1 An introduction to empirical modeling

1.1 Introduction

In an attempt to give some idea of what empirical modeling is all about, we begin the discussion with an epigrammatic demarcation of its intended scope:

> **Empirical modeling** is concerned with the parsimonious description of observable stochastic phenomena using statistical models.

The above demarcation is hardly illuminating because it involves the unknown terms *stochastic phenomenon* and *statistical model* which will be explained in what follows. At this stage, however, it suffices to note the following distinguishing features of empirical (as opposed to other forms of) modeling:

(a) the *stochastic* nature of the phenomena amenable to such modeling,
(b) the indispensability of the *observed data*, and
(c) the nature of the description in the form of a *statistical model*.

The primary objective of empirical modeling is to provide an *adequate description* of certain types of observable phenomena of interest in the form of stochastic mechanisms we call *statistical models.* A statistical model purports to capture the *statistical systematic information* (see sections 2–3), which is different from the theory information (see section 4). In contrast to a *theory model*, a statistical model is codified exclusively in terms of probabilistic concepts and it is descriptive and anti-realistic in nature (see chapter 10 for further discussion). The *adequacy* of the description is assessed by how well the postulated statistical model accounts for all the statistical systematic information in the data (see section 5). In section 6 we provide a preliminary discussion of certain important dimensions of the constituent element of empirical modeling, the observed data.

Empirical modeling in this book is considered to involve a wide spectrum of interrelated procedures including:

(i) *specification* (the choice of a statistical model),
(ii) *estimation* (estimation of the parameters of the postulated statistical model),

(iii) *misspecification testing* (assessing the validity of the probabilistic assumptions of the postulated statistical model), and
(iv) *respecification* (an alternative choice of a statistical model).

As argued below, these facets of modeling are particularly involved in the case of **observational data**. In the case of **experimental data** the primary focus is on estimation because facets (i) and (iv) constitute the other side of the *design* coin and (iii) plays a subsidiary role.

A quintessential example of empirical modeling using observational data is considered to be *econometrics*. An important thesis adopted in this book is that econometrics differs from mainstream statistics (dominated by the experimental design and the least-squares traditions), not so much because of the economic theory dimension of modeling, but primarily because of the particular modeling issues that arise due to the *observational nature* of the overwhelming majority of economic data. Hence, we interpret the traditional definition of econometrics "the estimation of relationships as suggested by economic theory" (see Harvey (1990), p. 1), as placing the field within the experimental design modeling framework. In a nutshell, the basic argument is that the traditional econometric textbook approach utilizes the experimental design modeling framework for the analysis of non-experimental data (see Spanos (1995b) for further details).

1.1.1 A bird's eye view of the chapter

The rest of this chapter elaborates on the distinguishing features of empirical modeling (a)–(c). In section 2 we discuss the meaning of **stochastic observable phenomena** and why such phenomena are amenable to empirical modeling. In section 3, we discuss the relationship between stochastic phenomena and **statistical models**. This relationship comes in the form of *statistical systematic information* which is nothing more than the formalization of the chance regularity patterns exhibited by the observed data emanating from stochastic phenomena. In section 4 we discuss the important notion of statistical adequacy: whether the postulated statistical model "captures" all the statistical systematic information in the data. In section 5 we contrast the statistical and theory information. In a nutshell, the theoretical model is formulated in terms of the behavior of economic agents and the statistical model is formulated exclusively in terms of probabilistic concepts; a sizeable part of the book is concerned with the question of: What constitutes statistical systematic information? In section 6 we raise three important issues in relation to **observed data**, their different *measurement scales*, their *nature*, and their *accuracy*, as they relate to the statistical methods used for their modeling.

The main message of this chapter is that, in assessing the validity of a theory, the modeler is required to ensure that the observed data constitute an unprejudiced witness whose testimony can be used to assess the validity of the theory in question. A statistical model purports to provide an adequate summarization of the statistical systematic information in the data in the form of a stochastic mechanism that conceivably gave rise to the observed data in question.

1.2 Stochastic phenomena, a preliminary view

As stated above, the intended scope of empirical modeling is demarcated by the stochastic nature of observable phenomena. In this section we explain intuitively the idea of a stochastic phenomenon and relate it to the notion of a statistical model in the next section.

1.2.1 Stochastic phenomena and chance regularity

A **stochastic phenomenon** is one whose observed data exhibit what we call *chance regularity patterns*. These patterns are usually revealed using a variety of graphical techniques.

