpretation a probability can be attached to any event, so ca are interested in "Caesar crossed the Rubicon" or "The next US presiden woman", it is always possible to indicate the degree of belief that a perso their occurrence, making such definition more flexible and versatile than t sical one, as it does not rely on the assumption that the events should be Nevertheless, frequentist reasoning can be used to inform the subjective ity, that is, information on the previous occurrence of the event can be used the degree of belief that an individual has got.

3.2.3 Conditional probability

Given the two events A and B, we can define the *conditional* event – denot $A|B$ – as the event A under the condition that the event B has already oc Then a probability measure can be associated to this event, measuring how the event A will occur *given that* the event B has already been observed. Th bility measure is called the *conditional probability*. In other words, if we con conditional probability, we are interested in the subset of the sample space in both A and B can occur and as basis for the comparison, we focus only on set where B can occur instead of considering the entire sample space Ω. Th conditional probability can be specified as follows:

$$Pr(A|B) = \frac{Pr(A \cap B)}{Pr(B)}.$$

Looking at the Venn diagram in Figure 3.1, this is equivalent to divide the p bility that both events happen (central part) by the sum of this probabilit one that only B occurs (right part). To fully appreciate the concept of cond probability, we present two simple examples.

Example 3.1: Consider a standard deck of 52 cards (13 cards for each seed); we are interested in calculating the probability of drawing two aces in a row. The event B can be defined as "an ace is extracted at the first draw" and as there are four aces in the deck:

$$Pr(B) = \frac{4}{52}.$$

The event A can be defined as "an ace is extracted at the second draw". This depends on the result of the first draw: if an ace did not occur then there are still four aces in the deck (and 51 cards in total), so $Pr(A|B^C) = \frac{4}{51}$, while if an ace was extracted in the first draw it means that only 3 are still available, so $Pr(A|B) = \frac{3}{51}$. This last result can also be obtained by using Eq. (3.1) as follows:

$$Pr(A|B) = \frac{4/52 \times 3/51}{4/52} = \frac{3}{51} = 0.059$$

concluding that around 6% of the time two consecutive aces will be drawn from a deck with 52 cards.

Example 3.2: Simon (1988) studied the 13-year cycle cicada, a noisy insect similar to the grasshopper which appears periodically in some southern US states. Table 3.1 describes a sample of cicada captured in Tennesee in 1998 by weight and gender.

Suppose that we are interested in the probability that a cicada weights more than 25 g (an event that we indicate as W) given that it is female (an event that we indicate as F). The event of interest can be written as follows:

$$W|F = \{\text{Weight} > 25 \text{ g } \textit{given that} \text{ Gender} = \text{Female}\}.$$

Using the formula in Eq. (3.1), the probability of interest $Pr(W|F)$ can be written as follows:

$$Pr(W|F) = \frac{Pr(W \cap F)}{Pr(F)}$$

and using the data in Table 3.1 we obtain

$$Pr(W|F) = \frac{11/104}{59/104} = 0.186.$$

Table 3.1 Number of Cicadas captured in Tennessee by weight and gender.

	Female (F)	Male (F^C)	Total
≤ 25 g (W^C)	48	43	91
> 25 g (W)	11	2	13
Total	59	45	104

Similarly it is possible to calculate the probability that the weight is above 25 g given that the gender of the cicada is male:

$$Pr(W|F^C) = \frac{2/104}{45/104} = 0.044,$$

so we can conclude that it is four times more likely that the weight of a cicada is above 25 g if it is a female than if it is a male.

3.3 Bayes theorem

Conditional probability plays an important role in Bayesian statistics (see Dawid, 1979, for a technical review on the topic) and Bayes theorem follows naturally from it: rearranging the definition given in Eq. (3.1), it is possible to write the probability of the intersection between the events A and B as follows:

$$Pr(A \cap B) = Pr(A|B) \times Pr(B). \tag{3.2}$$

If now we are interested in the probability of the event B, conditioning on A, applying again Eq. (3.1) we get

$$Pr(B|A) = \frac{Pr(A \cap B)}{Pr(A)} \tag{3.3}$$

and substituting Eq. (3.2) into Eq. (3.3), we obtain the so-called Bayes theorem

$$Pr(B|A) = \frac{Pr(A|B) \times Pr(B)}{Pr(A)}. \tag{3.4}$$

The interpretation of Bayes theorem follows from the inverse probability theory introduced by Thomas Bayes and Pierre Simon Laplace, as described in Section 3.1; the interest lies in the probability of the event B given that A occurs. $Pr(B)$ is calculated before we observe A; then the probability of observing A given B is used to update the original $Pr(B)$ so that $Pr(B|A)$ is obtained. If we frame Bayes theorem in an experimental setting, we can rephrase the abovementioned sentence saying that before running the experiment the researcher has some information about B whose level of uncertainty is expressed by $Pr(B)$, which is combined with the result of the experiment, expressed by $Pr(A|B)$, to obtain an updated probability for B, that is, $\Pr(B|A)$.

