7
Probability

7.1 Definition

The definition of ”probability” is closely linked to the intuitive definition of a random (or random) event. An event is random when its occurrence cannot be predicted in a deterministic way, due to an imperfect knowledge of the conditions that lead or not to its occurrence. Let’s clarify with a classic example: the toss of a coin produces the ”heads” event or the ”tails” event in a way that cannot be predicted: the result of the toss of a coin (or die, or the drawing of a card from a deck) is a random event. In reality, the event itself would be perfectly determined from a physical point of view: if the coin were thrown under the exact same conditions, the result would always be the same; therefore the event is random due to our ”ignorance” about the initial conditions. More complex examples can be given. The exact duration of a train journey is a random event, as it depends on ”accidental” factors which are, as such, not predictable. The weight of an object is random as, due to accidental errors, the scale produces a different result in a series of repeated measurements.

The definition of probability is still a controversial topic, which divides two different schools of thought. For our purposes, it is sufficient to introduce the concept of probability in the simplest situations, and then generalize the definition.

It is intuitive to assign a value of 0 to an event that cannot occur. Just as, in everyday language, we are used to assigning a probability 1 to a certain event (for example, the probability that the height of a human being is less than 3 meters is 100%, i.e. 1). To define the probability in intermediate cases, it is necessary to introduce some auxiliary concepts:

7.1.1 Definition of probability in terms of relative frequency

The definition of probability in terms of

$$P = \frac{M}{N}$$

and it can be calculated as proportion of favoreable cases (M) with respect to the total number of cases (N), makes sense only when the events are reduced to a ”pattern of cases” and, moreover, are finite in number. This situation rarely occurs in practice, and it is therefore necessary to introduce other definitions of probability. The analysis of this equation suggests a possible definition, even when events do not form a complete class etc. Let’s think, for example, of throwing a dice: even if we do not resort to any symmetry property (the one that allows us to assign p = 1/6 to the output of

each single face), supposing we roll the die a large number of times and to record the frequency of each result on a sheet, we will have obtained at the end of a total number of throws N1 times one, N2 times two,... .., N6 times six. It seems natural to define the probability of the various results through their relative frequency N_1/N, N_2/N,... .., N_6/N. We observe, in fact, that the relative frequency is a number between 0 and 1 and that it is clearly indicative of the "probability" of obtaining a certain result (if, for example, in 100,000 throws the number 1 never comes out, we can begin to think that the die is loaded).

The above is in summary the content of a theorem proved by Bernoulli and known as the "law of large numbers" described below, for which the relative frequency of an event tends to its probability when the number of "proofs" tends to infinity as shown above.

7.1.2 Axiomatic Definition of probability

The above definition is not entirely satisfactory. The main problem is that the definition in terms of relative frequency requires to be able to repeat an experiment a large number of times (even if only "conceptually"), and this is not always possible or reasonable. From a mathematical point of view, probability can be defined as follows: given an event E, we assign to E a number P(E) which we call probability of E, which satisfies the following conditions:

- P(E) is greater than or equal to 0
- P(E) = 1 only if E occurs with certainty
- Given two incompatible events A and B, the probability of their sum is equal to the sum of their probabilities: P(A + B) = P(A) + P(B)

These properties and the axioms will be defined in more details in the sections below.

7.2 Sample Space and Events

In random experiments the list of all possible outcomes is called the **sample space**, usually detoned by $\mathcal{S}$. Sample spaces can be finite or infinite, and the elements can be continuos or discrete. $\mathcal{F}$ is a set of subsets or better a *family* of sets or, with the event nomenclature, a family of events with given properties. A few examples are given here for the rolling of a die:

- What is the change of getting 1 when rolling a die. Assuming that the due is fair, the change is one in size, therefore 1/6.
- What is the chance of getting a 1 or 2. One and two are two of the equally possible one sixths, therefore the change of getting 1 or 2 is 2/6 = 1/3.
- What is the change of getting 1,2,3,4,5 or 6. Obviusly it is 100 %.
- What is the chance of not rolling a 6? This change can be obtained by knowing that the change of getting a 6 is 1/6, therefore the 16.6 %, as a consequence the change of not getting a 6 is 100 % - 16.6 % which is 83.3 % (equivalent to 5/6).

Further examples could be presented, including also the throwing of multiple dice. From a mathematical standpoint to describe changes and probability, set theory is employed. We refer to classic books for set theory, where the concept of set and its properties are described in details. For instance, the throws of dice is an example of events:

- A: "The dice displays an even number", identified by the set $A = \{2, 4, 6\}$
- B: "The dice displays a number $<$ 3. $B = \{1, 2\}$
- C: "Event $= 6$" . $C = 6$

The totality of events is named $\mathcal{S}$. The set $\{A, B, \{6\}, \mathcal{S}\}$ is a family of events, as $\{A, B\}$ is as well, $\{\Omega, \varnothing\}$, etc. Also the $\{\varnothing, \Omega\}$ sets are *all possible* subsets of $\mathcal{S}$ and they form a family. How many subsets of Ω are possible ? For instance, in one throw, how many are the possible events ?

A general classification of events is:

- Full group of events: a set of events forms a complete group when at least one of them occurs with certainty. Example: heads, tails in the flip of a coin.
- Incompatible events: two events are incompatible if they cannot occur simultaneously. Example: score 2 and score 3 when rolling a die.
- Equally likely events: two events are equally likely if there is no reason to assign them a different probability. This concept is somewhat unsatisfactory, because it suffers from "circularity" (that is, it requires having already defined the probability). However, it can be justified by resorting to considerations of symmetry (think of the faces of a die) or the principle of insufficient reason: if I know nothing about two events, I have no reason to consider one more probable than the other.
- Elementary events and compound events: an event is elementary when it cannot be decomposed into other events. For example, the event "the drawn card is a two of spades" is an elementary event. An event is said to be composed when it can be obtained as a "sum" of elementary events. For example: "the result in the roll of a die is an even number" corresponds to the sum of the events "a 2 comes up", "a 4 comes out" and "a 6 comes out".

Elements of a set can be *mutually exclusive* if two or more outcomes cannot occur simultaneously, and they are *exhaustive* if all possible outcomes are present in the list. It means that every time the experiments is performed one of the outcome of the sample space will occur. A collection of elements belonging to the sample space is called an *event* and **set theory** is employed to derive relationships between events and to derive probabilities.

In **set notation** the sample space $\mathcal{S}$ is called *universal set*, the set that includes all the possible outcomes, while an event A is called a *subset*. The main set operations are:

- **Union**. The union of two sets A and B is a set of all elements which belongs to A, to B or to both. It is written as $\mathsf{A} \cup \mathsf{B} = (x \in \mathsf{A} \text{ or } x \in \mathsf{B} \text{ or both})$.

- **Intersection**. The intersection of two sets A and B is a set which contains all the elements common to A and B. It is written as $A \cap B = (x \in A \text{ and } x \in B)$.
- **Complement**. The complement A^c of a set A is the set of elements which belongs to the universal set $\mathcal{S}$ but do not belong to A^c. It is written as $A^c = (x \notin A)$.

