Introduction to Probabilistic Modelling

Michael Gutmann
Institute for Adaptive and Neural Computation
School of Informatics, University of Edinburgh
michael.gutmann@ed.ac.uk

January 8, 2020

Abstract

We give a brief introduction to probability and probabilistic modelling. The document is a refresher; it is assumed that the reader has some prior knowledge about the topic.

Contents

1 Probability

1.1 Probability space

Random or uncertain phenomena can be mathematically described using probability theory where a fundamental quantity is the probability space. A probability space consists of three elements: the sample space Ω, the event space $\mathcal{F}$, and the probability (measure) $\mathbb{P}$.

1. Ω is the set of all possible elementary outcomes of the phenomenon of interest. In the literature, the random phenomenon of interest is usually considered to be some kind of "experiment" whose outcome is uncertain. Ω is then the collection of all possible elementary, i.e. finest-grain and distinguishable outcomes of the experiment.

2. $\mathcal{F}$ is the collection of all events (subsets of Ω) whose probability to occur one might want to compute.

3. The probability $\mathbb{P}$ measures the plausibility of each event $E \in \mathcal{F}$, assigning to it a number between zero (most implausible/improbable) and one (most plausible/probable).

Probabilities are thus non-negative,

$$\mathbb{P}(E) \geq 0 \quad \forall E \in \mathcal{F} \tag{1}$$

and normalised, which means that the maximal probability for an event to occur is one.

For the definition of the probability space to be consistent, the event space $\mathcal{F}$ needs to satisfy some basic conditions: If we can compute the probability for an event E to occur, we should also be able to compute the probability for the event E not to occur. That is:

$$\text{If } E \in \mathcal{F} \text{ then } \bar{E} = \Omega \setminus E \in \mathcal{F} \tag{2}$$

This property is called closure under complements. Further, if we can compute the probability for $E_1, E_2, \ldots$ individually to occur, we should also be able to compute the probability that any of the events occurs. That is:

$$\text{If } E_1 \in \mathcal{F}, E_2 \in \mathcal{F}, \ldots \text{ then } (\cup_i E_i) \in \mathcal{F}. \tag{3}$$

This property is called closure under countable unions. The third condition is that

$$\Omega \in \mathcal{F}, \tag{4}$$

which is a consequence of the above because $\Omega = E \cup (\Omega \setminus E)$.

Since Ω is the set of all possible outcomes, we can be certain that the event Ω occurs and the normalisation condition is

$$\mathbb{P}(\Omega) = 1. \tag{5}$$

This normalisation condition has a number of important consequences when we learn parameters of a model.

Finally, in order to avoid inconsistencies, the assignment of probabilities to events needs to satisfy an additivity condition: For every countable collection of pairwise disjoint events $E_1, E_2, \ldots$,

$$\mathbb{P}(\cup_i E_i) = \sum_i \mathbb{P}(E_i). \tag{6}$$

Note that on the left hand side, we have the probability of the event $\cup_i E_i$, while on the right hand side, we have the (possibly infinite) sum over all the probabilities of the individual events E_i. The equation says that the probability for $\cup_i E_i$ can be computed by summing up all $\mathbb{P}(E_i)$.

For two events $A, B \in \mathcal{F}$ with $A \subseteq B$, it follows from the above that we must have $\mathbb{P}(A) \leq \mathbb{P}(B)$. This is because for $A \subseteq B$ we can express B as $B = A \cup (B \setminus A)$ where A and $B \setminus A$ are disjoint, so that $\mathbb{P}(B) = \mathbb{P}(A) + \mathbb{P}(B \setminus A)$. Since $\mathbb{P}(B \setminus A)$ is non-negative, we must have $\mathbb{P}(A) \leq \mathbb{P}(B)$. This result is known as monotonicity of probability. A further consequence is that $\mathbb{P}(A \cap B) \leq \mathbb{P}(A)$ and $\mathbb{P}(A \cap B) \leq \mathbb{P}(B)$ for any $A, B \in \mathcal{F}$.

1.2 Conditional probability

After observing some event, we may want to update the probabilities that we assign to the events in $\mathcal{F}$ accordingly. The updated probability of event A after we learn that event B has occurred is the conditional probability of A given B. It is denoted by $\mathbb{P}(A|B)$. If $\mathbb{P}(B) > 0$, we compute this probability as

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}. \tag{7}$$

The conditional probability is not defined when $\mathbb{P}(B) = 0$. The conditional probability of A given B is thus the joint probability of A and B, $\mathbb{P}(A \cap B)$, re-normalised by the probability of B. We can think that the conditional probability $\mathbb{P}(.|B)$ defines a new probability $\mathbb{P}'(.)$ that may only take non-zero values on subsets of B and assigns probability one to B.

Since $\mathbb{P}(A \cap B) = \mathbb{P}(B \cap A)$, we have for $\mathbb{P}(A) > 0$

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)}, \tag{8}$$

so that $\mathbb{P}(A|B)$ and $\mathbb{P}(B|A)$ differ only by the normalisation in the denominator.

From the definition of conditional probability, it follows that

$$\begin{aligned} \mathbb{P}(A \cap B) =& \mathbb{P}(A|B)\mathbb{P}(B) \qquad \text{if } \mathbb{P}(B) > 0 & (9) \\ =& \mathbb{P}(B|A)\mathbb{P}(A) \qquad \text{if } \mathbb{P}(A) > 0, & (10) \end{aligned}$$

which means that we can use conditional probabilities to assign joint probabilities. Equations (9) and (10) are called the "product rule".

