Probability Theory and Statistical Inference:
Empirical Modeling with Observational Data

Aris Spanos
Wilson E. Schmidt Professor of Economics,
Virginia Tech

Chapter 1: An Introduction to Empirical Modeling

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 elucidate (understand, predict, explain) 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 form 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 5-6.

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 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 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 unveil-

ing the probabilistic assumptions (implicitly or explicitly) imposed on the data (the statistical model), and (ii) 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.

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 ostentatious display of pedantry. As a preemptive 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)-(d). In section 2 we discuss the meaning of *stochastic observable phenomena* and why such phenomena are amenable to empirical modeling. Section 3 focuses on the relationship between data from stochastic phenomena and *statistical models*. Section 5, discusses several important issues relating to *observed data*, including their different *measurement scales*, their *nature* and *accuracy*. In section 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 6 discusses briefly the connection between a statistical model and the substantive information of interest.

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.

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 confusions with the more commonly used term of '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 to 12, but no real patterns are apparent, at least at first sight. To bring out any chance regularity patterns we use a graph shown in fig. 1.1: **t-plot**: $\{(t, x_t),\ t{=}1, 2, ..., n\}$

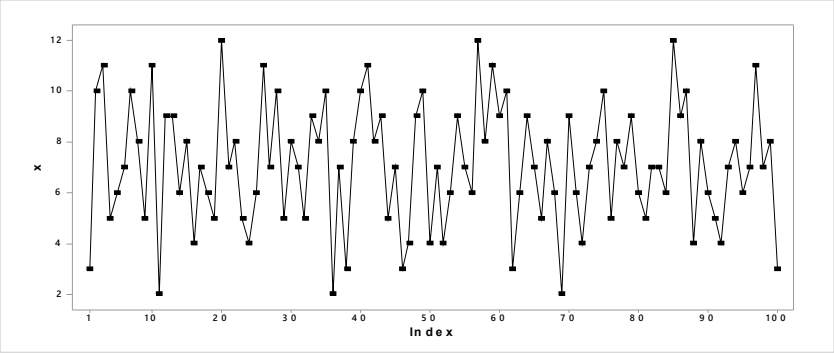

Fig. 1.1: t-plot of a sequence of 100 observations

The first distinction to be drawn is that between chance regularity patterns and deterministic regularities that are 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 the graphs of mathematical functions; see figure 1.2.

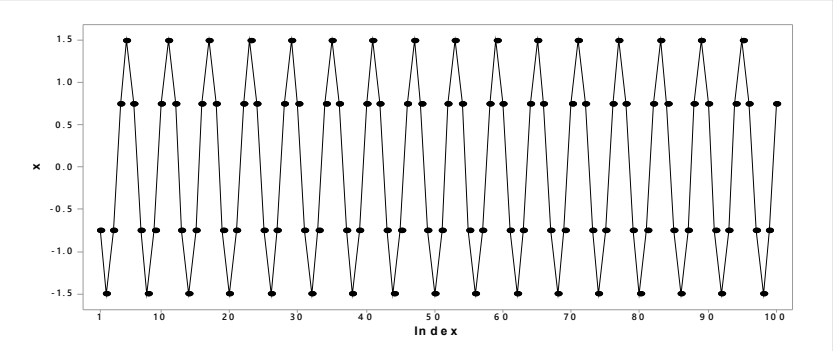

Fig. 1.2: The graph of x=1.5 cos((π/3)t+(π/3))

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

Thought experiment 1. Associated each observation with identical squares and rotate the figure 1.1 anti-clockwise by 90^o letting the squares fall vertically to form a pile on the x-axis. The pile represents the well known histogram (see figure 1.3).

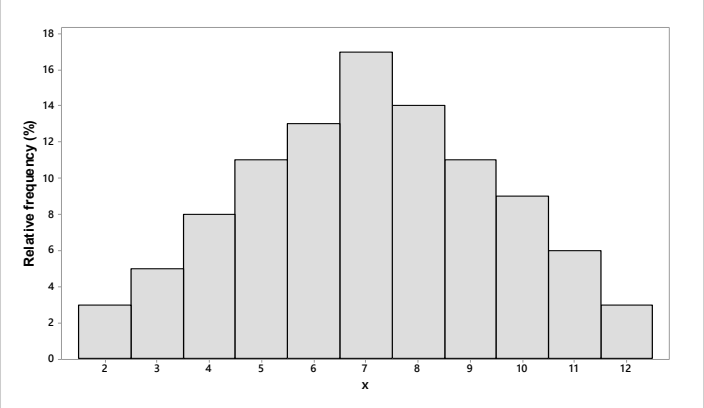

Fig. 1.3: Histogram of the data in fig. 1.1

The histogram exhibits a clear triangular shape reflecting a form of regularity often associated with *stable (unchanging) relative frequencies (RF)* expressed as percentages (%). Each bar of the histogram represents the frequency of each of the integers 2-12. For example, since the value 3 occurs 5 times in this data set its relative frequency is: RF(3)=$\frac{5}{100}$=.05. The relative frequency of the value 7 is: RF(7)=$\frac{17}{100}$ = .17, which is the highest among the values 2-12. For reasons that will become apparent shortly we name this discernible 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, one would 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:

[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 $\mathbf{x}_0{:=}(x_1, x_2, ..., x_n)$. These numbers are in turn interpreted as a *typical realization* of a finite initial segment $\mathbf{X}{:=}(X_1, X_2, ..., X_n)$ of a (possibly infinite) sequence of random variables $\{X_t,\ t{=}1, 2, ..., n, ...\}$, 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 $\{X_t,\ t{=}1,2,...,n,...\}$. 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 following simple form.

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:	$(X_1, X_2, ..., X_n)$ are Independent (I),
[H] Heterogeneity:	$(X_1, X_2, ..., X_n)$ 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.

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 sub-section 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 fig. 1.3? The first systematic account of the underlying mathematics behind figure 1.3 was given by Gerolamo Cardano (1501-1576) whos lived in Milan, Italy. He is an Italian polymath, whose wide interests ranged from being a mathematician, physician, biologist, chemist, astrologer/astronomer, and a gambler.

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

Probability:	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$
Number of dots:	1	2	3	4	5	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=$1, 2, ..., 6$; see table 1.3. That is, behind each one of the possible events $\{2, 3, ..., 12\}$ there is a combination of elementary outcomes, whose probability of occurrence could be use to explain the differences in their relative frequencies.

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)

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 $\frac{1}{36}$. These conditions are of paramount importance in modeling stochastic 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 come up. 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 $\Pr(\{1,1\})=\frac{1}{36}$. the same applies to the number 12: $\Pr(\{6,6\})=\frac{1}{36}$. On the other hand the number 3 can arise as the sum of two sets of faces: $\{(1,2),(2,1)\}$, hence $\Pr(\{(1,2),(2,1)\})=\frac{2}{36}$. The same applies to the number 11: $\Pr(\{(6,5),(5,6)\})=\frac{2}{36}$. If you do not find the above derivations straightforward do not feel too bad because a giant of 18th century mathematics, Gottfried Leibniz (1646-1716), who developed differential and integral calculus independently of Isaac Newton, made an elementary mistake when he argued that $\Pr(z{=}11)=\Pr(z{=}12)=\frac{1}{36}$; 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 to 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; 3 ways to lose and 1 to win. He knew that based on table 1.3: $\Pr(C)=\frac{6}{36}$. This meant that he would win most of 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 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 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.

