Joaquim P. Marques de Sá

Chance

Joaquim P. Marques de Sá

CHANCE

The Life of Games
&
the Game of Life

With 105 Figures and 10 Tables

Springer

Prof. Dr. Joaquim P. Marques de Sá

ISBN 978-3-540-74416-0 e-ISBN 978-3-540-74417-7

DOI 10.1007/978-3-540-74417-7

Library of Congress Control Number: 2007941508

Typesetting: LE-TEX Jelonek, Schmidt & Vöckler GbR, Leipzig, Germany
Production: LE-TEX Jelonek, Schmidt & Vöckler GbR, Leipzig, Germany
Cover design: eStudio Calamar S.L., F. Steinen-Broo, Girona, Spain
Cartoons by: Luís Gonçalves and Chris Madden

Printed on acid-free paper

9 8 7 6 5 4 3 2 1

springer.com

Preface

Our lives are immersed in a sea of chance. Everyone's existence is a meeting point of a multitude of accidents. The origin of the word 'chance' is usually traced back to the vulgar Latin word 'cadentia', meaning a befalling by fortuitous circumstances, with no knowable or determinable causes. The Roman philosopher Cicero clearly expressed the idea of 'chance' in his work *De Divinatione*:

> For we do not apply the words 'chance', 'luck', 'accident' or 'casualty' except to an event which has so occurred or happened that it either might not have occurred at all, or might have occurred in any other way. 2.VI.15.
>
> For if a thing that is going to happen, may happen in one way or another, indifferently, chance is predominant; but things that happen by chance cannot be certain. 2.IX.24.

In a certain sense chance is the spice of life. If there were no phenomena with unforeseeable outcomes, phenomena with an element of chance, all temporal cause–effect sequences would be completely deterministic. In this way, with sufficient information, the events of our daily lives would be totally predictable, whether it be the time of arrival of the train tomorrow or the precise nature of what one will be doing at 5.30 pm on 1 April three years from now. All games of chance, such as dice-throwing games for example, would no longer be worth playing! Every player would be able to control beforehand, deterministically, how many points to obtain at each throw. Playing the stock markets would no longer be hazardous. Instead, it would be a mere contractual operation between the players. Every experimental measurement would produce an exact result and as a consequence we would be far more advanced in our knowledge of the laws of the universe. To put it briefly, we would be engaged in a long and tedious walk of eternal predictability. Of course, there would also be some pleasant consequences. Car accidents would no longer take place, many deaths would be avoided by early therapeutic action, and economics would become an exact science.

In its role as the spice of life, chance is sweetness and bitterness, fortune and ruin. This double nature drove our ancestors to take it as a divine property, as the visible and whimsical emanation of the will of the gods. This imaginary will, supposedly conditioning every event, was called *sors* or *sortis* by the Romans. This word for fatalistic luck is present in the Latin languages ('sorte' in Portuguese, 'suerte' in Spanish, and 'sort' in French). But it was not only the Romans and peoples of other ancient civilizations who attributed this divine, supernatural feeling to this notion of chance. Who has never had the impression that they are going through a wave of misfortune, with the feeling that the 'wave' has something supernatural about it (while in contrast a wave of good fortune is often overlooked)? Being of divine origin, chance was not considered a valid subject of study by the sages of ancient civilizations. After all, what would be the sense in studying divine designs? Phenomena with an element of chance were simply used as a means of probing the will of the gods. Indeed, dice were thrown and entrails were inspected with this very aim of discerning the will of the gods. These were then the only reasonable experiments with chance. Even today fortune-telling by palm reading and tarot are a manifestation of this long-lived belief that chance offers a supernatural way of discerning the divine will.

Besides being immersed in a sea of chances, we are also immersed in a sea of regularity. But much of this regularity is intimately related to chance. In fact, there are laws of order in the laws of chance, whose discovery was initiated when several illustrious minds tried to clarify certain features that had always appeared obscure in popular games. Human intellect then initiated a brilliant journey that has led from probability theory to many important and modern fields, like statistical learning theory, for example. Along the way, a vast body of knowledge and a battery of techniques have been developed, including statistics, which has become an indispensable tool both in scientific research and in the planning of social activities. In this area of application, the laws of chance phenomena that were discovered have also provided us with an adequate way to deal with the risks involved in human activities and decisions, making our lives safer, even though the number of risk factors and accidents in our lives is constantly on the increase. One need only think of the role played by insurance companies.

This book provides a quick and almost chronological tour along the long road to understanding chance phenomena, picking out the most relevant milestones. One might think that in seven thousand years of history mankind would already have solved practically all the problems arising from the study, analysis and interpretation of chance phenomena. However, many milestones have only recently been passed and a good number of other issues remain open. Among the many roads of scientific discovery, perhaps none is so abundant in amazing and counterintuitive results. We shall see a wide range of examples to illustrate this throughout the book, many of which relate to important practical issues.

In writing the book I have tried to reduce to a minimum the mathematical prerequisites assumed of the reader. However, a few notes are included at the end of the book to provide a quick reference concerning certain topics for the sake of a better understanding. I also include some bibliographic references to popular science papers or papers that are not too technically demanding.

The idea of writing this book came from my work in areas where the influence of chance phenomena is important. Many topics and examples treated in the book arose from discussions with members of my research team. Putting this reflection into writing has been an exhilarating task, supported by the understanding of my wife and son.

Porto, *J.P. Marques de Sá*
October 2007

Contents

1

Probabilities and Games of Chance

1.1 Illustrious Minds Play Dice

A well-known Latin saying is the famous *alea jacta est* (the die has been cast), attributed to Julius Caesar when he took the decision to cross with his legions the small river Rubicon, a frontier landmark between the Italic Republic and Cisalpine Gaul. The decision taken by Julius Caesar amounted to invading the territory (Italic Republic) assigned to another triumvir, sharing with him the Roman power. Civil war was an almost certain consequence of such an act, and this was in fact what happened. Faced with the serious consequences of his decision, Caesar found it wise before opting for this line of action to allow the divine will to manifest itself by throwing dice. Certain objects appropriately selected to generate unpredictable outcomes, one might say 'chance' generators, such as astragals (specific animal bones) and dice with various shapes, have been used since time immemorial either to implore the gods in a neutral manner, with no human bias, to manifest their will, or purely for the purposes of entertainment.

The Middle Ages brought about an increasing popularization of dice games to the point that they became a common entertainment in both taverns and royal courts. In 1560, the Italian philosopher, physician, astrologist and mathematician Gerolamo Cardano (1501–1576), author of the monumental work *Artis Magnae Sive de Regulis Algebraicis* (*The Great Art or the Algebraic Rules*) and co-inventor of complex numbers, wrote a treatise on dice games, *Liber de Ludo Aleae* (*Book of Dice Games*), published posthumously in 1663, in which he introduced the concept of probability of obtaining a specific die face, although he did not explicitly use the word 'probability'. He also wove together several ideas on how to obtain specific face combinations when throwing dice.

Nowadays, Gerolamo Cardano (known as Jérôme Cardan, in French) is better remembered for one of his inventions: the cardan joint (originally used to keep ship compasses horizontal).

Dice provided a readily accessible and easy means to study the 'rules of chance', since the number of possible outcomes (chances) in a game with one or two dice is small and perfectly known. On the other hand, the incentive for figuring out the best strategy for a player to adopt when betting was high. It should be no surprise then that several mathematicians were questioned by players interested in the topic. Around 1600, Galileo Galilei (1564–1642) wrote a book entitled *Sopra le Scoperte dei Dadi* (*Analysis of Dice Games*), where among other topics he analyses the problem of how to obtain a certain number of points when throwing three dice. The works of Cardano and Galileo, besides presenting some counting techniques, also reveal concepts such as event equiprobability (equal likelihood) and expected gain in a game.

Sometime around 1654, Blaise Pascal (1623–1662) and Pierre-Simon de Fermat (1601–1665) exchanged letters discussing the mathematical analysis of problems related to dice games and in particular a series of tough problems raised by the Chevalier de Méré, a writer and nobleman from the court of French king Louis XIV, who was interested in the mathematical analysis of games. As this correspondence unfolds, Pascal introduces the classic concept of probability as the ratio between the number of favorable outcomes over the number of possible outcomes. A little later, in 1657, the Dutch physicist Christiaan Huygens (1629–1695) published his book *De Ratiociniis in Ludo Aleae* (*The Theory of Dice Games*), in which he provides a systematic discussion of all the results concerning dice games as they were understood in his day. In this way, a humble object that had been used for at least two thousand years for religious or entertainment purposes inspired human thought to understand chance (or random) phenomena. It also led to the name given to random phenomena in the Latin languages: 'aleatório' in Portuguese and 'aleatoire' in French, from the Latin word 'alea' for die.

1.2 The Classic Notion of Probability

Although the notion of probability – as a degree-of-certainty measure associated with a random phenomenon – reaches back to the works of Cardano, Pascal and Huygens, the word 'probability' itself was only used for the first time by Jacob (James) Bernoulli (1654–1705), professor of mathematics in Basel and one of the great scientific personalities of his day, in his fundamental work *Ars Conjectandi* (*The Art of Conjecture*), published posthumously in 1713, in which a mathematical theory of probability is presented. In 1812, the French mathematician Simon de Laplace (1749–1827) published his *Théorie Analytique des Probabilités*, in which the classic definition of probability is stated explicitly:

> Pour étudier un phénomène, il faut réduire tous les événements du même type à un certain nombre de cas également possibles, et alors la probabilité d'un événement donné est une fraction, dont le numérateur représente le nombre de cas favorables à l'événement et dont le dénominateur représente par contre le nombre des cas possibles.
>
> To study a phenomenon, one must reduce all events of the same type to a certain number of equally possible cases, and then the probability is a fraction whose numerator represents the number of cases favorable to the event and whose denominator represents the number of possible cases.

Therefore, in order to determine the probability of a given outcome (also called 'case' or 'event') of a random phenomenon, one must first reduce the phenomenon to a set of elementary and equally probable outcomes or 'cases'. One then proceeds to determine the ratio between the number of cases favorable to the event and the total number of possible cases. Denoting the probability of an event by P (event), this gives

$$P(\text{event}) = \frac{\text{number of favorable cases}}{\text{number of possible cases}} .$$

Let us assume that in the throw of a die we wish to determine the probability of obtaining a certain face (showing upwards). There are 6 possible outcomes or elementary events (the six faces), which we represent by the set $\{1, 2, 3, 4, 5\}$. These are referred to as *elementary* since they cannot be decomposed into other events. Moreover, assuming that we are dealing with a fair die, i.e., one that has not been tampered with, and also that the throws are performed in a fair manner, any of the six faces is equally probable. If the event that interests us is 'face 5 turned up', since there is only one face with this value (one favorable event), we have P (face 5 turned up) $= P(5) = 1/6$, and likewise for the other faces. Interpreting probability values as percentages, we may say that there is 16.7% certainty that any previously specified die face will turn up. Figure 1.1 displays the various probability values P for a fair

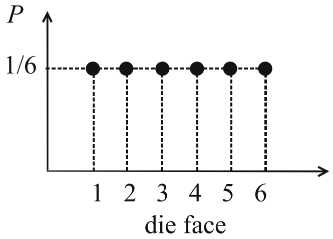


Fig. 1.1. Probability value of a given face turning up when throwing a fair die

die and throws, the so-called *probability function* of a fair die. The sum of all probabilities is $6 \times (1/6) = 1$, in accord with the idea that there is 100% certainty that one of the faces will turn up (a *sure event*).

The fairness of chance-generating devices, such as dice, coins, cards, roulettes, and so on, and the fairness of their use, will be discussed later on. For the time being we shall just accept that the device configuration and its use will not favor any particular elementary event. This is henceforth our basic assumption unless otherwise stated. When tossing a coin up in the air (a chance decision-maker much used in football matches), the set of possible results is $\{\text{heads}, \text{tails}\}$. For a fair coin, tossed in a fair way, we have $P(\text{heads}) = P(\text{tails}) = 1/2$. When extracting a playing card at random from a 52 card deck, the probability of a card with a specific value and suit turning up is 1/52. If, on the other hand, we are not interested in the suit, so that the set of elementary events is

$$\{\text{ace}, \text{king}, \text{queen}, \text{jack}, 10, 9, 8, 7, 6, 5, 4, 3, 2\} \ ,$$

the probability of a card of a given value turning up is 1/13.

Let us now consider a die with two faces marked with 5, perhaps as a result of a manufacturing error. In this case the event 'face 5 turned up' can occur in two distinct ways. Let us assume that we have written the letter a on one of the faces with a 5, and b on the other. The set of equiprobable elementary events is now $\{1, 2, 3, 4, 5a, 5b\}$ (and not $\{1, 2, 3, 4, 5\}$). Hence, $P(5) = 2/6 = 1/3$, although 1/6 is still the probability for each of the remaining faces and the sum of all elementary probabilities is still of course equal to 1.

1.3 A Few Basic Rules

The great merit of the probability measure lies in the fact that, with only a few basic rules directly deducible from the classic definition, one can compute the probability of any event composed of finitely many elementary events. The quantification of chance then becomes a feasible task. For instance, in the case of a die, the event 'face value higher than 4' corresponds to putting together (finding the union of) the elementary events 'face 5' and 'face 6', that is:

$$\text{face value higher than } 4 = \text{face } 5 \text{ or face } 6 = \{5, 6\} \ .$$

Note the word 'or' indicating the union of the two events. What is the value of $P(\{5, 6\})$? Since the number of favorable cases is 2, we have

$$P(\text{face value higher than } 4) = P(\{5, 6\}) = \frac{2}{6} = \frac{1}{3} \ .$$

It is also obvious that the degree of certainty associated with $\{5, 6\}$ must be the sum of the individual degrees of certainty. Hence,

$$P(\{5, 6\}) = P(5) + P(6) = \frac{1}{6} + \frac{1}{6} = \frac{1}{3} \ .$$

The probability of the union of elementary events computed as the sum of the respective probabilities is then confirmed. In the same way,

$$P(\text{even face}) = P(2) + P(4) + P(6) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2} \ ,$$

$$P(\text{any face}) = 6 \times \frac{1}{6} = 1 \ .$$

The event 'any face' is the *sure event* in the die-throwing experiment, that is, the event that always happens. The probability reaches its highest value for the sure event, namely 100% certainty.

Let us now consider the event 'face value lower than or equal to 4'. We have

$$\begin{aligned} P\begin{pmatrix} \text{face value lower than} \\ \text{or equal to } 4 \end{pmatrix} &= P(1) + P(2) + P(3) + P(4) \\ &= \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{4}{6} = \frac{2}{3} \ . \end{aligned}$$

But one readily notices that 'face value lower than or equal to 4' is the opposite (or negation) of 'face value higher than 4'. The event 'face value lower than or equal to 4' is called the *complement* of 'face value higher than 4' relative to the sure event. Given two complementary events denoted by A and $\bar{A}$, where $\bar{A}$ means the complement of A, if we know the probability of one of them, say $P(A)$, the probability of the other is obtained by *subtracting* it from 1 (the probability of the certain event): $P(\bar{A}) = 1 - P(A)$. This rule is a direct consequence of the fact that the sum of cases favorable to A and $\bar{A}$ is the total number of possible cases. Therefore,

$$\begin{aligned} P\left(\begin{array}{c}\text{face value less than}\\ \text{or equal to 4}\end{array}\right) &= 1 - P(\text{face value higher than 4}) \\ &= 1 - \frac{1}{3} = \frac{2}{3}\,, \end{aligned}$$

which is the result found above.

From the rule about the complement, one concludes that the probability of the complement of the sure event is $1 - P(\text{sure event}) = 1 - 1 = 0$, the minimum probability value. Such an event is called the *impossible event*, in the case of the die, the event 'no face turns up'.

Let us now consider the event 'even face and less than or equal to 4'. The probability of this event, composed of events 'even face' and 'face smaller than or equal to 4', is obtained by the *product* of the probabilities of the individual events. Thus,

$$\begin{aligned} P\left(\begin{array}{c}\text{even face and less than}\\ \text{or equal to 4}\end{array}\right) &= P(\text{even face}) \times P\left(\begin{array}{c}\text{face less than}\\ \text{or equal to 4}\end{array}\right) \\ &= \frac{1}{2} \times \frac{2}{3} = \frac{1}{3}\,. \end{aligned}$$

As a matter of fact there are two possibilities out of six of obtaining a face that is both even and less than or equal to 4 in value, the two possibilities forming the set $\{2, 4\}$. Note the word 'and' indicating event composition (or intersection).

These rules are easily visualized by the simple ploy of drawing a rectangle corresponding to the set of all elementary events, and representing the latter by specific domains within it. Such drawings are called Venn diagrams (after the nineteenth century English mathematician John Venn). Figure 1.2 shows diagrams corresponding to the previous examples, clearly illustrating the consistency of the classic definition. For instance, the intersection in Fig. 1.2c can be seen as taking away half of certainty (hatched area in the figure) from something that had only two thirds of certainty (gray area in the same figure), thus yielding one third of certainty.

Finally, let us consider one more problem. Suppose we wish to determine the probability of 'even face or less than or equal to 4' turning

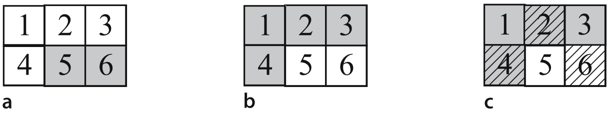


Fig. 1.2. Venn diagrams. (**a**) Union (*gray area*) of face 5 with face 6. (**b**) Complement (*gray area*) of 'face 5 or face 6'. (**c**) Intersection (*gray hatched area*) of 'even face' (*hatched area*) with 'face value less than or equal to 4' (*gray area*)

up. We are thus dealing with the event corresponding to the union of the following two events: 'even face', 'face smaller than or equal to 4'. (What we discussed previously was the intersection of these events, whereas we are now considering their union.) Applying the addition rule, we have

$$\begin{aligned} P\left(\begin{matrix}\text{even face or less than}\\ \text{or equal to 4}\end{matrix}\right) &= P(\text{even face}) + P\left(\begin{matrix}\text{face less than}\\ \text{or equal to 4}\end{matrix}\right)\\ &= \frac{1}{2}+\frac{2}{3}=\frac{7}{6}>1\,. \end{aligned}$$

We have a surprising and clearly incorrect result, greater than 1! What is going on? Let us take a look at the Venn diagram in Fig. 1.2c. The union of the two events corresponds to all squares that are hatched, gray or simultaneously hatched and gray. There are five. We now see the cause of the problem: we have counted some of the squares twice, namely those that are simultaneously hatched and gray, i.e., the two squares in the intersection. We thus obtained the incorrect result 7/6 instead of 5/6. We must, therefore, modify the addition rule for two events A and B in the following way:

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)\,.$$

Let us recalculate in the above example:

$$\begin{aligned} P\left(\begin{matrix}\text{even face or less than}\\ \text{or equal to 4}\end{matrix}\right) &= P(\text{even face}) + P\left(\begin{matrix}\text{face less than}\\ \text{or equal to 4}\end{matrix}\right)\\ &\quad -P\left(\begin{matrix}\text{even face and less}\\ \text{than or equal to 4}\end{matrix}\right)\\ &= \frac{1}{2}+\frac{2}{3}-\frac{2}{6}=\frac{5}{6}\,. \end{aligned}$$

In the first example of a union of events, when we added the probabilities of elementary events, we did not have to subtract the probability of the intersection because there was no intersection, i.e., the intersection was empty, and thus had null probability, being the impossible event. Whenever two events have an empty intersection they are called disjoint and the probability of their union is the sum of the probabilities. For instance,

$$P(\text{even or odd face}) = P(\text{even face}) + P(\text{odd face}) = 1 \;.$$

In summary, the classic definition of probability affords the means to measure the 'degree of certainty' of a random phenomenon on a convenient scale from 0 to 1 (or from 0 to 100% certainty), obeying the following rules in accordance with common sense:

- $P(\text{sure event}) = 1\,,$
- $P(\text{impossible event}) = 0\,,$
- $P(\text{complement event of A}) = 1 - P(A)\,,$
- $P(A \text{ and } B) = P(A) \times P(B)$ (rule to be revised later),
- $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)\,.$

1.4 Testing the Rules

There is nothing special about dice games, except that they may originally have inspired the study of random phenomena. So let us test our ability to apply the rules of probability to other random phenomena.

We will often talk about tossing coins in this book. As a matter of fact, everything we know about chance can be explained by reference to coin tossing! The first question is: what is the probability of heads (or tails) turning up when (fairly) tossing a (fair) coin in the air? Since both events are equally probable, we have $P(\text{heads}) = P(\text{tails}) = 1/2$ (50% probability). It is difficult to believe that there should be anything more to say about coin tossing. However, as we shall see later, the hardest concepts of randomness can be studied using coin-tossing experiments.

Let us now consider a more mundane situation: the subject of families with two children. Assuming that the probability of a boy being born is the same as that of a girl being born, what is the probability of a family with two children having at least one boy? Denoting the birth of a boy by M and the birth of a girl by F, there are 4 elementary events:

$$\{MM, MF, FM, FF\} \;,$$

where MF means that first a boy is born, followed later by a girl, and similarly for the other cases. If the probability of a boy being born is the same as that for a girl, each of these elementary events has probability 1/4. In fact, we can look at these events as if they were event intersections. For instance, the event MF can be seen as the intersection of the events 'first a boy is born' and 'second a girl is born'. MF is called a *compound event*. Applying the intersection rule, we get $P(MF) = P(M) \times P(F) = 1/2 \times 1/2 = 1/4$. The probability we wish to determine is then

$$P(\text{at least one boy}) = P(MM) + P(MF) + P(FM) = 3/4 .$$

A common mistake in probability computation consists in not correctly representing the set of elementary events. For example, take the following argument. For a family with two children there are three possibilities: both boys; both girls; one boy and one girl. Only two of these three possibilities correspond to having at least one boy. Therefore, the probability is 2/3. The mistake here consists precisely in not taking into account the fact that 'one boy and one girl' corresponds to two equiprobable elementary events: MF and FM.

Let us return to the dice. Suppose we throw two dice. What is the probability that the sum of the faces is 6? Once again, we are dealing with compound experiments. Let us name the dice as die 1 and die 2. For any face of die 1 that turns up, any of the six faces of die 2 may turn up. There are therefore $6 \times 6 = 36$ possible and equally probable outcomes (assuming fair dice and throwing). Only the following amongst them favor the event that interests us here, viz., sum equal to 6: $\{15, 24, 33, 42, 51\}$ (where 15 means die 1 gives 1 and die 2 gives 5 and likewise for the other cases). Therefore,

$$P(\text{sum of the faces is 6}) = 5/36 .$$

The number of throws needed to obtain a six with one die or two sixes with two dice was a subject that interested the previously cited Chevalier de Méré. The latter betted that a six would turn up with one die in four throws and that two sixes would turn up with two dice in 24 throws. He noted, however, that he constantly lost when applying this strategy with two dice and asked Pascal to help him to explain why. Let us then see the explanation of the De Méré paradox, which constitutes a good illustration of the complement rule. The probability of not obtaining a 6 (or any other previously chosen face for that matter) in a one-die throw is

$$P(\text{no } 6) = 1 - \frac{1}{6} = \frac{5}{6} .$$

Therefore, the probability of not obtaining at least one 6 in four throws is given by the intersection rule:

$$\begin{aligned} &P(\text{no 6 in 4 throws}) \\ &\qquad = P(\text{no 6 in first throw}) \times P(\text{no 6 in second throw}) \\ &\qquad\quad \times P(\text{no 6 in third throw}) \times P(\text{no 6 in fourth throw}) \\ &\qquad = \frac{5}{6} \times \frac{5}{6} \times \frac{5}{6} \times \frac{5}{6} = \left(\frac{5}{6}\right)^4 . \end{aligned}$$

Hence, by the complement rule,

$$\begin{aligned} P(\text{at least one 6 in 4 throws}) &= 1 - P(\text{no 6 in 4 throws}) \\ &= 1 - \left(\frac{5}{6}\right)^4 = 0.51775 . \end{aligned}$$

We thus conclude that the probability of obtaining at least one 6 in four throws is higher than 50%, justifying a bet on this event as a good tactic. Let us list the set of elementary events corresponding to the four throws:

$$\{1111, 1112, \ldots, 1116, 1121, \ldots, 1126, \ldots, 1161, \ldots, 1166, \ldots, 6666\} .$$

It is a set with a considerable number of elements, in fact, 1296. To count the events where a 6 appears at least once is not an easy task. We should have to count the events where only one six turns up (e.g., event 1126), as well as the events where two, three or four sixes appear. We saw above how, by thinking in terms of the complement of the event we are interested in, we were able to avoid the difficulties inherent in

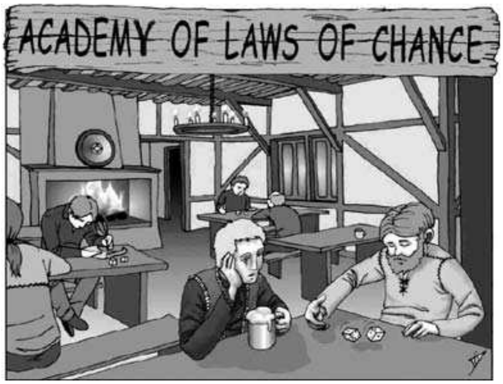


Fig. 1.3. Putting the laws of chance to the test

such counting problems. Let us apply this same technique to the two dice:

$$P(\text{no double 6 with two dice}) = 1 - \frac{1}{36} = \frac{35}{36} ,$$

$$P(\text{no double 6 in 24 throws}) = \left(\frac{35}{36}\right)^{24} ,$$

$$P(\text{at least one double 6 in 24 throws}) = 1 - \left(\frac{35}{36}\right)^{24} = 0.49141 .$$

We obtain a probability lower than 50%: betting on a double 6 in 24 two-die throws is not a good policy. One can thus understand the losses incurred by Chevalier de Méré!

1.5 The Frequency Notion of Probability

The notions and rules introduced in the previous sections assumed fair dice, fair coins, equal probability of a boy or girl being born, and so on. That is, they assumed specific probability values (in fact, equiprobability) for the elementary events. One may wonder whether such values can be reasonably assumed in normal practice. For instance, a fair coin assumes a symmetrical mass distribution around its center of gravity. In normal practice it is impossible to guarantee such a distribution. As to the fair throw, even without assuming a professional gambler gifted with a deliberately introduced flick of the wrist, it is known that some faulty handling can occur in practice, e.g., the coin sticking to a hand that is not perfectly clean, a coin rolling only along one direction, and so on. In reality it seems that the tossing of a coin in the air is usually biased, even for a fair coin that is perfectly symmetrical and balanced. As a matter of fact, the American mathematician Persi Diaconis and his collaborators have shown, in a paper published in 2004, that a fair tossing of a coin can only take place when the impulse (given with the finger) is exactly applied on a diameter line of the coin. If there is a deviation from this position there is also a probability higher than 50% that the face turned up by the coin is the same as (or the opposite of) the one shown initially. For these reasons, instead of assuming more or less idealized conditions, one would like, sometimes, to deal with true values of probabilities of elementary events.

The French mathematician Abraham De Moivre (1667–1754), in his work *The Doctrine of Chance: A Method of Calculating the Probabilities of Events in Play*, published in London in 1718, introduced the frequency or frequentist theory of probability. Suppose we wish to determine the probability of a 5 turning up when we throw a die in the air. Then we can do the following: repeat the die-throwing experiment a certain number of times denoted by n, let k be the number of times that face 5 turns up, and compute the ratio

$$f = \frac{k}{n} \; .$$

The value of f is known as the *frequency of occurrence* of the relevant event in n repetitions of the experiment, in this case, the die-throwing experiment.

Let us return to the birth of a boy or a girl. There are natality statistics in every country, from which one can obtain the birth frequencies of boys and girls. For instance, in Portugal, out of 113 384 births occurring in 1998, 58 506 were of male sex. Therefore, the male frequency of birth was $f = 58\,506/113\,384 = 0.516$. The female frequency of birth was the complement, viz., $1 - 0.516 = 0.484$.

One can use the frequency of an event as an *empirical probability measure*, a counterpart of the ideal situation of the classic definition. If the event always occurs (sure event), then k is equal to n and $f = 1$. In the opposite situation, if the event never occurs (impossible event), k is equal to 0 and $f = 0$. Thus, the frequency varies over the same range as the probability.

Let us now take an event such as ‘face 5 turns up’ when throwing a die. As we increase n, we verify that the frequency f tends to wander with decreasing amplitude around a certain value. Figure 1.4a shows how the frequencies of face 5 and face 6 might evolve with n in two distinct series of 2 000 throws of a fair die. Some wanderings are clearly visible, each time with smaller amplitude as n increases, reaching a reasonable level of stability near the theoretical value of 1/6. Figure 1.4b shows the evolution of the frequency for a new series of 2 000 throws, now relative to faces 5 or 6 turning up. Note how the stabilization takes place near the theoretical value of 1/3. Frequency does therefore exhibit the additivity required for a probability measure, for sufficiently large n.

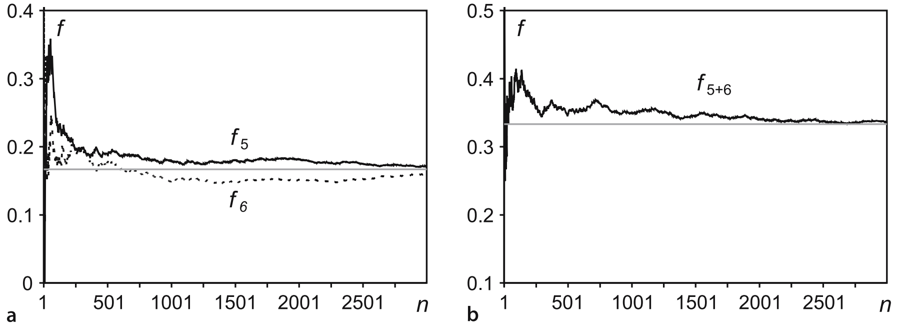


Fig. 1.4. Possible frequency curves in die throwing. (**a**) Face 5 and face 6. (**b**) Face 5 or 6

Let us assume that for a sufficiently high value of n, say 2 000 throws, we obtain $f = 0.16$ for face 5. We then assign this value to the probability of face 5 turning up. This is the value that we evaluate empirically (or *estimate*) for $P(\text{face } 5)$ based on the 2 000 throws. Computing the probabilities for the other faces in the same way, let us assume that we obtain:

$$P(\text{face } 1) = 0.17\ , \qquad P(\text{face } 2) = 0.175\ ,$$
$$P(\text{face } 3) = 0.17\ , \qquad P(\text{face } 4) = 0.165\ ,$$
$$P(\text{face } 5) = 0.16\ , \qquad P(\text{face } 6) = 0.16\ .$$

The elementary events are no longer equiprobable. The die is not fair (biased die) and has the tendency to turn up low values more frequently than high values. We also note that, since n is the sum of the different values of k respecting the frequencies of all the faces, the sum of the empirical probabilities is 1.

Figure 1.5 shows the probability function of this biased die. Using this function we can apply the same basic rules as before. For example,

$$P(\text{face higher than } 4) = P(5) + P(6) = 0.16 + 0.16 = 0.32\ ,$$

$$P\begin{pmatrix}\text{face lower than}\\ \text{or equal to } 4\end{pmatrix} = 1 - P\begin{pmatrix}\text{face higher}\\ \text{than } 4\end{pmatrix} = 1 - 0.32 = 0.68\ ,$$

$$P(\text{even face}) = P(2) + P(4) + P(6) = 0.175 + 0.165 + 0.16 = 0.5\ ,$$

$$P\begin{pmatrix}\text{even face and lower}\\ \text{than or equal to } 4\end{pmatrix} = 0.5 \times 0.68 = 0.34\ .$$

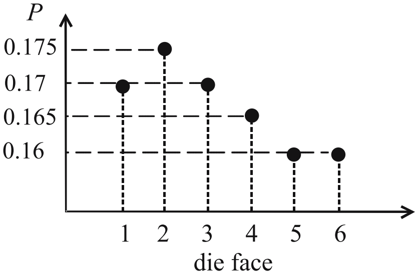


Fig. 1.5. Probability function of a biased die

In the child birth problem, frequencies of occurrence of the two sexes obtained from sufficiently large birth statistics can also be used as a probability measure. Assuming that the probability of boys and girls born in Portugal is estimated by the previously mentioned frequencies – $P(\text{boy}) = 0.516$, $P(\text{girl}) = 0.484$ – then, for the problem of families with two children presented before, we have for Portuguese families

$$\begin{aligned} P(\text{at least one boy}) &= P(MM) + P(MF) + P(FM) \\ &= 0.516 \times 0.516 + 0.516 \times 0.484 + 0.484 \times 0.516 \\ &= 0.766 \; . \end{aligned}$$

Some questions concerning the frequency interpretation of probabilities will be dealt with later. For instance, when do we consider n sufficiently large? For the time being, it is certainly gratifying to know that we can estimate degrees of certainty of many random phenomena. We make an exception for those phenomena that do not refer to constant conditions (e.g., the probability that Chelsea beats Manchester United) or those that are of a subjective nature (e.g., the probability that I will feel happy tomorrow).

1.6 Counting Techniques

Up to now we have computed probabilities for relatively simple events and experiments. We now proceed to a more complicated problem: computing the probability of winning the first prize of the football pools known as *toto* in many European countries. A toto entry consists of a list of 13 football matches for the coming week. Pool entrants have to select the result of each match, whether it will be a home win, an away win, or neither of these, typically by marking a cross over each of the three possibilities. In this case there is a single favorable event (the right template of crosses) in an event set that we can imagine to have quite a large number of possible events (templates). But how large is this set? To be able to answer such questions we need to know some counting techniques. Abraham de Moivre was the first to present a consistent description of such techniques in his book mentioned above. Let us consider how this works for toto. Each template can be seen as a list (or sequence) of 13 symbols, and each symbol can have one of three values: home win, home loss (away win) and draw, that we will denote, respectively, by W (win), L (loss), and D (draw). A possible template is thus

$$\text{WWLDDDWDDDLLW} .$$

How many such templates are there? First, assume that the toto has only one match. One would then have only three distinct templates, i.e., as many as there are symbols. Now let us increase the toto to two matches. Then, for each of the symbols of the first match, there are three different templates corresponding to the symbols of the second match. There are therefore $3 \times 3 = 9$ templates. Adding one more match to the toto, the next three symbols can be used with any of the previous templates, whence we obtain $3 \times 3 \times 3$ templates. This line of thought can be repeated until one reaches the 13 toto matches, which then correspond to $3^{13} = 1\,594\,323$ templates. This is a rather large number. The probability of winning the toto by chance alone is $1/1\,594\,323 = 0.000\,000\,627$. In general, the number of lists of n symbols from a set of k symbols is given by n^k.

We now consider the problem of randomly inserting 5 different letters into 5 envelopes. Each envelope is assumed to correspond to one and only one of the letters. What is the probability that all envelopes receive the right letter? Let us name the letters by numbers 1, 2, 3, 4, 5, and let us imagine that the envelopes are placed in a row, with no visible clues (sender and addressee), as shown in Fig. 1.6.

Figure 1.6 shows one possible letter sequence, namely, 12345. Another is 12354. Only one of the possible sequences is the right one. The question is: how many distinct sequences are there? Take the leftmost position. There are 5 distinct possibilities for the first letter. Consider one of them, say 3. Now, for the four rightmost positions, letter 3 is no longer available. (Note the difference in relation to the toto case:

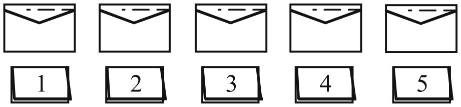


Fig. 1.6. Letters and envelopes

Table 1.1. Probability P of matching n letters

n	1	2	3	4	5	10	15
P	1	0.5	0.17	0.042	0.0083	0.000 000 28	0.000 000 000 000 78

the fact that a certain symbol was chosen as the first element of the list did not imply, as here, that it could not be used again in other list positions.) Let us then choose one of the four remaining letters for the second position, say, 5. Only three letters now remain for the last three positions. So for the first position we had 5 alternatives, and for the second position we had 4. This line of thought is easily extended until we reach a total number of alternatives given by

$$5 \times 4 \times 3 \times 2 \times 1 = 120 \; .$$

Each of the possible alternatives is called a *permutation* of the 5 symbols. Here are all the permutations of three symbols:

$$123 \quad 132 \quad 213 \quad 231 \quad 312 \quad 321 \; .$$

We have $3 \times 2 \times 1 = 6$ permutations. Note how the total number of permutations grows dramatically when one increases from 3 to 5 symbols. Whereas with 3 letters the probability of a correct match is around 17%, for 5 letters it is only 8 in one thousand.

In general, the number of *ordered sequences* (permutations) of n symbols is given by

$$n \times (n-1) \times (n-2) \times \ldots \times 1 \; .$$

This function of n is called the *factorial function* and written $n!$ (read n factorial). Table 1.1 shows the fast decrease of probability P of matching n letters, in correspondence with the fast increase of $n!$. For example, for $n = 10$ there are $10! = 3\,628\,800$ different permutations of the letters. If we tried to match the envelopes, aligning each permutation at a constant rate of one per second, we should have to wait 42 days in the worst case (trying all 10! permutations). For 5 letters, 2 minutes would be enough, but for 15 letters, 41 466 years would be needed!

Finally, let us consider tossing a coin 5 times in a row. The question now is: what is the probability of heads turning up 3 times? Denote heads and tails by 1 and 0, respectively. Assume that for each heads–tails configuration we take note of the position in which heads occurs, as in the following examples:

Sequence	Heads position
10011	1, 4, 5
01101	2, 3, 5
10110	1, 3, 4

Taking into account the set of all 5 possible positions, $\{1, 2, 3, 4, 5\}$, we thus intend to count the number of distinct subsets that can be formed using only 3 of those 5 symbols. For this purpose, we start by assuming that the list of all 120 permutations of 5 symbols is available and select as subsets the first 3 permutation symbols, as illustrated in the diagram on the right.

$$\begin{array}{cc} \overbrace{123}^{3} & 45 \\ 123 & 54 \\ 132 & 45 \\ 132 & 54 \\ \vdots & \vdots \\ 234 & 15 \\ 234 & 51 \\ \vdots & \vdots \end{array}$$

When we consider a certain subset, say $\{1, 2, 3\}$, we are not concerned with the order in which the symbols appear in the permutations. It is all the same whether it is 123, 132, 231, or any other of the $3! = 6$ permutations. On the other hand, for each sequence of the first three symbols, there are repetitions corresponding to the permutations of the remaining symbols; for instance, 123 is repeated twice, in correspondence with the $2! = 2$ permutations 45 and 54. In summary, the number of subsets that we can form using 3 of 5 symbols (without paying attention to the order) is obtained by dividing the 120 permutations of 5 symbols by the number of repetitions corresponding to the permutations of 3 and of $5 - 3 = 2$ symbols. The conclusion is that

$$\text{number of subsets of 3 out of 5 symbols} = \frac{5!}{3! \times (5-3)!} = \frac{120}{6 \times 2} = 10\,.$$

In general, the number of distinct subsets (the order does not matter) of k symbols out of a set of n symbols, denoted $\binom{n}{k}$, and called the number of *combinations of n, taking k at a time*, is given by

$$\binom{n}{k} = \frac{n!}{k! \times (n-k)!}\,.$$

Let us go back to the problem of the coin tossed 5 times in a row, using a fair coin with $P(1) = P(0) = 1/2$. The probability that heads turns up three times is $(1/2)^3$. Likewise, the probability that tails turns up twice

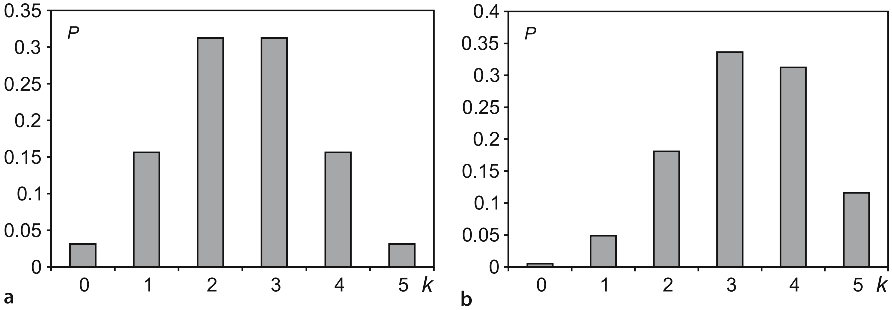


Fig. 1.7. Probability of obtaining k heads in 5 throws of a fair coin (**a**) and a biased coin (**b**)

is $(1/2)^2$. Therefore, the probability of a given sequence of three heads and two tails is computed as $(1/2)^3 \times (1/2)^2 = (1/2)^5$. Since there are $\binom{5}{3}$ distinct sequences on those conditions, we have (disjoint events)

$$P(3 \text{ heads in 5 throws}) = \binom{5}{3} \times \left(\frac{1}{2}\right)^5 = \frac{10}{32} = 0.31\ .$$

Figure 1.7a shows the probability function for several k values. Note that for $k = 0$ (no heads, i.e., the sequence 00000) and for $k = 5$ (all heads, i.e., the sequence 11111) the probability is far lower (0.031) than when a number of heads near half of the throws turns up (2 or 3).

Let us suppose that the coin was biased, with a probability of 0.65 of turning up heads (and, therefore, 0.35 of turning up tails). We have

$$P(3 \text{ heads in 5 throws}) = \binom{5}{3} \times (0.65)^3 \times (0.35)^2 = 0.336\ .$$

For this case, as expected, sequences with more heads are favored when compared with those with more tails, as shown in Fig. 1.7b. For instance, the sequence 00000 has probability 0.005, whereas the sequence 11111 has probability 0.116.

1.7 Games of Chance

Games of the kind played in the casino – roulette, craps, several card games and, in more recent times, slot machines and the like – together with any kind of national lottery are commonly called games of chance.

All these games involve a chance phenomenon that is only very rarely favorable to the player. Maybe this is the reason why in some European languages, for example, Portuguese and Spanish, these games are called games of bad luck, using for 'bad luck' a word that translates literally to the English word 'hazard'. The French also use 'jeux de hasard' to refer to games of chance. This word, meaning risk or danger, comes from the Arab *az-zahar*, which in its turn comes from the Persian *az zar*, meaning dice game.

Now that we have learned the basic rules for operating with probabilities, as well as some counting techniques, we will be able to understand the bad luck involved in games of chance in an objective manner.

1.7.1 Toto

We have already seen that the probability of a full template hit in the football toto is 0.000 000 627. In order to determine the probability of other prizes corresponding to 11 or 12 correct hits one only has to determine the number of different ways in which one can obtain the 11 or 12 correct hits out of the 13. This amounts to counting subsets of 11 or 12 symbols (the positions of the hits) of the 13 available positions. We have seen that such counts are given by $\binom{13}{12}$ and $\binom{13}{11}$, respectively. Hence,

$$\text{probability of 13 hits} \qquad (1/3)^{13} = 0.000\,000\,627\ ,$$

$$\text{probability of 12 hits} \qquad \binom{13}{12} \times (1/3)^{13} = 0.000\,008\,15\ ,$$

$$\text{probability of 11 hits} \qquad \binom{13}{11} \times (1/3)^{13} = 0.000\,048\,92\ .$$

In this example and the following, the reader must take into account the fact that the computed probabilities correspond to idealized models of reality. For instance, in the case of toto, we are assuming that the choice of template is made entirely at random with all 1 594 323 templates being equally likely. This is not usually entirely true.

As a comparison the probability of a first toto prize is about twice as big as the probability of a correct random insertion of 10 letters in their respective envelopes.

1.7.2 Lotto

The lotto or lottery is a popular form of gambling that often corresponds to betting on a subset of numbers. Let us consider the lotto that runs in several European countries consisting of betting on a subset of six numbers out of 49 (from 1 through 49). We already know how to compute the number of distinct subsets of 6 out of 49 numbers:

$$\binom{49}{6} = \frac{49!}{6! \times 43!} = \frac{49 \times 48 \times \ldots \times 1}{(6 \times 5 \times \ldots) \times (43 \times 42 \times \ldots)} = 13\,983\,816 \ .$$

Therefore, the probability of winning the first prize is

$$P(\text{first prize: 6 hits}) = \frac{1}{13\,983\,816} = 0.000\,000\,071\,5 \ ,$$

smaller than the probability of a first prize in the toto.

Let us now compute the probability of the third prize corresponding to a hit in 5 numbers. There are $\binom{6}{5} = 6$ different ways of selecting five numbers out of six, and the sixth number of the bet can be any of the remaining 43 non-winning numbers. Therefore, the probability of winning the third prize is $6 \times 43 = 258$ times greater than the probability of winning the first prize, that is,

$$P(\text{third prize: 5 hits}) = 258 \times 0.000\,000\,071\,5 = 0.000\,018\,4 \ .$$

The second prize corresponds to five hits plus a hit on a supplementary number, which can be any of the remaining 43 numbers. Thus, the probability of winning the second prize is 43 times smaller than that of the third prize:

$$\begin{aligned} P(\text{second prize: 5 hits + 1 supplementary hit}) &= 6 \times 0.000\,000\,071\,5 \\ &= 0.000\,000\,429 \ . \end{aligned}$$

In the same way one can compute

$$\begin{aligned} P(\text{fourth prize: 4 hits}) &= \binom{6}{4} \times \binom{43}{2} \times 0.000\,000\,071\,5 \\ &= 15 \times 903 \times 0.000\,000\,071\,5 = 0.000\,969 \ , \end{aligned}$$

$$\begin{aligned} P(\text{fifth prize: 3 hits}) &= \binom{6}{3} \times \binom{43}{3} \times 0.000\,000\,071\,5 \\ &= 20 \times 12\,341 \times 0.000\,000\,071\,5 = 0.017\,65 \ . \end{aligned}$$

The probability of winning the fifth prize is quite a lot higher than the probability of winning the other prizes and nearly twice as high as the probability of matching 5 letters to their respective envelopes.

1.7.3 Poker

Poker is played with a 52 card deck (13 cards of 4 suits). A player's hand is composed of 5 cards. The game involves a betting system that guides the action of the players (card substitution, staying in the game or passing, and so on) according to a certain strategy depending on the hand of cards and the behavior of the other players. It is not, therefore, a game that depends only on chance. Here, we will simply analyze the chance element, computing the probabilities of obtaining specific hands when the cards are dealt at the beginning of the game.

First of all, note that there are

$$\binom{52}{5} = \frac{52!}{5! \times 47!} = 2\,598\,960 \text{ different 5-card hands}\,.$$

Let us now consider the hand that is the most difficult to obtain, the royal flush, consisting of an ace, king, queen, jack and ten, all of the same suit. Since there are four suits, the probability of randomly drawing a royal flush (or any previously specified 5-card sequence for that matter), is given by

$$P(\text{royal flush}) = \frac{4}{2\,598\,960} = 0.000\,001\,5\ ,$$

almost six times greater than the probability of winning a second prize in the toto.

Let us now look at the probability of obtaining the hand of lowest value, containing a simple pair of cards of the same value. We denote the hand by AABCD, where AA corresponds to the equal-value pair of cards. Think, for instance, of a pair of aces that can be chosen out of the 4 aces in $\binom{4}{2} = 6$ different ways. As there are 13 distinct values, the pair of cards can be chosen in $6 \times 13 = 78$ different ways. The remaining three cards of the hand will be drawn out of the remaining 12 values, with each card being of any of the 4 suits. There are $\binom{12}{3}$ combinations of the values of the three cards. Since each of them can come from any of the 4 suits, we have

$$\binom{12}{3} \times 4 \times 4 \times 4 = 14\,080$$

different possibilities. Thus,

$$P(\text{pair}) = 13 \times \frac{14\,080}{2\,598\,960} = 0.423\ .$$

That is, a pair is almost as probable as getting heads (or tails) when tossing a coin. Using the same counting techniques, one can work out the list of probabilities for all interesting poker hands as follows:

Pair	AABCD	0.423
Two pairs	AABBC	0.048
Three-of-a-kind	AAABC	0.021
Straight	5 successive values	0.0039
Flush	5 cards of the same suit	0.0020
Full house	AAABB	0.0014
Four-of-a-kind	AAAAB	0.000 24
Straight flush	Straight + flush	0.000 015
Royal flush	Ace, king, queen, jack, ten of the same suit	0.000 001 5

1.7.4 Slot Machines

The first slot machines (invented in the USA in 1897) had 3 wheels, each with 10 symbols. The wheels were set turning until they stopped in a certain sequence of symbols. The number of possible sequences was therefore $10 \times 10 \times 10 = 1000$. Thus, the probability of obtaining a specific winning sequence (the jackpot) was 0.001. Later on, the number of symbols was raised to 20, and afterwards to 22 (in 1970), corresponding, in this last case, to a 0.000 093 914 probability of winning the jackpot. The number of wheels was also increased and slot machines with repeated symbols made their appearance. Everything was done to make it more difficult (or even impossible!) to obtain certain prizes.

Let us look at the slot machine shown in Fig. 1.8, which as a matter of fact was actually built and came on the market with the following list of prizes (in dollars):

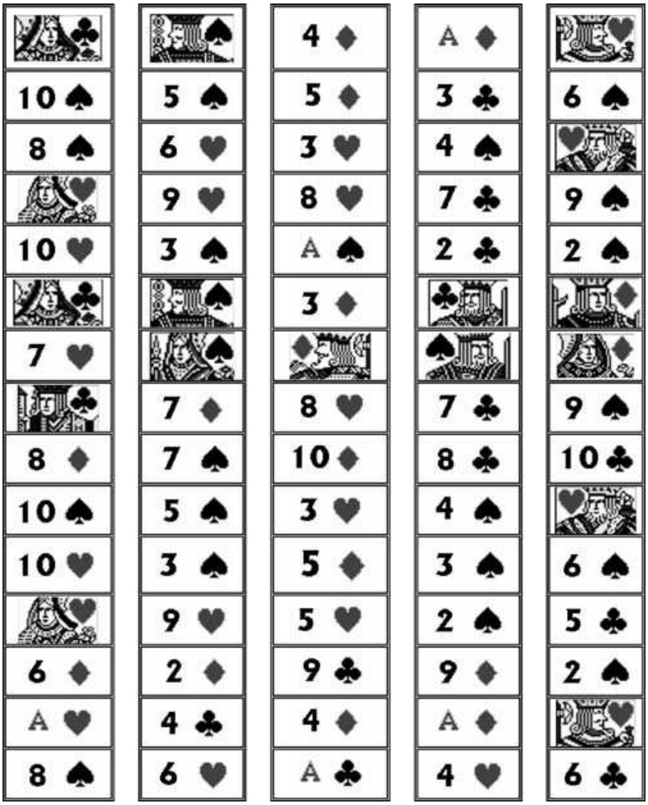


Fig. 1.8. Wheels of a particularly infamous slot machine

Pair (of jacks or better)	$ 0.05
Two pairs	$ 0.10
Three-of-a-kind	$ 0.15
Straight	$ 0.25
Flush	$ 0.30
Full house	$ 0.50
Four-of-a-kind	$ 1.00
Straight flush	$ 2.50
Royal flush	$ 5.00

The configuration of the five wheels is such that the royal flush can never show up, although the player may have the illusion when looking at the spinning wheels that it is actually possible. Even the straight flush only occurs for one single combination: 7, 8, 9, 10 and jack of diamonds. The number of possible combinations is $15^5 = 759\,375$. The number of favorable combinations for each winning sequence and the respective probabilities are computed as:

Pair (of jacks or better)	68 612	0.090 353
Two pairs	36 978	0.048 695
Three-of-a-kind	16 804	0.022 129
Straight	2 396	0.003 155
Flush	1 715	0.002 258
Full house	506	0.000 666
Four-of-a-kind	60	0.000 079
Straight flush	1	0.000 001
Royal flush	0	0

We observe that the probabilities of these 'poker' sequences for the pair and high-valued sequences are far lower than their true poker counterparts.

The present slot machines use electronic circuitry with the final position of each wheel given by a randomly generated number. The probabilities of 'bad luck' are similar.

1.7.5 Roulette

The roulette consists of a sort of dish having on its rim 37 or 38 hemispherical pockets. A ball is thrown into the spinning dish and moves around until it finally settles in one of the pockets and the dish (the roulette) stops. The construction and the way of operating the roulette are presumed to be such that each throw can be considered as a random phenomenon with equiprobable outcomes for all pockets. These are numbered from 0. In European roulette, the numbering goes from 0 to 36. American roulette has an extra pocket marked 00.

In the simplest bet the player wins if the bet is on the number that turns up by spinning the roulette, with the exception of 0 or 00 which are numbers reserved for the house. Thus, in European roulette the probability of hitting a certain number is $1/37 = 0.027$ (better than getting three-of-a-kind at poker), which does not look so bad, especially when one realises that the player can place bets on several numbers (as well as on even or odd numbers, numbers marked red or black, and so on). However, we shall see later that even in 'games of bad luck' whose elementary probabilities are not that 'bad', prizes and fees are arranged in such a way that in a long sequence of bets the player will lose, while maintaining the illusion that good luck really does lie just around the corner!

2

Amazing Conditions

2.1 Conditional Events

The results obtained in the last chapter, using the basic rules for operating with probabilities, were in perfect agreement with common sense. From the present chapter onwards, many surprising results will emerge, starting with the inclusion of such a humble ingredient as the assignment of conditions to events. Consider once again throwing a die and the events ‘face less than or equal to 4’ and ‘even face’. In the last chapter we computed $P(\text{face less than or equal to 4}) = 2/3$ and $P(\text{even face}) = 1/2$. One might ask: what is the probability that an even face turns up *if* we know that a face less than or equal to 4 has turned up? Previously, when referring to the even-face event, there was no prior condition, or information conditioning that event. Now, there is one condition: we know that ‘face less than or equal to 4’ has turned up. Thus, out of the four possible elementary events corresponding to ‘face less than or equal to 4’, we enumerate those that are even. There are two: face 2 and face 4. Therefore,

$$P(\text{even face if face less than or equal to 4}) = \frac{2}{4} = \frac{1}{2}\,.$$

Looking at Fig. 1.2c of the last chapter, we see that the desired probability is computed by dividing the cases corresponding to the event intersection (hatched gray squares) by the cases corresponding to the condition (gray squares). Briefly, from the point of view of the classic definition, it is as though we have moved to a new set of elementary events corresponding to compliance with the conditioning event, viz., $\{1, 2, 3, 4\}$.

At the same time notice that, for this example the condition did not influence the probability of 'even face':

$$P(\text{even face if face less than or equal to } 4) = P(\text{even face}) \ .$$

This equality does not always hold. For instance, it is an easy task to check that

$$P(\text{even face if face less than or equal to } 5) = \frac{2}{5} \neq P(\text{even face}) \ .$$

From the frequentist point of view, and denoting the events of the random experiment simply by A and B, we may write

$$P(A \text{ if } B) \approx \frac{k_{A \,\text{and}\, B}}{k_B} \ ,$$

where $k_{A \,\text{and}\, B}$ and k_B are, respectively, the number of occurrences of 'A and B' and 'B' when the random experiment is repeated n times with n sufficiently large. The symbol $\approx$ means 'approximately equal to'. It is a well known fact that one may divide both terms of a fraction by the same number without changing its value. Therefore, we rewrite

$$P(A \text{ if } B) \approx \frac{k_{A \,\text{and}\, B}}{k_B} = \frac{k_{A \,\text{and}\, B}/n}{k_B/n} \ ,$$

justifying, from the frequentist point of view, the so-called *conditional probability* formula:

$$P(A \text{ if } B) = \frac{P(A \text{ and } B)}{P(B)} \ .$$

For the above example where the random experiment consists of throwing a die, the formula corresponds to dividing 2/6 by 4/6, whence the above result of 1/2. Suppose we throw a die a large number of times, say 10 000 times. In this case, about two thirds of the times we expect 'face less than or equal to 4' to turn up. Imagine that it turned up 6 505 times. From this number of times, let us say that 'even face' turned up 3 260 times. Then the frequency of 'even face if face less than or equal to 4' would be computed as 3 260/6 505, close to 1/2. In other trials the numbers of occurrences may be different, but the final result, for sufficiently large n, will always be close to 1/2.