We now show that the restrictions $\mathbb{P}(B) > 0$ and $\mathbb{P}(A) > 0$ are actually not needed in the formulation of the product rule in Equations (9) and (10). This is because if, for example, $\mathbb{P}(B) = 0$ then we also must have $\mathbb{P}(A \cap B) \leq \mathbb{P}(B) = 0$ so that the product rule holds for all events A, B even if $\mathbb{P}(A|B)$ is left undefined for events B with $\mathbb{P}(B) = 0$.

Suppose that B can be partitioned into events $B_1, \ldots, B_k$,

$$B = \cup_{i=1}^k B_i, \qquad\qquad B_i \cap B_j = \varnothing \text{ if } i \neq j, \tag{11}$$

with $\mathbb{P}(B_i) > 0$. Then

$$\begin{aligned} \mathbb{P}(A \cap B) &= \mathbb{P}\left(A \cap (\cup_i^k B_i)\right) & (12) \\ &= \mathbb{P}\left(\cup_i^k (A \cap B_i)\right) & (13) \\ &= \sum_{i=1}^k \mathbb{P}(A \cap B_i) & (14) \\ &= \sum_{i=1}^k \mathbb{P}(A|B_i)\mathbb{P}(B_i), & (15) \end{aligned}$$

where we used Equation (9). The equation

$$\mathbb{P}(A \cap B) = \sum_{i=1}^{k} \mathbb{P}(A \cap B_i) \tag{16}$$

is known as the sum rule. Using Ω for B, we obtain

$$\mathbb{P}(A) = \sum_{i=1}^{k} \mathbb{P}(A|B_i)\mathbb{P}(B_i), \tag{17}$$

which is called the law of total probability.

1.3 Bayes' rule

From Equation (9) and (10), we have

$$\mathbb{P}(A|B)\mathbb{P}(B) = \mathbb{P}(B|A)\mathbb{P}(A) \tag{18}$$

from where the so called Bayes' rule follows (for $\mathbb{P}(A) > 0$):

$$\mathbb{P}(B|A) = \mathbb{P}(A|B)\frac{\mathbb{P}(B)}{\mathbb{P}(A)} \tag{19}$$

On the left hand side, the conditioning event is A while on the right hand side the conditioning event is B. Bayes' rule shows how to move from $\mathbb{P}(A|B)$ to $\mathbb{P}(B|A)$ and vice versa, that is to "revert the conditioning".

For a partition $B_1, \ldots, B_k$ of Ω, we obtain with the law of total probability a more general version of Bayes' rule,

$$\begin{aligned}\mathbb{P}(B_i|A) &= \frac{\mathbb{P}(B_i \cap A)}{\mathbb{P}(A)} && (20)\\ &= \frac{\mathbb{P}(A|B_i)\mathbb{P}(B_i)}{\mathbb{P}(A)} && (21)\\ &= \frac{\mathbb{P}(A|B_i)\mathbb{P}(B_i)}{\sum_{j=1}^{k} \mathbb{P}(A|B_j)\mathbb{P}(B_j)}. && (22)\end{aligned}$$

It can be seen that the posterior probability $\mathbb{P}(B_i|A)$ of B_i, after learning about event A, is larger than the prior probability $\mathbb{P}(B_i)$ (prior, or before observing A) if $\mathbb{P}(A|B_i)$ is larger than the weighted average of all $\mathbb{P}(A|B_j)$.

Bayes' rule has many important applications. A prototypical application is as follows: Suppose that we observe an event A whose occurrence might be due to a number of mutually exclusive causes $B_1, \ldots, B_k$. If it is known for each cause B_i how probable it is to observe A, that is, if the $P(A|B_i)$ are known, Bayes' rule enables us to compute the "reverse" conditional probability that B_i has indeed caused A, that is, $\mathbb{P}(B_i|A)$.

1.4 Examples

1.4.1 Coin tossing

A coin toss is perhaps the simplest example of a random experiment: The outcome space is $\Omega = \{H, T\}$, where H means that the outcome of the coin toss was heads while T means that the outcome was tails; the event space is $\mathcal{F} = \{\{H\}, \{T\}, \Omega, \varnothing\}$, and the probability (measure) is defined as

$$\mathbb{P}(E) = \begin{cases} \theta & \text{if } E = \{H\} \\ 1-\theta & \text{if } E = \{T\} \\ 1 & \text{if } E = \Omega \\ 0 & \text{if } E = \varnothing, \end{cases} \tag{23}$$

where $\theta \in [0, 1]$. This probability space can be used to describe any random phenomenon with a binary outcome: For example, whether the spin of a magnet is up or down, whether a bit is 0 or 1, or weather some statement holds or not.

1.4.2 Noisy measurements

Another simple example is the noisy measurement of some state of nature. We can think that nature may be randomly in a certain state or not ("Y" for "yes, it is in the state", and "N" for "no, it isn't"), and that we measure its state using some imperfect apparatus or test, returning Y or N. The outcome space is $\Omega = \{YY, YN, NY, NN\}$ where YY means that nature is truly in the state considered and the outcome of the test is correct (true positive) while YN would mean that the test is false (nature is in the state, "Y", while the test does not indicate so, "N"; this is called a false negative). Similarly, NN corresponds to a true negative while NY is a false positive. As event space we can take the set of all subsets of Ω. We can assign probabilities to all events by specifying the probabilities of the elementary outcomes $\omega \in \Omega$, that is by using three non-negative numbers θ_1,θ_2, θ_3 as in Table 1.