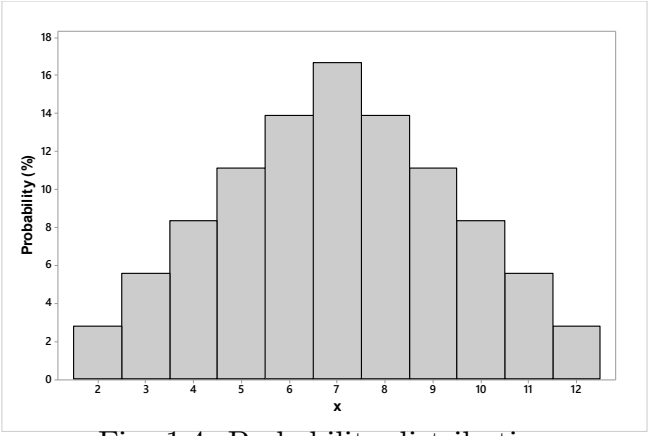

Fig. 1.4: Probability distribution

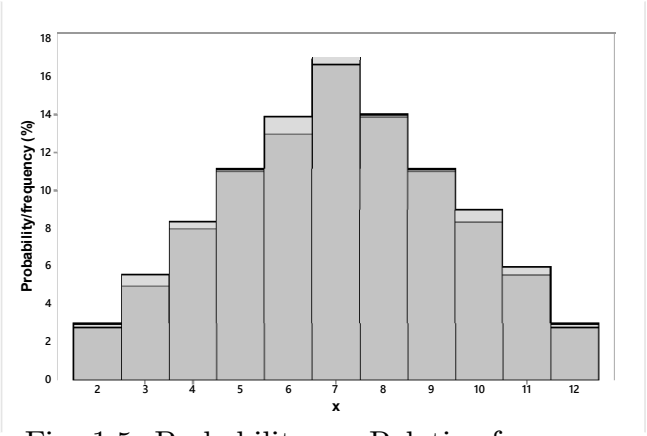

Fig. 1.5: Probability vs. Relative frequency

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 is an odd or an even number (the Greeks introduced these concepts at 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; 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:

$$\Pr(A)=\Pr(3)+\Pr(5)+\Pr(7)+\Pr(9)+\Pr(11)=\tfrac{2}{36}+\tfrac{4}{36}+\tfrac{6}{36}+\tfrac{4}{36}=\tfrac{1}{2}$$

$$\Pr(B)=\Pr(2)+\Pr(4)+\Pr(6)+\Pr(8)+\Pr(10)+\Pr(12)=\tfrac{1}{36}+\tfrac{3}{36}+\tfrac{5}{36}+\tfrac{5}{36}+\tfrac{3}{36}+\tfrac{1}{36}=\tfrac{1}{2}$$

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.

2.2.1 Example 1.3: Chevalier de Mere's paradox*

Historically, the connection between a stable (unchanging) law of relative frequencies is traced back to the middle of the 17th 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:

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

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 4 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, , . . . , and one element from S_k.

In the case of two dice the set of all possible outcomes is $6 \times 6 = 6^2 = 36$ (table 1.3). To explain the empirical regularity observed by de Mere, one needs to assume equal probability $(\frac{1}{36})$ for each *pair* of numbers from 1 to 6 in casting two dice, and argue as in table 1.6 below. 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 empirically discernable 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 ($\mathbb{P}(i)=\frac{1}{6},\ i=1,2,..,6$)	**Two dice** ($\mathbb{P}(i,j)=\frac{1}{36},\ i,j=1,2,..,6$)
$\mathbb{P}$(one 6)= $\frac{1}{6}$,	$\mathbb{P}$(one (6,6))= $\frac{1}{36}$,
$\mathbb{P}$(one 6 in n casts)= $\left(\frac{1}{6}\right)^n$,	$\mathbb{P}$(one (6,6) in n casts)=$\left(\frac{1}{36}\right)^n$,
$\mathbb{P}$(no 6 in n casts)= $\left(\frac{5}{6}\right)^n$,	$\mathbb{P}$(no (6,6) in n casts)=$\left(\frac{35}{36}\right)^n$,
$\mathbb{P}$(at least one 6 in n casts)=1-$(\frac{5}{6})^n$=q,	$\mathbb{P}$(at least one (6,6) in n 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.

2.2.2 Statistical models and substantive information

Having revealed that the data in figure 1.1 have been generated by casting two dice, the question is whether that information will change the statistical model in table 1.2, built exclusively on the statistical information gleaned from chance regularity patterns. In this case the substantive information simply confirms the appropriateness of assuming that the integers between 2-12 constitute all possible values that the generating mechanism can give rise to.

In practice, any substantive subject matter information, say the two dice are perfectly symmetric and homogeneous, should not be imposed on the statistical model at the outset. Instead, one should allow the data to confirm or deny the validity of such information.

2.3 Chance regularity patterns and real-world phenomena

In the case of the experiment of casting two dice the chance mechanism is explicit and most people will be willing to accept on faith that if this experiment is actually performed properly, the chance regularity patterns of IID will be present. The question that naturally arises is whether data generated by real-world stochastic phenomena also exhibit such patterns. It is argued that the overwhelming majority of observable phenomena in many disciplines can be viewed as stochastic, and thus amenable to statistical modeling.

Example 1.4. Consider an example from economics where the t-plot of $X{=}\Delta \ln(ER)$, i.e. log-changes of the Canadian/USA dollar exchange rate (ER), for the period 1973-1991 (weekly observations) is shown in figure 1.6.

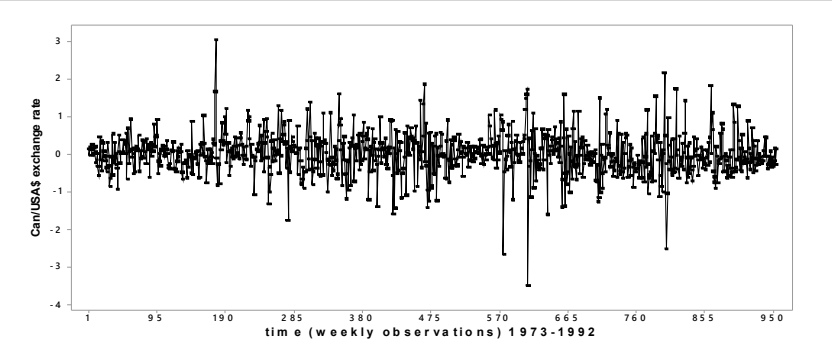

Fig. 1.6: Exchange rate returns

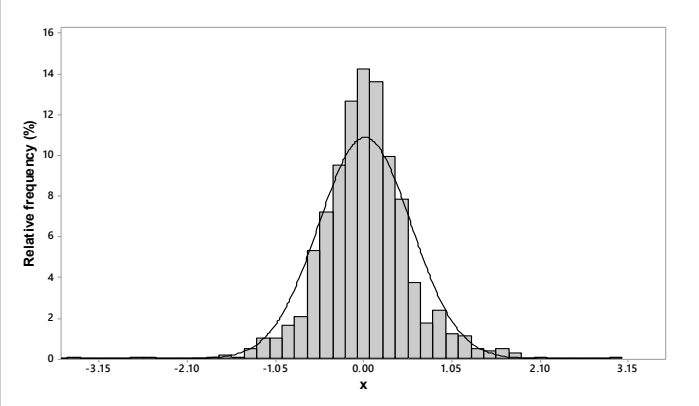

Fig. 1.7: Histogram of exchange rate returns

What is interesting about the data in fig. 1.6 is the fact that they do exhibit a number of *chance regularity* patterns very similar to those exhibited by the dice observations in figure 1.1, but some additional patterns are also discernible. The regularity patterns exhibited by both sets of data are:

(a) The arithmetic average over the ordering (*time*) appears to be constant.

(b) The band of variation around this average appears to be relatively constant.

