1 An Introduction to Empirical Modeling

1.1 Introduction

Empirical modeling, broadly speaking, refers to the process, methods, and strategies grounded on statistical modeling and inference whose primary aim is to give rise to "learning from data" about stochastic observable phenomena, using *statistical models.* Real-world phenomena of interest are said to be "stochastic," and thus amenable to statistical modeling, when the data they give rise to exhibit *chance regularity patterns*, irrespective of whether they arise from passive observation or active experimentation. In this sense, empirical modeling has three crucial features:

(a) it is based on observed data that exhibit chance regularities;
(b) its cornerstone is the concept of a statistical model that decribes a probabilistic generating mechanism that could have given rise to the data in question;
(c) it provides the framework for combining the statistical and substantive information with a view to elucidating (understanding, predicting, explaining) phenomena of interest.

Statistical vs. substantive information. Empirical modeling across different disciplines involves an intricate blending of *substantive* subject matter and *statistical information*. The substantive information stems from a theory or theories pertaining to the phenomenon of interest that could range from simple conjectures to intricate *substantive* (structural) models. Such information has an important and multifaceted role to play by demarcating the crucial aspects of the phenomenon of interest (suggesting the relevant variables and data), as well as enhancing the learning from data when it meliorates the statistical information without belying it. In contrast, statistical information stems from the *chance regularities* in data. Scientific knowledge often begins with substantive conjectures based on subject matter information, but it becomes knowledge when its veracity is firmly grounded in real-world data. In this sense, success in "learning from data" stems primarily from a harmonious blending of these two sources of information into an empirical model that is both statistically and substantively "adequate"; see Sections 1.5 and 1.6.

Empirical modeling as curve-fitting. The current traditional perspective on empirical modeling largely ignores the above distinctions by viewing the statistical problem as "quantifying theoretical relationships presumed true." From this perspective, empirical modeling is viewed as a *curve-fitting problem*, guided primarily by goodness-of-fit. The substantive model is often imposed on the data in an attempt to quantify its unknown parameters. This treats the substantive information as established knowledge, and not as tentative conjectures to be tested against data. The end result of curve-fitting is often an estimated model that is misspecified, both statistically (invalid probabilistic assumptions) and substantively; it doesn't elucidate sufficiently the phenomenon of interest. This raises a thorny problem in philosophy of science known as *Duhem's conundrum* (Mayo, 1996), because there is no principled way to distinguish between the two types of misspecification and apportion blame. It is argued that the best way to address this impasse is (i) to disentangle the statistical from the substantive model by unveiling the probabilistic assumptions (implicitly or explicitly) imposed on the data (the statistical model) and (ii) to separate the modeling from the inference facet of empirical modeling. The *modeling facet* includes specifying and selecting a statistical model, as well as appraising its adequacy (the validity of its probabilistic assumptions) using misspecification testing. The *inference facet* uses a statistically adequate model to pose questions of substantive interest to the data. Crudely put, conflating the modeling with the inference facet is analogous to mistaking the process of constructing a boat to preset specifications with sailing it in a competitive race; imagine trying to construct the boat while sailing it in a competitive race.

Early cautionary note. It is likely that some scholars in empirical modeling will mock and criticize the introduction of new terms and distinctions in this book as "mounds of gratuitous jargon," symptomatic of an ostentatious display of pedantry. As a pre-emptive response to such critics, allow me to quote R. A. Fisher's 1931 reply to Arne Fisher's [American mathematician/statistician] complaining about his

"introduction in statistical method of some outlandish and barbarous technical terms. They stand out like quills upon the porcupine, ready to impale the sceptical critic. Where, for instance, did you get that atrocity, a *statistic*?"