We can compute the conditional probabilities for the test to return a certain result given the state of nature. We denote the event that nature is in the state of interest by B_Y,

$$B_Y = \{YY, YN\}, \tag{24}$$

while B_N is the event that nature is not in the considered state

$$B_N = \{NY, NN\}. \tag{25}$$

The probabilities of the two events are $\mathbb{P}(B_Y) = \theta_1 + \theta_2$ and $\mathbb{P}(B_N) = \theta_3 + \theta_4 = 1 - \theta_1 - \theta_2$. Similarly, the event that the test returns Y corresponds to the set E_Y,

$$E_Y = \{YY, NY\}, \tag{26}$$

while the set E_N is the event that the test returns N,

$$E_N = \{YN, NN\}. \tag{27}$$

We thus have $\{YY\} = B_Y \cap E_Y$ and $\{NN\} = B_N \cap E_N$ and

$$\mathbb{P}(E_Y|B_Y) = \frac{\mathbb{P}(E_Y \cap B_Y)}{\mathbb{P}(B_Y)} \qquad \mathbb{P}(E_N|B_N) = \frac{\mathbb{P}(E_N \cap B_N)}{\mathbb{P}(B_N)} \tag{28}$$

or

$$\mathbb{P}(E_Y|B_Y) = \frac{\theta_1}{\theta_1 + \theta_2} \qquad \mathbb{P}(E_N|B_N) = \frac{\theta_4}{\theta_3 + \theta_4} \tag{29}$$

The probability $\mathbb{P}(E_Y|B_Y)$ is called the sensitivity of the test while the probability $\mathbb{P}(E_N|B_N)$ is called the specificity. Since conditional probabilities also have to sum to one, the probability $\mathbb{P}(E_Y|B_N)$ that the test says wrongly "Y" equals $1 - \mathbb{P}(E_N|B_N)$ and is called type 1 error. The probability $\mathbb{P}(E_N|B_Y)$ that the test says wrongly "N" equals $1 - \mathbb{P}(E_Y|B_Y)$ and is called type 2 error.

Meaning	Nature	Measurement	Probability
True positive	Y	Y	θ_1
False negative	Y	N	θ_2
False positive	N	Y	θ_3
True negative	N	N	θ_4

Table 1: A noisy measurement whether nature is in a certain state or not. The probabilities sum to one, $\theta_1 + \theta_2 + \theta_3 + \theta_4 = 1$. Only three of the four θ_i thus need to be specified to define the probability measure.

1.4.3 Bayes' rule

Measurements do generally not mirror reality perfectly, so that the sensitivity and specificity of a test are not equal to one. Bayes' rule allows us to compute the probability that nature is a certain state given the test result. For example, assume that the test returns Y. The probability that nature is indeed in the specific state is

$$\begin{aligned}
\mathbb{P}(B_Y|E_Y) &= \frac{\mathbb{P}(E_Y|B_Y)\mathbb{P}(B_Y)}{\mathbb{P}(E_Y|B_Y)\mathbb{P}(B_Y) + \mathbb{P}(E_Y|B_N)\mathbb{P}(B_N)} && (30)\\
&\equiv \frac{\text{sensitivity} \cdot \mathbb{P}(B_Y)}{\text{sensitivity} \cdot \mathbb{P}(B_Y) + \text{type 1 error} \cdot \mathbb{P}(B_N)} && (31)
\end{aligned}$$

or in terms of the specificity,

$$\begin{aligned}
\mathbb{P}(B_Y|E_Y) &= \frac{\mathbb{P}(E_Y|B_Y)\mathbb{P}(B_Y)}{\mathbb{P}(E_Y|B_Y)\mathbb{P}(B_Y) + (1 - \mathbb{P}(E_N|B_N))\mathbb{P}(B_N)} && (32)\\
&\equiv \frac{\text{sensitivity} \cdot \mathbb{P}(B_Y)}{\text{sensitivity} \cdot \mathbb{P}(B_Y) + (1 - \text{specificity}) \cdot \mathbb{P}(B_N)} && (33)
\end{aligned}$$

Assume, for example, that both the specificity and sensitivity of a test for a medical condition is 0.95. If the prior probability that a patient suffers from the medical condition is low, e.g. $\mathbb{P}(B_Y) = 0.001$ (one out of thousand), the posterior probability for the patient to have the condition given that the test was positive equals $\mathbb{P}(B_Y|E_Y) = 0.019$ only.

2 Random variables

Random variables are the main tools to model uncertain, or as the name suggests, random quantities.

2.1 Definition

Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, a random variable x is a real-valued function defined on Ω: $x : \Omega \to \Omega_x \subseteq \mathbb{R}$. It can be thought of as an outcome of a probing measurement made on a random phenomenon described by the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. While x is a known (deterministic) function from one space to another its uncertainty or randomness derives from the uncertainty or randomness about its inputs; if the inputs $\omega \in \Omega$ are not observed and selected randomly according to $\mathbb{P}$ the corresponding outputs $x(\omega) \in \Omega_x$ do indeed appear random even though x is a known function. It thus makes sense to talk about the probability for the occurrence of events like $x \leq \alpha$, that is of $x \in [-\infty, \alpha]$.