In contrast to the data in figure 1.2, the distributional pattern exhibited by the data in figure 1.5 is not a triangular. Instead,

(c) the graph of the relative frequencies (histogram) in figure 1.7 exhibits a certain bell-shape symmetry. The Normal density is inserted in order to show that it does not fit well at the tails, in the mid-section and the top, which is much higher than the Normal curve. As argued in chapter 5, the Student's t provides a more appropriate distribution for this data; see figure 3.19. In addition, the data in figure 1.6 exhibit another regularity pattern:

(d) there is a sequence of clusters of small and big changes in succession.

At this stage the reader might not have been convinced that the features noted above are easily discernible from t-plots. An important dimension of modeling in this book is to discuss how to *read* systematic information in data plots, which will begin in chapter 5.

3 Chance regularities and statistical models

Motivated by the desire to account for (model) these chance regularities, we look to probability theory to find ways to formalize in terms of probabilistic concepts. In particular, the stable relative frequencies regularity pattern (tables 1.3-1.5) will be formalized using the concept of a probability distribution (see chapter 5). The unpredictability pattern will be related to the concept of Independence ([2]) and the approximate 'sameness' pattern to the Identical Distribution concept ([3]). To render statistical model specification easier, the probabilistic concepts aiming to 'model' the chance regularities can be viewed as belonging to three broad categories:

(D) Distribution, (M) Dependence, (H) Heterogeneity.

These broad categories can be seen as defining the basic components of a statistical model in the sense that every statistical model is a blend of components from all three categories. The *first* recommendation to keep in mind in empirical modeling:

> 1. A statistical model is simply a set of (internally) consistent probabilistic assumptions from the three broad categories, (D),(M) and (H), defining a stochastic generating mechanism that could have given rise to the data.

The statistical model is chosen to represent a description of a chance mechanism that accounts for the systematic information (the chance regularities) in the data. The distinguishing feature of a statistical model is that it specifies a situation, a mechanism or a process in terms of a certain *probabilistic structure*. The main objective of chapters 2-8 is to introduce numerous probabilistic concepts and ideas that render the choice of an appropriate statistical model an educated guess and not a hit-or-miss decision.

The examples of casting dice, discussed above, are important not because of their intrinsic interest but because they represent examples of a simple stochastic phenomenon we refer to as a *Random Experiment* which will be used in chapters 2-4 to motivate the basic structure of simple statistical model. For the exchange rate data in figure 1.4 we will need to extend the scope of such models to account for dependence

and heterogeneity; this is the subject matter of chapters 6-8. Hence, the appropriate choice of a statistical model depends on:

(a) detecting the chance regularity patterns as exhibited by the observed data,

(b) accounting (modeling) for these patterns by selecting the appropriate probabilistic assumptions.

The first requires developing the skill to detect such patterns using a variety of graphical techniques. Hence, the *second* recommendation in empirical modeling is:

2. Graphical techniques constitute an indispensable tool in empirical modeling!

The interplay between chance regularities and probabilistic concepts using a variety of graphical displays is discussed in chapter 5.

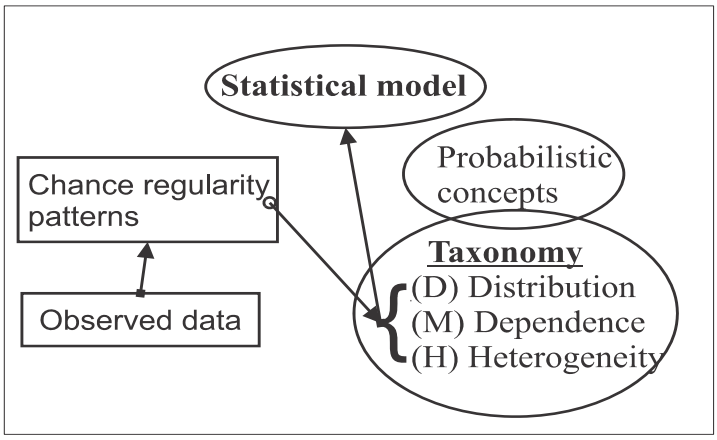

Fig. 1.8: Chance regularity patterns, probabilistic assumptions and a statistical model

Accounting for the statistical systematic information in the data presupposes a mathematical framework rich enough to model the detected chance regularity patterns. Figure 1.8 brings out the interplay between observable chance regularity patterns and formal probabilistic concepts used to construct statistical models.

The variety and intended scope of statistical models are constrained only by the scope of probability theory (as a modeling framework) and the training and the imagination of the modeler. Empirical modeling begins with choosing adequate statistical models with a view to account for the systematic statistical information in the data. The primary objective of modeling, however, is to learn from data by posing substantive questions of interest in the context of the selected statistical model. The *third* recommendation in empirical modeling is:

3. Statistical model specification is guided primarily by the probabilistic structure of the observed data, with a view to pose substantive questions of interest in its context.

Some of the issues addressed in the next few chapters are:

(i) How should one construe a statistical model?

(ii) Why is statistical information coded in probabilistic terms?

(iii) What information does one utilize when choosing a statistical model?

(iv) What is the relationship between the statistical model and the data?

(v) How does one detect the statistical systematic information in data

4 Observed data and empirical modeling

In this section we will attempt a preliminary discussion of a crucial constituent element of empirical modeling, the observed data. Certain aspects of the observed data play an important role in the choice of statistical models.

4.1 Experimental vs. observational data

In most sciences, such as physics, chemistry, geology and biology, the observed data are often generated by the modelers themselves in well-designed experiments. In econometrics the modeler is often faced with *observational* as opposed to *experimental* data. This has two important implications for empirical modeling. First, the modeler needs to develop better skills in validating the model assumptions because random (IID) sample realizations are rare with observational data. Second, the separation of the data collector and the data analyst requires the modeler to examine thoroughly the nature and structure of the data in question.

In economics, along with the constant accumulation of observational data collection grew the demand to analyze these data series with a view to a better understanding of economic phenomena such as inflation, unemployment, exchange rate fluctuations and the business cycle, as well improving our ability to forecast economic activity. A first step toward attaining these objectives is to study the available data by being able to answer questions such as:

(i) How were the data collected and compiled?
(ii) What is the subject of measurement and what do the numbers measure?
(iii) What are the measurement units and scale?
(iv) What is the measurement period?
(v) What is the link between the data and any corresponding theoretical concepts?

A *fourth* recommendation to keep in mind in empirical modeling is:

> 4. One needs to get to know all the important dimensions (i)-(v) of the particular data before any statistical modeling and inference is carried out.

4.2 Observed data and the nature of a statistical model

A data set comprising n observations will be denoted by: $\mathbf{x}_0{:=}(x_1, x_2, ..., x_n,)$.

REMARK: it is crucial to emphasize the value of mathematical symbolism when one is discussing Probability theory. The clarity and concision this symbolism introduces to the discussion is indispensable.

It is common to classify economic data according to the observation units into:

(i) **Cross-section:** $\{x_k,\ k{=}1, 2, ..., n\}$, k denotes individuals (firms, states, etc.),

(ii) **Time series:** $\{x_t,\ t{=}1, 2, ..., T\}$, t denotes time (weeks, months, years, etc.).

for example, observed data on consumption might refer to consumption of different households at the same point in time or aggregate consumption (consumers' expenditure) over time. The first will constitute cross-section, the second, time series data. By combining these two, e.g. observing the consumption of the same households over

time, we can define a third category:

(iii) **Panel (longitudinal):** $\{x_{\mathbf{k}},\ \mathbf{k}{:=}(k,t)\,,\ k{=}1,2,...,n,\ t{=}1,2,...,T\}$; where k and t denote the index for individuals and time, respectively.

NOTE that in this category the index $\mathbf{k}$ is two-dimensional but $x_{\mathbf{k}}$ is one-dimensional.