Set operations can be represented by the **Venn diagrams** as shown in Fig.7.1

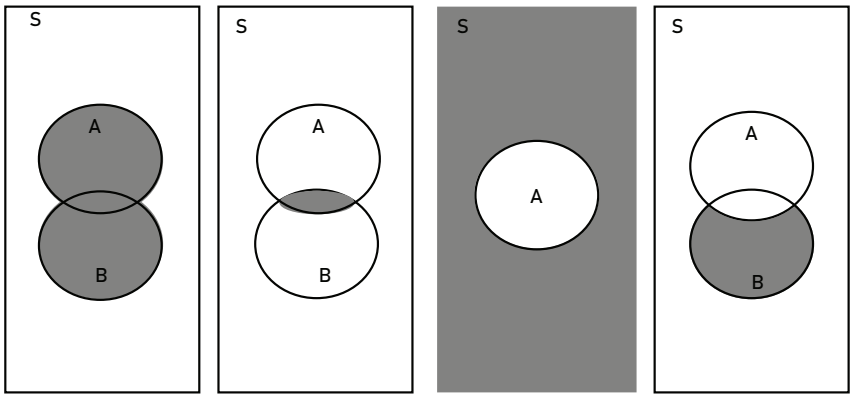

Fig. 7.1 S is the universal set. From left to right: the union $A \cup B$, the intersection $A \cap B$, the complement A^c of A and the complement of A^c intersection B, $A^c \cap B$

A set that contains no element is called **empty set** and it is denoted $\varnothing$.

As described above two events A and B are *mutually exclusive* if A and B have no elements in common, therefore their intersection is an empty set, $A \cap B = \varnothing$, in set theory these two sets are called *disjoint*.

The probability of any events must satisfy three axioms:

- **Axiom 1**: $0 \leq P(A) \leq 1$
- **Axiom 2**: $P(S) = 1$
- **Axiom 1**: $P(A \cup B) = P(A) + P(B)$ if (A) and (B) are mutually exclusive.

7.2.1 Dice with R

In this section examples of probabilities are presented, by simulating the throwing of dice with R.

The first example is to generate one number included between 1 and 6 with the instruction `sample`. This instruction will be used many times throughout this book. Sample generates random numbers and permutations. It takes a sample of the specified size from the elements of x using either with or without replacement. The function has the following arguments:

```
sample(x, size, replace = FALSE, prob = NULL)
```

where `x` is either a vector of one or more elements from which to choose, or a positive integer; `size` a non−negative integer giving the number of items to choose; `replace` defines if sampling should be with replacement? `prob` a vector of probability weights for obtaining the elements of the vector being sampled.

Sampling with replacement is used to find probability with *replacement*. In other words, the probability of some event is computed where for instance there is a number of balls, cards or other objects, and you replace the item each time you choose one, so the overall total number of elements in the set does not change. This method results in the possibility of choosing the same element multiple times. Indeed, by choosing replacement, it is possible to pick a number and then put the number back in the set, therefore the same number could be choose again. In this case the probability are independent.

Sampling *without replacement* is a way to figure out probability without replacing the element. In other words, the first chosen item is not replaced before choosing the second and so forth. This dramatically changes the odds of choosing sample items. Here the game of dice is implemented with some instructions. In the first example the instruction `sample` is used to roll once a dice with 6 faces.

```
sample(1:6,size= 1, replace=TRUE)
```

The outcome of this computation is the simulation of one roll, determining random numbers comprises between 1 and 6 as shown below.

```
> sample(1:6,size= 1, replace=TRUE)
[1] 4
> sample(1:6,size= 1, replace=TRUE)
[1] 3
> sample(1:6,size= 1, replace=TRUE)
[1] 4
```

The same instruction can be written as follows:

```
sample(1:6,1)
```

and the outcome will be the same. To roll the die ten times:

```
sample(1:6,size= 10, replace=TRUE)
```

and the outcome will be the ten numbers. Since `replace` is set to TRUE, there can be pairs, triplets and so on of the same number.

```
[1] 1 6 5 1 3 3 5 6 1 2
```

The function `samples` allows for generating more numbers. In this example, two numbers are generated (like rolling two dice), they are visualized and them and summed up. The same instruction can be placed into a function:

```
dice= function(n){
       sample(1:6,size=n, replace=TRUE)
}
```

```
outcome=dice(10)
outcome
```

and the same outcome is obtained.

```
tworoll= sample(1:6,size=2, replace=TRUE)
tworoll
sum(tworoll)
```

the outcome will be:

```
[1] 2 2
[1] 4
```

In the next example, a code is written to check the theoretical probabilities described above. The die is rolled 1000 times and the numbers of times the number 2 is obtained are counted.

```
s= sample(1:6,size=1000, replace=TRUE)
s==2
sum(s==2)
```

The outcome is

```
[1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE
FALSE FALSE FALSE FALSE FALSE FALSE FALSE
FALSE FALSE FALSE FALSE FALSE FALSE TRUE
FALSE FALSE FALSE..........
[1] 152
```

where the FALSE and TRUE are the results of the test `s==2` executed 1000 times and 152 is the number of times the outcome was 2. The proportion is 152/1000= 0.152, or 15.2 %. Since the probability for a single number should be 1/6 or 0.1666 or ≅ 16 %, the outcome is a little smaller. If the die is rolled 10000 times, the results obtained are 0.169, 0.160, 0.17, 0.168 and so forth. By increasing the number of rolls the number gets closer and closer to the theoretical probability value. With 100000 rolls the number is always 0.1669, 0.1668 and so forth, therefore it is correct to the second decimal.

In the example below the probability of obtaining the number 2 from a consecutive and incremental number of rolls is computed.

```
count=1
rolls<-vector()
x<-vector()
for(rolls in 1:1000){
       s=sample(1:6,size=rolls,replace=TRUE)
       #print(s)
       summation<-sum(s == 2)
       #print(summation)
```

```
        ratio<-summation/rolls
        print(ratio)
        x[rolls]<-ratio
}

plot(x, type = "s", col = "red", lwd = 1,
main = "",xlab = "Number of rolls",
ylab = "Probability [-]")
```

The output of this program is plotted in Fig. 7.2

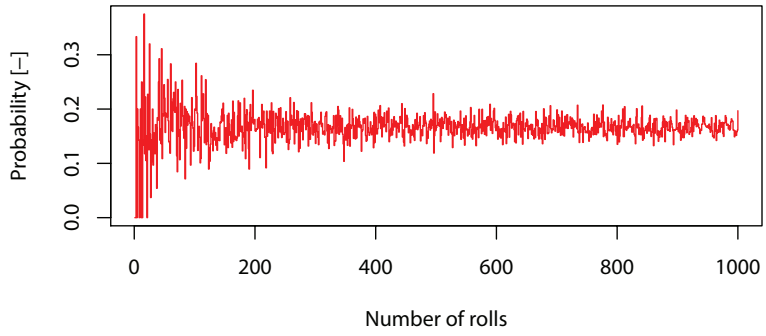

Fig. 7.2 Fraction of die rolls that are 2 at each roll in a simulation. The proportion tends to get closer to the probability $1/6 \cong 0.166$ as the number of rolls increases.

At increasing numbers of rolls the probability $(\overline{\mathsf{P}_\mathsf{n}})$ tends to stabilize around its theoretical value (P). This tendency is described by the **Law of Large Numbers**.