Let $\mathcal{F}_x$ be an event space defined for Ω_x. The probability that x takes a value in an event $E \in \mathcal{F}_x$ can be determined by computing the probability of the set of all $\omega \in \Omega$ which are mapped to E,

$$\mathbb{P}(x \in E) = \mathbb{P}(\{\omega \in \Omega : x(\omega) \in E\}) \tag{34}$$

For this equation to make sense, the set $\{\omega \in \Omega : x(\omega) \in E\}$ must be an element of $\mathcal{F}$. This puts a (mild) constraint on x. We assume that this condition is fulfilled for all mappings x and event spaces $\mathcal{F}_x$ which follow.[1]

The above definition can be extended to vector valued functions $\boldsymbol{x} = (x_1, \ldots, x_n)$: $\Omega \to \Omega_{\boldsymbol{x}} \subseteq \mathbb{R}^n$ where Equation (34) becomes

$$\mathbb{P}(\boldsymbol{x} \in E) = \mathbb{P}(\{\omega \in \Omega : \boldsymbol{x}(\omega) \in E), \tag{35}$$

where E is an element of the event space $\mathcal{F}_{\boldsymbol{x}}$. Such vector valued functions are sometimes called random vectors. They are essentially collections of random variables which are defined on the same underlying probability space $(\Omega, \mathcal{F}, \mathbb{P})$. In what follows we often do not verbally distinguish between random vectors and random variables, that is $\boldsymbol{x}$ may be called a random variable.

2.2 Distribution of random variables

Computing the probabilities in Equation (35) for all events $E \in \mathcal{F}_{\boldsymbol{x}}$ describes the (probabilistic) behaviour of $\boldsymbol{x}$, that is, its distribution, completely. Because of the properties of event spaces and probabilities, which we reviewed in Section 1, it is turns out that the computations can be restricted to events of the form $\{\boldsymbol{x} : x_1 \leq \alpha_1, \ldots, x_n \leq \alpha_n\}$. The corresponding probabilities define the cumulative distribution function (cdf) $F_{\boldsymbol{x}}$ of $\boldsymbol{x}$,

$$\begin{aligned} F_{\boldsymbol{x}}(\boldsymbol{\alpha}) &= \mathbb{P}\left(\{\boldsymbol{x} : x_1 \leq \alpha_1, \ldots, x_n \leq \alpha_n\}\right) && (36) \\ &= \mathbb{P}\left(\{\omega \in \Omega : x_1(\omega) \leq \alpha_1, \ldots, x_n(\omega) \leq \alpha_n\}\right). && (37) \end{aligned}$$

Knowing $F_{\boldsymbol{x}}$ allows one to compute the probability of any event $E \in \mathcal{F}_{\boldsymbol{x}}$. Importantly, this can be done without having to go back to the original probability space $(\Omega, \mathcal{F}, \mathbb{P})$ that was used to define $\boldsymbol{x}$ and its distribution. Ultimately, this results in a new "stand-alone" probability space.

2.3 Discrete and continuous random variables

A random variable $\boldsymbol{x}$ is called discrete if $\Omega_{\boldsymbol{x}}$ contains countably many elements only, that if $\boldsymbol{x}$ takes at most countable many different values $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots$. The distribution of $\boldsymbol{x}$ is then completely described by knowing the probabilities for each of the events $\boldsymbol{x} = \boldsymbol{\alpha}_k$. These probabilities define the probability mass function (pmf) $p_{\boldsymbol{x}}$,

$$p_{\boldsymbol{x}}(\boldsymbol{\alpha}) = \mathbb{P}(\boldsymbol{x} = \boldsymbol{\alpha}). \tag{38}$$

The probability for $\boldsymbol{x} \in E$, for some $E \in \mathcal{F}_{\boldsymbol{x}}$, can then be computed as

$$\mathbb{P}(\boldsymbol{x} \in E) = \sum_{\boldsymbol{\alpha} \in E} p_{\boldsymbol{x}}(\boldsymbol{\alpha}). \tag{39}$$

[1] Courses on probability and measure theory give more details.

By the normalisation condition for probabilities, we obtain the condition that $p_{\boldsymbol{x}}$ must sum to one,

$$\sum_{k=1}^{\infty} p_{\boldsymbol{x}}(\boldsymbol{\alpha}_k) = 1. \tag{40}$$

A random variable $\boldsymbol{x}$ is continuous if, for any event $E \in \mathcal{F}_{\boldsymbol{x}}$, the probability $\mathbb{P}(\boldsymbol{x} \in E)$ can be computed by an integral over a non-negative function $p_{\boldsymbol{x}}$ defined on $\Omega_{\boldsymbol{x}}$,

$$\mathbb{P}(\boldsymbol{x} \in E) = \int_E p_{\boldsymbol{x}}(\boldsymbol{\alpha}) \mathrm{d}\boldsymbol{\alpha}. \tag{41}$$

The function is called the probability density function (pdf) of $\boldsymbol{x}$. We use the same symbol for both the probability mass function and the probability density function. By the normalisation condition for probabilities, a pdf must integrate to one,

$$\int_{\Omega_{\boldsymbol{x}}} p_{\boldsymbol{x}}(\boldsymbol{\alpha}) \mathrm{d}\boldsymbol{\alpha} = 1. \tag{42}$$

The pdf can be determined from the cdf $F_{\boldsymbol{x}}$ by taking partial derivatives,

$$p_{\boldsymbol{x}}(\boldsymbol{\alpha}) = \frac{\partial^n F_{\boldsymbol{x}}(\alpha_1, \ldots, \alpha_n)}{\partial \alpha_1 \cdots \partial \alpha_n}. \tag{43}$$

In what follows, we may call $p_{\boldsymbol{x}}$ the pdf of $\boldsymbol{x}$ whether it is a continuous or discrete random variable.