We now look at the problem of determining the probability of 'face less than or equal to 4 if even face'. Applying the above formula we get

$$P\begin{pmatrix}\text{face less than or equal}\\ \text{to 4 if even face}\end{pmatrix} = \frac{P\begin{pmatrix}\text{face less than or equal}\\ \text{to 4 and even face}\end{pmatrix}}{P(\text{even face})}$$

$$= \frac{2/6}{3/6} = \frac{2}{3}\,.$$

Thus, changing the order of the conditioning event will in general produce different probabilities. Note, however, that the intersection is (always) *commutative*: $P(A \text{ and } B) = P(B \text{ and } A)$. In our example:

$$P(A \text{ and } B) = P(A \text{ if } B)P(B) = \frac{1}{2} \times \frac{2}{3} = \frac{1}{3}\,,$$

$$P(B \text{ and } A) = P(B \text{ if } A)P(A) = \frac{2}{3} \times \frac{1}{2} = \frac{1}{3}\,.$$

From now on we shall simplify notation by omitting the multiplication symbol when no possibility of confusion arises.

2.2 Experimental Conditioning

The literature on probability theory is full of examples in which balls are randomly extracted from urns. Consider an urn containing 3 white balls, 2 red balls and 1 black ball. The probabilities of randomly extracting one specific-color ball are

$$P(\text{white}) = 1/2\,, \qquad P(\text{red}) = 1/3\,, \qquad P(\text{black}) = 1/6\,.$$

If, whenever we extract a ball, we put it back in the urn – extraction with replacement – these probabilities will remain unchanged in the following extractions. If, on the other hand, we do not put the ball back in the urn – extraction without replacement – the experimental conditions for ball extraction are changed and we obtain new probability values. For instance:

- If a white ball comes out in the first extraction, then in the following extraction

  $$P(\text{white}) = 2/5\,, \qquad P(\text{red}) = 2/5\,, \qquad P(\text{black}) = 1/5\,.$$

- If a red ball comes out in the first extraction, then in the following extraction:

  $$P(\text{white}) = 3/5\,, \qquad P(\text{red}) = 1/5\,, \qquad P(\text{black}) = 1/5\,.$$

- If a black ball comes out in the first extraction, then in the following extraction:

$$P(\text{white}) = 3/5\ , \qquad P(\text{red}) = 2/5\ , \qquad P(\text{black}) = 0\ .$$

As we can see, the conditions under which a random experiment takes place may drastically change the probabilities of the various events involved, as a consequence of event conditioning.

2.3 Independent Events

Let us now consider throwing two dice, one white and the other green, and suppose someone asks the following question: what is the probability that an even face turns up on the white die if an even face turns up on the green die? Obviously, if the dice-throwing is fair, the face that turns up on the green die has no influence whatsoever on the face that turns up on the white die, and vice versa. The events 'even face on the white die' and 'even face on the green die' do not interfere. The same can be said about any other pair of events, one of which refers to the white die while the other refers to the green die. We say that the events are *independent*. Applying the preceding rule, we may work this out in detail as

$$\begin{aligned} P\begin{pmatrix}\text{even face of white die if}\\ \text{even face of green die}\end{pmatrix} &= \frac{P\begin{pmatrix}\text{even face of white die and}\\ \text{even face of green die}\end{pmatrix}}{P(\text{even face of green die})} \\ &= \frac{(1/2)\times(1/2)}{1/2} = \frac{1}{2}\ , \end{aligned}$$

which is precisely the probability of 'even face of the white die'. In conclusion, the condition 'even face of the green die' has no influence on the probability value for the other die. Consider now the question: what is the probability of an even face of the white die turning up if at least one even face turns up (on either of the dice or on both)? We first notice that 'at least one even face turns up' is the complement of 'no even face turns up', or in other words, both faces are odd, which may occur $3 \times 3 = 9$ times. Therefore,

$$\begin{aligned} P\begin{pmatrix}\text{even face of the white die}\\ \text{if at least one even face}\end{pmatrix} &= \frac{P\begin{pmatrix}\text{even face of the white die}\\ \text{and at least one even face}\end{pmatrix}}{P(\text{at least one even face})} \\ &= \frac{3\times 6/36}{(36-9)/36} = \frac{2}{3}\ . \end{aligned}$$

We now come to the conclusion that the events ‘even face of the white die’ and ‘at least one even face’ *are not* independent. Consequently, they are said to be *dependent*. In the previous example of ball extraction from an urn, the successive extractions are independent if performed with replacement and dependent if performed without replacement.

For independent events, $P(A \text{ if } B) = P(A)$ and $P(B \text{ if } A) = P(B)$. Let us go back to the conditional probability formula:

$$P(A \text{ if } B) = \frac{P(A \text{ and } B)}{P(B)} , \quad \text{or } P(A \text{ and } B) = P(A \text{ if } B)P(B) .$$

If the events are independent one may replace $P(A \text{ if } B)$ by $P(A)$ and obtain $P(A \text{ and } B) = P(A)P(B)$. Thus, the probability formula presented in Chap. 1 for event intersection is only applicable when the events are independent. In general, if $A, B, \dots , Z$ are independent,

$$P(A \text{ and } B \text{ and } \dots \text{ and } Z) = P(A) \times P(B) \times \dots \times P(Z) .$$

2.4 A Very Special Reverend

Sometimes an event can be obtained in distinct ways. For instance, in the die-throwing experiment let us stipulate that event A means ‘face less than or equal to 5’. This event may occur when either the event B = ‘even face’ or its complement $\overline{B}$ = ‘odd face’ have occurred. We readily compute

$$P(A \text{ and } B) = 2/6 , \qquad P(A \text{ and } \overline{B}) = 3/6 .$$

But the union of disjoint events ‘A and B’ and ‘A and $\overline{B}$’ is obviously just A. Applying the addition rule for disjoint events we obtain, without much surprise,

$$\begin{aligned} P(\text{face less than or equal to 5}) &= P(A \text{ and } B) + P(A \text{ and } \overline{B}) \\ &= \frac{2}{6} + \frac{3}{6} = \frac{5}{6} . \end{aligned}$$

In certain circumstances, we have to evaluate probabilities of intersections, $P(A \text{ and } B)$ and $P(A \text{ and } \overline{B})$, on the basis of conditional probabilities. Consider two urns, named X and Y, containing white and black balls. Imagine that X has 4 white and 5 black balls and Y has 3 white and 6 black balls. One of the urns is randomly chosen and afterwards a ball is randomly drawn out of that urn. What is the probability that a white ball is drawn? Using the preceding line of thought, we compute:

$$\begin{aligned}P(\text{white ball}) &= P(\text{white ball and urn } X) + P(\text{white ball and urn } Y)\\ &= P(\text{white ball if urn } X) \times P(\text{urn } X)\\ &\qquad\qquad +P(\text{white ball if urn } Y) \times P(\text{urn } Y)\\ &= \frac{4}{9} \times \frac{1}{2} + \frac{3}{9} \times \frac{1}{2} = \frac{7}{18} = 0.39\ .\end{aligned}$$

In the book entitled *The Doctrine of Chances* by Abraham de Moivre, mentioned in the last chapter, reference was made to the urn *inverse problem*, that is: What is the probability that the ball came from urn X if it turned out to be white? The solution to the inverse problem was first discussed in a short document written by the Reverend Thomas Bayes (1702–1761), an English priest interested in philosophical and mathematical problems. In Bayes' work, entitled *Essay Towards Solving a Problem in the Doctrine of Chances*, and published posthumously in 1763, there appears the famous Bayes' theorem that would so significantly influence the future development of probability theory. Let us see what Bayes' theorem tells us in the context of the urn problem. We have seen that

$$\begin{aligned}P(\text{white ball}) = P(\text{white ball if urn } X) \times P(\text{urn } X)\\ +P(\text{white ball if urn } Y) \times P(\text{urn } Y)\ .\end{aligned}$$

Now Bayes' theorem allows us to write

$$P(\text{urn } X \text{ if white ball}) = \frac{P(\text{urn } X) \times P(\text{white ball if urn } X)}{P(\text{white ball})}\ .$$

Note how the inverse conditioning probability $P(\text{urn } X \text{ if white ball})$ depends on the direct conditioning probability $P(\text{white ball if urn } X)$. In fact, it corresponds simply to dividing the 'urn X' term of its decomposition, expressed in the formula that we knew already, by $P(\text{white ball})$. Applying the concrete values of the example, we obtain

$$P(\text{urn } X \text{ if white ball}) = \frac{(4/9) \times (1/2)}{7/18} = \frac{4}{7}\ .$$

In the same way,

$$P(\text{urn } Y \text{ if white ball}) = \frac{(3/9) \times (1/2)}{7/18} = \frac{3}{7}\ .$$

Of course, the two probabilities add up to 1, as they should (the ball comes either from urn X or from urn Y). But care should be taken here: the two conditional probabilities for the white ball do *not* have to add up to 1. In fact, in this example, we have

$$P(\text{white ball if urn } X) + P(\text{white ball if urn } Y) = \frac{7}{9}\,.$$

In the event 'urn X if white ball', we may consider that the drawing out of the white ball is the effect corresponding to the 'urn X' cause. Bayes' theorem is often referred to as a theorem about the probabilities of causes (once the effects are known), and written as follows:

$$P(\text{cause if effect}) = \frac{P(\text{cause}) \times P(\text{effect if cause})}{P(\text{effect})}\,.$$

Let us examine a practical example. Consider the detection of breast cancer by means of a mammography. The mammogram (like many other clinical analyses) is not infallible. In reality, long studies have shown that (frequentist estimates)

$$P\begin{pmatrix}\text{positive mammogram}\\ \text{if breast cancer}\end{pmatrix} = 0.9\,, \quad P\begin{pmatrix}\text{positive mammogram}\\ \text{if no breast cancer}\end{pmatrix} = 0.1\,.$$

Let us assume the random selection of a woman in a given population of a country where it is known that (again frequentist estimates)

$$P(\text{breast cancer}) = 0.01\,.$$

Using Bayes' theorem we then compute:

$$P\begin{pmatrix}\text{breast cancer if}\\ \text{positive mammogram}\end{pmatrix} = \frac{0.01 \times 0.9}{0.01 \times 0.9 + 0.99 \times 0.1} = 0.08\,.$$

It is a probability that we can consider low in terms of clinical diagnostic, raising reasonable doubt about the *isolated* use of the mammogram (as for many other clinical analyses for that matter). Let us now assume that the mammogram is prescribed following a medical consultation. The situation is then quite different: the candidate is no longer a randomly selected woman from the population of a country, but a woman from the female population that seeks a specific medical consultation (undoubtedly because she has found that something is not as it should be). Now the probability of the cause (the so-called *prevalence*) is different. Suppose we have determined (frequentist estimate)

$$P\begin{pmatrix}\text{breast cancer (in women}\\ \text{seeking specific consultation)}\end{pmatrix} = 0.1\ .$$

Using Bayes' theorem, we now obtain

$$P\begin{pmatrix}\text{breast cancer if}\\ \text{positive mammogram}\end{pmatrix} = \frac{0.1\times 0.9}{0.1\times 0.9 + 0.9\times 0.1} = 0.5\ .$$

Observe how the prevalence of the cause has dramatic consequences when we infer the probability of the cause assuming that a certain effect is verified.

Let us look at another example, namely the use of DNA tests in assessing the innocence or guilt of someone accused of a crime. Imagine that the DNA of the suspect matches the DNA found at the crime scene, and furthermore that the probability of this happening by pure chance is extremely low, say, one in a million, whence

$$P(\text{DNA match if innocent}) = 0.000\,001\ .$$

But,

$$P(\text{DNA match if guilty}) = 1\ ,$$

another situation where the sum of the conditional probabilities is not 1. Suppose that in principle there is no evidence favoring either the innocence or the guilt (no one is innocent or guilty until there is proof), that is, $P(\text{innocent}) = P(\text{guilty}) = 0.5$. Then,

$$P(\text{innocent if DNA match}) = \frac{0.000\,001\times 0.5}{0.000\,001\times 0.5 + 1\times 0.5} \approx 0.000\,001\ .$$

In this case the DNA evidence is decisive. In another scenario let us suppose that, given the circumstances of the crime and the suspect, there are good reasons for presuming innocence. Let us say that, for the suspect to be guilty, one would require a whole set of circumstances that we estimate to occur only once in a hundred thousand times, that is, in principle, $P(\text{innocent}) = 0.999\,99$. Then,

$$P(\text{innocent if DNA match}) = \frac{0.000\,001\times 0.999\,99}{0.000\,001\times 0.999\,99 + 1\times 0.000\,01} \approx 0.09\ .$$

In this case, in spite of the DNA evidence, the presumption of innocence is still considerable (about one case in eleven).

Fig. 2.1. The risks of entering the PD if C clinic

2.5 Probabilities in Daily Life

The normal citizen often has an inaccurate perception of probabilities in everyday life. In fact, several studies have shown that the probability notion is one of the hardest to grasp. For instance, if after flipping a coin that turned up heads, we ask a child what will happen in the following flip, the most probable answer is that the child will opt for tails. The adult will have no difficulty in giving the correct answer to this simple question (that is, either heads or tails will turn up, meaning that both events are equally likely), but if the question gets more complicated an adult will easily fall into the so-called *gambler's fallacy*. An example of this is imagining that one should place a high bet on a certain roulette number that has not come up for a long time.

Let us illustrate the fallacy with the situation where a coin has been flipped several times and tails has been turning up each time. Most gamblers (and not only gamblers) then tend to consider that the probability of heads turning up in the following flips will be higher than before. The same applies to other games. Who has never said something like: 30 is bound to come now; it's time for black to show up; the 6 is hot; the odds are now on my side, and so on? It looks as if one would deem the information from previous outcomes as useful in determining the following outcome; in other words, it is as if there were some kind of memory in a sequence of trials. In fact, rigorously conducted experiments have even shown that the normal citizen has the tendency to believe that there is a sort of ‘memory’ phenomenon in chance devices. In this spirit, if a coin has been showing tails, why not let it ‘rest’ for a while before playing again? Or change the coin? Suppose a coin has produced tails 30 times in a row. We will find it strange and cast some doubts about the coin or the fairness of the player. On the other hand, if instead of being consecutive, the throws are spaced over one month (say, one throw per day), maybe we would not find the event so strange after all, and maybe we would not attribute the ‘cause’ of the event to the coin’s ‘memory’, but instead conjecture that we are having an ‘unlucky month’.

Consider a raffle with 1 000 tickets and 1 000 players, but only one prize. Once all the tickets have been sold – one ticket for each of the 1 000 players – it comes out that John Doe has won the prize. In a new repetition of the raffle John Doe wins again. It looks as if ‘luck’ is with John Doe. Since the two raffles are independent, the probability of John Doe winning twice consecutively is $(1/1000)^2 = 0.000\,001$. This is exactly the same probability that corresponds to John Doe winning the first time and Richard Roe the second time. Note that this is the same probability as if John Doe had instead won a raffle in which a million tickets had been sold, which would certainly seem less strange than the double victory in the two 1000-ticket raffles.

Let us slightly modify our problem, considering an only-one-prize raffle with 10 000 tickets and the participation of 1 000 players, each one buying the same number of tickets. The probability that a given player (say, John Doe) wins is 0.001 (the thousand players are equally likely to win). Meanwhile, the probability that *someone* wins the raffle is obviously 1 (someone has to win). Now suppose the raffle is repeated three times. The probability that the same player (John Doe) wins all 3 times is (independent events) $(0.001)^3 = 0.000\,000\,001$. However, the probability that *someone* (whoever he or she may be) wins the raffle three times is *not* 1, but at most $1000 \times (0.001)^3 = 0.000\,001$. In fact, any combination of three winning players has a probability $(0.001)^3$; if the same 1 000 players play the three raffles, there are 1 000 distinct possibilities of someone winning the three raffles, resulting in the above probability. If the players are not always the same, the probability will be smaller. Thus, what will amaze the common citizen is not so much that John Doe has a probability of one in a thousand million of winning, but the fact that the probability of *someone* winning in three repetitions is one in a million.

Let us now take a look at the following questions that appeared in tests assessing people's ability to understand probabilities:

1. A hat contains 10 red and 10 blue smarties. I pull out 10 and 8 are red. Which am I more likely to get next?
2. A box contains green and yellow buttons in unknown proportions. I take out 10 and 8 are yellow. Which am I more likely to get next?
3. In the last four matches between Mytholmroyd Athletics and Giggleswick United, Mytholmroyd have kicked off first every time, on the toss of a coin. Which team is more likely to kick off next time?

People frequently find it difficult to give the correct answer[1] to such questions. When dealing with conditional events and the application of Bayes' theorem, the difficulty in arriving at the right answer is even greater. Let us start with an apparently easy problem. We randomly draw a card out of a hat containing three cards: one with both faces red, another with both faces blue, and the remaining one with one face red and the other blue. We pull a card out at random and observe that the face showing is red. The question is: what is the probability that the other face will also be red? In general, the answer one obtains is based on the following line of thought: if the face shown is red, it can only be one of two cards, red–red or red–blue; therefore, the probability that it is red–red is 1/2. This result is wrong and the mistake in the line of thought has to do with the fact that the set of elementary events has been incorrectly enumerated. Let us denote the faces by R (red) and B (blue) and let us put a mark, say an asterisk , to indicate the face showing. The set of elementary events is then:

$$\{R^*R, RR^*, R^*B, RB^*, B^*B, BB^*\} .$$

In this enumeration, R^*B means that for the red–blue card the R face is showing, and likewise for the other cards. As for the RR card, any of the faces can be turned up: R^*R or RR^*. In this way,

$$P(RR \text{ if } R^*) = \frac{P(RR \text{ and } R^*)}{P(R^*)} = \frac{2}{3} .$$

In truth, for the three events containing R^*, namely R^*R, RR^* and R^*B, only two correspond to the RR card.

The following problem was presented by two American psychologists. A taxi was involved in a hit-and-run accident during the night.

[1] Correct answers are: blue, yellow, either.

Two taxi firms operate in the city: green and blue. The following facts are known:

- 85% of the taxis are green and 15% are blue.
- One witness identified the taxi as blue. The court tested the witness' reliability in the same conditions as on the night of the accident and concluded that the witness correctly identified the colors 80% of the time and failed in the other 20%.

What is the probability that the hit-and-run taxi was blue?

The psychologists found that the typical answer was 80%. Let us see what the correct answer is. We begin by noting that

$$P\begin{pmatrix}\text{witness is right}\\ \text{if taxi is blue}\end{pmatrix} = 0.8\ , \quad P\begin{pmatrix}\text{witness is wrong}\\ \text{if taxi is blue}\end{pmatrix} = 0.2\ .$$

Hence, applying Bayes' theorem:

$$\begin{aligned} P\begin{pmatrix}\text{taxi is blue}\\ \text{if witness is right}\end{pmatrix} &= \frac{0.15 \times 0.8}{0.15 \times 08 + 0.85 \times 0.2} \\ &= 0.41\ . \end{aligned}$$

We see that the desired probability is below 50% and nearly half of the typical answer. In this and similar problems the normal citizen tends to neglect the role of prevalences. Now the fact that blue taxis are less prevalent than green taxis leads to a drastic decrease in the probability of the witness being right.

2.6 Challenging the Intuition

There are problems on probabilities that lead to apparently counter-intuitive results and are even a source of embarrassment for experts and renowned mathematicians.

2.6.1 The Problem of the Sons

Let us go back to the problem presented in the last chapter of finding the probability that a family with two children has at least one boy. The probability is 3/4, assuming the equiprobability of boy and girl. Let us now suppose that we have got the information that the randomly selected family had a boy. Our question is: what is the probability that the other child is also a boy? Using the same notation as in Chap. 1, we have

$$P\big(\{MM\} \text{ if } \{MF, FM, MM\}\big) = \frac{1/4}{3/4} = \frac{1}{3}\,.$$

This result may look strange, since apparently the fact that we know the sex of one child should not condition the probability relative to the other one. However, let us not forget that, as in the previous three-card problem, there are two distinct situations for boy–girl, namely MF and FM.

We now modify the problem involving the visit to a random family with two children, by assuming that we knock at the door and a boy shows up. What is the probability that the other child is a boy? It seems that nothing has changed relative to the above problem, but as a matter of fact this is not true. The event space is now

$$\{M^*M, MM^*, M^*F, F^*M, FM^*, F^*F, FF^*\}\,,$$

where the asterisk denotes the child that opens the door. We are interested in the conditional probability of $\{M^*M, MM^*\}$ if $\{M^*M, MM^*, M^*F, FM^*\}$ takes place. The probability is 1/2. We conclude that the simple fact that a boy came to the door has changed the probabilities!

2.6.2 The Problem of the Aces

Consider a 5-card hand, drawn randomly from a 52-card deck, and suppose that an ace is present in the hand. Let us further specify two categories of hands: in one category, the ace is of clubs; in the other, the ace is any of the 4 aces. Suppose that in either category of hands we are interested in knowing the probability that at least one more ace is present in the hand. We are then interested in the following conditional probabilities:

$$P\begin{pmatrix}\text{two or more aces if}\\ \text{an ace of clubs was drawn}\end{pmatrix}\,, \qquad P\begin{pmatrix}\text{two or more aces}\\ \text{if an ace was drawn}\end{pmatrix}\,.$$

Are these probabilities equal? The intuitive answer is affirmative. After all an ace of clubs is an ace and it would seem to be all the same whether it be of clubs or of something else. With a little bit more thought, we try to assess the fact that in the second case we are dealing with any of 4 aces, suggesting a larger number of possible combinations and, as a consequence, a larger probability. But is that really so? Let us actually compute the probabilities. It turns out that for this problem the conditional probability rule is not of much help. It is preferable to resort to the classic notion and evaluate the ratio of favorable cases over possible cases. In reality, since the evaluation of favorable cases would amount to counting hand configurations with two, three and four aces, it is worth trying rather to evaluate *unfavorable* cases, i.e., hands with only one ace.

Let us examine the first category of hand with an ace of clubs. There remain 51 cards and 4 positions to fill in. There are therefore $\binom{51}{4}$ possible cases, of which $\binom{48}{4}$ are unfavorable, i.e., there is no other ace in the remaining 4 positions, whence

$$P\left(\begin{array}{c}\text{two or more aces if}\\ \text{an ace of clubs was drawn}\end{array}\right) = \frac{\binom{51}{4} - \binom{48}{4}}{\binom{51}{4}} = 0.22 \ .$$

In the second category – a hand with at least one ace – the number of possible cases is obtained by subtracting from all possible hands those with no ace, which gives

$$\text{possible cases} \quad \binom{52}{5} - \binom{48}{5} \ .$$

Concerning the unfavorable cases, once we fix an ace, we get the same number of unfavorable cases as before. Since there are four distinct possibilities for fixing an ace, we have

$$\text{unfavorable cases} \quad 4 \times \binom{48}{4} \ .$$

Hence,

$$P\left(\begin{array}{c}\text{two or more aces}\\ \text{if an ace was drawn}\end{array}\right) = \frac{\binom{52}{5} - \binom{48}{5} - 4 \times \binom{48}{4}}{\binom{52}{5} - \binom{48}{5}} = 0.12 \ .$$

We thus obtain a smaller (!) probability than in the preceding situation. This apparently strange result becomes more plausible if we inspect what happens in a very simple situation of two-card hands from a deck

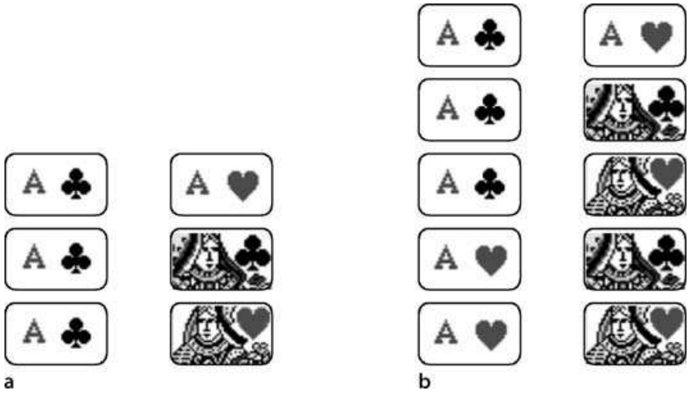


Fig. 2.2. The problem of the aces for a very simple deck

with only 4 cards: ace and queen of hearts and clubs (see Fig. 2.2). We see that, when we fix the ace of clubs, the probability of a second ace is 1/3 (Fig. 2.2a). If on the other hand we are allowed any of the two aces or both, the probability is 1/5 (Fig. 2.2b).

2.6.3 The Problem of the Birthdays

Imagine a party with a certain number of people, say n. The question is: how large must n be so that the probability of finding two people with the same birthday date (month and day) is higher than 50%?

At first sight it may seem that n will have to be quite large, perhaps as large as there are days in a year. We shall find that this is not so. The probability of a person's birthday falling on a certain date is 1/365, ignoring leap years. Since the birthdays of the n people are independent of each other, any list of n people's birthday dates has a probability given by the product of the n individual probabilities (refer to what was said before about the probability of independent events). Thus,

$$P\big((\text{list of})\ n\ \text{birthdays}\big) = \left(\frac{1}{365}\right)^n .$$

What is the probability that two birthdays do not coincide? Imagine the people labeled from 1 to n. We may arbitrarily assign a value to the first person's birthday, with 365 different ways of doing it. Once this birthday has been assigned there remain 364 possibilities for selecting a date for the second person with no coincidence; afterwards, 363 possibilities are available for the third person's birthday, and so on, up to

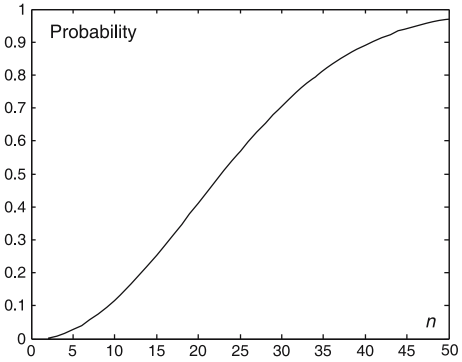


Fig. 2.3. Probability of two out of n people having the same birthday

the last remaining person with $365 - n + 1$ possibilities. Therefore,

$$P\begin{pmatrix} n \text{ birthdays without} \\ \text{two coincidences} \end{pmatrix} = \frac{365 \times 364 \times \ldots \times (365 - n + 1)}{365^n} .$$

It is now clear that $P(n$ birthdays with two coincidences) is the complement of the above probability with respect to 1. It can be checked that for $n = 23$ the probability of finding two coincident birthdays is already higher than 0.5. Indeed, $P(23$ birthdays with two coincidences) $= 0.5073$. Figure 2.3 shows how this probability varies with n. Note that, whereas for 10 people the probability is small (0.1169), for 50 people it is quite considerable (0.9704).

2.6.4 The Problem of the Right Door

The problem of the right door comes from the USA (where it is known as the Monty Hall paradox) and is described as follows. Consider a TV quiz where 3 doors are presented to a contestant and behind one of them is a valuable prize (say, a car). Behind the other two doors, however, the prizes are of little value. The contestant presents his/her choice of door to the quiz host. The host, knowing what is behind each door, opens one of the other two doors (one not hiding the big prize) and reveals the worthless prize. The contestant is then asked if he/she wants to change the choice of door. The question is: should the contestant change his/her initial choice? When a celebrated columnist of an American newspaper recommended changing the initial choice, the newspaper received several letters from mathematicians indignant over such a recommendation, stating that the probability of winning with such a strategy was 0.5. Let us see if that is true. We start by numbering the doors from 1 to 3 and assume that door 1 is the contestant's

Table 2.1. Finding the right door

Prize door	Initial choice	Host opens	New choice
Door 1	Door 1	Door 2	Door 3
Door 2	Door 1	Door 3	**Door 2**
Door 3	Door 1	Door 2	**Door 3**

initial choice. If the contestant does not change his/her initial choice, there is a 1/3 probability of choosing the right door. If, on the other hand, he/she decides to change his/her choice, we have the situations shown in Table 2.1.

The contestant wins in the two new choices written in bold in the table; therefore, the probability of winning when changing the choice doubles to 2/3! This result is so counter-intuitive that even the Hungarian mathematician Paul Erdös (1913–1996), famous for his contributions to number theory, had difficulty accepting the solution. The problem for most people in accepting the right solution is connected with the strongly intuitive and deeply rooted belief that probabilities are something immutable. Note, however, how the information obtained by the host opening one of the doors changes the probability setting and allows the contestant to increase his/her probability of winning.

2.6.5 The Problem of Encounters

We saw in the last chapter that there is a $1/n!$ probability of randomly inserting n letters in the right envelopes. We also noticed how this probability decreases very quickly with n. Let us now take a look at the event 'no match with any envelope'. At first sight, one could imagine that the probability of this event would also vary radically with n. Let us compute the probability. We begin by noticing that the one-match event has probability $1/n$, whence the no-match event for one letter has the probability $1 - 1/n$. If the events 'no match for letter 1', 'no match for letter 2', ..., 'no match for letter n' were independent, the probability we wish to compute would be $(1 - 1/n)^n$, which tends to 1/e, as n grows. The number denoted e, whose value up to the sixth digit is 2.718 28, is called the Napier constant, in honor of the Scottish mathematician John Napier, who introduced logarithmic calculus (1618).

However, as we know, these events are not independent since their complements are not. For instance, if $n - 1$ letters are in the right envelope, the nth letter has only one position (the right one) to be inserted in.

It so happens that the problem we are discussing appears under several guises in practical issues, generally referred to as the problem of encounters, which for letters and envelopes is formulated as: supposing that the letters and the envelopes are numbered according to the natural order $1, 2, \ldots, n$, what is the probability that, when inserting letter i, it will sit in its natural position? (We then say that an encounter has taken place.)

Consider all the possibilities of encounters of letters 1 and 3 in a set of 5 letters:

$$\underline{1}2\underline{3}45\ , \quad \underline{1}2\underline{3}54\ , \quad \underline{1}4\underline{3}25\ , \quad \underline{1}4\underline{3}52\ , \quad \underline{1}5\underline{3}24\ , \quad \underline{1}5\underline{3}42\ .$$

We see that for r encounters ($r = 2$ in the example), there are only $n-r$ letters that may permute (in the example, $3! = 6$ permutations). Any of these permutations has probability $(n-r)!/n!$. Finally, we can compute the probability of no encounter as the complement of the probability of at least one encounter. This one is computed using a generalization of the event union formula for more than two events. Assuming $n = 3$ and abbreviating 'encounter' to 'enc', the probability is computed as follows:

$$\begin{aligned}
P\begin{pmatrix}\text{at least one encounter}\\ \text{in 3 letters}\end{pmatrix}
&= P(\text{enc. letter 1}) + P(\text{enc. letter 2}) + P(\text{enc. letter 3})\\
&\quad - P(\text{enc. letters 1 and 2}) - P(\text{enc. letters 1 and 3})\\
&\quad - P(\text{enc. letters 2 and 3}) + P(\text{enc. letters 1, 2 and 3})\\
&= \binom{3}{1} \times \frac{(3-1)!}{3!} - \binom{3}{2} \times \frac{(3-2)!}{3!} + \binom{3}{3} \times \frac{(3-3)!}{3!}\\
&= 3 \times \frac{1}{3} - 3 \times \frac{1}{6} + 1 \times \frac{1}{6} = \frac{2}{3}\ .
\end{aligned}$$

Figure 2.4 shows the Venn diagram for this situation.

Hence,

$$P(\text{no encounter in 3 letters}) = \frac{1}{3}\ .$$

It can be shown that the above calculations can be carried out more easily as follows for n letters:

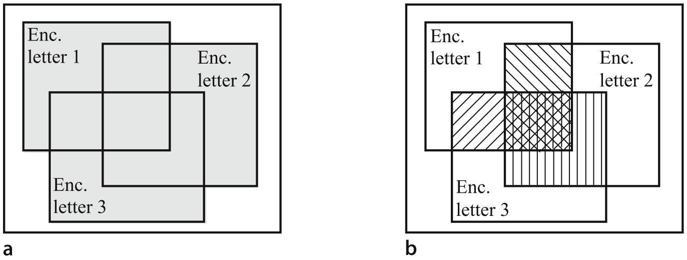


Fig. 2.4. The *gray area* in (**a**) is equal to the sum of the areas of the rectangles corresponding to the events 'enc. letter 1', 'enc. letter 2', and 'enc. letter 3', after subtracting the *hatched areas* in (**b**) and adding the *double-hatched central area*

Table 2.2. Relevant functions for the problem of encounters

n	2	3	4	5	6	7	8	9
$d(n)$	1	2	9	44	265	1854	14833	133496
$P(n)$	0.5	0.3333	0.3750	0.3667	0.3681	0.3679	0.3679	0.3679
$(1-1/n)^n$	0.25	0.2963	0.3164	0.3277	0.3349	0.3399	0.3436	0.3464

$$P(\text{no encounter in } n \text{ letters}) = \frac{d(n)}{n!}\ ,$$

where $d(n)$, known as the number of *derangements* of n elements (permutations without encounters) is given by

$$d(n) = (n-1) \times \big[d(n-1) + d(n-2)\big]\ ,$$

with $d(0) = 1$ and $d(1) = 0$. Table 2.2 lists the values of $d(n)$, the probability of no encounter in n letters denoted $P(n)$, and the value of the function $(1-1/n)^n$.

The probability of no encounter in n letters approaches $1/\mathrm{e} = 0.3679$, as in the incorrect computation where we assumed the independence of the encounters, but it does so in a much faster way than with the wrong computation. Above 6 letters, the probability of no match in an envelope is practically constant!

3

Expecting to Win

3.1 Mathematical Expectation

We have made use in the preceding chapters of the frequentist concept of probability. Let us assume that, in a sequence of 8 tosses of a coin, we have obtained:

$$\text{heads, tails, tails, heads, heads, heads, tails, heads}\,.$$

Assigning the value 1 to heads and 0 to tails, we may then add up the values of the sequence 10011101. We obtain 5, that is, an average value of $5/8 = 0.625$ for heads. This average, computed on the 0-1 sequence, corresponds to the frequency of occurrence $f = 0.625$ heads per toss. The frequentist interpretation tells us that, if we toss a coin a very large number of times, about half of the times heads will turn up. In other words, *on average* heads will turn up once every two tosses.

As an alternative to the tossing sequence in time, we may imagine a large number of people tossing coins in the air. On average heads would turn up once for every two people. Of course this *theoretical* value of 0.5 for the average (or frequency) presupposes that the coin-toss experiment either in time or for a set of people is performed a very large number of times. When the number of tosses is small one expects large deviations from an average value around 0.5. (The same phenomenon was already illustrated in Fig. 1.4 when throwing a die.)

We now ask the reader to imagine playing a card game with someone and using a pack with only the picture cards (king, queen, jack), the ace and the joker of the usual four suits. Let us denote the cards by K, Q, J, A and F, respectively. The $4 \times 5 = 20$ cards are fairly shuffled and an attendant extracts cards one by one. After extracting and showing

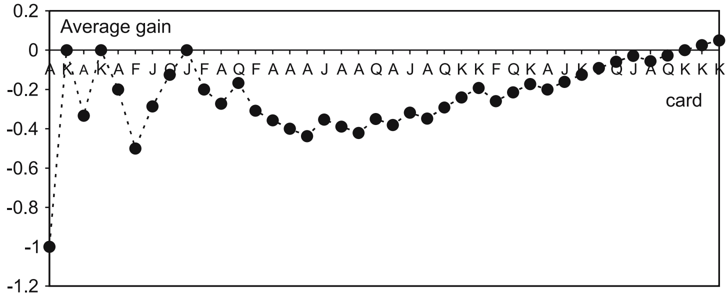


Fig. 3.1. Evolution of the average gain in a card game

the card it is put back in the deck, which is then reshuffled. For each drawn picture card the reader wins 1 euro from the opponent and pays 1 euro to the opponent if an ace is drawn and 2 euros if a joker is drawn.

Suppose the following sequence turned up:

$$\mathrm{AKAKAFJQJFAQFAAAJAAQAJAQKKFQKAJKJQJAQKKKF}\,.$$

For this sequence the reader's accumulated gain varies as: −1, 0, −1, 0, −1, −3, −2, and so on. By the end of the first five rounds the average gain per round would be −0.2 euros (therefore, a loss). Figure 3.1 shows how the average gain varies over the above 40 rounds.

One may raise the question: what would the reader's average gain be in a long sequence of rounds? Each card has an equal probability of being drawn, which is $4/20 = 1/5$. Therefore, the 1 euro gain occurs three fifths of the time. As for the 1 or 2 euro losses, they only occur one fifth of the time. Thus, one expects the average gain in a long run to be as follows:

$$\text{expected average gain} = 1 \times \frac{3}{5} + (-1) \times \frac{1}{5} + (-2) \times \frac{1}{5} = 0 \text{ euro}\,.$$

It thus turns out that the expected average gain (theoretical average gain) in a long run is zero euros. In this case the game is considered fair, that is, either the reader or the opponent has the same a priori chances of obtaining the same average gain. If, for instance, the king's value was 2 euros, the reader would have an a priori advantage over the opponent, expecting to win 0.2 euros per round after many rounds. The game would be biased towards the reader. If, on the other hand, the ace's value was −2 euros, we would have the opposite situation: the game would be biased towards the opponent, favoring him/her by 0.2 euros per round after many rounds.

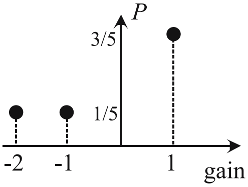


Fig. 3.2. The probability function of the gain for a card game

Let us take a closer look at how the average gain is computed. One can look at it as if one had to deal with a variable called gain, taking values in the set $\{-2, -1, 1\}$, and such that the appropriate probability is assigned to each value, as shown in Fig. 3.2. One then simply adds up the products of the values with the probabilities.

We say that the gain variable is a *random variable*. The *expected average gain*, expected after a large number of rounds, which, as we saw, is computed using the event probabilities, is called the *mathematical expectation* of the gain variable. Note the word 'mathematical'. We may have to play an incredibly large number of times before the average gain per round is sufficiently near the mathematical, theoretical, value of the expected average gain. At this moment the reader will probably suspect, and rightly so, that in the coin-tossing experiment the probability that heads turns up is the mathematical expectation of the random variable 'frequency of occurrence', denoted by f.

There is in fact a physical interpretation of the mathematical expectation that we shall now explain with the help of Fig. 3.2. Imagine that the 'gain' axis is a wire along which point masses have been placed with the same value and in the same positions as in Fig. 3.2. The equilibrium point (center of gravity) of the wire would then correspond to the mathematical expectation, in this case, the point 0.

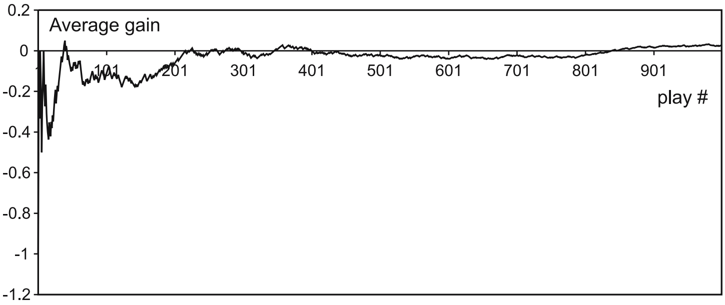


Fig. 3.3. Evolution of the average gain in 1 000 rounds

Fig. 3.4. The euphemistic average

Figure 3.3 shows how the average gain of the card game varies over a sequence of 1 000 rounds. Note the large initial wanderings followed by a relative stabilization. Another sequence of rounds would most certainly lead to a different average gain sequence. However, it would also stabilize close to

$$\begin{matrix}\text{mathematical}\\ \text{expectation}\end{matrix} = \sum \begin{pmatrix}\text{value of random}\\ \text{variable}\end{pmatrix} \times \begin{pmatrix}\text{respective}\\ \text{probability}\end{pmatrix} ,$$

where $\sum$ denotes the total sum.

3.2 The Law of Large Numbers

The time has come to clarify what is meant by a ‘sufficiently long run’ or ‘a large number of rounds’. For that purpose, we shall use the preceding card game example with null mathematical expectation. Denoting the expectation by E, we have

$$E(\text{AKQJF game}) = 0 .$$

On the other hand, the average gain after n rounds is computed by adding the individual gains of the n rounds – that is, the values of the corresponding random variable – and dividing the resulting sum by n. In Fig. 3.1, for instance, the values of the random variable for the first 10 rounds were

$$-1 ,\ 1 ,\ -1 ,\ 1 ,\ -1 ,\ -2 ,\ 1 ,\ 1 ,\ 1 ,\ -2 .$$

We thus obtain the following average gain in 10 rounds:

$$\begin{aligned}\text{average gain in 10 rounds} &= \frac{-1+1-1+1-1-2+1+1+1-2}{10} \\ &= -0.2 \,.\end{aligned}$$

Figure 3.3 shows a sequence of 1 000 rounds. This is only one sequence out of a large number of possible sequences. With regard to the number of possible sequences, we apply the same counting technique as for the toto: the number of possible sequences is 5^{1000}, which is closely represented by 4 followed by 71 zeros! (A number larger than the number of atoms in the Milky Way, estimated as 1 followed by 68 zeros.)

Figure 3.5 shows 10 curves of the average gain in the AKQJF game, corresponding to 10 possible sequences of 1 000 rounds. We observe a tendency for all curves to stabilize around the mathematical expectation 0. However, we also observe that some average gain curves that looked as though they were approaching zero departed from it later on. Let us suppose that we establish a certain deviation around zero, for instance 0.08 (the broken lines in Fig. 3.5). Let us ask whether it is possible to guarantee that, after a certain number n of rounds, all the curves will deviate from zero by less than 0.08?

Let us see what happens in the situation shown in Fig. 3.5. When $n = 200$, we have 4 curves (out of 10) that exceed the specified deviation. Near $n = 350$, we have 3 curves. After around 620, we only have one curve and finally, after approximately 850, no curve exceeds the specified deviation. This experimental evidence suggests, therefore, that the probability of finding a curve that deviates from the mathematical expectation by more than a pre-specified tolerance will become smaller and smaller as n increases.

Now, it is precisely such a law that the mathematician Jacob Bernoulli, already mentioned in Chap. 1, was able to demonstrate and

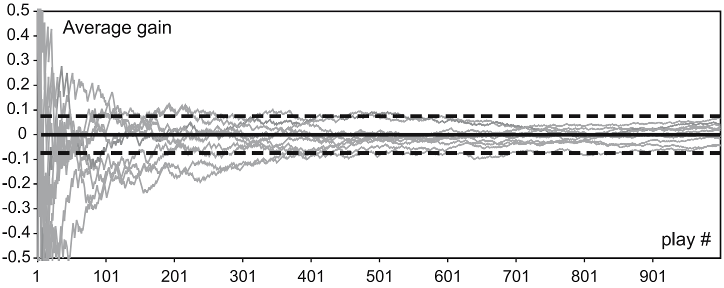


Fig. 3.5. Ten possible evolutions of the average gain in the card game

which has come to be known as the *law of large numbers* or *Bernoulli's law*: for a sequence of random variables with the same probability law and independent from each other (in our example, no round depends on the others), the probability that the respective average deviates from the mathematical expectation by more than a certain pre-established value tends to zero. Turning back to our previous question, we cannot guarantee that after a certain number of rounds all average gain curves will deviate from the mathematical expectation by less than a certain value, but it is nevertheless possible to determine a degree of certainty (compute the probability) of such an event, and that degree of certainty increases with the number of rounds.

This way, the probability measure (for instance, of heads turning up in coin tossing) does not predict the outcome of a particular sequence, but, in a certain sense, 'predicts' the average outcome of a large number of sequences. We cannot predict the *detail* of all possible evolutions, but we can determine a certain *global* degree of certainty.

In a game between John and Mary, if John bets a million with 1/1000 probability of losing and Mary bets 1000 with 999/1000 probability of losing, the gain expectation is:

$$1\,000 \times \frac{999}{1\,000} - 1\,000\,000 \times \frac{1}{1\,000} = -1 \quad \text{for John}\,,$$
$$1\,000\,000 \times \frac{1}{1\,000} - 1\,000 \times \frac{999}{1\,000} = +1 \quad \text{for Mary}\,.$$

Therefore, for sufficiently large n, the game is favorable to Mary (see Fig. 3.6). However, if the number of rounds is small, John has the advantage.

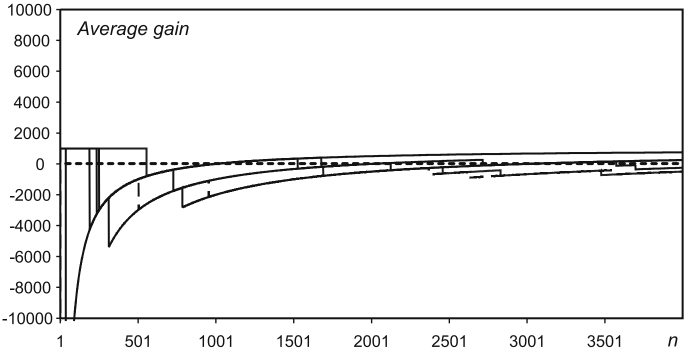


Fig. 3.6. Evolution of John's average gain with the number of rounds, in six possible sequences. Only in one sequence is chance unfavorable to John during the first 37 rounds

When we say 'tends to zero', in the law of large numbers, the meaning is that it is possible to find a sufficiently high value of n after which the probability of a curve exceeding a pre-specified deviation is below any arbitrarily low value we may wish to stipulate. We shall see later what value of n that is, for this *convergence in probability*.

3.3 Bad Luck Games

The notion of mathematical expectation will allow us to make a better assessment of the degree of 'bad luck' in so-called games of chance. Let us consider the European roulette game discussed in Chap. 1 and let us assume that a gambler only bets on a specific roulette number and for that purpose (s)he pays 1 euro to the casino. Let us further assume that if the gambler wins the casino will pay him/her 2 euros.

For a sufficiently large number of rounds the player's average gain will approach the mathematical expectation

$$E(\text{roulette}) = 2 \times \frac{1}{37} + 0 \times \frac{36}{37} = 0.054 \text{ euros}\,.$$

Since the player pays 1 euro for each bet, the roulette game is unfavorable to him/her, implying an average loss of $1-0.054 = 0.946$ euros per bet. The game would only be equitable if the casino paid 37 euros (!) for each winning number. On the other hand, for a two euro payment by the casino, the game would only be equitable for the player if (s)he paid the mathematical expectation for betting, namely 0.054 euros.

Let us now analyze the slot machine described in Chap. 1 (see Fig. 1.8). The question is: how much would a gambler have to pay for the right to one equitable round at the slot machine? To answer this question we compute the mathematical expectation using the values presented in Chap. 1:

$$\begin{aligned} E(\text{slot machine}) = {} & 0.05 \times 0.090\,353 + 0.10 \times 0.048\,695 \\ & +0.15 \times 0.022\,129 + 0.25 \times 0.003\,155 \\ & +0.30 \times 0.002\,258 + 0.50 \times 0.000\,666 \\ & +1.00 \times 0.000\,079 + 2.50 \times 0.000\,001 = \$\,0.01\ . \end{aligned}$$

Therefore the gambler should only pay 1 cent of a dollar per go.

The winning expectation in state games (lottery, lotto, toto, and the like) is also, as can be expected, unfavorable to the gambler. Consider the toto and let 0.35 euros be the payment for one bet. In state games, from the total amount of money amassed from the bets, only half is spent on prizes, and approximately in proportion to the probabilities of each type of hit. Now, as we saw in Chap. 1, for each first prize, one expects 43 second prizes, 258 third prizes, 13 545 fourth prizes and 246 820 fifth prizes. Consequently, 50% of the amassed money will be used to pay prizes with a probability of:

$$\begin{aligned} P(\text{prizes}) &= 0.000\,000\,071\,5 \times (1 + 43 + 258 + 13\,545 + 246\,820) \\ &= 0.018\,6 \,. \end{aligned}$$

Therefore,

$$E(\text{toto}) = 0.018\,6 \times 0.175 - 0.9814 \times 0.175 = -0.096 \text{ euros}\,.$$

That is, the toto gambler loses on average (in the sense of the law of large numbers), 0.096 euros per bet.

3.4 The Two-Envelope Paradox

The two-envelope paradox, also known as the swap paradox, seems to have been created by the Belgian mathematician Maurice Kraitchik (1882–1957) in 1953. There is probably no other paradox (or better, pseudo-paradox) of probability theory that has stimulated so much discussion in books and papers. It is interesting to note that its enunciation is apparently simple and innocuous. Let us consider a first version of the paradox. Suppose I show the reader two identical and closed envelopes with unknown contents. The only thing that is disclosed is that they contain certain amounts of money, with the amount in one of the envelopes being double the amount in the other one. The reader may choose one envelope telling me his/her choice. Having told me his/her choice, the reader is given the possibility of swapping his/her choice. What should the reader do? Swap or not swap? It seems obvious that any swap is superfluous. It is the same thing as if I offered the possibility of choosing heads or tails in the toss of a coin. The probability is always 1/2. But let us contemplate the following line of thought:

1. The envelope I chose contains x euros.
2. The envelope I chose contains either the smaller or the bigger amount, as equally likely alternatives.

3. If my envelope contains the smaller amount I will get $2x$ euros if I swap and I would therefore win x euros.
4. On the other hand, if my envelope contains the bigger amount, by swapping it I will get $x/2$ euros and I would thus lose $x/2$ euros.
5. Therefore, the expected gain if I make the swap is

$$E(\text{swap gain}) = 0.5 \times x - 0.5 \times (x/2) = 0.25x \text{ euros}\,.$$

That is, I should swap since the expected gain is positive. (The wisest decision is always guided by the mathematical expectation, even if it is not always the winning decision it is definitely the winning decision with arbitrarily high probability in a sufficiently large number of trials.) But, once the swap takes place the same argument still holds and therefore I should keep swapping an infinite number of times!

The crux of the paradox lies in the meaning attached to x in the formula E(swap gain). Let us take a closer look by attributing a value to x. Let us say that x is 10 euros and I am told that the amount in the other envelope is either twice this amount (20 euros) or half of it (5 euros), with equal probability. In this case I should swap, winning on average 2.5 euros with the swap. There is no paradox in this case. Or is there? If I swap, does the same argument not hold? For suppose that the other envelope contains y euros. Therefore, the amount of the initially chosen envelope is either $2y$ euros or $y/2$ euros and the average swap gain is $0.25y$ euros. Then the advantage in continuing to swap ad infinitum would remain, and so would the paradox. The problem with this argument is that, once we have fixed x, the value of y is *not* fixed: it is either 20 euros or 5 euros. Therefore we *cannot* use y in the computation of E(swap gain), as we did above. This aspect becomes clear if we consider the *standard* version of the paradox that we analyze in the following.

In the standard version of the paradox what is considered fixed is the total amount in the two envelopes (the total amount does not depend on my original choice). Let $3x$ euros be the total amount. Then, my initial choice contains either x euros or $2x$ euros, and the calculation in step 5 of the above argument is not legitimate, since *it assumes a fixed value* for x. If, for instance, the initial envelope contains the lower amount, x euros, I win x euros with the swap; but if it is the higher amount, $2x$ euros, and I do not lose $x/2$ euros as in the formula of step 5, but in fact x euros. Therefore, $E(\text{swap gain}) = 0.5x - 0.5x = 0$ euros and I do not win or lose with the swap.

Other solutions to the paradox have been presented, based on the fact that the use of the formula in step 5 would lead to impossible probabilistic functions for the amount in one of the envelopes (see, e.g., Keith Devlin, 2005).

Do the two-envelope paradox and its analysis have any practical interest or is it simply an 'entertaining problem'? In fact, this and other paradoxes do sometimes have interesting practical implications. Consider the following problem: let x be a certain amount in euros that we may exchange today in dollars at a rate of 1 dollar per euro, and later convert back to euros, at a rate of 2 dollars or 1/2 dollar per euro, with equal probability. Applying the envelope paradox reasoning we should perform these conversions hoping to win $0.25x$ euros on average. But then the dollar holder would make the opposite conversions and we arrive at a paradoxical situation known as the *Siegel paradox*, whose connection with the two-envelope paradox is obvious (see Friedel Bolle, 2003).

3.5 The Saint Petersburg Paradox

We now discuss a coin-tossing game where a hypothetical casino (extremely confident in the gambler's fallacy) accepts to pay 2 euros if heads turns up at the first throw, 4 euros if heads only turns up at the second throw, 8 euros if it only turns up at the third throw, 16 euros if it only turns up at the fourth throw, and so on and so forth, always doubling the previous amount. The question is: how much money should the player give to the casino as initial down-payment in order to guarantee an equitable game?

The discussion and solution of this problem were first presented by Daniel Bernoulli (nephew of Jacob Bernoulli) in his *Commentarii Academiae Scientiarum Imperialis Petropolitanae* (Comments of the Imperial Academy of Science of Saint Petersburg), published in 1738. The problem (of which there are several versions) came to be known as the Saint Petersburg paradox and was the subject of abundant bibliography and discussions. Let us see why. The probability of heads turning up in the first throw is 1/2. For the second throw, taking into account the compound event probability rule for independent events, we have

$$P(\text{first time tails}) \times P(\text{second time heads}) = \left(\frac{1}{2}\right)^2 ,$$

and likewise for the following throws, always multiplying the previous value by 1/2. In this way, as shown in Fig. 3.7b, the probability rapidly decreases with the throw number (the random variable) when heads turns up for the first time. It then comes as no surprise that the casino proposes to pay the player an amount that rapidly grows with the throw number. Applying the rule that the player must pay the mathematical expectation to make the game equitable, this would correspond to the player having to pay

$$\begin{aligned} E(\text{Saint Petersburg game}) &= 2 \times \frac{1}{2} + 4 \times \frac{1}{4} + 16 \times \frac{1}{16} + \cdots \\ &= 1 + 1 + 1 + 1 + \cdots = \infty \,. \end{aligned}$$

The player would thus have to pay an infinite (!) down-payment to the casino in order to enter the game.

On the other hand, assuming that for entering the game the player would have to pay the prize value whenever no heads turned up, when heads did finally turn up at the nth throw (s)he would have disbursed

$$2 + 4 + 8 + \cdots + 2^{n-1} = 2^n - 2 \text{ euros}\,.$$

At the nth throw (s)he would finally get the 2^n euro prize. Thus, only a net profit of 2 euros is obtained. One is unlikely to find a gambler who would pay out such large amounts for such a meager profit.

The difficulty with this paradox lies in the fact that the mathematical expectation is infinite. Strictly, this game has no mathematical expectation and the law of large numbers is not applicable. The paradox is solvable if we modify the notion of equitable game. Let us consider

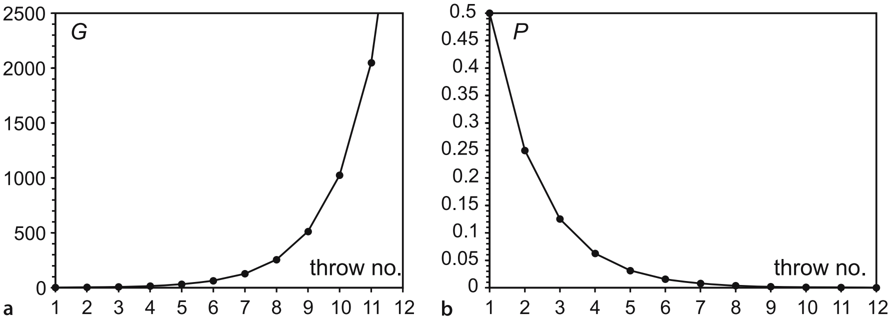


Fig. 3.7. (**a**) The exponential growth of the gain in the paradoxical Saint Petersburg game. (**b**) The decreasing exponential probability in the same game

Table 3.1. Relevant functions for the Saint Petersburg paradox

n	$P(n)$	Gain f [euro]	Utility $\log f$	Expected utility
1	1/2	2	0.693 147	0.346 574
2	1/4	4	1.386 294	0.693 147
3	1/8	8	2.079 442	0.953 077
4	1/16	16	2.772 589	1.126 364
5	1/32	32	3.465 736	1.234 668
6	1/64	64	4.158 883	1.299 651
7	1/128	128	4.852 030	1.337 557
8	1/256	256	5.545 177	1.359 218
9	1/512	512	6.238 325	1.371 403
10	1/1024	1024	6.931 472	1.378 172

the sum of the gains after n throws, which we denote S_n. Moreover, let us consider that, instead of a fixed amount per throw, the gambler's payments are variable, with an accumulated payment denoted P_n after n throws. We now modify the notion of an equitable game by imposing the convergence in probability towards 1 of the ratio S_n/P_n. In other words, we would like any imbalance between what the casino and the gambler have already paid to become more and more improbable as the number of throws increases. In this sense it is possible to guarantee the equitability of the Saint Petersburg game if the player pays in accordance with the following law:

$$P_n = n \times \log_2 n + 1 \ ,$$

corresponding to paying the following amounts:

$$1, 2, 2.76, 3.25, 3.61, 3.9, 4.14, 4.35, 4.53, 4.69, 4.83, 4.97, 5.09, 5.2, \ldots \ .$$

When Daniel Bernoulli presented the Saint Petersburg paradox in 1738, he also presented a solution based on what he called the utility function. The idea is as follows. Any variation of an individual's fortune, say f, has different meanings depending on the value of f.