7.2.2 Law of large numbers

Be F_n the relative frequency of success in n independent trials, and p the probability of success in each trial. Then F_n, for $n \to \infty$, *converge in probability* to p. Here with it "converges in probability" means that the probability of the event: $\{|F_n - p| < \epsilon\}$ approximate 1, with large n and small ϵ. The law of large numbers is a theorem about the probability of an event, valid in the framework of the adopted theoretical model. It can be stated as:

The proportion $(\overline{\mathsf{P}_\mathsf{n}})$ *of occurrences with a particular outcome converges to the probability* (P) *of that outcome, as increasing numbers are collected*

It exists an empirical law that postulate: the frequency of success approximate the probability P and the approximation tends to improve at the increasing number of trials or experiments.

7.2.3 Addition rule

Two probabilities are called *mutually exclusive* or *disjoint* if they cannot happen simultaneously. For instance the probability of obtaining 1 and an even number. The

probability can be summed. For instance the probability of getting 1 or 2:

$$P(1 \text{ or } 2) = P(1) + P(2) = 1/6 + 1/6 = 1/3 \tag{7.1}$$

Similarly, all the outcomes can be summed:

$$P(1 \text{ or } 2, \text{or } 3, \text{or } 4, \text{or } 5, \text{or } 6) = \\ P(1) + P(2) + P(3) + P(4) + P(5) + P(6) = \\ 1/6 + 1/6 + 1/6 + 1/6 + 1/6 + 1/6 = 1 \tag{7.2}$$

This law is called addition rule, and it applies when the outcomes are mutually exclusive.

7.2.4 Probability of mutually inclusive outcomes

In many instances, probabilities are not disjointed such for a deck of 52 cards.

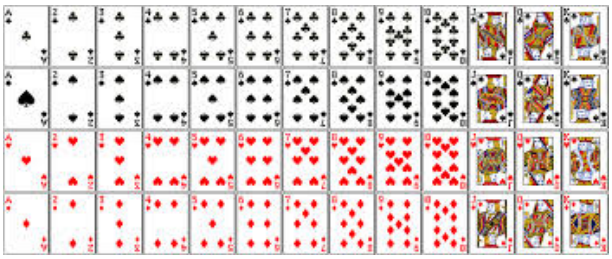

Fig. 7.3 Card deck with 52 cards.

The 52 cards are split into four suits: club (♣), diamond (♦), spade (♠) and heart (♥). Each suit has its 13 cards labeled: A (ace), 2, 3, ..., 10, J (jack), Q (queen) and K (king). Thus, each card is a unique combination of a suit and a label. The 12 cards represented by the jacks, queens, and kings are called face cards. The cards that are club and spade are typically colored black while the other two suits are typically colored red.

The question could be what is the probability that a randomly selected card is a heart ? and what is the probability that a randomly selected card is a face ? In this case the probabilities are not disjointed since it is possible to obtain a card that is a heart and it is facecard.

To describe the probability it is convenient to graphically represent it with set theory concept, the **Venn Diagram**. This diagram allows for representation of a number of elements in a set that are shared with another set. On the left the oval encompass the set of hearts (♥), which are thirteen cards, while the oval on the right represents the set for face card, which are twelve cards. When a card is both a heart and a face card it falls into the intersection of the ovals since it belongs to both sets. In set theory this is called an intersection set. The intersection of two sets, A and B, denoted by $A \cap B$, is the set containing all elements of A that also belong to B (or equivalently, all elements of B that also belong to A.

If a card is a heart but not a face card than it falls on the left part of the left oval (10 cards), likewise if a card is face card but not a heart it falls in the right part of the right-oval (9 cards). The probability that a card is a heart but not a face card is $10/52 = 0.19$ and the probability that a card is a face card but not a heart is is $9/52 = 0.17$.

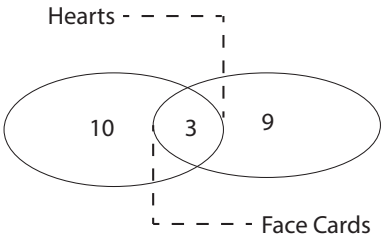

Fig. 7.4 Venn diagram for hearts and face cards.

The event that a randomly selected card is a heart is represented by A and the event that it is a face card represented by B.

How is the event $P(A \text{ or } B)$ computed ? These events are not disjointed, since the three cards J(♥), Q(♥) and K(♥) are present in both categories. In this case the addition rule for disjoint probabilities cannot be used. The Vienn Diagram is used. Firsr the probabilities of the two events are computed:

$$\mathsf{P(A)} + \mathsf{P(B))} = \mathsf{P((P(A)))} + \mathsf{P(facecards)} = 13/52 + 12/52 = 0.25 + 0.23 = 0.48 \quad (7.3)$$

Using this calculation the three cards were counted twice, once for the hearts and one for the face cards. The correction of this error is performed by using:

$$\begin{aligned} \mathsf{P(A \text{ or } B)} = \mathsf{P(}\heartsuit \text{ or facecard)} = \\ \mathsf{P(}\heartsuit + \text{facecard)} - \mathsf{P(}\heartsuit \text{ and facecard)} = \\ = 13/52 + 12/52 - 3/52 = 22/52 = 11/26 = 0.42 \end{aligned} \quad (7.4)$$

The general equation for two events A and B , disjoint or not, for the probability that at least one of them will occur is

$$\mathsf{P(A \text{ or } B)} = \mathsf{P(A)} + \mathsf{P(B)} - \mathsf{P(A \text{ and } B)} \quad (7.5)$$

In statistics the term *or* means *and* − *or* unless it is explicitly stated otherwise. Thus, A or B occurs means A, B, or both A and B occur.

7.3 Conditional probability and Bayes Theorem

Often data present interesting relationships between two or more variables that are useful to investigate. For example a car insurance company will consider information about a person's driving history to assess the risk that they will be responsible for an accident. These types of relationships are the realm of conditional probabilities.

7.3.1 Contingency Table

Let's assume a dataset called `photo_leaf_classification` represent a sample of 1822 photos of leaves with a yellow spots from a fungal disease. Researchers have been working to develop an image analysis algorithm based on machine learning (ML) to improve the automatic classification of leaves having or not having yellow spots. The 1822 photos represents a test for their classification. Each photo gets two classifications: the first is called ML and gives a classification from a machine learning (ML) system of either pred yellow spots or not (TRUE or FALSE). The algorithm presents a system to classify a leaf without spots even if it has few spots. The independent test was performed by a group of researchers with visual inspection and these data are considered the reference source or truth.

The following *contingency table* summarizes the results of the study:

	Truth		
MachineLearning	TRUE	FALSE	Total
PredTRUE	197	22	219
PredFALSE	112	1,491	1,603
Total	309	1,513	1,822

The question is now: If a leaf has the presence of spots by human inspection (TRUE), what is the chance the ML classified correctly the photo as having spots ? The probability can be estimated using the data. Of the 309 leaves presenting spots, the ML algorithm correctly classified 197 of the photos having spots. Therefore the probability is

$$P = \frac{197}{309} = 0.638$$

What is the probability that a leaf that did not have the presence of spots was correctly classified as FALSE from the ML algorithm. In analogy with the previous computation:

$$P = \frac{112}{1603} = 0.069$$

7.3.2 Marginal and joint probabilities

The data presented in the contingency table lists row and column totals for each variable separately in the photo classify data set. These totals represent *marginal probabilities* for the sample, which are the probabilities based on a single variable. For instance, the probability based solely on the ML variable is a marginal probability:

$$P = \frac{219}{1822} = 0.12$$

This probability is the probability of predicted presence of spots over the photographed leaves analyzed with machine learning.