It is often the case that notation is simplified and the pdf or pmf of $\boldsymbol{x}$ is denoted by $p(\boldsymbol{x})$. In this convention, $\boldsymbol{x}$ takes both the role of the random variable and the values it may take. The context often makes it clear which is meant, but if not, the convention can lead to considerable confusion so that one then better resorts to the more verbose notation $p_{\boldsymbol{x}}$ to denote the pdf or pmf, and $p_{\boldsymbol{x}}(\boldsymbol{\alpha})$ to denote the value of $p_{\boldsymbol{x}}$ at $\boldsymbol{\alpha}$.

There are also mixed-type of random variables which cannot be classified as either discrete or continuous. The probability that $\boldsymbol{x} \in E$ is then computed by a combination of summation and integration.

2.4 Conditional distributions and Bayes' rule

Knowing the cumulative distribution function $F_{\boldsymbol{x}}$ in (36), that is the probabilities for events of the form $\{\boldsymbol{x} : x_1 \leq \alpha_1, \ldots, x_n \leq \alpha_n\}$, defines the distribution of the random variables $\boldsymbol{x}$ completely. Typically, we have some information about some of the random variables $\boldsymbol{x} = (x_1, \ldots, x_n)$, or we learn about them over time. We thus want to be able to adjust the probabilities of the events $\{\boldsymbol{x} : x_1 \leq \alpha_1, \ldots, x_n \leq \alpha_n\}$ in light of new evidence. This is the purpose of conditional distributions.

If new information about $\boldsymbol{x}$ comes in the form of general events $E \in \mathcal{F}_{\boldsymbol{x}}$, updating of the distribution of $\boldsymbol{x}$ is best done by working with the cumulative distribution function $F_{\boldsymbol{x}}$ and the rules for conditioning of probabilities reviewed in Section 1.2.

If we observe the values of a subset of the random variables, it is generally easier to work with the pmf or pdf $p_{\boldsymbol{x}}$. Let $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2)$. For discrete random variables, the conditional pmf $p_{\boldsymbol{x}_2|\boldsymbol{x}_1}$ is defined as

$$p_{\boldsymbol{x}_2|\boldsymbol{x}_1}(\boldsymbol{\alpha}_2|\boldsymbol{\alpha}_1) = \frac{p_{\boldsymbol{x}}(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2)}{p_{\boldsymbol{x}_1}(\boldsymbol{\alpha}_1)} \tag{44}$$

for all $\boldsymbol{x}_1$ where the marginal pmf $p_{\boldsymbol{x}_1}(\boldsymbol{\alpha}_1) = \sum_{\boldsymbol{\alpha}_2} p_{\boldsymbol{x}}(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2) > 0$. The conditional is left undefined for $\boldsymbol{\alpha}_1$ where $p_{\boldsymbol{x}_1}(\boldsymbol{\alpha}_1) = 0$. This definition is in line with the definition of conditional probability in Section 1.2. We also have a corresponding product rule

$$p_{\boldsymbol{x}}(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2) = p_{\boldsymbol{x}_2|\boldsymbol{x}_1}(\boldsymbol{\alpha}_2|\boldsymbol{\alpha}_1) p_{\boldsymbol{x}_1}(\boldsymbol{\alpha}_1). \tag{45}$$

As in Section 1.2, the product rule is valid for all $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2$ even though $p_{\boldsymbol{x}_2|\boldsymbol{x}_1}$ is left undefined for those $\boldsymbol{\alpha}_1$ where $p_{\boldsymbol{x}_1}(\boldsymbol{\alpha}_1) = 0$.

For continuous random variables, the conditional pdf $p_{\boldsymbol{x}_2|\boldsymbol{x}_1}$ is defined as

$$p_{\boldsymbol{x}_2|\boldsymbol{x}_1}(\boldsymbol{\alpha}_2|\boldsymbol{\alpha}_1) = \frac{p_{\boldsymbol{x}}(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2)}{p_{\boldsymbol{x}_1}(\boldsymbol{\alpha}_1)} \tag{46}$$

for all $\boldsymbol{\alpha}_1$ where the marginal pdf $p_{\boldsymbol{x}_1}(\boldsymbol{\alpha}_1) = \int p_{\boldsymbol{x}}(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2) \mathrm{d}\boldsymbol{\alpha}_2 > 0$. For $\boldsymbol{\alpha}_1$ where $p_{\boldsymbol{x}_1}(\boldsymbol{\alpha}_1) = 0$, the conditional pdf is left unspecified. In fact, we are free to define it as we wish as long as it is a valid pdf. As before, we obtain the product rule

$$p_{\boldsymbol{x}}(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2) = p_{\boldsymbol{x}_2|\boldsymbol{x}_1}(\boldsymbol{\alpha}_2|\boldsymbol{\alpha}_1) p_{\boldsymbol{x}_1}(\boldsymbol{\alpha}_1). \tag{47}$$

Since $p_{\boldsymbol{x}}(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2) \geq 0$, it follows that $p_{\boldsymbol{x}_1}(\boldsymbol{\alpha}_1) = 0$ implies that $p_{\boldsymbol{x}}(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2) = 0$ for all $\boldsymbol{\alpha}_2$. Hence, as before, the product rule is valid for all $(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2)$ and the non-uniqueness of $p_{\boldsymbol{x}_2|\boldsymbol{x}_1}$ for those $\boldsymbol{\alpha}_1$ where $p_{\boldsymbol{x}_1}(\boldsymbol{\alpha}_1) = 0$ is irrelevant.