Now if we extend this concept and consider a set of events $B_1, \ldots, B_K$, they are defined *mutually exclusive* if $B_i \cap B_j = \emptyset$, for $i \neq j$ and *exhaustive* if $U_{i=1}^K B_i = \Omega$. Then $Pr(U_{i=1}^K B_i) = \sum_{i=1}^K Pr(B_i) = 1$, so that the probability of A can be written as

$$Pr(A) = \sum_{i=1}^{K} Pr(A \cap B_i), \text{ as } A \cap B_i \text{ are mutually exclusive events}$$

or alternatively

$$Pr(A) = \sum_{i=1}^{K} Pr(A|B_i) \times Pr(B_i), \text{ applying Eq. (3.2).}$$

These two formulations are called the law of total probabilities and can be used to rewrite Bayes theorem as follows:

$$Pr(B_i|A) = \frac{Pr(A|B_i)Pr(B_i)}{\sum_{i=1}^{K} Pr(A|B_i)Pr(B_i)}. \tag{3.5}$$

Example: A new HIV test is supposed to have 95% sensitivity (defined as probability that the test is positive given that a person has HIV) and 98% specificity (defined as probability that the test is negative given that a person has not HIV). In the English population, HIV prevalence (proportion of a population with HIV) is 0.0015 and we are interested in evaluating the probability that a patient has HIV given that the test is positive.

Let A be the event that the patient is truly HIV positive and A^C be the event that they are truly HIV negative. As either A or A^C will definitively happen, but they cannot happen at the same time, we can conclude that A and A^C are exhaustive and mutually exclusive events. Let B be the event that the patient tests positive, we want to assess the probability of the event A given B: $Pr(A|B)$.

From the available information on the test, "95% sensitivity" can be translated in probability terms as $Pr(B|A) = 0.95$, while "98% specificity" can be written as $Pr(B|A^C) = 0.02$. Then applying Bayes theorem in its formulation presented in Eq. (3.5), we obtain

$$\begin{aligned} Pr(A|B) &= \frac{Pr(B|A)Pr(A)}{Pr(B|A)Pr(A) + Pr(B|A^C)Pr(A^C)} \\ &= \frac{0.95 \times 0.0015}{0.95 \times 0.0015 + 0.02 \times 0.9985} = 0.067. \end{aligned}$$

Thus, we can conclude that about 6.7% of those testing positive will in fact have HIV.

3.4 Prior and posterior distributions

The HIV example shows that despite the high sensitivity (95%) and specificity (98%) of the test, the fact that the proportion of HIV positive in a population is very small (0.15%) has a strong impact on the final probability of being HIV positive, given that the test returns a positive result. Assuming now that the proportion of

HIV positives in the population is 20%, the probability $Pr(A|B)$ would increase as follows:

$$Pr(A|B) = \frac{0.95 \times 0.2}{0.95 \times 0.2 + 0.02 \times 0.8} = 0.92.$$

This can be explained if we think that the disease prevalence $Pr(A)$ is the information on the event of interest (being HIV positive in the example) available *a priori*, that is before carrying out any experiment (also called *prior information*). Observing a positive result through the diagnostic test updates the prior, leading to an increased *posterior* probability of having the disease. So the posterior probability will depend on the prior information as well as on the results of the experiment: when the prevalence of the disease is very low ($Pr(A) = 0.0015$), the posterior probability after observing a positive result will increase to $Pr(A|B) = 0.067$, while when the prevalence is higher ($Pr(A) = 0.2$), the posterior probability after observing a positive result will reach $Pr(A|B) = 0.92$.

3.4.1 Bayesian inference

Bayes theorem applied to observable events (as in diagnostic testing) is uncontroversial and well established, on the other hand more controversial is its use in general statistical analyses, where *parameters* are unknown quantities and their prior distribution needs to be specified, in order to obtain the corresponding *posterior* distribution. This process is known as *Bayesian inference* and it makes fundamental distinction between observable and unknown quantities.

Consider a random variable[2] Y; its uncertainty is modeled using a probability distribution or a density function (according to whether Y is a discrete or continuous random variable, respectively) indexed by a generic parameter θ. Let

$$L(\theta) = p(Y = y|\theta) \tag{3.6}$$

be the so-called *likelihood* function which specifies the distribution of the data y under the model defined by θ. Notice that $p(\cdot)$ is used to indicate the probability distribution or density function for a random variable, while $Pr(\cdot)$ is used for the probability of events (see Section 3.2). For instance, the random variable Y could be the number of deaths for respiratory diseases, we observe y and we are interested in studying the death rate θ in the population. From now on, for the sake of simplicity, we will refer to the likelihood as $p(y|\theta)$.

