Summer Seminar: Philosophy of Statistics
Lecture Notes 2: Probability Theory as a Modeling Framework

Aris Spanos [Summer 2019]

1 Introduction

1.1 Primary objective

The primary objective of chapters 2-8 is to introduce probability theory as a mathematical framework for modeling observable stochastic phenomena (chapter 1). Center stage in this modeling framework is occupied by the concept of a *statistical model*, denoted by $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$, that provides the cornerstone of a model-based inductive process underlying empirical modeling.

1.2 Descriptive vs. inferential statistics

The first question we need to consider is:

▶ **Why do we need probability theory**?

The brief answer is that it frames both the **modeling** as well as the relevant inference procedures for empirical modeling. What distinguishes *statistical inference* proper from *descriptive statistics* is the fact that the former is grounded in probability theory. In descriptive statistics one aims to summarize and bring out the important features of a particular data set in a readily comprehensible form. This usually involves the presentation of the data in tables, graphs, charts, and histograms, as well as the computation of summary 'statistics', such as measures of central tendency and dispersion. Descriptive statistics, however, has one very crucial limitation:

conclusions from the data description cannot be extended beyond the data in hand.

A serious problem during the early 20th century was that statisticians would use descriptive summaries of the data, and then proceed to claim generality for their inferences beyond the data in hand.

The conventional wisdom at the time is summarized by Mills (1924) who distinguishes between 'statistical description' and 'statistical induction', where the former is always valid, and "may be used to perfect confidence, as accurate descriptions of the given characteristics" (p. 549), but the latter is only valid when the inherent assumptions of (a) 'uniformity' for the *population* and (b) the 'representativeness' of the *sample* (pp. 550-2) are appropriate for the particular data. **Not really**!

The fine line between *statistical description* and *statistical induction* was blurred until the 1920s, and as a result there was (and, unfortunately, still is) a widespread belief that statistical description *does not* require any *assumptions* because 'it's just a summary of the data'. The reality is that there are *appropriate* and *inappropriate* (misleading) summaries.

Example 2.1. Consider a particular data set data $\mathbf{x}_0{:=}(x_1, x_2, \ldots, x_n)$ whose descriptive statistics for the mean and variance yield the following values:

$$\overline{x}{=}\tfrac{1}{n}\textstyle\sum_{k=1}^{n} x_k{=}12.1, \text{ and } s_x^2{=}\tfrac{1}{n}\textstyle\sum_{k=1}^{n}(x_k{-}\overline{x})^2{=}34.21. \tag{1}$$

There is no empirical justification to conclude from (1) that these numbers are typical of the broader population from where $\mathbf{x}_0$ was observed, and thus representative of the 'population' mean and variance $(E(X),\ Var(X))$; such an inference is *unwarranted*. This is because such inferences presuppose that $\mathbf{x}_0$ satisfies certain probabilistic assumptions that render $(\overline{x}, s_x^2)$ appropriate estimators (appraisers) of $(E(X),\ Var(X))$, but these assumptions need to be empirically validated before such inference becomes warranted. In the case of the formulae behind $(\overline{x}{=}12.1,\ s_x^2{=}34.21)$ the assumptions needed are Independence and Identically Distributed (IID) (chapter 1).

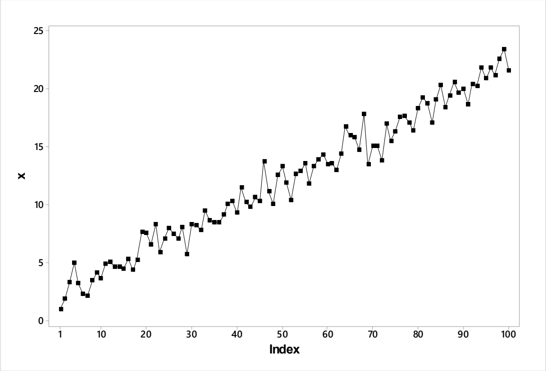

Fig. 2.1: t-plot of data $\mathbf{x}_0$

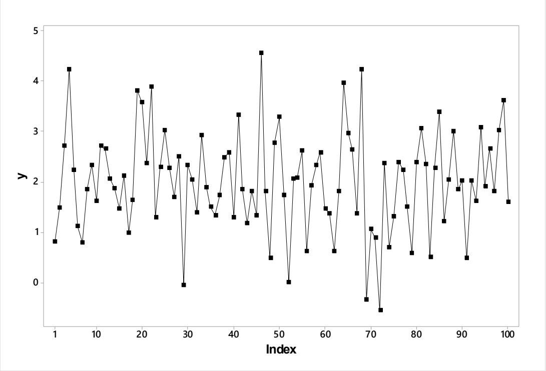

Fig. 2.2: Typical realization of NIID data

Looking at the t-plot of $\mathbf{x}_0$ in figure 2.1, it is clear that the ID assumption is invalid because the arithmetic average of $\mathbf{x}_0$ is increasing with t (the index). This renders the formulae in (1) completely inappropriate for estimating $(E(X),\ Var(X))$, whose *true* values are:

$$E(X){=}2{-}.2t,\ Var(X){=}1, \tag{2}$$

where $t{=}1, 2, ..., n$ is the index; these values are known because the data were created by simulation. The summary statistics in (1) have nothing to do with the true values in (2), because the chance regularities exhibited by the data in fig. 2.1 indicate clearly that the mean is changing with t and the evaluation of the variance using s_x^2 is erroneous when the deviations are evaluated from a fixed $\overline{x}$.

On the other hand, if the data in $\mathbf{x}_0$ looked like the data $\mathbf{y}_0{:=}(y_1, y_2, \ldots, y_n)$ shown in figure 2.2, the formulae in (1) would have given reliable summaries statistics:

$$\overline{y}{=}\tfrac{1}{n}\textstyle\sum_{k=1}^{n} y_k{=}2.01, \text{ and } s_y^2{=}\tfrac{1}{n}\textstyle\sum_{k=1}^{n}(y_k{-}\overline{y})^2{=}1.02,$$

since the true values are $E(Y){=}2,\ Var(Y){=}1$. The lesson from this example is that there is no such a thing as summary statistics that invoke no probabilistic assumptions. There are reliable and unreliable descriptive statistics depending on the validity of probabilistic assumptions implicitly invoked. Indeed, the crucial change pioneered by Fisher (1922a) in recasting descriptive into modern statistics is to bring out these implicit pre-suppositions in the form of a statistical model and render them testable. In this sense, *statistical inference* proper views data $\mathbf{x}_0$ through the prism of a pre-specified statistical model $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$. That is, the data $\mathbf{x}_0$ are being viewed as a typical realization of the stochastic mechanism specified by $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$. The presumption is that $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$ could have generated data $\mathbf{x}_0$. This presumption can be validated vis-a-vis $\mathbf{x}_0$ by testing the probabilistic assumptions comprising $\mathcal{M}_{\boldsymbol{\theta}}(\mathbf{x})$.