In simplified notation, the above equations become

$$p(\boldsymbol{x}_2|\boldsymbol{x}_1) = \frac{p(\boldsymbol{x}_1, \boldsymbol{x}_2)}{p(\boldsymbol{x}_1)} \tag{48}$$

$$p(\boldsymbol{x}_1, \boldsymbol{x}_2) = p(\boldsymbol{x}_2|\boldsymbol{x}_1) p(\boldsymbol{x}_1). \tag{49}$$

The equations remain valid for mixed type of random variables.

The conditional pdf (pmf) is a valid pdf (pmf) for $\boldsymbol{x}_2$. In particular it satisfies the normalisation condition for all values of $\boldsymbol{x}_1$,

$$\sum_{\boldsymbol{\alpha}_2 \in \Omega_2} p(\boldsymbol{\alpha}_2|\boldsymbol{x}_1) = 1 \qquad \int_{\Omega_2} p(\boldsymbol{\alpha}_2|\boldsymbol{x}_1) \mathrm{d}\boldsymbol{\alpha}_2 = 1, \tag{50}$$

where Ω_2 denotes the sample space of $\boldsymbol{x}_2$. It is thus possible to use the product rule to define the joint distribution of $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ by separately specifying the marginal distribution of $\boldsymbol{x}_1$ and the conditional distribution of $\boldsymbol{x}_2$ given $\boldsymbol{x}_1$.

Finally, as in Section 1, the product rule yields

$$p(\boldsymbol{x}_1|\boldsymbol{x}_2) = \frac{p(\boldsymbol{x}_2|\boldsymbol{x}_1) p(\boldsymbol{x}_1)}{p(\boldsymbol{x}_2)}, \tag{51}$$

which is the Bayes rule for random variables.

2.5 Examples

2.5.1 Bernoulli random variable

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be the probability space from the coin tossing example (Section 1.4.1) and x the mapping from Ω to $\{-1, 1\}$ where "head" is assigned to 1 and "tails" to -1. The mapping is deterministic but since we do not see the coin flips, the output appears random: $\mathbb{P}(x = 1) = p(1) = \mathbb{P}(H) = \theta$ and $\mathbb{P}(x = -1) = p(-1) = \mathbb{P}(T) = 1 - \theta$.

This is not the only way to construct binary random variables. There are other mappings x' and probability spaces $(\Omega', \mathcal{F}', \mathbb{P}')$ which result in the same probability distribution. For example, let $\Omega = [0, 1]$, $\mathcal{F}'$ the event space containing all intervals of the form $[a, b)$, $0 \leq a \leq b \leq 1$, and $\mathbb{P}'([a, b)) = b - a$. Then, the mapping x' with $\omega \mapsto 1$ if $\omega \leq \theta$ and $\omega \mapsto -1$ if $\omega > \theta$ has $\mathbb{P}(x' = 1) = \mathbb{P}'(\{\omega : 0 \leq \omega < \theta\}) = \theta$ as well.

2.5.2 Beta random variable

Figure 1 shows a plot of the function $b(u, v)$,

$$b(u, v) = \frac{u}{u + v}, \tag{52}$$

for $u > 0$ and $v > 0$. It can be seen that $b \in (0, 1)$. While the mapping is deterministic, if we let u and v take random values, b will take random values as well. Let u and v be the sum of squared independent Gaussian (standard normal) random variables,

$$u = \sum_{i=1}^{2\alpha} Z_i^2, \qquad\qquad v = \sum_{i=1}^{2\beta} Z_i'^2, \tag{53}$$

where the first sum goes over 2α terms and the second over 2β terms. Figure 2(a) shows a scatter plot of some random values that (u, v) take and Figure 2(b) shows a histogram of the corresponding values of b. The distribution of b is called the "beta distribution". It takes different shapes for different values of α and β because u and v behave differently even though the mapping $(u, v) \mapsto b(u, v)$ stays the same. If a random variable x has a beta distribution, it is denoted by $x \sim \text{Beta}(\alpha, \beta)$. Its pdf is given by

$$p(x|\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1 - x)^{\beta-1}, \tag{54}$$

for $x \in (0, 1)$, and where $\Gamma(.)$ is the gamma function,

$$\Gamma(t) = \int_0^\infty y^{t-1} \exp(-y) \mathrm{d}y. \tag{55}$$

The term $\Gamma(\alpha + \beta)/(\Gamma(\alpha)\Gamma(\beta))$ is needed to normalise $p(x|\alpha, \beta)$.

3 Models

We here explain the (fine) difference between probabilistic, statistical, and Bayesian models. They are often confounded and people may use "statistical model" or "probabilistic model" to refer to any one of them.