If the estimation of the machine learning method and the direct observations are combined, the probability is called *joint probability*

$$P = \frac{197}{1822} = 0.11$$

In this case it is the probability that both methods will reveal a presence of spots on the 1822 leaves, in this case the probability is slightly smaller.

Summarizing, if a probability is based on a single variable, it is a *marginal probability*, on the other hand if the probability of outcomes for two or more variables or processes is called a *joint probability*.

	Truth		
MachineLearning	TRUE	FALSE	Total
PredTRUE	0.11	$1.21 \cdot 10^{-2}$	0.12
PredFALSE	$6.15 \cdot 10^{-2}$	0.82	0.88
Total	0.17	0.83	1

The probability can then be summarized into a table

Table 7.1 Joint probability distribution for the leaves classification data set.

Joint outcome	Probability
ML is TRUE and truth is TRUE	0.1081
ML is TRUE and truth is FALSE	0.0121
ML is FALSE and truth is TRUE	0.0615
ML is FALSE and truth is FALSE	0.8183

7.3.3 Conditional probability and Bayes Theorem

Given two events A and B, the occurrence of A may or may not affect the likelihood of B occurring. For example, if a box contains one white ball and ten black ones, drawing the white ball makes certain ($P = 1$) the next drawing of a black ball. Conversely, if the balls are returned to the box after the draw, the draw of the white ball has no influence on the subsequent draw.

We denote by $P(B|A)$ the probability of B "conditional" on the occurrence of event A. We now define the probability of the product between two events, meaning that event $A * B$ consists of "A occurs and B occurs" (for example, in the roll of a die, both A = "2 comes out" and B = "comes 3", the event $A * B$ = "comes 2" on the first roll and "comes 3" on the second roll or with a second die). It is evident that $P(A * B) = P(A) * P(B)$ only if the two events A and B are independent. It can be shown that, in general:

$$P(A * B) = P(A) * P(B|A) \tag{7.6}$$

therefore

$$\mathsf{P(A * B) \leq P(A) * P(B)} \tag{7.7}$$

with the equality that holds only if $P(B|A) = P(B)$, that is, if B is independent of A. It can also be shown (and it is intuitive) that if B is independent of A also A is independent of B, and therefore the following holds:

$$\mathsf{P(A * B) = P(B) * P(A|B)} \tag{7.8}$$

Finally, from equations 7.6 and 7.8 above it follows that:

$$\mathsf{P(A \mid B) = \frac{P(A) * P(B|A)}{P(B)}} \tag{7.9}$$

From equation 7.9 it is possible to rigorously define what is defined as "learning from experience" and which is the basis of all sciences. Let's see how. Suppose we are dealing with a complete set of incompatible events $H1, H2, \ldots, H_N$, which we will call "hypothesis" for reasons that will be clear shortly. For each of the hypotheses Hi we can apply the equations above:

$$\mathsf{P(H_i \mid E) = \frac{P(H_i) * P(E|H_i)}{P(E)}} \tag{7.10}$$

where E is any event. Since E occurs in conjunction with only one of the H_i (by definition, since the H_i are incompatible) and we can therefore write

$$\mathsf{E = E * H_1 + E * H_2 + ... + E * H_N} \tag{7.11}$$

from the addition formula we have

$$\mathsf{P(E) = P(E * H_1) + P(E * H_2) + ... + P(E * H_N)} \tag{7.12}$$

and from 7.7:

$$\mathsf{P(E) = P(H_1) * P(E|H_1) + P(H_2) * P(E|H_2) + ... + P(H_N) * P(E|H_N)} \tag{7.13}$$

therefore

$$\mathsf{P(E) = \sum_{i=1}^{N} P(H_i) * P(E|H_i)} \tag{7.14}$$

Therefore equation 7.3.4 becomes:

$$\mathsf{P(H_i \mid E) = \frac{P(H_i) * P(E|H_i)}{\sum_{i=1}^{N} P(H_i) * P(E|H_i)}} \tag{7.15}$$

Equation 7.15 is known as Bayes equation, and it allows to compute how our confidence is updated in the hypothesis H every tume a given event has occurred.

Overall, the Bayesian inference scheme is formally quite simple: the uncertainty about a parameter θ after data y have been observed, is computed simply by specifying $P(y|\theta)$ (usually referred to as the likelihood $l(\theta, y)$ when viewed as a function of θ) and $P(\theta)$, and normalizing their product to make it a probability distribution.
In the example above the machine learning predicted if a photo of a leaf displayed spots from a disease. The estimation is not perfect (it has error) but it can be a very useful and fast method to classify leaves based on automatized methods, rather than from visual inspection. It is of interest to better understand how to use this information to improve the ability to estimate a second variable, which for the example above is the truth measurement.

7.3.4 Examples

Spotted leaves. We go back to the example of the spotted leaves. The probability that a random photo from the data set present spots is about 0.17 (309/1822). The question is: if we know that machine learning predicted that the leaf had spots, can we get a better estimate of the probability that the leaf has indeed spots ? The answer is yes. As an example a subset is selected of 219 cases where the ML classified the leaf as TRUE (having spots):

$$\text{P(truth is TRUE given ML has predicted TRUE)} = \frac{197}{219} = 0.9$$

In this case the probability increased from 0.638 to 0.9 since the denominator was smaller. It is called a *conditional probability* because the probability was computed under a condition: the ML classifier prediction determine that the photo had spots (TRUE). The conditional probability is made of two parts: the outcome of interest and the condition.

It is important to consider the condition as information we know to be true and this information can be described as a known outcome or event.

It is useful to think of the condition as information we know to be true, and this information usually can be described as a known outcome or event. We generally separate the text inside our probability notation into the outcome of interest and the condition with a vertical bar:

$$\text{P(truth is TRUE} \mid \text{ML has predicted TRUE)} = \frac{197}{219} = 0.9$$

It simply reads that the visual observation detects a spot since the machine learning detected a spot, therefore the probability is 0.9.

In many cases, marginal and joint probabilities are provided instead of count data. For example, disease rates are commonly listed in percentages rather than in a count format.

Shooter. Two shooters each fire one shot at a target. The probability of the first shooter hitting the target is 0.8, that of the second is 0.4. It is observed that the target has been hit; what is the probability that it was the first shooter to hit him?
Solution:
Before the "target hit" event, there are four possible hypotheses:

- H_1 = both shooters miss
- H_2 = both hit the mark
- H_3 = the first hits the mark and the second doesn't
- H_4 = the second hits the mark and the first doesn't

Given $B1$ = "first shooter to score", and $B2$ = "second shooter to score", the probabilities of the four hypotheses are as follows (note that $B1$ and $B2$ are independent). We denote, as before, with $B^{'}$ 'the event opposite to B.