His serene response was:

I use special words for the best way of expressing special meanings. Thiele and Pearson were quite content to use the same words for what they were estimating and for their estimates of it. Hence the chaos in which they left the problem of estimation. Those of us who wish to distinguish the two ideas prefer to use different words, hence 'parameter' and 'statistic'. **No one who does not feel this need is under any obligation to use them**. Also, to Hell with pedantry. (Bennett, 1990, pp. 311–313) [emphasis added]

A bird's-eye view of the chapter. The rest of this chapter elaborates on the crucial features of empirical modeling (a)–(c). In Section 1.2 we discuss the meaning of *stochastic observable phenomena* and why such phenomena are amenable to empirical modeling. Section 1.3 focuses on the relationship between data from stochastic phenomena and *statistical models*. Section 1.4, discusses several important issues relating to *observed data*, including their different *measurement scales*, *nature*, and *accuracy*. In Section 1.5 we discuss the important notion of statistical adequacy: whether the postulated statistical model "accounts fully for"

the statistical systematic information in the data. Section 1.6 discusses briefly the connection between a statistical model and the substantive information of interest.

1.2 Stochastic Phenomena: A Preliminary View

This section provides an intuitive explanation for the notion of a stochastic phenomenon as it relates to the concept of a statistical model, discussed in the next section.

1.2.1 Chance Regularity Patterns

The *chance regularities* denote patterns that are usually revealed using a variety of graphical techniques and careful preliminary data analysis. The essence of *chance regularity*, as suggested by the term itself, comes in the form of two entwined features:

chance an inherent uncertainty relating to the occurrence of particular outcomes;
regularity discernible regularities associated with an aggregate of many outcomes.

TERMINOLOGY: The term "chance regularity" is used in order to avoid possible confusion with the more commonly used term "randomness."

At first sight these two attributes might appear to be contradictory, since "chance" is often understood as the *absence* of order and "regularity" denotes the *presence* of order. However, there is no contradiction because the "disorder" exists at the level of individual outcomes and the order at the aggregate level. The two attributes should be viewed as inseparable for the notion of chance regularity to make sense.

Example 1.1 To get some idea about "chance regularity" patterns, consider the data given in Table 1.1.

Table 1.1 *Observed data*

3	10	11	5	6	7	10	8	5	11	2	9	9	6	8	4	7	6	5	12
7	8	5	4	6	11	7	10	5	8	7	5	9	8	10	2	7	3	8	10
11	8	9	5	7	3	4	9	10	4	7	4	6	9	7	6	12	8	11	9
10	3	6	9	7	5	8	6	2	9	6	4	7	8	10	5	8	7	9	6
5	7	7	6	12	9	10	4	8	6	5	4	7	8	6	7	11	7	8	3

A glance at Table 1.1 suggests that the observed data constitute integers between 2 and 12, but no real patterns are apparent, at least at first sight. To bring out any chance regularity patterns we use a graph as shown in Figure 1.1, **t-plot**: $\{(t, x_t),\ t = 1, 2, \ldots, n\}$.

The first distinction to be drawn is that between chance regularity patterns and deterministic regularities that is easy to detect.

Deterministic regularity. When a t-plot exhibits a clear pattern which would enable one to predict (guess) the value of the next observation *exactly*, the data are said to exhibit *deterministic* regularity. The easiest way to think about deterministic regularity is to visualize