The variability on y depends only on the sampling selection (sampling variability). In other words we assume that the data are a random sample from the study population and the uncertainty originates by the fact that we only observe that sample instead of all the other possible ones. Conversely, the parameter θ is an unknown

[2] Given the outcomes of an experiment defined in the sample space Ω, a random variable is a variable which maps each outcome of Ω to real numbers. There are two types of random variables: discrete, if they can only assume integer values, e.g., $0, 1, \ldots, n, \ldots$; continuous if they can assume any values in a specified range, e.g., $(-\infty, +\infty)$.

quantity modeled through a suitable *prior* probability distribution $p(\theta)$ before we observe any realization y of the random variable Y and reflects our *prior* belief on θ. In the presence of a hierarchical structure or spatial (or temporal) dependence between the parameters, it would be more common to express the knowledge on θ through *hyperparameters* ψ, so that its distribution would become $p(\theta|\psi)$. The concept of hyperparameters will be treated extensively in Chapters 5–7.

Given the two components (likelihood and prior), in a Bayesian perspective the inferential problem is solved using Bayes theorem, where instead of probability of events we now include probability distributions (for the parameter θ and for the data y):

$$p(\theta|y) = \frac{p(y|\theta) \times p(\theta)}{p(y)} \tag{3.7}$$

to obtain the *posterior* distribution $p(\theta|y)$, which represents the uncertainty about the parameter of interest θ after having observed the data, thus the conditioning on y. Note that $p(y)$, in the denominator of Eq. (3.7) is the *marginal distribution* of the data and it is considered a normalization constant as it does not depend on θ, so the Bayes theorem is often reported as

$$p(\theta|y) \propto p(y|\theta) \times p(\theta),$$

where the *equal to* sign (=) is replaced by the *proportional to* sign ($\propto$).

The process to obtain the marginal distribution $p(y)$ is similar to what we showed in Eq. (3.5), where we applied the law of total probabilities for mutually exclusive and exhaustive events. To explain this point, assume that θ is a discrete parameter which can only assume values 0 and 1. We first consider the conditional probability $p(y|\theta = 0)$, weighting it by the probability that θ will assume value 0, $p(\theta = 0)$; then consider the conditional probability $p(y|\theta = 1)$, weighting it by the probability that θ will assume value 1, $p(\theta = 1)$; finally $p(y)$ will simply be the sum of the two weighted probabilities:

$$p(y) = p(y|\theta = 0) \times p(\theta = 0) + p(y|\theta = 1) \times p(\theta = 1).$$

In other words, marginalizing $p(y)$ means *integrating out* all the uncertainty on θ. This process can easily be extended to a case when θ can assume discrete values in Θ, leading to

$$p(y) = \sum_{\theta \in \Theta} p(y|\theta)p(\theta).$$

When θ is a continuous variable, the sum in the previous equation is replaced by the integral calculation

$$p(y) = \int_{\theta \in \Theta} p(y|\theta)p(\theta)\mathrm{d}\theta.$$

Note that as the posterior distribution is a combination of prior and likelihood, it is always somewhere in between these two distributions. We will see in Section 3.6 what this means in practice with two examples based on two different prior distributions.

3.5 Working with the posterior distribution

A great advantage of working in a Bayesian framework is the availability of the entire posterior probability distribution for the parameter(s) of interest. Obviously, it is always possible and useful to summarize it through some suitable synthetic indicators. The summary statistic typically used is the posterior mean, which, for a hypothetical continuous parameter of interest θ, is

$$E(\theta|y) = \int_{\theta \in \Theta} \theta p(\theta|y) \mathrm{d}\theta,$$

where Θ are all the possible values that the variable θ can assume. Note that if θ is discrete the integral is replaced by the sum.

In a similar way, it is also easy to calculate indicators which divide conveniently the probability distribution. For instance, the posterior median $\theta_{0.5}$ is defined as the value which divides the probability distribution in two equal halves, so that

$$p(\theta \leq \theta_{0.5}|y) = 0.5 \text{ and } p(\theta \geq \theta_{0.5}|y) = 0.5,$$

while the 95% credibility interval (CI) is defined as the pair of θ values ($\theta_{0.025}$ and $\theta_{0.975}$) so that

$$p(\theta \leq \theta_{0.025}|y) = 0.025 \text{ and } p(\theta \geq \theta_{0.975}|y) = 0.025.$$

Note that despite the apparent similarity of the credibility interval to the confidence intervals as defined in the frequentist approach, the interpretation of the two is completely different.

In the frequentist philosophy, the $(100 - \alpha)\%$ confidence interval suggests that if we could repeat the same experiment, under the same conditions, for a large number M of times, then the real value of θ would fall out of the intervals only $\alpha\%$ of the times. This convoluted statement is not equivalent to asserting that the probability of θ lying within the confidence interval is $(100 - \alpha)\%$, since the parameter is considered a fixed, unknown value, and it is not a random variable characterized by a probability distribution.