At first sight the two primary categories do not seem to differ substantively because the index sets appear identical; the index sets are subsets of the set of natural numbers. A moment's reflection, however, reveals that there is more to an index set than meets the eye. In the case where the index set $\mathbb{N}{:=}\{1,2,...,n\}$ refers to particular households, the index might stand for the names of the households, say:

$$\{\text{Jones, Brown, Smith, Johnson,...}\}. \tag{1}$$

For time series the index $\mathbb{T}{:=}\{1,2,...,T,...\}$ might refer to particular dates, say:

$$\{1972, 1973, ..., 2017\}. \tag{2}$$

Comparing the two index sets we note immediately that they have very different mathematical structures. The most apparent difference is that the set (1) does not have a natural ordering, whether we put Brown before Smith is immaterial, but in the case of the index set (2) the ordering is a crucial property of the set.

In the above example the two index sets appear identical but they turn out to be very different. This difference renders the two data sets qualitatively dissimilar to the extent that the statistical analysis of one set of data will be distinctively different from that of the other. The reason for this will become apparent in later chapters. At this stage it is sufficient to note that a number of concepts such as *dependence* and *heterogeneity* are inextricably bound up with the ordering of the index set.

The mathematical structure of the index set is not the only criterion for classifying dissimilar data sets. The mathematical structure of the range of values of observations themselves constitutes another even more important criterion. For example the "number of children" in different households can take values: $\{0,1,2,...,100\}$; 100 is an assumed upper bound. The set of values of the variable consumption would be $\mathbb{R}_+{=}(0,\infty)$. The variable *religion* (Christian, Muslim, Buddhist, Other) cannot be treated in the same way because there is no natural way to measure religion. Even if we agree on a measurement scale for religion, say $\{1,2,3,4\}$, the ordering is irrelevant and the difference between these numbers is meaningless.

The above discussion raised important issues in relation to the measurement of observed data. The first is whether the numerical values can be thought of as being values from a certain interval on the real line, say $[0,1]$ or they represent a set of discrete values, say $\{0,1,2,3,4,5,6,7,8,9\}$. The second is whether these values have a natural ordering or not.

Collecting these comments together we can see that the taxonomy which classifies the data into cross-section and time series is inadequate because there are several additional classifications which are ignored. These classifications are important from the modeling viewpoint because they make a difference in so far as the applicable

statistical techniques are concerned. In its abstract formulation a generic data set is designated by $\{x_k,\ k{\in}\mathbb{N},\ x_k{\in}\mathbb{R}_X\}$, where $\mathbb{N}$ denotes the index set and $\mathbb{R}_X$ the range of values of x; NOTE that both sets $\mathbb{N}$ and $\mathbb{R}_X$ are subsets of the real line, denoted by $\mathbb{R}{:=}(-\infty,\infty)$. Depending on the mathematical structure of these two sets different classifications arise. Indeed, the mathematical structure of the sets $\mathbb{N}$ and $\mathbb{R}_X$ plays a very important role in the choice of the statistical model (see sections 3-5). $\mathbb{R}_X$ can be a **discrete** (countable) **subset** of $\mathbb{R}$, such as $\mathbb{R}_X{=}\{0,1,2,...\}$, or a **continuous** (uncountable) **subset** of $\mathbb{R}$, such as $\mathbb{R}_X{=}[0,\infty)$. The same *discrete-continuous* classification can also be applied to the index set $\mathbb{N}$ leading to a four way classification of variables and the corresponding data. As shown in chapters 3-4, the nature of both sets $\mathbb{N}$ (the index set) and $\mathbb{R}_X$ (the range of values of the data), play an important role in selecting the statistical model.

4.3 Measurement scales and data

A very important dimension of any observed data is the **measurement scale** of the individual data series. The measurement scales are traditionally classified into four broad categories (table 1.7) together with the mathematical operations that are meaningful (legitimate) for different scales.

Ratio scale. Variables in this category enjoy the richest mathematical structure in their range of values, where for any two values along the scale, say x_1 and x_2, all the mathematical operations (i)-(iv) are meaningful. Length, weight, consumption, investment and Gross Domestic Product (GDP) belong to this category.

Table 1.7: Scales and mathematical operations

Scale	(i) $\left(\frac{x_1}{x_2}\right)$	(ii) $(x_2{-}x_1)$	(iii) $x_2 \gtrless x_1$	(iv) $x_2{\neq}x_1$	**Transformation**
Ratio:	✓	✓	✓	✓	scalar multiplication
Interval:	×	✓	✓	✓	linear function
Ordinal:	×	×	✓	✓	increasing monotonic
Nominal:	×	×	×	✓	one-to-one replacement

Interval scale. For a variable measured on an *interval* scale the operations (ii)-(iv) are meaningful but (i) is not. The index set (2) (calendar time) is measured on the interval scale because the difference (1970-1965) is a meaningful magnitude but the ratio ($\frac{1965}{1970}$) is not. Additional examples of variables of interval scale are: temperature (Celsius, Fahrenheit), systolic blood pressure.

Ordinal scale. For a variable measured on an *ordinal* scale the operations (iii)-(iv) are meaningful but (i)-(ii) are not, e.g. grading (excellent, very good, good, failed), income class (upper, middle, lower). For such variables the ordering exists but the distance between categories is not meaningfully quantifiable.

Nominal scale. For a variable measured on an *nominal* scale the operation (iv) is meaningful but (i)-(iii) are not. Such a variable denotes categories which do not have

a natural ordering, e.g. marital status (married, unmarried, divorced, separated), gender (male, female, other), employment status (employed, unemployed, other).

It is important to note that statistical concepts and methods do not apply to all variables irrespective of scale of measurement (see chapter 6).

TERMINOLOGY. In the statistical literature there is some confusion between the measurement scales and three different categorizations of variables:

discrete/continuous, qualitative/quantitative, categorical/non-categorical.

Discrete variables can be measured on all four scales and continuous variables can sometimes be grouped into a small number of categories. Categorical variables are only variables that can be measured on either the ordinal or the nominal scales, but the qualitative variables category is less clearly defined in several statistics books.

Measurement scales and the index set. The examples of measurement scales used in the above discussion referred exclusively to the set $\mathbb{R}_X$: the range of values of a variable X. However, the discussion is also relevant for the index set $\mathbb{N}$. In the case of the variable names of households (1) is measured on a nominal scale. On the other hand in the case of GDP the index set (2) is measured on the interval scale (time). This is because time does not have a natural origin (zero) and in statistical analysis the index set (2) is often replaced by a set of the form $\mathbb{T}:=\{1,2,...,T,...\}$. We note that the time series/cross-section distinction is often based on the measurement scale of the index set. The index set of time series is of interval scale, but that of cross-section can vary from nominal scale (gender), to ratio scale (age).

In view of the fact that in addition to the discrete/continuous dichotomy we have four different measurement scales for the range of values of the observed variable itself ($\mathbb{R}_X$) and another four for the index set $\mathbb{N}:=\{1,2,...,n,...\}$, a wide variety of data types can be defined. Our concern is with the kind of statistical methods that can be meaningfully applied to the particular data in light of their nature and features.