- $\mathsf{P}(\mathsf{H}_1) = \mathsf{P}(\mathsf{B}_1^{'}) * \mathsf{P}(\mathsf{B}_2^{'}) = 0.2 * 0.6 = 0.12$
- $\mathsf{P}(\mathsf{H}_2) = \mathsf{P}(\mathsf{B}_1) * \mathsf{P}(\mathsf{B}_2) = 0.8 * 0.4 = 0.32$
- $\mathsf{P}(\mathsf{H}_3) = \mathsf{P}(\mathsf{B}_1) * \mathsf{P}(\mathsf{B}_2^{'}) = 0.8 * 0.6 = 0.48$
- $\mathsf{P}(\mathsf{H}_4) = \mathsf{P}(\mathsf{B}_1^{'}) * \mathsf{P}(\mathsf{B}_2) = 0.2 * 0.4 = 0.08$

The probabilities of $E =$ "*hit target*" conditional on the 4 hypotheses are, of course:

- $\mathsf{P}(\mathsf{E}|\mathsf{H}_1) = 0$
- $\mathsf{P}(\mathsf{E}|\mathsf{H}_2) = 0$
- $\mathsf{P}(\mathsf{E}|\mathsf{H}_3) = 1$
- $\mathsf{P}(\mathsf{E}|\mathsf{H}_4) = 1$

Once E occurs, H_1 and H_2 become impossible and the probabilities of H_3 and H_4 - conditional on the occurrence of E - become:

$$\mathsf{P}(\mathsf{H}_3 \mid \mathsf{E}) = \frac{0.48 * 1}{0.48 * 1 + 0.08 * 1} = \frac{6}{7} \tag{7.16}$$

and

$$\mathsf{P}(\mathsf{H}_4 \mid \mathsf{E}) = \frac{0.08 * 1}{0.48 * 1 + 0.08 * 1} = \frac{1}{7} \tag{7.17}$$

The core of Bayes' formula lies in 7.16 and 7.17. Before experiment (E) the probability of hypothesis H_3 was 48%, after the experiment its probability increased to 86%.

Although artificial, the example shows how experience changes our knowledge of a phenomenon.

Equation says that the "a posteriori probability" $\mathsf{P}(\mathsf{H}|\mathsf{E})$ of a hypothesis H given the occurrence of the event E can be calculated by multiplying the "a priori probability" $\mathsf{P}(\mathsf{H})$ of the same hypothesis by the "likelihood" of the experimental datum E, $\mathsf{P}(\mathsf{E}|\mathsf{H})$, divided by the probability of the event E.

The likelihood of event E (or experiment) given a hypothesis H measures how much this event is "compatible" with the hypothesis itself. In the simple example above, the likelihood was 0 or 1, but this is generally not the case. Let's see it with another, more realistic example.

Clinical test. A clinical test gives a 90% accurate result, i.e. 90% positive and 10% negative when performed on people with a certain disease M. Let's assume the test is 95% negative and 5% positive, when performed on healthy people. It can also be described having 10% of false negative and 5% of false positive.

The incidence of M disease in the population is known, for example this incidence is 2%. The test is performed on a patient, which is positive. How likely is the patient to really have M?

Solution:

The hypotheses H = "the patient is sick" and H' "the patient is healthy" have, before the test, the following probabilities: $\mathsf{P}(\mathsf{H}) = 0.02$ and $\mathsf{P}(\mathsf{H}') = 0.98$. After the T test (positive), we have:

$$\mathsf{P}(\mathsf{H}|\mathsf{T}) = \frac{\mathsf{P}(\mathsf{H}) * \mathsf{P}(\mathsf{T}|\mathsf{H})}{\mathsf{P}(\mathsf{T})} = \frac{0.02 * 0.9}{0.02 * 0.9 + 0.98 * 1} = 0.27$$

If we have no other indications that the patient is suffering from M, the probability that he is ill - once the test is positive - is 27 %. This surprising result is due to the low a priori probability of the disease M.

What if the test gives more "false positives", for example if the probability of the test being positive on healthy people is 10%? In this case, the numerator does not change, but the denominator becomes:

$$\mathsf{P}(\mathsf{H}|\mathsf{T}) = \frac{\mathsf{P}(\mathsf{H}) * \mathsf{P}(\mathsf{T}|\mathsf{H})}{\mathsf{P}(\mathsf{T})} = \frac{0.02 * 0.9}{0.02 * 0.9 + 0.98 * 0.1} = 0.15$$

from which: $\mathsf{P}(\mathsf{H}|\mathsf{T})$ decrease to 15.5%.

A simple code in R is shown below.

```
content# H = sick patient
# A priori Probability
p.H <- 0.02
p.notH <- 1 - p.H
p.T.givenH <- 0.9
p.T.given.notH <- 0.1

p.T <- p.H*p.T.givenH + p.notH*p.T.given.notH

(p.H.givenT <- p.H*p.T.givenH/p.T)...
```

Lighthouse

This example is taken and reworked from *Gull (1988) in "Bayesian inductive inference and maximum entropy", in Maximum entropy and Bayesian methods in science and engineering, Kluwer.*

The example is remarkable for at least three reasons. (1) It shows how the Bayesian approach is a " paradigm " that includes techniques normally used in a more or less uncritical way. (2) It shows how the mean is not **always** the best estimate of a random quantity, and how the solution exists even when the central limit theorem does not

hold. (3) Finally, it shows how the information present in the data always wins in the end, if it is sufficient, and how the influence of the first choice of the *a priori* probability becomes negligible as the experimental data grows.

Description

A lighthouse is in a certain position on a straight stretch of coastline, at a position X_0 along the beach, measured from an arbitrarily chosen origin, and at a distance of Y_0 from the sea. The lighthouse is in constant rotation and emits short collimated flashes, at random time intervals (and therefore θ angles). Photo detectors on the beach record the flash, but not the θ angle from which the beam is coming.

The experimental data is the set of $\{x_k\}$ positions of the photo - detectors that have been activated by a flash.

Suppose, for simplicity (just so as not to have to infer two parameters, but only one) we know the distance Y_0. What is the X_0 position? How can we estimate it from the data?

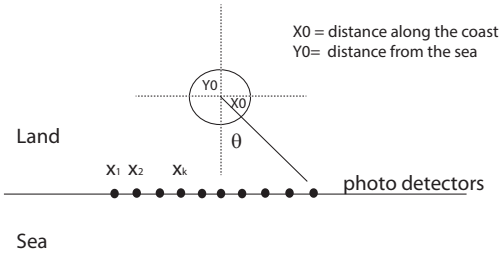

Fig. 7.5 Schematic of the problem

Solution

At each x_k correspond an azimuth value θ_k. For the light beam to be visible, the angle must be between $-\pi/2$ and $\pi/2$ (extremes not included), i.e .:

$$\theta_k \in \left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$$

Obviously, if the angle is exactly $\pm\pi/2$ the light beam does not hit the coast. Realistically we can assign a uniform probability density to the azimuth θ_k, that is to the $k-th$ datum. If we denote by X the unknown position (while Y_0 is known):

$$p(\theta_k|X, Y_0) = \frac{1}{\pi}$$

The value $\frac{1}{\pi}$ comes from the integration of the uniform probability distribution:

$$\int_{-\infty}^{+\infty} p(\theta) d\theta = 1$$

In this case, the integral is definite:

$$p(\theta_k|X,Y_0) = \int_{-\pi/2}^{\pi/2} p(\theta)d\theta = \frac{1}{\pi/2-(-\pi/2)} = \frac{1}{\pi}$$