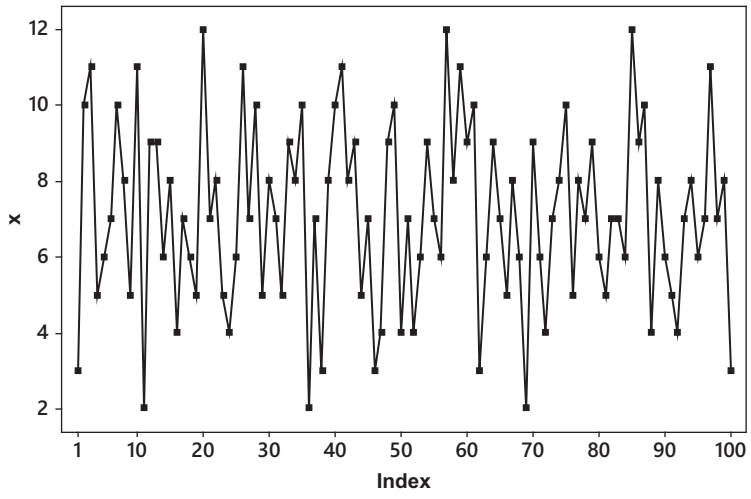

Fig. 1.1 t-Plot of a sequence of 100 observations

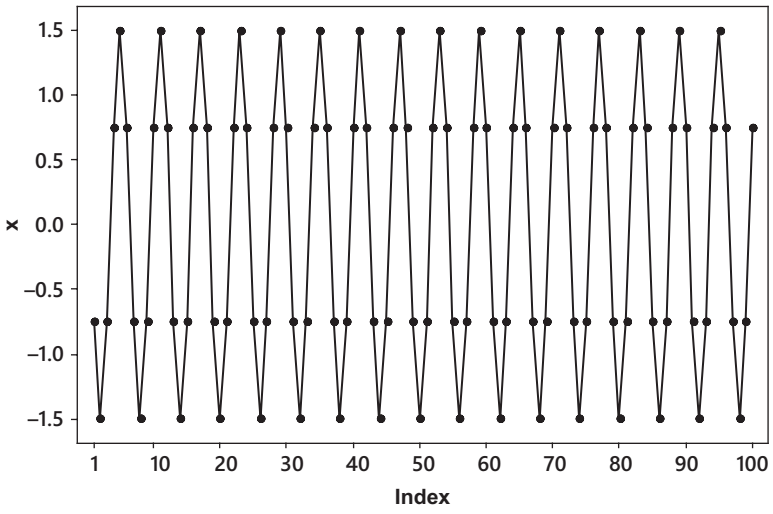

Fig. 1.2 Graph of $x = 1.5\cos((\pi/3)t+(\pi/3))$

the graphs of mathematical functions. If a t-plot of data can be depicted by a mathematical function, the numbers exhibit deterministic regularity; see Figure 1.2.

In contrast to deterministic regularities, to detect chance patterns one needs to perform a number of thought experiments.

Thought experiment 1–Distribution regularity. Associate each observation with identical squares and rotate Figure 1.1 anti-clockwise by 90°, letting the squares fall vertically to form a pile on the x-axis. The pile represents the well-known histogram (see Figure 1.3).

The histogram exhibits a clear triangular shape, reflecting a form of regularity often associated with *stable (unchanging) relative frequencies (RF)* expressed as percentages

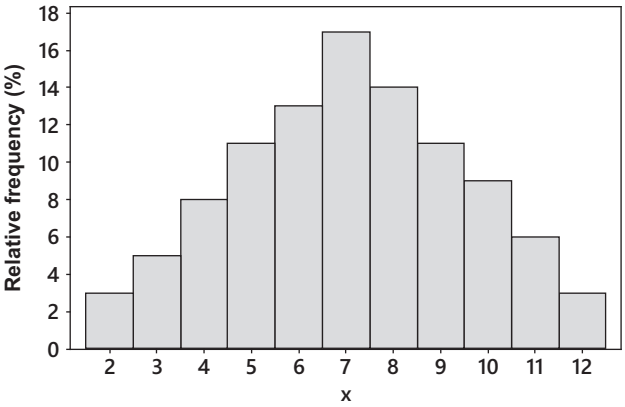

Fig. 1.3 Histogram of the data in Figure 1.1

(%). Each bar of the histogram represents the frequency of each of the integers 2–12. For example, since the value 3 occurs five times in this data set, its relative frequency is RF(3)=5/100 = .05. The relative frequency of the value 7 is RF(7)=17/100 = .17, which is the highest among the values 2–12. For reasons that will become apparent shortly, we name this discernible distribution regularity.

[1] Distribution: After a large enough number of trials, the relative frequency of the outcomes forms a seemingly stable distribution shape.