Let us denote the variation of f by Δf. If, for instance, an individual's fortune varies by $\Delta f = 10\,000$ euros, the meaning of this variation for an individual whose fortune is $f = 10\,000$ euros (100% variation) is quite different from the meaning it has for another individual whose fortune is $f = 1\,000\,000$ euros (1% variation). Let us then consider that any transaction between individuals is deemed equivalent only when the values of $\Delta f/f$ are the same. This amounts to saying that there is a utility function of the fortune f expressed by $\log(f)$. (Effectively, the base e logarithm function – see the appendix at the end of the book – enjoys the property that its variation is expressed in that way, viz., $\Delta f/f$.) Table 3.1 shows the expected utilities for the Saint Petersburg game, where the gain represents the fortune f. The expected utilities converge to a limit (1.386 294) corresponding to 4 euros. The rational player from the money-utility point of view would then pay an amount lower than 4 euros to enter the game (but we do not know whether the casino would accept it).

3.6 When the Improbable Happens

In Chap. 1 reference was made to the problem of how to obtain heads in 5 throws, either with a fair coin – $P(\text{heads}) = P(\text{tails}) = 0.5$ – or with a biased one – $P(\text{heads}) = 0.65$, $P(\text{tails}) = 0.35$. We are now able to compute the mathematical expectation in both situations. Taking into account the probability function for 0 through 5 heads turning up for a fair coin (see Fig. 1.7a), we have

$$\begin{aligned} E(k \text{ heads in 5 throws}) &= 0 \times 0.031\,25 + 1 \times 0.156\,25 + 2 \times 0.312\,5 \\ &\quad +3 \times 0.312\,5 + 4 \times 0.156\,25 + 5 \times 0.031\,25 \\ &= 2.5\ . \end{aligned}$$

We have obtained the entirely expected result that, for a fair coin, heads turns up on average in half of the throws (for 5-throw sequences or any other number of throws).

There are a vast number of practical situations requiring the determination of the probability of an event occurring k times in n independent repetitions of a certain random experiment. Assuming that the event that interests us (success) has a probability p of occurring in any of the repetitions, the probability function to use is the same as the one found in Chap. 1 for the coin example, that is,

$$P(k \text{ successes in } n) = \binom{n}{k} \times p^k \times (1-p)^{n-k}\ .$$

The distribution of probability values according to this formula is called the *binomial distribution*. The name comes from the fact that the probabilities correspond to the terms in the binomial expansion of $(p+q)^n$, with $q = 1-p$. It can be shown that the mathematical expectation of the binomial distribution (the *distribution mean*) is given by

$$E(\text{binomial distribution}) = n \times p\,.$$

We can check this result for the above example: $E = 5 \times 0.5 = 2.5$. For the biased coin we have $E = 5 \times 0.65 = 3.25$. Note how the expectation has been displaced towards a higher number of heads. Note also (see Fig. 1.7) how the expectation corresponds to a central position (center of gravity) of the probability distribution. Incidentally, from $f = k/n$ and $E(k) = np$, we arrive at the result mentioned in the first section of this chapter that the mathematical expectation of an event frequency is its probability, i.e., $E(f) = p$.

Consider a sequence of 20 throws of a fair coin in which 20 heads turned up. Looking at this sequence

$$1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1\,,$$

one gets the feeling that something very peculiar has happened. As in the gambler's fallacy, there is a great temptation to say that, at the twenty-first throw, the probability of turning up tails will be higher. In other situations one might be tempted to attribute a magical or divine cause to such a sequence. The same applies to other highly organized sequences, such as

$$0,0,0,0,0,0,0,0,0,0,1,1,1,1,1,1,1,1,1,1\,,$$

or

$$1,1,0,0,1,1,0,0,1,1,0,0,1,1,0,0,1,1,0,0\,.$$

Nevertheless all these sequences have the same probability of occurring as any other sequence of 20 throws, namely, $(1/2)^{20} = 0.000\,000\,953\,67$. If we repeat the 20-throw experiment a million times, we expect to obtain such sequences (or any other sequence for that matter) a number of times given by $n \times p = 1\,000\,000 \times 0.000\,000\,95 \approx 1$. That is, there is a high certainty that once in a million such a sequence may occur. In the case of the toto game, where the probability of the first prize is 0.000 000 627, we may consider what happens if there are 1.6 million bets. The expectation is 1, i.e., on average, we will find (with high certainty) a first prize among 1.6 million bets. The conclusion to be drawn is that, given a sufficiently large number of experiments, the 'improbable' will happen: highly organized sequences, emergence of life on a planet, or our own existence.

3.7 Paranormal Coincidences

Imagine yourself dreaming about the death of someone you knew. When reading the newspaper the next day you come across the column reporting the death of your acquaintance. Do you have a premonitory gift? People ‘studying’ so-called paranormal phenomena often present various experiences of this sort as evidence that ‘something’ such as telepathy or premonition is occurring.

Many reports on ‘weird’ cases, supposedly indicating supernatural powers, contain, in fact, an incorrect appreciation of the probabilities of coincident events. We saw in the birthday problem that it suffices to have 23 people at a party in order to obtain a probability over 50% that two of them share the same birthdays. If we consider birthday dates separated by at most one day, the number 23 can be dropped to 14!

Let us go back to the problem of dreaming about the death of someone we knew and subsequently receiving the information that (s)he has died. Let us assume that in 30 years of adult life a person dreams, on average, only once about someone he/she has known during that 30 year span. (This is probably a low estimate.) The probability of dreaming about the death of an acquaintance on the day preceding his/her death (one hit in 30×365.25 days) is then $1/(30 \times 365.25)$. Let us further assume that in the 30 year span there are only 15 acquaintances (friends and relatives) whose death is reported in the right time span, so that the coincidence can be detected. With this assumption, the probability of a coincidence taking into account the 15 deaths is $15/(30 \times 365.25) = 1/730.5$ in 30 years of one’s life. The probability of one coincidence per year is estimated as $1/(30 \times 730.5) = 0.000\,045\,6$. Finally, let us take, for instance, the Portuguese adult population estimated at 7 million people. We may then use the binomial distribution formula for the mean in order to estimate how many coincidences occur per year, on average, in Portugal:

$$\begin{aligned}\text{average number of coincidences per year} &= n \times p \\ &= 7\,000\,000 \times 0.000\,045\,6 \\ &= 319\ .\end{aligned}$$

Therefore, a ‘paranormal’ coincidence will occur, on average, about once per day (and of course at a higher rate in countries with a larger population). We may modify some of the previous assumptions without significantly changing the main conclusion: ‘paranormal’ coincidences are in fact entirely normal when we consider a large population. Once again, given a sufficiently large number, the ‘improbable’ will happen.

3.8 The Chevalier de Méré Problem

On 29 July 1654, Pascal wrote a long letter to Fermat presenting the solution of a game problem raised by the Chevalier de Méré, a member of the court of Louis XIV already mentioned in Chap. 1. The problem was as follows. Two players (say Primus and Secundus) bet 32 pistoles (the French name for a gold coin) in a heads or tails game. The 64 pistoles would stay with the player that first obtained 3 successes (a success might be, for instance, heads for Primus and tails for Secundus), either consecutive or not. In the first round Primus wins; at that moment the two players are obliged to leave without ever finishing the match. How should the bet be divided in a fair way between them?

Fermat, in a previous letter to Pascal, had proposed a solution based on the enumeration of all possible cases. Pascal presented Fermat with another, more elegant solution, based on a hierarchical description of the problem and introducing the mathematical concept of expectation. Let us consider Pascal’s solution, based on the enumeration of all the various paths along which the game might have evolved. It is useful to represent the enumeration in a tree diagram, as shown in Fig. 3.8.

The circle marked 44 corresponds to the initial situation, where we assume that Primus won the first round. For the moment we do not know the justification for this and other numberings of the circles. Suppose that Primus wins again; we then move to the circle marked 56, following the branch marked P (Primus wins). If, on the other hand, Secundus wins the second round, we move downwards to the circle marked 32, in the tree branch marked S. Thus, upward branches correspond to Primus wins (P) and downward branches to Secundus wins (S). Note that, since a Primus loss is a Secundus win and vice versa, the match goes on for only 5 rounds. Let us now analyze the tree from right to left (from the end to the beginning), starting with the gray circles. The circle marked 64 corresponds to a final P win, with three heads for P according to the sequence PPSSP. P gets the 64 pistoles. The circle marked 0 corresponds to a final win for S with three tails for S according to the sequence PPSSS. Had the match been interrupted immediately before this final round, what would P’s expectation be? We have already learned how to do this calculation:

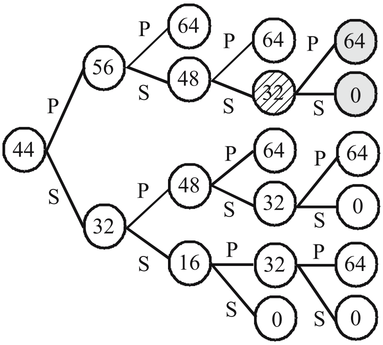


Fig. 3.8. Tree diagram for Pascal's solution

$$E(P) = 64 \times \frac{1}{2} + 0 \times \frac{1}{2} = 32 \; .$$

We thus obtain the situation represented by the hatched circle. For the paths that start with PPS, we obtain in the same way, at the third round,

$$E(P) = 64 \times \frac{1}{2} + 32 \times \frac{1}{2} = 48 \; .$$

The process for computing the expectation repeats in the same way, progressing leftward from the end circles, until we arrive at the initial circle to which the value 44 has been assigned. In this way, when he leaves, Primus must keep 44 pistoles and Secundus 20. A problem that seemed difficult was easily solved using the concept of mathematical expectation. As a final observation, let us note that there are 10 end circles: 10 possible sequences of rounds. Among these, Primus wins 6 and Secundus 4. In spite of this, the money at stake must not be shared in the ratio 6/10 (38.4 pistoles) for Primus and 4/10 (25.6 pistoles) for Secundus!

Once more a simple and entertaining problem has important practical consequences. In fact, Pascal's reasoning based on an end-to-start analysis, is similar to the reasoning that leads to the price calculation of a stock market option, particularly in the context of a famous formula established by Fisher Black and Myron Scholes in 1973 (which got them the Nobel Prize for Economics in 1997).

3.9 Martingales

The quest for strategies that allow one to win games is probably as old an endeavor as mankind itself. Let us imagine the following game between two opponents A and B. Both have before them a stack of coins, all of the same type. Each piles them up and takes out a coin from their stack at the same time. If the drawn coins show the same face, A wins from B in the following way: 9 euros if it is heads–heads and 1 euro if it is tails–tails. If the upper faces of the coins are different, B wins 5 euros from A no matter whether it is heads–tails or tails–heads. Player A will tend to pile the coins up in such a way as to exhibit heads more often, expecting to find many heads–heads situations. On the other hand, B, knowing that A will try to show up heads, will try to show tails.

One might think that, since the average gain in the situation that A wins, viz., $(9+1)/2$, and in the situation that B wins, viz., $(5+5)/2$, are the same, it will not really matter how the coins are piled up. However, this is not true. First, consider A piling up the coins so that they always show up heads. Let t represent the fraction of the stack in which B puts heads facing up and let us see how the gain of B varies with t. Let us denote that gain by G. B must then pay 9 euros to A during a fraction t of the game, but will win 5 euros during a fraction $1-t$ of the game. That is, the gain of B is

$$G(\text{stack A: heads}) = -9t + 5(1-t) = -14t + 5 \ .$$

If the coins of stack A show tails, one likewise obtains

$$G(\text{stack A: tails}) = +5t - 1(1-t) = 6t - 1 \ .$$

As can be seen from Fig. 3.9, the B gain lines for the two situations intersect at the point $t = 0.3$, with a positive gain for B of 0.8 euros. What has been said for a stack segment where A placed the same coin face showing up applies to the whole stack. This means that if B stacks his coins in such a way that three tenths of them, randomly distributed in the stack, show up heads, he will expect to win, during a long sequence of rounds, an average of 0.8 euros per round.

The word 'martingale' is used in games of chance to refer to a winning strategy, like the one just described. The scientific notion of *martingale* is associated with the notion of an equitable game. This notion appeared for the first time in the book *The Doctrine of Chance* by Abraham de Moivre, already mentioned earlier, and plays an important role in probability theory and other branches of mathematics.

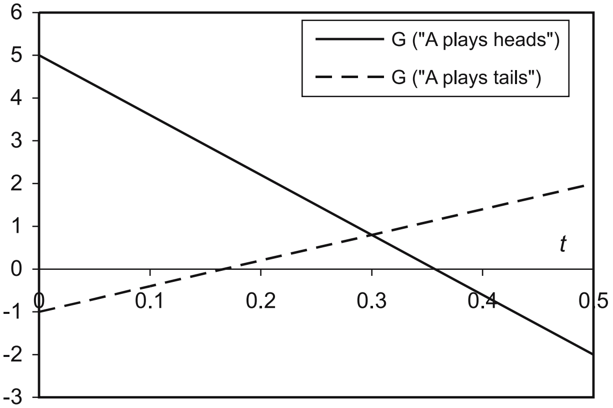


Fig. 3.9. How the gains of B vary with a fraction t of the stack where he plays heads facing up

The word comes from the Provençal word 'martegalo' (originating from the town of Martigues), and described a leather strap that prevented a horse from lifting up its head. A martingale is similarly intended to hold sway over chance.

One must distinguish true and false martingales. Publications sometimes appear proposing winning systems for the lotto, toto, and so on. One such system for winning the lotto is based on the analysis of graphs representing the frequency of occurrence of each number as well as combinations of numbers. Of course such information does not allow one to forecast the winning bet; this is just one more version of the gambler's fallacy. Other false martingales are more subtle, like the one that we shall describe at the end of Chap. 4.

3.10 How Not to Lose Much in Games of Chance

Let us suppose that the reader, possessing an initial capital of I euros, has decided to play roulette with even–odd- or red–black-type bets, set at a constant value of a euros. The gain–loss sequence constitutes, in this case, a coin-tossing type of sequence, with probability $p = 18/37 = 0.486$ of winning and $q = 1-p$ of losing. Let us further suppose that the reader decides at the outset to keep playing either until he/she wins F euros or until ruined by losing the initial capital. Jacob Bernoulli demonstrated in 1680 that such a game ends in a finite number of steps with a winning probability given by

Table 3.2. Roulette for different size bets. F represents the winnings at which the player stops and $P(F)$ is the probability of reaching that value

50 euro bets		100 euro bets		200 euro bets	
F	$P(F)$	F	$P(F)$	F	$P(F)$
1050	0.9225	1100	0.8826	**1200**	**0.8100**
1100	0.8527	**1200**	**0.7853**	1400	0.6747
1150	0.7896	1300	0.7034	1600	0.5736
1200	**0.7324**	1400	0.6337	1800	0.4952
1250	0.6804	1500	0.5736	2000	0.4328
1300	0.6330	1600	0.5215	2200	0.3820
1350	0.5896	1700	0.4758	2400	0.3399
1400	0.5498	1800	0.4356	2600	0.3045

$$P(\text{winning } F) = P(F) = \frac{1-(q/p)^{I/a}}{1-(q/p)^{F/a}} \; .$$

In the case of roulette, $q/p = 1.0556$. If the reader starts with an initial capital of 1 000 euros and makes 50-euro bets, the probability of reaching 1 200 euros is given by

$$P(F) = \frac{1-1.0556^{20}}{1-1.0556^{24}} = 0.7324 \; .$$

Table 3.2 shows what happens when we make constant bets of 50, 100 and 200 euros.

Looking at the values for situations with $F = 1200$ euros, printed in bold in the table, we see that the gambler would be advised to stay in the game the least time possible, whence he must *risk the highest possible amount in a single move*. In the above example one single 200-euro bet has 81% probability of reaching the goal set by the gambler. This probability value is near the theoretical maximum value $P(F) = 1000/1200 = 83.33\%$. If the bet amounts can vary, there are then betting strategies to obtain a probability close to that maximum value. One such strategy, proposed by Lester Dubins and Leonard Savage in 1956, is inspired by the idea of always betting the highest possible amount that allows one to reach the proposed goal. In the example we have been discussing, we should then set the sequence of bets as shown in Fig. 3.10.

The following probabilities can be computed from the diagram:

$$P(\text{going from 1000 to 1200}) = p + q \times P(\text{going from 800 to 1200}) \; ,$$

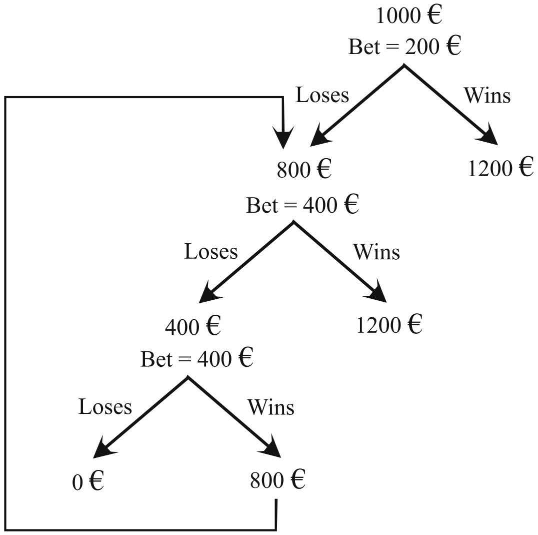


Fig. 3.10. The daring bet strategy proposed by Dubins and Savage

$$P(\text{going from 800 to 1200}) = p + q \times p \times P(\text{going from 800 to 1200}) .$$

From this one concludes that

$$P(\text{going from 1000 to 1200}) = p + \frac{qp}{1 - qp} = 0.8195 ,$$

closer to the theoretical maximum.

As a final remark let us point out that the martingale of Dubins and Savage looks quite close to a martingale proposed by the French mathematician Jean Le Rond d'Alembert (1717–1783), according to which the gambler should keep doubling his/her bet until he/she wins or until he/she has enough money to double the bet. For the above example, this would correspond to stopping when reaching 400 euros, with a probability of winning of 0.7363.

Needless to say, these martingales should not be viewed as an incitement to gambling – quite the opposite. As a matter of fact, there is even a result about when one should stop a constant-bet game of chance in order to maximize the winning expectation. The result is: don't play at all! In the meantime we would like to draw the reader's attention to the fact that all these game-derived results have plenty of practical applications in different areas of human knowledge.

4

The Wonderful Curve

4.1 Approximating the Binomial Law

The binomial distribution was introduced in Chap. 1 and its mathematical expectation in the last chapter. We saw how the binomial law allows us to determine the probability that an event, of which we know the individual probability, occurs a certain number of times when the corresponding random experiment is repeated.

Take for instance the following problem. The probability that a patient gets infected with a certain bacterial strain in a given hospital environment is $p = 0.0055$. What is the probability of finding more than 10 infected people in such a hospital with 1000 patients? Applying the binomial law, we have

$$\begin{aligned} P\begin{pmatrix}\text{more than 10}\\ \text{events in 1000}\end{pmatrix} &= P(11\text{ events}) + P(12\text{ events}) \\ &\qquad + \cdots + P(1000\text{ events}) \\ &= \binom{1000}{11} \times 0.0055^{11} \times 0.9945^{989} \\ &\quad + \binom{1000}{12} \times 0.0055^{11} \times 0.9945^{988} \\ &\quad + \cdots + \binom{1000}{1000} \times 0.0055^{1000} \times 0.9945^{0}\ . \end{aligned}$$

The difficulty in carrying out these calculations can be easily imagined. For instance, calculating $\binom{1000}{500}$ involves performing the multiplication $1000 \times 999 \times \ldots \times 501$ and then dividing by $500 \times 499 \times \ldots \times 2$.

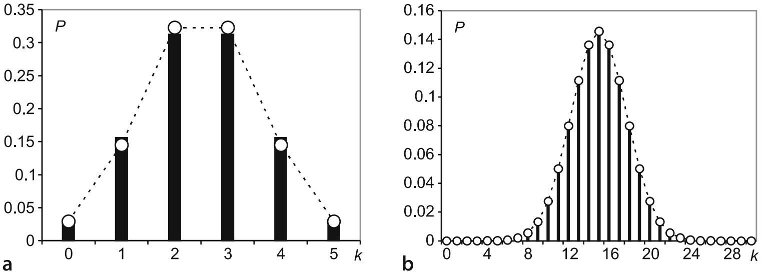


Fig. 4.1. De Moivre's approximation (*open circles*) of the binomial distribution (*black bars*). (**a**) $n = 5$. (**b**) $n = 30$

Even if one applies the complement rule – subtracting from 1 the sum of the probabilities of 0 through 10 events – one still gets entangled in painful calculations. When using a computer, it is not uncommon for this type of calculation to exceed the number representation capacity.

Abraham de Moivre, mentioned several times in previous chapters, was the first to find the need to apply an easier way of computing factorials. (In his day, performing the above calculations was an extremely tedious task.) That easier way (known as Stirling's formula) led him to an approximation of the binomial law for high values of n. Denoting the mathematical expectation of the binomial distribution by m ($m = np$) and the quantity $np(1-p)$ by v, the approximation obtained by Abraham de Moivre, assuming that m is not too small, was the following:

$$\binom{n}{k} \times p^k \times (1-p)^{n-k} \approx \frac{1}{\sqrt{2\pi v} \times \mathrm{e}^{(k-m)^2/2v}} \ .$$

We have already found Napier's number, universally denoted by the letter e, with the approximate value 2.7, when discussing the problem of encounters (Chap. 2). Let us see if De Moivre's approximation works for the example in Chap. 1 of obtaining 3 heads in 5 throws of a fair coin:

$$m = np = \tfrac{5}{2} = 2.5 \ , \quad v = np(1-p) = \tfrac{2.5}{2} = 1.25 \ ,$$
$$k - m = 3 - 2.5 = 0.5 \ ,$$

whence

$$P(3 \text{ heads in 5 throws}) \approx \frac{1}{\sqrt{7.85} \times 2.7^{0.25/2.5}} = \frac{1}{2.8 \times 2.7^{0.1}} = 0.32 \ ,$$

which is really quite close to the value of 0.31 we found in Chap. 1. Figure 4.1 shows the probability function of the binomial distribution (bars) for $n = 5$ and $n = 30$ and Abraham de Moivre's approximation (open circles). For $n = 5$, the maximum deviation is 0.011; for $n = 30$, the maximum deviation comes down to a mere 0.0012.

4.2 Errors and the Bell Curve

Abraham de Moivre's approximation (described in a paper published in 1738, with the second edition of the book *The Doctrine of Chances*) was afterwards worked out in detail by the French mathematician Pierre-Simon Laplace (1749–1827) in his book *Théorie Analytique des Probabilités* (1812) and used to analyze the errors in astronomical measurements. More or less at the same time (1809), the German mathematician Johann Carl Friedrich Gauss (1777–1855) rigorously justified the distribution of certain sequences of measurement errors (particularly geodesic measurements) in agreement with that formula.

It is important at this point to draw the reader's attention to the fact that, in science, the word 'error' does not have its everyday meaning of 'fault' or 'mistake'. This semantic difference is often misunderstood. When measuring experimental outcomes, the word 'error' refers to the *unavoidable uncertainty* associated with any measurement, however *aptly* and *carefully* the measurement procedure has been applied. The error (uncertainty) associated with a measurement is, therefore, the effect of chance phenomena; in other words, it is a random variable.

Consider weighing a necklace using the spring balance with 2 g minimum scale resolution shown in Fig. 4.2. Suppose you measure 45 g for the weight of the necklace. There are several uncertainty factors influencing this measurement:

- The *measurement system resolution*: the balance hand and the scale lines have a certain width. Since the hand of Fig. 4.2 has a width almost corresponding to 1 g, it is difficult to discriminate weights differing by less than 1 g.
- The *interpolation error* that leads one to measure 45 g because the scale hand was *approximately* half-way between 44 g and 46 g.
- The *parallax error*, related to the fact that it is difficult to guarantee a perfect vertical alignment of our line of sight and the scale plane; for instance, in Fig. 4.2, if we look at the hand with our line of sight

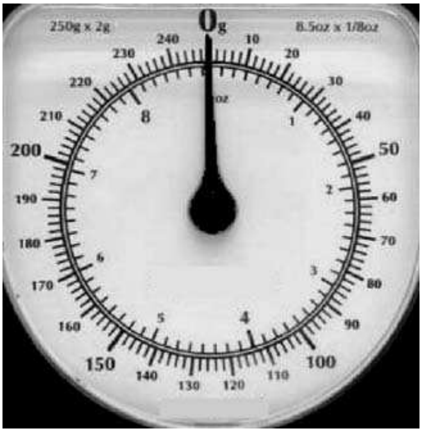


Fig. 4.2. Spring balance with 2 g minimum resolution

to its right, we read a value close to 250 g; if we look to its left, we read a value close to 2 g.

- The *background noise*, corresponding in this case to spring vibrations that are transmitted to the hand.

If we use a digital balance, we eliminate the parallax and interpolation errors, but the effects of limited resolution (the electronic sensor does not measure pressures below a certain value) and background noise (such as the thermal noise in the electronic circuitry) are still present.

Conscious of these uncertainty factors, suppose we weighed the necklace fifty times and obtained the following sequence of fifty measurements:

$$44, 45, 46, 43, 47, 47, 45, 48, 48, 44, 43, 45, 46, 44, 43, 44, 45, 49, 46, 48,$$

$$45, 43, 46, 45, 44, 44, 45, 45, 46, 45, 46, 48, 47, 44, 49, 44, 47, 45, 46, 48,$$

$$46, 46, 47, 49, 45, 45, 47, 46, 46, 45\ .$$

The average weight is 45.68 g. It seems reasonable to consider that since the various error sources act independently of one other, with contributions on either side of the true value, the errors should have a tendency to cancel out when the weight is determined in a long series of measurements. Things happen as though the values of a random variable (the ‘error’) were added to the true necklace weight, this random variable having zero mathematical expectation. In a long sequence of measurements we expect, from the law of large numbers, the average of the errors to be close to zero, and we thus obtain the true weight value.

Let us therefore take 45.68 g as the true necklace weight and consider now the percentage of times that a certain error value will occur (by first subtracting the average from each value of the above sequence). Figure 4.3 shows how these percentages (vertical bars) vary with the error value. This graphical representation of frequencies of occurrence of a given random variable (the error, in our case) is called a *histogram*. For instance, an error of 0.68 g occurred 13 times, which is 26% of the time.

Laplace and Gauss discovered independently that, for many error sequences of the most varied nature (relating to astronomical, geodesic, geomagnetic, electromagnetic measurements, and so on), the frequency distributions of the errors were well approximated by the values of the function

$$f(x) = \frac{1}{\sqrt{2\pi\nu}} \mathrm{e}^{-x^2/2\nu} ,$$

where x is the error variable and ν represents the so-called mean square error. Let us try to understand this. If we subtract the average weight from the above measurement sequence, we obtain the following error sequence: -1.68, -0.68, 0.32, -2.68, 1.32, 1.32, -0.68, 2.32, 2.32, -1.68, -2.68, etc. We see that there are positive and negative errors. If we get many errors of high magnitude, no matter whether they are positive or negative, the graph in Fig. 4.3 will become significantly stretched. In the opposite direction, if all the errors are of very small magnitude the whole error distribution will be concentrated around zero. Now, one way to focus on the contribution of each error to the larger or smaller

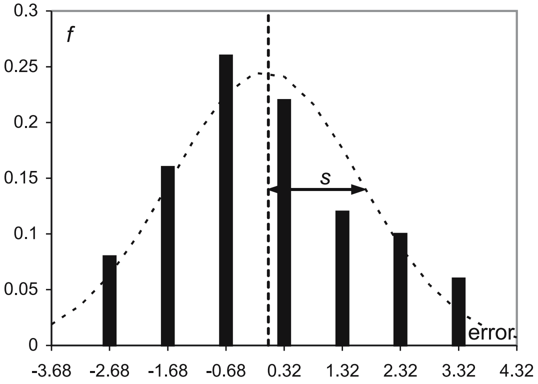


Fig. 4.3. Error histogram when weighing a necklace

stretching of the graph but regardless of its sign (positive or negative) consists in taking its square. So when we add the squares of all errors and divide by their number, we obtain a measure of the *spread* of the error distribution. The mean square error is precisely such a measure of the spread:

$$\nu = \frac{1.68^2 + 0.68^2 + 0.32^2 + 2.68^2 + 1.32^2}{50} = 2.58 \text{ g}^2 \ .$$

The mean square error ν, is also called the *variance*. The square root of the variance, viz., $s = \sqrt{\nu}$, is called the *standard deviation*. In Fig. 4.3, the value of the standard deviation ($\sqrt{2.58} = 1.61$ g) is indicated by the double-ended arrow.

The dotted line in Fig. 4.3 also shows the graph of the function $f(x)$ found by Laplace and Gauss, and computed with $\nu = 2.58$ g^2. We see that the function seems to provide a reasonable approximation to the histogram values. As we increase the number of measurement repetitions the function gradually fits better until values of bars are practically indistinguishable from values taken from the dotted line. This bell-shaped curve often appears when we study chance phenomena. Gauss called it the *normal distribution* curve, because error distributions *normally* follow such a law.

4.3 A Continuum of Chances

Let us picture a rifle shooting at a target like the one shown in Fig. 4.4. The rifle is assumed to be at a perfectly fixed position, aiming at the target center. Despite the fixed rifle position, the position of the bullet impact is random, depending on chance phenomena such as the wind, rifle vibrations, barrel heating, barrel grooves wearing out, and so on.

If we use the frequentist interpretation of probability, we may easily obtain probability estimates of hitting a certain target area. For instance, if out of 5000 shots hitting the target 2000 fall in the innermost circle, an estimate of P(shot in innermost circle) is 2/5. We may apply the same procedure to the outermost circle; suppose we obtained 3/5. The estimated probability of a hit in the circular ring is then 1/5 and of a hit outside the outermost circle is $1 - 3/5 = 2/5$. We are assuming that by construction there can be no shots falling outside the target.

We now proceed to consider a small circular area of the target (in any arbitrarily chosen position) for which we determine by the above procedure the probability of a hit. Suppose we go on decreasing the

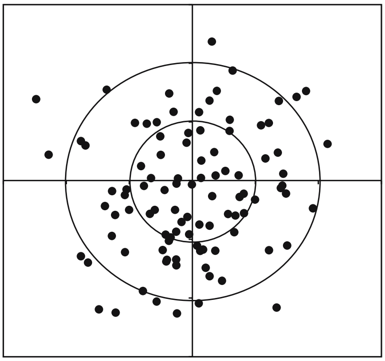


Fig. 4.4. Shooting at a target

circle radius, therefore shrinking the circular region. The probability of a hit will also decrease in a way approximately proportional to the area of the circle when the area is sufficiently small and the shots are uniformly spread out over the circle. In other words, there comes a time when P(shot in the circle) is approximately given by $k \times a$, where a is the area of the circle and k is a proportionality factor. In order to obtain arbitrarily good estimates of P(shot in the circle) for a certain value of a, we would of course need a large number of shots on the target, and all the more as a gets smaller.

We thus draw two conclusions from our imaginary experiment of shooting at a target: the hit probability in a certain region *continually* decreases as the region shrinks; for sufficiently small regions, the probability is practically proportional to the area of the region. It is as if the total probability, 1, of hitting the target was finely smeared out over the whole target. In a very small region we can speak of *probability density per unit area*, and our previous k factor is

$$k = \frac{P(\text{shot in region})}{a} \; .$$

When the region shrinks right down to a point, the probability of a hit is zero because the area is zero:

$$P(\text{hit a point}) = 0 \qquad (\text{always, in continuous domains})\,.$$

We are now going to imagine that the whole rifle contraption was mounted in such a way that the shots could only hit certain pre-established points in the target and nothing else. If we randomly select a small region and proceed to shrink it, we will almost always obtain *discontinuous* variations of P(shot in region), depending on whether the region includes the pre-established points or not. In this case the probability of hitting a point is not zero for the pre-established points. The total probability is not continuously smeared out on the target; it is discretely divided among the pre-established points. The domain of possible impact points is no longer continuous but discrete. We have in fact the situation of point-like probabilities of discrete events (the pre-established points) that we have already considered in the preceding chapters.

Let us now go back to the issue of P(hit a point) equalling zero for continuous domains. Does this mean that the point-bullet never hits a specified point? Surely not! But it does it a finite number of times when the number of shots is infinite, so that, according to the frequentist approach, the value of P(hit a point) is 0. On the other hand, for discrete domains (shots at pre-established points), hitting a specific point would also happen an infinite number of times.

Operating with probabilities in a continuum of chances demands some extension of the notions we have already learned (not to mention some technical difficulties that we will not discuss). In order to do that, let us consider a similar situation to shooting at a target, but rather simpler, with the 'shot' on a straight-line segment. Once again the probability of hitting a given point is zero. However, the probability of the shot falling into a given straight line interval is not zero. We may divide the interval into smaller intervals and still obtain non-zero probabilities. It is as if our degree of certainty that the shot falls in a given interval were finely smeared out along the line segment, resulting in a probability density per unit length. Thinking of the probability measure as a mass, there would exist a mass density along the line segment (and along any segment interval). Figure 4.5a illustrates this situation: a line segment, between 0 and 1, has a certain probability density of a hit. Many of the bullets shot by the rifle fall near the center (dark gray); as we move towards the segment ends, the hit probability density decreases. The figure also shows the *probability density function*, abbreviated to p.d.f. We see how the probability density is large at the center and drops towards the segment ends. If we wish to determine the probability of a shot hitting the interval from 0.6 to 0.8, we have only to determine the hatched area of Fig. 4.5a. (In the mass analogy, the hatched area is the mass of the interval.) In Fig. 4.5b, the probability density is constant. No interval is favored with more hits than any other equal-length interval. This is the so-called *uniform* distribution. Uniformity is the equivalent for continuous domains of equiprobability (equal likelihood) for discrete domains.

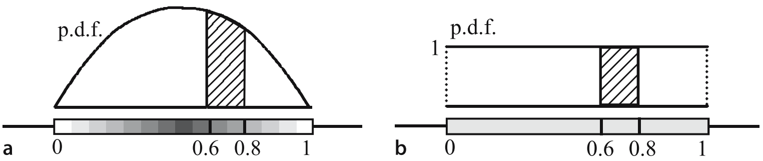


Fig. 4.5. Two cases of the probability density of shots on [0, 1]. (**a**) The density is larger at the center. (**b**) The density is uniform

In the initial situation of shooting at a target, the p.d.f. is a certain surface, high at the center and decaying towards the periphery. The probability in a certain region corresponds to the volume enclosed by the surface on that region. In the two situations, either the total area (Fig. 4.5) or the total volume (Fig. 4.4) represents the probability of the sure event, that is, they are equal to 1. Previously, in order to determine the probability of an event supported by a discrete domain, we added probabilities of elementary events. Now, in continuous domains, we substitute sums by areas and volumes. With the exception of this modification, we can continue to use the basic rules of Chap. 1.

The normal probability function presented in the last section is in fact a probability density function. In the measurement error example, the respective values do not constitute a continuum; it is easy to see that the values are discrete (for instance, there are no error values between 0.32 g and 0.68 g). Moreover, the error set is not infinite. Here as in other situations one must bear in mind that a continuous distribution is merely a mathematical model of reality.

4.4 Buffon’s Needle

Georges-Louis Leclerc, Count of Buffon (1707–1788), is known for his remarkable contributions to zoology, botany, geology and in particular for a range of bold ideas, many ahead of his time. For instance, he argued that natural causes formed the planets and that life on Earth was the consequence of the formation of organic essences resulting from the action of heat on bituminous substances. Less well known is his important contribution to mathematics, and particularly to probability theory.

Buffon can be considered a pioneer in the experimental study and application of probability theory, a specific example of which is the famous Buffon’s needle. It can be described as follows. Consider randomly throwing a needle of length l on a floor with parallel lines separated by

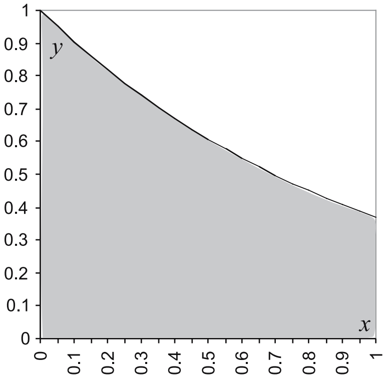


Fig. 4.6. Area under the curve $y = \mathrm{e}^{-x}$

$d \geq l$ (think, for instance, of the plank separations of a wooden floor). What is the probability that the needle falls over a line?

We are dealing here with a problem of continuous probabilities suggested to Buffon by a fashionable game of his time known as *franc-carreau*. The idea was to bet that a coin tossed over a tile floor would fall completely inside a tile. For the needle problem, Buffon demonstrated that the required probability is given by

$$P = \frac{2l}{\pi\lambda} .$$

For $d = l$, one then obtains $2/\pi$, i.e., Buffon's needle experiment offers a way of computing π ! The interested reader may find several Internet sites with simulations of Buffon's needle experiment and estimates of π using this method.

4.5 Monte Carlo Methods

It seems that Buffon did indeed perform long series of needle throws on a floor in order to verify experimentally the result that he had concluded mathematically. Nowadays, we hardly would engage in such labor, since we dispose of a superb tool for carrying out any experiments relating to chance phenomena: the computer.

The Monte Carlo method – due to the mathematician Stanislaw Ulam (1909–1984) – refers to any problem-solving method based on the use of random numbers, usually with computers.

Consider the problem of computing the shaded area of Fig. 4.6 circumscribed by a unit square. In this example, the formula of the curve delimiting the area of interest is assumed to be known, in this case $y = \mathrm{e}^{-x}$. This allows us, by mathematical operations, to obtain the value of the area as $1 - 1/\mathrm{e}$, and hence the value of e. However, it often happens that the curve formula is such that one cannot directly determine the value of the area, or the formula may even be unknown. In such cases one has to resort to numerical methods. A well-known method for computing area is to divide the horizontal axis into sufficiently small intervals and add up the vertical 'slices' of the relevant region, approximating each slice by a trapezium. Such computation methods are *deterministic*: they use a well defined rule.

In certain situations it may be more attractive (in particular, with regard to computation time) to use another approach that is said to be *stochastic* (from the Greek 'stokhastikós' for a phenomenon or method whose individual cases depend on chance) because it uses probabilities. Let us investigate the stochastic or Monte Carlo method for computing area. For this purpose we imagine that we randomly throw points in a haphazard way onto the domain containing the interesting area (the square domain of Fig. 4.6). We also assume that the randomly thrown points are uniformly distributed in the respective domain; that is, for a sufficiently large number of points any equal area subdomain has an equal probability of being hit by a point. Figure 4.7 shows sets of points obtained with a uniform probability density function, thrown onto a unit square. Let d be the number of points falling inside the relevant region and n the total number of points thrown. Then, for sufficiently large n, the ratio d/n is an estimate of the relevant area with given accuracy (according to the law of large numbers).

This method is easy to implement using a computer. We only need a *random number generator*, that is, a function that supplies num-

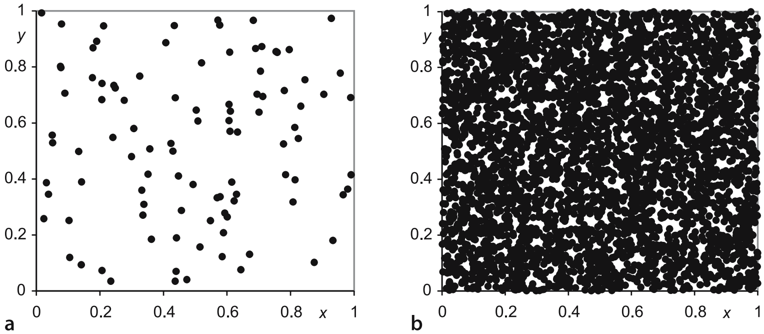


Fig. 4.7. Examples of uniformly distributed points. (**a**) 100 points. (**b**) 4000 points

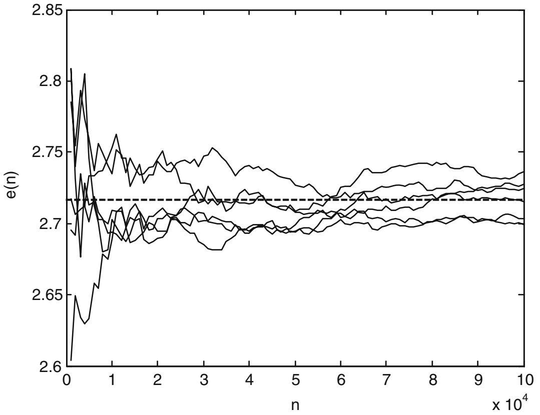


Fig. 4.8. Six Monte Carlo trials for the estimation of e

bers as if they were generated by a random process. (One could also use roulette-generated numbers, whence the name Monte Carlo.) Many programming languages provide random (in fact, pseudo-random) number generating functions. The randomness of the generated numbers (the quality of the generator) has an influence on the results of Monte Carlo methods.

Figure 4.8 shows the dependence $e(n)$ of the estimate e on n in six Monte Carlo trials corresponding to Fig. 4.6. By inspection of Fig. 4.8, one sees that the law of large numbers is also at work here. The estimate tends to stabilize around the value of e (dotted line) for sufficiently large n.

Monte Carlo methods as described here are currently used in many scientific and technical areas: aerodynamic studies, forecasts involving a large number of variables (meteorology, economics, etc.), optimization problems, special effects in movies, video games, to name but a few.

4.6 Normal and Binomial Distributions

If the reader compares the formula approximating the binomial distribution, found by Abraham de Moivre, and the probability density function of the normal distribution, found by Laplace and Gauss, (s)he will immediately notice their similarity. As a matter of fact, for large n, the binomial distribution is well approximated by the normal one, if one uses

$$x = k - m \,, \qquad m = np \,, \qquad \nu = np(1-p) \,.$$

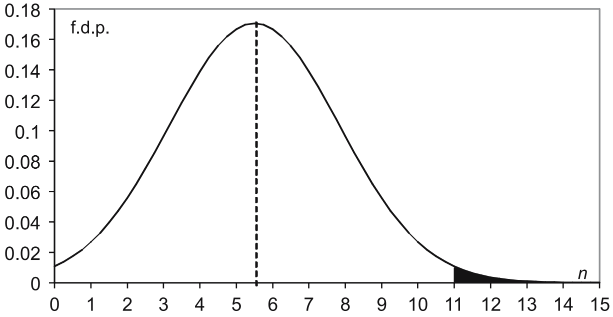


Fig. 4.9. The value of P(more than 10 events in 1000) (*shaded area*) using the normal approximation

The mathematical expectation (mean) of the normal distribution is the same as that of the binomial distribution: $m = np$. The variance is $np(1-p)$, whence the standard deviation is $s = \sqrt{np(1-p)}$.

We can now go back to the initial problem of this chapter. For $n = 1000$ the approximation of the normal law to the binomial law is excellent (it is usually considered good for most applications when n is larger than 20 and the product np is not too low, conditions satisfied in our example). Let us then consider the normal law using

$$m = np = 1000 \times 0.0055 = 5.5 \ ,$$

$$\nu = np(1-p) = 5.5 \times 0.9945 = 5.47 \ ,$$

whence $s = 2.34$.

We then have the value of P(more than 10 events in 1000) corresponding to the shaded area of Fig. 4.9. There are tables and computer programs that supply values of the normal p.d.f. areas.[1] In this way, one can obtain the value of the area as 0.01. The probability of finding more than 10 infected people among the 1000 hospital patients is therefore 1%.

4.7 Distribution of the Arithmetic Mean

When talking about the measurement error problem we mentioned the arithmetic mean as a better estimate of the desired measurement. We are now in a position to analyze this matter in greater detail.

Consider a sequence of n values, obtained independently of each other and originating from identical random phenomena. Such is the

[1] Assuming zero mean and unit variance. Simple manipulations are required to find the area for other values.

usual measurement scenario: no measurement depends on the others and all of them are under the influence of the same random error sources. An important result, independently discovered by Laplace and Gauss, is the following: if each measurement follows the normal distribution with mean (mathematical expectation) m and standard deviation s, then, under the above conditions, the arithmetic mean (the computed average) random variable also follows the normal distribution with the same mean (mathematical expectation) and with standard deviation given by $s/\sqrt{n}$.

Let us assume that in the necklace weighing example some divinity had informed us that the true necklace weight was 45 g and the true standard deviation of the measurement process was 1.5 g. In Fig. 4.10, the solid line shows the normal p.d.f. relative to one measurement while the dotted line shows the p.d.f. of the arithmetic mean of 10 measurements. Both have the same mathematical expectation of 45 g. However, the standard deviation of the arithmetic mean is $s/\sqrt{n} = 1.5/\sqrt{10} = 0.47$ g, about three times smaller; the p.d.f. of the arithmetic mean is therefore about three times more concentrated, as illustrated in Fig. 4.10. As we increase the number of measurements the distribution of the arithmetic mean gets gradually more concentrated, forcing it to converge rapidly (in probability) to the mathematical expectation. We may then interpret the mathematical expectation as the 'true value of the arithmetic mean' ('the true weight value for the necklace') that would be obtained had we performed an infinite number of measurements.

Suppose we wish to determine an interval of weights around the true weight value into which one measurement would fall with 95% probability. It so happens that, for a normal p.d.f., such an interval starts (to a good approximation) at the distribution mean minus two

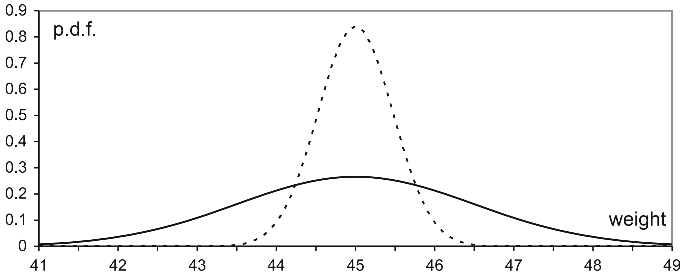


Fig. 4.10. Weighing a necklace: p.d.f. of the error in a single measurement (*solid curve*) and for the average of 10 measurements (*dotted curve*)

standard deviations and ends at the mean plus two standard deviations. Thus, for the necklace example, the interval is 45 ± 3 g ($3 = 2 \times 1.5$). If we take the average of 10 measurements, the interval is then 45 ± 0.94 g ($0.94 = 2 \times 0.47$).

In conclusion, whereas for one measurement one has a 95% degree of certainty that it will fall in the interval 45 ± 3 g, for the average of 10 measurements and the same degree of certainty, the interval is reduced to 45 ± 0.94 g. Therefore, the weight estimate is clearly better: in the first case 95 out of 100 measurements will produce results that may be 3 g away from the true value; in the second case 95 out of 100 averages of 10 measurements will produce results that are at most about 1 gram away. Computing averages of repeated measurements is clearly an advantageous method in order to obtain better estimates, and for this reason it is widely used in practice.

The difficulty with the above calculations is that no divinity tells us the true values of the mean and standard deviation. (If it told us, we would not have to worry about this problem at all.) Supposing we computed the arithmetic mean and the standard deviation of 10 measurements many times, we would get a situation like the one illustrated in Fig. 4.11, which shows, for every 10-measurement set, the arithmetic mean (solid circle) plus or minus two standard deviations (double-ended segment) computed from the respective mean square errors. The figure shows 15 sets of 10 measurements, with trials numbered 1 through 15. The arithmetic means and standard deviations vary for each 10-measurement set; it can be shown, however, that in a large number of trials nearly 95% of the intervals 'arithmetic mean plus or minus two standard deviations' contain the true mean value (the value that only the gods know). In Fig. 4.11, 14 out of the 15 segments, i.e., 93% of the segments, intersect the line corresponding to the true mean value.

What can then be said when computing the arithmetic mean and standard deviation of a set of measurements? For the necklace example one computes for the fifty measurements:

$$m = 45.68 \text{ g} , \qquad s = 1.61 \text{ g} .$$

The standard deviation of the arithmetic mean is therefore $1.61/\sqrt{50} = 0.23$ g. The interval corresponding to two standard deviations is thus 45.68 ± 0.46 g. This interval is usually called the 95% *confidence interval* of the arithmetic mean, and indeed 95% certainty is a popular value for confidence intervals.

What does it mean to say that the 95% confidence interval of the necklace's measured weight is 45.68 ± 0.46 g? Does it mean that we are

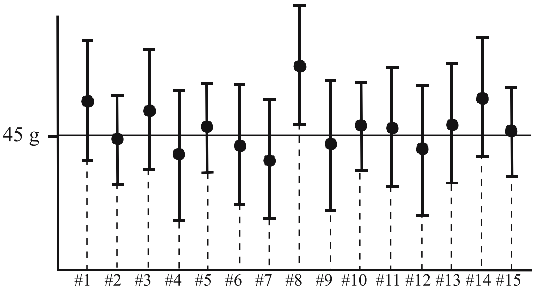


Fig. 4.11. Computed mean ±2 standard deviations in several trials (numbered 1 to 15) of 10 measurements

95% sure that the true value is inside that interval? This is an oft-stated but totally incorrect interpretation. In fact, if the true weight were 45 g and the true standard deviation were 3 g, we have already seen that, with 95% certainty, the average of the 50 measurements would fall in the interval 45 ± 0.94 g, and not in 45.68 ± 0.46 g. Whatever the true values of the mean and standard deviation may be, there will always exist 50-measurement sets with confidence intervals distinct from the confidence interval of the true mean value. The interpretation is quite different and much more subtle.

Figure 4.11 can help us to work out the correct interpretation. When computing the 95% confidence interval, we are computing an interval that will not contain the true mean value only 5% of the time. We may have had such bad luck that the 50-measurement set obtained is like set 8 in Fig. 4.11, a set that does not contain the true mean value. For this set there is 100% certainty that it does not contain the true mean (the true weight). We may say that it is an atypical measurement set. The interpretation to be given to the confidence interval is thus as follows: when proposing the 95% confidence interval for the true mean, we are only running a 5% risk of using an atypical interval that does not contain the true mean.

4.8 The Law of Large Numbers Revisited

There is one issue concerning the convergence of means to mathematical expectations that has still not been made clear enough: when do we consider a sequence of experiments to be sufficiently large to ensure an acceptable estimate of a mathematical expectation?

The notion of confidence interval is going to supply us with an answer. Consider the necklace-weighing example. What is an ‘acceptable estimate’ of the necklace weight? It all depends on the usefulness of our estimate. If we want to select pearl necklaces, maybe we can tolerate a deviation of 2 g between our estimate and the true weight and maybe one single weighing is enough. If the necklaces are made of diamonds we will certainly require a smaller tolerance of deviations, say 0.05 g (1/4 of carat). We may then ask: how many measurements of a diamond necklace must we perform to ensure that the 95% confidence interval of the measurements does not deviate by more than 0.05 g? The answer is now easy to obtain. Assuming a 0.1 g standard deviation, we must have $2 \times 0.1/\sqrt{n} = 0.05$, corresponding to $n = 16$ measurements.

Let us now consider the problem of estimating the probability of an event. Consider, for instance, a die we suspect has been tampered with, because face 6 seems to turn up more often than would be expected. To clear this matter up we wish to determine how many times we must fairly throw the die in order to obtain an acceptable estimate of face 6 turning up with 95% confidence. We assume that an acceptable estimate is one that differs from the true probability value by no more than 1%. Suppose we throw the die n times and face 6 turns up k times. We know that k follows the binomial distribution, which for large n is well approximated by the normal distribution with mean np and variance $np(1-p)$, where p is the probability we wish to estimate. The variable we use to estimate p is the frequency of occurrence of face 6: $f = k/n$. Now, it is possible to show that, when k is well approximated by the normal distribution, then so is f, and with mean p and standard deviation $\sqrt{p(1-p)}/\sqrt{n}$. (Note once again the division by $\sqrt{n}$.)

We thus arrive at a vicious circle: in order to estimate p, we need to know the standard deviation which depends on p. The easiest way out of this vicious circle is to place ourselves in a worst case situation of the standard deviation. Given its formula, it is easy to check that the worst case (largest spread) occurs for $p = 0.5$. Since we saw that a 95% confidence level corresponds to two standard deviations we have

$$2 \times \frac{\sqrt{0.25}}{\sqrt{n}} = \frac{1}{\sqrt{n}} \,.$$

We just have to equate this simple $1/\sqrt{n}$ formula to the required 1% tolerance: $1/\sqrt{n} = 0.01$, implying $n = 10\,000$. Hence, only after 10 000 throws do we attain a risk of no more than 5% that the confidence interval $f \pm 0.01$ is atypical, i.e., does not contain the true p value.

4.9 Surveys and Polls

Everyday we see reports of surveys and polls in the newspapers and media. Consider, for instance, the following report of an investigation carried out amongst the population of Portugal:

> Having interviewed 880 people by phone, 125 answered ‘yes’, while the rest answered ‘no’. The ‘yes’ percentage is, therefore, 14.2% with an error of 3.4% for a confidence degree of 95%.

The vast majority of people have trouble understanding what this means. Does it mean that the ‘yes’ percentage is 14.2%, but 3.4% of the people interviewed filled in the survey forms in the wrong way or gave ‘incorrect’ answers to the interviewers? Is it the case that only $14.2 - 3.4\%$, that is, 10.8% of those interviewed deserve 95% confidence (i.e., will not change their opinion)? Maybe it means that only 3.4% of the Portuguese population will change the 14.2% percentage? Or does it perhaps mean that we are 95% sure (however this may be interpreted) that the ‘yes’ vote of the Portuguese population will fall in the interval from 10.8% to 17.6%? Indeed, many and varied interpretations seem to be to hand, so let us analyze the meaning of the survey report.

First of all, everything hinges on estimating a certain variable: the proportion of Portuguese that vote ‘yes’ (with regard to some question). That estimate is based on an analysis of the answers from a certain set of 880 people, selected for that purpose. One usually refers to a set of cases from a population, studied in order to infer conclusions applicable to the whole population, as a *sample*. In the previous section we estimated the probability of turning up a 6 with a die by analyzing the results of 10 000 throws. Instead of throwing the same die 10 000 times, we could consider 10 000 people throwing a die, as far as we can guarantee an identical throwing method and equivalent dice. In other words, it makes no difference whether we consider performing experiments in time or across a set of people.

If the throws are made under the same conditions, we may of course, imagine robots performing the throws instead of people. But one difficulty with surveys of this kind is that people are not robots. They are unequal with regard to knowledge, standard of living, education, geographical position, and many other factors. We must therefore assume that these are sources of uncertainty and that the results we hope to obtain *reflect* all these sources of uncertainty. For instance, if the survey only covers illiterate people, its result will not reflect the source of uncertainty we refer to as ‘level of education’. Any conclusion drawn from the survey would only apply to the illiterate part of the Portuguese population. This is the *sampling problem* which arises when one must select a sample.