Trigonometric considerations, when X_0 is known, allow us to say that:

$$Y_0\tan(\theta_k) = x_k - X_0$$

By knowing X_0, Y_0 and x_k, the $\tan(\theta_k)$ is obtained and the angle is obtained with the arc tangent function.
The position of the photo−detector activated by the light beam is therefore:

$$x_k = X_0 + Y_0\tan(\theta_k)$$

and the tangent of the angle is:

$$\tan(\theta_k) = \frac{x_k - X_0}{Y_0}$$

therefore the angle is

$$\theta_k = \operatorname{atan}\left(\frac{x_k - X_0}{Y_0}\right)$$

We use the variable transformation from θ to x to get the probability density of x_k. If $x = x(\theta)$, for dx and $d\theta$ infinitesimal:

$$p(x)dx = p(\theta)d\theta$$

therefore

$$p(x) = p(\theta)\left|\frac{d\theta}{dx}\right|$$

(the reason for the absolute value is that it must be a length ratio, i.e. always positive). The derivative $d\theta/dx$ is given by the derivative of the arc tangent with respect to x which is:

$$\frac{d\theta}{dx} = \frac{Y_0}{\left[Y_0^2 + (x_k - X_0)^2\right]}$$

while $p(\theta) = 1/\pi$ as shown above.
Finally:

$$p(x_k|X_0,Y_0) = \frac{Y_0}{\pi\left[Y_0^2 + (x_k - X_0)^2\right]}$$

In summary, if we know the (X_0, Y_0) position of the lighthouse, the probability of recording a flash at the x_k position has the Cauchy distribution. The Cauchy distribution is explained in details in the next chapter about probability distributions. The Cauchy distribution is often used in statistics as an example of a "pathological"

distribution since both its expected value and its variance are undefined. The Cauchy distribution does not have finite moments of order greater than or equal to one.

NOTE: If we did not know the distance Y_0 of the lighthouse from the sea, we would have to estimate two parameters, a somewhat more complex problem (the solution of which risks overshadowing what we are interested in showing here).

To estimate (infer) the X parameter (the position of the light beam), we need to estimate the posterior probability of X, given Y_0 and the records $\{x_k\}$:

$$p(X|x_k, Y_0)$$

From Bayes' theorem:

$$p(X|\{x_k\}, Y_0) = \frac{p(\{x_k\}|X, Y0)p(X|Y_0)}{p(\{x_k\}|Y_0)}$$

The term in the denominator does not depend on the parameter sought. The first term of the numerator is what is called the " likelihood " of the data, the second is the a priori probability of X. When we have no idea what a priori distribution a variable has, it is reasonable to take it uniform in a sensible range $[X_{min}, X_{max}]$ and zero outside:

$$p(X|Y_0) = p(X) = \begin{cases} \alpha, \; X \in [X_{min}, X_{max}] \\ 0, \; X \notin [X_{min}, X_{max}] \end{cases} \tag{7.18}$$

If the data x_k are independent, as it is reasonable to assume, the probability $p(\{x_k\}|X, Y0)$ is the product of the probabilities of the single events x_k, therefore:

$$p(\{x_k\}|X, Y_0) = \prod_{k=1}^{N} p(x_k|X, Y_0)$$

Take the logarithm of the posterior probability $p(\{x_k\}|X, Y0)$

$$L = \log(p(\{x_k\}|X, Y_0) = \beta - \sum_{k=1}^{N} \log(Y_0^2 + (x_k - X)^2)$$

where β includes everything that does not depend on the X parameter. The estimate of the position X_0 is obtained by looking for the maximum of the *a posteriori* distribution, that is, theoretically looking for the value of X which is a solution of:

$$\frac{dL}{dX} = 2 \sum_{k=1}^{N} \frac{x_k - X}{Y_0^2 + (x_k - X)^2} = 0 \tag{7.19}$$

Numerical procedure

The explicit solution of (7.19) is not analytically feasible. Instead of solving it numerically, it is more instructive to see how the posterior probability $\exp(L)$ behaves as the number of detections $\{x_k\}$ changes. This is what the following R code does, where we assume $Y_0 = 1$ km and the " true " value of X_0 2 km. The code generates N angles θ_k and from these it calculates x_k, since the true value of X_0 is known.

```
##Ch7_3.R
## Where is the light ?

# distance from sea
Y0 <- 1
# distance along the coast
X0 <- 2
# possible values of X (positions of photo-detectors)
dx <- 0.05
X <- seq(-5,5,dx)
Nx <- length(X)

########################################################################
# numero di rilevazioni
# da variare per osservare l'effetto sulla distribuzione a posteriori
N <- 10
########################################################################

# tetak <- runif(k,-pi/2,pi/2)
# instead of taking theta between -pi/2 and pi/2
# theta is selected such to determine "possible" x
# included between -x_max and x_max
x.max <- 50
# Here the angle is obtained using the arctan function, where x.max in an angle
tetak.max <- atan(x.max)
tetak <- runif(N,-tetak.max,tetak.max)
tetak
# compute the positions of the detectors activated by flash light
xk <- X0+Y0*tan(tetak)
# What is the distribution of the positions ?
hist(xk,main="")

L <- rep(0,Nx)
for (i in 1:Nx){
lk <- log(Y0^2+(xk-X[i])^2)
L[i] <- sum(lk)
}

hist(lk)

# posterior probability
post <- exp(-L)
plot(X,dx*post/sum(post),t="l",ylab="p(X|x,Y0)")
abline(v=2,col="blue",lty=2)
abline(v=mean(xk),col="red")
```

Results

This is what happens (typically, the calculation is stochastic ...) as the number of detections changes.

The true value (in blue) and the average value of x_k (in red) are superimposed on the probability density curve. Due to the choice of a uniform distribution for θ (which is reflected in a Cauchy distribution for likelihood), and a uniform a priori probability, with little data the maximum posterior probability rarely hits the real position X_0.

For $N = 100$ detections, the maximum a posteriori probability starts hitting the true value (figure 7.8) practically always, but the average value of x_k can be very far

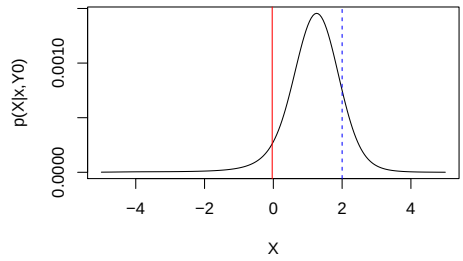

Fig. 7.6 N = 4

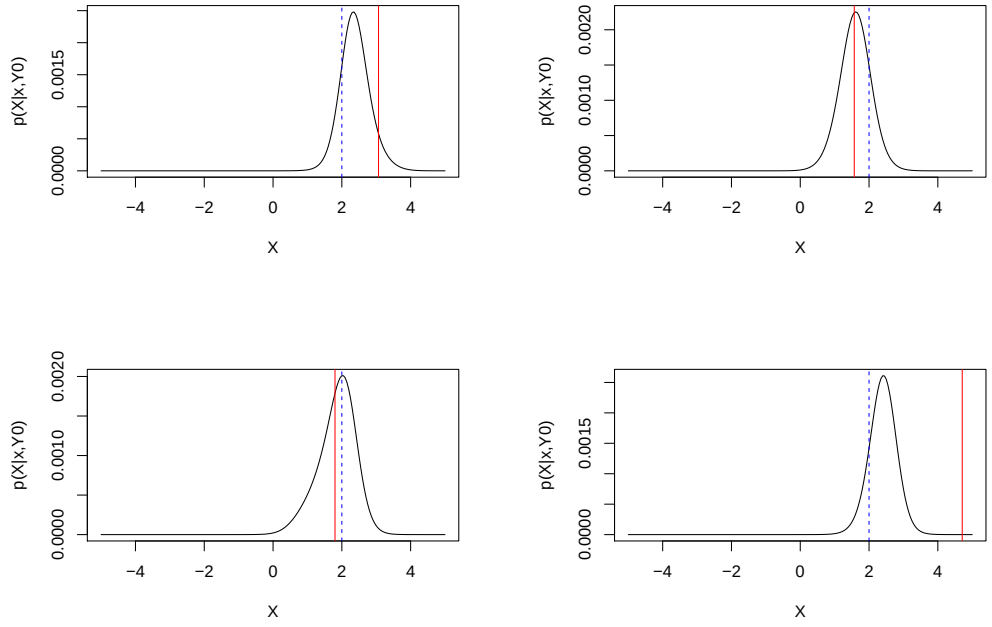

Fig. 7.7 N = 10

from it.