The essence of *chance regularity,* as suggested by the term itself, comes in the form of two entwined characteristics:

> *chance*: an inherent uncertainty relating to the occurence of particular outcomes,
> *regularity*: an abiding regularity in relation to the occurence of many such outcomes.

TERMINOLOGY: the term chance regularity is introduced in order to avoid possible confusion and befuddlement which might be caused by the adoption of the more commonly used term known as **randomness**; see chapter 10 for further discussion.

At first sight these two attributes might appear to be contradictory in the sense that *chance* refers to 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. Indeed, the essence of chance regularity stems from the fact that the disorder at the individual level creates (somehow) order at the aggregate level. The two attributes should be viewed as inseparable for the notion of chance regularity to make sense. When only one of them is present we cannot talk of chance regularity.

Any attempt to define formally what we mean by the term *chance regularity* at this stage will be rather pointless because one needs several mathematical concepts that will be developed in what follows. Instead, we will attempt to give some intuition behind the notion of chance regularity using a simple example and postpone the formal discussion until chapter 10.

Example
Consider the situation of casting two dice and adding the dots on the sides facing up. The *first* crucial feature of this situation is that at each trial (cast of the two dice) the outcome (the sum of the dots of the sides) cannot be guessed with any certainty. The only thing one can say with certainty is that the outcome will be one of the numbers:

$$\{2,3,4,5,6,7,8,9,10,11,12\},$$

we exclude the case where the dice end up standing on one of the edges! All 36 possible combinations behind the outcomes are shown in table 1.1. The *second* crucial feature of

the situation is that under certain conditions, such as the dice are symmetric, we know that certain outcomes are more likely to occur than others. For instance, we know that the number 2 can arise as the sum of only one set of faces: {1,1} – each die comes up with 1; the same applies to the number 12 with faces: {6,6}. On the other hand, the number 3 can arise as the sum of two sets of faces: {(1,2), (2,1)}; the same applies to the number 11 with faces: {(6,5),(5,6)}. In the next subsection we will see that this line of combinatorial reasoning will give rise to a *probability distribution* as shown in table 1.3.

Table 1.1. *Outcomes in casting two dice*

	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)

At this stage it is interesting to pause and consider the notions of chance regularity as first developed in the context of such games of chance. This is, indeed, the way probabilities made their first appearance. Historically, probabilities were introduced as a way to understand the differences noticed empirically between the likely occurrence of different betting outcomes, as in table 1.1. Thousands of soldiers during the medieval times could attest to the differences in the empirical relative frequencies of occurrence of different events related to the outcomes in table 1.1. While waiting to attack a certain town, the soldiers had thousands of hours with nothing to do and our historical records suggest that they indulged mainly in games of chance like casting dice. After thousands of trials they knew intuitively that the number 7 occurs more often than any other number and that 6 occurs less often than 7 but more often than 5. Let us see how this intuition was developed into something more systematic that eventually led to probability theory.

Table 1.2 reports 100 actual trials of the random experiment of casting two dice and adding the number of dots turning up on the uppermost faces of the dice. A look at the table confirms only that the numbers range from 2 to 12 but no real patterns are apparent, at least at first sight.