Conversely, within the Bayesian approach, a credibility interval will explicitly indicate the posterior probability that θ lies within its boundaries, $p(\theta \in CI|y)$; this is made possible by the fact that the parameter of interest is associated with a probability distribution, so that we can make probabilistic statements and take the underlying uncertainty into account. To highlight this difference, we talk about credibility (as opposed to confidence) intervals.

Following the same rationale, Bayesian inference can provide any probability statements about parameters; for instance, it is easy to calculate the posterior probability that θ is larger (or smaller) than an assigned threshold (e.g., $p(\theta > 1|y)$, see Figure 3.2). This probability is commonly computed in epidemiological studies,

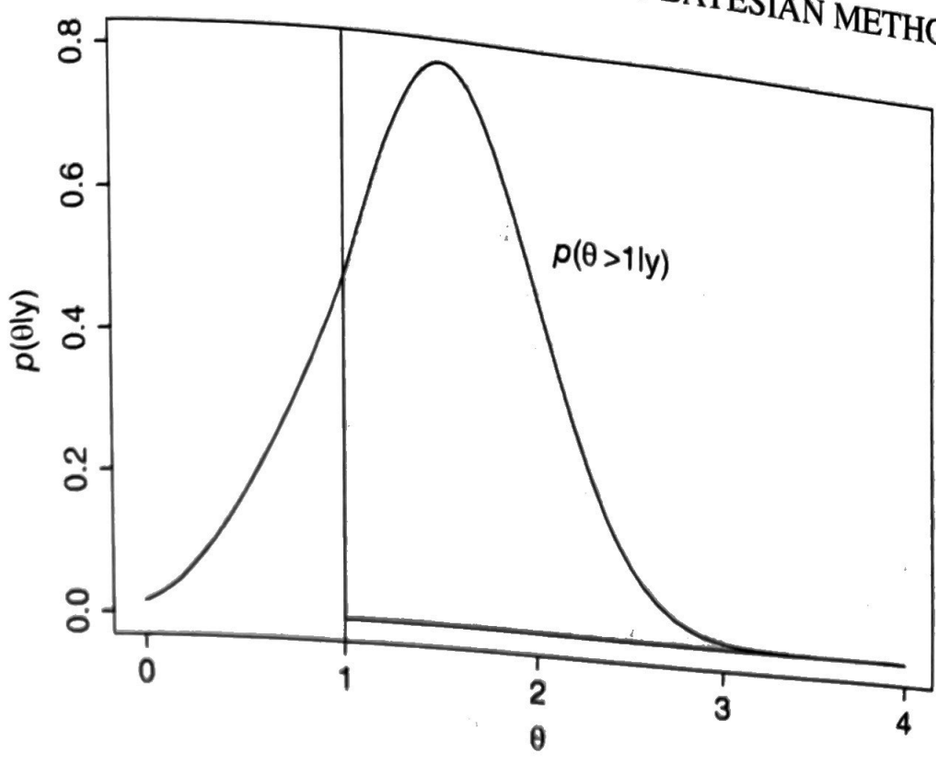

Figure 3.2 Posterior distribution for the θ parameter. The gray area corresponds to the posterior probability that $\theta > 1$.

assuming that θ is the relative risk[3] of disease (or death), and it is particularly interesting to evaluate the probability of an increased relative risk ($\theta > 1$). In comparison, inference under the frequentist approach could only lead to the estimate of the p-value for an increased relative risk. This is not equivalent to $p(\theta > 1|y)$, as it only provides a quantification of how extreme is the estimated θ, thus giving a statement about the plausibility of the data under the hypothesis that the true value of θ is 1 (Casella and Berger, 2002).

3.6 Choosing the prior distribution

When performing Bayesian inference, the choice of prior distribution is a vital issue, as it represents the information that is available for the parameters of interest. In particular, there are two aspects which need to be taken into account: (i) the type of distribution, which should be representative of the nature of the parameters, and (ii) the hyperparameters, which would make the distribution more or less informative, thus providing the level of information (or ignorance) available for the parameters.

[3] The relative risk (RR) is a measure commonly used in epidemiology to determine the change in the probability of experiencing a particular outcome (e.g., disease or death) for exposed and unexposed groups. A different measure equally common in epidemiology and used throughout the book is the odds ratio (OR), which compares the odds of the outcome amongst exposed and nonexposed. In case of rare events, e.g., when the probability of disease or death is small, then RR $\approx$ OR.