Thought experiment 2. In Figure 1.1, one would hide the observations beyond a certain value of the index, say $t = 40$, and try to guess the next outcome on the basis of the observations up to $t = 40$. Repeat this along the x-axis for different index values and if it turns out that it is more or less impossible to use the previous observations to narrow down the potential outcomes, conclude that there is *no dependence* pattern that would enable the modeler to guess the next observation (within narrow bounds) with any certainty. In this experiment one needs to exclude the extreme values of 2 and 12, because following these values one is almost certain to get a value greater and smaller, respectively. This type of predictability is related to the *distribution regularity* mentioned above. For reference purposes we name the chance regularity associated with the unpredictability of the next observation given the previous observations.

[2] Independence: In a sequence of trials, the outcome of any one trial does not influence and is not influenced by the outcome of any other.

Thought experiment 3. In Figure 1.1 take a wide enough frame (to cover the spread of the fluctuations) that is also long enough (roughly less than half the length of the horizontal axis) and let it slide from left to right along the horizontal axis, looking at the picture inside the frame as it slides along. In cases where the picture does not change significantly, the data exhibit the chance regularity we call *homogeneity*, otherwise *heterogeneity* is present; see

Chapter 5. Another way to view this pattern is in terms of the arithmetic average and the *variation* around this average of the observations as we move from left to right. It appears as though this *sequential average* and its *variation* are relatively constant around 7. Moreover, the *variation* around this constant average value appears to be within fixed bands. This chance regularity can be intuitively described by the notion of homogeneity.

[3] Homogeneity: The probabilities associated with all possible outcomes remain the same for all trials.

In summary, the data in Figure 1.1 exhibit the following chance regularity patterns:

[1] A triangular distribution; [2] Independence; [3] Homogeneity (ID).

It is important to emphasize that these patterns have been discerned directly from the observed data without the use of any *substantive* subject matter information. Indeed, at this stage it is still unknown what these observations represent or measure, but that does not prevent one from discerning certain chance regularity patterns. The information conveyed by these patterns provides the raw material for constructing statistical models aiming to adequately account for (or model) this (statistical) information. The way this is achieved is to develop probabilistic concepts which aim to formalize these patterns in a mathematical way and provide canonical elements for constructing statistical models.

The formalization begins by representing the data as a set of n ordered numbers denoted generically by $\mathbf{x}_0:=(x_1,x_2,\ldots,x_n)$. These numbers are in turn interpreted as a *typical realization* of a finite initial segment $\mathbf{X}:=(X_1,X_2,\ldots,X_n)$ of a (possibly infinite) sequence of random variables $\{X_t,\ t=1,2,\ldots,n,\ldots\}$ we call a *sample* $\mathbf{X}$; note that the random variables are denoted by capital letters and observations by small letters. The chance regularity patterns exhibited by the data are viewed as reflecting the probabilistic structure of $\{X_t,\ t=1,2,\ldots,n,\ldots\}$. For the data in Figure 1.1, the structure one can realistically ascribe to sample $\mathbf{X}$ is that they are independent and identically distributed (IID) random variables, with a triangular distribution. These probabilistic concepts will be formalized in the next three chapters to construct a statistical model that will take the simple form shown in Table 1.2.

Table 1.2 *Simple statistical model*

[D] Distribution	$X_t \backsim \Delta(\mu,\sigma^2),\ x_t \in \mathbb{N}_X:=(2,\ldots,12)$, discrete triangular
[M] Dependence	$(X_1,X_2,\ldots,X_n)$ are independent (I)
[H] Heterogeneity	$(X_1,X_2,\ldots,X_n)$ are identically distributed (ID)

Note that $\mu=E(X_t)$ and $\sigma^2=E(X_t-\mu)^2$ denote the mean and variance of X_t, respectively; see Chapter 3.

It is worth emphasizing again that the choice of this statistical model, which aims to account for the regularities in Figure 1.1, relied exclusively on the chance regularities, without invoking any substantive subject matter information relating to the actual mechanism that gave rise to the particular data. Indeed, the generating mechanism was deliberately veiled in the discussion so far to make this point.