In order to reflect what is going on in the Portuguese population, the people in our sample must reflect all possible *nuances* in the sources of uncertainty. One way to achieve this is to select the people *randomly*. Suppose we possess information concerning the 7 million identity cards of the Portuguese adult population. We may number the identity cards from 1 to 7 000 000. We now imagine seven urns, 6 with 10 balls numbered 0 through 9 and the seventh with 8 balls numbered 0 through 7, corresponding to the millionth digit. Random extraction with replacement from the 7 urns would provide a number selecting an identity card and a person. For obvious practical reasons, people are not selected in this way when carrying out such a survey. It is more usual to use other 'almost random' selection rules; for instance, by random choice of phone numbers. The penalty is that in this case the survey results do not apply to the whole Portuguese population but only to the population that has a telephone (the source of uncertainty referred to as 'standard of living' has thus been reduced).

In second place, the so-called 'error' is not an error. It is half of the width of the estimated confidence interval: $1/\sqrt{n} = 1/\sqrt{880} = 0.0337$. We are not therefore 95% sure that the probability of a 'yes' falls into the interval from 10.8% to 17.6%, in the sense that if we repeated the survey many times, 95% of the time the percentage estimate of 'yes' answers would fall in that interval. The interpretation is completely different, as we have already seen in the last section. We are only 95% sure that our confidence interval is a typical interval (one of the many, 95%, of the intervals that in many random repetitions of the survey would contain the true value of the 'yes' proportion of voters). We are not rid of the fact that, by bad luck (occurring 5% of the time in many repetitions), an atypical set of 880 people may have been surveyed, for which the confidence interval does not contain the true value of the proportion.

Suppose now that in an electoral opinion poll, performed under the same conditions as in the preceding survey, the following voting percentages are obtained for political parties A and B, respectively: 40.5% and 33.8%. Since the difference between these two values is well above the 3.4% 'error', one often hears people drawing the conclusion that it is almost certain (95% certain) that party A will win over party B. But is this necessarily so? Assuming typical confidence intervals, which therefore intersect the true values of the respective percentages, it may happen that the true percentage voting A is $40.5 - 3.4 = 37.1\%$, whereas the true percentage voting B is $33.8 + 3.4 = 37.2\%$. In other words, it may happen that votes for B will outnumber votes for A! In conclusion, we will only be able to discriminate the voting superiority of A over B if the respective estimates are set apart by more than the width of the confidence interval (6.8% in this case). If we want a tighter discrimination, say 2%, the 'error' will then have to be 1%, that is, $1/\sqrt{n} = 0.01$, implying a much bigger sample: $n = 10\,000$!

4.10 The Ubiquity of Normality

Many random phenomena follow the normal distribution law, namely those that can be considered as emerging from the addition of many causes acting independently from one another; even when such causes have non-normal distributions. Why does this happen? The answer to this question has to do with a famous result, known as the *central limit theorem*, whose first version was presented by Gauss. Afterwards the theorem was generalized in several ways.

There is a geometrical construction that can quickly convince us of this result. First, we have to know how to add two random variables. Take a very simple variable, with only two values, 0 and 1, exactly as if we tossed a coin in the air. Let us denote this variable by x_1 and let us assume that the probabilities are

$$P(x_1 = 0) = 0.6\ , \qquad P(x_1 = 1) = 0.4\ .$$

Supposing we add to x_1 an identical and independent variable, denoted x_2, what would be the probability function of $x_1 + x_2$? The possible values of the sum are 0, 1 and 2. The value 0 only occurs when the value of both variables is 0. Thus, as the variables are independent, the corresponding probability of $x_1 + x_2 = 0$ is $0.6 \times 0.6 = 0.36$. The value 1 occurs when $x_1 = 0$ and $x_2 = 1$ or, in the opposite way, $x_1 = 1$ and $x_2 = 0$; thus, $P(x_1 + x_2 = 1) = 0.6 \times 0.4 + 0.4 \times 0.6 = 0.48$. Finally, $x_1 + x_2 = 2$ only occurs when the value of both variables is 1; thus, $P(x_1 + x_2 = 2) = 0.16$. One thus obtains the probability function of Fig. 4.12b.

Let us now see how we can obtain $P(x_1 + x_2)$ with a geometric construction. For this purpose we imagine that the $P(x_1)$ graph is fixed, while the $P(x_2)$ graph, *reflected* as in a mirror, 'walks' in front of the $P(x_1)$ graph in the direction of increasing x_1 values. Along the 'walk'

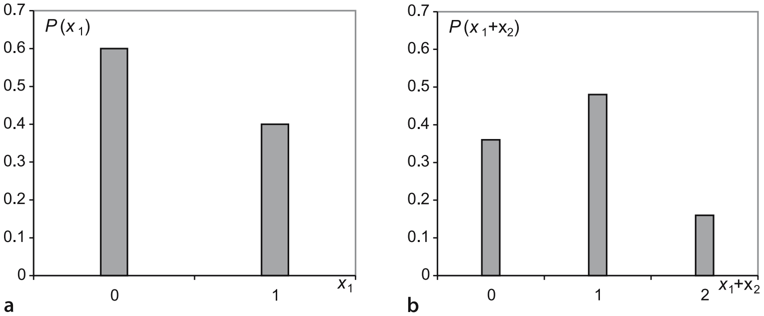


Fig. 4.12. Probability function of (**a**) a random variable with only two values and (**b**) the sum of two random variables with the same probability function

we add the products of the overlapping values of $P(x_1)$ and the reflected $P(x_2)$. Figure 4.13 shows, from left to right, what happens. When the $P(x_2)$ reflected graph, shown with white bars in Fig. 4.13, lags behind the $P(x_1)$ graph, there is no overlap and the product is zero. In the left-most graph of Fig. 4.13 we reached a situation where, along the 'walk', the $P(x_2)$ reflected graph overlaps the $P(x_1)$ graph at the point 0; one thus obtains the product 0.6×0.6, which is the value of $P(x_1 + x_2)$ at point 0. The central graph of Fig. 4.13 corresponds to a more advanced phase of the 'walk'. There are now two overlapping values whose sum of products is $P(x_1+x_2 = 1) = 0.4\times0.6+0.6\times0.4$. In the rightmost graph of Fig. 4.13, the $P(x_2)$ reflected graph only overlaps at point 2, and one obtains, by the same process, the value of $P(x_1 + x_2 = 2) = 0.4 \times 0.4$. From this point on there is no overlap and $P(x_1 + x_2)$ is zero. The operation we have just described is called *convolution*. If the random variables are continuous, the convolution is performed with the probability density functions (one of them is reflected), computing the area below the graph of the product in the overlap region.

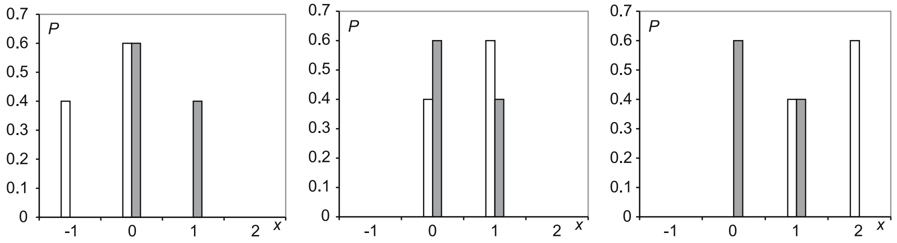


Fig. 4.13. The geometric construction leading to Fig. 4.12b

But what has this to do with the central limit theorem? As unbelievable as it may look, if we repeat the convolution with *any* probability functions or *any* probability density functions (with rare exceptions), we will get closer and closer to the Gauss function!

Figure 4.14 shows the result of adding variables as in the above example, for several values of the number of variables. The scale on the horizontal axis runs from 0 to 1, with 1 corresponding to the maximum value of the sum. Thus, for instance in the case $n = 3$, the value 1 of the horizontal axis corresponds to the situation where the value of the three variables is 1 and their sum is 3. The circles signaling the probability values are therefore placed on $0 \times (1/3)$, $1 \times (1/3)$, $2 \times (1/3)$, and $3 \times (1/3)$. The construction is the same for other values of n. Note that in this case (variables with the same distribution), the mean of the sum is the mean of any distribution multiplied by the number of added variables: $3 \times 0.6 = 1.8$, $5 \times 0.6 = 3$, $10 \times 0.6 = 6$, $100 \times 0.6 = 60$. In the graphs of Fig. 4.14, these means are placed at 0.6, given the division by n (standardization to the interval 0 through 1).

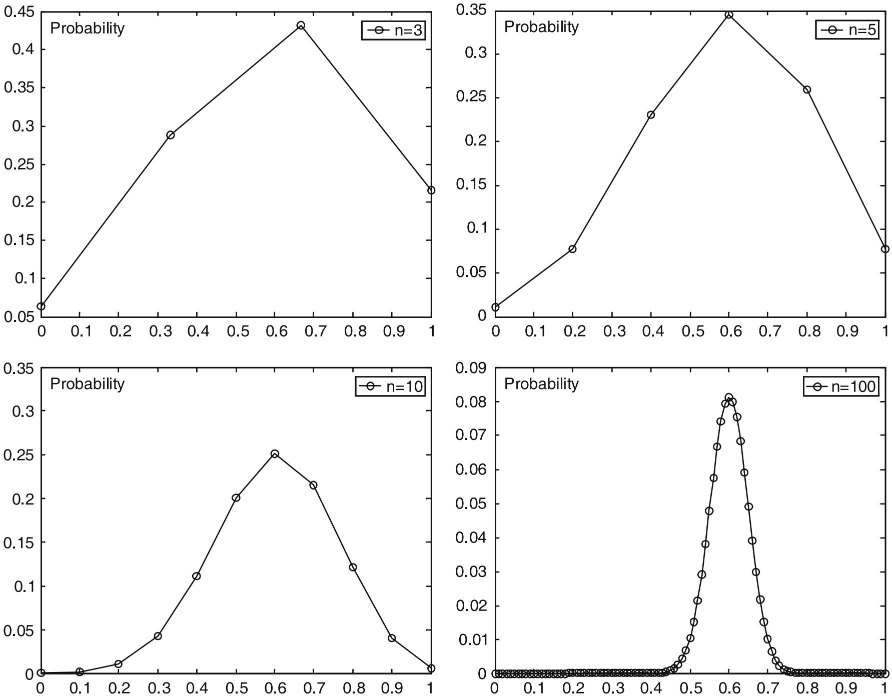


Fig. 4.14. Probability functions (p.f.) of the sum of random variables (all with the p.f. of Fig. 4.12a). Note how the p.f. of the sum gets closer to the Gauss curve for increasing n

Fig. 4.15. The normal distribution on parade

In conclusion, it is the convolution law that makes the normal distribution ubiquitous in nature. In particular, when measurement errors depend on several sources of uncertainty (as we saw in the section on measurement errors), even if each one does not follow the normal law, the sum of all of them is in general well approximated by the normal distribution. Besides measurement errors, here are some other examples of data whose distribution is well approximated by the normal law: heights of people of a certain age group; diameters of sand grains on the beach; fruit sizes on identical trees; hair diameters for a certain person; defect areas on surfaces; chemical concentrations of certain wine components; prices of certain shares on the stock market; and heart beat variations.

4.11 A False Recipe for Winning the Lotto

In Chap. 1 we computed the probability of winning the first prize in the lotto as 1/13 983 816. Is there any possibility of devising a scheme to change this value? Intuition tells us that there cannot be, and rightly so, but over and over again there emerge 'recipes' with differing degrees of sophistication promising to raise the probability of winning. Let us describe one of them. Consider the sum of the 6 randomly selected numbers (out of 49). For reasons already discussed, the distribution of this sum will closely follow the normal distribution around the mean value $6 \times \left[(1+49)/2\right] = 150$ (see Fig. 4.16). It then looks a good idea to choose the six numbers in such a way that their sum falls into a certain interval around the mean or equal to the mean itself, i.e., 150. Does this work?

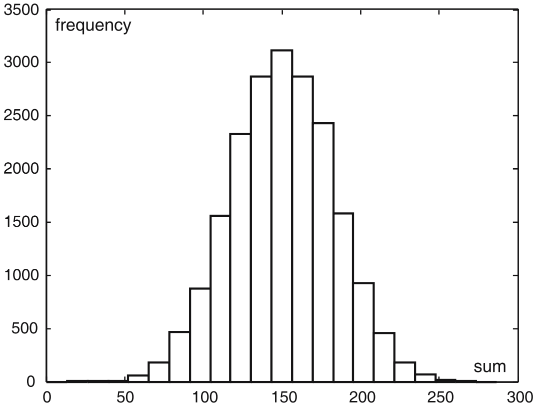


Fig. 4.16. Histogram of the sum of the lotto numbers in 20 000 extractions

First of all, let us note that it is precisely around the mean that the number of combinations of numbers with the same sum is huge. For instance, suppose we chose a combination of lotto numbers with sum precisely 150. There are 165 772 possible combinations of six numbers between 1 and 49 whose sum is 150. But this is a purely technical detail. Let us proceed to an analysis of the probabilities by considering the following events:

A The sum of the numbers of the winning set is 150.

B The sum of the numbers of the reader's bet is 150.

C The reader wins the first prize (no matter whether or not the sum of the numbers is 150).

Let us suppose the reader chooses numbers whose sum is 150. The probability of winning is then

$$\begin{aligned}P(\mathrm{C\ and\ A}) &= P(\mathrm{C\ if\ A}) \times P(\mathrm{A}) \\ &= \frac{1}{165\,772} \times \frac{165\,772}{13\,983\,816} = \frac{1}{13\,983\,816}\,.\end{aligned}$$

Therefore, the fact that the reader has chosen particular numbers did not change the probability of the reader winning the first prize. On the other hand, we can also see that

$$\begin{aligned}P(\mathrm{C\ if\ B}) &= \frac{P(\mathrm{B\ and\ C})}{P(B)} = \frac{P(\mathrm{B\ if\ C}) \times P(C)}{P(B)} \\ &= \frac{(165\,772/13\,983\,816) \times (1/13\,983\,816)}{165\,772/13\,983\,816} = \frac{1}{13\,983\,816}\,.\end{aligned}$$

Thus $P(\mathrm{C\ if\ B}) = P(C)$, that is, the events ‘the sum of the numbers of the reader’s bet is 150’ and ‘the reader wins the first prize’ are independent. The probability of winning is the same whether or not the sum of the numbers is 150! As a matter of fact, this recipe is nothing else than a sophisticated version of the gambler’s fallacy, using the celebrated normal curve in an attempt to make us believe that we may forecast the outcomes of chance.

5

Probable Inferences

5.1 Observations and Experiments

We are bombarded daily with news and advertisements involving inferences based on sets of observations: healthy children drink yoghurt A; regular use of toothpaste B reduces dental caries; detergent X washes whiter; recent studies leave no doubt that tobacco smoke may cause cancer; a US investigation revealed that exposure to mobile phone radiation causes an increase in cancer victims; pharmaceutical product C promotes a 30% decrease in infected cells. All these statements aim to generalize a conclusion drawn from the analysis of a certain, more or less restricted data set. For instance, when one hears 'healthy children drink yoghurt A', one may presuppose that the conclusion is based on a survey involving a limited number of children. The underlying idea of all these statements is as follows: there are two variables, say X and Y, and one hopes to elucidate a measure of the relationship between them. Measurements of both variables are influenced by chance phenomena. Will it be possible, notwithstanding the influence of chance, to draw some general conclusion about a possible relation between them?

Consider the statement 'healthy children drink yoghurt A'. Variable X is 'healthy children' and variable Y is 'drink yoghurt A'. The values of X can be established by means of clinical examinations of the children and consequent assignment of a rank to their state of health. However, such ranks are not devoid of errors because any clinical analysis always has a non-zero probability of producing an incorrect value. In the same way, variable Y has an obvious random component with regard to the ingested quantities of yoghurt. In summary, X and Y are random variables and we wish to elucidate whether or not there is a relation between them. If the relation exists, it can either be causal, i.e., X causes Y or Y causes X, or otherwise. In the latter case, several situations may occur, which we exemplify as follows for the statement 'healthy children drink yoghurt A':

1. A third variable is involved, e.g., Z = standard of living of the child's parents. Z causes X and Y.
2. There are multiple causes, e.g., Z = standard of living of the child's parents, U = child's dietary regime. Y, Z and U together cause X.
3. There are remote causes in causal chains, e.g., G = happy parents causes 'happy children', which causes 'good immune system in the children'; G = happy parents causes 'joyful children', which causes 'children love reading comics', which causes 'children love comics containing the advertisement for yoghurt A'.

In all the above advertisement examples, with the exception of the last one, the data are usually *observational*, that is, resulting from observations gathered in a certain population. In fact, many statements of the kind found in the media are based on observational data coming from surveys. The influences assignable to situations 1, 2 and 3 cannot be avoided, often for ethical reasons (we cannot compel non-smokers to smoke or expose people with or without cancer to mobile phone radiation).

Regarding the last statement, viz., 'pharmaceutical product C promotes a 30% decrease in infected cells', there is a different situation: it is based on *experimental data*, in which the method for assessing the random variables is subject to a control. In the case of the pharmaceutical industry, it is absolutely essential to establish experimental conditions that allow one either to accept or to reject a causal link. In this way, one takes measures to avoid as far as possible the influence of further variables and multiple causes. It is current practice to compare the results obtained in a group of patients to whom the drug is administered with those of another group of patients identical in all relevant aspects (age, sex, clinical status, etc.) to whom a placebo is administered (control group).

The data gathered in physics experiments are also frequently experimental data, where the conditions underlying the experiment are rigorously controlled in order to avoid the influence of any factors foreign to those one hopes to relate. In practice, it is always difficult to guarantee the absence of perturbing factors, namely those of a remote causality. The great intellectual challenge in scientific work consists precisely in elucidating the interaction between experimental conditions and alternative hypotheses explaining the facts. Statistics makes an important contribution to this issue.

5.2 Statistics

The works of Charles Darwin (1809–1882) on the evolution of species naturally aroused the interest of many zoologists of the day. One of them, the English scientist Walter Weldon (1822–1911), began the painstaking task of measuring the morphological features of animals, seeking to establish conclusions based on those measurements, and thus pioneering the field of biometrics. He once said: "... the questions raised by the Darwinian hypothesis are purely statistical". In his day, 'statistics' meant a set of techniques for representing, describing and summarizing large sets of data. The word seems to have come from the German 'Statistik' (originally from the Latin 'statisticum') or 'stat künde', meaning the gathering and description of state data (censuses, taxes, administrative management, etc.). Coming back to Walter Weldon, the desire to analyze biometrical data led him to use some statistical methods developed by Francis Galton (1822–1911), particularly those based on the concept of correlation, or deviations from the mean. (Galton is also known for his work on linear regression, a method for fitting straight lines to data sets. The idea here is to minimize the squares of the deviations from the straight line. This is the mean square error method, already applied by Abraham de Moivre, Laplace and Gauss.)

The studies of Weldon and Galton were continued by Karl Pearson (1857–1936), professor of applied Mathematics in London, generally considered to be the founder of modern statistics. Statistics is based on probability theory and, besides representational and descriptive techniques, its main aim is to establish general conclusions (inferences) from data sets. In this role it has become an inescapable tool in every area of science, as well as in other areas of study: social, political, entrepreneurial, administrative, etc. The celebrated English writer H.G. Wells (1866–1946) once said: "Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write."

Statistics can be descriptive (e.g., the average salary increased by 2.5% in ..., the main household commodities suffered a 3.1% inflation in ..., child mortality in third world countries has lately decreased although it is still on average above 5%, a typical family produces five kilos of domestic waste daily) or inferential.

We saw in Chap. 4, how an estimate of the mean allows one to establish conclusions (to infer) about the true value of the mean. We then drew the reader's attention to two essential aspects: the sampling process (e.g., was the enquiry on domestic waste sufficiently random to allow conclusions about the typical family?) and the sample size (e.g., was the number of people surveyed sufficiently large in order to extract a conclusion with a high degree of certainty, i.e., 95% or more?). Many misuses of statistics are related to these two essential features. There are, however, many other types of misuse, in particular when what is at stake is to establish a relation involving two or more chance phenomena.

5.3 Chance Relations

Let us consider the height and weight of adults. Figure 5.2 shows the values of these two random variables (in kilos and meters, respectively) measured for a hundred people. We all know that there is some relation between height and weight. In general, taller people are heavier than shorter ones; but there are exceptions, of course. In Fig. 5.2, the average weight and height are marked by broken lines. Let us now observe the two cases lying on the dotted line: both have approximately the same above-average height; however, one of the cases has a weight above the average and the other a weight below the average.

If there were no relation between height and weight, we would get many pairs of cases like the ones lying on the dotted line. That is,

Fig. 5.1. Statistical uncertainty

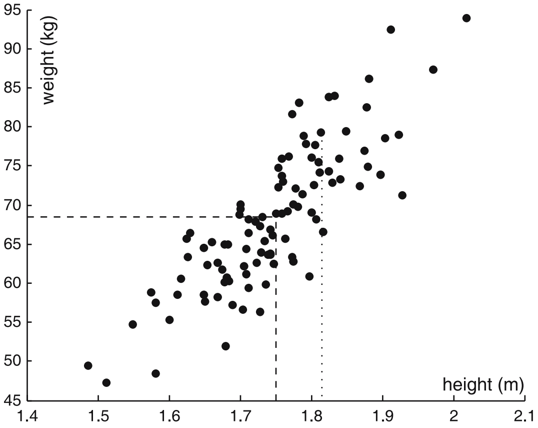


Fig. 5.2. Height and weight of one hundred adults

the simple fact of knowing that someone has an above-average height would not tell us, on average, anything about the person's weight. The situation would then be as pictured in Fig. 5.3: a cloud of points of almost circular shape. But that is not what happens. Figure 5.2 definitely shows us that there is a relationship of some kind between the two variables. Indeed, the points tend to concentrate around an upward direction. If for the hundred people we were to measure the waist size, we would also find a certain relation between this variable and the other two. Relations between two or more chance phenomena are encountered when examining natural phenomena or everyday events. There is than a strong temptation to say that one phenomenon is 'the cause' and the other 'the effect'. Before discussing this issue, let us first get acquainted with the usual way of measuring relationships between two chance phenomena.

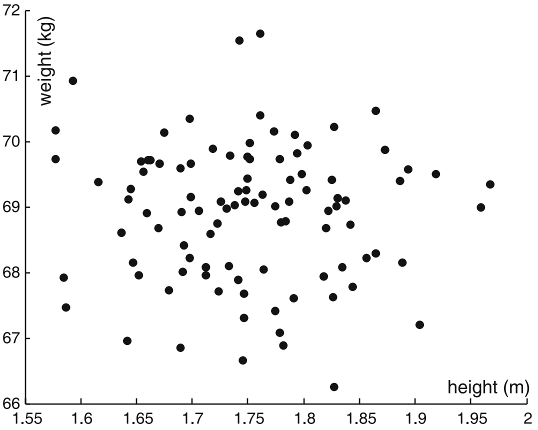


Fig. 5.3. Fictitious height and weight

5.4 Correlation

One of the most popular statistical methods for analyzing the relation between variables is the correlation-based method. We shall now explain this by considering simple experiments in which balls are drawn out of one urn containing three balls numbered 1, 2 and 3. Figure 5.4a shows a two-dimensional histogram for extraction of two balls, with replacement, repeated 15 times. The vertical bars in the histogram show the number of occurrences of each two-ball combination. Let us denote the balls by B1 and B2, for the first and second extraction, respectively. We may assume that this was a game where B1 represented the amount in euros that a casino would have to pay the reader and B2 the amount in euros the reader would have to pay the casino.

In the 15 rounds, represented in Fig. 5.4a, the total number of times that each value from 1 to 3 for B1 occurs, is 6, 5, 4. For B2, it is 6, 4, 5. We may then compute the reader's average gain:

$$\text{average gain (B1)} = 1 \times \frac{6}{15} + 2 \times \frac{5}{15} + 3 \times \frac{3}{15} = \frac{25}{15} = 1.67 \text{ euros}\,.$$

In a similar way one would compute the average loss (B2) to be 1.93 . Let us make use of the idea, alluded to in the last section, of measuring the relation between B1 and B2 through their deviations from the mean. For that purpose, let us multiply together those deviations. For instance, when B1 has value 1 and B2 has value 3, the product of the deviations is

$$(1 - 1.87) \times (3 - 1.93) = -0.93 \;\text{(square euros)}\,.$$

In this case the deviations of B1 and B2 were discordant (one above and the other below the mean); the product is negative. When the values of B1 and B2 are simultaneously above or below the mean, the product is positive; the deviations are concordant. For Fig. 5.4a, we obtain the following sum of the products of the deviations:

$$2 \times (1 - 1.87) \times (1 - 1.93) + \cdots + 2 \times (3 - 1.87) \times (3 - 1.93) = 1.87\,.$$

The average product of the deviations is 1.87/15 = 0.124. If gains and losses were two times greater (2, 4 and 6 euros instead of 1, 2 and 3 euros), the average product of the deviations would be nearly 0.5. This means that, although the histogram in Fig. 5.4a would be similar in both cases, evidencing an equal relationship, we would find a different value for the average product of the deviations. This is clearly not

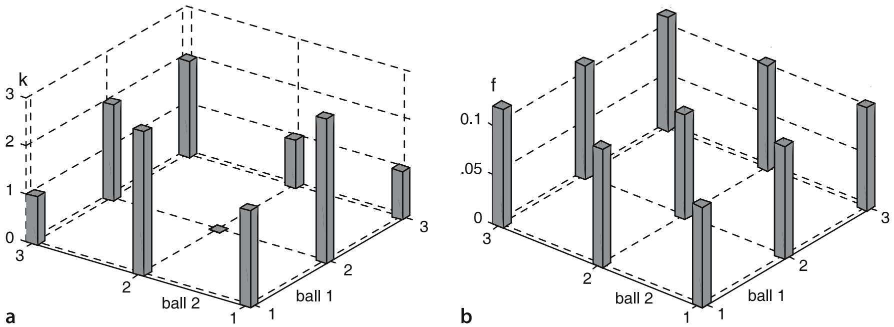


Fig. 5.4. (**a**) Histogram of 15 extractions, with replacement, of two balls from one urn. (**b**) The same, for 2000 extractions

what we were hoping for. We solve this difficulty (making our relation measure independent of the measurement scale used) by dividing the average product of the deviations by the product of the standard deviations. Carrying out the computations of the standard deviations of B1 and B2, we get $s(\mathrm{B1}) = 0.81$ euros, $s(\mathrm{B2}) = 0.85$ euros. Denoting our measure by C, which is now a dimensionless measure (no more square euros), we get

$$\begin{aligned} C(\mathrm{B1}, \mathrm{B2}) &= \frac{\text{average product of deviations of B1 and B2}}{\text{product of standard deviations}} \\ &= \frac{0.124}{0.81 \times 0.85} = 0.18 \,. \end{aligned}$$

The histogram of Fig. 5.4b corresponds to $n = 2000$, with the vertical bars representing *frequencies* of occurrence $f = k/n$, where k denotes the number of occurrences in the $n = 2000$ extractions. We see that the values of f are practically the same. In fact, as we increase the number of extractions n, the number of occurrences of each pair approaches $np = n/9$ (see Sect. 3.6). The measure C, which we have computed/estimated above for a 15-sample case, then approaches the so-called *correlation coefficient* between the two variables. We may obtain (the exact value of) this coefficient, of which the above value is a mere estimate, by calculating the mathematical expectation of the product of the deviations from the true means:

$$E\Big((B1-2)\times(B2-2)\Big) = \frac{1}{9}\Big[(1-2)(1-2)+(3-2)(3-2) + (1-2)(3-2)+(3-2)(1-2)\Big]$$
$$= \frac{1}{9}\times(1+1-1-1)=0\ .$$

Thus, for extraction with replacement of the two balls, the correlation is zero and we say that variables B1 and B2 are uncorrelated. This is also the case in Fig. 5.3 (there is no 'privileged direction' in Figs. 5.3 and 5.4b). Incidentally, note that the variables B1 and B2 are independent, and it so happens that, whenever two variables are independent, they are also uncorrelated (the opposite is not always true).

Let us now consider ball extraction without replacement. For $n = 15$, we obtained in a computer simulation the histogram of occurrences shown in Fig. 5.5a. Repeating the previous computations, the estimate of the correlation between the two variables is -0.59. The exact value is (see Fig. 5.5b)

$$C\Big(\text{B1,B2 (without replacement)}\Big) = \frac{E\Big((B1-2)\times(B2-2)\Big)}{\sqrt{2/3}\times\sqrt{2/3}} = \frac{-1/3}{2/3} = -0.5\ .$$

In this situation the values of the balls are inversely related: if the value 3 (high) comes first, the probability that 1 (low) comes out the following time increases and vice versa. B1 and B2 are negatively correlated.

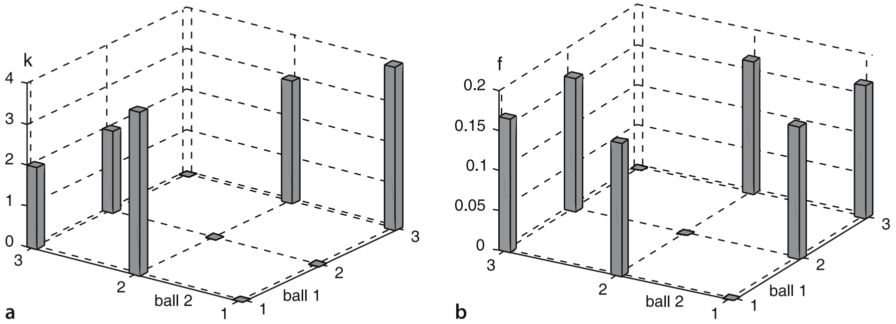


Fig. 5.5. (**a**) Histogram of 15 extractions, without replacement, of two balls from one urn. (**b**) The corresponding probability function

(Note the diagonal with zeros in the histograms of Fig. 5.5, representing impossible combinations.)

It is easily shown that the correlation measure can only take values between -1 and $+1$. We have already seen what a zero correlation corresponds to. The -1 and $+1$ values correspond to a *total relation*, where the data graph is a straight line. In the case of weight and height, when the correlation is $+1$, it would be as though one could tell the exact weight of a person simply by knowing his/her height, and vice versa. A -1 correlation reflects a perfect inverse relation; this would be the case if the second ball extraction from the urn only depended on the first one according to the formula $\mathrm{B2} = 4 - \mathrm{B1}$.

5.5 Significant Correlation

An important contribution of Karl Pearson to statistics was precisely the study of correlation and the establishment of criteria supporting the significance of the correlation coefficient (also known as *Pearson's correlation coefficient*). Let us go back to the correlation between adult weight and height. The computed value of the correlation coefficient is 0.83. Can this value be attributed any kind of significance? For instance, in the case of the urn represented in Fig. 5.4a, we got a correlation estimate of 0.18, when in fact the true correlation value was 0 (uncorrelated B1 and B2). The 0.18 value suggesting some degree of relation between the variables could be called an artifact, due only to the small number of experiments performed.

Karl Pearson approached the problem of the 'significance' of the correlation value in the following way: assuming the variables to be normally distributed and uncorrelated how many cases n would be needed to ensure that, with a given degree of certainty (for instance, in 95% of the experiments with n cases), one would obtain a correlation value at least as high as the one computed in the sample? In other words, how large must the sample be so that with high certainty (95%) the measured correlation is not due to chance alone (assuming normally governed chance)?

The values of a significant estimate of the correlation for a 95% degree of certainty are listed in Table 5.1 for several values of n. For instance, for $n = 100$, we have 95% certainty that a correlation at least as high as 0.2 is not attributable to chance. In the height–weight data set, n is precisely 100. The value $C(\text{height, weight}) = 0.83$ is therefore

Table 5.1. Significant estimate of the correlation for a 95% degree of certainty

n	5	10	15	20	50	100	200
Significant correlation with 95% certainty	0.88	0.63	0.51	0.44	0.27	0.20	0.14

significant. (In fact, the normality assumption is not important for n above 50.)

Note once again the expression of a degree of certainty assigned to conclusions of random data analysis. This is an important aspect of *statistical proof*. If, for instance, we wish to prove the relation between tobacco smoke and lung cancer using the correlation, we cannot expect to observe all cases of smokers that have developed lung cancer. Basing ourselves on the correlation alone, there will always exist a non-null probability that such a relation does not hold. Of course, in general there is other evidence besides correlation. For instance, in physics the statistical proof applied to experimental results is often only one among many proofs. As an example, we may think of (re)determining Boyle–Mariotte's law for ideal gases through the statistical analysis of a set of experiments producing results on the pressure–volume variation, but the law itself – inverse relation between pressure and volume – is supported by other evidence beyond the statistical evidence. However, it should not be forgotten that the statistical proof so widely used in scientific research is very different from standard mathematical proof.

5.6 Correlation in Everyday Life

Correlation between random variables is one of the most abused topics in the media and in careless research and technical reports. Such abuse already shows up with the idea that, if the correlation between two variables is zero, then there is no relation between them. This is false. As mentioned before, correlation is a *linear* measure of association, so there may be some relation between the variables even when their Pearson correlation is zero.

The way that results of analyses involving correlations are usually presented leads the normal citizen to think that a high correlation implies a cause–effect association. The word 'correlation' itself hints at some 'intimate' or deep relation between two variables, whereas it is nothing other than a simple measure of linear association, which may be completely fortuitous.

In truth, an intrinsic cause–effect type relation may sometimes exist. For instance, the graph of Fig. 5.6 shows the area (in ha) of forest burnt annually in Portugal during the period 1943 through 1978, as a function of the total number of annual forest fires. There is here an obvious cause–effect relationship which translates into a high correlation of 0.9.

Let us now consider the tobacco smoke issue. Figure 5.7 is based on the results of a study carried out in the USA in 1958, which involved over a hundred thousand observations of adult men aged between 50 and 69 years. The horizontal axis shows the average number of cigarettes smoked per day, from 0 up to 2.5 packets. The vertical axis represents the frequency of occurrence of lung cancer relative to the total number of detected cancer cases. The correlation is high: 0.95. The probability of obtaining such a correlation due to chance alone (in a situation of non-correlation and normal distributions) is only 1%. Does the significantly high value of the correlation have anything to do with a cause–effect relation? In fact, several complementary studies concerning certain substances (carcinogenic substances) present in the tobacco smoke point in that direction, although lung cancer may occur in smokers in a way that is not uniquely caused by tobacco smoke,

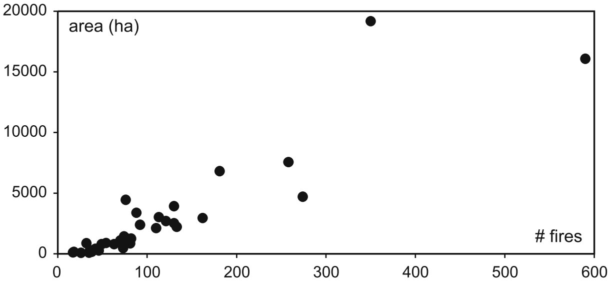


Fig. 5.6. Area (in ha) of annual burnt forest plotted against the number of forest fires in the period 1943 through 1978 (Portugal)

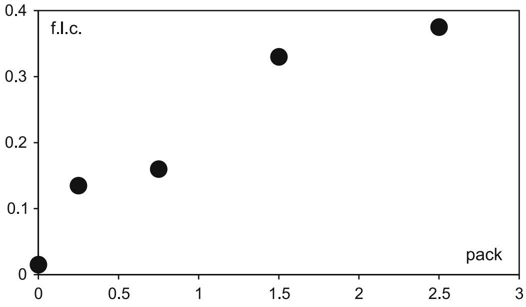


Fig. 5.7. Frequency of lung cancer (f.l.c.) and average number of smoked cigarettes in packets (USA study, 1958)

but instead by the interaction of smoke with the ingestion of alcoholic drinks, or other causes. Therefore, in this case, the causal link rests upon many other complementary studies. The mere observation of a high correlation only hints that such a link could be at work.

In many other situations, however, there is no direct 'relation' between two random variables, despite the fact that a significant correlation is observed. This is said to be a *spurious correlation*. Among the classic examples of spurious correlations are the correlation between the number of ice-creams sold and the number of deaths by drowning, or the correlation between the number of means used to fight city fires and the total amount of damage caused by those fires. It is obvious that ice-creams do not cause death by drowning or vice-versa; likewise, it is not the means used to fight fires that cause the material damage or vice versa. In both cases there is an obvious third variable, viz., the weather conditions and the extent of the fires, respectively.

Finally, let us present the analysis of a data set that gives a good illustration of the kind of abuse that arises from a superficial appreciation of the notion of correlation. The data set, available through the Internet, concerns housing conditions in Boston, USA, and comprises 506 observations of several variables, evaluated in different residential areas.

We may apply the preceding formula to compute the correlations between variables. For instance, the correlation between the percentage of poor people (lower status of the population) and the median value of the owner-occupied homes is negative: -0.74. This is a highly significant value for the 506 cases (the probability of such a value due to chance alone is less than $1/1\,000\,000$). Figure 5.8a shows the graph for these two variables. It makes sense for the median value of the houses to decrease with an increase in the percentage of poor people. The causal link is obvious: those with less resources will not buy expensive houses.

Let us now consider the correlation between the percentage of area used for industry and the index of accessibility to highways (measured on a scale from 1 to 24). The correlation is high, viz., 0.6. However, looking at Fig. 5.8b, one sees that there are a few atypical cases with accessibility 24 (the highest level corresponding to being close to the highways). Removing these atypical cases one obtains an almost zero correlation. We see how a few atypical cases may create a false idea of correlation. An examination of graphs relating the relevant variables is therefore helpful in order to assess this sort of situation. (Incidentally, even if we remove atypical cases from Fig. 5.6, the correlation will remain practically the same.)

Imagine now that we stumble on a newspaper column like the following: Enquiry results about the quality of life in Boston revealed that it is precisely the lower status of the population that most contributes to town pollution, measured by the concentration of nitric oxides in the atmosphere. To add credibility to the column we may even imagine that the columnist went to the lengths of including the graph of the two variables (something we very rarely come across in any newspaper!), as shown in Fig. 5.8c, in which one clearly sees that the concentration of nitric oxides (in parts per 10 million) increases with the percentage of poor people in the studied residential area.

In the face of this supposed causal relation – poor people cause pollution! – we may even imagine politicians proposing measures for solving the problem, e.g., compelling poor people to pay an extra pollution tax. Meanwhile, the wise researcher will try to understand whether that correlation is real or spurious. He may, for instance, study the relation

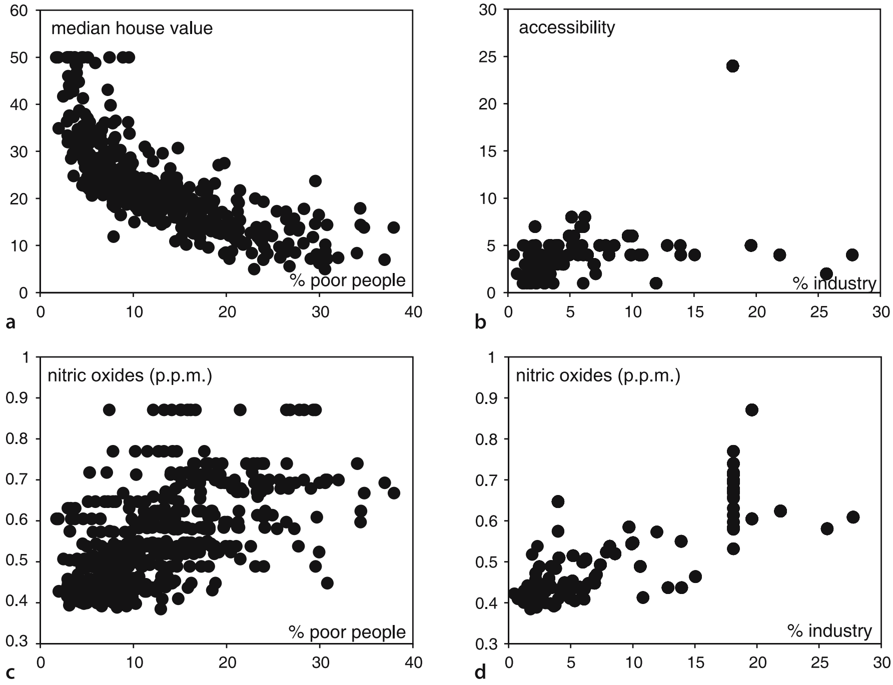
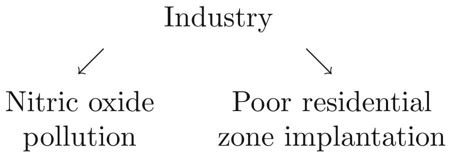


Fig. 5.8. Graphs and correlations for pairs of variables in the Boston housing data set: (**a**) −0.74, (**b**) 0.6, (**c**) 0.59, (**d**) 0.76

between percentage of area used by industry and percentage of poor people in the same residential area and verify that a high correlation exists (0.6). Finally, he may study the relation between percentage of area used by industry and concentration of nitric oxides (see Fig. 5.8d) and verify the existence of yet another high correlation, viz., 0.76. Complementary studies would then suggest the following causality diagram:



In summary, the correlation triggering the comments in the newspaper column is spurious and politicians should think instead of suitably taxing the polluting industries.

There are many other examples of correlation abuse. Maybe the most infamous are those pertaining to studies that aim to derive racist or eugenic conclusions, interpreting in a causal way skull measurements and intelligence coefficients. Even when using experimental data as in physics research, correlation alone does not imply causality; it only hints at, suggests, or opens up possibilities. In order to support a causality relation, the data will have to be supplemented by other kinds of evidence, e.g. the right sequencing in time.

Besides the correlation measure, statistics also provides other methods for assessing relations between variables. For instance, in the Boston pollution case, one might establish average income ranks in the various residential areas and obtain the corresponding averages of nitric oxides. One could then use statistical methods for comparing these averages. The question is whether this or other statistical methods can be used to establish a causality inference. Contrary to a widely held opinion, the answer is negative. The crux of the problem lies not in the methods, but in the data.

6

Fortune and Ruin

6.1 The Random Walk

The coin-tossing game, modest as it may look, is nevertheless rather instructive when we wish to consider what may be expected in general from games, and in particular the ups and downs of a player's fortunes. But let us be quite clear, when we speak of games, we are referring to games that depend on chance alone. We exclude games in which a player can apply certain logical strategies to counter the previous moves of his or her opponent (as happens in many card games, checkers, chess, etc.). On the other hand, we nevertheless understand 'games' in a wide sense: results derived from the analysis of coin-tossing sequences are applicable to other random sequences occurring in daily life.

The analysis of the coin-tossing game can be carried out in an advantageous manner by resorting to a specific interpretation of the meaning of each toss. For this purpose we consider arbitrarily long sequences of throws of a fair coin and instead of coding heads and tails respectively by 1 and 0, as we did in Chap. 4, we code them by $+1$ and -1. We may consider these values as winning or losing 1 euro. Let us compute the average gain along the sequence as we did in Chap. 4. For every possible tossing sequence the average gain will evolve in a distinct way but always according to the law of large numbers: for a sufficiently large number of throws n, the probability that the average gain deviates from the mathematical expectation (0) by more than a preset value is as small as one may wish. Figure 6.1 shows six possible sequences of throws, exhibiting the stabilization around zero, with growing n, according to the law of large numbers.

Instead of average gains we may simply plot the gains, as in Fig. 6.2. In the new interpretation of the game, instead of thinking in terms of

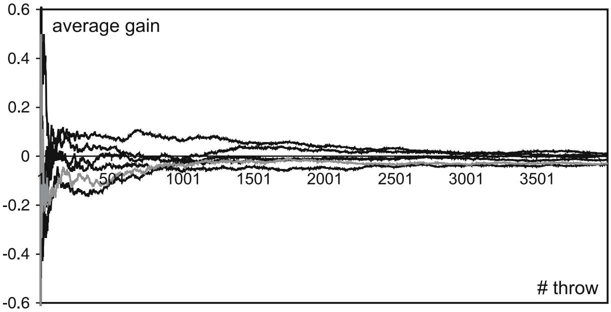


Fig. 6.1. Six sequences of average gain in 4000 throws of a fair coin

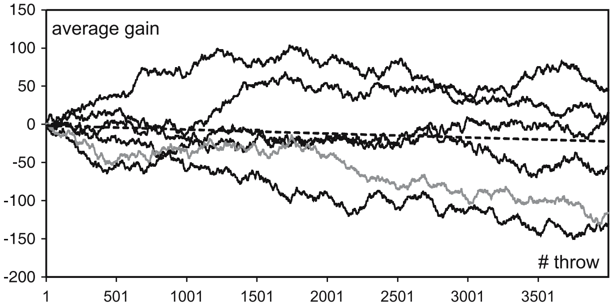


Fig. 6.2. Gain curves for the six sequences of Fig. 6.1

gain we think in terms of upward and downward displacements of a particle, along the vertical axis. Let us substantiate this idea by considering the top curve of Fig. 6.1, which corresponds to the following sequence in the first 10 throws:

tails, heads, heads, tails, heads, heads, heads, heads, tails, heads .

We may then imagine, in the new interpretation, that the throws are performed at the end of an equal time interval, say 1 second, and that the outcome of each throw will determine how a point particle moves from its previous position: +1 (in a certain space dimension) if the outcome is heads, and −1 if it is tails. At the beginning the particle is assumed to be at the origin. The particle trajectory for the 10 mentioned throws is displayed in Fig. 6.3a. We say that the particle has performed a *random walk*.

We may consider the six curves of Fig. 6.2 as corresponding to six random walks in one dimension: the particle either moves upwards or downwards. We may also group the six curves of Fig. 6.2 in three pairs of curves representing random walks of three particles in a two-axis system: one axis represents up and down movements and the other one right and left movements. Figure 6.3b shows the result for 1000 throws: the equivalent of three particles moving on a plane. We may also group

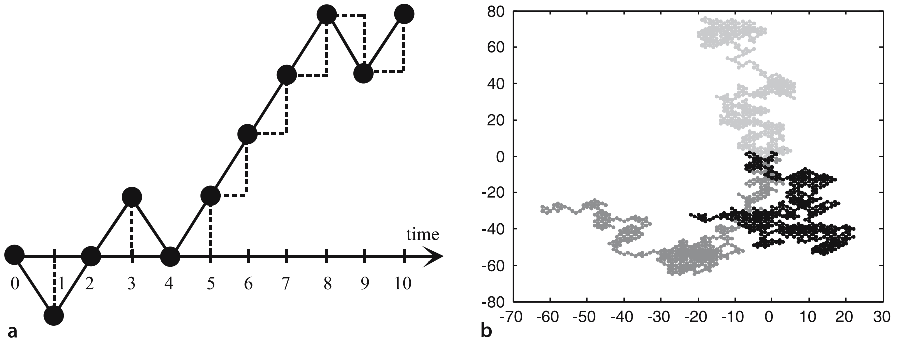


Fig. 6.3. (**a**) Random walk for ten throws. (**b**) Random walk of three particles in two dimensions

the curves in threes, obtaining three-dimensional random walks, as in Brownian motion (first described by the botanist Robert Brown for pollen particles suspended in a liquid).

The random walk, in one or more dimensions, is a good mathematical model of many natural and social phenomena involving random sequences, allowing an enhanced insight into their properties. Random walk analysis thus has important consequences in a broad range of disciplines, such as physics and economics.

6.2 Wild Rides

The random walk, despite its apparent simplicity, possesses a great many amazing and counterintuitive properties, some of which we shall now present.

Let us start with the property of obtaining a given walk for n time units (walk steps). As there are 2^n distinct and equiprobable walks, the probability is $1/2^n$. This probability quickly decreases with n. For only 100 steps, the probability of obtaining a given walk is

$$0.000\,000\,000\,000\,000\,000\,000\,000\,000\,000\,8\ .$$

For 160 steps it is roughly the probability of a random choice of one atom from all the matter making up the Earth, viz., roughly $1/8 \times 10^{49}$. For 260 steps (far fewer than in Fig. 6.2), it is roughly the probability of a random choice of one atom from all known galaxies, viz., roughly $1/4 \times 10^{79}$!

Let us go back to interpreting the random walk as a heads–tails game, say between John and Mary. If heads turns up, John wins 1 euro from Mary; if tails turns up, he pays 1 euro to Mary. The game is thus equitable. We may assume that the curves of Fig. 6.2 represent John's accumulated gains. However, we observe something odd. Contrary to common intuition, there are several gain curves (5 out of 6) that stay above or below zero for long periods of time. Let us take a closer look at this. We note that the number of steps for reaching zero gain will have to be even (the same number of upward and downward steps), say $2n$. Now, it can be shown that the probability of reaching zero gain *for the first time* in $2n$ steps is given by

$$P(\text{zero gain for the first time in } 2n \text{ steps}) = \frac{\binom{2n}{n}}{2^{2n}(2n-1)} \ .$$

Using this formula one observes that the probability of zero gain for the first time in 2 steps is 0.5, and for the first time in four steps is 0.125. Therefore, the probability that a zero gain has occurred in four steps, whether or not it is the first time, is 0.625.

Let us denote by $P(2n)$ the sum

$$\begin{aligned} &P(\text{zero gain for the first time in 2 steps}) \\ &\qquad +P(\text{zero gain for the first time in 4 steps}) \\ &\qquad\quad +\cdots+P(\text{zero gain for the first time in } 2n \text{ steps}) \ . \end{aligned}$$

This sum represents the probability of John's and Mary's gains having equalized at least once in $2n$ steps. It can be shown that $P(2n)$ converges to 1, which means that at least one equalization is sure to occur for sufficiently high n, and this certainly agrees with common intuition. However, looking at the $P(2n)$ curve in Fig. 6.4a we notice that the convergence is rather slow. For instance, after 64 steps the probability that no gain equalization has yet occurred is $1-0.9=0.1$. This means that, if ten games are being played simultaneously, the most probable situation (on average terms) is that in one of the games *the same player will always be winning*. After 100 steps the situation has not improved much: the probability that no equalization has yet occurred is $1-0.92=0.08$, which is still quite high!

The random walk really does look like a wild ride. The word 'random' coming from the old English 'randon', has its origin in an old Frank word, 'rant', meaning a frantic and violent ride. Game rides really are frantic and violent.

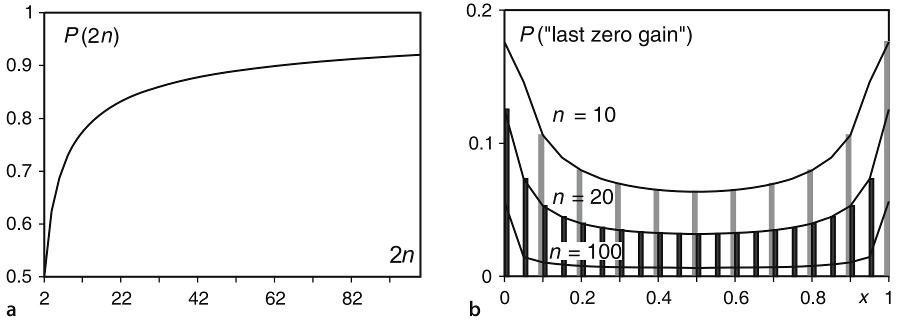


Fig. 6.4. (**a**) Probability of at least a zero gain in $2n$ steps. (**b**) Probability of the last zero gain occurring after x

6.3 The Strange Arcsine Law

We have seen above how the probability of not obtaining gain equalization in $2n$ steps decreases rather slowly. However, random walk surprises become even more dramatic when one studies the probability of obtaining zero gain for the last time after a certain number of rounds, say $2k$ in a sequence of $2n$ rounds. Letting $x = k/n$ represent the proportion of the total rounds corresponding to the $2k$ rounds (x varies between 0 and 1), it can be shown that this probability is given by the following formula:

$$P(\text{last zero gain at time } x \text{ in } 2n \text{ rounds}) \approx \frac{1}{n\pi\sqrt{x(1-x)}} \; .$$

For instance, the probability that the last zero gain has occurred in the middle of a sequence of 20 rounds ($n = 10$) is computed as follows:

$$\begin{aligned} P(\text{last zero gain at the tenth of 20 rounds}) &\approx \frac{1}{10 \times 3.14\sqrt{0.5 \times 0.5}} \\ &= 0.064 \; . \end{aligned}$$

The curves of Fig. 6.4b show this probability for several values of n (the probabilities for $x = 0$ and $x = 1$ are given by another formula). Note that, for each n, only certain values of x are admissible (vertical bars shown in the figure for $n = 10$ and $n = 20$). The bigger n is, the more values there are, with spacing $1/n$.

Common intuition would tell us that in a long series of throws John and Mary would each be in a winning position about half of the time and that the name of the winner should swap frequently. Let us now see what the curves in Fig. 6.4b tell us. Consider the events $x < 0.5$ and $x > 0.5$, which is the same as saying $2k < n$ and $2k > n$, respectively, or in more detail, 'the last zero gain occurred somewhere in the first half of $2n$ rounds' and 'the last zero gain occurred somewhere in the second half of $2n$ rounds'. Since the curves are symmetric, $P(x < 0.5)$ and $P(x > 0.5)$ are equal [they correspond to adding up the vertical bars from 0 to 0.5 (exclusive) and from 0.5 (exclusive) to 1]. That is, the events 'the last zero gain occurred somewhere in the first half of $2n$ rounds' and 'the last zero gain occurred somewhere in the second half of $2n$ rounds' are equiprobable; in other words, with probability close to 1/2, no zero gain occurs during half of the game, independently of the total number of rounds! Thus, the most probable situation is that one of the two, John or Mary, will spend half of the game permanently in the winning position. Moreover, the curves in Fig. 6.4b show that the last time that a zero gain occurred is more likely to happen either at the beginning or at the end of the game!

In order to obtain the probability that the last zero gain occurred *before and including* a certain time x, which we will denote by $P(x)$, one only has to add up the preceding probabilities represented in Fig. 6.4b, for values smaller than or equal to x. For sufficiently large n (say, at least 100), it can be shown that $P(x)$ is well approximated by the inverse of the sine function, called the *arcsine function* and denoted arcsin (see Appendix A):

$$P(\text{last zero gain occurred at a time} \leq x) = P(x) \approx \frac{2}{\pi}\arcsin(\sqrt{x}) \; .$$

The probability of the last zero gain occurring before a certain time x is said to follow the arcsine law, displayed in Fig. 6.5. Table 6.1 shows some values of this law.

Imagine that on a certain day several pairs of players passed 8 hours playing heads–tails, tossing a coin every 10 seconds. The number of

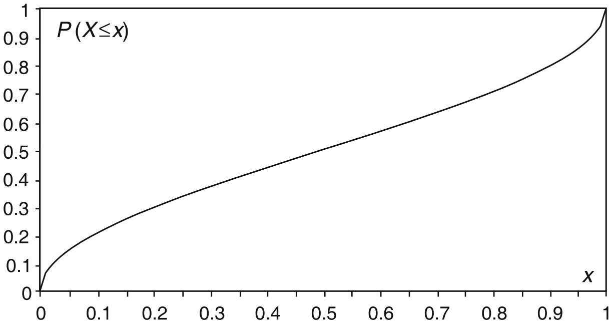


Fig. 6.5. Probability of the last zero gain occurring before or at x, according to the arcsine law

Table 6.1. Values of the arcsine law

x	0.005	0.01	0.015	0.02	0.025	0.03	0.095	0.145	0.205	0.345
$P(x)$	0.045	0.064	0.078	0.090	0.101	0.111	0.199	0.249	0.299	0.400

tosses ($8 \times 60 \times 6 = 2880$) is sufficiently high for a legitimate application of the arcsine law. Consulting the above table, we observe that on average one out of 10 pairs of players would obtain the last gain equalization during the first 12 minutes ($0.025 \times 2880 = 72$ tosses) passing the remaining 7 hours and 48 minutes without having any swap of winner position! Moreover, on average, one of each 3 pairs would pass the last 6 hours and 22 minutes without changing the winner!