Table 1.2. *Observed data on dice casting*

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

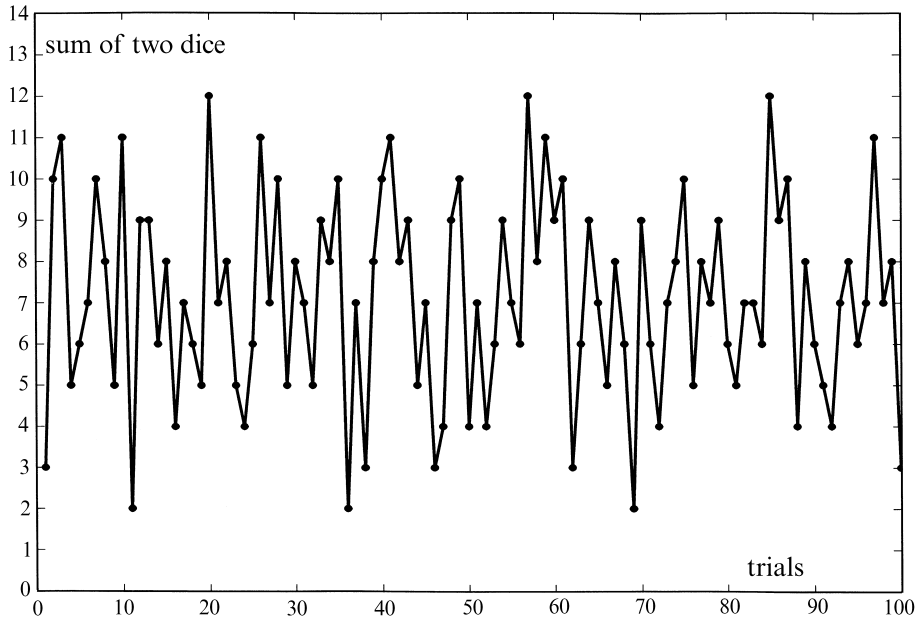

Figure 1.1 A sequence of 100 throws of two dice

In figure 1.1 the data are plotted over the index of the number of the trial. At the first casting of the dice the sum was 3, at the second the sum was 10, at the third the sum of 11 etc. Joining up these outcomes (observations) gives the viewer a better perspective with regard to the sequential nature of the observations. NOTE that the ordering of the observations constitutes an important dimension when discussing the notion of chance regularity.

Historically, the first chance regularity pattern discerned intuitively by the medieval soldiers was that of *a stable law of relative frequencies* as suggested by the histogram in figure 1.2 of the data in table 1.2; without of course the utilization of graphical techniques but after numerous casts of the dice. The question that naturally arises at this stage is:

How is the histogram in figure 1.2 related to the data in figure 1.1?

Today, c*hance regularity* patterns become discernible by performing a number of thought experiments.

Thought experiment 1 Think of the observations as little squares with equal area and rotate the figure 1.1 clockwise by 90° and let the squares representing the observations fall vertically creating a pile on the x-axis. The pile represents the well-known histogram as shown in figure 1.2. This histogram exhibits a clear triangular shape that will be related to a probability distribution derived by using arguments based on combinations and permutations in the next sub-section. For reference purposes we summarize this regularity in the form of the following intuitive notion:

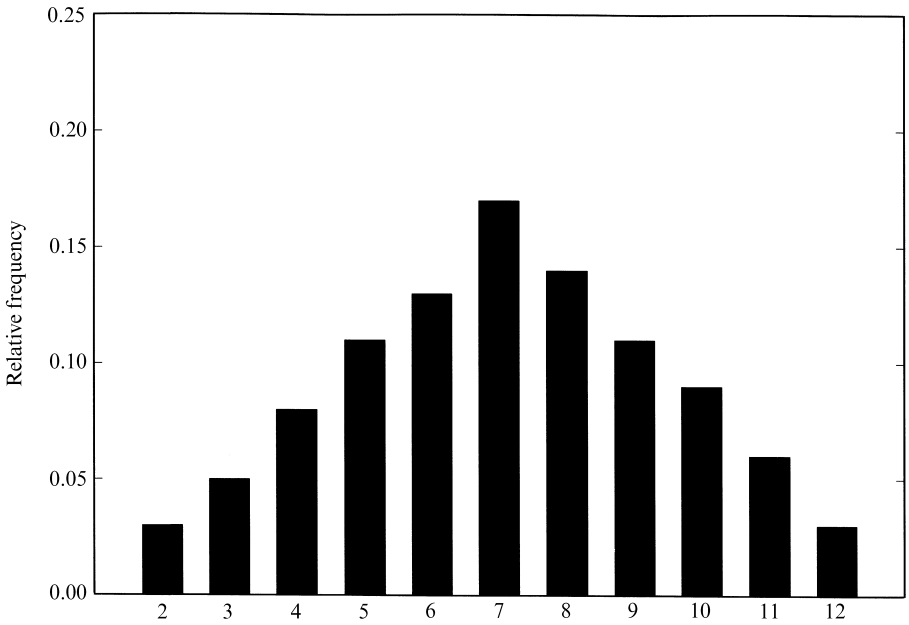

Figure 1.2 Histogram of the sum of two dice data

[1] *Distribution*: after several trials the outcomes form a (seemingly) stable law.

Thought experiment 2 Hide the observations following a certain value of the index, say $t = 40$, and try to guess the next outcome. Repeat this along the observation index axis and if it turns out that it is impossible to use the previous observations to guess the value of the next observation, excluding the extreme cases 2 and 12, then the chance regularity pattern we call *independence* is present. It is important to note that in the case of the extreme outcomes 2 and 12 one is almost sure that after 2 the likelihood of getting a number greater than that is much higher, and after 12 the likelihood of getting a smaller number is close to one. As argued below, this type of predictability is related to the regularity component of chance known as a stable relative frequencies law. Excluding these extreme cases, when looking at the previous observations, one cannot discern a pattern in figure 1.1 which helps narrow down the possible alternative outcomes, enabling the modeler to guess the next observation (within narrow bounds) with any certainty. Intuitively, we can summarize this notion in the form of:

[2] *Independence*: in any sequence of trials the outcome of any one trial does not influence and is not influenced by that of any other.