3.1 Probabilistic models

A probabilistic, or probability model of some random phenomenon formally corresponds to a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Most often, one works with random variables and the probability density (mass) function that corresponds to $\mathbb{P}$. Furthermore, the sample space and the event space are also often not explicitly indicated but implied by the context.

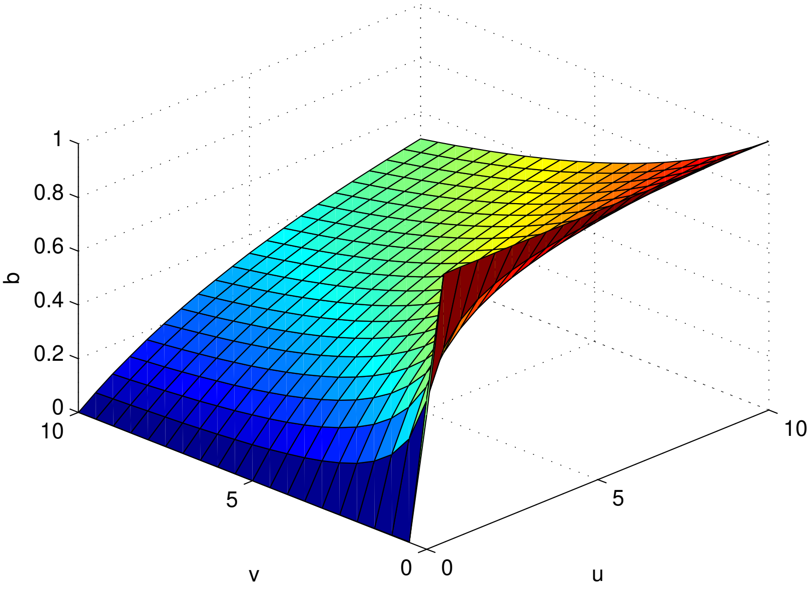
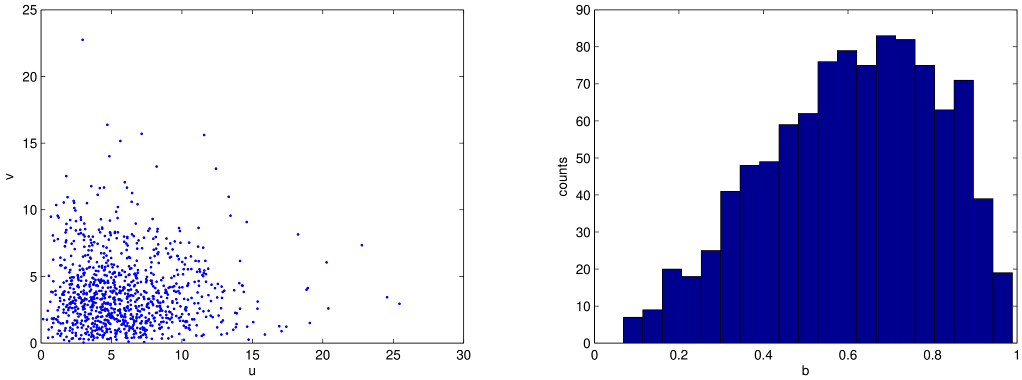


Figure 1: A plot of the function $b(u, v) = u/(u+v)$ which can be used to generate beta-distributed random variables.



(a) Scatter plot of u and v

(b) Histogram of $b = u/(u + v)$

Figure 2: If u is obtained by squaring and summing 2α standard normal random variables, and v in the same way by squaring and summing other 2β standard normal random variables, $b(u, v) = u/(u + v)$ follows a beta distribution with parameters (α, β).

For example, the Bernoulli random variable from Section 1.4.1 with success probability $1/2$ is a probabilistic model of coin tosses. A probabilistic model for a wide range of random phenomena is the Gaussian random variable with probability density function

$$p(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right). \tag{56}$$

More generally, a Gaussian random variable with known mean μ_0 and known variance σ_0^2 is also a probabilistic model. The corresponding density is

$$p(x) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{(x-\mu_0)^2}{2\sigma_0^2}\right). \tag{57}$$

3.2 Statistical models

A statistical model is a collection, or set, of probability measures defined on the same sample space Ω. In other words, a statistical model is a set of random variables that are defined on the same domain.

A parametric statistical model is a set of random variables $\boldsymbol{x}_{\boldsymbol{\theta}}$ parametrised by $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^d$. This means that for each value of the parameters $\boldsymbol{\theta}$, $\boldsymbol{x}_{\boldsymbol{\theta}}$ is a (possibly vector valued) random variable as defined in Section 2 with probability density (mass) function $p(.|\boldsymbol{\theta})$.[2] Furthermore, it is common to drop the "parametric" and just refer to statistical models instead of parametric statistical models.

For example, the collection of Bernoulli random variables parametrised by the success probability θ is a statistical model of coin tosses. The set of Gaussian random variables parametrised by the mean μ and, possibly, the variance σ^2 is also a statistical model. The set (or family) of probability density functions $p(.|\boldsymbol{\theta})$, with $\boldsymbol{\theta} = (\mu, \sigma^2)$, is

$$p(x|\boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \tag{58}$$

For a probabilistic model the mean and variance are fixed, but for the statistical model, they are free parameters. We typically use data to pick a member of the family $\{p(.|\boldsymbol{\theta})\}_{\boldsymbol{\theta}\in\Theta}$ by determining a suitable value for $\boldsymbol{\theta}$. We then say that we learn the parameters, or equivalently, that we estimate the (statistical) model. The outcome of the learning or estimation process is a probabilistic model.