6.4 The Chancier the Better

An amazing aspect of the convergence of averages to expectations is their behavior when the event probabilities change over a sequence of random experiments. Let us suppose for concreteness that the probabilities of a heads–tails type of game were changed during the game. For instance, John and Mary did a first round with a fair coin, a second round with a tampered coin, producing on average two times more heads than tails, and so on. Denote the probability of heads turning up in round k by p_k. For example, we would have $p_1 = 0.5$, $p_2 = 0.67$, and so on. The question is, will the arithmetic mean still follow the law of large numbers? The French mathematician Siméon Poisson (1781–1840) proved in 1837 that it does indeed, and that the arithmetic mean of the gains converges to the arithmetic mean of the p values, viz.,

$$\frac{p_1 + p_2 + \cdots + p_n}{n} \, .$$

(Poisson is better known for a probability law that carries his name and corresponds to an approximation of the binomial distribution for a very low probability of the relevant event.) Meanwhile, since the probability of turning up heads changes in each round, common intuition would tell us that it might be more difficult to obtain convergence. Yet the Russian mathematician Vassily Hoeffding (1914–1991) demonstrated more than a century after Poisson's result that, contrary to intuition, the probability of a given deviation of the average of the x values from the average of the p values reaches a maximum when the p_k are constant. In other words, the larger the variability of the p values the better

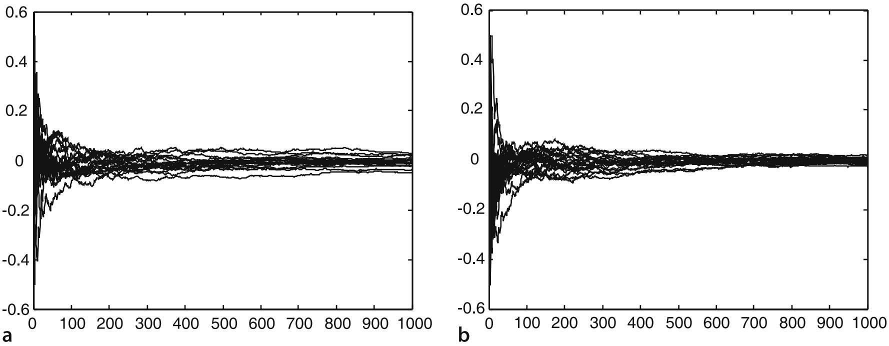


Fig. 6.6. Average gains in 20 sequences of 1000 rounds of the heads–tails game. (**a**) Constant probability $p_k = 0.5$. (**b**) Variable probability p_k uniformly random between 0.3 and 0.7

the convergence is! It is as if one said that by increasing the 'disorder' during the game more quickly, one achieves order in the limit!

This counterintuitive result is illustrated in Fig. 6.6, where it can be seen that the convergence of the averages is faster with variable probabilities randomly selected between 0.3 and 0.7.

6.5 Averages Without Expectation

We saw how the law of large numbers assured a convergence of the average to the mathematical expectation. There is an aspect of this law that is easily overlooked. In order to talk of convergence, the mathematical expectation must exist. This fact looks so trivial at first sight that it does not seem to deserve much attention. In fact, at first sight, it does not seem possible that there could be random distributions without a mathematical expectation. But is it really so?

Consider two random variables with normal distribution and with zero expectation and unit variance. To fix ideas, suppose the variables are the horizontal position x and the vertical position y of rifle shots at a target, as in Chap. 4. Since x and y have zero expectation, the rifle often hits the bull's-eye on average terms. We assume that the bull's-eye is a very small region around the point $(x = 0, y = 0)$. It may happen that a systematic deviation along y is present when a systematic deviation along x occurs. In order to assess whether this is the case, let us suppose that we computed the ratio $z = x/y$ for 500 shots and plotted the corresponding histogram. If the deviation along the vertical direction is strongly correlated with the deviation along the horizontal direction, we will get a histogram that is practically one single bar placed at the correlation value. For instance, if y is approximately equal to x, the histogram is practically a bar at 1. Now, consider the situation where x are y uncorrelated. We will get a broad histogram such as the one in Fig. 6.7a.

Looking at Fig. 6.7a, we may suspect that the probability density function of z is normal and with unit variance, like the broken curve also shown in Fig. 6.7a. However, our suspicion is wrong and the mistake becomes clearly visible when we look at the histogram over a sufficiently large interval, as in Fig. 6.7b. It now becomes clear that the normal curve is unable to approximate the considerably extended histogram tails. For instance, for z exceeding 5 in absolute value, we would expect to obtain according the normal law about 6 shots in 10 million. That is, for the 500 shots we do not expect to obtain any with z exceeding 5. Now, in reality that is not what happens, and we obtain a considerable number of shots with z in excess of 5. In the histogram of Fig. 6.7b, we obtained 74 such shots! In fact, z does not follow the normal law, but another law discovered by the French mathematician Augustin Louis Cauchy (1789–1857) and called the Cauchy distribution in his honor. The Cauchy probability density function

$$f(z) = \frac{1}{\pi(1+z^2)}$$

is represented by the solid line in Fig. 6.7b.

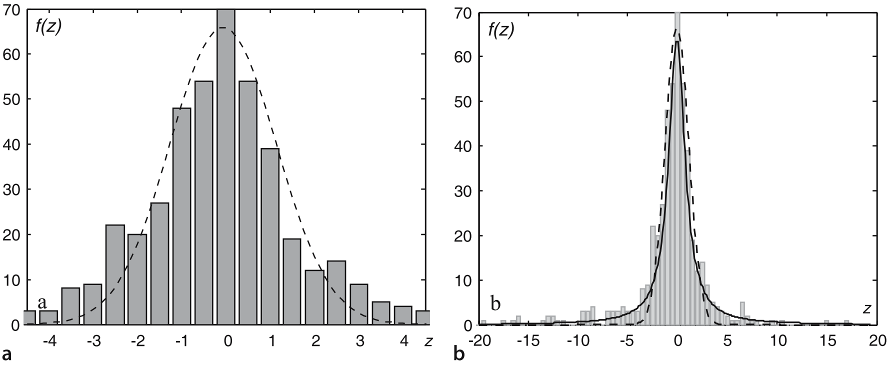


Fig. 6.7. Histogram of the ratio of normal deviations in 500 shots: (**a**) with the normal probability density function (*broken curve*), (**b**) with the Cauchy probability density function (*continuous curve*)

Visually, Cauchy's curve represented in Fig. 6.7b does not look much of a monster; it even looks like the bell-shaped curve and there is a tendency to think that its expectation is the central 0 point (as in the analogy with the center of gravity). But let us consider, in Fig. 6.8a, what happens with a sequence of values generated by such a distribution. We see that the sequence exhibits many occurrences of large values (sometimes, very large), clearly standing out from most of the values in the sequence. In contrast, a normally distributed sequence is characterized by the fact that it very rarely exhibits values that stand out from the 'background noise'. The impulsive behavior of the Cauchy distribution is related to the very slow decay of its tails for very small or very large values of the random variable; this slow decay is reflected in a considerable probability for the occurrence of extreme values.

Suppose now that we take the random values with Cauchy distribution as wins and losses of a game and compute the average gain. Figure 6.8b shows 6 average gain sequences for 1000 rounds. It seems that there is some stabilization around 0 for some sequences, but this impression is deceptive. For example, we observe a sequence close to 400 rounds that exhibits a large jump with a big deviation from zero. Had we carried out a large number of experiments with a large number of rounds, we would then confirm that the averages do not stabilize around zero. The reason is that the Cauchy law does not really have a mathematical expectation. In some everyday phenomena, such as risk assessment in financial markets, sequences obeying such strange laws as the Cauchy law do emerge. Since such laws do not have a mathematical expectation, it has no meaning to speak of the law of large numbers for those sequences, and it makes no sense to speak of convergence or long term behavior; fortune and ruin may occur abruptly.

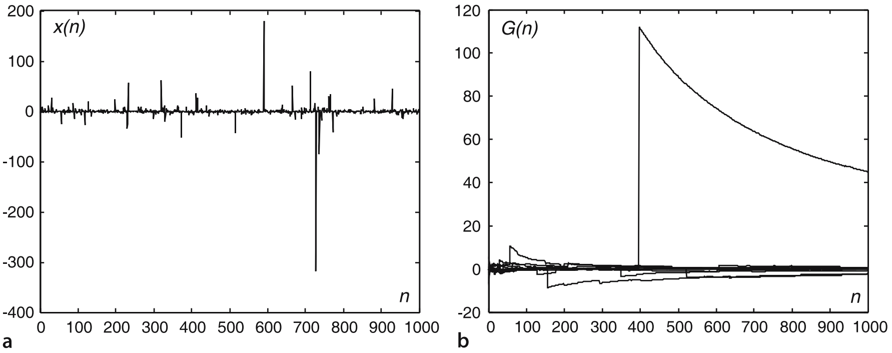


Fig. 6.8. (**a**) A sequence of rounds with Cauchy distribution. (**b**) Six average gain sequences for 1000 rounds

6.6 Borel's Normal Numbers

We spoke of confidence intervals in Chap. 4 and, on the basis of the normal law, we were able to determine how many throws of a die were needed in order to be 95% sure of the typicality of a ±0.01 interval around the probability p for the frequency of occurrence f in the worst case (worst standard deviation). This result (inverse of $\sqrt{n}$) constituted an illustration of the law of large numbers. Denoting the number of throws by n, we could write

$$P(\text{deviation of } f \text{ from } p \text{ is below } 0.01) = 0.95 \;, \qquad \text{for } n > 10\,000 \;.$$

In fact, the law of large numbers stipulates that whatever the tolerance of the deviation might be (0.01 in the above example), it is always possible to find a minimum value of n above which one obtains a degree of certainty as large as one may wish (0.95 in the above example). In the limit, i.e., for arbitrarily large n, the degree of certainty is 1. Therefore, denoting any preselected tolerance by t, we have

$$\text{limit of } P(\text{deviation of } f \text{ from } p \text{ is below } t) = 1 \;.$$

In general, using the word 'deviation' to denote the absolute value of the deviation of an average (f, in the example) from the mathematical expectation (p, in the example), we may write

$$\text{limit of } P(\text{deviation is below } t) = 1 \;.$$

Consequently, when applying the law of large numbers, we first compute a probability and thereafter its limit. Now, it does frequently happen that we do not know how many times the random experiment will be repeated, and we are more interested in *first* computing the limit of the deviations between average and expectation and only afterward determining the probability of the limit. That is, right from the start we must consider *infinite sequences* of repetitions.

Setting aside the artificiality of the notion of an 'infinite sequence', which has no practical correspondence, the study of infinite sequences is in fact rather insightful. Let us take a sequence of n tosses of a coin. Any of the 2^n possible sequences is equiprobable, with probability $1/2^n$. Imagine then that the sequence continued without limit. Since $1/2^n$ decreases with n, we would thus be led to the paradoxical conclusion that no sequence is feasible; all of them would have zero probability. This difficulty can be overcome if we only assign zero probability to certain sequences.

Table 6.2. Assigning numbers to sequences

Sequence	Binary	Decimal value
Tails, tails, tails, tails, heads	0.00001	0.03125
Heads, heads, heads, heads, heads	0.11111	0.96875
Heads, tails, heads, tails, heads	0.10101	0.65625

Let us see how this is possible, omitting some technical details. For this purpose, consider representing heads and tails respectively by 1 and 0, as we have already done on several occasions. For instance, the sequence heads, tails, heads, heads is represented by 1011. We may assume that 1011 represents a number in a binary numbering system. Instead of the decimal system that we are accustomed to from an early age, and where the position of each digit represents a power of 10 (e.g., $375 = 3 \times 10^2 + 7 \times 10^1 + 5 \times 10^0$), in the binary numbering system the position of each digit (there are only two, 0 and 1), called a *bit*, represents a power of 2. So 1011 in the binary system is evaluated as $1 \times 2^3 + 0 \times 2^2 + 1 \times 2^1 + 1 \times 2^0$, that is, 11 (eleven) in the decimal system. Let us now take the 1011 sequence as the representation of a number in the interval $[0, 1]$. It is thus 0.1011; we just omit the 0 followed by the decimal point. To compute its value in the decimal system we shall have to multiply each bit by a power of 1/2 (by analogy, in the decimal system each digit is multiplied by a power of 1/10). That is, 1011 is then evaluated as

$$1 \times \frac{1}{2} + 0 \times \frac{1}{4} + 1 \times \frac{1}{8} + 1 \times \frac{1}{16} = 0.6875 \ .$$

Table 6.2 shows some more examples.

It is easy to understand that all infinite heads–tails sequences can thus be put in correspondence with numbers in the interval $[0, 1]$, with 1 represented by $0.1111111111\ldots$. In the case of 1/2, which may be represented by both $0.1000000\ldots$ and $0.0111111111\ldots$, we adopt the convention of taking the latter (likewise for 1/4, 1/8, etc.).

We may visualize this binary representation by imagining that each digit corresponds to dividing the interval [0, 1] into two halves. The first digit represents the division of [0, 1] into two intervals: from 0 to 0.5 (excluding 0.5) and from 0.5 to 1; the second digit represents the division of each of these two intervals into halves; and so on, in succession, as shown in Fig. 6.9, assuming a sequence of only $n = 3$ digits.

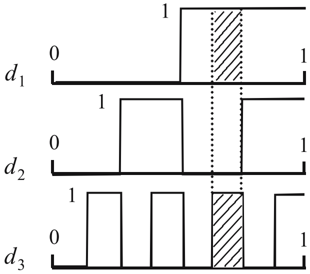


Fig. 6.9. The first 3 digits (d_1, d_2, d_3) and the intervals they determine in $[0, 1]$. The value 0.625 corresponds to the *dotted intervals*

Figure 6.9 shows the intervals that 0.625 $(1 \times 1/2 + 0 \times 1/4 + 1 \times 1/8)$ 'occupies'. We now establish a mapping between probabilities of binary sequences and their respective intervals in $[0, 1]$. Take the sequence heads, tails, heads represented by 0.101. We may measure the probability that this sequence occurs by taking the width of the corresponding interval in $[0, 1]$: the small interval between 0.625 and 0.75 with width 1/8 shown in Fig. 6.9. That is, from the point of view of assigning probabilities to the sequence heads, tails, heads, it comes to the same thing as randomly throwing a point onto the straight line segment between 0 and 1 (excluding 0), with the segment representing a kind of target cut into slices 0.125 units wide. If the point-throwing process has a uniform distribution, each number has probability 1/8, as required.

Now, consider infinite sequences. To each sequence corresponds a number in $[0, 1]$. Let us look at those sequences where 0 and 1 show up the same number of times (thus, the probability of finding 0 or 1 is 1/2), as well as any of the four two-bit subsequences 00, 01, 10 and 11 (each showing up 1/4 of the time), any of the three-bit subsequences 000, 001, 010, 011,100, 101, 110 and 111 (each showing up 1/8 of the time), and so forth. The real numbers in $[0, 1]$ corresponding to such sequences are called *Borel's normal numbers*.

The French mathematician Emile Borel (1871–1956) proved in 1909 that the set of non-normal numbers occupies such a meaningless 'space' in $[0, 1]$ that we may consider their probability to be zero. Take, for instance, the sequence 0.101100000... (always 0 from the fifth digit onward). It represents the rational number 11/16. Now take the sequence 0.1010101010... (repeating forever). It represents the rational number 2/3. Obviously, no rational number (represented by a sequence that either ends in an infinite tail of zeros or ones, or by a periodic sequence) is a Borel normal number. The set of all sequences corresponding to rational numbers thus has zero probability; in other words, the probability that we get an infinite sequence corresponding to a rational number when playing heads–tails is zero. On the other hand, since the set of all normal numbers has probability 1, we may then say that practically all real numbers in $[0, 1]$ are normal (once again, a result that may seem contrary to common intuition). Yet, it has been found to be a difficult task to prove that a given number is normal. Are the decimal expansions of $\sqrt{2}$, e or π normal? No one knows (although there is some evidence that they are). A few numbers are known to be normal. One of them, somewhat obvious, was presented by the economist David Champernowne (1920–2000) and corresponds to concatenating the distinct binary blocks of length 1, 2, 3, etc.:

$$0.0100011011000001010011100101110111\ldots\ .$$

The so-called Copeland–Erdös constant

$$0.235\,711\,131\,719\,232\,931\,373\,941\ldots\ ,$$

obtained by concatenating the prime numbers, is also normal in the base 10 system.

We end with another amazing result. It is well known that the set of all infinite 0–1 sequences is uncountable; that is, it is not possible to establish a one-to-one mapping between these sequences and the natural numbers. Well, although the set of non-normal numbers has zero probability, it is also uncountable!

6.7 The Strong Law of Large Numbers

Let us look at sequences of n zeros and ones as coin tosses, denoting by f the frequency of occurrence of 1. The normal numbers satisfy, by definition, the following property:

$$P(\text{limit of the deviation of } f \text{ from } 1/2 \text{ is } 0) = 1\ .$$

This result is a particular case of a new version of the law of large numbers. In Chap. 3, the Bernoulli law of large numbers stated that

$$\text{limit of } P(\text{deviation is below } t) = 1\ ,$$

where the term ‘deviation’ refers to the absolute value of the difference between the average and the expectation (1/2, for coin throwing), as already pointed out in the last section. The new law, demonstrated by the Russian mathematician Andrei Kolmogorov (1903–1987) in 1930 (more than two centuries after the Bernoulli law), is stated as follows:

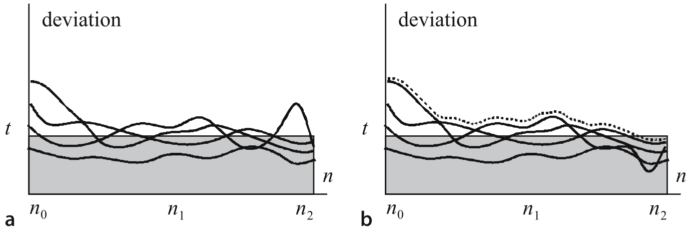


Fig. 6.10. Convergence conditions for the weak law (**a**) and the strong law (**b**) of large numbers

$$P(\text{limit of the deviation is } 0) = 1 .$$

Let us examine the difference between the two formulations using Fig. 6.10, which displays several deviation curves as a function of n, with t the tolerance of the deviations. In the first version, illustrated in Fig. 6.10a, the probability of a deviation larger than t must converge to 0; that is, once we have fixed the tolerance t (the gray band), we find fewer curves above t as we proceed toward larger values of n. In the case of Fig. 6.10a, there are three curves above t at n_0, two at n_1 and one at n_2. The law says nothing about the deviation values. In Fig. 6.10a, the deviation at n_2 is even more prominent than the deviation at n_1. In the second version of the law, in order for the limit of the average to equal the limit of the expectation with high probability, large wanderings above any preselected t must become less frequent each time, as Fig. 6.10b is intended to illustrate. An equivalent formulation consists in saying that, in the *worst case* – the dotted line in Fig. 6.10b – the probability of exceeding t converges to zero. Thus, the second version of the law of large numbers imposes a stronger restriction on the behavior of the averages than the first version. It is thus appropriate to call it the *strong law* of large numbers. The Bernoulli law is then the *weak law*. Understanding the laws of large numbers is an essential prerequisite to understanding the convergence properties of averages of chance sequences.

6.8 The Law of Iterated Logarithms

We spoke of the central limit theorem in Chap. 4, which allowed us to compute the value of n above which there is a high certainty of obtaining a frequency close to the corresponding probability value. We were thus able to determine that, when $p = 1/2$ (the coin-tossing example), only after 10 000 throws was there a risk of at most 5% that the $f \pm 0.01$ confidence interval was not typical, i.e., did not contain the true value of p. On the other hand, the strong law of large numbers shows us that there is an almost sure convergence of f toward p. How fast is that convergence? Take a sequence of experiments with only two possible outcomes, as in the coin-tossing example, and denote the probability of one of the outcomes by p. Let $S(n)$ be the accumulated gain after n outcomes and consider the absolute value of the discrepancy between $S(n)$ and its expectation, which we know to be np (binomial law). Now, the Russian mathematician Alexander Khintchine demonstrated the following result in 1924:

$$P(\text{limit, in the worst case, of a } \sqrt{2npq\log\big(\log n\big)} \text{ deviation}) = 1\ .$$

Note that the statement is formulated in terms of a worst case deviation curve as in Fig. 6.10b. This result, known as the *theorem of iterated logarithms* (the logarithm is repeated), allows one to conclude that $f = S(n)/n$ will almost certainly be between p plus or minus a deviation of $\sqrt{2pq\log\big(\log n\big)/n}$, where 'almost certainly' means that, if we add an infinitesimal increase to that quantity, only a finite subset of all possible infinite sequences will deviate by more than that quantity. For instance, for 10 000 throws of a coin we are almost sure that all sequences will deviate by less than $\sqrt{0.5\log\big(\log 10\,000\big)/10\,000} = 0.0105$.

Figure 6.11 shows 50 sequences of 10 000 throws of a coin with the deviation of the iterated logarithm (black curves). Note how, after 10 000 throws, practically all curves are inside 0.5 plus or minus

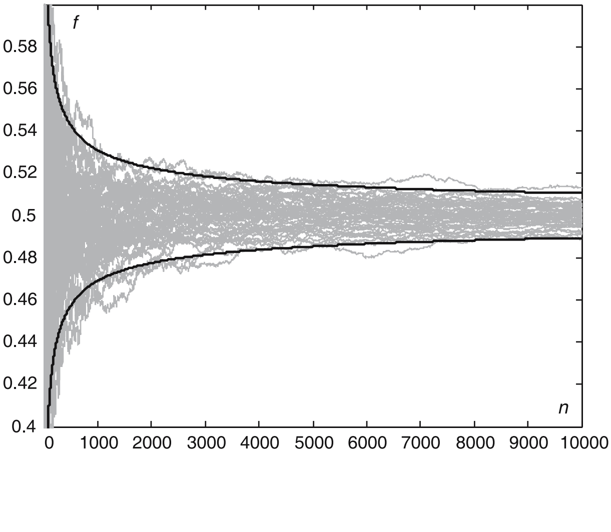


Fig. 6.11. Frequencies of heads or tails in 50 sequences of 10 000 throws of a coin, with the deviation of the strong law

Fig. 6.12. Pushing the poor souls to the limit

the deviation of the iterated logarithm. Note also the slow decay of the iterated logarithm. The strong law of large numbers and the law of iterated logarithms show us that, even when we consider infinite sequences of random variables, there is order in the limit.

7

The Nature of Chance

7.1 Chance and Determinism

The essential property characterizing chance phenomena is the *impossibility of predicting any individual outcome*. When we toss a fair die or a fair coin in a fair way we assume that that condition is fulfilled. We assume the impossibility of predicting any individual outcome. The only thing we may perhaps be able to tell with a given degree of certainty, based on the laws of large numbers, is the *average outcome* after a large number of experiments. Yet one might think that it would be possible in principle to predict any individual outcome. If one knew all the initial conditions of motion, either of the die or of the coin, together with the laws of motion, one might think it possible to forecast the outcome. Let us look at the example of the coin. The initial conditions that would have to be determined are the force vector applied to the coin, the position of its center of gravity relative to the ground, the shape of the coin and its orientation at the instant of throwing, the viscosity of the air, the value of the acceleration due to gravity at the place of the experiment, and the friction and elasticity factors of the ground. Once all these conditions are known, writing and solving the respective motion equations would certainly be a very complex task, although not impossible. The problem is thus *deterministic*. Every aspect of the problem (coin fall, coin rebound, motion on the ground with friction, etc.) obeys perfectly defined and known laws.

At the beginning of the twentieth century there was a general conviction that the laws of the universe were totally deterministic. Probability theory was considered to be a purely mathematical trick allowing us to deal with our ignorance of the initial conditions of many phenomena. According to Pierre-Simon Laplace in his treatise on probability theory:

> If an intelligence knew at a given instant all the forces animating matter, as well as the position and velocity of any of its molecules, and if moreover it was sufficiently vast to submit these data to analysis, it would comprehend in a single formula the movements of the largest bodies of the Universe and those of the tiniest atom. For such an intelligence nothing would be irregular, and the curve described by an air or vapor molecule would appear to be governed in as precise a way as is for us the course of the Sun. But due to our ignorance of all the data needed for the solution of this grand problem, and the impossibility, due to our limited abilities, of subjecting most of the data in our possession to calculation, even when the amount of data is very limited, we attribute the phenomena that seem to us to occur and succeed without any particular order to changeable and hidden causes, whose action is designated by the word *chance*, a word that after all is only the expression of our ignorance. Probability relates partly to this ignorance and in other parts to our knowledge.

The development of quantum mechanics, starting with the pioneering work of the physicists Max Planck (research on black body radiation, which established the concept of the emission of discrete energy quantities, or energy *quanta*), Albert Einstein (research on the photoelectric effect, establishing the concept of photons as light *quanta*) and Louis de Broglie (dual wave–particle nature of the electron), at the beginning of the twentieth century, came to change the Laplacian conception of the laws of nature, revealing the existence of many so-called quantum phenomena whose probabilistic description is not simply the result of resorting to a comfortable way of dealing with our ignorance about the conditions that produced them. For quantum phenomena the probabilistic description is not just a handy trick, but an absolute necessity imposed by their intrinsically random nature. Quantum phenomena are ‘pure chance’ phenomena.

Quantum mechanics thus made the first radical shift away from the classical concept that everything in the universe is perfectly determined. However, since quantum phenomena are mainly related to elementary particles, i.e., microscopic particles, the conviction that, in the case of macroscopic phenomena (such as tossing a coin or a die), the probabilistic description was merely a handy way of looking at things and that it would in principle be possible to forecast any individual outcome, persisted for quite some time. In the middle of the twentieth century several researchers showed that this concept would also have to be revised. It was observed that in many phenomena governed by perfectly deterministic laws, there was in many circumstances such a high sensitivity to initial conditions that one would never be able to guarantee the same outcome by properly adjusting those initial conditions. However small a deviation from the initial conditions might be, it could sometimes lead to an unpredictably large deviation in the experimental outcomes. The only way of avoiding such disorderly behavior would be to define the initial conditions with infinite accuracy, which amounts to an impossibility. Moreover, it was observed that much of this behavior, highly sensitive to initial conditions, gave birth to sequences of values – called chaotic orbits – with no detectable periodic pattern. It is precisely this chaotic behaviour that characterizes a vast number of chance events in nature. We can thus say that chance has a threefold nature:

- A consequence of our incapacity or inability to take into account *all* factors that influence the outcome of a given phenomenon, even though the applicable laws are purely deterministic. This is the coin-tossing scenario.
- A consequence of the extreme sensitivity of many phenomena to certain initial conditions, rendering it impossible to forecast future outcomes, although we know that the applicable laws are purely deterministic. This is the nature of *deterministic chaos* which is manifest in weather changes, for instance.
- The intrinsic random nature of phenomena involving microscopic particles, which are described by quantum mechanics. Unlike the preceding cases, since the random nature of quantum phenomena is an intrinsic feature, the best that one can do is to determine the probability of a certain outcome.

In what follows we first focus our attention on some amazing aspects of probabilistic descriptions of quantum mechanics. We then proceed to a more detailed analysis of deterministic chaos, which plays a role in a vast number of macroscopic phenomena in everyday life.

7.2 Quantum Probabilities

The description of nature at extremely small scales, as is the case of photons, electrons, protons, neutrons and other elementary particles, is ruled by the laws of quantum mechanics. Such particles are generically named quanta. There is a vast popular science literature on this topic (a few indications can be found in the references), describing in particular the radical separation between classical physics, where laws of nature are purely deterministic, and quantum mechanics where laws of nature at a sufficiently small, microscopic, scale are probabilistic. In this section we restrict ourselves to outlining some of the more remarkable aspects of the probabilistic descriptions of quantum mechanics (now considered to be one of the best validated theories of physics).

A first remarkable feature consists in the fact that, for a given microscopic system, there are variables whose values are objectively undefined. When the system undergoes a certain evolution, with or without human intervention, it is possible to obtain values of these variables which only depend on purely random processes: objective, pure, randomness. More clearly, the chance nature of the obtained outcome does not depend on our ignorance or on incomplete specifications as is the case for chaotic phenomena. In quantum mechanics, event probabilities are *objective*.

An experiment illustrating the objective nature of chance in quantum mechanics consists in letting the light of an extremely weak light source fall upon a photographic plate, for a very short time. Suppose that some object is placed between the light source and the plate. Since the light intensity is extremely weak, only a small number of photons falls on the plate, which will then reveal some spots distributed in a haphazard way around the photographed object. If we repeat the experiment several times the plates will reveal different spot patterns. Moreover, it is observed that there is no regularity in the process, as would happen for instance if the first half of the photons impacting the plate were to do so in the middle region and the other half at the edges. The process is genuinely random. On the other hand, if we combine the images of a large number of plates, the law of large numbers comes into play and the image (silhouette) of the photographed object then emerges from a 'sequence of chance events'. It is as though we had used a Monte Carlo method as a way of defining the contour of the object, by throwing tiny points onto the plate.

One strange aspect of quantum mechanics lies in the fact that certain properties of a microscopic particle, such as its position, are characterized by a probability wave (described by the famous wave equation discovered by the Austrian physicist Erwin Schrödinger in 1926). We cannot, for instance, say at which point a given particle is at a given instant. In the absence of a measurement there is a non-null probability that the particle is 'positioned' at several points at the same time. We write 'positioned' between quotes because in fact it does not have the usual meaning; it really is as though the particle were 'smeared out' in space, being present partly 'here', with a certain probability, and partly 'there', with another probability. The particle is said to be in a *superposition state*. This superposition state aspect and the wave nature, and therefore *vector* nature of the quantum description, give rise to counterintuitive results which are impossible to obtain when dealing with uncertainties in everyday life, such as the outcomes of games, forecasts in the stock market, weather forecasts, and so on, characterized by a *scalar* probability measure.

There is a much discussed experiment, the two-slit experiment, that illustrates the strange behavior of the probabilistic vector nature of quantum phenomena. A narrow beam of light shines on a target, separated from the light source by a screen with two small holes (or slits), as shown in Fig. 7.1.

Let us interpret the experiment as in Chap. 4 for the example of shooting at a target, treating the photons like tiny bullets. With this assumption the probability of a certain target region being hit by a bullet is the sum of the probabilities for each of the two possible bullet trajectories; either through A or through B. We may even assume a normal probability density for each of the two possible trajectories, as in Fig. 7.1a. In any small region of the target the bullet-hit probability is the sum of the probabilities corresponding to each of the normal

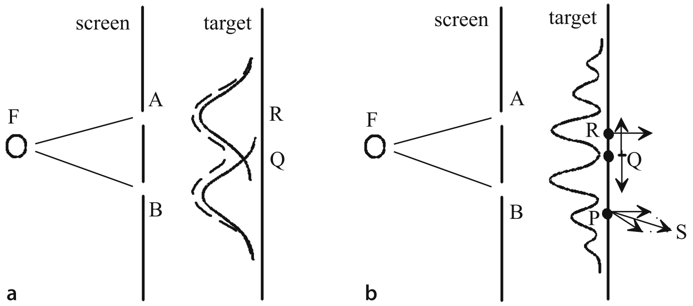


Fig. 7.1. Probability density functions in the famous two-slit experiment. (**a**) According to classical mechanics. (**b**) What really happens

distributions. Thus, the probability density of the impact due to the two possible trajectories is obtained as the sum of the two densities, shown with a dotted line in Fig. 7.1a. This would certainly be the result with macroscopic bullets. Now, that is not what happens with photons, where a so-called interference phenomenon like the one shown in Fig. 7.1b is observed, with concentric rings of smaller or larger light intensity than the one corresponding to a single slit. This result can be explained in the framework of classical wave theory. Quantum theory also explains this phenomenon by associating a probability wave to each photon.

In order to get an idea as to how the calculations are performed, we suppose that an arrow (vector) is associated with each possible trajectory of a photon, with the base of the arrow moving at the speed of light and its tip rotating like a clock hand. At a given trajectory point the probability of finding the photon at that point is given by the *square* of the arrow length. Consider the point P. Figure 7.1b shows two arrows corresponding to photon trajectories passing simultaneously (!) through A and B; this is possible due to the state superposition. When arriving at P, the clock hands exhibit different orientations, asymmetric in relation to the horizontal bearing, since the lengths of the trajectories FAP and FBP are different. The total probability is obtained as follows: the two arrows are added according to the parallelogram rule (the construction illustrated in Fig. 7.1b), and the arrow ending at S is obtained as a result. The square of the length of the latter arrow corresponds to the probability of an impact at P. At the central point Q, the arrows corresponding to the two trajectories come up with symmetric angles around the horizontal bearing, since the lengths of FAQ and FBQ are now equal. Depending on the choice of the distance between the slits, one may obtain arrows with the same bearing but opposite directions, which cancel each other when summing. There are also points, such as R, where the arrows arrive in a concordant way and the length of the sum arrow is practically double the length of each vector. It is then possible to observe null probabilities and probabilities that are four times larger than those corresponding to a single slit. We thus arrive at the surprising conclusion that the quantum description does not always comply with

$$\begin{aligned}&P(\text{passing through A or passing through B})\\&\qquad\qquad = P(\text{passing through A}) + P(\text{passing through B})\ ,\end{aligned}$$

as would be observed in a classical description!

What happens when a single photon passes through a single slit? Does it smear out in the target as a blot with a bell-shaped intensity? No. The photon ends at an exact (but unknown) point. The bell-shaped blot is a result of the law of large numbers. There is no intrinsic feature deciding which photons hit the center and which ones hit the periphery. The same applies to the interference phenomenon. Only the probabilistic laws that govern the phenomena are responsible for the final result. Incidentally, if we place photon detectors just after the two slits A and B, the interference phenomenon no longer takes place. Now, by acting with a detector, the state superposition is broken and we can say which slit a given photon has passed through.

The interference phenomenon has also been verified experimentally with other particles of reduced dimensions such as electrons, atoms, and even with larger molecules, as in recent experiments with the buckminsterfullerene molecule, composed of 60 carbon atoms structured like a football.

Another class of experiments illustrating the strangeness of quantum probabilities relates to properties of microscopic particles exhibiting well defined values. It is a known fact that light has a wave nature as well as a particle nature, represented by the oscillation of an electromagnetic field propagating as a wave, with the electric field vector vibrating perpendicularly to the propagation direction. Common light includes many wave components with various vibration directions, called *polarizations*. However, with the help of a polarizing device (for instance, a sheet of a special substance called polaroid or a crystal of the mineral calcite), one may obtain a single light component with a specific direction of polarization.

In Fig. 7.2b, horizontally polarized light (the electric field vibrates perpendicularly to the page), denoted by h, shines on a polarizer whose direction of polarization makes a 45-degree angle with the horizontal

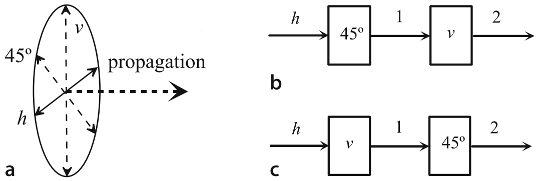


Fig. 7.2. (**a**) Polarization directions in a plane perpendicular to the propagation direction. (**b**) Horizontally polarized light passing through two polaroids. (**c**) As in (**b**) but in the opposite order

direction (see Fig. 7.2a). Let E_h be the electric field intensity of h. The intensity coming out of the 45° polarizer corresponds to the field projection along this direction, that is,

$$E_1 = E_h \cos 45^\circ = \frac{E_h}{\sqrt{2}} \ .$$

The light passes next through a vertical polarizer (the electric field vibrates in the page perpendicularly to the direction of propagation), and likewise

$$E_2 = E_1 \sin 45^\circ = \frac{E_1}{\sqrt{2}} = \frac{E_h}{2} \ .$$

Let us now suppose that we change the order of the polarizers as in Fig. 7.2c. We obtain

$$E_1 = E_h \cos 90^\circ = 0 \ ,$$

hence, $E_2 = 0$. These results, involving electric field projections along polarization directions, are obvious in the framework of the classical wave theory. Let us now interpret them according to quantum mechanics. In fact, light intensity, which is proportional to the square of the electric field amplitude, measures the probability of a photon passing through the two-polarizer sequence. But we thus arrive at the conclusion, contrary to common sense, that in the quantum description the probabilities of a chain of events are not always commutative, independent of the order of events, since the probability associated with the experimental result of Fig. 7.2b is different from the one associated with Fig. 7.2c. Suppose that instead of light we had a flow of sand grains passing through a sequence of two sieves: sieve 1 and sieve 2. In the classical description, the commutative property holds:

$$P(\text{passing through sieve 1}) \times P\begin{pmatrix}\text{passing through sieve 2} \\ \text{if it has passed through sieve 1}\end{pmatrix}$$

$$= P(\text{passing through sieve 2}) \times P\begin{pmatrix}\text{passing through sieve 1} \\ \text{if it has passed through sieve 2}\end{pmatrix} .$$

In the quantum description the probability of the intersection depends on the order of the events! This non-commutative feature of quantum probabilities has also been observed in many other well-defined properties of microscopic particles, such as the electron *spin*.

The laws of large numbers also come into play in all quantum phenomena, as we saw earlier in the case where an object was photographed with very weak light. In the interference example, if the source light only emits a small number of photons per second, one initially observes a poorly defined interference pattern, but it will begin to build up as time goes by. Another particularly interesting illustration of the law of large numbers acting on quantum phenomena is obtained in an experiment in which the intensity of a light beam with a certain polarization, say h, is measured after it has passed through a polarizer with polarization axis at an angle θ to h. We have already seen that the light intensity is then proportional to the square of $\cos\theta$. In fact, it is proportional to the number of photons passing through the polarizer, viz., $N_{\mathrm{p}} = N\cos^2\theta$, where N is the number of photons hitting the polarizer. If we carry out the experiment with a light beam of very weak intensity (in such a way that only a small number of photons hits the polarizer) and, for each value of θ, we count how many photons pass through (this can be done with a device called a photomultiplier), we obtain a graph like the one shown in Fig. 7.3. This graph clearly shows random deviations in the average number of photons N_{p}/N passing through the polarizer from the theoretical value of $\cos^2\theta$. An experiment with a normal light beam, with very high N, can be interpreted as carrying out multiple experiments such as the one portrayed in Fig. 7.3. But in this case the law of large numbers comes into play, and in such a way that the average number of photons passing through the polarizer exhibits a standard deviation proportional to $1/\sqrt{N}$ (as already seen in Chap. 4). In summary, for an everyday type of light beam, the law of large numbers explains the $\cos^2\theta$ proportionality, as deduced in the framework of classical physics for the macroscopic object we call a light beam.

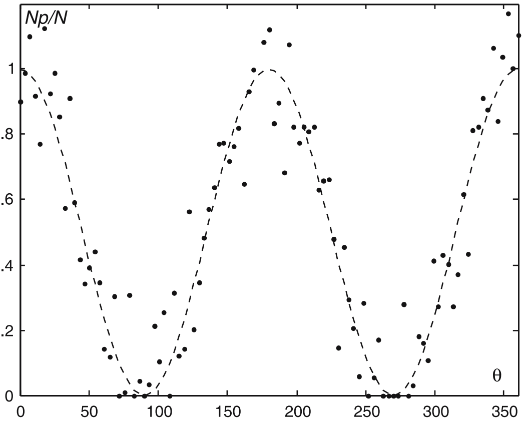


Fig. 7.3. Average number of photons passing through a polarizer for several polarizer angles (simulation)

Another strange aspect of quantum mechanics, which still raises much debate, concerns the transition between a quantum description and a classical description. In other words, when is a particle large enough to ensure that quantum behavior is no longer observed? The modern decoherence theory presents an explanation consistent with experimental results, according to which for sufficiently large particles, where sufficient means that an interaction with other particles in the surrounding space is inevitable, the state superposition is no longer preserved.

The intrinsically random nature of quantum phenomena was a source of embarrassment for many famous physicists. It was even suggested that instead of probability waves, there existed in the microscopic particles some internal variables, called *hidden variables*, that were deterministically responsible for quantum phenomena. However, a result discovered in 1964 by the physicist John Bell (known as Bell's theorem) proved that no theory of local hidden variables would be able to reproduce in a consistent way the probabilistic outcomes of quantum mechanics. The random nature of quanta is therefore the stronghold of *true randomness*, of true chance. Quanta are the quintessential chance generators, putting dice, coins and the like to shame.

7.3 A Planetary Game

Playing dice in space may not be a very entertaining activity. In weightless conditions, astronauts would see the dice floating in the air, without much hope of deciding what face to read off. Imagine a space game for astronauts, using balls instead of dice, and making use of the weightless conditions in order to obtain orbital ball trajectories. One of the balls, of largest mass, plays the role of the Sun in a mini-planetary system. In this planetary game two balls are used, playing the role of planets with co-planar orbits.

Let us first take a single planet P_1 and to fix ideas let us suppose that the Sun and planet have masses 10 kg and 1 kg, respectively. The mass of the planet is reasonably smaller than the mass of the Sun so that its influence on the Sun's position can be neglected. We may then consider that the Sun is fixed at the point $(0, 0)$ of a coordinate system with orthogonal axes labelled x and y in the orbital plane. If an astronaut places P_1 1.5 m away from the Sun with an initial velocity of 7 cm/hr along one of the orthogonal directions, (s)he will be able to confirm the universal law of gravitation, observing P_1 moving round in a slightly elliptic orbit with an orbital period of about 124 hr (assuming that the astronaut is patient enough). We now suppose that a second planet P_2, also with mass 1 kg, is placed farther away from the Sun than P_1 and with an initial velocity of 6 cm/hr along one of the orthogonal directions. The game consists in predicting where (to within some tolerance) one of the planets can be found after a certain number of hours. If P_2 is placed at an initial distance of $r = 3$ m from the Sun, we get the stable orbit situation shown in Fig. 7.4a and the position of the planets is predictable. In this situation a slight deviation in the position of P_2 does not change the orbits that much, as shown in Fig. 7.4b. On the other hand, if P_2 is placed 2 m from the Sun, we get the chaotic orbits shown in Fig. 7.5a and the positions of the planets can no longer be predicted. Even as tiny a change in the position as an increment of 0.00000001 m (a microbial increment!) causes a large perturbation, as shown in Fig. 7.5b. One cannot even predict the answer to such a simple question as: Will P_1 have made an excursion beyond 3 m after 120 hr? Faced with such a question, the astronauts are in the same situation as if they were predicting the results of a coin-tossing experiment.

The chaotic behavior of this so-called three-body problem was first studied by the French mathematician Jules Henri Poincaré (1854–1912), pioneer in the study of chaotic phenomena. Poincaré and other researchers have shown that, by randomly specifying the initial conditions of three or more bodies, the resulting motion is *almost always* chaotic, in contrast to what happens with two bodies, where the motion is never chaotic. For three or more bodies no one can tell, in general, which trajectories they will follow. This is the situation with certain

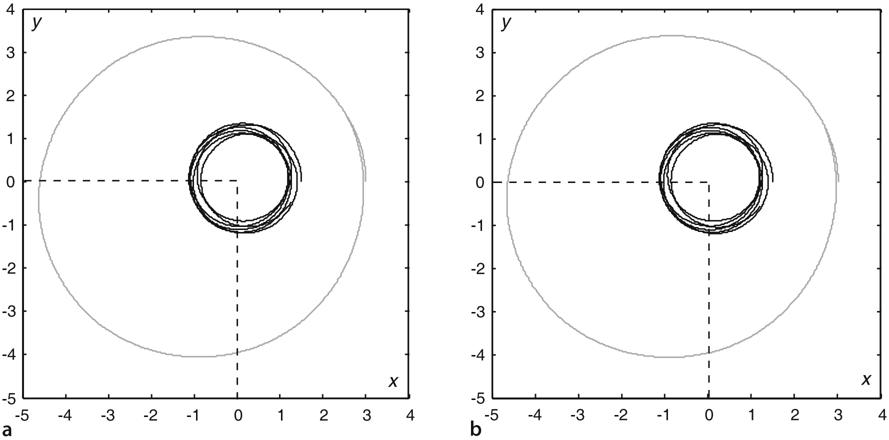


Fig. 7.4. Stable orbits of the planets, with the 'Sun' at (0, 0). (**a**) $r = 3$ m. (**b**) $r = 3.01$ m

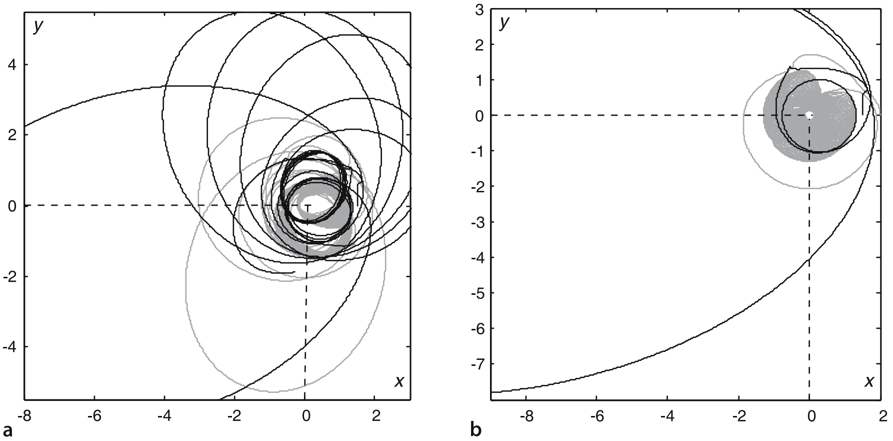


Fig. 7.5. Chaotic orbits of the planets, with the 'Sun' at $(0,0)$. (**a**) $r = 2$ m. (**b**) $r = 2.000\,000\,01$ m

asteroids. The problem can only be treated for certain specific sets of conditions. In the following sections, we shall try to cast some light on the nature of these phenomena.

7.4 Almost Sure Interest Rates

Consider the calculation of accumulated capital after n years, in a certain bank with an annual constant interest rate of j. Assume an initial capital, denoted by C_0, of 1 000 euro, and an interest rate of 3% per year. After the first year the accumulated capital C_1 is

$$C_1 = C_0 + C_0 \times j = 1\,000 + 1\,000 \times 0.03 = C_0 \times 1.03\ .$$

After the second year we get in the same way, $C_2 = C_1 \times 1.03 = C_0 \times 1.03^2$. Thus, after n years, iteration of the preceding rule shows that the capital is $C_n = C_0 \times 1.03^n$.

After carrying out the computations with a certain initial capital C_0, suppose we came to the conclusion that its value was wrong and we should have used $C_0 + E_0$. For instance, instead of 1 000 euro we should have used 1 001 euro, corresponding to a relative error of $E_0/C_0 = 1/1\,000$. Do we have to redo all calculations or is there a simple rule for determining how the initial $E_0 = 1$ euro error propagates after n years? Let us see. After the first year the capital becomes

$$C_0 \times 1.03 + E_0 \times 1.03\ .$$

That is, we get a difference relative to C_1 of $E_0 \times 1.03$. The relative error (relative to C_1) is again E_0/C_0 (in detail, $E_0 \times 1.03/C_0 \times 1.03$). Repeating the same reasoning, it is easy to conclude that the initial error E_0 propagates in a fairly simple way over the years: the propagation is such that the relative error remains *constant*. If we suppose a scenario with $n = 10$ years and imagine that we had computed $C_{10} = 1\,343.92$ euro, then given the preceding rule, the relative error $1/1\,000$ will remain constant and we do not therefore need to redo the whole calculation; we just have to add to C_{10} one thousandth of its value in order to obtain the correct value. Note how this simple rule is the consequence of a *linear relation* between the capital C_n after a given number of years and the capital C_{n+1} after one further year: $C_{n+1} = C_n \times 1.03$. (In a graph of C_{n+1} against C_n, this relation corresponds to a straight line passing through the origin and with slope 1.03, which is what justifies use of the word 'linear'.)

Let us now look at another aspect of capital growth. Suppose that another bank offers an interest rate of 3.05% per year. Is this small difference very important? Let us see by what factor the capital grows in the two cases after 10 years:

$$\begin{aligned} 3\% \text{ interest rate after 10 years:} &\quad 1.343\,916\ , \\ 3.05\% \text{ interest rate after 10 years:} &\quad 1.350\,455\ . \end{aligned}$$

If the initial capital is 100 euro, we get a difference that many of us would consider negligible, namely 65 cents. But, if the initial capital is one million euros, the difference will no doubt look quite important: 6 538 euro. This is of course a trivial observation and it seems that not much remains to be said about it. But let us take a closer look at how the difference of capital growth evolves when the interest rates are close to each other. As n increases, that difference also increases. Figure 7.6 shows what happens for two truly usurious interest rates: 50% and 51%. The vertical axis shows the capital C_{n+1} after one further year, corresponding to the present capital C_n, represented on the horizontal axis: $C_{n+1} = C_n(1 + j)$. The graphical construction is straightforward. Starting with $C_0 = 1$, we extend this value vertically until we meet the straight line $1 + j$, in this case the straight line with slope 1.5 or 1.51. We thus obtain the value C_1, which we extend horizontally until we reach the line at 45°, so that we use the same C_1 value on the horizontal axis when doing the next iteration. The process is repeated over and over again. The solid staircase shows how the capital grows for the 51% interest rate. One sees that as the number of years increases

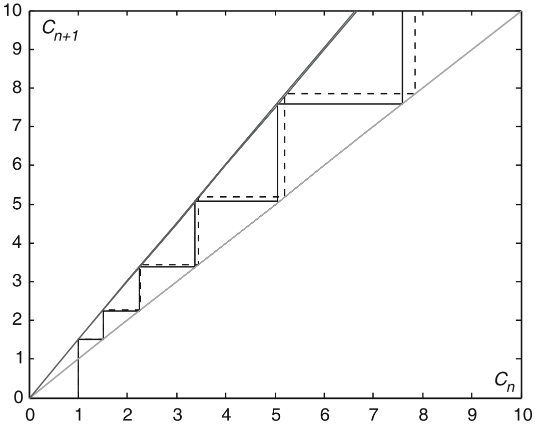


Fig. 7.6. Capital growth for (usurious!) interest rates: 50% and 51%

(see Fig. 7.6) the difference between the two capital growth factors increases in proportion to the number of years. In fact, Fig. 7.6 clearly illustrates that the difference in values of the capital is proportional to n, i.e., changes linearly with n. For the 3% and 3.05% interest rates, the difference in values of the capital varies with n as a straight line with slope $(0.030\,5 - 0.03)/1.03 = 0.000\,485$ per year.

This linearly growing difference in capital for distinct (and constant) interest rates is also well known. However, what most people do not realize is that numerical computations are always performed with *finite accuracy*. That is, it does not matter how the calculations are carried out, e.g., by hand or by computer, the number of digits used in the numerical representation is necessarily finite. Therefore, the results are *generally* approximate. For instance, we never get exact calculations for an interest rate of 1/30 per year, since $1/30 = 0.033\,333\,3\ldots$ (continuing infinitely).

Banks compute accumulated capitals with computers. These may use different numbers of bits in their calculations, corresponding to different numbers of decimal digits. Personal computers usually perform the calculations using 32 bits (the so-called simple precision representation, equivalent to about 7 significant decimal digits) and in 64 bits (the so-called double precision representation, equivalent to about 15 significant decimal digits). Imagine a firm deciding to invest one million euros in a portfolio of shares with an 8% annual interest rate. Applying this constant 8% interest rate the firm's accountant calculates the accumulated capital on the firm's computer with 32-bit precision. The bank performs the same calculation but with 64-bit precision. They compare their calculations and find that after 25 years there is a 41

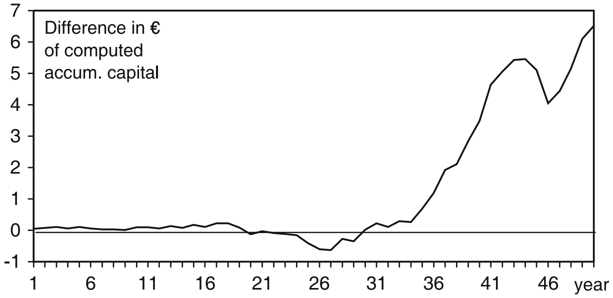


Fig. 7.7. Annual evolution of the difference of accumulated capital computed by two different computers

cent difference in an accumulated capital of about six million eight hundred thousand euros. In order to clear up this issue they compare the respective calculations for a 50-year period and draw the graph of the differences between the two calculations, as in Fig. 7.7. After 50 years the difference is already 6.51 euro. Moreover, they verify that the difference does not progress in a systematic, linear way; it shows chaotic fluctuations that are difficult to forecast. If there is a noticeable difference when using a 64-bit computer, why not use a 128-bit or a 256-bit computer instead? In fact, for problems where numerical precision is crucial one may use supercomputers that allow a very large number of bits for their computations. But the bottom line is this: no matter how super the computer is, we will always have to deal with uncertainty factors in our computations due to the finite precision of number representation and error propagation. In some cases the mentioned uncertainty is insignificant (the firm would probably not pay much attention to a 41 cent difference in six million eight hundred thousand euros). In other cases error propagation can have dramatic consequences.

7.5 Populations in Crisis

In the above interest rate problem, the formula $C_{n+1} = C_n(1 + j)$ for capital growth is a linear formula, i.e., C_{n+1} is directly proportional to C_n. This linear dependency is reflected in the well-behaved evolution of the capital growth for a small 'perturbation' of the initial capital. Unfortunately, linear behavior is more the exception than the rule, either in natural phenomena or in everyday life.

We now consider a non-linear phenomenon, relating to the evolution of the number of individuals in a population after n time units. Let us denote the number of individuals by N_n and let N be the maximum reference value allowed by the available food resources. We may, for instance, imagine a population of rabbits living inside a fenced field. If our time unit is the month, we may further assume that after three months we have $N_3 = 60$ rabbits living in a field that can only support up to $N = 300$ rabbits. Let $P_n = N_n/N$ denote the fraction of the maximum number of individuals that can be supported by the available resources. For the above-mentioned values, we have $P_3 = 0.2$. If the population grew at a constant rate r, we would get a situation similar to the one for the interest rate: $P_{n+1} = P_n(1 + r)$. However, with an increasing population, the necessary resources become scarcer and one may reach a point where mortality overtakes natality owing to lack of resources; P_n will then decrease. A formula reflecting the need to impose a growth rate that decreases as one approaches a maximum number of individuals, becoming a dwindle rate if the maximum is exceeded, is

$$P_{n+1} = P_n\big[1 + r(1 - P_n)\big] \; .$$

According to this rule the population grows or dwindles at the rate of $r(1-P_n)$. If P_n is below 1 (the P_n value for the population maximum), the population grows, and all the more as the deviation from the maximum is greater. As soon as P_n exceeds 1, a dwindling phase sets in. Note that the above formula can be written as $P_{n+1} = P_n(1+r) - rP_n^2$, clearly showing a non-linear, quadratic, dependency of P_{n+1} on P_n.

Suppose we start with $P_0 = 0.01$ (one per cent of the reference maximum) and study the population evolution for $r = 1.9$. We find that the population keeps on growing until it stabilizes at the maximum (Fig. 7.8a). If $r = 2.5$, we get an oscillating pattern (Fig. 7.8b) with period 4 (months).

Sometimes the periodic behavior is more complex than the one in Fig. 7.8b. For instance, in the case of Fig. 7.9, the period is 32 and it takes about 150 time units until a stable oscillation is reached. If we set a 0.75 threshold, as in Fig. 7.9, interpreting as 1 the values above the threshold and as 0 the values below, we get the following periodic sequence:

$$111010101110101011101010111011101110101011101010111\ldots$$

But strange things start to happen when r approaches 3. For $r = 3$ and $P_0 = 0.1$, we get the behavior illustrated in Fig. 7.10a. A very tiny error in the initial condition P_0, for instance $P_0 = 0.100\,001$, is enough to trigger a totally different evolution, as shown in Fig. 7.10b.