In contrast to descriptive statistics, the primary objective of statistical modeling and inference proper is to model (represent in terms of a probabilistic framing) the stochastic mechanism that gave rise to the particular data, and not to describe the data itself. This provides a built in *inductive argument* which enables one to draw inferences and establish generalizations and claims about the *mechanism* itself, including observations beyond the particular data set. This is known as the *ampliative* dimension of inductive inference: reasoning whose conclusions go beyond what is contained in the premises.

2 Simple statistical model: a preliminary view

As mentioned above, the notion of a statistical model takes center stage in the mathematical framework for modeling stochastic phenomena. In this section we attempt an informal discussion of the concept of a simple statistical model at an intuitive level with a healthy dose of hand waving. The main objective of this preliminary discussion is twofold. Firstly, for the less mathematically inclined readers, the discussion, although incomplete, will provide an adequate description of the primary concept of statistical modeling. Secondly, this preliminary discussion will help the reader keep an eye on the forest, and not get distracted by the trees, as the formal argument in sections 3-8 unfolds. The formalization of the notion of a generic random experiment will be completed in chapter 4.

2.1 The basic structure of a simple statistical model

The *simple statistical model*, pioneered by Fisher (1922a), has two components:

$$\begin{array}{ll}\text{[i] Probability model:} & \Phi{=}\left\{f(x;\boldsymbol{\theta}),\ \boldsymbol{\theta}{\in}\Theta,\ x{\in}\mathbb{R}_X\right\},\\ \text{[ii] Sampling model:} & \mathbf{X}{:=}(X_1,X_2,...,X_n)\text{ is a random sample.}\end{array}$$

The *probability model* specifies a family of *densities* $(f(x;\theta),\ \theta{\in}\Theta)$, defined over the range of values ($\mathbb{R}_X$) of the random variable X; one density function for each value of the *parameter* θ, as the latter varies over its range of values Θ: *the parameter space*; hence the term *parametric* statistical model.

Example 2.2. The best way to visualize a probability model is in terms of figure 2.3. This diagram represents several members of a particular family of densities known as the one parameter *Gamma* family and takes the explicit form:

$$\Phi{=}\left\{f(x;\boldsymbol{\theta}){=}\tfrac{\beta^{-1}}{\Gamma[\alpha]}\left(\tfrac{x}{\beta}\right)^{\alpha-1}\exp\left\{-\left(\tfrac{x}{\beta}\right)\right\},\ \boldsymbol{\theta}{:=}(\alpha,\beta){\in}\mathbb{R}_+^2,\ x{\in}\mathbb{R}_+\right\}, \tag{3}$$

where $\Gamma[\alpha]$ denotes the gamma function $\Gamma[\alpha]=\int_0^\infty \exp(-u)\cdot u^{\alpha-1}du$.

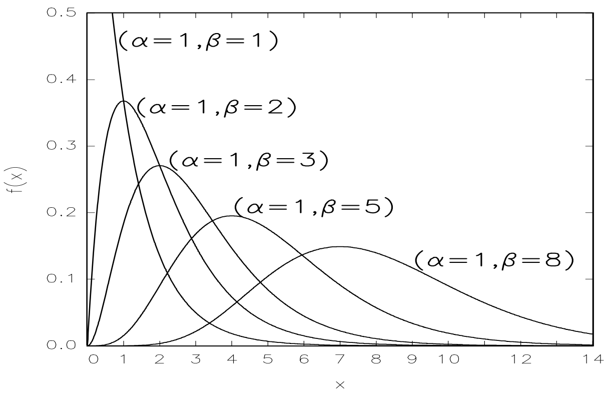

Fig. 2.3: The Gamma probability model

2.2 A random sample: a preliminary view

2.2.1 A statistical model with a random sample

What makes the generic statistical model specified in section 2.2 *simple* is the form of the sampling model, the *random sample* assumption. This assumption involves two interrelated notions known as *Independence* and *Identical Distribution*. These notions can be explained intuitively as a prelude to the more formal discussion that follows.

Independence. The random variables $(X_1, X_2, ..., X_n)$ are said to be *independent* if the occurrence of any one, say X_i, does not influence and is not influenced by the occurrence of any other random variable in the set, say X_j, for $i \neq j,\ i,j{=}1,2,...,n$.

Identical Distribution. The independent random variables $(X_1, X_2, ..., X_n)$ are said to be *identically distributed* if their density functions are identical in the sense:

$$f(x_1;\theta){=}f(x_2;\theta){=}\cdots{=}f(x_n;\theta).$$

For observational data the validity of the IID assumptions can often be assessed using a battery of graphical techniques discussed in chapters 5-6.

2.2.2 Experimental data: sampling and counting techniques

Sampling refers to a procedure to select a number of objects (balls, cards, persons), say r, from a larger set, we call the target 'population', with n $(n \geq r)$ such objects. The sampling procedure gives rise to a random sample (IID) when:

(i) the probability of selecting any one of the population objects is the same, and

(ii) the selection of the i-th object does not affect and it is not affected by the selection of the j-th object for all $i{\neq}j,\ i,j{=}1,2,...,n$.

Two features of the selection procedure matter, whether we replace an object after being selected or we do not, and whether the order of the selected objects matters or not. This give rise to the four way classification in table 2.1 for which the assignment of the common probability of an object being selected is different.

Table 2.1: Sampling procedure probabilities

$O \setminus R$	replacement (R)	no replacement $(\bar{R})$
order (O)	$\left(\frac{1}{n^r}\right)$	$\left(\frac{1}{P_r^n}\right)$, where $P_r^n=\frac{n!}{(n-r)!}$
no order $(\bar{O})$	$\left(\frac{1}{C_n^{n+r-1}}\right)$	$\left(\frac{1}{C_r^n}\right)$, where $C_k^n=\frac{n!}{r!(n-r)!}$

NOTE that all the formulae in table 2.1 assume that the probabilies come from a single distribution $f(x)$, $x{\in}R_X$, i.e. the assumption of **Identically Distributed** (ID) is implicitly imposed.

To shed light on the formulae in table 2.1 let us state a key counting rule.

Multiplication counting rule. Consider the sets $S_1,\ S_2, \ldots, S_k$ with $n_1,\ n_2, \ldots, n_k$ elements, respectively. Then the number of ways one can choose k elements, one from each of these sets, is: $n_1 \times n_2 \times \ldots \times n_k$.

Example 2.8. The number of ways one can choose r elements from a set of n elements is: n^r.

Combinations. An *unordered* subset of r elements from a set S containing n elements $(0{<}r{\leq}n)$ is said to constitute an r-element combination of S. The number of such r-element combinations is equal to:

$$C_r^n{:=}\binom{n}{r}{=}\frac{n!}{r!(n-r)!}$$

Permutations. An *ordered* subset of r elements from a set S containing n elements $(0{<}r{\leq}n)$ is said to constitute an r-element permutation of S. The number of such r-element permutation is equal to:

$$P_r^n{=}n{\cdot}(n{-}1){\cdot}(n{-}2){\cdots}(n{-}r{+}1){=}\frac{n!}{(n-r)!}.$$

For a better understanding of the notion of a random sample, it is worth considering the question of ensuring the appropriateness of IID assumptions in the case of sample survey data using a simple Bernoulli model.