4.4 Measurement scale and statistical analysis

The measurement scales are of interest in statistical modeling because data measured on different scales need different statistical treatment. To give an idea of what that involves, consider data $\mathbf{x}_0:=(x_1,x_2,...,x_n)$ on religious affiliation under the categories:

$$\text{Christian (1)}, \quad \text{Jewish (2)}, \quad \text{Muslim (3)}, \quad \text{Other (4)},$$

and decide to attach to these four groups the numbers 1-4. How can one provide a set of *summary statistics* for such data in the context of **descriptive statistics**? The set of data for such a variable will look like: $(1,4,3,1,1,2,2,2,1,2,3,3,1,1,1)$
It is clear that for such data the notion of the arithmetic mean:

$$\overline{x}=\tfrac{1}{n}\sum_{k=1}^{n}x_k=(\tfrac{1}{15})(1+4+3+1+1+2+2+2+1+2+3+3+1+1+1)=1.867 \qquad (3)$$

makes no substantive or statistical sense because the numbers we attached to these groups could easily have been: 10, 20, 30, 40. How can one provide a measure of location for such data? A more appropriate descriptive measure is that of the **mode**:

the value in the data that has highest relative frequency. In this case, the mode is x_m=1, since this value occurs 7 out of 15 data values; see table 1.8.

Table 1.8: Scales and location measures

	nominal	**ordinal**	**interval**	**ratio**
mean	×	×	✓	✓
median	×	✓	✓	✓
mode	✓	✓	✓	✓

Consider data on an ordinal variable that measures a teacher's performance:

Excellent (1), Good (2), Average (3), Poor (4), Very poor (5)

The data for a particular teacher will look like: $(1, 5, 3, 1, 1, 2, 4, 2, 1, 2, 2, 5, 3, 1, 4)$ What measures of location are statistically meaningful for this data in the context of descriptive statistics? The histogram and the mode are clearly meaningful, but so is the **median**: the middle value when the data are arranged in an ascending or descending order of magnitude. For the above data: $(1, 1, 1, 1, 1, 2, , 2, \boxed{2}, 2, 3, 3, 4, 4, 5, 5)$ Again, the arithmetic average (mean) in (3) because of the arbitrariness of the values we chose; we could equally use the values: 5,7,11,17,19.

For nominal and ordinal data a number of measures of variation like:

$$\text{variance: } s_x^2=\tfrac{1}{n}\textstyle\sum_{k=1}^{n}(x_k-\overline{x})^2, \quad \text{or} \quad \text{standard deviation: } s_x,$$

are also statistically questionable because of the arbitrariness of the values given to the underlying variables. The same is true for the notion of a covariance between two nominal/ordinal variables. For instance, if one suspects that the teacher's performance is related to their academic rank (Y):

Assistant (1), Associate (2), Full professor (3),

one could collect such data and evaluate the *covariance* between performance and rank:

$$c_{xy}=\tfrac{1}{n}\textstyle\sum_{k=1}^{n}(x_k-\overline{x})(y_k-\overline{y}).$$

This statistic, however, will also be statistically spurious, and so will be the *correlation coefficient*:

$$r_{xy}=\frac{\sum_{k=1}^{n}(x_k-\overline{x})(y_k-\overline{y})}{\sqrt{\sum_{k=1}^{n}(x_k-\overline{x})^2\sum_{k=1}^{n}(y_k-\overline{y})^2}}=\frac{c_{xy}}{s_x\cdot s_y}.$$

In practice, researchers often abuse such data indirectly when used in the context of regression analysis. Estimating the regression line:

$$y_k=\beta_0+\beta_1 x_k+u_k,\ k=1,2,...,n,$$

using the least-squares method, gives rise to the estimated coefficients:

$$\widehat{\beta}_0=\overline{y}-\widehat{\beta}_1\overline{x},\quad \widehat{\beta}_1=\frac{\sum_{k=1}^{n}(x_k-\overline{x})(y_k-\overline{y})}{\sum_{k=1}^{n}(x_k-\overline{x})^2},$$

which involve the means of X and Y, the variance of X and their covariance.

The only general rule for the methods of analysis of different measurement-scale variables one can state at this stage is that a method appropriate for a certain measurement scale in the hierarchy is also appropriate for the scales above but not below. There are several books which discuss the methods of analysis of the so-called *categorical data*: data measured on the *nominal* or *ordinal* scale; see Bishop, Fienberg and Holland (1975), Agresti (2013) inter alia.

A cursory look at the applied econometrics literature reveals that variables from very different measurement scales are involved in the same regression equation (see chapter 7), rendering some of these results problematic. Hence:

5. The scale of measurement of different data series should be taken into account for statistical analysis purposes to avoid meaningless statistics and inference results.

4.5 Cross-section vs. time-series, is that the question?

In concluding this chapter it is important to return to the traditional cross-section/ time series taxonomy to warn the reader against highly misleading claims. The conventional wisdom in econometrics is that *dependence or/and heterogeneity are irrelevant for cross-section data* because we know how to select 'random samples' from populations of individual units such as people, households, firms, cities, states, countries, etc.; see Wooldridge (2013) inter alia.

It turns out that this distinction stems from insufficient appreciation of the notion of 'random sampling'. When defining a random sample as a set of random variables $X_1, X_2, ..., X_n$ which are Independent and Identically Distributed (IID), the *ordering* of $X_1, X_2, ..., X_n$, based on the index k=1, 2, ..., n, provides the key to this definition; see chapter 6. The IID is defined relative to this ordering; without the ordering this definition as well as the broader notions of dependence and heterogeneity make little sense. How does this render the distinction between statistical models for cross-section and time-series data superfluous?

In time series there is a generic ordering (time) that suggests itself when talking about dependence and heterogeneity. The fact that in cross-section data there is no *one* generic ordering that suggests itself does not mean that the ordering of such samples is irrelevant. The opposite is true. Because of the diversity of the individual units, in cross-section data there is often more than one ordering of interest. For instance, in a medical study the gender or the age of individuals might be orderings of interest. In a sample of cities, geographical position and population size might be such orderings of interest. For each of these different orderings one can define dependence and heterogeneity in a statistically meaningful way.

Despite claims to the contrary, the notions dependence and heterogeneity are equally applicable to modeling cross-section or time series data. The only differences arise in the measurement scale of the relevant ordering(s). For time series data, the time ordering is measured on an interval scale, and thus it makes sense to talk about serial correlation (a particular form of temporal dependence) and trending mean and variance (particular forms of heterogeneity). In the case of a sample of individuals

used in a medical study, the gender ordering is measured on a nominal scale and thus it makes sense to talk about heterogeneity (shift) in the mean or the variance of male vs. female units; see data plots in chapter 5. In the case of a cross-section of cities or states, geographical position might be a relevant ordering in which case one can talk about spatial heterogeneity or/and dependence.

A data set can always be represented in the form $\mathbf{x}_0{:=}(x_1, x_2, ..., x_n)$ and viewed as a finite realization of the sample $\mathbf{X}{:=}(X_1, X_2, ..., X_n)$ of a stochastic processes $\{X_k,\ k{\in}\mathbb{N},\ x_k{\in}\mathbb{R}_X\}$ (chapter 8), where $\mathbb{N}$ denotes the index set and $\mathbb{R}_X$ the range of values of x, irrespective of whether the data constitute a cross-section or a time-series; their only differences might lie in the mathematical structure of $\mathbb{N}$ and $\mathbb{R}_X$. Statistical modeling and inference begins with viewing data $\mathbf{x}_0$ as a finite realization of an underlying stochastic process $\{X_k,\ k{\in}\mathbb{N},\ x_k{\in}\mathbb{R}_X\}$, and the statistical model constitutes a particular parameterization of this process. A closer look at the formal notion of a random sample (IID) reveals that it presupposes a built-in ordering. Once the ordering is made explicit, both notions of dependence and heterogeneity become as relevant in cross-section as they are for time series data. If anything, cross-section data are often much richer in terms of ordering structures, which is potentially more fruitful in learning from data. Moreover, the ordering of a sample renders the underlying probabilistic assumptions, such as IID, potentially testable in practice. The claim that we know how to select a random sample from a population, and thus we can take the IID assumptions as valid at face value is misguided.