3.6.1 Type of distribution

Based on the nature of the parameters of interest there is usually a "natural" candidate for the type of prior distribution. For instance, if the parameter under study is a proportion (e.g., the probability of death in the population or the proportion of responses for a particular drug), the uncertainty on the parameter should be represented by a distribution varying between 0 and 1; if the parameter of interest is a continuous symmetric variable (e.g., the average age in the population, the daily temperature in a year or the log transformed concentration of air pollutants), the prior distribution should be allowed to vary between 0 and $+\infty$ or between $-\infty$ and $+\infty$; if the parameter of interest is a continuous positive variable (e.g., the hospitalization or mortality rate in the population), the prior distribution should only be allowed to vary between 0 and $+\infty$. Starting from three examples, we now look at commonly used models, describing the typical choice of prior distribution and the Bayesian inferential process which leads to the posterior distribution.

Binomial-Beta model

A study is carried out to evaluate the prevalence of high blood pressure among people living within 5 miles of Heathrow airport (London, UK). For each individual, the outcomes are the following:

$$1 = \text{high blood pressure; } 0 = \text{normal or low blood pressure.}$$

Defining a "success" if the ith person has high blood pressure and considering n people sampled independently from the reference population, the number of successes y represents the data in the study and can be described by a Binomial distribution

$$y|\pi \sim \text{Binomial}(\pi, n),$$

characterized by the probability function

$$p(y|\pi) = \binom{n}{y} \pi^y (1-\pi)^{n-y} \propto \pi^y (1-\pi)^{n-y},$$

where π corresponds to the generic parameter θ and represents the proportion (or probability) of successes; $\binom{n}{y}$ is called *binomial coefficient* and represents all the possible combinations of y successes in n trials. It can also be expressed through the factorial function: $\binom{n}{y} = \frac{n!}{y!(n-y)!}$, where $n!$ stands for "n factorial" and can be calculated as $n \times (n-1) \times (n-2) \times \cdots \times 2 \times 1$.

To complete a Bayesian model, we need to specify a suitable prior distribution for the proportion parameter π, such that it can only assume values between 0 and 1. A good candidate is the Beta distribution. This is a continuous distribution defined in $[0, 1]$, denoted by

$$\pi \sim \text{Beta}(a, b)$$

and characterized by two hyperparameters usually indicated as a and b. The density function of the Beta distribution is the following:

$$p(\pi) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\,\pi^{a-1}(1-\pi)^{b-1} \qquad a, b > 0, \tag{3.8}$$

where $\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}$ is the Beta function and $\Gamma(a)$ is the Gamma function, defined as $\int_0^\infty \exp(-t)t^{a-1}\mathrm{d}t$, which, if a is an integer, can also be expressed using the factorial function: $\Gamma(a) = (a-1)!$.

We stress here that a *probability function* indicates the probability distribution for discrete random variables (e.g., the Binomial distribution of the data in this example), while *density function* is used for continuous random variables (e.g., the Beta distribution of the π parameter in this example). Nevertheless, for the sake of simplicity in the rest of the book we will refer to them interchangeably.

It can be shown that the mean and the variance for the Beta distribution are functions of the hyperparameters:

$$E(\pi) = \frac{a}{a+b} \tag{3.9}$$

$$\mathrm{Var}(\pi) = \frac{ab}{(a+b)^2(a+b+1)}. \tag{3.10}$$

Changing the value for a and b leads to different shapes and scales for the Beta distribution. This circumstance renders this distribution quite flexible and capable of handling different situations. See Figure 3.3 for an example of Beta density functions changing the hyperparameters a and b.

To obtain the posterior distribution for π, we need to apply Bayes theorem, combining the Binomial likelihood and the Beta prior:

$$\begin{aligned} p(\pi|y) &\propto p(y|\pi) \times p(\pi) \\ &\propto \pi^y(1-\pi)^{n-y} \times \pi^{a-1}(1-\pi)^{b-1} \\ &\propto \pi^{y+a-1}(1-\pi)^{n-y+b-1}, \end{aligned} \tag{3.11}$$

which is the functional form of another Beta random variable:

$$\pi|y \propto \mathrm{Beta}(y+a,\ n-y+b),$$

where the parameters are a function of the prior (a and b) and of the observations (y and n). In particular, the mean of the posterior Beta distribution (also the so-called *posterior mean*) is

$$E(\pi|y) = \frac{y+a}{n+a+b}, \tag{3.12}$$

which indicates how the posterior average proportion of successes is a function of the prior average proportion, depending only on a and b, updated by the number of observed successes and the total number of trials (individuals).