Example 2.12. Consider the problem of designing a sample survey in order to evaluate the voting intentions of the USA electorate in a forthcoming presidential election. Assuming that there are only two candidates, the Democrat (D) and Republican (R) nominees, we can define the random variable:

$$X(\text{D}){=}1,\ X(\text{R}){=}0,\ \text{with}\ \mathbb{P}(X{=}1){=}\theta,\ \mathbb{P}(X{=}0){=}1{-}\theta.$$

This enables us to use the Bernoulli distribution and the question which arises is how to design a sample survey, of size $n{=}1000$, so as to ensure the randomness of the sample realization. To get some idea on what the notion of a random sample entails, let us consider a number of ways to collect sample surveys which *do not* constitute a random sample:

(a) Picking "at random" 1000 subscribers from the **local telephone directory** and ask them to declare their voting intentions.

(b) Sending a team of students to the local **shopping mall** to ask the first 1000 potential voters entering the mall.

(c) Driving through all 51 states, stop outside the **main post office of the state capital** and ask as many voters as the ratio of the voters of that state, to the total voting population allows.

In all three cases our action will not give rise to a **random sample** because:

(i) it does not give every potential voter the same likelihood of being asked; not everybody has a phone or goes to the mall, and

(ii) the local nature of the selection in cases (a) and (b) excludes the majority of the voting population; this induces some potential *heterogeneity* and *dependence* into the sample; asking people from the same family is likely to introduces dependence.

2.2.3 Sample survey procedures

Simple random sampling.

Stratified sampling.

Cluster sampling.

Quota sampling.

The **chance set-up** assumed in all these cases of **survey sampling** is one of a target population where each selected unit is assumed to come from the same distribution (univariate). This invariably imposes Identically Distributed (ID) and often Idependence (I). Hence, the terminology associated with **sampling from a target population** can be misleading in cases of observing phenomena that are changing during the sampling process. As argued in chapter 1, what renders data amenable to statistical modeling and inference is whether they exhibit *chance regularity patterns*, and not whether they can be thought of as a sample (random or not) from a population. The most appropriate metaphor for such as broader framework is the (notional) existence of a stochastic generating mechanism (a statistical model) that could have given rise to the data.

3 Probability theory: an introduction

3.1 Kolmogorov's axiomatic framing of probability

This is the approach adopted when **teaching students of mathematics**!

Primitive notions: $(S, \Im, \mathbb{P}(.))$ – a probability space

- S is a set of elementary outcomes; $S \neq \varnothing$.
- $\Im$ is a collection of subsets of S (we call **events**), assumed to be a $\sigma-$*field*, i.e. it satisfies the following conditions:

$$\begin{array}{ll} \text{(i)} & S \in \Im, \\ \text{(ii)} & \text{if } A \in \Im, \text{ then } \overline{A} \in \Im, \\ \text{(iii)} & \text{if } A_i \in \Im \text{ for } i=1,2,...,n,... \text{ the set } (\bigcup_{i=1}^{\infty} A_i) \in \Im. \end{array}$$

- $\mathbb{P}(.)$ is *a probability set function*:

$$\mathbb{P}(.)\colon \Im \rightarrow [0,1],$$

assigning probabilities to **events** which satisfies the following **axioms**:

[**1**] $\mathbb{P}(S)=1$, for any outcomes set S,

[**2**] $\mathbb{P}(A) \ \geq 0$, for any event $A \in \Im$,

[**3**] *Countable Additivity.* For a countable sequence of mutually exclusive events, i.e. $A_i \in \Im, i=1,2,...,n,..$ such that $A_i \cap A_j = \emptyset$, for all $i \neq j$, $i,j=1,2,...,n,...,$

then $\quad \mathbb{P}(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$.

3.2 Mathematical deduction

As a deductive science, mathematics begins with a set of fundamental statements we call axioms (the premises) and ends with other fundamental statements we call theorems which are derived from the axioms using deductive logical inference.

Accepting the axioms [**A1**]-[**A3**] (table 2.9) as "true" we can proceed to derive certain corollaries which provide a more complete picture of the mathematical framework.

Theorem 1. $\mathbb{P}(\bar{A})=1-\mathbb{P}(A)$, for any $A \in \Im$.

Since $\bar{A} \cup A = S$, and $\bar{A} \cap A = \varnothing$, we can use axioms [**A1**] and [**A3**] to deduce that:

$$\mathbb{P}(S)=1=\mathbb{P}(\bar{A} \cup A)=\mathbb{P}(\bar{A})+\mathbb{P}(A).$$

The first equality is axiom [**A1**], the second follows from the fact that $\bar{A} \cup A = S$, and the third from the fact that $\bar{A} \cap A = \varnothing$ and axiom [**A3**].

Example 2.44. In the case of tossing a coin twice let $A{=}\{(HH),(HT),(TH)\}$. Given that $\bar{A}{=}\{(TT)\}$, using theorem 1 we can deduce that $\mathbb{P}(\bar{A}){=}\frac{1}{4}$.

The next result is almost self-evident but in mathematics we need to ensure that it follows from the axioms. Using theorem 1 for $A{=}S$ (and hence $\bar{A}{=}\varnothing$), we deduce:

Theorem 2. $\mathbb{P}(\varnothing){=}0$.

The next theorem extends axiom [**A3**] to the case where $(A\cap B){\neq}\varnothing$.

Theorem 3. $\mathbb{P}(A\cup B){=}\mathbb{P}(A)+\mathbb{P}(B)-\mathbb{P}(A\cap B)$, for any $A{\in}\Im,\ B{\in}\Im$.

The way to prove this is to define $A\cup B$ in terms of mutually exclusive events and then use [**A3**]. It is not difficult to see that the events $C{=}\{A-(A\cap B)\}$ and B are mutually exclusive and $C\cup B{=}A\cup B$. Hence, by axiom [**A3**]:

$$\mathbb{P}(A\cup B){=}\mathbb{P}(C\cup B){=}\mathbb{P}\{A{-}(A\cap B)\}{+}\mathbb{P}(B){=}\mathbb{P}(A){+}\mathbb{P}(B){-}\mathbb{P}(A\cap B).$$

The **usefulness** of the axiomatic approach stems from the fact that it renders probability theory part of mathematics proper, securing the validity of all its theorems, irrespective of any particular *interpretation of probability* one prefers.

Questions that naturally arise:

Why these particular primitive notions $(S,\Im,\mathbb{P}(.))$?

Why is the set of events of interest $\Im$ a sigma-field?

Why these particular axioms?

To answer these questions the discussion that follows motivates the axiomatic framing of probability theory given above, by formalizing (mathematizing) **a simple stochastic mechanism**, we call a Random Experiment defined in terms of *plain English*!

3.3 The notion of a random experiment

The notion of an experiment is used to denote any process, actual or hypothetical, whose possible outcomes are known at the outset. A special case of that is a *random experiment* $\mathcal{E}$, is defined as a simple chance mechanism which satisfies conditions [a]-[c] in table 2.2.