Thought experiment 3 Take a wide frame (to cover the spread of the fluctuations in a t-plot such as figure 1.1) 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 the case where the picture does not change significantly, the data exhibit *homogeneity*, otherwise *heterogeneity* is present; see chapter 5. Another way to view this pattern is in terms of the average and the *variation* around this average of the numbers as we move from left to right. It appears as though this *sequential average* and its *variation* are relatively constant around 7. The *variation* around this constant average value appears to be within constant bands. This chance regularity can be intuitively summarized by the following notion:

[3] *Homogeneity*: the probabilities associated with the various outcomes remain identical for all trials.

NOTE that in the case where the pattern in a *t*-plot is such so as to enable the modeler to guess the next observation *exactly*, the data do not exhibit any chance pattern, they exhibit what is known as *deterministic* regularity. The easiest way to think about deterministic regularity is to visualize the graphs of mathematical functions from elementary (polynomial, algebraic, transcendental) to more complicated functions such as Bessel functions, differential and integral equations. If we glance at figure 1.1 and try to think of a function that can describe the zig-zag line observed, we will realize that no such mathematical function exists; unless we use a polynomial of order 99 which is the same as listing the actual numbers. The patterns we discern in figure 1.1 are chance regularity patterns.

1.2.2 Chance regularity and probabilistic structure

The step from the observed regularities to their formalization (mathematization) was prompted by the distribution regularity pattern as exemplified in figure 1.2. The formalization itself was initially very slow, taking centuries to materialize, and took the form of simple combinatorial arguments. We can capture the essence of this early formalization if we return to the dice casting example.

Example
In the case of the experiment of casting two dice, we can continue the line of thought that suggested differences in the likelihood of occurrences of the various outcomes in {2,3,4,5,6,7,8,9,10,11,12} as follows. We already know that 3 occurs twice as often as 2 or 11. Using the same common sense logic we can argue that since 4 occurs when any one of {(1,3), (2,2), (3,1)} occurs, its likelihood of occurrence is three times that of 2. Continuing this line of thought and assuming that the 36 combinations can occur with the same probability, we discover a distribution that relates each outcome with a certain likelihood of occurrence shown below in figure 1.3; first derived by Coordano in the 1550s. As we can see, the outcome most likely to occur is the number 7; it is no coincidence that several games of chance played with two dice involve the number 7. We think of the likelihoods of occurrence as *probabilities* and the overall pattern of such probabilities associated with each outcome as a *probability distribution*; see chapter 3.

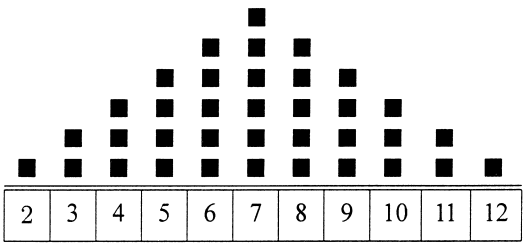

Figure 1.3 Regularity at the aggregate

Table 1.3. *The sum of two dice: a probability distribution*

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

The probability distribution in table 1.3 represents a probabilistic concept formulated by mathematicians in order to capture the chance regularity in figure 1.1. A direct comparison between figures 1.2 and 1.3 confirms the soldiers' intuition. The empirical relative frequencies in figure 1.2 are close to the theoretical probabilities shown in figure 1.3. 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 1.2 as an empirical realization of the probability distribution in figure 1.3 (see chapter 5 for further discussion).

Example

In the case of the experiment of casting two dice, the medieval soldiers used to gamble on whether the outcome is an odd or an even number (the Greeks introduced these concepts at around 300 BC). That is, soldier A would bet on the outcome being $A = \{3,5,7,9,11\}$ and soldier B on being $B = \{2,4,6,8,10,12\}$. At first sight it looks as though soldier B will be a definite winner because there are more even than odd numbers. The medieval soldiers, however, knew by empirical observation that this was not true! Indeed, if we return to table 1.3 and evaluate the probability of event A occurring, we discover that the soldiers were indeed correct: the probability of both events is $\frac{1}{2}$; the probability distribution is given in table 1.4.