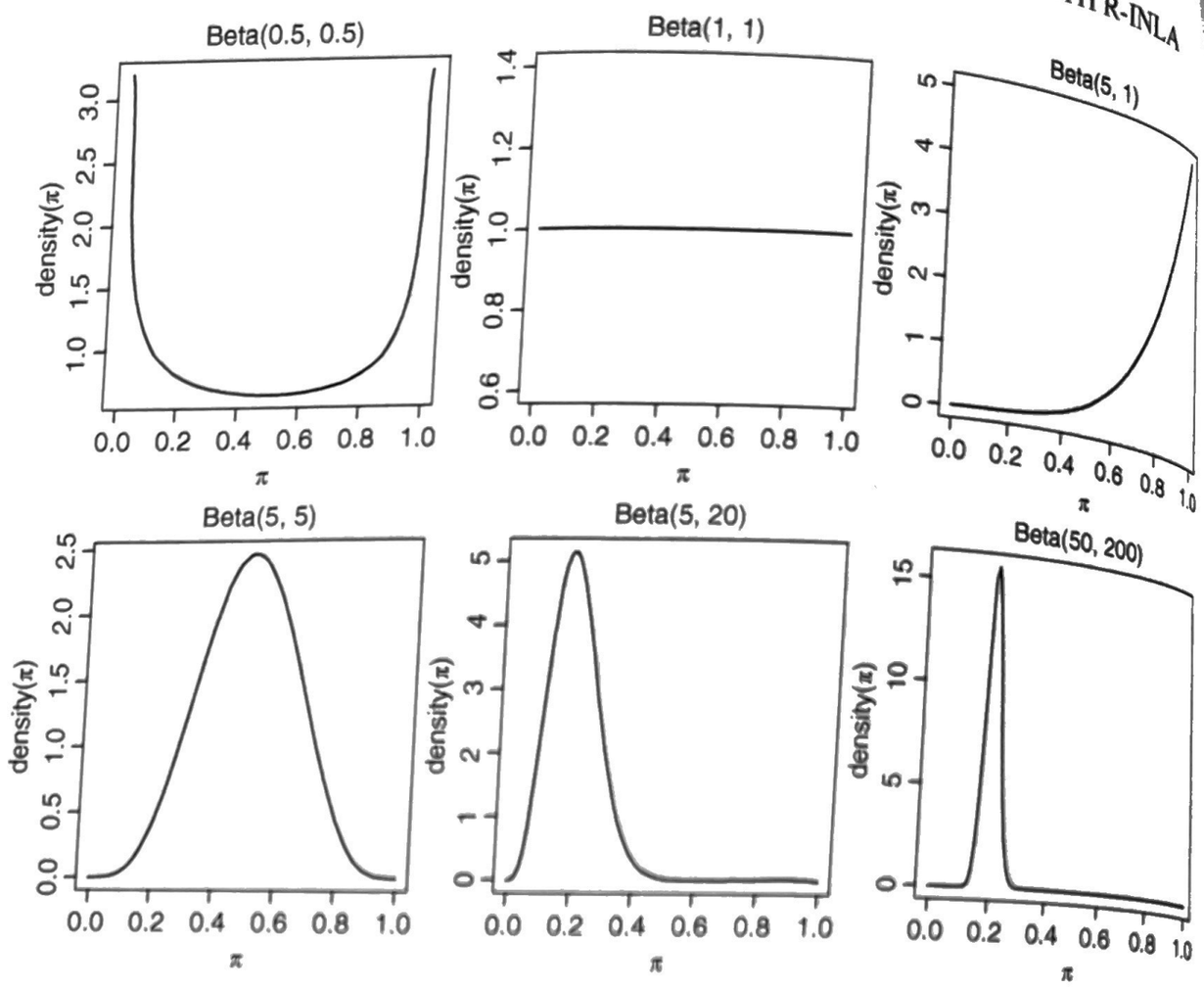

Figure 3.3 Density function for a Beta random variable for different values of the a and b hyperparameters.

Normal–Normal model

To estimate the annual mean of air pollution in Greater London (μ), daily measures of ozone (O_3) are collected in the study area using monitors for 1 year. As the values of O_3 can only be positive and generally are characterized by a long right tail, they are log transformed to resemble a Normal distribution. Note that this is typically done for skewed variables (see Gelman and Hill, 2007, Page 59). For each day ($i = 1, \ldots, n = 365$), the distribution of the log concentration (likelihood) is specified as follows:

$$y_i = \log(O_{3i})$$

$$y_i|\mu, \sigma^2 \sim \text{Normal}(\mu, \sigma^2)$$

$$p(\mathbf{y}|\mu, \sigma^2) = \prod_{i=1}^{n} p(y_i|\mu, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2\right),$$

with $\mathbf{y} = (y_1, \ldots, y_n)$, $-\infty < \mu < +\infty$. In this model, the parameter $\sigma^2 > 0$ is the sampling variability, which for the sake of simplicity we assume to know, and μ corresponds to the generic parameter θ.[4]

[4] Note that in the likelihood π is the mathematical constant that is the ratio of a circle's circumference to its diameter, equal to 3.14159...