Table 2.2: Random Experiment ($\mathcal{E}$)

[a]	All possible distinct outcomes are known at the outset.
[b]	In any particular trial the outcome is not known in advance, but there exist discernible regularities pertaining to the frequency of occurrence associated with different outcomes.
[c]	The experiment can be repeated under identical conditions.

The purpose of introducing $\mathcal{E}$ is twofold. First, to give a verbal description of a simple stochastic phenomenon we have in mind, that is amenable to statistical modeling.

Second, to bring out its essential features and proceed to formalize them in a precise mathematical form to motivate the introduction of needed probabilistic concepts.

Example 2.13. [i] Toss a coin and note the outcome. Assuming that we can repeat the experiment under identical conditions, this is a random experiment because the above conditions are satisfied. The possible distinct outcomes are: $\{H, T\}$, where (H) and (T) stand for "Heads" and "Tails", respectively.

[ii] Toss a coin twice and note the outcome. The possible distinct outcomes are:

$$\{(HH), \quad (HT), \quad (TH), \quad (TT)\}.$$

[iii] Toss a coin thrice and note the outcome. The possible distinct outcomes are:

$$\{(THH), (HHH), (HHT), (HTH), (TTT), (HTT), (THT), (TTH)\}.$$

[iv] Tossing a coin until the first "H" occurs. The possible distinct outcomes are:

$$\{(H), \quad (TH), \quad (TTH), \quad (TTTH), \quad (TTTTH), \quad (TTTTTH), \ldots\}.$$

[v] A hacker is repeatedly and persistently trying to break into a company's computer server. Count the number of attempts needed for a successful break-in. This represents a more realistic case of a stochastic phenomenon but it can be viewed as a random experiment since the above conditions can be ensured in practice. The possible distinct outcomes include all natural numbers: $\mathbb{N}$:=$\{1, 2, 3, ...\}$.

[vi] Count the number of calls arriving in a telephone exchange over a period of time. The possible distinct outcomes include all integers from 0 to infinity: $\mathbb{N}_0$:=$\{0, 1, 2, 3, ...\}$.

[vii] Measure the lifetime of a light bulb in a typical home environment. In theory the possible distinct outcomes include any real number from zero to infinity: $[0, \infty)$.

Let us also mention an observable stochastic phenomenon which *does not* constitute a random experiment.

[viii] Observe the closing daily price of IBM shares on the New York stock exchange. The conditions [a]-[b] of a random experiment are easily applicable. [a] The possible distinct outcomes are real numbers between zero and infinity: $[0, \infty)$. [b] The closing IBM share price in a particular day is not known in advance. Condition [c], however, is inappropriate because the circumstances from one day to the next change and today's share prices are related to yesterday's. Millions of people use this information in an effort to "buy low" and "sell high" to make money.

The forest: the formalization of a random experiment into a simple statistical model is the main objective of this and the next two chapters. The notion of a random experiment is given a mathematical formulation in the form of a *simple statistical model* in the next two chapters.

The trees: the introduction of numerous mathematical concepts which enable us to formalize $\mathcal{E}$ into the simple statistical model providing the basis for a more general mathematical framework that underpins empirical modeling.

4 Formalizing condition [a]: the outcomes set

4.1 The concept of a set in set theory

The first step in constructing a mathematical model for a RE ($\mathcal{E}$) is to formalize the notion of all distinct outcomes. We do this by collecting the outcomes together and defining a set. The naive (as opposed to the axiomatic) notion of *a set* is used informally as a well-defined collection of distinct objects which we call its *elements*.

Example 2.14. Let $S=\{\blacklozenge, \heartsuit, \clubsuit, \spadesuit\}$ be a set with elements the card suits: diamonds, hearts, clubs and spades. If S is a set and $\clubsuit$ one its elements, it is denoted by $\clubsuit \in S$, where '$\in$' reads 'belongs to'. If $\blacktriangle$ is not an element of S, it is denoted by $\blacktriangle \notin S$, and is read '$\blacktriangle$ does not belong to S'. The card suits are distinct objects when viewed separately, but they form a single entity S when viewed collectively $\{\blacklozenge, \heartsuit, \clubsuit, \spadesuit\}$.

REMARKS:

(i) The 'membership' ($\in$) notion is one of the crucial primitive concepts of set theory.

(ii) Mathematically speaking what the objects defining a set denote is irrelevant.

(iii) The notion of 'distinct elements' is very important.

4.2 The outcomes set

A set S which includes *all possible distinct outcomes* of the experiment in question is called an *outcomes set*. NOTE: that the established terminology refers to S as the 'sample space'. This term is avoided because S is neither a 'space' nor has anything to do with term 'sample' as used later in chapter 4.

Condition [a] of a random experiment $\mathcal{E}$ is formalized using the idea of *a set*. In set theoretic language the outcomes set S is called the *universal set*. This might seem like a trivial step but in fact it provides the key to the rest of the formalization. In particular, set theory will be instrumental in formalizing condition [b].

Example 2.15. The outcomes sets for the random experiments in example **2.13**:

$$\begin{aligned}
S_1 =& \ \{H, T\}, \\
S_2 =& \ \{(HH), (HT), (TH), (TT)\}, \\
S_3 =& \ \{(THH), (HHH), (HHT), (HTH), (TTT), (HTT), (THT), (TTH)\}, \\
S_4 =& \ \{(H), (TH), (TTH), (TTTH), (TTTTH), (TTTTTH), ...\}.
\end{aligned}$$

In order to utilize the notion of *the outcomes set* effectively, we need to introduce some set theoretic notation which will be used extensively in this book. The way we defined *a set* in the above examples was by listing its elements.

An alternative way to define a set is to use a *property* shared by all the elements of the set. For example, the outcomes set for experiment [v] can be written as:

$$S_5=\{x\colon\ x\in\mathbb{N}:=\{1,2,3,...\}\},$$

which reads "S_5 is the set of all x'*s* such that x belongs to $\mathbb{N}$," i.e., x is a *natural number*. Similarly, the set of all *real numbers* can be written as:

$$\mathbb{R}=\{x\colon\ x \text{ a real number}, \ -\infty < x < \infty\}.$$

Using this set we can write the outcomes set for experiment [**vii**] as:

$$S_7=\{x\colon\ x\in\mathbb{R},\ 0\leq x<\infty\}.$$

NOTE: a shorter notation for this set is: $S_7=[0,\infty)$. NOTE that when a square bracket is used, the adjacent element is included in the set, but when an ordinary bracket is used it is excluded. Table 2.3 lists some of the most important intervals on the real line.

Table 2.3: Types of intervals on the real line

(i)	singleton:	$\{a\}$,
(ii)	closed interval:	$[a,b]=\{x\colon\ x\in\mathbb{R},\ a\leq x\leq b\}$,
(iii)	open interval:	$(a,b)=\{x\colon\ x\in\mathbb{R},\ a<x<b\}$,
(iv)	half-closed interval:	$(-\infty,a]=\{x\colon\ x\in\mathbb{R},\ -\infty<x\leq a\}$.