At first sight this thing is surprising, because we are used to attribute almost " magical " properties to the average value, by virtue of the central limit theorem which asserts that for a sample $\{x_1 ... x_n\}$, trait from a distribution with mean μ and variance σ^2 the distribution of the mean value $\bar{x}$ tends to a normal distribution with mean μ and variance σ^2/n, for $n \to \infty$.

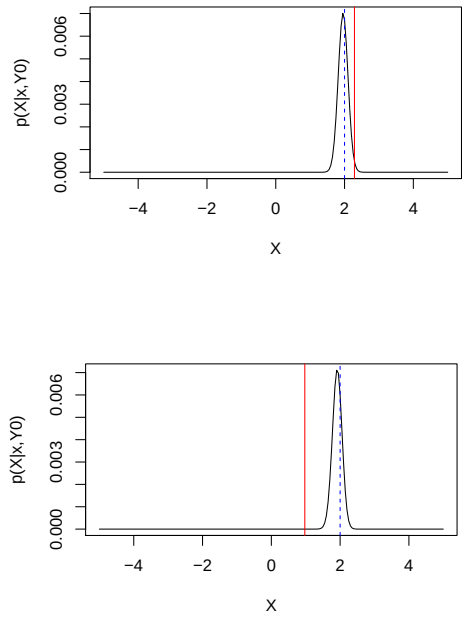

Fig. 7.8 N = 100

The problem here is that the Cauchy distribution violates the validity conditions of this theorem, because it has large tails and such as to give a moment of order 2 of infinite value. From this example we can observe that although the central limit theorem is not valid, and therefore the average is not a sensible estimate of the position of the lighthouse, we can still calculate the a posteriori distribution and the maximum of this gives us the correct position of the lighthouse. We note, incidentally, that this procedure coincides with the search for the maximum likelihood! In addition to shedding light on the latter, whose motivation is often unclear in any other way, it is evident that nothing prevents us from using information on the position of the lighthouse, if we have it, and using it by imposing a different form of a priori probability, and making the calculation of the true position correct and faster.

7.4 Probability and determinism: the Buffon's needle

An interesting example of a problem that can be solved with either a deterministic approach or a probabilistic one is the Buffon's needle. Buffon's needle was the earliest problem in geometric probability to be solved.

We draw on a piece of paper a series of parallel lines with distance among each other of `2a`. A needle, of length `2L`, with `L<a` is dropped from above. The position of the needle is identified based on the distance `x` from the centre of the needle to the closest line and with the angle θ generated from the intersection of the needle with the lines.

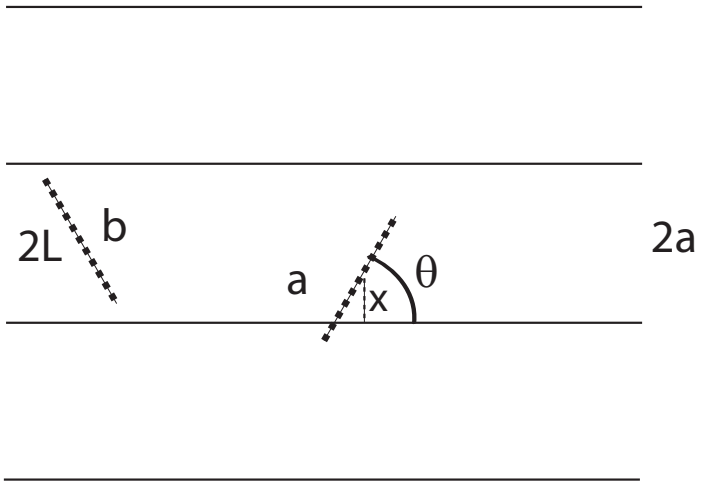

Fig. 7.9 Example of a needle (a) that lies across a line, and of needle (b) that does not.

The needle intersect the line if $x \leq L\sin\theta$. The space Ω of the possible cases is the "rectangle" of sides (a,π), since we consider the distance of the closest line and the angle comprises between 0 and π.

To estimate the probability that the needle will intersect the line we must define a space of favoreable cases F. From the condition defined above this space is comprised between the line $x = 0$ and $x = L\sin\theta$:

We define the areas Ω and of F as $a(\Omega)$ and $a(F)$ respectively.

$$a(\Omega) = \pi a$$

while:

$$a(F) = \int_0^{\pi} L\sin\theta d\theta = 2L$$

The ratio $P = a(F)/a(\Omega)$ is the seeked *estimated probability* P. Substituting the values of $a(\Omega)$ and $a(F)$ it leads to:

$$P = \frac{2L}{\pi a}$$

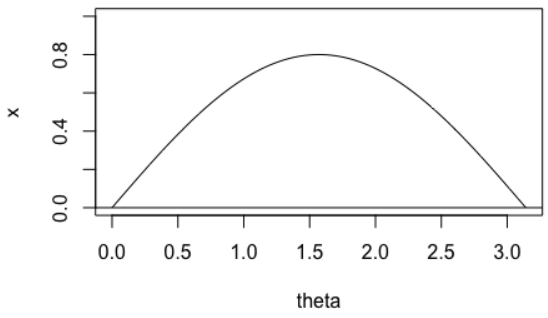

Fig. 7.10 "Rectangle" of sides (a,π)

Therefore, if we experimentally obtain p , from a sufficiently large number of throws, it is possible to estimate the value of π:

$$\pi = \frac{2L}{aP}$$

The R code below estimates the distribution of π in `Nruns` repetitions of the experiment that consist in throwing the needle `N` times.

```
##Ch7_4.R
## Buffon's needle problem

# a is the half-distance between lines
a <- 1
# L is the half-length of the needle
L <- 0.8

# repeat the experiment Nruns times
Nruns <- 200
result <- rep(0,Nruns)

for (k in 1:Nruns){
       #N is the number of throws of the needle
       N <- 100000
       #This instruction runif generates numbers with uniform distribution
       #with intervals from min (0) to max (2a)
       x <- runif(N,0,2*a)

       # The distance x is defined as that from the nearest line
       for (i in 1:N)
       if(x[i]>a) x[i] <- 2*a-x[i]
```

```
        #This instruction runif generates angles with uniform distribution
        #with intervals from min (0) to max (pi)
        theta <- runif(N,0,pi)

        L_sin_theta <- L*sin(theta)

        # compute frequency of line-crosses
        x<=L_sin_theta -> test
        p <- sum(test==TRUE)/N

        result[k] <-2*L/(a*p)
}

hist(result,main="Distribution of PI")
```