1.2.2 From Chance Regularities to Probabilities

The question that naturally arises is whether the available substantive information pertaining to the mechanism that gave rise to the data in Figure 1.1 would affect the choice of a statistical model. Common sense suggests that it should, but it is not clear what its role should be. Let us discuss that issue in more detail.

The *actual* data-generating mechanism (DGM). It turns out that the data in Table 1.1 were generated by a sequence of $n = 100$ trials of *casting two dice* and adding the dots of the two sides facing up. This game of chance was very popular in medieval times and a favorite pastime of soldiers waiting for weeks on end outside the walls of European cities they had under siege, looking for the right opportunity to assail them. After thousands of trials these illiterate soldiers learned empirically (folk knowledge) that the number 7 occurs more often than any other number and that 6 occurs less often than 7 but more often than 5; 2 and 12 would occur the least number of times. One can argue that these soldiers had an instinctive understanding of the empirical relative frequencies summarized by the histogram in Figure 1.3.

In this subsection we will attempt to reconstruct how this intuition was developed into something more systematic using mathematization tools that eventually led to probability theory. Historically, the initial step from the observed regularities to their probabilistic formalization was very slow in the making, taking centuries to materialize; see Chapter 2.

The *first* crucial feature of the generating mechanism is its stochastic nature: at each trial (the casting of two dice), the outcome (the sum of the dots of the sides) cannot be predicted with any certainty. The only thing one can say with certainty is that the result of each trial will be one of the numbers $\{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$. It is also known that these numbers do *not* occur equally often in this game of chance.

How does one *explain* the differences in the empirical relative frequency of occurrence for the different numbers as shown in Figure 1.3? The first systematic account of the underlying mathematics behind Figure 1.3 was given by Gerolamo Cardano (1501–1576), who lived in Milan, Italy. He was an Italian polymath, whose wide interests ranged from being a mathematician, physician, biologist, chemist, astrologer/astronomer, to gambler.

The mathematization of chance regularities. Cardano reasoned that since each die has six faces $(1, 2, \ldots, 6)$, if the die is symmetric and homogeneous, the probability of each outcome is equal to $1/6$, i.e.

Number of dots	1	2	3	4	5	6
Probability	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

When casting two dice (D_1, D_2), one has 36 possible outcomes associated with the different pairings of these numbers (i, j), $i, j = 1, 2, \ldots, 6$; see Table 1.3. That is, behind each one of the possible events $\{2, 3, \ldots, 12\}$ there is a combination of elementary outcomes, whose probability of occurrence could be used to explain the differences in their relative frequencies.

The *second* crucial feature of the generating mechanism is that, under *certain conditions,* all elementary outcomes (x, y) are equally likely to occur; each elementary outcome occurs with probability $1/36$. These conditions are of paramount importance in modeling stochastic

Table 1.3 *Elementary outcomes: casting two dice*

$D_1 \backslash D_2$	**1**	**2**	**3**	**4**	**5**	**6**
1	(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
2	(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
3	(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
4	(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
5	(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
6	(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

phenomena, because they constitute the premises of inference. In this case they pertain to the physical *symmetry* of the two dice and the *homogeneity* (sameness) of the replication process. In the actual experiment giving rise to the data in Table 1.1, the dice were cast in the same wooden box to secure a certain form of nearly identical conditions for each trial.

Going from these elementary outcomes to the recorded result $z = x+y$, it becomes clear that certain events are more likely to occur than others, because they occur when different combinations of the elementary outcomes arise (see Table 1.4). For instance, we know that the number 2 can arise as the sum of a single combination of faces: $\{1, 1\}$ – each die comes up 1, hence $\Pr(\{1, 1\}) = 1/36$. The same applies to the number 12: $\Pr(\{6, 6\}) = 1/36$. On the other hand, the number 3 can arise as the sum of two sets of faces: $\{(1, 2), (2, 1)\}$, hence $\Pr(\{(1, 2), (2, 1)\}) = 2/36$. The same applies to the number 11: $\Pr(\{(6, 5), (5, 6)\}) = 2/36$. If you do not find the above derivations straightforward do not feel too bad, because a giant of eighteenth-century mathematics, Gottfried Leibniz (1646–1716), who developed differential and integral calculus independently of Isaac Newton, made an elementary mistake when he argued that $\Pr(z = 11) = \Pr(z = 12) = 1/36$; see Todhunter (1865, p. 48). The reason? Leibniz did not understand clearly the notion of "the set of all possible distinct outcomes" (Table 1.3)!