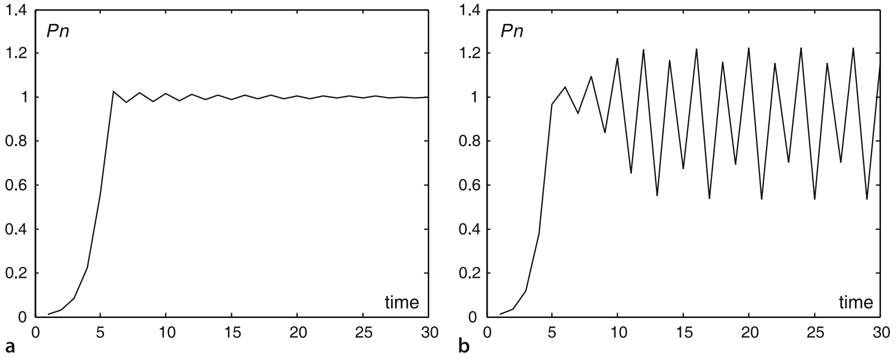


Fig. 7.8. Population evolution starting from $P_0 = 0.01$ and with two different growth rates. (**a**) $r = 1.9$. (**b**) $r = 2.5$

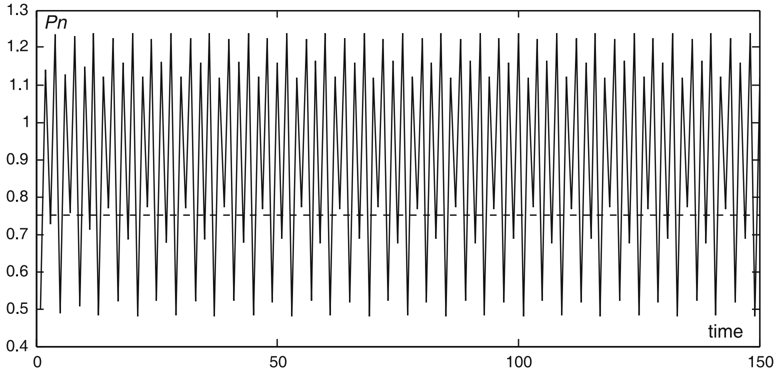


Fig. 7.9. Population evolution starting from 0.5 and with $r = 2.569$

This is a completely different scenario from the error propagation issue in the interest rate problem. There, a small deviation in the initial conditions was propagated in a controlled, linear, way. In the present case, as in the planetary game, error propagation is completely chaotic.

What we have here is not only a problem of high sensitivity to changes in initial conditions. The crux of the problem is that the law governing such sensitivity is unknown! To be precise, the graphs of Figs. 7.5 and 7.10 were those obtained on *my* computer. On this basis, the reader might say, for instance, that the population for $r = 3$ and $P_0 = 0.1$ reaches a local maximum at the 60th generation. However, such a statement does not make sense. Given the high sensitivity to initial conditions, only by using an infinitely large number of digits would one be able to reach a conclusion as to whether or not the value attained at the 60th generation is a local maximum. Figure 7.11 shows

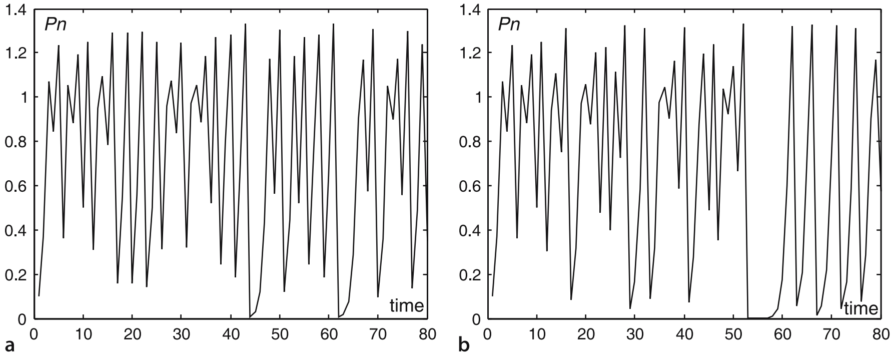


Fig. 7.10. Chaotic population evolution for $r = 3$. (**a**) $P_0 = 0.1$. (**b**) $P_0 = 0.100\,001$

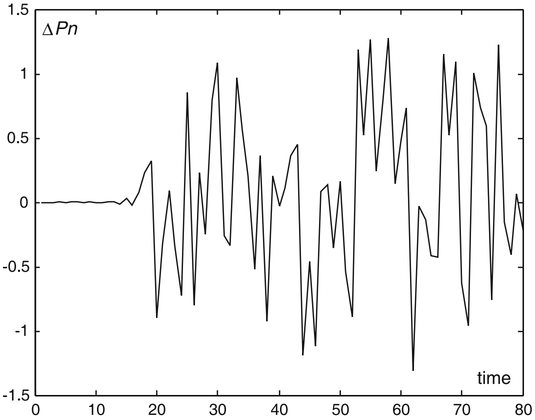


Fig. 7.11. Time evolution of the difference in the Fig. 7.10 populations

the difference of P_n values between the evolutions shown in Fig. 7.10. It is clear that after a certain time no rule allows us to say how the initial perturbation evolves with n. This is the nature of *deterministic chaos*. In the case of the interest rate, we had no chaos and a very simple rule was applied to a 'perturbation' of the initial capital: the perturbation was propagated in such a way that the relative differences were kept constant. In deterministic chaos there are no rules. The best one can do is to study *statistically* how the perturbations are propagated, on average.

7.6 An Innocent Formula

There is a very simple formula that reflects a quadratic dependency of the next value of a given variable x relative to its present value. It is called a quadratic iterator and is written as follows:

$$x_{n+1} = ax_n(1 - x_n) \ .$$

The name 'iterator' comes from the fact that the same formula is applied repeatedly to any present value x_n (as in the previous formulas). Since the formula can be rewritten $x_{n+1} = ax_n - ax_n^2$, the quadratic dependency becomes evident. (We could say that the formula for the interest rate is the formula for a linear iterator.)

The formula is so simple that one might expect exemplary deterministic behavior. Nevertheless, and contrary to one's expectations, the quadratic iterator allows a simple and convincing illustration of the deterministic chaos that characterizes many non-linear phenomena. For instance, expressing $a = r + 1$ and $x_n = rP_n/(r + 1)$, one obtains the preceding formula for the population growth.

Let us analyze the behavior of the quadratic iterator as in the case of the interest rate and population growth. For this purpose we start by noticing that the function $y = ax(1 - x)$ describes a parabola passing through the points $(0, 0)$ and $(0, 1)$, as shown in Fig. 7.12a. The parameter a influences the height of the parabola vertex occurring at $x = 0.5$ with value $a/4$. If a does not exceed 4, any value of x in the interval $[0, 1]$ produces a value of y in the same interval.

Figure 7.12a shows the graph of the iterations, in the same way as for the interest rate problem, for $a = 2.5$ and the initial value $x_0 = 0.15$. Successive iterations generate the following sequence of values: 0.15, 0.3187, 0.5429, 0.6204, 0.5888, 0.6053, 0.5973, 0.6013, 0.5993,

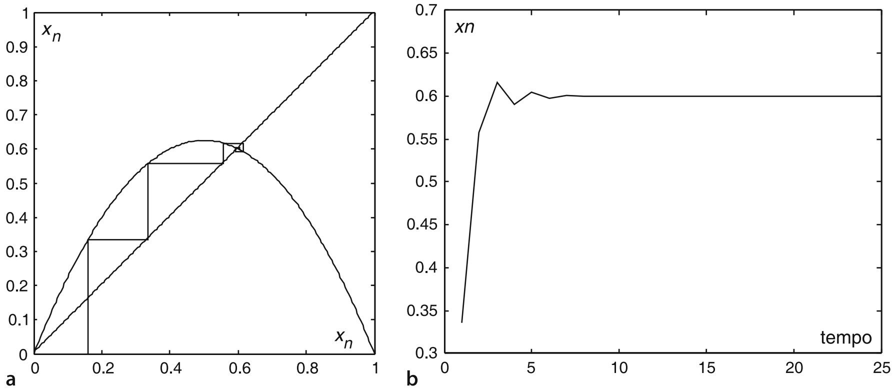


Fig. 7.12. Quadratic iterator for $a = 2.5$ and $x_0 = 0.15$. (**a**) Graph of the iterations. (**b**) Orbit

0.6003, 0.5998, and so on. This sequence, converging to 0.6, is shown in Fig. 7.12b. As in the previous examples we may assume that the iterations occur in time. The graph of Fig. 7.12b is said to represent one *orbit* of the iterator. If we introduce a small perturbation of the initial value (for instance, $x_0 = 0.14$ or $x_0 = 0.16$), the sequence still converges to 0.6.

Figure 7.13 shows a non-convergent chaotic situation for $a = 4$. A slight perturbation of the initial value produces very different orbits, as shown in Fig. 7.14. Between the convergent case and the chaotic case the quadratic iterator also exhibits periodic behavior, as we had already detected in the population growth model.

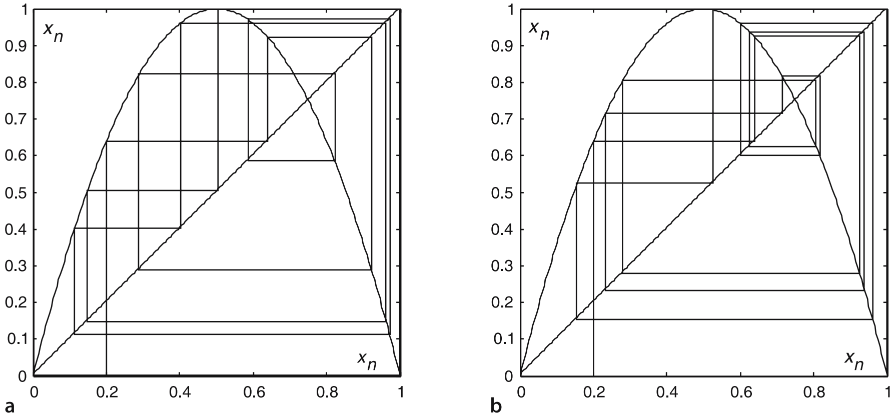


Fig. 7.13. Fifteen iterations of the quadratic iterator for $a = 4$. (**a**) $x_0 = 0.2$. (**b**) $x_0 = 0.199$

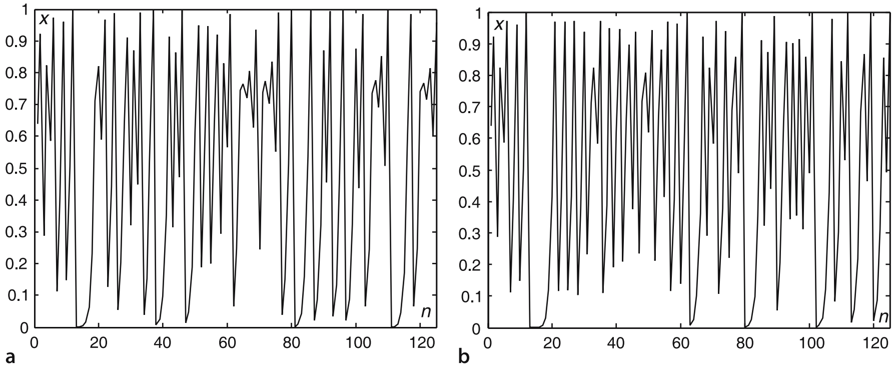


Fig. 7.14. Quadratic iterator orbits for $a = 4$. (**a**) $x_0 = 0.2$. (**b**) $x_0 = 0.199999$

Now suppose we have divided the $[0, 1]$ interval into small subintervals of width 0.01 and that we try to find the frequency with which any of these subintervals is visited by the quadratic iterator orbit, for $a = 4$ and $x_0 = 0.2$. Figure 7.15 shows what happens after 500 000 iterations. As the number of iterations grows the graph obtained does not change when we specify different initial values x_0. Interpreting frequencies as probabilities, we see that the probability of obtaining a visit of a subinterval near 0 or 1 is rather larger than when it is near 0.5. In fact, it is possible to prove that for infinitesimal subintervals one would obtain the previously mentioned arcsine law (see Chap. 6) for the frequency of visits of an arbitrary point. That is, the probability of our chaotic orbit visiting a point x is equal to the probability of the last zero gain having occurred before time x in a coin-tossing game!

Consider the following variable transformation for the quadratic iterator:

$$x_n = \sin^2(\pi y_n) \ .$$

For $a = 4$, it can be shown that this transformation allows one to rewrite the iterator formula as

$$y_{n+1} = 2y_n \pmod 1 \ ,$$

where the mod 1 (read modulo 1) operator means removing the integer part of the expression to which it applies ($2y_n$ in this case). The chaotic orbits of the quadratic iterator are therefore mapped onto chaotic orbits of y_n. We may also express the y_n value relative to an initial value y_0. As in the case of the interest rate, we get

$$y_n = 2^n y_0 \pmod 1 \ .$$

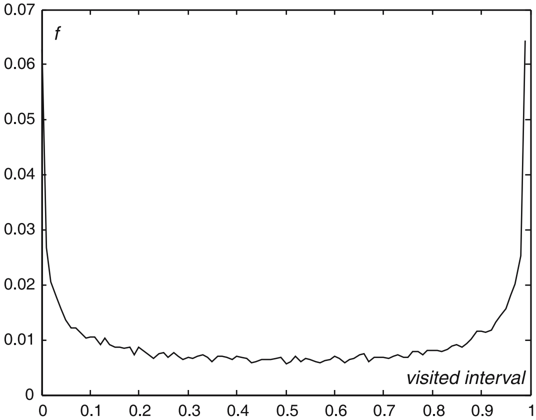


Fig. 7.15. Probability of the chaotic quadratic iterator ($a = 4$) visiting a subinterval of $[0, 1]$ of width 0.01, estimated in 500 000 iterations

Let us see what this equation means, by considering y_0 expressed in binary numbering, for instance $y_0 = 0.101\,100\,100\,010\,111\,1$. In the first iteration we have to multiply by 2 and remove the integer part. In binary numbering, multiplying by 2 means a one-bit leftward shift of the whole number, with a zero carried in at the right (similarly to what happens when we multiply by 10 in decimal numbering). We get $1.011\,001\,000\,101\,111\,0$. Removing the integer part we get $0.011\,001\,000\,101\,111\,0$. The same operations are performed in the following iterations. Suppose we only look at the successive bits of the removed integer parts. What the iterations of the above formula then produce are the consecutive bits of the initial value, which can be interpreted as an infinite sequence of coin throws. Now, we already know that, with probability 1, a Borel normal number can be obtained by random choice in $[0, 1]$. As a matter of fact, we shall even see in Chap. 9 that, with probability 1, an initial value y_0 randomly chosen in $[0, 1]$ turns out to be a number satisfying a strict randomness criterion, whose bits cannot be predicted either by experiment or by theory. Chaos already exists in the initial y_0 value! The quadratic iterator does not compute anything; it only reproduces the bits of y_0. The main issue here is not an infinite precision issue. The main issue is that we do not have a means of producing the bits of the sequence other than having a prestored list of them (we shall return to this point in Chap. 9). However, not all equations of the above type produce chaotic sequences, in the sense that they do not propagate chaos to the following iterations. For instance, the equation $x_{n+1} = x_n + b \pmod 1$ does not propagate chaos into the following iteration. We see this because, if x_0 and x'_0 are very close initial conditions, after n iterations we will get $x'_n - x_n = x'_0 - x_0$. That is, the difference of initial values continues indefinitely.

7.7 Chance in Natural Systems

Many chance phenomena in natural systems composed of macroscopic objects are of a chaotic nature. These systems generally involve several variables to characterize the state of the system as time goes by, rather than one as in the example of population evolution:

- for the 3-sphere system described above, the characterizing variables are the instantaneous values of two or three components of position and velocity,
- for a container where certain reagents are poured, the characterizing variables are the concentrations of reagents and reaction products,

- for weather forecasting, the characterizing variables are the pressure and temperature gradients,
- for cell culture evolution the relevant variables may be nutrient and cell concentrations,
- for financial investments, they may be rental rates and other market variables,

and so on and so forth.

Let x denote a variable characterizing the system and let x_n and x_{n+1} represent the values at consecutive instants of time, assumed close enough to each other to ensure that the difference $x_{n+1} - x_n$ reflects the instantaneous time evolution in a sufficiently accurate way. If these differences vary linearly with x_n, no chaotic phenomena take place. Chaos in the initial conditions does not propagate. Chaos requires non-linearities in order to propagate, as was the case for the quadratic dependence in the population growth example. Now it just happens that the time evolution in a huge range of natural systems depends in a non-linear way on the variables characterizing the system. Consider the example of a sphere orbiting in the xy plane around another sphere of much larger mass placed at $(0, 0)$. Due to the law of gravitation, the time evolution of the sphere's coordinates depends on the reciprocal of its distance $\sqrt{x^2 + y^2}$ to the center. We thus get a non-linear dependency. In the case of a single orbiting sphere no chaotic behavior will arise. However, by adding a new sphere to the system we are adding an extra non-linear term and, as we have seen, chaotic orbits may arise.

The French mathematician Poincaré carried out pioneering work on the time evolution of non-linear systems, a scientific discipline which has attracted great interest not only in physics and chemistry, but also in the fields of biology, physiology, economics, meteorology and others. Many non-linear systems exhibit chaotic evolution under certain conditions. There is an extensive popular science literature describing the topic of chaos, where many bizarre kinds of behavior are presented and characterized. What interests us here is the fact that a system with chaotic behavior is a chance generator (like tossing coins or dice), in the sense that it is not possible to predict the values of the variables. The only thing one can do is to characterize the system variables probabilistically.

An example of a chaotic chance generator is given by the thermal convection of the so-called Bénard cells. The generator consists of a thin layer of liquid between two glass plates. The system is placed over a heat source. When the temperature of the plate directly in contact with the heat source reaches a certain critical value, small domains are formed in which the liquid layer rotates by thermal convection. These are the so-called Bénard cells. The direction of rotation of the cells (clockwise or counterclockwise) is unpredictable and depends on microscopic perturbations. From this point of view, one may consider that the rotational state of a Bénard cell is a possible substitute for coin tossing. If after obtaining the convection cells we keep on raising the temperature, a new critical value is eventually reached where the fluid motion becomes chaotic (turbulent), corresponding to an even more complex chance generator. Many natural phenomena exhibit some similarity with the Bénard cells, e.g., meteorological phenomena.

Meteorologists are perfectly familiar with turbulent atmospheric phenomena and with their hypersensitivity to small changes in certain conditions. In fact, it was a meteorologist (and mathematician) Edward Lorenz who discovered in 1956 the lack of predictability of certain deterministic systems, when he noticed that weather forecasts based on certain computer-implemented models were highly sensitive to tiny perturbations (initially Lorenz mistrusted the computer and repeated the calculations with several machines!). As Lorenz says in one of his writings: "One meteorologist remarked that if the theory were correct, one flap of a seagull's wings could change the course of weather forever." This high sensitivity to initial conditions afterwards came to be called the butterfly effect.

The Lorenz equations, which are also applied to the study of convection in Bénard cells, are

$$x_{n+1} - x_n = \sigma(y_n - x_n) \;, \qquad y_{n+1} - y_n = x_n(\tau - z_n) - y_n \;,$$

$$z_{n+1} - z_n = x_n y_n - \beta z_n \;.$$

In these equations, x represents the intensity of the convective motion, y the temperature difference between ascending and descending fluid, and z the deviation from linearity of the vertical temperature profile. Note the presence of the non-linear terms $x_n z_n$ and $x_n y_n$, responsible for chaotic behavior for certain values of the control parameters, as illustrated in Fig. 7.16. This three-dimensional orbit is known as the Lorenz attractor.

Progress in other studies of chaotic phenomena has allowed the identification of many other systems of inanimate nature with chaotic behavior, e.g., variations in the volume of ice covering the Earth, vari-

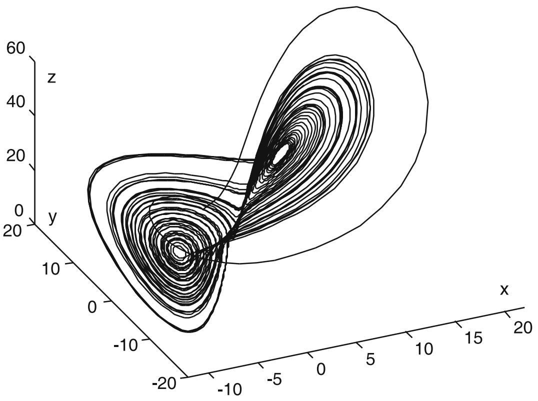


Fig. 7.16. Chaotic orbit of Lorenz equations for $\sigma = 10$, $\tau = 28$, $\beta = 8/3$

ations in the Earth's magnetic field, river flows, turbulent fluid motion, water dripping from a tap, spring vibrations subject to non-linear forces, and asteroid orbits.

7.8 Chance in Life

While the existence of chaos in inanimate nature may not come as a surprise, the same cannot be said when we observe chaos in nature. In fact, life is ruled by laws that tend to maintain equilibrium of the biological variables involved, in order to guarantee the development of metabolic processes compatible with life itself. For instance, if we have been running and we are then allowed to rest, the heart rate which increased during the run goes back to its normal value after some time; concomitantly, the energy production mode that was adjusted to suit the greater muscular requirements falls back into the normal production mode which prevails in the absence of muscular stress. One then observes that the physiological variables (as well as the biological variables at cellular level) fall back to the equilibrium situation corresponding to an absence of effort. If on the other hand we were not allowed to rest and were instead obliged to run indefinitely, there would come a point of death by exhaustion, when it would no longer be possible to sustain the metabolic processes in a way compatible with life.

Life thus demands the maintenance of certain basic balances and one might think that such balances demand a perfectly controlled, non-chaotic, evolution of the physiological and biological variables. Surprisingly enough this is not the case. Consider the heart rate of a resting adult human. One might think that the heart rate should stay at a practically constant value or at most exhibit simple, non-chaotic fluctuations. However, it is observed that the heart rate of a healthy adult (in fact, of any healthy animal and in any stage of life) exhibits a chaotic behavior, even at rest, as shown in Fig. 7.17. Heart rate variability is due to two factors: the pacemaker and the autonomic nervous system. The pacemaker consists of a set of self-exciting cells in the right atrium of the heart. It is therefore an oscillator that maintains a practically constant frequency in the absence of stimuli from the autonomic nervous system. The latter is responsible for maintaining certain involuntary internal functions approximately constant. Beside the heart beat, these include digestion, breathing, and the metabolism. Although these functions are usually outside voluntary control they can be influenced by the mental state. The autonomic nervous system is composed of two subsystems: the sympathetic and the parasympathetic. The first is responsible for triggering responses appropriate to stress situations, such as fear or extremes of physical activity. Increase in heart rate is one of the elicited responses. The parasympathetic subsystem is responsible for triggering energy saving and recovery actions after stress. A decrease in heart rate is one of the elicited actions.

The variability one observes in the heart rate is therefore influenced by the randomness of external phenomena causing stress situations. However, it has also been observed that there is a chaotic behavior of the heart rate associated with the pacemaker itself, whose causes are not yet fully understood. (However, some researchers have been able to explain the mechanism leading to the development of chaotic phenomena in heart cells of chicken, stimulated periodically.) Whatever its causes, this chaotic heart rate behavior attenuates with age, reflecting the heart's loss of flexibility in its adaptation to internal and external stimuli.

Chaotic phenomena can also be observed in electrical impulses transmitted by nervous cells, the so-called action potentials. An action potential is nothing other than an electrical impulse sent by a neuron to

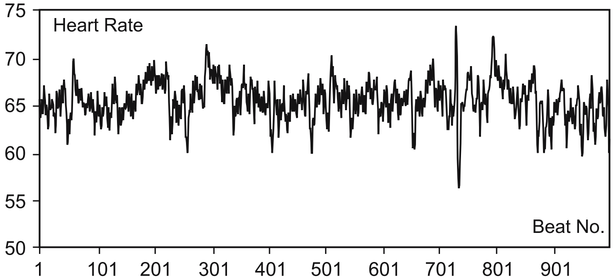


Fig. 7.17. Heart rate of an adult man. The frequency is expressed in beats per minute

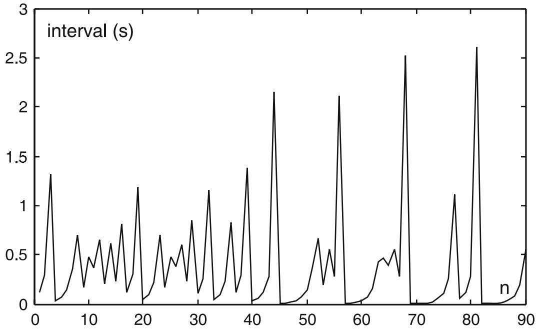


Fig. 7.18. Evolution of successive action potentials in nervous cells ($t_0 = 0.1$ s, $R = 3\ \mathrm{s}^{-1}$)

other neurons. These impulses have an amplitude of about 0.1 volt and last about 2 milliseconds. Two researchers (Glass and Mackey in 1988) were able to present a deterministic model for the time intervals t_n between successive peaks of action potentials. The model is non-linear and stipulates that the following inter-peak interval is proportional to the logarithm of the absolute value of $1 - 2\mathrm{e}^{-Rt_n}$, where R is a control parameter of the model. This model exhibits chaotic behavior of the kind shown in Fig. 7.18 for certain values of the control parameter, and this behavior has in fact been observed in practice.

Many other physiological phenomena exhibit chaotic behavior, e.g., electrical potentials in cell membranes, breathing tidal volumes, vascular movements, metabolic phenomena like glucose conversion in adenosine triphosphate, and electroencephalographic signals.

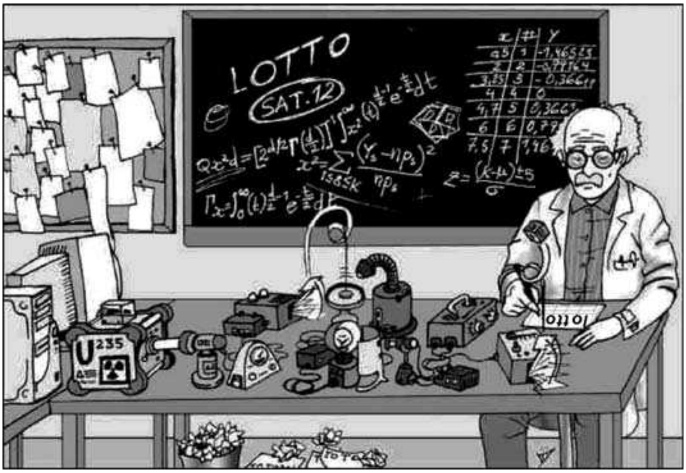


Fig. 7.19. Will it work?

8

Noisy Irregularities

8.1 Generators and Sequences

Imagine that you were presented with two closed boxes named C_1 and C_2, both producing a tape with a sequence of sixteen zeros and ones whenever a button is pressed. The reader is informed that one of the boxes uses a random generator to produce the tape, while the other box uses a periodic generator. The reader's task is to determine, using only the tapes produced by the boxes, which one, C_1 or C_2, contains the random generator. The reader presses the button and obtains:

$$\begin{aligned} C_1 \text{ sequence :} &\quad 0011001100110011 \;, \\ C_2 \text{ sequence :} &\quad 1101010111010101 \;. \end{aligned}$$

The reader then decides that box C_2 contains the random generator. The boxes are opened and to the reader's amazement the random generator is in box C_1 (where, for instance, a mechanical device tosses a coin) whereas box C_2 contains the periodic generator corresponding to Fig. 7.9. This example illustrates the *impossibility* of deciding the nature (deterministic or random) of a sequence generator on the basis of observation, and in fact on the basis of any randomness measurement, of one of its sequences (necessarily finite). For the moment the 'measure' is only subjective. The reader decides upon C_2 because the corresponding sequence looks irregular. Later on, we present a few objective methods for measuring sequence irregularity. Note that segments of periodic sequences characterized by long and complex periods tend to be considered as random, independently of whatever method we choose to measure irregularity. For instance, the following sequence:

8, 43, 26, 5, 28, 15, 14, 9, 48, 51, 2, 13, 4, 23, 54, 17, 24, 59, 42, 21, 44, 31, 30, 25, 0, 3, 18, 29, 20, 39, 6, 33, 40, 11, 58, 37, 60, 47, 46, 41, 16, 19, 34, 45, 36, 55, 22, 49, 56, 27, 10, 53, 12, 63, 62, 57, 32, 35, 50, 61, 52, 7, 381, 8, 43, 26, 5, 28, 15, 14, 9, 48, 51, 2, 13, 4, 23, 54, 17, 24, 59, 42, 21, 44, 31, 30, 25, 0, 3, 18, 29, 20, 39, 6, 33, 40, 11, 58, 37

is generated by a periodic generator of period 64 of the population evolution type described in Chap. 7. It is difficult, however, to perceive the existence of periodicity. If we convert the sequence into binary numbers, viz.,

```
001000101011011010000101011100001111001110001001110000
110011000010001101000100010111110110010001011000111011
101010010101101100011111011110011001000000000011010010
011101010100100111000110100001101000001011111010100101
111100101111101110101001010000010011100010101101100100
110111010110110001111000011011001010110101001100111111
111110111001100000100011110010111101110100000111100110
000001001000101011011010000101011100001111001110001001
110000110011000010001101000100010111110110010001011000
111011101010010101101100011111011110011001000000000011
010010011101010100100111000110100001101000001011111010
100101
```

it becomes even more difficult.

Now, consider sequences generated by random generators such as coin-tossing. We may, of course, obtain sequences such as the one generated by C_1 or other, even more anomalous sequences, such as 0000000000000000 or 0000000011111111. In fact, as we know already, any sequence of n bits has the same probability of occurring, namely $1/2^n$. We also know that only after generating a large number of sequences of n bits can we expect to elucidate the nature of the generator. If the sequences are generated by a coin-tossing process, we expect about 95% of the generated sequences to exhibit average values S/n within $0.5 \pm 1/\sqrt{n}$. For $n = 16$, this interval is 0.5 ± 0.25. For instance, the sequences 0000000000000000 ($S/n = 0$) and 0111111111111110 ($S/n = 0.87$) are outside that interval. If more than 5% of the sequences are outside that interval, we then have strong doubts (based on the usual 95% confidence level) that our generator is random. Moreover, if the generator produces independent outcomes – a usual assumption in fair coin-tossing – we then hope to find an equal rate of occurrence of 00, 01, 10 and 11 in a large number of sequences. The same applies to all distinct, equal-length subsequences. It is also possible in this case to establish confidence intervals that would exclude sequences such as 0111111111111110, where 01 and 10 only occur once and 00 never occurs. There are also statistical tests that allow one to reject the randomness assumption, with a certain degree of certainty, when the number of equal-value subsequences (called *runs*) is either too large – as in 0100110011001100, exhibiting 9 equal-value subsequences –, or too small, as in 0000000011111111.

Briefly, given a large number of sequences there are statistical tests allowing one to infer something about the randomness of the generator, with a certain degree of certainty. The important point is this: the probability measure associated with the generator outcomes is only manifest in sets of a large number of experiments (laws of large numbers). Probability (generator randomness) is an ensemble property.

In daily life, however, we often cannot afford the luxury of repeating experiments. We cannot repeat the evolution of share values in the stock market; we cannot repeat the historical evolution of economic development; we cannot repeat the evolution of the universe in its first few seconds; we cannot repeat the evolution of life on Earth and assess the impact of gene mutations. We are then facing *unique sequences* of events on the basis of which we hope to judge determinism or randomness. From all that has been said, an important observation emerges: generator randomness and sequence randomness are distinct issues. From this chapter onwards we embark on a quite different topic from those discussed previously: is it possible to speak of sequence randomness? What does it mean?

8.2 Judging Randomness

How can one judge the randomness of a sequence of symbols? The Russian mathematicians Kolmogorov and Usphenski established the following criteria of sequence randomness, based on a certain historical consensus and a certain plausibility as described in probability theory:

- Typicality: the sequence must belong to a majority of sequences that do not exhibit obvious patterns.

- Chaoticity: there is no simple law governing the evolution of the sequence terms.
- Frequency stability: equal-length subsequences must exhibit the same frequency of occurrence of the different symbols.

In any case, the practical assessment of sequence randomness is based on results of statistical tests and on values of certain measures, some of which will be discussed later on. No set of these tests or measures is infallible; that is, it is always possible to come up with sequences that pass the tests or present ‘good’ randomness measures and yet do not satisfy one or another of the Kolmogorov–Usphenski criteria. Moreover, as we shall see in the next chapter, although there is a good definition of sequence randomness, there is no definite proof of randomness. In this sense, the best that can be done in practice using the available tests and measures is to judge not randomness itself, in the strict and well-defined sense presented in the next chapter, but rather to judge a certain degree of irregularity or chaoticity in the sequences.

It is an interesting fact that the human specification of random sequences is a difficult task. Psychologists have shown that the notion of randomness emerges at a late phase of human development. If after tossing a coin which turned up heads we ask a child what will come up next, the most probable answer the child will give is tails. Psychologists Maya Bar-Hillel and Willem Wagenaar performed an extensive set of experiments with volunteers who were required to write down random number sequences. They came to the conclusion that the sequences produced by humans exhibited:

- too few symmetries and too few long runs,
- too many alternating values,
- too much symbol frequency balancing in short segments,
- a tendency to think that deviations from equal likelihood confer more randomness,
- a tendency to use local representations of randomness, i.e., people believe in a sort of ‘law of small numbers’, according to which even short sequences must be representative of randomness.

It was also observed that these shortcomings are manifest, not only in lay people, but also in probability experts as well. It seems as though we each have an internal prototype of randomness; if a sequence is presented which does not fit the prototype, the most common tendency is to reject it as non-random.

8.3 Random-Looking Numbers

So-called ‘random number’ sequences are more and more often used in science and technology, where many analyses and studies of phenomena and properties of objects are based on stochastic simulations (using the Monte Carlo method, mentioned in Chap. 4). The simulation of random number sequences in computers is often based on a simple formula:

$$x_{n+1} = ax_n \mod m \ .$$

The mod symbol (read modulo) represents the remainder after integer division (in this case of ax_n by m; see also what was said about the quadratic iterator in the last chapter). Suppose we take $a = 2$, $m = 10$ and that the initial value x_0 of the sequence is 6. The following value is then 2×6 mod 10, i.e., the remainder of the integer division of 12 by 10. We thus obtain $x_1 = 2$. Successive iterations produce the periodic sequence 48624862, shown in Fig. 8.1a. Since the sequence values are given by the remainder after integer division, they necessarily belong to the interval from 0 to $m-1$ and the sequence is necessarily periodical: as soon as it reaches the initial value, in at most m iterations, the sequence repeats itself. One can obtain the maximum period by using a new version of the above formula, $x_n = (ax_{n-1} + c) \bmod m$, setting appropriate values for a and c. Figure 8.1b shows 15 iterations using the same m and initial value, with $a = 11$ and $c = 3$. The generated sequence is the following, with period 10:

$$25814703692581470369\ldots \ .$$

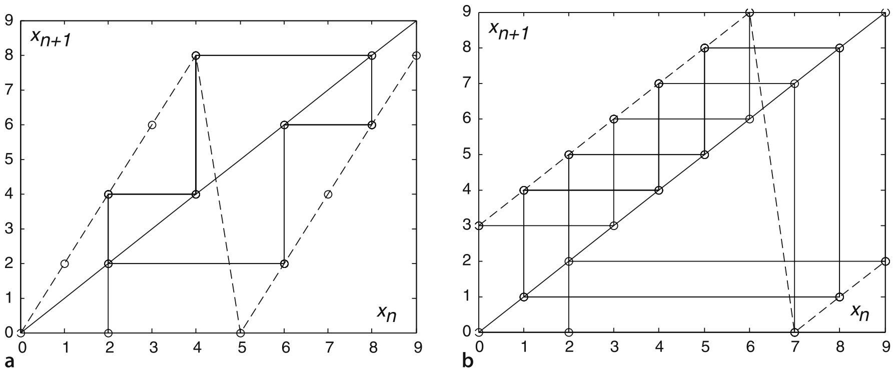


Fig. 8.1. Pseudo random-number generator starting at 2 and using $m = 10$. (**a**) Period 4, with $a = 2$, $c = 0$. (**b**) Period 10, with $a = 11$ and $c = 3$

Note how the latter sequence looks random when we observe the first 10 values, although it is deterministic and periodic. For this reason it is customary to call the sequence pseudo-random. This type of chaotic behavior induced by the mod operator is essentially of the same nature as the chaoticity observed in Fig. 7.7 for the difference in interest rates. For sufficiently large m, the periodic nature of the sequence is of little importance for a very wide range of practical applications. Figure 8.2 illustrates one such sequence obtained by the above method and with period 64. In computer applications the period is often at least as large as $2^{16} = 65\,536$ numbers.

There are other generators that use sophisticated versions of the above formula. For instance, the next iterated sequence number instead of depending only on the preceding number may depend on other past values, as in the formula

$$x_n = (ax_{n-1} + bx_{n-2} + c) \quad \mathrm{mod}\, m\ .$$

A generator of this type that has been and continues to be much used for the generation of pseudo-random binary sequences is

$$x_n = (x_{n-p} + x_{n-q}) \quad \mathrm{mod}\ 2\ , \qquad \text{with integer } p \text{ and } q\ .$$

In 2004, on the basis of a study that involved the analysis of sequences of 26 000 bits, the researchers Heiko Bauke, from the University of Magdeburg, Germany, and Stephan Mertens, from the Centre for Theoretical Physics, Trieste, Italy, showed (more than twenty years after the invention of this generator and its worldwide adoption!), that this

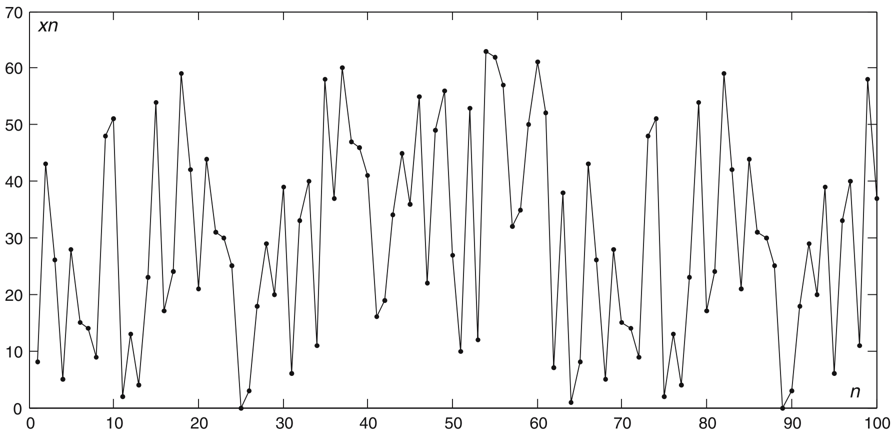


Fig. 8.2. Pseudo-random sequence of period 64 ($a = 5$, $c = 3$, $m = 64$)

Fig. 8.3. Drunk and disorderly, on a regular basis

generator was strongly biased, producing certain subsequences of zeros and ones exhibiting strong deviations from what would be expected from a true random sequence. And so a reputable random generator that had been in use for over 20 years without much suspicion was suddenly discredited.

8.4 Quantum Generators

There are some applications of random number sequences that are quite demanding as far as their randomness is concerned. One example is the use of random numbers in cryptography. Let us take the following message:

ANNARRIVESTODAY .

This message can be coded using a random sequence of letters, known as a *key*, such as KDPWWYAKWDGBBEFAHJOJJJXMLI. The procedure is as follows. We consider the letters represented by their number when in alphabetic order. For the English language the alphabet has 26 letters, which we number from 0 through 25. We then take the letters corresponding to the message and the key and add the respective order numbers. For the above message and key the first letters are respectively, A and K, with order numbers 0 and 10. The sum is 10. We next take the remainder of the integer division by 26 (the number of letters in the alphabet), which is 10. The letter of order 10 is K and it will be the first letter of the coded message. Repeating the process one obtains:

$$\text{KQCWNPIFAVZPEED} .$$

If the key is completely random each letter of the key is unpredictable and the 'opponent' does not possess any information which can be used to decipher it. If it is pseudo-random there are regularities in the key that a cryptanalyst may use in order to decipher it.

For this sort of application, requiring high quality sequence randomness, there are commercialized random number generators based on quantum phenomena. Here are some examples of quantum phenomena presently used by random number generators:

- Radioactive emission.
- Atmospheric noise.
- Thermal noise generated by passing an electric current through a resistance.
- Noise detected in the computer supply line (with thermal noise component).
- Noise detected at the input of the sound drive of a computer (essentially thermal noise).
- Arrival event detection of photons in optical fibers.

8.5 When Nothing Else Looks the Same

We mentioned the existence of statistical methods for randomness assessment, or more rigorously for assessing sequence irregularity. Methods such as the confidence interval of S/n (arithmetic average) or the runs test allow one to assess criterion 3 of Kolmogorov and Usphenski. However, these tests have severe limitations allowing only the rejection of the irregularity hypothesis in obvious cases. Consider the sequence 00110011001100110011. It can be checked that it passes the single proportion test and the runs test; that is, on the basis of those tests, one cannot reject the irregularity hypothesis with 95% confidence. However, the sequence is clearly periodic and therefore does not satisfy criterion 2 of Kolmogorov and Usphenski.

There is an interesting method for assessing irregularity which is based on the following idea: if there is no rule that will generate the following sequence elements starting with a given element, then a certain sequence segment will not resemble any other equal-length segment. On the other hand, if there are similar segments, an element-generating rule may then exist. We now present a 'similarity' measure consisting of adding up the product of corresponding elements between the original sequence and a shifted version of it. Take the very simple sequence 20102. Figure 8.4a shows this sequence (open bars) and another version of it (gray bars) perfectly aligned (for the purposes of visualization, bars in the same position are represented side by side; the reader must imagine them superimposed). In the case of Fig. 8.4a, the similarity between the two sequences is perfect. The sum of products is

$$2 \times 2 + 1 \times 1 + 2 \times 2 = 9 \; .$$

Figures 8.4b and c show what happens when the replica of the original sequence is shifted left (say, corresponding to positive shift values). In the case of Fig. 8.4b, there is a one-position shift and the sum of products is 0 (when shifting we assume that zeros are carried in). In the case of Fig. 8.4c, there is a two-position shift and the sum of products is 4. The sum of products thus works as a similarity measure. When the two sequences are perfectly aligned the similarity reaches its maximum. When the shift is only by one position, the similarity reaches its minimum (in fact zero values of one sequence are in this case in correspondence with non-zero values of the other).

Figure 8.5a shows, for the 20102 sequence, the curve of the successive values of this similarity measure known as *autocorrelation*. The

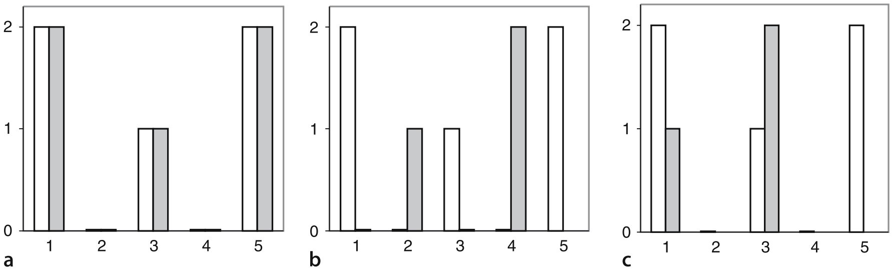


Fig. 8.4. Correlation of 20102 with itself for 0 (**a**), 1 (**b**) and 2 shifts (**c**)

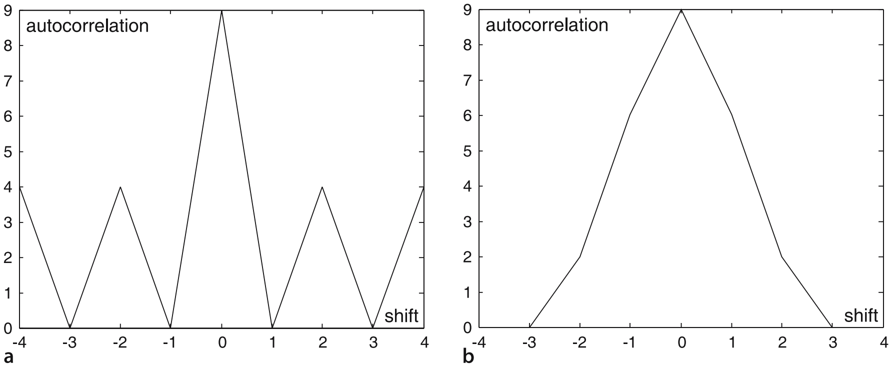


Fig. 8.5. Autocorrelation curves of (**a**) 20102, (**b**) 01220

name is due to the fact that we are (co)relating a sequence with itself. The sequence exhibits a certain zero–non-zero periodicity. The autocorrelation also exhibits that periodicity. Now suppose we shuffled the sequence values so that it looked like a random sequence of 0, 1 and 2. For instance, 01220. Now, the autocorrelation curve, shown in Fig. 8.5b, does not reveal any periodicity.

For a random sequence we expect the autocorrelation to be different from zero only when the sequence and its replica are perfectly aligned. Any shift must produce zero similarity (total dissimilarity): nothing looks the same. As a matter of fact, we are interested in a canceling effect between concordant and discordant values, through the use of the products of the deviations from the mean value, as we did with the correlation measure described in Chap. 5. Figure 8.6 shows the autocor-

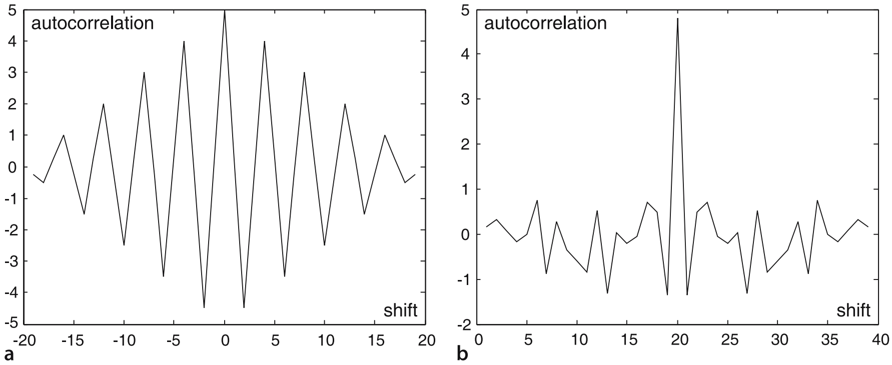


Fig. 8.6. Autocorrelation curves of (**a**) 00110011001100110011 and (**b**) 00100010101101101000

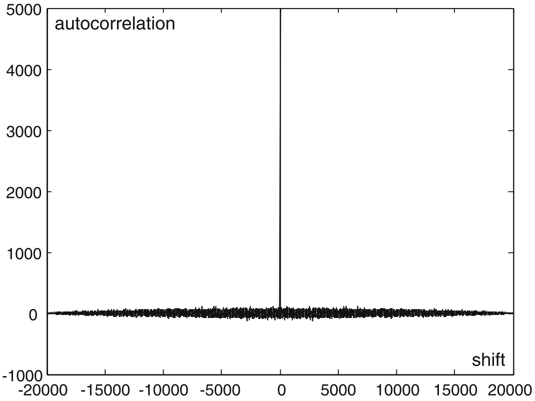


Fig. 8.7. Autocorrelation of a 20 000-bit sequence generated by a radioactive emission phenomenon

relations for the previous periodic binary sequence (at the beginning of this section) and for the first twenty values of the long-period sequence presented at the beginning of this chapter, computed using the deviations from the mean. The irregularity of the latter is evident.

Finally, Fig. 8.7 shows the autocorrelation of a 20 000-bit sequence generated by a quantum phenomenon (radioactive emission). Figure 8.7 is a good illustration that, off perfect alignment, nothing looks the same.

8.6 White Noise

Light shining from the Sun or from the lamps in our homes, besides its corpuscular nature (photons), also has a wave nature (electromagnetic fields) decomposable into sinusoidal vibrations of the respective fields. Newton's decomposition of white light reproducing the rainbow phenomenon is a well-known experiment. The French mathematician Joseph Fourier (1768–1830) introduced functional analysis based on sums of sinusoids (the celebrated Fourier transform) which became a widely used tool for analysis in many areas of science. We may also apply it to the study of random sequences.

Suppose we add two sinusoids (over time t) as shown in Fig. 8.8a. We obtain a periodic curve. We may vary the amplitudes, the phases (deviation from the origin of the sinusoid starting point), and the frequencies (related to the inter-peak interval) and the result is always periodic. In Fourier analysis the inverse operation is performed: given the signal (solid curve of Fig. 8.8a), one obtains the sinusoids that compose it (the rainbow). Figure 8.8b shows the Fourier analysis result

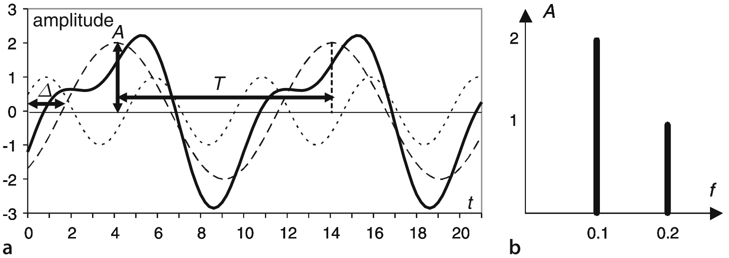


Fig. 8.8. (**a**) Sum of two sinusoids (*dotted curves*) resulting in a periodic curve (*solid curve*). One of the sinusoids has amplitude $A = 2$, frequency 0.1 (reciprocal of the period $T = 10$) and phase delay $\Delta \approx 2$. (**b**) Spectrum of the periodic signal

(only the amplitudes and the frequencies) of the signal in Fig. 8.8a. This is the so-called Fourier *spectrum*.

What happens if we add up an arbitrarily large number of sinusoids of the same frequency but with random amplitudes and phases? Well, strange as it may seem at first, we obtain a sinusoid with the same frequency (this result is easily deducible from trigonometric formulas).

Let us now suppose that we keep the amplitudes and phases of the sinusoids constant, but randomly and uniformly select their frequencies. That is, if the highest frequency value is 1, then according to the uniform law mentioned in Chap. 4, no subinterval of $[0, 1]$ is favored relative to others (of equal length) in the random selection of a frequency value. The result obtained is now completely different. We obtain a random signal! Figure 8.9 shows a possible example for the sum of 500 sinusoids with random frequencies. As we increase the number of sinusoids, the

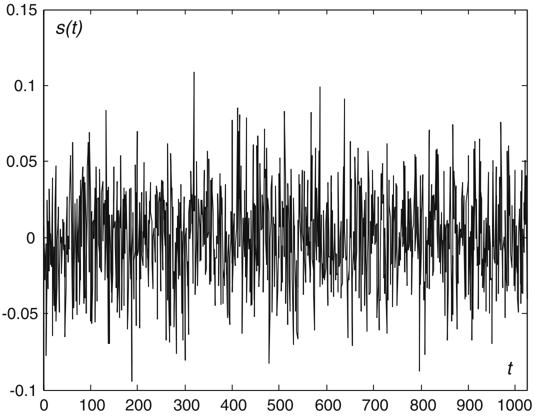


Fig. 8.9. An approximation of white noise (1 024 values) obtained by adding up 500 sinusoids with uniformly distributed frequencies

autocorrelation of the sum approaches a single impulse and the signal spectrum reflects an even sinusoidal composition for all frequency intervals; that is, the spectrum tends to a constant value (1, for unit amplitudes of the sinusoids). This result is better observed in the autocorrelation spectrum. Figure 8.10 shows the autocorrelation and its spectrum for the Fig. 8.9 sequence. The same result (approximately) is obtained for any other similar experiment of sinusoidal addition.

A signal of the type shown in Fig. 8.9 is called *white noise* (by analogy with white light). If we hear a signal of this type, it resembles the sound of streaming water. An interesting aspect is that, although the frequencies are uniformly distributed, the amplitudes of the white noise are *not* uniformly distributed. Their distribution is governed by the wonderful Gauss curve (see Fig. 8.11). In fact, it does not matter what the frequency distribution is, uniform or otherwise! Once again, the central limit theorem is at work here, requiring in the final result that the amplitudes be distributed according to the normal law.

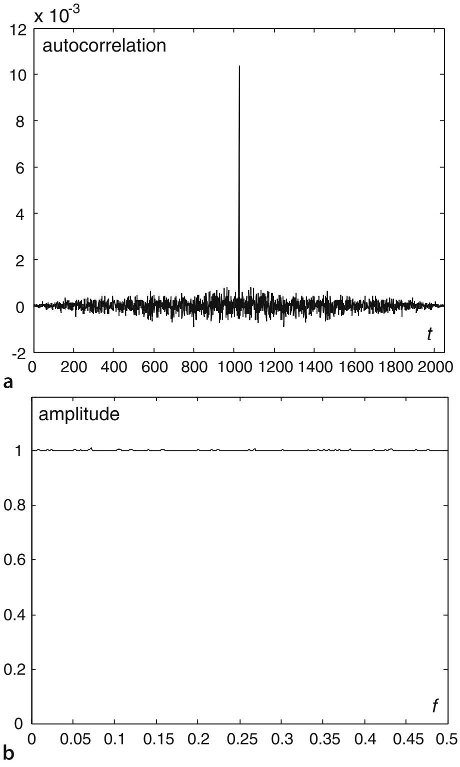


Fig. 8.10. Autocorrelation and its spectrum for the sequence of Fig. 8.9

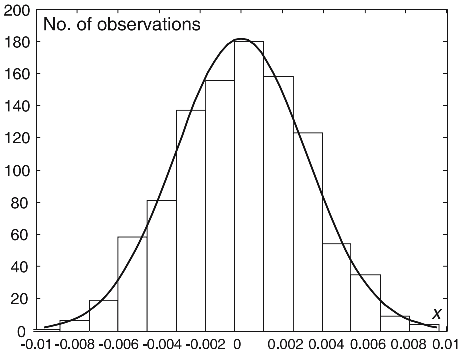


Fig. 8.11. Histogram of a white noise sequence (1 024 values) obtained by adding up 50 000 sinusoids with random amplitudes

White noise is, therefore, nothing other than a random sequence in which the different values are independent of each other (zero autocorrelation for any non-zero shift) produced by a Gaussian random generator.

We thus see that the autocorrelation and spectrum analysis methods tell us something about the randomness of sequences. However, these are not infallible either. For instance, the spectrum of the sequence 0000000010000000, clearly not random, is also 1. On the other hand, the spectrum of Champernowne's normal number (mentioned in Chap. 6) suggests the existence of five periodic regimes which are not visually apparent.

8.7 Randomness with Memory

White noise behaves like a sequence produced by coin-tossing. In fact, if we compute the autocorrelation and the spectrum of a sequence of coin throws (0 and 1), we obtain results that are similar to those obtained for white noise (Fig. 8.10). Let us now calculate the accumulated gain by adding the successive values of white noise as we did in the coin-tossing game. Figure 8.12a shows the accumulated gain of a possible instance of white noise. We obtain a possible instance of the one-dimensional Brownian motion already known to us. On the basis of this result, we may simulate Brownian motions in two (or more) dimensions (see Fig. 8.13), in a more realistic way than using the coin-tossing generator, which imposed a constant step in every direction (see Chap. 6).