The results is depicted in Figure 7.4

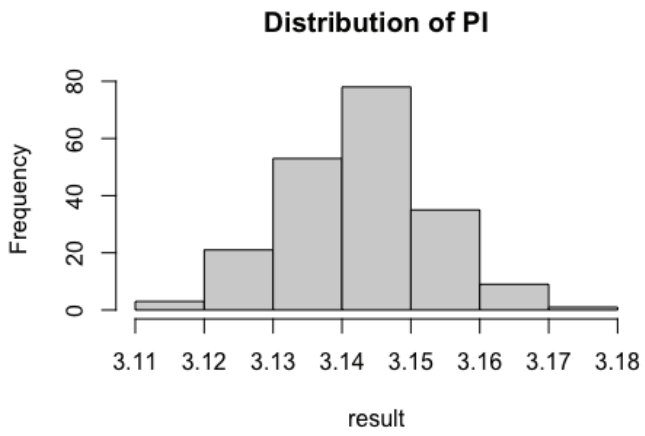

Fig. 7.11

7.5 Probability Distribution

A probability distribution is a table of all disjoint outcomes and their associated probabilities. The questions could be: what are the elements of the sample space $\mathcal{S}$? What is the probability that the sum of two faces is equal to 5 ? and so forth. The total number of possible combinations, therefore the sample space is $\mathcal{S} = 6 \times 6 = 36$.

A continuous probability distribution is a function that describes the assignment of probabilities to the occurences of values taken by a variable. It is usually called **probability density function (PDF)**. The area of the probability density function is

$$\int_{-\infty}^{+\infty} f(x)dx = 1 \tag{7.20}$$

Figure 7.5 shows the probability density function for a normal distribution with $\mu = 0$ and $\sigma = 1$. The area under the curve is equal to 1.

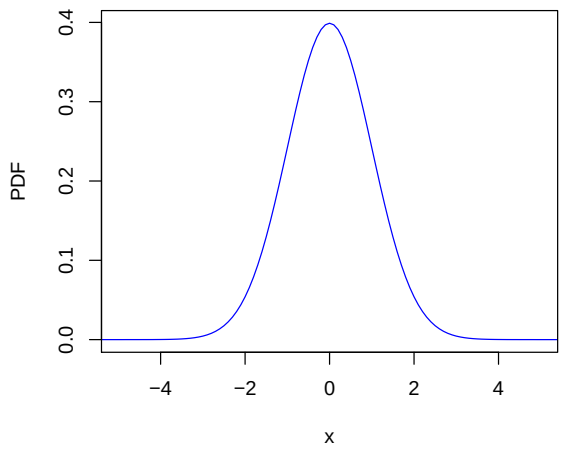

Fig. 7.12 Probability density function for a normal distribution having mean=0 and standard deviation = 1.

The balancing point of the curve is called **Expectation (E)**:

$$E[x] = \int_{-\infty}^{+\infty} xf(x)dx \tag{7.21}$$

Expectation works on a random number. It defines the **centroidal axis** of the PDF and is also known as the **mean** of x, $m(x) = E(x)$. A common measure of the dispersion of the distribution of x is the **variance** of x:

$$var[x] = E\left([x - m(x)]^2\right) = \int_{\text{all } x} [x - m(x)]^2 f(x)dx \tag{7.22}$$

The **covariance** is defined as:

$$\text{Cov}[x, y] = E\left([x - E(x)][y - E(y)]\right) \tag{7.23}$$

the values of $E(x)$ and $E(y)$ are compareable to the mean, the center of gravity of the distribution.

The PDF can be integrated to obtain a cumulative distribution curve (CDF) as depicted in Figure 7.5

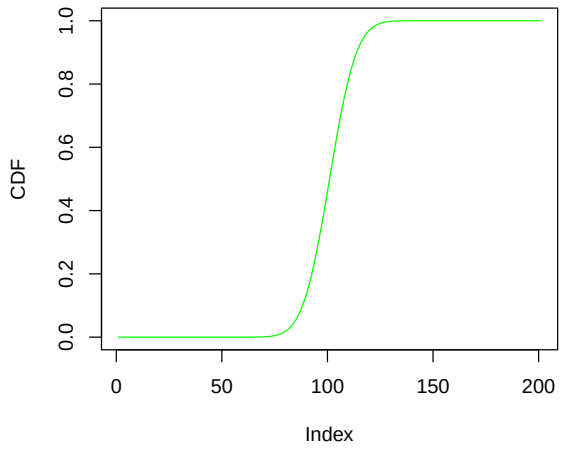

Fig. 7.13 Cumulative density function for a normal distribution having mean=0 and standard deviation = 1. Note that the cumulative curve reaches a value of 1, corresponding to the total area of the PDF.

Random does not mean uniform, normal or other distribution. Many types of distribution are possible for a random variable.

The code to plot the graphs for the PDF and CDF is shown below.

```
# Create a sequence of numbers between -10 and 10 incrementing by 0.1.
x <- seq(-10, 10, by = .1)

# Choose the mean as 0 and standard deviation as 1.
y <- dnorm(x, mean = 0, sd = 1)
y_cum<-pnorm(x, mean = 0, sd = 1)

plot(x,y,type="l",col="blue",xlim = c(-5,5),ylab=("PDF"))
plot(y_cum,type="l",col="green",ylab=("CDF"))
```

In the next chapter several distribution functions will be described.

Example. The probability of the occurrences of the sum of two dice is presented in Table 7.2. The total number of possible combinations is given by $6 \times 6 = 36$. Clearly the probability of obtaining 2 is only $1/36$ since it can be obtained only by $1 + 1$. On the other hand the combinations to obtain 3 are given by $2 + 1$ or $1 + 2$, therefore it is $2/36$ and so forth.

Table 7.2 Data of income depending on 5 measured features for 20 individuals

Dice Sum	Probability
2	1/36
3	2/36
4	3/36
5	4/36
6	5/36
7	6/36
8	5/36
9	4/36
10	3/36
11	2/36
12	1/36

In the code below, two vectors for the sum and the probabilities as reported in Table 7.2, are created and binded into a vector called `prob.dist`.

```
sum<-c(2,3,4,5,6,7,8,9,10,11,12)
prob<-c(1/36,2/36,3/36,4/36,5/36,6/36,5/36,4/36,3/36,2/36,1/36)
prob.dist<-c(sum,prob)
barplot(prob,names.arg=sum, main="",
xlab="Sum of two Dice",col="green")
```

Then a bar plot of the distribution is plotted in Fig. 7.5

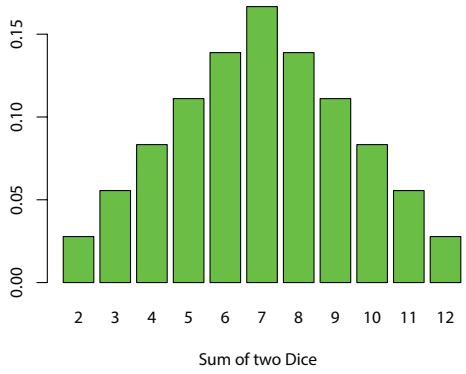

Fig. 7.14 Probability distribution for the sum of two dice.

7.6 Exercises

1. From a pack of well-shuffled 52 cards randomly pick a card. Compute the probability to pick a heart or an ace. Hint: the card ace of heart is common to both sets, therefore the sets are not mutually exclusive. content...