As the parameter of interest μ is an average value of normally distributed observations, its prior distribution should be defined between $-\infty$ and $+\infty$. A good candidate is the Normal distribution characterized by the hyperparameters μ_0 and σ_0^2:

$$\mu \sim \text{Normal}(\mu_0, \sigma_0^2).$$

Combining the likelihood and the prior distribution leads to a posterior distribution, which is again a Normal random variable:

$$\mu|y \sim \text{Normal}\left(\frac{\sigma_0^2 n\bar{y} + \sigma^2\mu_0}{n\sigma_0^2 + \sigma^2}, \frac{\sigma^2\sigma_0^2}{n\sigma_0^2 + \sigma^2}\right), \tag{3.13}$$

where its mean is a weighted average between the prior mean (μ_0) and the mean of the data (sample mean $\bar{y} = \sum_{i=1}^{n} y_i/n$), with weights equal to the prior and the sampling variance:

$$E(\mu|y) = \frac{\sigma_0^2 n\bar{y} + \sigma^2\mu_0}{n\sigma_0^2 + \sigma^2}.$$

Poisson–Gamma model

Consider a study to evaluate ρ, the average rate of leukemia in Milan (Italy) in 2010. The data are the count of cases (y), i.e., a discrete variable that can assume values between 0 and $+\infty$ and can be modeled using a Poisson distribution

$$y|\lambda \sim \text{Poisson}(\lambda),$$

where λ can be written as ρE, and E represents the expected number of leukaemia cases in the study area and period. Note that this formulation is typically used when the parameter of interest is the relative risk ρ (as in our case) as opposed to when the interest lays on the mean λ of the Poisson distribution and is the basis for small area studies which will be extensively treated in Chapter 6.

The distribution of the data (likelihood) can be specified as follows:

$$p(y|\rho) = \frac{(E\rho)^y \exp(-E\rho)}{y!} \qquad \rho > 0.$$

The parameter of interest ρ corresponds to the generic θ introduced previously. As it is a rate, its prior distribution should be a continuous variable defined on the positive axis, so that ρ can only assume values between 0 and $+\infty$. One of the most used random variables in this case is the Gamma, characterized by hyperparameters (usually identified with a and b) representing the shape and the scale of the following density function[5]:

$$\rho \sim \text{Gamma}(a, b)$$

$$p(\rho) = \frac{b^a}{\Gamma(a)}\rho^{a-1}\exp(-b\rho) \qquad a, b > 0, \tag{3.14}$$

[5] Alternatively, to use the specification implemented in R, the scale parameter could be replaced by the so-called *rate*, defined as $1/b$.

where $\Gamma(a)$ is the Gamma function. The mean and variance for the Gamma distribution are functions of the hyperparameters:

$$E(\rho) = \frac{a}{b} \tag{3.15}$$

$$\text{Var}(\rho) = \frac{a}{b^2}. \tag{3.16}$$

Similarly to what we have seen for the Beta distribution, changing the values of a and b impacts on the shape and scale of the distribution, so that also the Gamma distribution is very flexible. Figure 3.4 shows the Gamma density function for different values of a and b.

The posterior distribution is once again obtained combining the likelihood and the prior:

$$p(\rho|y) \propto \frac{b^a}{\Gamma(a)} \rho^{a-1} \exp(-b\rho) \frac{(\rho E)^y \exp(-\rho E)}{y!}.$$

Leaving aside $\frac{b^a}{\Gamma(a)}$ and $\frac{1}{y!}$ which do not depend on ρ, thus can be seen as normalization factors, we obtain