4.3 Special types of sets

In relation to the above examples, it is useful to make two distinctions. The first is the distinctions between finite and infinite sets and the second is the further division of infinite sets into countable and uncountable. A set A is said to be *finite* if it can be expressed in the following form:

$$A=\{a_1,a_2,...,a_n\}\text{ for some integer }n.$$

A set that is not finite is said to be *infinite*.

Example 2.16. (a) The set $C=\{\clubsuit,\diamondsuit,\heartsuit\}$ is finite.

(b) The intervals (ii)-(iv) in table 2.3 define infinite sets of numbers.

(c) Table 2.4 lists several important infinite sets of *numbers*.

Table 2.4: Different sets of numbers on the real line

(i)	Natural numbers:	$\mathbb{N}=\{1,2,3,...\}$,
(ii)	Integers:	$\mathbb{Z}=\{0,\ \pm1,\ \pm2,\ \pm3,\ ....\}$,
(iii)	Real numbers	$\mathbb{R}=\{x\colon\ x\in\mathbb{R},-\infty<x<\infty\}$,

Among the infinite sets we need to distinguish between the ones whose elements we can arrange in a sequence and those whose elements are so many and so close together that no such ordering is possible. For obvious reasons we call the former countable and the latter uncountable. More formally, a set A is said to be *countable* if it's either finite or infinite and each element of A can be matched with a distinct natural number, i.e., there is a one-to-one matching of the elements of A with the elements of $\mathbb{N}$.

Example 2.17. The set of even natural numbers is *countable* because we can define the following one-to-one correspondence between $\mathbb{N}_{even}$ and $\mathbb{N}$:

$$\begin{array}{rcccccccc}\mathbb{N}_{even}:=\{ & 2 & 4 & 6 & 8 & 10 & \cdots & 2n & \cdots\} \\ & \updownarrow & \updownarrow & \updownarrow & \updownarrow & \updownarrow & & \updownarrow & \\ \mathbb{N}:=\{ & 1 & 2 & 3 & 4 & 5 & \cdots & n & \cdots\}.\end{array}$$

In view of the fact that between any two natural numbers, say $[1, 2]$, there is an infinity of both rational and real numbers, intuition might suggest that the two sets $\mathbb{Q}$ and $\mathbb{R}$ have roughly speaking the same number of elements. In this case intuition is wrong! The set of real numbers is more numerous than the set of rational numbers**:**

$$\aleph_1 := [\text{number of elements of } \mathbb{R}] > \aleph_0 := [\text{number of elements } \mathbb{Q}] .$$

The different magnitudes of infinite sets we call their *cardinality;* see Binmore (1980) An infinite set whose number of elements is of the same cardinality as that of $\mathbb{R}$, is called *uncountable,* but there exist infinite sets with greater cardinality than $\aleph_1$.

Example 2.18. The sets $\mathbb{R}$, $\mathbb{R}^n$, $[a, b]$, (a, b), $(-\infty, x]$ are *uncountable.*

Historical aside. The father of modern set theory, **Georg Cantor** (1845-1918), introduced the idea of infinite sets with different cardinality beginning with $\aleph_0$ and $\aleph_1$. Most of the mathematicians in the late 19th century met Cantor's ideas with open hostility. Poincaré (1854–1912) referred to his ideas as a 'grave disease' infecting the discipline of mathematics, and Kronecker (1823–1891) voiced numerous criticisms that degenerated into personal attacks describing Cantor as a 'charlatan' and 'renegade'. The open hostility from his peers is often blamed for Cantor's recurring bouts of depression from 1884 to the end of his life. He died in 1918, in the sanatorium where he had spent the final five years of his life; see Dauben (1990).

5 Formalizing condition [b]: events & probabilities

Having formalized condition [a] of **random experiment** ($\mathcal{E}$) in the form of an outcomes set, we can proceed to formalize the second condition:

[b] In any particular trial the outcome is not known in advance,
but there exist discernible regularities pertaining to the
frequency of occurrence associated with different outcomes.

This condition entails two dimensions which appear contradictory at first sight. The first dimension is that individual outcomes are largely unpredictable but the second is that there exists some knowledge about their occurrence. In tossing a coin twice we have no idea which of the four outcomes will occur but we know that there exists some regularity associated with these outcomes. The way we deal with both of these dimensions is to formalize the perceptible regularity at the aggregate level. This formalization will proceed in two steps. The first involves the formalization of the notion of *events of interest* and the second takes the form of *attaching probabilities* to these events.

In this introduction we used a number of new notions which will be made more precise in what follows. One of these notions is that of an *event.* Intuitively, an event is a statement in relation to a random experiment for which the only thing that matters is whether in a particular trial an event has occurred or not. So far the only such statements we encountered are the *elementary outcomes.* For modeling purposes, however, we need to broaden this set of statements to include not just elementary outcomes but also combinations of them.

▶ **How do events differ from elementary outcomes?**

Example 2.19. In the context of the random experiment **[ii]:** tossing a coin twice with the outcomes set $S_2{:=}\{(HH),(HT),(TH),(TT)\}$ we might be interested in the following events:

(a) A- at least one H: $A{=}\{(HH),(HT),(TH)\}$,

(b) B- two of the same: $B{=}\{(HH),(TT)\}$,

(c) C-at least one T: $C{=}\{(HT),(TH),(TT)\}$.

In general, events are formed by *combining elementary outcomes* using set theoretic operations, and we say that an event A has occurred when any one of its elementary outcomes occurs. In order to make this more precise we need to take a detour into set theory to define certain basic set theoretic notions and operations.

5.1 Set theoretic operations

Subsets. The concept of an event is formally defined using the notion of a subset.

If A and S are sets, we say that A is a *subset* of S and denote it by $A\subset S$ if every element of A is also an element of S. More formally:

$$A\subset S \text{ if for each } a{\in}A \text{ implies } a{\in}S.$$

Example 2.20. (a) The set $D_1{=}\{\clubsuit,\heartsuit\}$ is said to be a *subset* of $D{=}\{\clubsuit,\diamondsuit,\heartsuit\}$, and denoted by $D_1{\subset}D$, because every element of D_1 is also an element of D.

(b) The sets $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}_+$ introduced above are all subsets of $\mathbb{R}$.

(c) In the case of the outcomes set

$$S_2{:=}\{(HH),(HT),(TH),(TT)\}$$

there are four elementary outcomes. By combining these we can form events such as:

$$A{=}\{(HH),(HT),(TH)\},\ B{=}\{(HH),(TT)\},\quad C{=}\{(HH)\},\ D{=}\{(HT),(TH)\}.$$

More formally *events* are subsets of S formed by applying the following set theoretic operations: *Union* ($\cup$), *Intersection* ($\cap$) and *Complementation* ($^{-}$), among the elements of S. It is worth noting that all these operations are defined in terms of the primitive notion of membership ($\in$) of a set.

Union. The *union* of A and B, denoted by $A\cup B$, is defined as follows:

$A\cup B$: the set of outcomes that are either in A or B (or both).

More formally: $A\cup B{:=}\{x\text{: } x{\in}A \text{ or } x{\in}B\}$; see figure 2.4.