In 1951, the hydraulic engineer Harold Hurst (1880–1970) presented the results of his study on the river Nile floods (in the context of the Assuan dam project, in Egypt). At first sight one might think that the

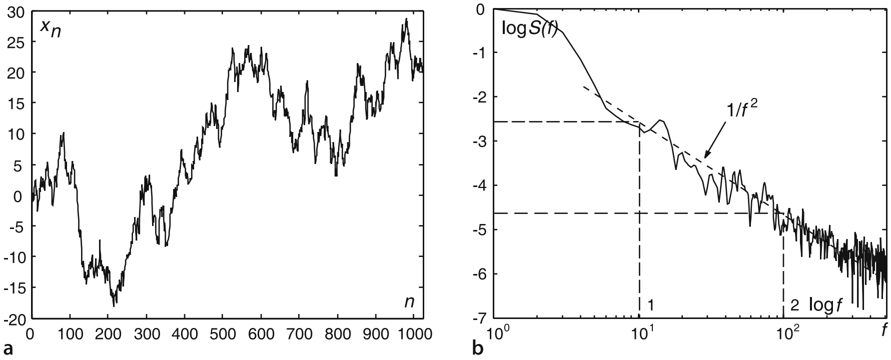


Fig. 8.12. Brownian motion sequence (**a**) and its autocorrelation spectrum (**b**). The *dotted line* with slope -2 corresponds to the $1/f^2$ evolution

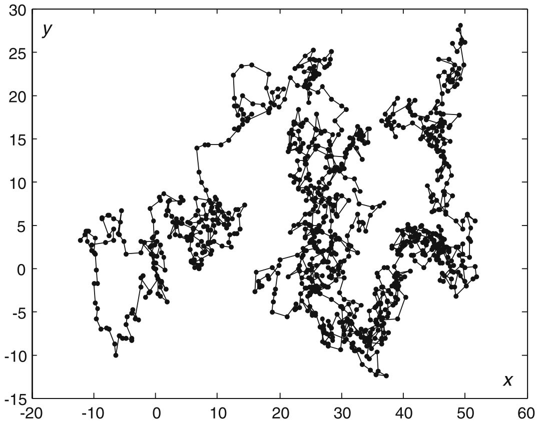


Fig. 8.13. Simulation of Brownian motion using white noise

average yearly flows of any river constitute a random sequence such as the white noise sequence. However, in that study and on the basis of flow records spanning 800 years, Hurst discovered that there was a correlation among the flows of successive years. There was a tendency for a year of good (high) flow rate to be followed by another year of good flow rate and for a year of bad (low) flow rate to be followed by another year of bad flow rate. It is almost as if there were a 'memory' in the flow rate evolution.

The study also revealed that the evolution of river flows was well approximated by a mathematical model, where instead of simply adding up white noise increments, one would add up the increments multiplied by a certain weighting factor influencing the contribution of past increments. Let t be the current time (year) and s some previous time (year). The factor in the mathematical model is given by $(t-s)^{H-0.5}$,

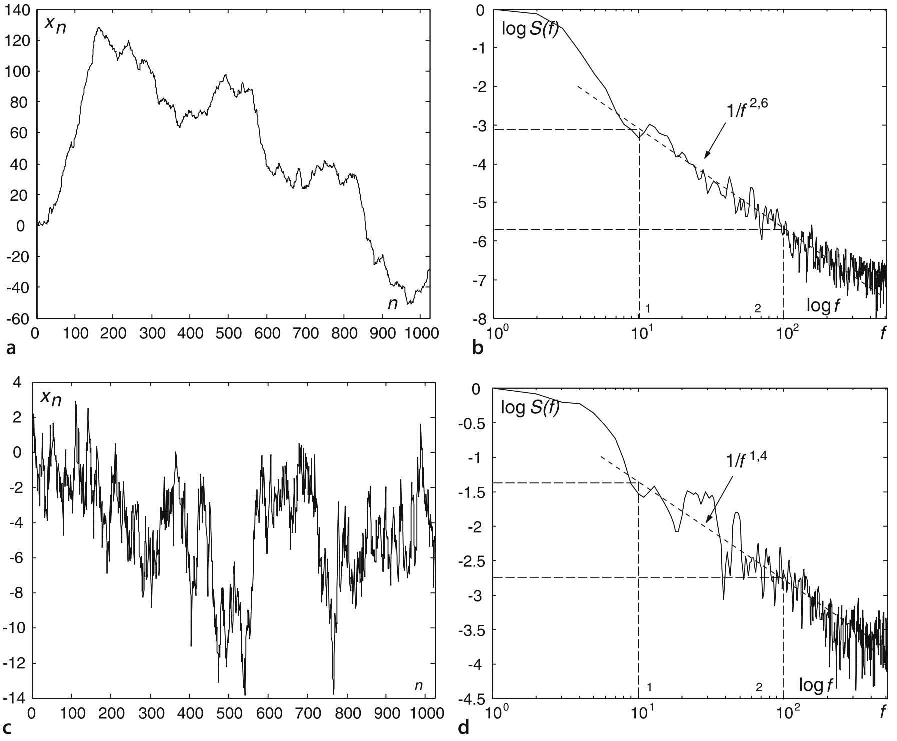


Fig. 8.14. Examples of fractional Brownian motion and their respective autocorrelation spectra (*right*). (**a**) $H = 0.8$. (**c**) $H = 0.2$. The spectral evolutions shown in (**b**) and (**d**) reveal the $1/f^{2H+1}$ decay, 2.6 and 1.4 for (**a**) and (**c**), respectively

where H (known as *Hurst exponent*) is a value greater than zero and less than one.

Consider the case with $H = 0.8$. The weight $(t - s)^{0.3}$ will then inflate the contribution of past values of s and one will obtain smooth curves with a 'long memory of the past', as in Fig. 8.14a. The current flow value tends to stray away from the mean in the same way as the previous year's value. Besides river flows, there are other sequences that behave in this persistent way, such as currency exchange rates and certain share values on the stock market.

On the other hand, if $H = 0.2$, the $(t - s)^{-0.3}$ term will quickly decrease for past increments and we obtain curves with high irregularity, with only 'short-term memory', such as the one in Fig. 8.14c. We observe this *anti-persistent* behavior, where each new value tends to revert to the mean, in several phenomena, such as heart rate tracings.

Between the two extremes, for $H = 0.5$, we have a constant value of $(t - s)^{H-0.5}$, equal to 1. There is no memory in this case, corresponding to Brownian motion. The cases with H different from 0.5 are called fractional Brownian motion.

8.8 Noises and More Than That

Besides river flows there are many other natural phenomena, involving chance factors, well described by fractional Brownian motion. An interesting example is the heart rate, including the fetal heart rate, a tracing of which is shown in Fig. 8.15. In this, as in other fields, the application of mathematical models reveals its usefulness.

Let us consider how chance phenomena are combined in nature and daily life. We have already seen that phenomena resulting from the addition of several random factors, acting independently of one another, are governed by the normal distribution law. There are also specific laws for phenomena resulting from multiplication (instead of addition) of random factors, as we shall see in a moment.

If we record a phrase of violin music, similar to a sum of sinusoids as in Fig. 8.8a, and we play it at another speed, the sound is distorted and the distinctive quality of the violin sound is lost. There are, however, classes of sounds whose quality does not change in any appreciable way when heard at other speeds. Such sounds are called scalable noises. The increase or decrease of the playing speed used to hear the sound is an observation scale. If we record white noise, such as in Fig. 8.12a, and play it at another speed, it will still sound like white noise (the sound of

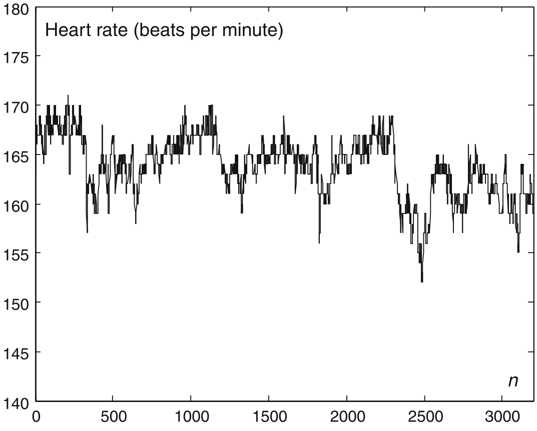


Fig. 8.15. Sequence of (human) fetal heart rate in calm sleep ($H = 0.35$, $D = 1.65$)

streaming water or ‘static’ radio noise). The spectrum is still a straight line and the autocorrelation is zero except at the origin. White noise remains white at any observation scale.

Let us now consider Brownian motion sequences. In this case the standard deviation of the noise amplitude can be shown to vary with the observation scale. Related to this fact, it can also be shown that the fluctuations of Brownian motion are correlated in a given time scale and the corresponding (autocorrelation) spectrum varies as $1/f^2$ (see Fig. 8.12b). Brownian ‘music’ sounds less chaotic than white noise and, when played at different speeds, sounds different while nevertheless retaining its Brownian characteristic. We have seen before that sequences of fractional Brownian motion also exhibit correlations. As a matter of fact, their spectra vary as $1/f^a$ with $a = 1 + 2H$ (see Figs. 8.14b and d) taking values between 1 and 3, that is, fractional Brownian noises have a spectrum above the $1/f$ spectrum. What kind of noise does this spectrum have? Well, it has been shown that music by Bach and other composers also has a $1/f$ spectrum. If we play Brownian music with a $1/f^2$ spectrum, we get an impression of monotony, whereas if we play ‘noise’ with a $1/f$ spectrum, we get a certain amount of aesthetic feeling.

Do chance phenomena generating $1/f$ spectrum sequences occur that often? In Chap. 3 we mentioned the Bernoulli utility function $U(f) = \log f$. Utility increments also obey the $1/f$ rule, which is nothing other than instantaneous variation of the logarithm. Imagine that I pick up the temporal evolution of the fortune of some individual and I multiply all the values by a scale factor a. What happens to the evolution of the respective utilities? Since $\log(af) = \log a + \log f$, the evolution of the scaled utility shows the same pattern with the minor difference that there is a shift of the original utility by the constant value $\log a$. Furthermore, the relative variation, i.e.,

$$\frac{\text{variation of } U}{\text{variation of } f}\,,$$

is maintained. Briefly, the utility function exhibits *scale invariance*. There are many natural phenomena exhibiting scale invariance, known as $1/f$ phenomena. For instance, it has long been known that animals do not respond to the absolute value of some stimuli (such as smell, heat, electric shock, skin pressure, etc.), but rather to the logarithm of the stimuli. There are also many socio-economical $1/f$ phenomena, such as the wealth distribution in many societies, the frequency that words are used, and the production of scientific papers.

The existence of $1/f$ random phenomena has been attributed to the so-called multiplicative effect of their causes. For instance, in the case of the production of scientific papers, it is well known that a senior researcher tends to co-author many papers corresponding to works of supervised junior researchers, more papers than he would individually have produced. In the same way, people that are already rich tend to accumulate extra wealth by a multiplicative process, mediated by several forms of investment and/or financial speculation. Consider the product f of several independent random variables acting in a multiplicative way. The probability law of the logarithm of the product is, in general, well approximated by the normal law, given the central limit theorem and the fact that the logarithm of a product converts into a sum of logarithms. Now, if $\log f$ is well approximated by the normal law, f is well approximated by another law (called the lognormal law) whose shape tends to $1/f$ as we increase the number of multiplicative causes. Hence, $1/f$ phenomena also tend to be ubiquitous in nature and everyday life.

8.9 The Fractality of Randomness

There is another interesting aspect of the irregularity of Brownian sequences. Suppose that I am asked to measure the total length of the curve in Fig. 8.12a and for that purpose I am given a very small non-graduated ruler, of length 1 in certain measurement units, say 1 mm. All I can do is to see how many times the ruler (the whole ruler) ‘fits’ into the curve end-to-end. Suppose that I obtain a length which I call $L(1)$, e.g., $L(1) = 91$ mm. Next, I am given a ruler of length 2 mm and I am once again asked to measure the curve length. I find that the ruler ‘fits’ 32 times, and therefore $L(2) = 64$ mm. Now, 64 mm is approximately $L(1)/\sqrt{2}$. In fact, by going on with this process I would come to the conclusion that the following relation holds between the measurement scale r (ruler length), and the total curve length, $L(r)$: $L(r) = L(1)/\sqrt{r}$, which we may rewrite as $L(r) = L(1)r^{-0.5}$. We thus have an object (a Brownian motion sequence) with a property (length) satisfying a *power law* (r^{α}) relative to the scale (r) used to measure the property. In the present case, $\alpha = -0.5$. Whenever such a power law is satisfied, we are in the presence of a *fractal property*. Thus, the length of a sequence of Brownian motion is a fractal.

Much has been written about fractals and there is a vast popular literature on this topic. The snowflake curve (Fig. 8.16) is a well-known

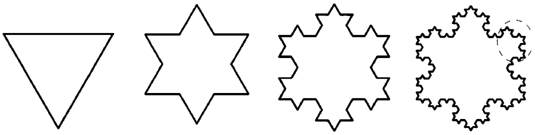


Fig. 8.16. The first four iterations of the snowflake curve. As one increases the number of iterations, the curve length tends to infinity, while staying confined inside a bounded plane region

fractal object characterized by an infinite length, although contained in a finite area. In fact, the number of snowflake segments $N(r)$ measured with resolution r is given by $N(r) = (1/r)^{1.26} = r^{-1.26}$. Were it a 'common', non-fractal curve, we would obtain $N(r) = (1/r)^1 = r^{-1}$ (if, for instance, a 1-meter straight line is measured in centimeters, we find $1/0.01 = 100$ one-centimeter segments). In the same way, for a 'common', non-fractal area, we would obtain $N(r) = (1/r)^2 = r^{-2}$ [if, for instance, I have to fill up a square with side 1 meter with tiles of side 20 cm, I need $(1/0.2)^2 = 25$ tiles]. The snowflake curve is placed somewhere between a flat 'common' curve of dimension 1 and a plane region of dimension 2. For 'common' objects, the exponent of r is always an integer. However, for fractal objects such as the snowflake, the exponent of r often contains a fractional part leading to the 'fractal' designation. (As a matter of fact there are also fractal objects with integer exponents.) We say that 1.26 is the *fractal dimension* of the snow flake. Suppose that, instead of counting segments, we measure the curve length. We then obtain the relation $L(r) = L(1)(1/r)^{D-1}$, where D is the fractal dimension. For the snowflake, we get $L(r) = L(1)(1/r)^{0.26}$. The length grows *with* $1/r$, but not as quickly *as* $1/r$.

We thus arrive at the conclusion that the length of the Brownian motion sequence behaves like the length of the snowflake curve, the only difference being that the growth is faster: $L(r) = L(1)(1/r)^{0.5}$. The Brownian motion sequence has fractal dimension $D = 1.5$ (since $D - 1 = 0.5$).

When talking about fractal objects it is customary to draw attention to the self-similarity property of these objects. For instance, when observing a flake bud at the fourth iteration (enclosed by a dotted ball in Fig. 8.16), one sees an image corresponding to several buds of the previous iteration. Does something similar happen with the Brownian motion sequence? Let us take the sequence in Fig. 8.12a and zoom in on one of its segments. Will the zoomed segment resemble the original sequence? If it did, something repetitive would be present, as in the snowflake, and the sequence would not have a random behavior. Does self-similarity reside in the type of sequence variability? We know that the standard deviation of a sequence of numbers is a measure of its degree of variability. Let us then proceed as follows: we divide the sequence into segments of length (number of values) r and determine the average standard deviation for all segments of the same length. Figure 8.17 shows how the average standard deviation $s(r)$ varies with r. It varies approximately as $s(r) = \sqrt{r}$. The conclusion is therefore affirmative, that there does in fact exist a 'similarity' of the standard deviations – and hence of the variabilities – among segments of Brownian motion observed at different scales. It does come as a surprise that a phenomenon entirely dependent on chance should manifest this statistical self-similarity.

It can be shown that the curves of fractional Brownian motion have fractal dimension $D = 2 - H$. Thus, whereas the curve of Fig. 8.14a has dimension 1.2 and is close to a 'common' curve, without irregularities, the one in Fig. 8.14c has fractal dimension 1.8, and is therefore close to filling up a certain plane region. In the latter case, there is a reduced self-similarity and instead of Fig. 8.17 one would obtain a curve whose standard deviation changes little with r (varying as $r^{0.2}$). In the first case the self-similarity varies as $r^{0.8}$. For values of H very close to 1, the variation with r is almost linear and the self-similarity is perfect, corresponding to an absence of irregularity.

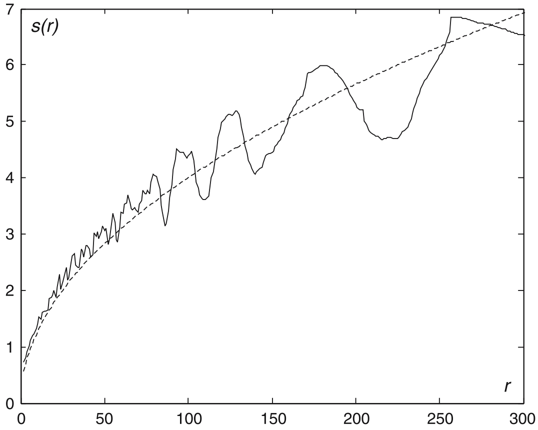


Fig. 8.17. Evolution of the average standard deviation for segments of length r of the curve in Fig. 8.12a

9

Chance and Order

9.1 The Guessing Game

The reader is probably acquainted with the guess-the-noun game, where someone thinks of a noun, say an animal, and the reader has to guess which animal their interlocutor thought of by asking him/her closed questions, which can be answered either “Yes” or “No”. A possible game sequence could be:

– Is it a mammal?

– Yes.

– Is it a herbivore?

– No.

– Is it a feline?

– Yes.

– Is it the lion?

– No.

And so on and so forth.

The game will surely end, because the number of set elements of the game is finite. It may, however, take too much time for the patience and knowledge of the player who will end up just asking the interlocutor which animal (s)he had thought of. Just imagine that your interlocutor thought of the Atlas beetle and the reader had to make in the worst case about a million questions, as many as the number of classified insects!

Let us move on to a different guessing game, independent of the player’s knowledge, and where one knows all the set elements of the game. Suppose the game set is the set of positions on a chess board where a queen is placed, but the reader does not know where. The reader is invited to ask closed questions aiming at determining the queen's position. The most efficient way of completing the game, the one corresponding to the smallest number of questions, consists in performing successive dichotomies of the board. A dichotomous question divides the current search domain into halves. Let us assume the situation pictured in Fig. 9.1. A possible sequence of dichotomous questions and their respective answers could be:

– Is it in the left half?

– Yes.

– Is it in the upper half? (In this and following questions it is always understood half of the previously determined domain; the left half in this case.)

– No.

– Is it in the left half?

– No.

– Is it in the upper half?

– Yes.

– Is it in the left half?

– Yes.

– Is it in the upper half?

– No. (And the game ends because the queen's position has now been found.)

Whatever the queen's position may be, it will always be found in at most the above six questions. Let us analyze this guessing game in more detail. We may code each "Yes" and "No" answer by 1 and 0, respectively, which we will consider as in Chap. 6 (when talking about Borel's normal numbers) as representing binary digits (bits). The sequence of

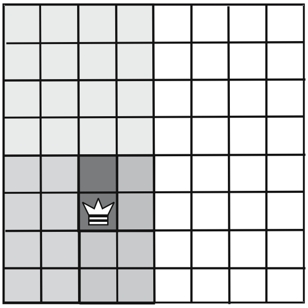


Fig. 9.1. Chessboard with a queen

answers is therefore a sequence of bits. The difference between this second game and the first lies in the fact that the correspondence between the position of each bit in the sequence and a given search domain is known and utilized.

Let us now consider the following set of letters which we call E:

$$E = \{a, b, c, d, e, f, g, h\} .$$

We may code the set elements by $\{000, 001, 010, 011, 100, 101, 110, 111\}$. Finding a letter in E corresponds to asking 3 questions, as many as the number of bits needed to code the elements of E. For instance, finding the letter c corresponds to the following sequence of answers regarding whether or not the letter is in the left half of the previous domain: Yes (c is in the left half of E), No (c is in the right half of $\{a, b, c, d\}$), Yes (c is in the left half of $\{c, d\}$). Let N be the number of elements in the set. We notice that in the first example of the chessboard we had $N = 2^6 = 64$ distinct elements and needed (at most) 6 questions; in this second example we have $N = 2^3 = 8$ distinct elements and need 3 questions. Briefly, it is not hard to see that the number of questions needed is in general given by $\log_2 N$ (the base 2 logarithm of N, which is the exponent to which one must raise 2 in order to obtain N; a brief explanation of logarithms can be found in Appendix A at the end of the book).

In 1920, the electronic engineer Ralph Hartley (1888–1970), working at the Bell Laboratories in the United States, was interested by the problem of defining a measure quantifying the amount of information in a message. The word ‘information’ is used here in an abstract sense. Concretely, it does not reflect either the form or the contents of the message; it only depends on the symbols used to compose the message. Let us start with a very simple situation where the message sender can only send the symbol “Yes”. Coding “Yes” as 1, the sender can only send sequences of ones. There is then no possible information in the sense that, from the mere fact that we received “Yes”, there has been no decrease in our initial uncertainty (already zero) as to the symbol being sent. The simplest case from the point of view of getting non-zero information is the one where the emitter uses with equal probability the symbols from the set $\{\text{Yes}, \text{No}\}$ and sends one of them. We then say that the information or degree of uncertainty is of 1 bit (a binary digit codes $\{\text{Yes}, \text{No}\}$, e.g., Yes = 1, No = 0).

Now, suppose that the sender sends each of the symbols of the above set $E = \{a, b, c, d, e, f, g, h\}$ with equal probability. The information obtained from one symbol, reflecting the degree of uncertainty relative to the symbol value, is equivalent to the number of Yes/No questions one would have to make in order to select it. Hartley's definition of the information measure when the sender uses an equally probable N-symbol set is precisely $\log_2 N$. It can be shown that even when N is not a power of 2, as in the previous examples, this measure still makes sense. The information unit is the bit and the measure satisfies the additivity property, as it should in order to be useful. We would like to say that we need 1 bit of information to describe the experiment in which one coin is tossed and 2 bits of information to describe the experiment in which two coins are tossed. Suppose that in the example of selecting an element of E, we started with the selection of an element from $E_1 = \{\text{left half of } E, \text{ right half of } E\}$ and, according to the outcome, we then proceeded to choose an element from the left or right half of E. The choice in E_1 implies 1 bit of information. The choice in either the left or right half of E implies 2 bits. According to the additivity rule the choice of an element of E then involves 3 bits, exactly as expected.

Now, consider the following telegram: ANN SICK. This message is a sequence of capital letters and spaces. In the English alphabet there are 26 letters. When including other symbols such as a space, hyphen, comma, and so forth, we arrive at a 32-symbol set. If the occurrence of each symbol were equiprobable, we would then have an information measure of 5 bits per symbol. Since the message has 8 symbols it will have $5 \times 8 = 40$ bits of information in total. What we really mean by this is that ANN SICK has the same information as ANN DEAD or as any other of the $N = 2^{40}$ possible messages using 8 symbols from the set. We clearly see that this information measure only has to do with the number of possible choices and not with the usual meaning we attach to the word 'information'. As a matter of fact, it is preferable to call this information measure an uncertainty measure, i.e., uncertainty as to the possible choices by sheer chance.

9.2 Information and Entropy

In the last section we assumed that the sending of each symbol was equally likely. This is not often the case. For instance, the letter e occurs about 10% of the time in English texts, whereas the letter z only occurs about 0.04% of the time. Suppose that we are only using two symbols, 0 and 1, with probabilities $P_0 = 3/4$ and $P_1 = 1/4$, and that we want to determine the amount of information when the sender sends one of the two symbols. In 1948 the American mathematician Claude Shannon

(1916–2001), considered as the father of information theory, presented the solution to this problem in a celebrated paper. In fact, he gave the solution not only for the particular two-symbol case, but for any other set of symbols. We present here a simple explanation of what happens in the above two-symbol example. We first note that if $P(0) = 3/4$, this means that, in a long sequence of symbols, 0 occurs three times more often than 1. We may look at the experiment as if the sender, instead of randomly selecting a symbol in $\{0, 1\}$, picked it from a set whose composition reflects those probabilities, say, $E = \{0, 0, 0, 0, 0, 0, 1, 1\}$, as shown in Fig. 9.2.

We thus have a set E with 8 elements which is the union of two sets, E_1 and E_2, containing 6 zeros and two ones, respectively. The information corresponding to making a random selection of a symbol from E, which we know to be $\log_2 8$, decomposes into two terms (additivity rule):

- The information corresponding to selecting E_1 or E_2, which is the term we are searching for and which we denote by H.
- The average information corresponding to selecting an element in E_1 or E_2.

We already know how to compute the average (mathematical expectation) of any random variable, in this case the quantity of information for E_1 and E_2 ($\log_2 6$ and $\log_2 2$, respectively):

$$\begin{aligned}\begin{array}{c}\text{Average information in the selection}\\ \text{of an element in } E_1 \text{ or } E_2\end{array} &= P_0 \times \log_2 6 + P_1 \times \log_2 2\\ &= P_0 \log_2(8P_0) + P_1 \log_2(8P_1)\ .\end{aligned}$$

Hence,

$$\log_2 8 = H + P_0 \log_2(8P_0) + P_1 \log_2(8P_1)\ .$$

Taking into account the fact that we may multiply $\log_2 8$ by $P_0 + P_1$, since it is equal to 1, one obtains, after some simple algebraic manipulation,

$$H = -P_0 \log_2 P_0 - P_1 \log_2 P_1\ .$$

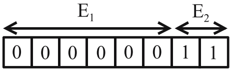


Fig. 9.2. A set with three times more 0s than 1s

For the above-mentioned probabilities, one obtains $H = 0.81$. Shannon called this information measure *entropy*. It is always non-negative and constitutes a generalization of Hartley's formula. If $P_0 = P_1 = 1/2$ (total disorder), one obtains $H = 1$, the maximum entropy and information in the selection of an element from E. On the other hand, if P_0 or P_1 is equal to 1 (total order), one obtains $H = 0$, the previously mentioned case of zero information, corresponding to the entropy minimum (by convention $0 \log_2 0 = 0$). Between the two extremes, we get an entropy value between 0 and 1. The general definition of entropy, for any set E of elementary events with probabilities P, is a straightforward generalization of the above result:

$$H = -\text{sum of } P \log_2 P \text{ for all elementary events of } E\ .$$

If E has k elements, the entropy maximum is $\log_2 k$ and occurs when all elements are equally probable (uniform distribution). As a matter of fact, the entropy maximum of an unrestricted discrete distribution occurs when the distribution is uniform.

Shannon proved that the entropy H is the minimum number of bits needed to code the complete statistical description of a symbol set. Let us once again take the set $E = \{a, b, c, d, e, f, g, h\}$ and assume that each symbol is equiprobable. Then $H = \log_2 8 = 3$ and we need 3 bits in order to code each symbol. Now imagine that the symbols are not equiprobable and that the respective probabilities are $\{0.5, 0.15, 0.12, 0.10, 0.04, 0.04, 0.03, 0.02\}$. Applying the above formula one obtains $H = 2.246$, less than 3. Is there any way to code the symbols so that an average information of 2.246 bits per symbol is attained? There is in fact such a possibility and we shall give an idea how such a code may be built. Take the following coding scheme which assigns *more* bits to the symbols occurring *less* often: $a = 1$, $b = 001$, $c = 011$, $d = 010$, $e = 00011$, $f = 00010$, $g = 00001$, $h = 00000$. Note that, in a sequence built with these codes, whenever a single 0 occurs (and not three consecutive 0s), we know that the following two bits discriminate the symbol; whenever three consecutive 0s occur we know that they are followed by two discriminating bits. There is then no possible confusion. Consider the following message obtained with a generator that used those probabilities: *bahaebaaaababaaacge*. This message would be coded as

$$0011000001000110011111100110011110100011$$

and would use 45 bits instead of the 57 bits needed if we had used three bits per symbol. Applying the formula for the average, one easily verifies that an average of 2.26 bits per symbol are used with the above coding scheme, close to the entropy value.

9.3 The Ehrenfest Dogs

The word ‘entropy’ was coined for the first time in 1865 by the German physicist Rudolf Clausius (1822–1888) in relation to the concept of transformation/degradation of energy, postulated by the second principle of thermodynamics. In fact ‘entropy’ comes from the Greek word ‘entropê’ meaning ‘transformation’. Later on, in 1877, the Austrian physicist Ludwig Boltzmann (1844–1906) derived a formula for the entropy of an isolated system which reflects the number of possible states at molecular level (microstates), establishing a statistical description of an isolated system consistent with its thermodynamic properties. Boltzmann’s formula for the entropy, designated by S, is $S = k \log \varOmega$, where $\varOmega$ is the number of equiprobable microstates needed to describe a macroscopic system. In Boltzmann’s formula the logarithm is taken in the natural base (number e) and k is the so-called Boltzmann constant: $k \approx 1.38 \times 10^{-23}$ joule/kelvin. Note the similarity between the physical and mathematical entropy formulas. The only noticeable difference is that the physical entropy has the dimensions of an energy divided by temperature, whereas mathematical entropy is dimensionless. However, if temperature is measured in energy units (say, measured in boltzmanns instead of K, °C or °F), the similarity is complete (the factor k can be absorbed by taking the logarithm in another base).

In 1907, the Dutch physicists Tatiana and Paul Ehrenfest proposed an interesting example which helps to understand how entropy evolves in a closed system. Suppose we have two dogs, Rusty and Toby, side-by-side, exchanging fleas between them, and no one else is present. Initially, Rusty has 50 fleas and Toby has none. Suppose that after this perfectly ordered initial state, a flea is randomly chosen at each instant of time to move from one dog to the other. What happens as time goes by? Figure 9.3 shows a possible result of the numerical simulation of the (isolated) dog–flea system. As time goes by the distribution of the fleas between the two dogs tends to an equilibrium situation with small fluctuations around an equal distribution of the fleas between the two dogs: this is the situation of total disorder.

At each instant of time, one can determine the probability distribution of the fleas by performing a large number of simulations. The result is shown in Fig. 9.4. The distribution converges rapidly to a binomial

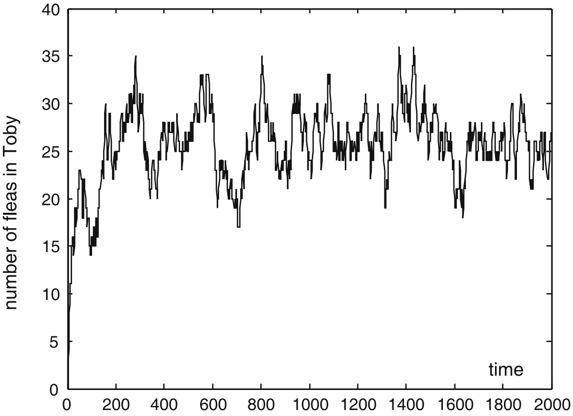


Fig. 9.3. Time dependence of the number of fleas on Toby

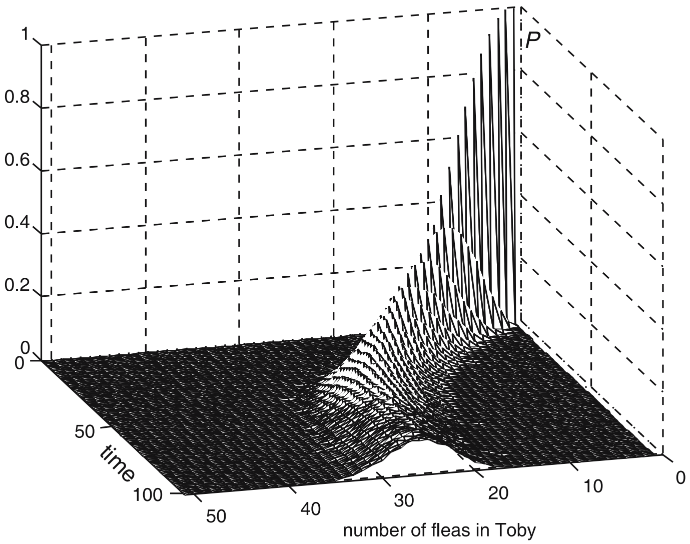


Fig. 9.4. The flea distribution as it evolves in time. The *vertical axis* represents the probability of the number of fleas for each time instant (estimated in 10 000 simulations)

distribution very close to the normal one. Effectively, the normal distribution is, among all possible continuous distributions with constrained mean value (25 in this case), the one exhibiting maximum entropy. It is the 'most random' of the continuous distributions (as the uniform distribution is the most random of the discrete distributions).

Figure 9.5 shows how the entropy of the dog–flea system grows, starting from the ordered state (all fleas on Rusty) and converging to the disordered state.

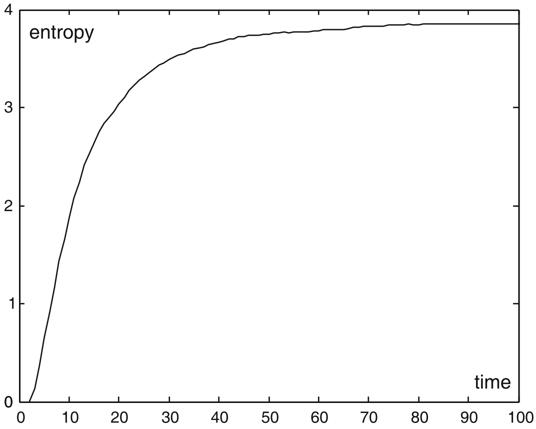


Fig. 9.5. Entropy evolution in the isolated dog–flea system

The entropy evolution in the Ehrenfest dogs is approximately what one can observe in a large number of physical systems. Is it possible for a situation where the fleas are all on Rusty to occur? Yes, theoretically, it may indeed occur. If the distribution were uniform, each possible flea partition would have a $1/2^{50} = 0.000\,000\,000\,000\,000\,89$ probability of occurring. For the binomial distribution, this is also the probability of getting an ordered situation (either all fleas on Toby or all fleas on Rusty). If the time unit for a change in the flea distribution is the second, one would have to wait about 2^{50} seconds, which amounts to more than 35 million years (!), in order to have a reasonable probability of observing such a situation.

Similarly to the dog–flea system, given a recipient initially containing luke warm water, we do not expect to observe water molecules with higher kinetic energy accumulating at one side of the recipient, which would result in an ordered partition with the emergence of hot water at one side and cold water at the other. If instead of water we take the simpler example of one liter of dilute gas at normal pressure and temperature, we are dealing with something like 2.7×10^{22} molecules, a huge number compared to the 50-flea system! Not even the whole lifetime of the Universe would suffice to reach a reasonable probability of observing a situation where hot gas (molecules with high kinetic energy) would miraculously be separated from cold gas (molecules with low kinetic energy)!

It is this tendency-to-disorder phenomenon and its implied entropy rise in an isolated system that is postulated by the famous second principle of thermodynamics. The entropy rise in many isolated physical systems is the trademark of time's 'arrow'. (As a matter of fact, no system is completely isolated and the whole analysis is more complex than the basic ideas we have presented; in a strict sense, it would have to embrace the whole Universe.)

9.4 Random Texts

We now take a look at 'texts' written by randomly selecting lower-case letters from the English alphabet plus the space character. We are dealing with a 27-symbol set. If the character selection is governed by the uniform distribution, we obtain the maximum possible entropy for our random-text system, i.e., $\log_2 27 = 4.75$ bits. A possible 85-character text produced by such a random-text system with the maximum of disorder, usually called a *zero-order model*, is:

> rlvezsfipc icrkbxknwk awlirdjokmgnmxdlfaskjn jmxckjvar
> jkzwhkuulsk odrzguqtjf hrlywwn

No similarity with a real text written in English is apparent. Now, suppose that instead of using the uniform distribution, we used the probability distribution of the characters in (true) English texts. For instance, in English texts the character that most often occurs is the space, occurring about 18.2% of the time, while the letter occurring most often is e, about 10.2% of the time, and z is the least frequent character, with a mere 0.04% of occurrences. A possible random text for this *first-order model* is:

> vsn wehtmvbols p si e moacgotlions hmelirtrl ro esnelh
> s t emtamiodys ee oatc taa p invoj oetoi dngasndeyse ie
> mrtrtlowh t eehdsafbtre rrau drnwsr

In order to generate this text we used the probabilities listed in Table 9.1, which were estimated from texts of a variety of magazine articles made available by the British Council web site, totalizing more than 523 thousand characters.

Using this first-order model, the entropy decreases to 4.1 bits, reflecting a lesser degree of disorder. We may go on imposing restrictions on the character distribution in accordance with what happens in English texts. For instance, in English, the letter j never occurs before a space, i.e., it never ends a word (with the exception of foreign words). In addition, in English q is always followed by u. In fact, given a character c, we can take into account the probabilities of c being followed by

Table 9.1. Frequency of occurrence of the letters of the alphabet and the space in a selection of magazine articles written in English

space	0.1818	g	0.0161	n	0.0567	u	0.0231
a	0.0685	h	0.0425	o	0.0635	v	0.0091
b	0.0128	i	0.0575	p	0.0171	w	0.0166
c	0.0239	j	0.0012	q	0.0005	x	0.0016
d	0.0291	k	0.0061	r	0.0501	y	0.0169
e	0.1020	l	0.0342	s	0.0549	z	0.0004
f	0.0180	m	0.0207	t	0.0750		

another character, say c_1, written $P(c_1 \text{ if } c)$, when generating the text. For instance, $P(\text{space if j}) = 0$ and $P(\text{u if q}) = 1$. Any texts randomly generated by this probability model, the so-called *second-order model*, have a slight resemblance with a text in English. Here is an example:

> litestheeneltr worither chethantrs wall o p the junsatinlere moritr istingaleorouson atout oven olle trinmobbulicsopexagender leturorode meacs iesind cl whall meallfofedind f cincexilif ony o agre fameis othn d ailalertiomoutre pinlas thorie the st m

Note that, in this ‘phrase’, the last letter of the ‘words’ is a valid end letter in English. Moreover, some legitimate English words do appear, such as ‘wall’, ‘oven’ and ‘the’. The entropy for this model of conditional probability is known as conditional entropy and is computed in the following way:

$$H(\text{letter 1 if letter 2}) = H(\text{letter 1 and letter 2}) - H(\text{letter 2}) \; .$$

Performing the calculations, one may verify that the entropy of this second-order model is 3.38 bits. Moving to *third-* and *fourth-order models*, we use the conditional probabilities estimated in English texts of a given character being followed by another two or three characters, respectively. Here is a phrase produced by the *third-order model*:

> l min trown redle behin of thrion sele ster doludicully to in the an agapt weace din ar be a ciente gair as thly eantryoung othe of reark brathey re the strigull sucts and cre ske din tweakes hat he eve id iten hickischrit thin ficated on bel pacts of shoot ity numany prownis a cord we danceady suctures acts i he sup fir is pene

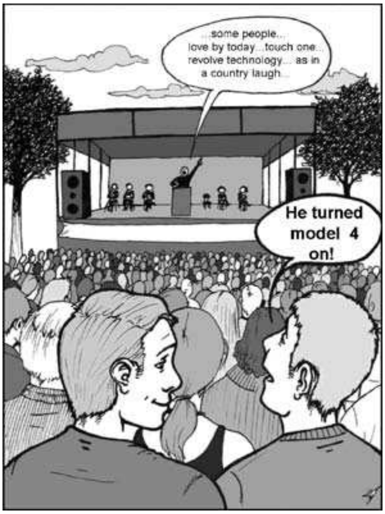


Fig. 9.6. Automating the art of public speaking

And a phrase produced by the *fourth-order model*:

> om some people one mes a love by today feethe del purestiling touch one from hat giving up imalso the dren you migrant for mists of name dairst on drammy arem nights abouthey is the the dea of envire name a collowly was copicawber beliver aken the days dioxide of quick lottish on diffica big tem revolvetechnology as in a country laugh had

The entropies are 2.71 and 2.06, respectively. Note that in all these examples we are only restricting probability distributions, with the result that the texts are only vaguely reminiscent of English texts. It would be possible to add syntactic and semantic rules imposing further restrictions on the generated phrases, thus enhancing the resemblance to legitimate English phrases. These additional rules would of course contribute to further lowering the entropy. As a matter of fact, each added restriction means added information, therefore a decrease in the uncertainty of the text.

As we go on imposing restrictions and as a consequence increasing the order of the model, the entropy steadily decreases. The question is: how far will it go? It is difficult to arrive at an accurate answer, since for high orders one would need extraordinarily large sequences in order to estimate the respective probabilities with some confidence. For instance, for a fifth-order model there are 27^5 possible 5-letter sequences, that is, more than 14 million combinations (although a large number of them never occur in English texts)! There is, however, a possibility

Table 9.2. Relative frequencies of certain words in English as estimated from a large selection of magazine articles

the	0.06271	it	0.01015	have	0.00631	from	0.00482
and	0.03060	for	0.00923	people	0.00586	their	0.00465
of	0.03043	are	0.00860	I	0.00584	can	0.00443
to	0.02731	was	0.00819	with	0.00564	at	0.00442
a	0.02702	you	0.00712	or	0.00559	there	0.00407
in	0.02321	they	0.00705	be	0.00541	one	0.00400
is	0.01427	on	0.00632	but	0.00521	were	0.00394
that	0.01157	as	0.00631	this	0.00513	about	0.00390

of estimating that limit using the so-called Zipf law. The American linguist George Zipf (1902–1950) studied the statistics of word occurrence in several languages, having detected an inverse relation between the frequency of occurrence of any word and its rank in the corresponding frequency table. If for instance the word ‘the’ is the one that most often occurs in English texts and the word ‘with’ is ranked in twentieth position, then according to Zipf’s law (presented in 1930), the word ‘with’ occurs about $1/20^a$ of the times that ‘the’ does. The same happens to other words. The parameter a has a value close to 1 and is characteristic of the language.

Let n denote the rank value of the words and $P(n)$ the corresponding occurrence probability. Then Zipf’s law prescribes the following dependency of $P(n)$ on n:

$$P(n) = \frac{b}{n^a} \, .$$

This relation is better visualized in a graph if one performs a logarithmic transformation: $\log P(n) = \log b - a \log n$. Hence, the graph of $\log P(n)$ against $\log n$ is a straight line with slope $-a$.

Table 9.2 shows the relative frequencies of some words estimated from the previously mentioned English texts. Figure 9.7 shows the application of Zipf’s law using these estimates and the above-mentioned logarithmic transformation. Figure 9.7 also shows a possible straight line fitting the observed data.

According to the straight line in Fig. 9.7, we obtain

$$P(n) \approx \frac{0.08}{n^{0.96}} \, .$$

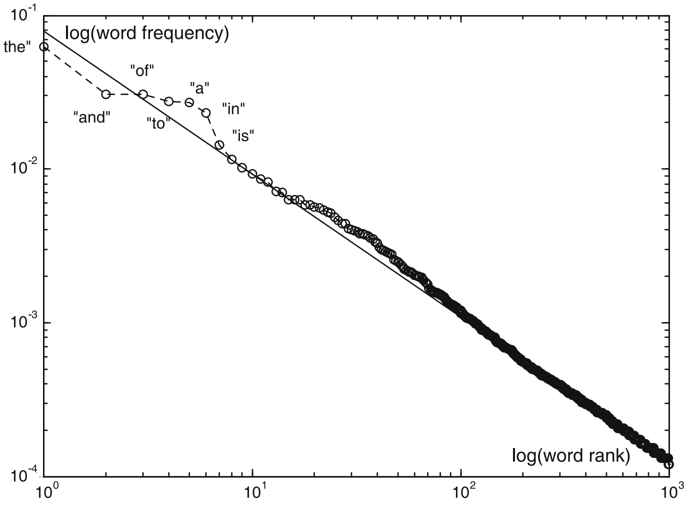


Fig. 9.7. Zipf's law applied to English texts

One may then apply the entropy formula to the $P(n)$ values for a sufficiently large n. An entropy $H \approx 9.2$ bits per word is then obtained. Given the average size of 7.3 letters per word in English (again estimated from the same texts), one finally determines a letter entropy estimate for English texts as 1.26 bits per letter. Better estimates based on larger and more representative texts yield smaller values; about 1.2 bits per letter. Other languages manifest other entropy values. For instance, entropy is higher in Portuguese texts; about 1.9 bits per letter. However, in Portuguese the number of characters to consider is not 27 as in English but 37 (this is due to the use of accents and of ç).

Zipf's law is empirical. Although its applicability to other types of data has been demonstrated (for instance, city populations, business volumes and genetic descriptions), a theoretical explanation of the law remains to be found.

9.5 Sequence Entropies

We now return to the sequence randomness issue. In a random sequence we expect to find k-symbol blocks appearing with equal probability for all possible symbol combinations. This is, in fact, the property exhibited by Borel's normal numbers presented in Chap. 6.

Let us assume that we measure the entropy for the combinations of k symbols using the respective probabilities of occurrence $P(k)$, as we did before in the case of texts:

$$H(k) = -\text{sum of } P(k) \times \log_2 P(k) \ .$$

Furthermore, let us take the limit of the $H(k) - H(k-1)$ differences for arbitrarily large k. One then obtains a number known as the *sequence (Shannon) entropy*, which measures the average uncertainty content per symbol when all symbol dependencies are taken into account. For random sequences there are no dependencies and $H(k)$ remains constant. In the case of an infinite binary sequence of the fair coin-tossing type, $H(k)$ is always equal to 1; if it is a tampered coin with $P(1) = 0.2$, a value of $H(k)$ equal to

$$-0.2 \times \log_2 0.2 - 0.8 \times \log_2 0.8 = 0.72$$

is always obtained.

In the last section, we obtained an entropy estimate for an arbitrarily long text sequence in English. If there are many symbols, we will also need extremely long sequences in order to obtain reliable probability estimates. For instance, in order to estimate the entropy of fourth-order models of English texts, some researchers have used texts with 70 million letters!

The sequence length required for entropy estimation reduces drastically when dealing with binary sequences. For instance, the sequence consisting of indefinitely repeating 01 has entropy 1 for one-symbol blocks and zero entropy for more than one symbol. For two-symbol blocks, only 01 and 10 occur with equal probability. There is not, therefore, an increase in uncertainty when progressing from 1 to 2 symbols. The sequence consisting of indefinitely repeating 00011011 has entropy 1 for blocks of 1 and 2 symbols. However, for 3 symbols, the entropy falls to 0.5, signaling an uncertainty drop for 3-symbol blocks, and it falls again to 0.25 for 4-symbol blocks. As a matter of fact, the reader may confirm that all 2-symbol combinations are present in the sequence, whereas there are 2 combinations of 3 symbols and 8 of 4 symbols that never occur (and there is a deficit for other combinations).

It is interesting to determine and compare the entropy values of several binary sequences. Table 9.3 shows estimated values for several sequences (long enough to guarantee acceptable estimates) from $H(1)$ up to $H(3)$. The sequences named e, π and $\sqrt{2}$, are the sequences of the fractional parts of these constants, expressed in binary. Here are the first 50 bits of these sequences:

Table 9.3. Estimated entropy values for several sequences. See text for explanation

Sequence	$H(1)$	$H(2)$	$H(3)$
e	1	1	0.9999
π	0.9999	0.9999	0.9999
$\sqrt{2}$	1	1	1
Normal number	1	1	1
Quantum number	1	0.9998	0.9997
Quadratic iterator	0.9186	0.6664	0.6663
Normal FHR	0.9071	0.8753	0.8742
Abnormal FHR	0.4586	0.4531	0.4483

e	01000101100101101000100000100011011101000100101011
π	01010000011011001011110111011010011100111100001110
$\sqrt{2}$	00100100001101000011000010100011111111101100011000

The other sequences are as follows:

- normal number refers to Champernowne's binary number mentioned in Chap. 6,
- quantum number is a binary sequence obtained with the quantum generator mentioned in Chap. 8 (atmospheric noise),
- quadratic iterator is the chaotic sequence shown in Fig. 7.14a, expressed in binary,
- normal FHR is a sequence of heart rate values (expressed in binary) of a normal human fetus sleeping at rest (FHR stands for fetal heart rate),
- abnormal FHR is the same for a fetus at severe risk.

Note the irregularity of e, π, $\sqrt{2}$, normal number and quantum number according to the entropy measure: constant $H(k)$ equal to 1. The FHR sequences, referring to fetal heart rate, both look like coin-tossing sequences with tampered coins, where the abnormal FHR is far more tampered with than the normal FHR. Finally, note that the quadratic iterator sequence is less irregular than the normal fetus sequence!

The American researcher Steve Pincus proposed, in 1991, a new irregularity/chaoticity measure combining the ideas of sequence entropy and autocorrelation. In this new measure, called *approximate entropy*, the probabilities used to calculate entropy, instead of referring to the possible k-symbol combinations, refer to occurrences of autocorrelations of k-symbol segments, which are below a specified threshold. This new measure has the important advantage of not demanding very long sequences in order to obtain an adequate estimate. For this reason, it has been successfully applied to the study of many practical problems, e.g., variability of muscular tremor, variability of breathing flow, anesthetic effects in electroencephalograms, assessment of the complexity of information transmission in the human brain, regularity of hormonal secretions, regularity of genetic sequences, dynamic heart rate assessment, early detection of the sudden infant death syndrome, the effects of aging on the heart rate, and characterization of fetal heart rate variability.

9.6 Algorithmic Complexity

We saw that the e, π, and $\sqrt{2}$ sequences, as well as the sequence for Champernowne's normal numbers, have maximum entropy. Should we, on the basis of this fact, consider them to be random? Note that all these sequences obey a perfectly defined rule. For instance, Champernowne's number consists of simply generating the successive blocks of k bits, which can easily be performed by (binary) addition of 1 to the preceding block. The digit sequences of the other numbers can also be obtained with any required accuracy using certain formulas, such as:

$$e = 1 + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \frac{1}{4!} + \frac{1}{5!} + \frac{1}{6!} + \frac{1}{7!} + \frac{1}{8!} + \cdots \;,$$

$$\frac{\pi}{2} = 1 + \frac{1}{3}\left(1 + \frac{2}{5}\left(1 + \frac{3}{7}\left(1 + \frac{4}{9}(1 + \cdots)\right)\right)\right) \;,$$

$$\sqrt{2} = 1 + \cfrac{1}{2 + \cfrac{1}{2 + \cfrac{1}{2 + \cdots}}} \;.$$

That is, in spite of the 'disorder' one may detect using the entropy measures of the digit sequences of these numbers (the empirical evidence that they are Borel normal numbers, as mentioned in Chap. 6), each digit is perfectly predictable. This being so it seems unreasonable to qualify such numbers as random. In fact, for a random sequence one would not expect any of its digits to be predictable. In around 1975, the mathematicians Andrei Kolmogorov, Gregory Chaitin and Ray Solomonof developed, independently of each other, a new theory for classifying a number as random, based on the idea of its computability. Suppose we wished to obtain the binary sequence consisting of repeating 01 eight times: 0101010101010101. We may generate this sequence in a computer by means of a program with the following two instructions:

REPEAT 8 TIMES: PRINT 01 .

Imagine that the REPEAT and PRINT instructions are coded with 8 bits and that the '8 TIMES' and '01' arguments are coded with 16 bits. We then have a program composed of 48 bits for the task of printing a 16-bit sequence. It would have been more economical if the program were simply: PRINT 0101010101010101, in which case it would have only 24 bits. However, if the sequence is composed of 1000 repetitions of 01, we may also use the above program replacing 8 by 1000 (also codable with 16 bits). In this case, the 48-bit program allows one to *compact* the information of a 2000-bit sequence. Elaborating this idea, Kolmogorov, Chaitin and Solomonof defined the *algorithmic complexity*[1] of a binary sequence s as the length of the smallest program s^*, also a binary sequence, that outputs s (see Fig. 9.8).

In this definition of algorithmic complexity, it does not matter in which language the program is written (we may always translate from one language to another), nor which type of computer is used. (In fact, the definition can be referred to a certain type of minimalist computer, but this is by and large an irrelevant aspect.) In the case where s is a thousand repeats of 01, the algorithmic complexity is clearly low: the rather simple program shown above reproduces the 2000 bits of s using only 48 bits. Even for eight repeats of 01, the algorithmic complexity may be low in a computer where the instructions are coded with a smaller number of bits. Instruction codes represent a constant surplus, dependent on the computer being used, which we may neglect

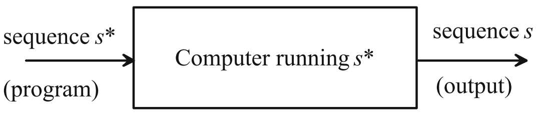


Fig. 9.8. Computer generation of a sequence

[1] Not to be confused with computational complexity, which refers to an algorithm's computational efficiency in run time.

when applying the definition; one has only to take a sufficiently long length for s to justify neglecting the surplus.

We may apply the definition of algorithmic complexity to the generation of any number (independently of whether it is expressed in binary or in any other numbering system). Consider the generation of e. Its digits seem to occur in a haphazard way, but as a matter of fact by using the above formula one may generate an arbitrarily large number of digits of e using a much shorter program than would correspond to printing out the whole digit sequence. In brief, the algorithmic complexity of e, and likewise of π and $\sqrt{2}$, is rather low.

From the point of view of algorithmic complexity, a random sequence is one which cannot be described in a more compact way than by simply instructing the computer to write down its successive digits. The instruction that generates the sequence is then the sequence itself (neglecting the PRINT surplus). Thus, *algorithmically random* means irreducible, or incompressible.

Algorithmic complexity has a deep connection with information theory. In fact, consider a binary sequence s with n bits and a certain proportion p of 1s. Then, as we saw previously when talking about entropy-based coding, there exists a sequence s_c of length nH which codes s using Shannon's coding by exploring the redundancies of s. If p is different from 1/2, the entropy is then smaller than 1; hence, $nH < n$. But if s can be compressed we may build the s^* program simply as a decoding of s_c. Neglecting the decoding surplus, the algorithmic complexity is $nH < n$. Therefore, in order for s to be algorithmically random, the proportions of 0s and 1s must be equal. Moreover, every possible equal-length block of 0s and 1s must occur in equal proportions; otherwise, once again by exploring the conditional entropies, one would be able to shorten s^* relative to s. In conclusion, infinite and algorithmically random sequences must have maximum entropy ($\log_2 k$, for sequences using a k-symbol set) and must be normal numbers! However, not all normal numbers are algorithmically random. For instance, Champernowne's number is not algorithmically random.

As mentioned above, the maximum complexity of an n-bit sequence, and therefore the length of its minimal program (the smallest length program outputting s), is approximately equal to n (equal to n, removing instruction surplus). At the other extreme, a sequence composed of either 0s or 1s has minimum complexity, approximately equal to $\log_2 n$, and its pattern is obvious. If $P(0) = 3/4$ and $P(1) = 1/4$, we have a middle situation: the complexity is approximately $nH = 0.8n$ (see Sect. 9.1).

An interesting fact is that it can be demonstrated that there are fewer than $1/2^c$ sequences whose complexity is smaller than $n - c$. For instance, for $c = 10$, fewer than 1/1024 sequences have complexity less than $n - 10$. We may conclude that, if one considers the interval $[0, 1]$ representing infinite sequences of 0s and 1s, as we did in Chap. 6, there is an infinite subset of Borel normal numbers that is algorithmically random. It is then not at all surprising that the quadratic iterator has infinitely many chaotic orbits; a chaotic orbit is its shortest description and its fastest computer. Unfortunately, although there are infinitely many algorithmically random sequences, it is practically impossible to prove whether or not a given sequence is algorithmically random. Incidentally, note that a minimal program s^* must itself be algorithmically random. If it were not, there would then exist another program s^{**} outputting s^*, and we would be able to generate s with a program instructing the generation of s^* with s^{**}, followed by the generation of s with s^*. The latter program would only have a few more bits than s^{**} (those needed for storing s^* and running it); therefore, s^* would not be a minimal program, which is a contradiction.

9.7 The Undecidability of Randomness

The English mathematician Alan Turing (1912–1954), pioneer of computer theory, demonstrated in 1936 a famous theorem known as the *halting problem theorem*. In simple terms it states the following: a general program deciding in finite time whether a program for a finite input finishes running or will run forever (solving the halting problem) cannot exist for all possible program–input pairs. Note that the statement applies to *all* possible program–input (finite input) pairs. One may build programs which decide whether or not a given program comes to a halt. The impossibility refers to building a program which would decide the halting condition for every possible program–input pair. The halting problem theorem is related to the famous theorem about the incompleteness of formal systems, demonstrated by the German mathematician Kurt Gödel (1906–1978): there is no consistent and complete system of axioms, i.e., allowing one to decide whether any statement built with the axioms of the system is true or false.

From these results it is possible to demonstrate the impossibility of proving whether or not a sequence is algorithmically random. To see this, suppose that for every positive integer n, we wish to determine whether there is a sequence whose complexity exceeds n. For that purpose we suppose that we have somehow got a program that can assess the algorithmic complexity of sequences. This program works in such a way that, as soon as it detects a sequence with complexity larger than n, it prints it and stops. The program would then have to include instructions for the evaluation of algorithmic complexity, occupying k bits, plus the specification of n, occupying $\log_2 n$ bits. For sufficiently large n the value of $k + \log_2 n$ is smaller than n (see Appendix A) and we arrive at a contradiction: a program whose length is smaller than n computes the first sequence whose complexity exceeds n. The contradiction can only be avoided if the program never halts! That is, we will never be able to decide whether or not a sequence is algorithmically random.