Example 1.5: Sleep aid Ambien. A real-life example of this form of misspecification is the case of the sleep aid Ambien (zolpidem) that was FDA approved in 1992. After a decade on the market and more than 40 million prescriptions, it was discovered (retrospectively) that women are more susceptible to the risk of 'next day impairment' because they metabolize zolpidem more slowly than men. This discovery was the result of thousands of women experiencing sleep-driving and getting involved in numerous accidents in early morning driving. The potential problem was initially raised by Cubała et al. (2008), who recounted the probing of potential third factors such as age, ethnicity, and prenatal exposure to drugs, but questioned why gender was ignored. After a more careful re-evaluation of the original pre-approval trials data and a some additional post-approval trials, the U.S. Food and Drug Administration (FDA) issued a Safety Communication [1-10-2013] recommending lowering the dose of Ambien for women; 3.5 milligram for men and 1.75 milligram for women.

Table 1.9: Test scores - alphabetical order

98	43	77	51	93	85	76	56	59	62	67	79	66	98	57	80	73	68
71	74	83	75	70	76	56	84	80	53	70	67	100	78	65	77	88	81
66	72	65	58	45	63	57	87	51	40	70	56	75	92	73	59	81	85
62	93	84	68	76	62	65	84	59	60	76	81	69	95	66	87		

Example 1.6. Consider the data given in the table 1.9 that refer to the test

scores (y-axis) in a multiple choice exam on Principles of Economics, reported in alphabetical order using the students' surnames (x-axis).

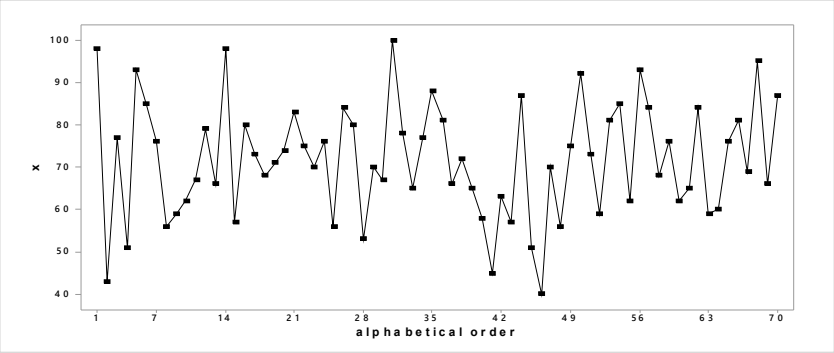

Fig. 1.9: Exam scores data in alphabetical order

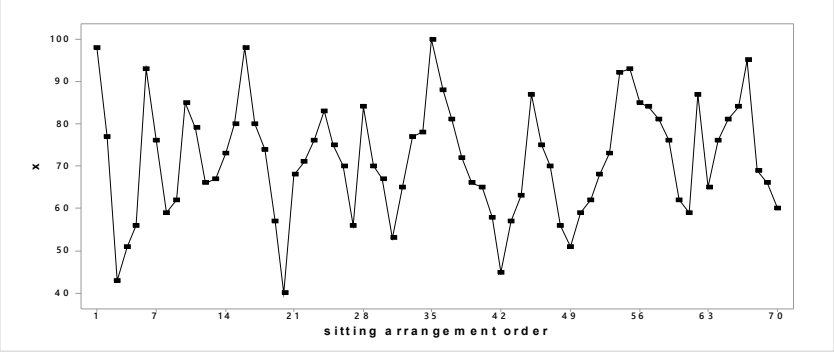

Fig. 1.10: Exam scores data in sitting order

The data in the t-plot (fig. 1.9) appears to exhibit the independence and homogeneity seen fig. 1.1. On the other hand, ordering the observations according to the *sitting arrangement* during the exam, as shown in figure 1.10, seems to exhibit very different chance regularity patters than figure 1.9. The ups and downs of the latter graph are a bit more orderly than those of figure 1.9. In particular, fig. 1.10 exhibits some sort of varying cyclical behavior that renders predicting the next observation easier. As explained in chapter 5, this pattern of irregular cycles reveals that the data exhibit some form of positive *dependence* related to the sitting arrangement. In plain English this means that there was cheating taking place during the examination by glancing at the answers of one's neighbors!

The main lesson from examples 1.5-1.6 is that ordering one's data is a must because it enables the modeler to test dependence and heterogeneity with respect to each ordering of interest. Hence:

> 6. Statistical models for cross section data do admit dependence and heterogeneity assumptions that need to be tested by selecting natural orderings (often more than one) for the particular data.

Statistical models should take into consideration a variety of different dimensions and features of the data.

4.6 Limitations of economic data

In relation with the limitations of economic data we will consider two important issues: (i) what they actually measure and (ii) how accurately. Morgenstern (1963) disputed the accuracy of published economic data and questioned the appropriateness of such data for inferences purposes. In cases where the accuracy and quality of the data raise problems, the modeler should keep in mind that no statistical procedure can extract information from observed data when it is not there in the first place:

7. "Garbage in garbage out" (GIGO): no statistical procedure or substantive information can salvage bad quality data that do not contain the information sought.

The accuracy of economic data has improved substantially since the 1960s and in developed countries data collected by governments and international institutions are sufficiently accurate. The need for different statistical techniques and procedures arises partly because of 'what is being measured' by the available data vs. what information is being sought. The primary limitation of the available economic data arises from the fact that there is a sizeable gap between what the theoretical variables denote and what the available data measure. Economic theory, via the *ceteris paribus* clauses, assumes a *nearly-isolated* system driven by the plans and intentions of optimizing agents, but the observed data are the result of an on-going multidimensional process with numerous influencing factors beyond the control of particular agents.

In what follows we assume that the modeler has checked the observed data thoroughly and deemed them accurate enough to be considered reliable for posing substantive questions of interest. This includes due consideration of the sample size n being large enough for the testing procedures to have adequate capacity to detect any discrepancies of interest; see chapter 13. Hence, a crucial recommendation in empirical modeling is:

8. Familiarize oneself thoroughly with the nature and the accuracy of the data to ensure that they *do* contain the information sought.

This will inform the modeler about what questions can and cannot be posed to a particular data set.

5 Statistical adequacy

The crucial message from the discussion in the previous sections is that probability theory provides the mathematical foundations and the over-arching framework for modeling observable stochastic phenomena of interest. The modus operandi of empirical modeling is the concept of a statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$, that mediates between the data $\mathbf{x}_0$ and the real-world phenomenon of interest at two different levels [A] and [B] (figure 1.11).

[A] **From a phenomenon of interest to a statistically adequate model**. The statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ is chosen so that the observed data $\mathbf{x}_0$ constitute a truly typical realization of the stochastic process $\{X_t,\ t\in\mathbb{N}\}$ underlying $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$. Validating the model assumptions requires trenchant *M-S testing*. The validity of these

assumptions secures the soundness of the inductive premises of inference ($\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$) and renders inference reliable in *learning from data* $\mathbf{x}_0$ about phenomena of interest. The notion of statistical adequacy is particularly crucial for empirical modeling because it can provide the basis for establishing *stylized facts* stemming from the data which theory needs to account for.

[B] **From the inference results to the substantive questions of interest**. This nexus raises issues like statistical vs. substantive significance and how one assesses substantive information. As argued in chapter 13, most of these issues can be addressed using the post-data severity evaluation of the accept/reject rules of testing by establishing the discrepancy from the null warranted by data $\mathbf{x}_0$ and test T_α.