Example 2.21. For the sets $A{=}\{(HH),(TT)\}$ and $B{=}\{(TT),(TH)\}$:

$$A\cup B{=}\{(HH),(TH),(TT)\}.$$

Intersection. The *intersection* of A and B, denoted by $A\cap B$, is defined as follows:

$A\cap B$: the set of outcomes that are in both A and B.

More formally: $A\cap B{:=}\{x\text{: } x{\in}A \text{ and } x{\in}B\}$; see figure 2.4.

Example 2.22. For events A and B defined in example 2.20-(c): $A{\cap}B{=}\{(TT)\}$.

Complementation. The *complement* of an event A, relative to the universal set S, denoted by $\bar{A}$, is defined by:

$\bar{A}$: the set of outcomes in the universal set S which are not in A.

More formally: $\bar{A}{:=}\{x{:}\ x{\in}S$ and $x{\notin}A\}$; see figure 2.4.

All three operations are illustrated in figure 2.4 using Venn diagrams. Note that the rectangle in the Venn diagrams represents, by definition, the outcomes set S.

Example 2.23. (a) For events A and B defined in example 2.20-(c): $\bar{A}{=}\{(TT)\}$, $\bar{B}{=}\{(HT),(TH)\}$. The union of A and $\bar{A}$ gives S i.e., $A{\cup}\bar{A}{=}S$ and $A{\cap}\bar{A}{=}\{\ \}{:=}\varnothing$, i.e. their intersection is the empty set. Also, $\bar{S}{=}\varnothing$ and $\overline{\varnothing}{=}S$.

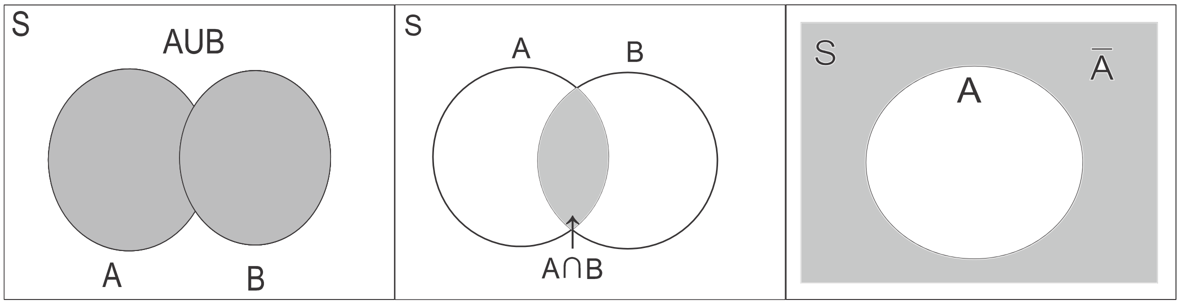

Figure 2.4: Venn diagrams depicting the basic set theoretic operations

(b) For $A{=}\{(HH),(HT),(TH)\},\ B{=}\{(HH),(TT)\},\ C{=}\{(HH)\},\ D{=}\{(HT),(TH)\}$:

$$A\cap B{=}\{(HH)\}{=}C \text{ and } B\cap D{=}\varnothing,$$

where $\varnothing$ denotes the *empty set*.

Using complementation we can define a duality result between unions and intersections in the form of *de Morgan's laws*:

$$[1]\ (\overline{A\cup B\cup C}){=}\bar{A}\cap\bar{B}\cap\bar{C},\ \ [2]\ (\overline{A\cap B\cap C}){=}\bar{A}\cup\bar{B}\cup\bar{C}.$$

Example 2.24. For the sets $A{=}\{(HH),(HT)\}$, and $C{=}\{(HH)\}$, $(A{\cup}C){=}A$ and thus:

$$(\overline{A{\cup}C}){=}\bar{A}{=}\{(TT)\}.$$

On the other hand $\bar{C}{=}\{(HT),(TH),(TT)\}$. Hence, $\bar{A}{\cap}\bar{C}{=}\{(TT)\}{=}(\overline{A{\cup}C})$.

Example 2.25. For the sets A and C defined above, $(A\cap C){=}C$ and thus $(\overline{A\cap C}){=}\bar{C}$. In contrast $\bar{A}\cup\bar{C}{=}\{(HT),(TH),(TT)\}{=}\bar{C}$.

Equality of sets. Two sets are equal if they have the same elements. We can make this more precise by using the notion of a subset to define equality between two sets. In the case of two sets A and B if:

$$A\subset B \text{ and } B\subset A \text{ then } A{=}B.$$

Example 2.26. For the sets $A{=}\{\diamondsuit,\heartsuit\}$ and $B{=}\{\heartsuit,\diamondsuit\}$, we can state that $A{=}B$; NOTE that the order of the elements in a set is unimportant.

5.2 Events vs. outcomes

In set-theoretic language, an *event* is a *subset* of the outcomes set S i.e.

If $A \subset S,\ A$ is an *event.*

In contrast, an *elementary outcome* s is an element of S, i.e.

If $s{\in}S,\ s$ is an *elementary outcome.*

That is, an outcome is also an event but the converse is not necessarily true. In order to distinguish between a subset and an element of a set consider the following example.

Example 2.27. For sets $D{=}\{\clubsuit, \diamondsuit, \heartsuit\}$ and $C{=}\{\clubsuit, \heartsuit\}$ it is true that: $C{\subset}D$ but $C{\notin}D$. In contrast, the set: $E{=}\,\{(\clubsuit, \heartsuit), \diamondsuit\}$ has two elements $(\clubsuit, \heartsuit)$ and $\diamondsuit$: $C{\in}E$.

The crucial property of an event is whether in a trial it has occurred or not. We say that $A{=}\{a_1, a_2, ..., a_k\}$ has *occurred* if one of its elements $a_1, ..., a_k$ has occurred.

5.2.1 Special events

In the present context there are two important events we need to introduce. The first is S itself (the universal set), referred to as the *sure event*: whatever the outcome, S occurs. In view of the fact that S is always a subset of itself $(S \subset S)$, we can proceed to consider the empty set: $\varnothing{=}S - S$, called the *impossible event*: whatever the outcome, $\varnothing$ does not occur. NOTE that $\varnothing$ is always a subset of every S.

Mutually exclusive events. Using the impossible event we can define an important relation between two sets. Any two events A and B are said to be *mutually exclusive* if: $A \cap B{=}\varnothing$.

Using the notion of mutually exclusive events in conjunction with S we define an important family of events.

Partition. The events $A_1, A_2, ..., A_m$ constitute a *partition* of S if they satisfy (i)-(ii) in table 2.5.

Table 2.5: Definition of a Partition of S

(i) **mutually exclusive**:	$A_i \cap A_j{=}\varnothing,$ for $i{\neq}j,\ i,j{=}1, 2, ..., m,$
(ii) **exhaustive**:	$\bigcup_{i=1}^{m} A_i{=}S.$

5.3 Event space

As argued at the beginning of this section the way we handle uncertainty relating to the outcome of a particular trial is first to define the events of interest and then to articulate it in terms of probabilities attached to different events of interest. Having formalized the notion of an event as a subset of the outcomes set, we can proceed to make more precise the notion of *events of interest.*

An **event space** $\Im$ is a set of the events of interest and related events; those we get by combining the events of interest using set theoretic operations. It is necessary to include such events because if we are interested in events A and B, we are also interested (indirectly) in the related events $\bar{A}$, $\bar{B}$, $A\cup B$, $A\cap B$, $(\bar{A}_1\cap\bar{A}_2)$, etc.,

$\bar{A}$: denotes the non-occurrence of A.
$(A\cup B)$: denotes the event that at least one of the events A or B occurs.
$(A\cap B)$: denotes the event that both A and B occur simultaneously.