9.8 A Ghost Number

In 1976, the American mathematician Gregory Chaitin defined a truly amazing number (presently known as Chaitin's constant). No one knows its value (or better, one of its possible values), but we know what it represents. It is a Borel normal number that can be described but not computed; in other words, there is no set of formulas (as for instance the above formulas for the calculation of e, π, or $\sqrt{2}$) and/or rules (as, for instance, the one used to generate Champernowne's number) which implemented by a program would allow one to compute it. This number, denoted by Ω, corresponds to the probability of a randomly chosen program coming to a halt in a given computer. The random choice of the program means that the computer is fed with binary sequences interpreted as programs, and with those sequences generated by a random process of the coin-tossing type. Assuming that the programs are delimited (with begin–end or a prefix indicating the length), the calculation of Ω is made by adding up, for each program of k bits coming to a halt, a contribution of $1/2^k$ to Ω. With this construction one may prove that Ω is between 0 and 1, excluding these values. In fact, 0 would correspond to no program coming to a halt and 1 to all programs halting, and both situations are manifestly impossible.

Imagine that some goddess supplied us the sequence of the first n bits of Ω, which we denote by Ω_n. We would then be able to decide about the halting or no halting of all programs with length shorter than or equal to n, as follows: we would check, for each k between 1 and n, whether any of the possible 2^k programs of length k stopped or not. We would then feed the computer with all sequences of $1, 2, 3, \ldots, n$ bits. In order to avoid getting stuck with the first non-halting program, we would have to run one step of all the programs in shared time. Meanwhile, each halting program contributes $1/2^k$. When the accumulated sum of all these contributions reaches the number Ω_n, which was miraculously given to us, we have solved the halting problem for all programs of length up to n bits (the computation time for not too small values of n would be enormous!). In fact, another interpretation of Ω is the following: if we repeatedly toss a coin, noting down 1 if it turns up heads and 0 if it turns up tails, the probability of eventually reaching a sequence representing a halting program is Ω.

Let us now attempt to determine whether or not Ω_n is algorithmically random. Knowing Ω_n, we also know all programs up to n bits that come to a halt; among them, the programs producing random sequences of length up to and including n bits. If the algorithmic complexity of Ω_n were smaller than n, there would exist a program of length shorter than n outputting Ω_n. As a consequence, using the process described above, the program would produce all random sequences of length n, which is a contradiction. Effectively, Ω is algorithmically random; it is a sort of quintessence of randomness.

Recently, Cristian Calude, Michael Dinneen, and Chi-Kou Shu have computed the first 64 bits of Ω in a special machine (they actually computed 84 bits, but only the first 64 are trustworthy). Here they are:

$$\begin{gathered}0.000\,000\,100\,000\,010\,000\,011\,000\,100\,001\,101\,000\,111\,111\,001\,011\\101\,110\,100\,001\,000\,0\end{gathered}$$

The corresponding decimal number is 0.078 749 969 978 123 844. Besides the purely theoretical interest, Chaitin's constant finds interesting applications in the solution of some mathematical problems (in the area of Diophantine equations).

9.9 Occam's Razor

William of Occam (or Ockham) (1285–1349?) was an English Franciscan monk who taught philosophy and theology at the University of Oxford. He became well known and even controversial (he ended up being excommunicated by Pope John XXII) because he had ideas well ahead of his time, that ran against the traditionalist trend of medieval religious thought. Occam contributed to separating logic from metaphysics, paving the way toward an attitude of scientific explanation. A famous principle used by Occam in his works, known as the principle of parsimony, principle of simplicity, or principle of economy, has been raised to the level of an assessment/comparison criterion for the scientific reasonableness of contending theories. The principle is popularly known as Occam's razor. It states the following: one should not increase the number of explanatory assumptions for a hypothesis or theory beyond what is absolutely necessary (unnecessary assumptions should be shaved off), or in a more compact form, the simpler the explanation, the better it is. When several competing theories are equal in other respects, the principle recommends selecting the theory that introduces the fewest assumptions and postulates the fewest hypothetical entities.

The algorithmic complexity theory brings interesting support for Occam's razor. In fact, a scientific theory capable of explaining a set of observations can be viewed as a minimal program reproducing the observations and allowing the prediction of future observations. Let us then consider the probability $P(s)$ of a randomly selected program producing the sequence s. For that purpose and in analogy with the computation of Ω, each program outputting s contributes $1/2^r$ to $P(s)$, where r is the program length. It can be shown that $P(s) = 1/2^{C(s)}$, where $C(s)$ represents the algorithmic complexity of s. Thus, sequences with low complexity occur more often than those with high complexity. If the assumptions required by a theorem are coded as sequences of symbols (for instance, as binary sequences), the most probable ones are the shortest.

Another argument stands out when considering the production of binary sequences of length n. If they are produced by a chance mechanism, they have probability $1/2^n$ of showing up. Meanwhile, a possible cause for the production of s could be the existence of another sequence s^* of length $C(s)$ used as input in a certain machine. The probability of obtaining s^* by random choice among all programs of length m is $1/2^{C(s)}$. The ratio $2^{-C(s)}/2^{-n} = 2^{n-C(s)}$ thus measures the probability that s occurred due to a definite cause rather than pure luck. Now, if $C(s)$ is much smaller than n, the probability that s occurred by pure luck is very small. One more reason to prefer simple theories!

10

Living with Chance

10.1 Learning, in Spite of Chance

Although immersed in a vast sea of chance, we have so far demonstrated that we are endowed with the necessary learning skills to solve the problems arising in our daily life, including those whose solution is a prerequisite to our survival as a species. Up to now, we have also shown that we are able to progress along the path of identifying the laws (regularities) of nature. But it is not only the human species that possesses learning skills, i.e., the capability of detecting regularities in phenomena where chance is present. A large set of living beings also manifests the learning skills needed for survival. Ants, for instance, are able to determine the shortest path leading to some food source. The process unfolds as follows. Several ants leave their nest to forage for food, following different paths; those that, by chance, reach a food source along the shortest path are sooner to reinforce that path with pheromones, because they are sooner to come back to the nest with the food; those that subsequently go out foraging find a higher concentration of pheromones on the shortest path, and therefore have a greater tendency (higher probability) to follow it, even if they did not follow that path before. There is therefore a random variation of the individual itineraries, but a learning mechanism is superimposed on that random variation – an increased probability of following the path with a higher level of pheromone –, which basically capitalizes on the law of large numbers. This type of social learning, which we shall call *sociogenetic*, is common among insects.

Other classes of animals provide, not only examples of social learning, but also individual learning, frequently depending on a scheme of punishment and reward. For instance, many animals learn to distinguish certain food sources (leaves, fruits, nuts, etc.) according to whether they have a good taste (reward) or a bad taste (punishment). Human beings begin a complex learning process at a tender age, where social and individual mechanisms complement and intertwine with each other. Let us restrict the notion of learning here to the ability to classify events/objects: short or long path, tasty or bitter fruit, etc. There are other manifestations of learning, such as the ability to assign values to events/objects or the ability to plan a future action on the basis of previous values of events/objects. However, in the last instance, these manifestations can be reduced to a classification: a fine-grained classification (involving a large number of classes) in the case of assigning values to events/objects, or a fine-grained and extrapolated classification in the case of future planning.

Many processes of individual learning, which we shall call *ontogenetic*, are similar to the guessing game mentioned in Chap. 9. Imagine a patient with suspicion of cardiovascular disease going to a cardiologist. The event/object the cardiologist has to learn is the classification of the patient in one of several diagnostic classes. This classification depends on the answer to several questions:

- Does the patient have a high cholesterol level?
- Is the patient overweight?
- Is the electrocardiogram normal?
- Does s(he) have cardiac arrhythmia?

We may call this type of learning, in which the learner is free to put whatever questions s(he) wants, *active learning*. As seen in Chap. 9, active learning guarantees that the target object, in a possible domain of n objects (in the case of the cardiologist, n is the number of possible diagnostic classes), is learned after presenting at most $\log_2 n$ questions. (Of course things are not usually so easy in practice. At the start of the diagnostic process one does not have the answer to all possible questions and some of the available answers may be ambiguous.)

We may formalize the guessing game by assuming that one intends to guess a number x belonging to the interval $[0, 1]$, that some person X thought of. To guess the number, the reader L may randomly chose numbers in $[0, 1]$, following the guessing game strategy. Imagine that the number that X thought of is 0.3 and that the following sequence of questions and answers takes place:

L: Is it greater than 0.5?

X: No.

L: Is it greater than 0.25?

X: Yes.

And so on and so forth. After n questions, L comes to know x with a deviation that is guaranteed not to exceed $1/2^n$. Learning means here classifying x in one of the 2^n intervals of $[0, 1]$ with width $1/2^n$.

Things get more complicated when we are not free to present the questions we want. Coming back to the cardiologist example, suppose that the consultation is done in a remote way and the cardiologist has only an abridged list of the 'answers'. In this case the cardiologist is constrained to learn the diagnosis using only the available data. We are now dealing with a *passive learning* process. Is it also possible to guarantee learning in this case? The answer is negative. (As a matter of fact, were the answer positive, we would probably already have learned everything there is to learn about the universe!)

Passive learning can be exemplified by learning the probability of turning up heads in the coin-tossing game, when the results of n throws are known. L is given a list of the results of n throws and L tries to determine, on the basis of this list, the probability P of heads, a value in $[0, 1]$. For this purpose L determines the frequency f of heads. The learning of P is obviously not guaranteed, in the sense that the best one can obtain is a degree of certainty, for some tolerance of the deviation between f and P. In passive learning there is therefore a new element: there is no longer an absolute certainty of reaching a given deviation between what we have learned and the ideal object/value to be learned – as when choosing a number x in $[0, 1]$. Rather, we are dealing with degrees of certainty (confidence intervals) associated with the deviations. When talking about Bernoulli's law of large numbers in Chap. 4, we saw how to establish a minimum value for n for probability estimation. The dotted line in Fig. 10.1 shows the minimum values of n when learning a probability, for several values of the deviation tolerance between f and P, and using a 95% certainty level.

Let us once again consider learning a number x in the interval $[0, 1]$, but in a passive way this time. Let us assume that X has created a sequence of random numbers, all with the same distribution in $[0, 1]$ and independent of each other, assigning a label or code according to the numbers belonging (say, label 1) or not (say, label 0) to $[0, x]$. The distribution law for the random numbers is assumed unknown to L. To

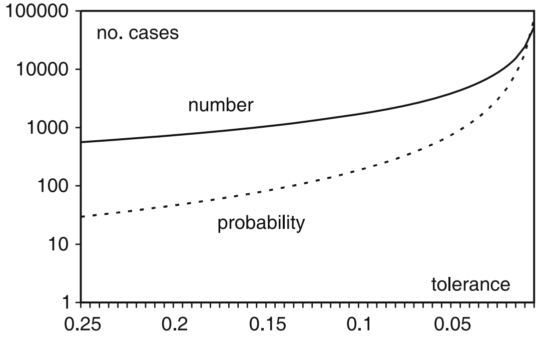


Fig. 10.1. Learning curves showing the minimum number of cases (logarithmic scale) needed to learn a probability (*dotted curve*) and a number (*solid curve*) with a given tolerance and 95% certainty

fix ideas, suppose that X thought of $x = 0.3$ and L was given the list:

$$(0.37, 0) \qquad (0.29, 1) \qquad (0.63, 0) \qquad (0.13, 1) \; .$$

A list like this one, consisting of a set of labeled points (the objects in this case), is known as a *training set*. On the basis of this training set, the best that L can do is to choose the value 0.29, with a deviation of 0.01. Now suppose that the random numbers were such that L received the list:

$$(0.78, 0) \qquad (0.92, 0) \qquad (0.83, 0) \qquad (0.87, 0) \; .$$

The best that L can do now is to choose 0.78, with a deviation of 0.48. In general, assuming that the distribution of the random numbers generated by X is uniform, there is a $1/2^n$ probability that all numbers fall in $[0.5, 1]$. If $x < 0.5$, the supplied list does not help L to learn a number. In the worst case we may get a deviation of 0.5 with probability at least $1/2^n$.

It can be shown that in this learning of number x, in order to get a tolerance interval with 95% certainty, we need more cases than when learning the probability value in the coin-tossing experiment. The minimum values of n in the passive learning of a number, for several values of the deviation tolerance and using 95% certainty, are shown by the solid curve in Fig. 10.1. Here is a concrete example: in order to learn a probability with 0.1 tolerance and 95% certainty, 185 cases are enough. To learn a number under the same conditions requires at least 1 696 trials. Learning a number when the probability law of the training set is unknown is more difficult than learning a probability. It requires larger training sets in order to be protected against atypical training sets (for instance, all points in $[0.5, 1]$).

Finally, there is a third type of learning which we shall call *phylogenetic*. It corresponds to the mechanisms of natural selection of the species, which grants a higher success to certain genomes to the detriment of others. Learning works as follows here: several gene combinations are generated inside the same species or phylum, via mutations and hereditary laws. As a consequence, a non-homogeneous population of individuals becomes available where a genomic variability component is present, as well as the individual traits that reflect it. Some individuals are better off than others in the prosecution of certain vital goals, tending therefore to a better success in genomic transmission to the following generation. Thus, learning corresponds here to better performance in the face of the external environment.

In all these learning models, chance phenomena are present and, in the last instance, it is the laws of large numbers that make learning possible. In the simplest sociogenetic learning model, we have a homogeneous set of individuals, each of them performing a distinct random experiment. Everything happens as if several individuals were throwing coins, with the throws performed under the same conditions. Here learning resembles the convergence of an 'ensemble average'. The more individuals there are, the better the learning will be.

In ontogenetic learning, we have a single individual carrying out or being confronted with random experiments as time goes by. This is the case of 'temporal average' convergence that we have spoken about so often. The more experiments there are, the better the learning will be.

In phylogenetic learning, the individuals are heterogeneous. We may establish an analogy with the coin-tossing case by assuming a living being, called a 'coin', whose best performance given the external environment corresponds to fairness: $P(\text{heads}) = 1/2$. At some instant there is a 'coin' population with different genes to which correspond different values of $P(\text{heads})$. In the next generation, the 'coins' departing from fairness get a non-null probability of being eliminated from the population, and with higher probability the more they deviate from fairness. Thus, in the following generation the distribution of $P(\text{heads})$ is different from the one existing in the previous generation. We thus have a temporal convergence of the average resembling the one in Sect. 6.4. Meanwhile, the decision about which 'coins' to eliminate from each generation depends on the departure of $P(\text{heads})$ from the better performance objective (fairness); that is, it depends on 'ensemble average' convergence.

These different learning schemes are currently being used to build so-called intelligent systems. Examples are *ant colony optimization algorithms* for sociogenetic learning, *genetic algorithms* for phylogenetic learning, and *artificial neural networks* for ontogenetic learning.

10.2 Learning Rules

In the preceding examples, learning consisted in getting close to a specified goal (class label), close enough with a given degree of certainty, after n experiments, that is, for a training set having n cases. This notion of learning is intimately related to the Bernoulli law of large numbers and entirely in accordance with common sense. When a physician needs to establish a diagnosis on the basis of a list of observations, s(he) uses a specific rule (or set of rules) with which s(he) expects to get close to the exact diagnosis with a given degree of certainty. This rule tries to value or weight the observations in such a way as to get close to an exact diagnosis – which may be, for instance, having or not having some disease – with a degree of certainty that increases with the number of diagnoses performed (the physician's training set). The observations gathered in n clinical consultations and the evidence obtained afterwards about whether the patient really had some disease constitute the physician's training set. As another example, when learning how to pronounce a word, say in French, by pronouncing it n times and receiving the comments of a French teacher, we expect to get close to the correct pronunciation, with a given degree of certainty, as n increases. For that purpose, we use some rule (or set of rules) for adjusting the pronunciation to ensure the above-mentioned convergence, in probability, to the correct pronunciation.

In all examples of task learning, there is the possibility of applying one rule to the same observations (supposedly a learning rule), chosen from several possible rules. Let us go back to the problem of determining a number x in $[0, 1]$, in order to illustrate this aspect. We may establish a practical correspondence for this problem. For instance, the practical task of determining the point of contact of two underground structures (or rocks) on the basis of drilling results along some straight line. The drillings are assumed to be made at random points and with uniform distribution in some interval, as illustrated in Fig. 10.2. They reveal whether, at the drilling point, we are above structure 1 or above structure 2. If we are able to devise a rule behaving like the Bernoulli law of large numbers in the determination of the contact point, we say that the rule is a *learning rule* for the task at hand.

There are several feasible rules that can serve as learning rules for this task. Let d be the estimate of the unknown contact point Δ supplied by the rule. If the rule is $d = x_{\mathrm{M}}$, that is, select the maximum point (revealed as being) from structure 1 (rule 1), or $d = x_{\mathrm{m}}$, select the minimum point from structure 2 (rule 2), or again $d = (x_{\mathrm{M}} + x_{\mathrm{m}})/2$

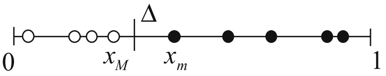


Fig. 10.2. Drillings in some interval, designated by $[0, 1]$, aiming to determine the contact point Δ between structure 1 (*open circles*) and structure 2 (*solid circles*)

(rule 3), the rule is a learning rule, as illustrated in Fig. 10.3. The three rules produce a value d converging in probability (Bernoulli law of large numbers) to Δ.

We now consider another version of the problem of finding a point on a line segment. In contrast to what we had in the above problem, where there was a separation between the two categories of points, we now suppose that this separation does not exist. Let us concretize this scenario by considering a clinical analysis for detecting the presence or absence of some disease. Our intention is to determine which value, produced by the analysis, separates the two classes of cases: the negative ones (without disease) and the positive ones (with disease), which we denote N and P, respectively. Usually, clinical analyses do not allow a total class separation, and instead one observes an overlap of the respective probabilistic distributions. Figure 10.4 shows an example of this situation.

In Fig. 10.4 we assume equal-variance normal distributions for the two classes of diagnosis. Whatever the threshold value d we may choose on the straight line, we will always have to cope with a non-null clas-

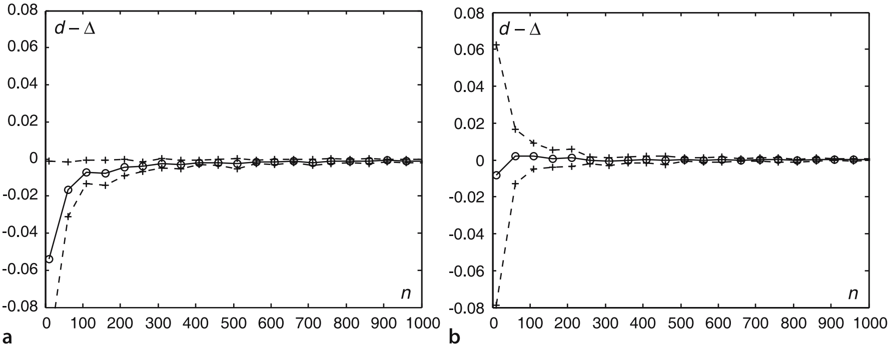


Fig. 10.3. Average deviation $d - \Delta$ curves (*solid lines*) with ±1 standard deviation bounds (*dotted lines*), depending on the training set size n, computed in 25 simulations for $\Delta = 0.5$ and two learning rules: (**a**) $d = x_{\mathrm{M}}$, (**b**) $d = (x_{\mathrm{M}} + x_{\mathrm{m}})/2$

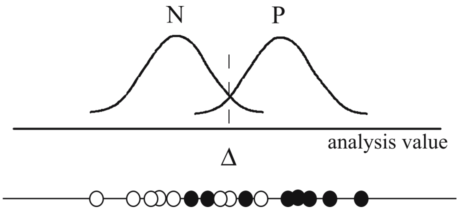


Fig. 10.4. *Upper*: Probability densities of a clinical analysis value for two classes, N and P. *Lower*: Possible training set

sification error. In the case of Fig. 10.4, the best value for d – the one yielding a minimum error – is the one indicated in the figure (as suggested by the symmetry of the problem): $d = \Delta$. The wrongly classified cases are the positive ones producing analysis values below Δ and the negative ones producing values above Δ. Unfortunately, in many practical problems where one is required to learn the task 'what is the best d separating positive from negative cases?', the probability distributions are unknown. We are only given a training set like the one represented at the bottom of Fig. 10.4. What is then the behavior of the previous rules 1, 2 and 3? Rules 1 and 2 for infinite tailed distributions as in Fig. 10.4 converge to values that are clearly far away from Δ, and hence they do not behave as learning rules in this case. In fact, in this case we will get a non-null probability of obtaining d values arbitrarily far from Δ. As for rule 3, it is still a learning rule for symmetrical distributions, since it will lead on average and for the wrongly classified cases to a balance between positive and negative deviations from Δ. Had we written rule 3 as $d = (x_{\mathrm{M}} + x_{\mathrm{m}})/3$, that is, changing the weight of the rule *parameters* $(x_{\mathrm{M}}, x_{\mathrm{m}})$ from $(1/2, 1/2)$ to $(1/3, 2/3)$, the rule would no longer be a learning rule. In fact, even with parameters $(1/2, 1/2)$, rule 3 is not, in general, a learning rule.

There are, however, more sophisticated rules that behave as learning rules for large classes of problems and distributions. One such rule is based on a technique that reaches back to Abraham de Moivre. It consists in adjusting d so as to minimize the sum of squares of the deviations for the wrongly classified cases. Another rule is based on the notion of entropy presented in the last chapter. It consists in adjusting d so as to minimize the entropy of the classification errors and, therefore, their degree of disorder around the null error. (For certain classes of problems, my research team has shown that, surprisingly enough, it may be preferable to maximize instead of minimizing the error entropy.)

10.3 Impossible Learning

As strange as it may look at first sight, there exist scenarios where learning is impossible. Some have to do with the distribution of the data; others correspond to stranger scenarios. Let us first take a look at an example of impossible learning for certain data distributions. Let us again consider the previous example of finding out a threshold value Δ separating two classes, N and P. The only difference is that we now use the following rule 4: we determine the averages of the values for classes N and P, say m_{N} and m_{P}, and thereafter their average, that is,

$$d = \frac{m_{\mathrm{N}} + m_{\mathrm{P}}}{2} \; .$$

This is a learning rule for a large class of distributions, given the convergence in probability of averages. But we now consider the situation where the distributions in both classes are Cauchy. As we saw in Chap. 6, the law of large numbers is not applicable to Cauchy distributions, and the convergence in probability of averages does not hold in this case. The situation we will encounter is the one illustrated in Fig. 10.5. Learning is impossible.

There are other and stranger scenarios where learning is impossible. Let us examine an example which has to do with expressing numbers of the interval $[0, 1]$ in binary, as we did in the section on Borel's normal numbers in Chap. 6 and illustrated in Fig. 6.9. Let us assume that X chooses an integer n that L has to guess using active learning. L does not know any rule that might lead X to favor some integers over others. For L everything happens as if X's choice were made completely at random. Since the aim is to learn/guess n, the game stipulates that

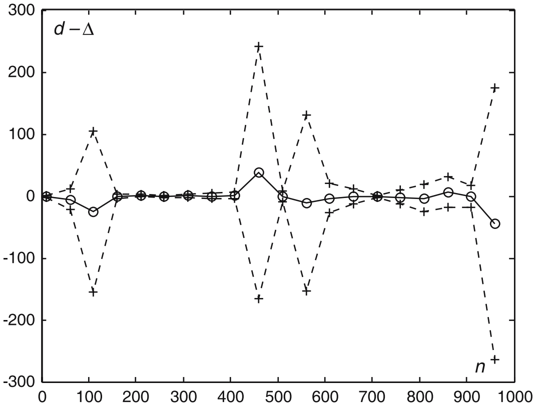


Fig. 10.5. The application of rule 4, under the same conditions as in Fig. 10.3, for two classes with Cauchy distributed values

L should ask X if a certain number in the interval $[0,1]$ contains the binary digit d_n. For instance, the number 0.3 represented in binary by 0.01001100110011... (repetition of 0011) does not contain d_3, 0.001, which corresponds in decimal to 0.125. Here is an example of questions and answers when X chooses $n = 3$ (d_3):

> L: 0.1?
>
> X : 0 [(the binary representation of) 0.1 does not contain (the binary representation of) 0.125].
>
> L: 0.2?
>
> X : 1 (0.2 contains 0.125).
>
> L: 0.3?
>
> X : 0 (0.3 does not contain 0.125).
>
> L: 0.22?
>
> X : 1 (0.22 contains 0.125).

The question now is: will L be able to guess n after a finite number of questions?

We first note that whatever d_i the learner L decides to present as a guess, s(he) always has 50% probability of hitting upon the value chosen by X. This is due to the fact that, for a uniformly distributed x in $[0,1]$, the average of the d_i values for any selected value of x converges to 0.5 independently of the chosen d_i. In other words, any random choice made by L has always 50% probability of a hit. Is it possible that, by using a sufficiently 'clever' rule, we will be able to substantially improve over this 50% probability? The surprising aspect comes now. Consider the average of the absolute differences between what is produced by the rule and what is produced by the right choice. In the above example, the values produced by the right choice 0.3 correspond to the 0101 sequence of answers. Supposing that L, using some rule, decided 0.25, that is d_2, we would get the 0010 sequence. The average of the absolute differences between what is produced by the rule and what is produced by the right choice is $(0 + 1 + 1 + 1)/4 = 0.75$. Now, this value always converges to 0.5 with an increasing number of questions because the difference between *any* pair (d_n, d_i) encloses a 0.5 area (as the reader may check geometrically). We may conclude that either L makes a hit, and the error is then zero, or s(he) does not make a hit, and the error is then still 50%. There is no middle term in the learning process, as happens in a vast number of scenarios, including the previous examples. Is it at least possible to arrive at a sophisticated rule allowing us to reach the right solution in a finite number of steps? The answer is negative. It is impossible to learn n from questions about the d_i set.

10.4 To Learn Is to Generalize

Let us look back at the example in Fig. 10.4. In a previous section we were interested in obtaining a learning rule that would determine a separation threshold with minimum classification error. That is, we are making the assumption that the best *type of classifier* for the two classes (N and P) is a point. If we use rule 3, we may summarize the learning task in the following way:

- Task to be learned: classify the value of a single variable (e.g., clinical analysis) in one of two classes.
- Classifier type: a number (threshold) x.
- Learning rule: rule 3.

This is a *deductive approach* to the classification problem. Once the type of classifier has been decided, as well as the weighting of its parameters, we apply the classifier in order to classify/explain new facts. We progress from the general (classifier type) to the particular (explaining the observations). What can provide us with the guarantee that the surmised classifier type is acceptable? It could well happen that the diagnostic classes were 'nested', as shown in Fig. 10.6. In the case of Fig. 10.6a, if we knew the class configuration, instead of considering one threshold we would consider three and the classifier would be: If x is below d_1 or between d_2 and d_3, it belongs to class N; otherwise, it belongs to class P. In the case of Fig. 10.6b, we would consider five thresholds and the following classifier: If x is below d_1 or between d_2 and d_3, or between d_4 and d_5, then x belongs to class N; otherwise, it belongs to class P. Classifiers involving one, three or five thresholds are linear in the respective parameters. In other scenarios it could well happen that the most appropriate classifier is quadratic, as in Fig. 10.7, where we separate the two classes with: If x^2 is greater than d, then x belongs to class N; otherwise, it belongs to class P. We may go on imagining the need to use ever more complex types of classifier, in particular when a larger number of parameters is involved.

Meanwhile, we never know, in general, if we are confronted with the configuration of Figs. 10.6a, 10.6b, 10.7, or some other situation. All we usually have available are simply the observations constituting

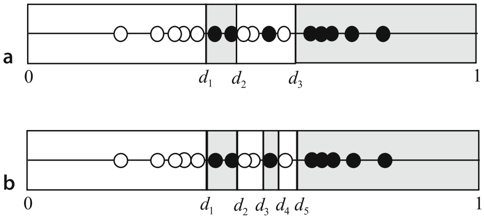


Fig. 10.6. Situation of classes N (*white*) and P (*grey*) separated by three (**a**) and five (**b**) thresholds. The training set is the same as in Fig. 10.4

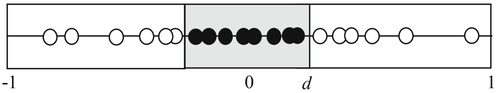


Fig. 10.7. Quadratic classifier

our training set. How can we then choose which type of classifier to use? In this situation we intend to determine the classifier type and its parameters simply on the basis of the observations, without any assumptions, or with only a minimum of assumptions. Learning is now *inductive*. We progress from the particular (the observations) to the general (classifier type). The answer to the question of what type of classifier to use is related to the difficult *generalization capability* issue of *inductive learning*.

Suppose from now on that in the above example the distributions of the analysis value x for the two classes of diagnosis did correspond, in fact, to the four-structure situation of Fig. 10.6a. However, the person who designed the classifier does not know this fact; s(he) only knows the training set. The classification rule based on three thresholds is then adequate. The only thing we have to do is to use a learning rule for this type of classifier, which adjusts the three thresholds to the three unknown values Δ_1, Δ_2, and Δ_3, in order to minimize the number of errors. Let us further suppose that in the case of Fig. 10.6a we were fortunate with the training set supplied to us and we succeeded in determining d_1, d_2, and d_3 quite close to the true thresholds Δ_1, Δ_2, and Δ_3. Our solution would then behave equally well, on average terms, for other (new) sets of values, affording the smallest probability of error. The designed classifier would allow *generalization* and would generalize well (with the smallest probability of error, on average). We could have been unfortunate, having to deal with an atypical training set, leading to the determination of d_1, d_2, and d_3 far away from the true values. In this case, the classifier designed on the basis of the training set would not generalize to our satisfaction.

In summary, we can never be certain that a given type of classifier allows generalization. The best we can have is some degree of certainty. But what does this degree of certainty depend on? In fact, it depends on two things: the number of elements of the training set and the *complexity* of the classifier (number of parameters and type of classifier). As far as the number of training set elements is concerned, it is easily understandable that, as we increase that number, the probability of the training set being typical also increases, in the sense of the strong law of large numbers mentioned in Chap. 6.

Let us now focus on the complexity issue. In the present example the complexity has to do with the number of thresholds used by the classifier. The simpler the classifier, the easier the generalization. For the above example, the classifier with a single threshold does not need such large training sets as the three-threshold classifier, in order to generalize; however, its generalization is always accompanied by a non-null classification error. It generalizes, but does so rather poorly. We are thus in a sub-learning situation. At the other extreme, the five-threshold classifier does not generalize in any reasonable way, since its performance behaves erratically; in training sets like those of Fig. 10.6, it attains zero errors, whereas in new data sets it may produce very high error rates. We are thus in an over-learning situation, where the classifier, given its higher complexity, gets so stuck to the training set details that it will not behave in a similar way when applied to other (new) data sets. We may say that the one-threshold classifier only describes the coarse-grained structure of the training set, while the five-threshold classifier describes a structure that is too finely grained, very much influenced by the random idiosyncrasies of the training set. The three-threshold classifier is the *best compromise* solution for the learning task. Finding such a best compromise is not always an easy task in practice.

10.5 The Learning of Science

Occam's razor told us that, in the presence of uncertainty, the simplest explanations are the best. Consider the intelligibility of the universe, or in other words the scientific understanding of material phenomena. If, in order to explain a given set of observations of these phenomena, say n observations, we needed to resort to a rule using n parameters, we could say that no intelligibility was present. There is no difference between knowing the rule and knowing the observations. In fact, to understand is to compress. The intelligibility of the universe exists only if infinitely many observations can be explained by laws with a reduced number of parameters. It amounts to having a small program that is able to produce infinitely many outcomes. This is the aim of the 'theory of everything' that physicists are searching for, encouraged by the appreciation that, up to now, simple, low complexity (in the above sense) laws have been found that are capable of explaining a multitude of observations. We will never be certain (fortunately!) of having found the theory of everything. On the other hand, inductively learning the laws of the universe – that is, obtaining laws based on observations alone –, presents the problem of their generalization: can we be certain, with a given level of certainty, that the laws we have found behave equally well for as yet unobserved cases?

The Russian mathematician Vladimir Vapnik developed pioneering work in statistical learning theory, establishing (in 1991) a set of theoretical results which allow one in principle to determine a degree of certainty with which a given classifier (or predictive system) generalizes. Statistical learning theory clarifies what is involved in the inductive learning model, supplying an objective foundation for inductive validity. According to this theory we may say that the objective foundation of inductive validity can be stated in a compact way as follows: in order to have an acceptable degree of certainty about the generalization of our inductive model (e.g., classifier type), one must never use a more complex type of model than is allowed by the available data. Now, one of the interesting aspects of the theory developed by Vapnik is the characterization of the generalizing capability by means of a single number, known as the *Vapnik–Chervonenkis dimension* (often denoted by D_{VC}). A learning model can only be generalized if its D_{VC} is finite, and all the more easily as its value is smaller. There are certain models with infinite D_{VC}. The parameters of such models can be learned so that a complete separation of the values of any training set into the respective classes can be achieved, no matter how complex those values are and even if we deliberately add incorrect class-labeled values to the training set.

The fact that an inductive model must have finite D_{VC} in order to be valid and therefore be generalizable is related to the falsifiability criterion proposed by the Austrian philosopher Karl Popper for demarcating a theory as scientific. As a matter of fact, in the same way as classifiers with infinite D_{VC} will never produce incorrect classifications on any training set (even if deliberately incorrectly labeled observations are added), it can be said that non-scientific theories will never err; they always explain the observations, no matter how strange they are. On the other hand, if scientific laws are used – thus, capable of generalization –, we are in a situation corresponding to the use of finite

D_{VC} classifiers; that is, it is possible to present observational data that is not explainable by the theory (*falsifying* the theory).

Occam's razor taught us to be parsimonious in the choice of the number of parameters of the classifier (more generally, of the learning system). Vapnik's theory goes a step further. It tells us the level of complexity of the classifier (or learning system) to which we may go for a certain finite number of observations, in such a way that learning with generalization is achievable.

10.6 Chance and Determinism: A Never-Ending Story

It seems that God does play dice after all, contradicting Albert Einstein's famous statement. It seems that the decision of where a photon goes or when a beta emission takes place is of divine random origin, as is the knowledge of the infinitely many algorithmically complex numbers, readily available to create a chaotic phenomenon.

The vast majority of the scientific community presently supports the vision of a universe where chance events of objective nature do exist. We have seen that it was not always so. Suffice it to remember the famous statement by Laplace, quoted in Chap. 7. There is, however, a minority of scientists who defend the idea of a computable universe. This idea – perhaps better referred to as a belief, because those that advocate it recognize that it is largely speculative, and not supported by demonstrable facts – is mainly based on some results of computer science: the computing capacity of the so-called cellular automata and the algorithmic compression of information. Cellular automata are somewhat like two- and three-dimensional versions of the quadratic iterator presented in Chap. 7, in the sense that, by using a set of simple rules, they can give rise to very complex structures, which may be used as models of such complex phenomena as biological growth. Algorithmic compression of information deals with the possibility of a simple program being able to generate a sequence that passes all irregularity tests, like those mentioned in Chap. 9 concerning the digit sequences of π, e and $\sqrt{2}$. On the basis of these ideas, it has been proposed that everything in the universe is computable, with these proposals ranging from an extreme vision of a virtual reality universe – a universe without matter, some sort of hologram –, to a less extreme vision of an inanimate universe behaving as if it were a huge computer.

There are of course many relevant and even obvious objections to such visions of a computable universe [42]. The strongest objec-

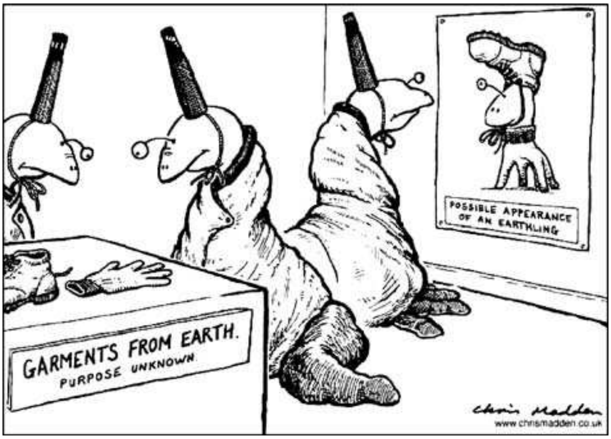


Fig. 10.8. When further observations are required

tion to this neo-deterministic resurrection has to do with the non-computability of pure chance, i.e., of quantum randomness. The most profound question regarding the intelligibility of the universe thus asks whether it is completely ordered or chaotic; in other words, whether chance is a deterministic product of nature or one of its fundamental features. If the universe has infinite complexity, as believed by the majority of scientists, it will never be completely understood. Further, we do not know whether the complexity of a living being can be explained by a simple program. We do not know, for instance, which rules influence biological evolution, in such a way that we may forecast which path it will follow. Are mutations a phenomenon of pure chance? We do not know. Life seems to be characterized by an infinite complexity which we will never be able to learn in its entirety.

As a matter of fact, as we said already in the preface, if there were no chance events, the temporal cause–effect sequences would be deterministic and, therefore, computable, although we might need huge amounts of information for that purpose. It remains to be seen whether one would ever be able to compute those vast quantities of information in a useful time. Notwithstanding, the profound conviction of the scientific community at present is that objective chance is here to stay. We may nevertheless be sure of one thing – really sure with a high degree of certainty – that chance and determinism will forever remain a subject of debate, since objective chance will always remain absolutely beyond the cognition of human understanding.

A

Some Mathematical Notes

A.1 Powers

For integer n (i.e., a whole number, without fractional part), the expression x^n, read x to the power n, designates the result of multiplying x by itself n times. x is called the *base* and n the *exponent*. If n is positive, the calculation of x^n is straightforward, e.g.,

$$2.5^3 = 2.5 \times 2.5 \times 2.5 = 15.625 \; .$$

From the definition, one has

$$x^n \times x^m = x^{n+m} \; , \quad (x^n)^m = x^{nm} \; ,$$

where nm means $n \times m$, e.g.,

$$2.5^5 = 2.5^3 \times 2.5^2 = 2.5 \times 2.5 \times 2.5 \times 2.5 \times 2.5 = 97.65625 \; ,$$

$$(0.5^2)^3 = (0.5)^6 = 0.015\,625 \; .$$

In the book, expressions such as $(2/3)^6$ sometimes occur. We may apply the definition directly, as before, or perform the calculation as

$$(2/3)^6 = 2^6/3^6 = 64/729 \; .$$

Note also that 1^n is 1.

From $x^n \times x^m = x^{n+m}$ comes the convention $x^0 = 1$ (for any x). Moreover, for a negative exponent, $-n$, we have

$$x^{-n} = \frac{1}{x^n} \; ,$$

since

$$x^n \times x^{-n} = x^0 = 1 \; .$$

For example,

$$2^{-3} = \frac{1}{2^3} = \frac{1}{8} \; .$$

Powers with fractions as exponents and non-negative bases are calculated as a generalization of $(x^n)^m = x^{nm}$. For instance, from $(x^{1/2})^2 = x$, one has $x^{1/2} = \sqrt{x}$.

A.2 Exponential Function

The figure below shows the graphs of functions $y = x^4$ and $y = 2^x$. Both are calculated with the above power rules. In the first function, the exponent is constant; the function is a *power function*. In the second function the base is constant; it is an *exponential function*. Figure A.1 shows the fast growth of the exponential function in comparison with the power function. Given any power and exponential functions, it is always possible to find a value x beyond which the exponential is above the power.

Given an exponential function a^x, its growth rate (measured by the slope at each point of the curve) is given by Ca^x, where C is a constant. C can be shown to be 1 when a is the value to which $(1 + 1/h)^h$ tends when the integer h increases arbitrarily. Let us see some values of this quantity:

$$\left(1 + \frac{1}{2}\right)^2 = 1.5^2 = 2.25 \; ,$$

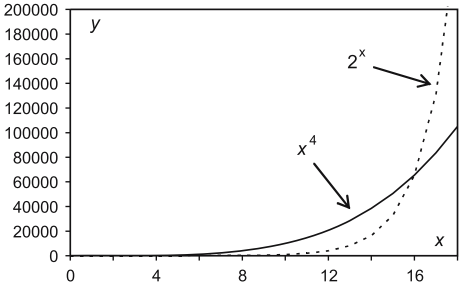


Fig. A.1. Comparing the growth of the exponential function and the power function

$$\left(1+\frac{1}{10}\right)^{10} = 1.1^{10} = 2.593\,742\ ,$$

$$\left(1+\frac{1}{1000}\right)^{1000} = 1.001^{1000} = 2.716\,924\,2\ .$$

These quantities converge to the *natural* or *Napier number*, denoted by e. The value of e to 6 significant figures is 2.718 28. The natural exponential e^x thus enjoys the property that its growth rate is equal to itself.

A.3 Logarithm Function

The logarithm function is the inverse of the exponential function; that is, given $y = a^x$, the logarithm of y in base a is x. The word 'logarithm', like the word 'algorithm', comes from Khwarizm, presently Khiva in Uzbekistan, birth town of Abu Já'far Muhammad ibn Musa Al-Khwarizmi (780–850), author of the first treatise on algebra.

This function is usually denoted by $\log_a$. Here are some examples:

- $\log_2(2.25) = 1.5$, because $1.5^2 = 2.25$.
- $\log_{10}(2.593\,742) = 1.1$, because $1.1^{10} = 2.593\,742$.
- $\log_a(1) = 0$, because as we have seen, $a^0 = 1$ for any a.

The logarithm function in the natural base, viz., $\log_e(x)$, or simply $\log(x)$, is frequently used. An important property of $\log(x)$ is that its growth rate is $1/x$. The graphs of these functions are shown in Fig. A.2.

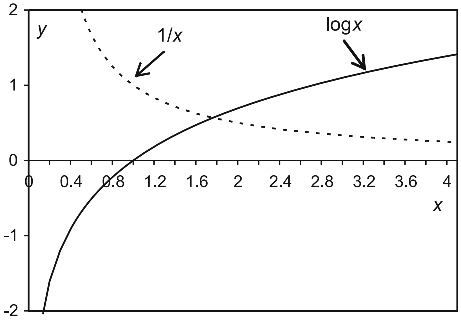


Fig. A.2. The logarithm function

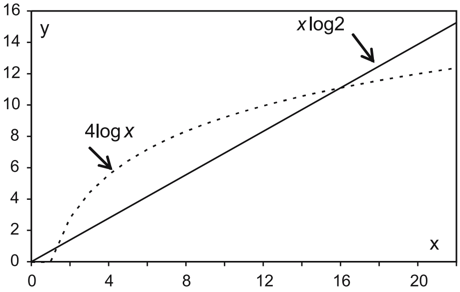


Fig. A.3. The behavior of the functions $\log(x^4)$ and $\log(2^x)$

The following results are a consequence of the properties of the power function:

$$\log(xy) = \log(x) + \log(y) \ , \qquad \log(x^y) = y \log(x) \ .$$

The logarithm function provides a useful transformation in the analysis of fast growth phenomena. For instance, in Fig. A.1 it is not possible to compare the growth of the two functions for very small or very high values of x. If we apply the log transformation, we get

$$\log(x^4) = 4 \log(x) \ , \qquad \log(2^x) = x \log(2) \ ,$$

whose graphs are shown in Fig. A.3. The behavior of the two functions is now graphically comparable. It becomes clearly understandable that, in the long run, any exponential function overtakes any power function.

A.4 Factorial Function

The factorial function of a positive integer number n, denoted $n!$, is defined as the product of all integers from n down to 1:

$$n! = n \times (n-1) \times (n-2) \times \ldots \times 2 \times 1 \ .$$

The factorial function is a very fast-growing function, as can be checked in the graphs of Fig. A.4. In the left-hand graph, $n!$ is represented together with x^4 and 2^x, while the respective logarithms are shown on the right. In the long run, the factorial function overtakes any exponential function.

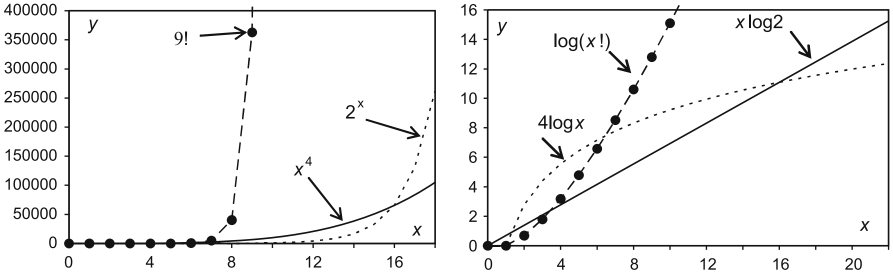


Fig. A.4. *Left*: The functions $n!$, x^4 and 2^x. *Right*: Logarithms of the three functions on the left

A.5 Sinusoids

The sine function $\sin(x)$ (sine of x) is defined by considering the radius OP of a unit-radius circle enclosing the angle x, measured in the anticlockwise direction in Fig. A.5. The function represents the length of the segment AP drawn perpendicularly from P to the horizontal axis.

The length can be either positive or negative, depending on whether the segment is above or below the horizontal axis, respectively. With the angle measured in radians, that is, as an arc length represented by a certain fraction of the unitary radius, and since 2π is the circumference, one has

$$\sin(0) = \sin(\pi) = \sin(2\pi) = 0 \,, \qquad \sin(\pi/2) = 1 \,,$$
$$\sin(3\pi/2) = -1 \,, \qquad \sin(\pi/4) = \sqrt{2}/2 \,.$$

The inverse function of $\sin(x)$ is $\arcsin(x)$ (arcsine of x). As an example, $\arcsin(\sqrt{2}/2) = \pi/4$ (for an arc between 0 and 2π).

Considering point P rotating uniformly around O, so that it describes a whole circle every T seconds, the segment AP will vary as $\sin(2\pi t/T)$, where t is the time, or equivalently $\sin(2\pi f t)$, where $f = 1/T$ is the frequency (cycles per second). If the movement of P starts at the position corresponding to the angle ϕ, called the phase, the segment will vary as $\sin(2\pi f t + \phi)$ (see Fig. 8.8).

The cosine function $\cos(x)$ (cosine of x) corresponds to the segment OA in Fig. A.5. Thus,

$$\cos(0) = \cos(2\pi) = 1 \,, \qquad \cos(\pi) = -1 \,,$$
$$\cos(\pi/2) = \cos(3\pi/2) = 0 \,, \qquad \cos(\pi/4) = \sqrt{2}/2 \,.$$

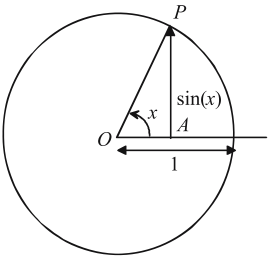


Fig. A.5. Unit circle and definition of the sine and cosine functions

From Pythagoras' theorem, one always has

$$\sin^2(x) + \cos^2(x) = 1 \ ,$$

for any x.

A.6 Binary Number System

In the decimal system which we are all accustomed to, the value of any number is obtained from the given representation by multiplying each digit by the corresponding power of ten. For instance, the digit sequence 375 represents

$$3 \times 10^2 + 7 \times 10^1 + 5 \times 10^0 = 300 + 70 + 5 \ .$$

The rule is the same when the number is represented in the binary system (or any other system base for that matter). In the binary system there are only two digits (commonly called bits): 0 and 1. Here are some examples: 101 has value

$$1 \times 2^2 + 0 \times 2^1 + 1 \times 2^0 = 4 + 1 = 5 \ ,$$

while 0.101 has value

$$1 \times 2^{-1} + 0 \times 2^{-2} + 1 \times 2^{-3} = 1/2 + 1/8 = 0.5 + 0.125 = 0.625 \ .$$

References

The following list of references includes works of popular science and other sources that are not not technically demanding.

1. Alvarez, L.: A pseudo-experience in parapsychology (letter), Science **148**, 1541 (1965)
2. Amaral, L., Goldberger, A., Ivanov, P., Stanley, E.: Modeling heart rate variability by stochastic feedback, Comp. Phys. Communications **121**–**2**, 126–128 (1999)
3. Ambegaokar, V.: *Reasoning about Luck. Probability and Its Uses in Physics*, Cambridge University Press (1996)
4. Bar-Hillel, M., Wagenaar, W.A.: The perception of randomness, Advances in Applied Mathematics **12**, 428–454 (1991)
5. Bassingthwaighte, J.B., Liebovitch, L., West, B.: *Fractal Physiology*, Oxford University Press (1994)
6. Beltrami, E.: *What is Random?*, Springer-Verlag (1999)
7. Bennet, D.J.: *Randomness*, Harvard University Press (1998)
8. Bernstein, P.: *Against the Gods. The Remarkable Story of Risk*, John Wiley & Sons (1996)
9. Blackmore, S.: The lure of the paranormal, New Scientist **22**, 62–65 (1990)
10. Bolle, F.: The envelope paradox, the Siegel paradox, and the impossibility of random walks in equity and financial markets, `www.econ.euv-frankfurto.de/veroeffentlichungen/bolle.html` (2003)
11. Bouchaud, J.-P.: Les lois des grands nombres, La Recherche **26**, 784–788 (1995)
12. Brown, J.: Is the universe a computer? New Scientist **14**, 37–39 (1990)

13. Casti, J.: Truly, madly, randomly, New Scientist **155**, 32–35 (1997)

14. Chaitin, G.: Randomness and mathematical proof, Scientific American **232**, No. 5, 47–52 (1975)

15. Chaitin, G.: Randomness in arithmetic, Scientific American **259**, No. 7, 52–57 (1988)

16. Chaitin, G.: Le hasard des nombres, La Recherche **22**, 610–615 (1991)

17. Chaitin, G.: L'Univers est-il intelligible? La Recherche **370**, 34–41 (2003)

18. Crutchfield, J., Farmer, J.D., Packard, N., Shaw, R.: Chaos, Scientific American **255**, No. 6, 38–49 (1986)

19. Deheuvels, P.: *La probabilité, le hasard et la certitude*, collection Que Sais-Je?, Presses Universitaires de France (1982)

20. Devlin, K.: The two envelopes paradox. Devlin's angle, MAA Online, The Mathematical Association of America, `www.maa.org/devlin/devlin_0708_04.html` (2005)

21. Diaconis, P., Mosteller, F.: Methods for studying coincidences, Journal of the American Statistical Association **84**, 853–861 (1989)

22. Droesbeke, J.-J., Tassi, P.: *Histoire de la statistique*, collection Que Sais-Je?, Presses Universitaires de France (1990)

23. Everitt, B.S.: *Chance Rules. An Informal Guide to Probability, Risk and Statistics*, Springer-Verlag (1999)

24. Feynman, R.: *QED. The Strange Theory of Light and Matter*, Penguin, London (1985)

25. Ford, J.: What is chaos that we should be mindful of? In: *The New Physics*, ed. by P. Davies, Cambridge University Press (1989)

26. Gamow, G., Stern, M.: *Jeux Mathématiques. Quelques casse-tête*, Dunod (1961)

27. Gell-Mann, M.: *The Quark and the Jaguar*, W.H. Freeman (1994)

28. Ghirardi, G.: *Sneaking a Look at God's Cards. Unraveling the Mysteries of Quantum Mechanics*, Princeton University Press (2005)

29. Gindikin, S.: *Tales of Physicists and Mathematicians*, Birkhäuser, Boston (1988)

30. Goldberger, A., Rigney, D., West, B.: Chaos and fractals in human physiology, Scientific American **262**, No. 2, 35–41 (1990)

31. Haken, H., Wunderlin, A.: Le chaos déterministe, La Recherche **21**, 1248–1255 (1990)

32. Hanley, J.: Jumping to coincidences: Defying odds in the realm of the preposterous, The American Statistician **46**, 197–202 (1992)

33. Haroche, S., Raimond, J.-M., Brune, M.: Le chat de Schrödinger se prête à l'expérience, La Recherche **301**, 50–55 (1997)

34. Hénon, M.: La diffusion chaotique, La Recherche **209**, 490–498 (1989)

35. Idrac, J.: *Mesure et instrument de mesure*, Dunod (1960)

36. Jacquard, A.: *Les Probabilités*, collection Que Sais-Je?, Presses Universitaires de France (1974)

37. Kac, M.: What is Random?, American Scientist **71**, 405–406 (1983)

38. Layzer, D.: The arrow of time, Scientific American **233**, No. 6, 56–69 (1975)

39. Mandelbrot, B.: A multifractal walk down Wall Street, Scientific American **280**, No. 2, 50–53 (1999)

40. Martinoli, A., Theraulaz, G., Deneubourg, J.-L.: Quand les robots imitent la nature, La Recherche **358**, 56–62 (2002)

41. McGrew, T.J., Shier, D., Silverstein, H.S.: The two-envelope paradox resolved, Analysis **57.1**, 28–33 (1997)

42. Mitchell, M.: L'Univers est-il un calculateur? Quelques raisons pour douter, La Recherche **360**, 38–43 (2003)

43. Nicolis, G.: Physics of far-from-equilibrium systems and self-organisation. In: *The New Physics*, ed. by P. Davies, Cambridge University Press (1989)

44. Ollivier, H., Pajot, P.: La décohérence, espoir du calcul quantique, La Recherche **378**, 34–37 (2004)

45. Orléan, A.: Les désordres boursiers, La Recherche **22**, 668–672 (1991)

46. Paulson, R.: Using lottery games to illustrate statistical concepts and abuses, The American Statistician **46**, 202–204 (1992)

47. Peitgen, H.-O., Jürgens, H., Saupe, D.: *Chaos and Fractals. New Frontiers of Science*, Springer-Verlag (2004)

48. Perakh, M.: Improbable probabilities, `www.nctimes.net/~mark/bibl_science/probabilities.htm` (1999)

49. Pierce, J.: *An Introduction to Information Theory. Symbols, Signals and Noise*, Dover Publications (1980)

50. Pincus, S.M.: Assessing serial irregularity and its implications for health, Proc. Natl. Acad. Sci. USA **954**, 245–267 (2001)

51. Polkinghorne, J.C.: *The Quantum World*, Princeton University Press (1984)

52. Postel-Vinay, O.: L'Univers est-il un calculateur? Les nouveaux démiurges, La Recherche **360**, 33–37 (2003)

53. Rae, A.: *Quantum Physics: Illusion or Reality?*, 2nd edn., Cambridge University Press (2004)

54. Rittaud, B.: L'Ordinateur à rude épreuve, La Recherche **381**, 28–35 (2004)

55. Ruelle, D.: *Hasard et chaos*, collection Points, Editions Odile Jacob (1991)

56. Salsburg, D.: *The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century*, W.H. Freeman (2001)

57. Schmidhuber, J.: A computer scientist's view of life, the universe, and everything, `http://www.idsia.ch/~juergen/everything` (1997)

58. Shimony, A.: Conceptual foundations of quantum mechanics. In: *The New Physics*, ed. by P. Davies, Cambridge University Press (1989)

59. Steyer, D.: *The Strange World of Quantum Mechanics*, Cambridge University Press (2000)

60. Taylor, J.: *An Introduction to Error Analysis*, 2nd edn., University Science Books (1997)

61. Tegmark, M., Wheeler, J.A.: 100 years of quantum mysteries, Scientific American **284**, 54–61 (2001)

62. UCI Machine Learning Repository (The Boston Housing Data): Website `www.ics.uci.edu/~mlearn/MLSummary.html`

63. West, B., Goldberger, A.: Physiology in fractal dimensions, American Scientist **75**, 354–364 (1987)

64. West, B., Shlesinger, M.: The noise in natural phenomena, American Scientist **78**, 40–45 (1990)

65. Wikipedia: Ordre de grandeur (nombre), `fr.wikipedia.org/wiki/Ordre_de_grandeur_(nombre)` (2005)