Continuing this line of thought, one can construct a *probability distribution* that relates each event of interest with a certain probability of occurrence (see Figure 1.4). As we can see, the outcome most likely to occur is the number 7. We associate the relative frequency of occurrence with the underlying probabilities defining a probability distribution over all possible results; see Chapter 3.

Table 1.4 *Probability distribution: sum of two dice*

Outcome	2	3	4	5	6	7	8	9	10	11	12
Probability	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

One can imagine Cardano sitting behind a makeshift table at a corner of Piazza del Duomo in Milan inviting passers-by to make quick money by betting on events like C – the sum of two dice being bigger than 9, and offering odds 3-to-1 against; three ways to lose and one to win. He knew that based on Table 1.3, $\Pr(C) = 6/36$. This meant that he would win most of the time, since the relevant odds to be a fair game should have been 5-to-1. Probabilistic knowledge meant easy money for this avid gambler and he was not ready to share

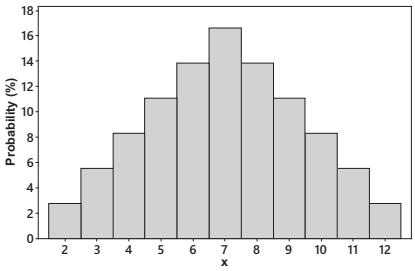

Fig. 1.4 Probability distribution

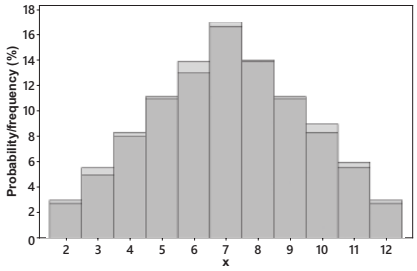

Fig. 1.5 Probability vs. relative frequency

it with the rest of the world. Although he published numerous books and pamphlets during his lifetime, including his autobiography in lurid detail, his book about games of chance, *Liber de Ludo Aleae*, written around 1564, was only published posthumously in 1663; see Schwartz (2006).

The probability distribution in Table 1.4 represents a mathematical concept formulated to model a particular form of chance regularity exhibited by the data in Figure 1.1 and summarized by the histogram in Figure 1.3. A direct comparison between Figures 1.3 and 1.4, by superimposing the latter on the former in Figure 1.5, confirms the soldiers' intuition: the empirical relative frequencies are very close to the theoretical probabilities. Moreover, if we were to repeat the experiment 1000 times, the relative frequencies would have been even closer to the theoretical probabilities; see Chapter 10. In this sense we can think of the histogram in Figure 1.3 as an empirical instantiation of the probability distribution in Figure 1.4.

Let us take the above formalization of the two-dice example one step further.

Example 1.2 When playing the two-dice game, the medieval soldiers used to gamble on whether the outcome would be an odd or an even number (the Greeks introduced these concepts around 300 BC), by betting on odd $A = \{3,5,7,9,11\}$ or even $B = \{2,4,6,8,10,12\}$ numbers. At first sight it looks as though the soldier betting on B would have had a clear advantage since there are more even than odd numbers. The medieval soldiers, however, had folk knowledge that this was a fair bet! We can confirm that $\Pr(A)=\Pr(B)$ using the probabilities in Table 1.4 to derive those in Table 1.5:

$$\Pr(A)=\Pr(3)+\Pr(5)+\Pr(7)+\Pr(9)+\Pr(11)=\tfrac{2}{36}+\tfrac{4}{36}+\tfrac{6}{36}+\tfrac{4}{36}+\tfrac{2}{36}=\tfrac{1}{2};$$
$$\Pr(B)=\Pr(2)+\Pr(4)+\Pr(6)+\Pr(8)+\Pr(10)+\Pr(12)=\tfrac{1}{36}+\tfrac{3}{36}+\tfrac{5}{36}+\tfrac{5}{36}+\tfrac{3}{36}+\tfrac{1}{36}=\tfrac{1}{2}.$$

Table 1.5 *Odd and even sum*

Outcome	A	B
Probability	.5	.5

The historical example credited with being the first successful attempt to go from empirical relative frequencies (real world) to probabilities (mathematical world) is discussed next.