In set theoretic language, an event space $\Im$ is *a set of subsets of* S which is *closed under the set theoretic operations* of union, intersection and complementation; when these operations are applied to any elements of $\Im$, the result is also an element of $\Im$.

For any outcomes set S we can consider two extreme event spaces:

(a) $\Im_0=\{S,\varnothing\}$: the *trivial* event space,

(b) $\mathcal{P}(S)=\{A\colon A\subset S\}$, the *power set* is the set of all possible subsets of S, including S and $\varnothing$.

Neither of these extreme cases is very interesting for several reasons.

(i) The **trivial** event space $\Im_0$ is not very interesting because it contains no information; S and $\varnothing$ are known in advance.

(ii) The **power set**. At first sight the set of all subsets of S seems to be an obvious choice for the event space, since it includes all the relevant events and is closed under the set theoretic operations of union, intersection and complementation.

Example 2.28. In the case of the random experiment of tossing a coin twice with outcomes set

$$S_2=\{(HH),(HT),(TH),(TT)\},$$

the power set takes the form:

$$\mathcal{P}(S_2)=\{S_2,\ \varnothing,\ \{(HH)\},\ \{(HT)\},\ \{(TH)\},\ \{(TT)\},\ \{(HH),(HT)\},\ \{(HH),(TH)\},$$
$$\{(HH),(TT)\},\ \{(HT),(TH)\},\ \{(HT),(TT)\},\ \{(TH),(TT)\},\ \{(HH),(HT),(TH)\},$$
$$\{(HH),(HT),(TT)\},\ \{(HH),(TH),(TT)\},\ \{(TT),(HT),(TH)\}\}\,;$$ see figure 2.3.

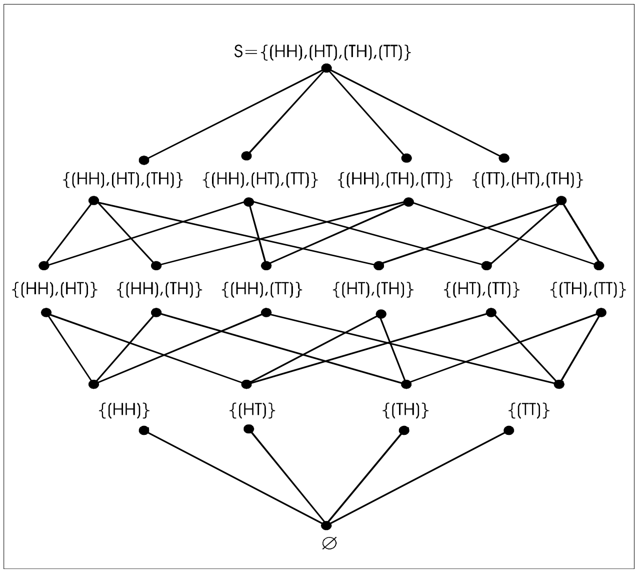

Fig. 2.3: Constructing a power set

Despite its obvious allure, the power set $\mathcal{P}(S)$ runs into practical and technical difficulties.

The general counting principle (chapter 1) can be used to evaluate the number of elements in $\mathcal{P}(S)$.

Case 1: S is finite. For $S{=}\{s_1, s_2, \ldots, s_n\}$ the power set has 2^n elements (events). Why? Each of the elements of S might (1) or might not (0) belong to a particular subset. More formally, there is a one-to-one mapping between the subsets of S and the sequence of 0's and 1's of length n. The general counting principle implies that the number sequences of 0's and 1's of length n is: $\overbrace{2 \times 2 \times \cdots \times 2}^{n \text{ times}} = 2^n$.

Example 2.29. Consider the case of tossing a coin three times. The outcomes set S_3 has 8 elements which implies that its power set has $2^8{=}256$ elements; too many to enumerate.

Case 2: S is infinite but countable. By the same counting principle, when $S{=}\{s_1, s_2, \ldots, s_n, \ldots\}$ the power set has $2^{\mathbb{N}}$ elements. It turns out, however, that $2^{\mathbb{N}}$ is uncountably infinite; see Binmore (1980).

Case 3: S is uncountably infinite. In this case the power set has more elements than $\mathbb{R}$.

These results suggest that the power set is impossible to handle in cases 2 and 3, as well as impractical for case 1 when n large; see Billingsley (1995).

Kolmogorov (1933) showed a way to circumvent these practical and technical difficulties, associated with the power set, by bestowing to the event space a specific mathematical structure (a field or a σ-field).

Example 2.30. If we return to the random experiment of tossing a coin three time, if the events of interest are only, say $A_1{=}\{(HHH)\}$ and $A_2{=}\{(TTT)\}$, there is no need to use the power set as the event space. Instead, we can define:

$$\Im_3{=}\{S, \varnothing, A_1, A_2, (A_1 \cup A_2), \bar{A}_1, \bar{A}_2, (\bar{A}_1 \cap \bar{A}_2)\},$$

which has only 8 elements. We can verify that $\Im_3$ is closed under the set theoretic operations:

$$(S_3 \cup \varnothing){=}S_3{\in}\Im_3,\ (S_3 \cap \varnothing){=}\varnothing{\in}\Im_3,\quad \bar{S}_3{=}\varnothing{\in}\Im_3,\ \overline{\varnothing}{=}S_3{\in}\Im_3,$$
$$(\bar{A}_1 \cup \bar{A}_2){=}(\overline{A_1 \cap A_2}){\in}\Im_3,\ \text{etc.}$$

The concept of an event space plays an important role in the formalization of condition [b] defining a random experiment by providing the necessary mathematical structure for a coherent assignment of probabilities to events. This is crucial for our purposes because if A and B are events of interest then the related events are also of interest because their occurrence or not is informative for the occurrence of A and B and thus we cannot ignore them when attaching probabilities.

Field. A collection $\Im$ of subsets of S, is said to be a *field* if it satisfies conditions (i)-(iii) in table 2.6.

Table 2.6: Definition of a field

(i)	$S{\in}\Im$,
(ii)	if $A{\in}\Im$ then $\bar{A}$ also belong to $\Im$,
(iii)	if $A, B{\in}\Im$, then $(A \cup B){\in}\Im$.

This means that $\Im$ is non-empty (due to (i)) and it's closed under complementation (due to (ii)), finite unions (due to (iii)) and finite intersections (due to (ii)-(iii)).

Example 2.31. (a) The power set of any finite S_2, $\mathcal{P}(S_2)$, is a field.

(b) $\Im_0{=}\{S, \varnothing\}$ is the trivial field for any outcomes set S.

▶ How does one generate a field of events when S is very large but finite?