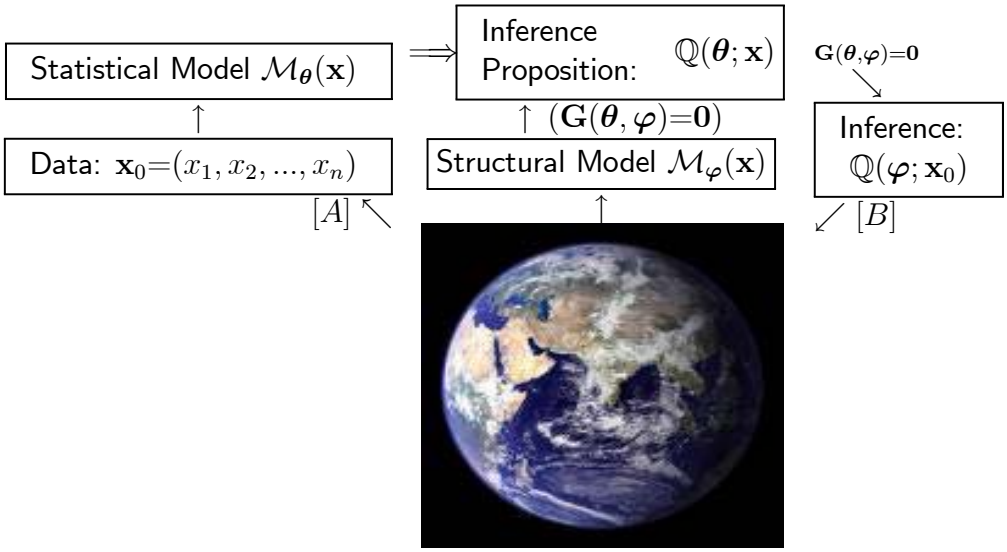

Fig. 1.11: Model-based frequentist statistical induction

These points of nexus with the real world are often neglected in traditional statistics textbooks, but the discussion that follows will pay special attention to the issues they raise and how they can be addressed.

Statistical inference is often viewed as the quintessential form of *inductive inference*: learning form a particular set of data $\mathbf{x}_0$ about the stochastic phenomenon that gave rise to the data. However, it is often insufficiently recognized that this inductive procedure is embedded in a *deductive argument*: if $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$, then $\mathbb{Q}(\boldsymbol{\theta};\mathbf{x})$ where $\mathbb{Q}(\boldsymbol{\theta};\mathbf{x})$ denotes inference propositions (estimation, testing, prediction, policy simulation). The procedure from $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ (the premise) to $\mathbb{Q}(\boldsymbol{\theta};\mathbf{x})$ is *deductive*. Estimators and tests are pronounced *optimal* based on a purely deductive reasoning. In this sense, the reliability (soundness) of statistical inference depends crucially on *the validity of the premises* $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$. The *ninth* recommendation in empirical modeling is:

9. Choose a statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ with a view to ensure that data $\mathbf{x}_0$ constitute a truly typical realization of the stochastic mechanism defined by $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$.

On the basis of the premise $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ we proceed to derive statistical inference results $\mathbb{Q}(\boldsymbol{\theta};\mathbf{x}_0)$ using a deductively valid argument ensuring that *if the premises are valid*,

the conclusions are necessarily (statistically) reliable. To secure the *soundness* of such results one need to establish the adequacy of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ vis-a-vis data $\mathbf{x}_0$. By the same token, if $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ is misspecified the inference results $\mathbb{Q}(\boldsymbol{\theta};\mathbf{x}_0)$ are generally unreliable. Indeed, the *ampliative* (going beyond the premises) dimension of statistical induction relies on the statistical adequacy of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$. The substantive questions of interest are framed in the context of $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{x})$, which is parametrically nested within $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ via the restrictions $\mathbf{G}(\boldsymbol{\theta},\boldsymbol{\varphi})=\mathbf{0}$. When the substantive parameters $\boldsymbol{\varphi}$ are uniquely defined as functions of $\boldsymbol{\theta}$, one can proceed to derive inferential propositions pertaining to $\boldsymbol{\varphi}$, say $\mathbb{Q}(\boldsymbol{\varphi};\mathbf{x})$. These can be used to test any substantive questions of interest, including the substantive adequacy of $\mathcal{M}_{\boldsymbol{\varphi}}(\mathbf{x})$. Hence, the *tenth* recommendation in empirical modeling is:

10. No statistical inference result can be presumed trustworthy unless the statistical adequacy of the underlying model has been secured.

The initial and most crucial step in establishing statistical adequacy is a complete list of the probabilistic assumptions comprising$\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$. Hence, the next several chapters pay particular attention to the problem of statistical model specification.

Departures from the postulated statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ are viewed as systematic information in the data that $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ does not account for that can be detected using *Mis-Specification (M-S) testing*. The statistical model needs to be respecified in order to account for such systematic information. Hence, the procedure is supplemented with the *respecification* stage. Fig. 1.12 depicts the proposed procedure with the added stages, indicated in circular and elliptical shapes, supplementing the traditional perspective. The M-S testing raises an important issue that pertains to the sample size n. For an adequate probing of the validity of $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$, one requires a 'large enough' n for the M-S tests to have sufficient capacity (power) to detect any departures from these assumptions.

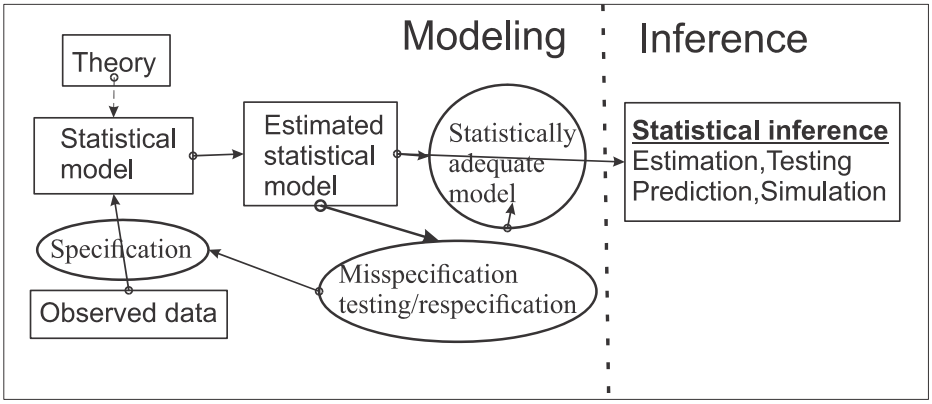

Fig. 1.12: Statistical adequacy and Inference

As shown in chapter 15, even the simplest statistical models that assume a random sample, such as the simple Normal and Bernoulli models call for $n > 40$. This leads to the following recommendation in empirical modeling:

11. If the sample size n is not large enough for a comprehensive testing of the model assumptions, n is not large enough for inference purposes.

6 Statistical vs. substantive information*

In an attempt to provide a more balanced view of empirical modeling and avoid any hasty indictments of the type: "the approach adopted in this book ignores the theory", this section will bring out briefly the proper role of substantive information in empirical modeling (see also Spanos, 1986,1995a, 2010a).

Despite the fact that the statistical model is specified after the relevant data have been chosen, it does not render either the data or the statistical model 'theory-laden'. In addition to the fact that the variables envisioned by the theory often differ from the available data, the chance regularities in the particular data exist independently from any substantive information a modeler might have. Indeed, in detecting the chance regularities one does not need to know what substantive variable the data measure. This is analogous to Shannon's (1948) framing of information theory: "Frequently the messages have meaning; that is, they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem." (p. 379). In direct analogy to that 'the semantic aspects of data are irrelevant to the statistical problem'. In addition, the statistical model is grounded on probabilistic assumptions aiming to account for the chance regularities in the particular data, and is related to the relevant substantive model in so far as it facilitates the posing of the substantive questions of interest after they have been re-framed in statistical terms. Hence, when a statistical model is viewed as a parsimonious description of the stochastic mechanism that gave rise to the particular data, it has 'a life of its own', providing the *inductive premises* for inferences stemming from the data; see Spanos (2006b, 2010a,c).