1.2.2.1 Example 1.3: Chevalier de Mere's Paradox*

Historically, the connection between a stable (unchanging) law of relative frequencies can be traced back to the middle of the seventeenth century in an exchange of letters between Pascal and Fermat; see Hacking (2006).

Chevalier de Mere's paradox was raised in a letter from Pascal to Fermat on July 29, 1654 as one of the problems posed to him by de Mere (a French nobleman and a studious gambler). De Mere observed the following empirical regularity:

$$P(\textit{at least one } 6 \text{ in 4 casts of 1 die}) > \tfrac{1}{2} > P(\textit{a double } 6 \text{ in 24 casts with 2 dice})$$

on the basis of numerous repetitions of the game. This, however, seemed to contradict his reasoning by analogy; hence the paradox.

De Mere's false reasoning. He reasoned that the two probabilities should be identical because one 6 in four casts of one die should be the same event as a double 6 in 24 casts of two dice, since 4 is to 6 as 24 is to 36. False! Why?

Multiplication counting principle. Consider the sets $S_1, S_2, \ldots, S_k$ with $n_1, n_2, \ldots, n_k$ elements, respectively. Then there are $n_1 \times n_2 \times \ldots \times n_k$ ways to choose one element from S_1, then one element from S_2, ..., then one element from S_k.

In the case of two dice, the set of all possible outcomes is $6 \times 6 = 6^2 = 36$ (see Table 1.3). To explain the empirical regularity observed by de Mere, one needs to assume equal probability (1/36) for each *pair* of numbers from 1 to 6 in casting two dice, and argue as in Table 1.6. The two probabilities $p = 0.4914039$ and $q = 0.5177469$ confirm that de Mere's empirical frequencies were correct but his reasoning by analogy was erroneous. What rendered the small difference of .026 in the two probabilities of empirical discernability is the very large number of repetitions under more or less identical conditions. The mathematical result underlying such stable long-run frequencies is known as the Law of Large Numbers (Chapter 9).

Table 1.6 *Explaining away de Mere's paradox*

One die ($\mathbb{P}(i) = \frac{1}{6},\ i = 1, 2, \ldots, 6$)	**Two dice** ($\mathbb{P}(i,j) = \frac{1}{36},\ i,j = 1, 2, \ldots, 6$)
$\mathbb{P}(\text{one } 6) = \frac{1}{6}$	$\mathbb{P}(\text{one } (6,6)) = \frac{1}{36}$
$\mathbb{P}(\text{one 6 in } n \text{ casts}) = \left(\frac{1}{6}\right)^n$	$\mathbb{P}(\text{one } (6,6) \text{ in } n \text{ casts}) = \left(\frac{1}{36}\right)^n$
$\mathbb{P}(\text{no 6 in } n \text{ casts}) = \left(\frac{5}{6}\right)^n$	$\mathbb{P}(\text{no } (6,6) \text{ in } n \text{ casts}) = \left(\frac{35}{36}\right)^n$
$\mathbb{P}(\text{at least one 6 in } n \text{ casts}) = 1 - (\frac{5}{6})^n = q$	$\mathbb{P}(\text{at least one } (6,6) \text{ in } n \text{ casts}) = 1 - (\frac{35}{36})^n = p$
For $n = 4,\ q = 1 - \left(\frac{5}{6}\right)^4 = 0.5177469$	For $n = 24,\ p = 1 - \left(\frac{35}{36}\right)^{24} = 0.4914039$