$$p(\rho|y) \propto \rho^{a+y-1} \exp(-(b+E)\rho)$$

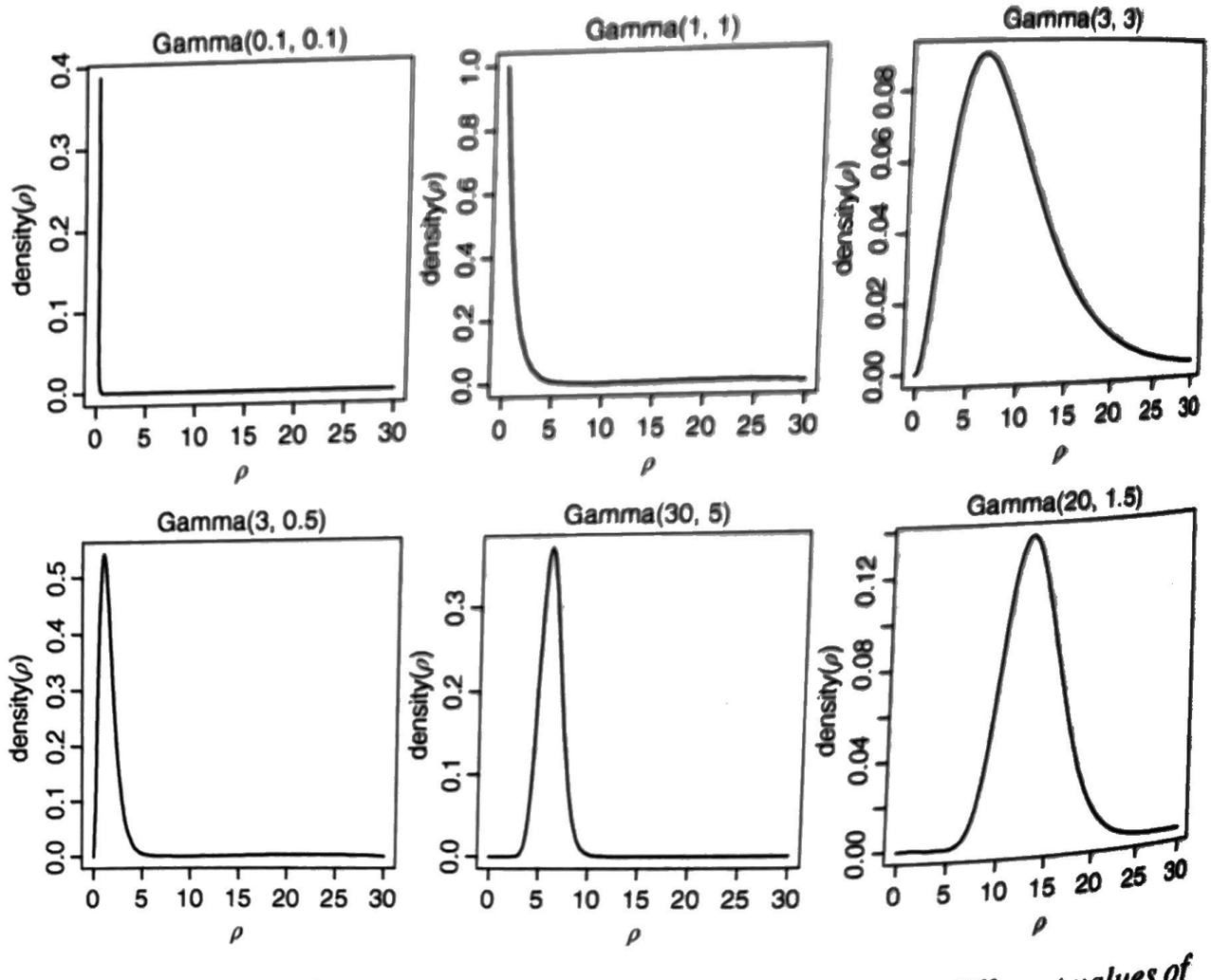

Figure 3.4 Density function for a Gamma random variable for different values of the shape and scale hyperparameters.

which is the functional form of a Gamma random variable:

$$\rho|y \sim \text{Gamma}(a + y, b + E).$$

It is characterized by a posterior mean which is a function of the hyperparameters (a and b) as well as of the data (y and E):

$$E(\rho|y) = \frac{a + y}{b + E}. \tag{3.17}$$

3.6.2 Conjugacy

The three models presented in Section 3.6.1 have one thing in common, that is the posterior distribution belongs to the same family as the prior distribution (Beta, Normal, and Gamma, respectively). This property is called *conjugacy* in Bayesian terms. Equivalently it can be said that the prior *is conjugated* to the likelihood. Conjugacy is a convenient property: as the functional form of the posterior distribution is known, as well as its hyperparameters, it is easy to extract summary statistics (like the posterior mean or median, 95% credibility intervals, etc.) or derive analytically any other quantities of interest. For instance, the posterior means for the three models described above are reported in Eq. (3.12) for the Beta distribution, Eq. (3.13) for the Normal distribution, and in Eq. (3.17) for the Gamma distribution. Table 3.2 gives a list of conjugate models and provides their prior and posterior probability distributions.

Nevertheless, it is important to notice that conjugate models are not often used in practice as they have very limited flexibility: not all the likelihoods have an associated conjugate prior or conjugacy is broken when generalized linear regression models are specified (see Chapter 5 for a detailed description of regression models in a Bayesian framework), that is the class of models typically used to assess the presence of an association between predictors (e.g., risk factors) and outcomes (e.g., health end points). In these cases, it is always possible to perform Bayesian inference, but appropriate simulative or approximation methods (MCMC, INLA) need to be applied, which will be the focus of Chapter 4.

3.6.3 Noninformative or informative prior

Once the functional form of the prior distribution has been specified, the definition of its parameters should be informed by whatever knowledge is available. This has always been a critical issue in Bayesian inference and a source of major criticism from the frequentist school. We briefly present the most used noninformative and informative priors and refer the reader to Lesaffre and Lawson 2012, Chap. 5, for a more detailed review.

Noninformative prior

The specification of a so-called *noninformative* prior has always been very appealing, particularly for applied researchers who often lack information on the