Observational data are often compiled by governments agencies and private organizations, but such data rarely coincide with the 'ideal' data needed when posing specific substantive questions of interest. Hence, a key recommendation is:

> 12. Never assume that the available data measure the theoretical concepts one has in mind just because the names are very similar (or even coincide)!

A striking example is the theoretical concept *demand* (intentions to buy given a range of hypothetical prices) vs. data on actual *quantities transacted*; see Spanos (1995a). As a result of this gap, empirical modeling in practice attempts to answer substantive questions of interest by utilizing data which contain no such information.

A clear distinction between statistical and substantive information constitutes one of the basic pillars of the empirical modeling methodology advocated in this book; see also Spanos 2006c; 2010c; 2012a). The theory influences the choice of an appropriate statistical model in two indirect ways. First, it demarcates the observable aspects of the phenomena of interest and that determines the relevant data. Second, the theory influences the *parameterization* of the statistical model in so far as the latter enables one to pose substantive questions of interest in its context. Hence, the misspecification testing and respecification facets of empirical modeling are purely statistical procedures guided by statistical information. Hence:

13. No theory, however sophisticated, can salvage a misspecified statistical model, unless it suggests a new statistical model that turns out to be statistically adequate.

As argued in chapter 7, the statistical and substantive perspectives provide very different but complementary viewing angles for modeling purposes; see Spanos (2007).

A statistically adequate $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ accounts for the statistical information in the data, but is often not the ultimate objective of empirical modeling. More often than not, the modeler is interested in appraising the validity of particular substantive information, such as 'is there a causal connection between inflation and money in circulation?' The statistical reliability of such inferences can only be secured when the question is posed in the context of a statistically adequate model. Hence:

14. Success of empirical modeling depends crucially on the skillful synthesizing of the statistical and substantive information, without undermining the credibility of either.

7 Looking ahead

The main objective of the next seven chapters (2-8) is to introduce the necessary probabilistic framework for relating the chance regularity patterns exhibited by data to the proper probabilistic assumptions, with a view to select an appropriate stastistical model. The discussion in chapters 2-4 present a *simple* statistical model as a formalization of a stochastic phenomenon known as a *random experiment*. The interplay between chance regularity patterns and the probabilistic concepts defining a simple statistical model is brought out in chapter 5 using a variety of graphical techniques. The primary objective of chapter 6 is to extend the simple statistical model in directions which enable the modeler to capture certain forms of dependence. Chapter 7 continues the theme of chapter 6 with a view to show that the key to modeling dependence and certain forms of heterogeneity in data is the notion of conditioning, leading naturally to regression and related models. Extending the simple statistical model in directions which enable the modeler to capture several forms of dependence and heterogeneity is completed in chapter 8.

Additional references: Spanos (1989a; 1990a; 2006a; 2010b; 2014b, 2015), Granger (1990), Hendry (2009; 2010)m Mayo and Spanos (2010).

Important concepts

Substantive information, statistical information, stochastic phenomena, chance regularity patterns, deterministic regularity, distribution regularity, dependence regularity, heterogeneity regularity, statistical adequacy, measurement scales, time-series data, cross-section data, panel data.

Crucial distinctions

Statistical vs. substantive subject matter information, chance vs. deterministic regularity patterns, statistical modeling vs. statistical inference, curve-fitting vs. statistical modeling, statistical vs. substantive adequacy chance regularity patterns

vs. probabilistic assumptions, relative frequencies vs. probabilities, induction vs. deduction, time-series vs. cross-section data, variables in substantive models vs. observed data, theoretical concepts vs. data.

Essential ideas

- The primary aim of empirical modeling is to learn from data about phenomena of interest by blending substantive subject matter and statistical information (chance regularity patterns).
- A statistical model comprises a set of internally consistent probabilistic assumptions that defines a stochastic generating mechanism. These assumptions are chosen to account for the chance regularities exhibited the data.
- The traditional metaphor of viewing data as a 'sample from a population' is only appropriate for real-world data that exhibit IID patterns. Hence, the notion of a 'population' is replaced with the concept of a statistical model.
- Chance regularities and the probabilistic assumptions aiming to account for such regularities can be classified into three broad categories: Distribution, Dependence and Heterogeneity.
- Graphical techniques provide indispensable tools for empirical modeling because they can be used to bring out the chance regularities exhibited by data.
- Time series and cross-section data differ only with respect to their ordering of interest. Time, an interval scale variable, is the natural ordering for the former but often cross-section data have several such orderings of interest, whose potential orderings span all four categories of scaling.
- Claims that one does not have to worry about dependence and heterogeneity when modeling cross-section data are highly misleading and misguided.
- Establishing the statistical adequacy of an estimated model is the most crucial step in securing the trustworthiness of the evidence stemming from the data.
- If the sample size is not large enough for properly testing the statistical model assumptions, it is not large enough for inference purposes.
- Assuming that a data series quantifies the variable used in a substantive model just because the names coincide, or very similar, is not a good strategy.

8 Questions and Exercises

1. What determines what phenomena are amenable to empirical modeling?

2. (a) Explain intuitively why statistical information, in the form of chance regularity patterns, is different from substantive subject-matter information.

(b) Explain how these two types of information can be separated, ab initio, by viewing the statistical model as a probabilistic construct specifying the stochastic mechanism that gave rise to the particular data.

(c) The perspective in (b) ensures that data and the statistical model are not 'theory laden'. Discuss.

3. Compare and contrast the notions of chance vs. deterministic regularities.

4. Explain why the slogan "All models are wrong, but some are useful" conflates two different types of being wrong using the distinction between statistical and substantive inadequacy.

5. In relation to the experiment of casting two dice (table 1.3) evaluate the probability of events A-the sum of the two dice is greater than 9, and B-the difference of the two dice is less than 3.

6. Discuss the connection between observed frequencies and the probabilistic reasoning that accounts for those frequencies.

7. In relation to the experiment of casting two dice, explain why focusing on (i) adding up the two faces and (ii) odds and evens constitute two different probability models stemming from the same experiment.

8. Explain the connection between a histogram and the corresponding probability distribution using de Mere's paradox.

9. Give four examples of variables measured on each of the different scales, beyond the ones given in the discussion above.

10. (a) Compare the different scales of measurement.

(b) Why do we care about measurement scales in empirical modeling?

11. Beyond the measurement scales, what features of the observed data are of interest from the empirical modeling viewpoint?

12. (a) In the context of *descriptive statistics*, explain briefly the following concepts: (i) mean, (ii) median, (iii) mode, (iv) variance, (v) standard deviation, (vi) covariance, (vii) correlation coefficient, (viii) regression coefficient.

(b) Explain which of the concepts (i)-(ix) make statistical sense when the data in question are measured on different scales: nominal, ordinal, interval and ratio.

13. Compare and contrast time-series, cross-section and panel data as they relate to heterogeneity and dependence.

14. Explain how the different features of observed data can be formalized in the context of expressing a data series in the form of $\{x_k,\ x_k{\in}\mathbb{R}_X,\ k{\in}\mathbb{N}\}$.

15. Explain briefly the connection between chance regularity patterns and probability theory concepts.

16. Explain the connection between chance regularities and statistical models.

17. Explain the notion of statistical adequacy and discuss its importance for statistical inference.

18. Under what circumstances can the modeler claim that the observed data constitute unprejudiced evidence in assessing the empirical adequacy of a theory?

19. 'Statistical inference is a hybrid of a deductive and an inductive procedure.' Explain and discuss.

20. Discuss the claim: 'if the sample size is not large enough for validating the model assumptions, it is not large enough for reliable inference'.