Table 1.4. *The sum of two dice: odd and even*

outcomes	$A = \{3,5,7,9,11\}$	$B = \{2,4,6,8,10,12\}$
probabilities	$\frac{1}{2}$	$\frac{1}{2}$

We conclude this subsection by reiterating that the stochastic phenomenon of casting two dice gave rise to the observed data depicted in figure 1.1, which exhibit the three different forms' chance regularity patterns:

[1] Distribution (triangular), [2] Independence, and [3] Homogeneity.

For reference purposes, it is important to note that the above discernible patterns, constitute particular cases of chance regularity patterns related to three different broad categories of probabilistic assumptions we call **Distribution**, **Dependence**, and **Heterogeneity**, respectively; see chapter 5. The concepts underlying these categories of probabilistic assumptions will be defined formally in chapters 3–4.

A digression – Chevalier de Mere's paradox

Historically, the connection between a stable law of relative frequencies and probabilities was forged in the middle of the 17th century in an exchange of letters between Pascal and Fermat. In order to get a taste of this early formulation, let us consider the following historical example.

The 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:

> the probability of getting at least one 6 in 4 casts of a die is greater than $\frac{1}{2}$, but
>
> the probability of getting a double 6 in 24 casts with *two* dice is less than $\frac{1}{2}$.

De Mere established this empirical regularity and had no doubts about its validity because of the enormous number of times he repeated the game. He was so sure of its empirical validity that he went as far as to question the most fundamental part of mathematics, arithmetic itself. Reasoning by analogy, de Mere argued that the two probabilities should be identical because one 6 in 4 casts of one die is the same as a double 6 in 24 casts of two dice since, according to his way of thinking: 4 is to 6 as 24 is to 36.

The statistical distribution in table 1.4 can be used to explain the empirical regularity observed by de Mere. Being a bit more careful than de Mere, one can argue as follows (the manipulations of probabilities are not important at this stage):

> Probability of one double six $= \frac{1}{36}$,
>
> Probability of one double six in n throws $= \left(\frac{1}{36}\right)^n$,
>
> Probability of no double six in n throws $= \left(\frac{35}{36}\right)^n$,
>
> Probability of at least one double six in n throws $= 1-\left(\frac{35}{36}\right)^n = p$,
>
> For $n = 24$, $p = 1-\left(\frac{35}{36}\right)^{24} = 0.4914039$.

It is interesting to note that in the above argument going from the probability of one double six in one trial to that of n trials we use the notion of *independence* to be defined later.

Using a statistical distribution for the case of *one* die, whose probability distribution is given in table 1.5, one can argue analogously as follows:

Table 1.5. *One die probability distribution*

outcomes	1	2	3	4	5	6
probabilities	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Probability of one six $= \left(\frac{1}{6}\right)$,

Probability of one six in n throws $= \left(\frac{1}{6}\right)^n$,

Probability of no six in n throws $= \left(\frac{5}{6}\right)^n$,

Probability of at least one six in n throws $= 1 - \left(\frac{5}{6}\right)^n = q$,

For $n = 4$, $q = 1 - \left(\frac{5}{6}\right)^4 = 0.5177469$.

The two probabilities $p = 0.4914039$ and $q = 0.5177469$ confirm de Mere's empirical regularity and there is no paradox of any kind! This clearly shows that de Mere's empirical frequencies were correct but his reasoning by analogy was faulty.

The chance regularity patterns of *unpredictability*, which we related to the probability concept of [2] *Independence* and that of sameness we related to [3] *homogeneity* using figure 1.1, are implicitly used throughout the exchange between Pascal and Fermat. It is interesting to note that these notions were not formalized explicitly until well into the 20th century. The probabilistic assumptions of Independence and homogeneity (Identical Distribution) underlay most forms of statistical analysis before the 1920s.

At this stage it is important to emphasize that the notion of probability underlying the probability distributions in tables 1.3–1.5, is one of *relative frequency* as used by de Mere to establish his regularity after a huge number of trials. There is nothing controversial about this notion of probability and the use of statistical models to discuss questions relating to games of chance, where the chance mechanism is explicitly an integral part of the phenomenon being modeled. It is not, however, obvious that such a notion of probability can be utilized in modeling other observable phenomena where the chance mechanism is not explicit.

1.2.3 Chance regularity in economic phenomena

In the case of the experiment of casting dice, the chance mechanism is explicit and most people will be willing to accept on faith that if this experiment is actually performed, the chance regularity patterns [1]–[3] noted above, will be present. The question which naturally arises is:

> Is this chance regularity conceivable in stochastic phenomena beyond games of chance?

In the case of stochastic phenomena where the chance mechanism is not explicit, we often:

(a) cannot derive a probability distribution a priori using some physical symmetry argument as in the case of dice or coins, and