3.3 Bayesian models

A Bayesian model is obtained by combining a statistical model with a (prior) probability distribution for the parameters $\boldsymbol{\theta}$. Each member of the family $\{p(.|\boldsymbol{\theta})\}_{\boldsymbol{\theta}\in\Theta}$ is considered to be a conditional pdf (or pmf), as implied by the notation $p(.|\boldsymbol{\theta})$. Together with the prior pdf for $\boldsymbol{\theta}$, the family of conditional pdfs thus defines a single joint pdf $p(.|\boldsymbol{\theta})p_{\boldsymbol{\theta}}$. Assuming that the conditional pdfs $p(.|\boldsymbol{\theta})$ are defined on $\Omega_{\boldsymbol{x}}$, a Bayesian model thus formally corresponds to a probabilistic model defined on $\Omega_{\boldsymbol{x}} \times \Theta$.

For statistical models, we used $\boldsymbol{x}_{\boldsymbol{\theta}}$ to denote the random variable that corresponds to the pdf $p(.|\boldsymbol{\theta})$ for a given value of $\boldsymbol{\theta}$. When $\boldsymbol{\theta}$ is considered a random variable, we often use the notation $\boldsymbol{x}|\boldsymbol{\theta}$ instead. We thus say that $\boldsymbol{x}|\boldsymbol{\theta}$ has the conditional pdf $p(.|\boldsymbol{\theta})$. Moreover, we associate the random variables $(\boldsymbol{x}, \boldsymbol{\theta})$ with the joint pdf $p(.|\boldsymbol{\theta})p_{\boldsymbol{\theta}}$, which is often written more simply as $p(\boldsymbol{x}, \boldsymbol{\theta}) = p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})$.

[2] The notation $p(.;\boldsymbol{\theta})$ is often used instead of $p(.|\boldsymbol{\theta})$, in particular in a non-Bayesian context.

3.4 Examples

3.4.1 Statistical model for binary data

Assume you have developed a new procedure which you think is faster or in other ways better than existing ones. The procedure could, for example, be a new classification algorithm, a new kind of measurement protocol, or a new treatment of an ailment. However, you also noticed that sometimes, the new procedure performs worse than the existing ones. The situation can be modelled using a binary random variable x where $x = 1$ means that the new procedure is performing better, while $x = 0$ means that it is performing worse. The probability that the procedure is a success is here the parameter θ, and the statistical model is specified by $p(x|\theta)$,

$$p(x|\theta) = \theta^x(1-\theta)^{(1-x)} \tag{59}$$

with $x \in \{0, 1\}$ and $\theta \in [0, 1]$.

3.4.2 Statistical model for proportions

You decide to prepare a number of tests and compute the proportion $f \in [0, 1]$ of successes to assess the efficacy of the new procedure. A popular model for the unknown proportion f is a beta random variable parametrised by α and β, see Section 2.5.2. The family of pdfs is

$$p(f|\alpha,\beta) = \frac{1}{Z(\alpha,\beta)} f^{\alpha-1}(1-f)^{\beta-1}, \tag{60}$$

where

$$Z(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \tag{61}$$

is called the partition function. It ensures that $p(f|\alpha,\beta)$ integrates to one for all values of the parameters α and β.

3.4.3 Bayesian model for binary data

We can turn the statistical model in (59) into a Bayesian model by attaching a probability distribution to θ. A common choice is to use a beta-distribution, i.e. we assume that

$$p(\theta) = p(\theta|\alpha_0,\beta_0) = \frac{1}{Z(\alpha_0,\beta_0)}\theta^{\alpha_0-1}(1-\theta)^{\beta_0-1}, \tag{62}$$

where α_0 and β_0 are assumed fix. The joint distribution $p(x,\theta)$ of (x,θ) is thus

$$\begin{aligned} p(x,\theta) &= \theta^x(1-\theta)^{(1-x)}\frac{1}{Z(\alpha_0,\beta_0)}\theta^{\alpha_0-1}(1-\theta)^{\beta_0-1} && (63)\\ &= \frac{1}{Z(\alpha_0,\beta_0)}\theta^x\theta^{\alpha_0-1}(1-\theta)^{(1-x)}(1-\theta)^{\beta_0-1} && (64)\\ &= \frac{1}{Z(\alpha_0,\beta_0)}\theta^{x+\alpha_0-1}(1-\theta)^{(\beta_0-x)}, && (65) \end{aligned}$$

and it is defined on $\{0, 1\} \times [0, 1]$.

We here assumed that α_0 and β_0 are known and fixed. If they are unknown and we consider them free parameters, we can formulate a statistical model

$$p(x,\theta|\alpha,\beta) = \frac{1}{Z(\alpha,\beta)}\theta^{x+\alpha-1}(1-\theta)^{(\beta-x)}, \tag{66}$$

which one can again turn into a Bayesian or probabilistic model by attaching a (prior) probability distribution to α and β. The parameters α and β are sometimes called hyperparameters, the assumed prior a hyperprior, and the resulting model a hierarchical Bayesian model.