Strategy. Define the m events of interest by partitioning S, say $\{A_1, A_2, ..., A_m\}$, i.e. $S{=}A_1 \cup A_2 \cup \cdots \cup A_m$ and $A_i \cap A_j{=}\varnothing$, for $i \neq j$, $i, j{=}1, 2, .., m$, and the set $\{\varnothing, A_1, A_2, ..., A_m\}$ with all possible unions of its elements form a field; no need to worry about intersections or complementations!!!

Example 2.35. In the case of tossing a coin three times with outcomes set S_3 let the events of interest be:

$$A_1{=}\{(HHH), (HHT), (HTT)\}, A_2{=}\{(HTH), (TTT), (TTH)\},\ A_3{=}\{(THT), (THH)\}.$$

The set $\{A_1, A_2, A_3\}$ is clearly a partition of S_3, and thus the relevant field can be generated as:

$$\Im_3^{\ddagger}{=}\{S_3, \varnothing, A_1, A_2, A_3, (A_1 \cup A_2), (A_1 \cup A_3), (A_2 \cup A_3)\}.$$

▶ How does one generate a sigma-field of events when S is infinite?

σ-field. A collection $\Im$ of subsets of S, is said to be a σ-field if it satisfies conditions (i)-(iii) in table 2.7.

Table 2.7: Definition of a sigma-field (σ-field)

(i)	$S{\in}\Im$,
(ii)	if $A{\in}\Im$, then $\overline{A}{\in}\Im$,
(iii)	if $A_i{\in}\Im$ for $i{=}1, 2, ..., n, ...$ the set $\bigcup_{i=1}^{\infty} A_i \in \Im$.

A σ-field $\Im$ is a non-empty set of subsets of S that is closed under countable unions. In addition, De Morgan's law $\overline{\bigcup_{i=1}^{\infty} A_i}{=}\bigcap_{i=1}^{\infty} \overline{A}_i$ implies that: (iv) $\bigcap_{i=1}^{\infty} A_i{\in}\Im$. That is, $\Im$ is also closed under countable intersections.

Borel σ-field. The most important σ-field in probability theory is the one defined on the real line $\mathbb{R}$, known as a Borel σ-field, or *Borel-field* for short, and denoted by $\mathcal{B}(\mathbb{R})$.

Given that the real line $\mathbb{R}$ has an uncountably infinite number of elements the question which naturally arises is:

▶ **how do we define the Borel field $\mathcal{B}(\mathbb{R})$?**

As shown above, the most effective way to define a σ-field over an infinite set is to define it via the elements that can generate this set. It turns out that the half-infinite interval $(-\infty, x]$ is particularly convenient for this purpose. The Borel-field generated by $B_x=\{(-\infty, x]\colon x\in\mathbb{R}\}$, includes all subsets we encounter in practice, including:

$$\{a\},\ (-\infty, a),\ (-\infty, a],\ (a, \infty),\ [a, \infty),\ [a, b],\ (a, b],\ [a, b),\ (a, b),$$

for any real numbers $a<b$, in the sense that $\sigma(B_x)=\mathcal{B}(\mathbb{R})$; see Galambos (1995).

Example 2.37. Consider the following intervals:

(a) $(a, \infty)\ =\overline{(-\infty, a]}\ \Rightarrow\ (a, \infty)\in\mathcal{B}(\mathbb{R})$,

(b) $(a, b]\ =(a, \infty)\cap(-\infty, b]$ for $b>a\ \Rightarrow\ (a, b]\in\mathcal{B}(\mathbb{R})$,

(c) $\{a\}\ =\cap_{n=1}^{\infty}(a-\frac{1}{n}, a]\ \Rightarrow\ \{a\}\in\mathcal{B}(\mathbb{R})$,

(d) $(a, b)\ =\cup_{n=1}^{\infty}(a, b-\frac{1}{n}]\ \Rightarrow\ (a, b)\in\mathcal{B}(\mathbb{R})$,

(e) $[a, b]\ =\cap_{n=1}^{\infty}(a-\frac{1}{n}, b]\ \Rightarrow\ [a, b]\in\mathcal{B}(\mathbb{R})$.

At this stage, it is crucial collect the terminology introduced so far (table 2.8), to bring out the connection between the set theoretic and the probabilistic terms.

Table 2.8. Set-theoretic vs. Probabilistic terminology

Set theoretic	Probabilistic
universal set S	sure event S
empty set $\varnothing$	impossible event $\varnothing$
B is a subset of A: $B\subset A$	when event B occurs event A occurs
set $A\cap B$	events A *and* B occur at the same time
set $A\cup B$	events A *or* B occur
set $\overline{A}$:=$S-A$	event A does not occur
disjoint sets: $A\cap B=\varnothing$	mutually exclusive events $A,\ B$
subset of S	event
element of S	elementary outcome
field	event space
σ-field	event space

The formalization so far. Summarizing in symbols the argument so far below.

$$\mathcal{E}:=([\mathsf{a}],[\mathsf{b}],[\mathsf{c}])\hookrightarrow\Big(\ [\mathsf{a}]\Rightarrow S\ ,\quad [\mathsf{b}]\Rightarrow(\Im,?)\ ,\quad [\mathsf{c}]\Rightarrow ?\ \Big)$$

In the next section we formalize the notion of probability, and proceed to show how we attach probabilities to elements of an event space $\Im$.

5.4 The mathematical notion of probability

The next step in formalizing condition [b] of a random experiment ($\mathcal{E}$), is to assign probabilities to the events of interest as specified by the event space.

5.4.1 Probability set function

The major breakthrough that led to the axiomatization of probability theory in 1933 by Kolmogorov was the realization that $\mathbb{P}(.)$ is a special type of as a *measure* in the newly developed advanced integration theory called *measure theory.* This realization enabled Kolmogorov to develop an axiomatic probability theory:

The idea behind the axiomatization of any field is to specify the fewest *independent* (not derivable from the other) axioms that specify a formal system which is *complete* (every statement that involves probabilities can be shown to true or false within the formal system) and *consistent* (no contradictions stem within the system). The main objective is for the axioms to be used in conjunction with deductive logic to derive theorems that unpack the information contained in the axioms.

$\mathbb{P}(.)$ is defined as a *function* from an event space $\Im$ to the real numbers between 0 and 1 which satisfies certain axioms. That is, the domain of the function $\mathbb{P}(.)$ is a set of subsets of S. To be more precise:

$$\mathbb{P}(.)\colon\ \Im \rightarrow [0,1],$$

is said to be *a probability set function* if it satisfies the axioms in table 2.9. That is, $\mathbb{P}(.)$ assigns probabilities *only* to events that belong to $\Im$; a set of subsets of S that is closed under the set theoretic operations of union, intersection and complementation, i.e. if $A{\in}\Im$ and $B{\in}\Im$, $(A\cup B){\in}\Im$, $(A\cap B){\in}\Im$, $\overline{A}{\in}\Im$, $\overline{B}{\in}\Im$, etc.

Table 2.9: Kolmogorov Axioms of Probability

[A1]	$\mathbb{P}(S)=1$, for any outcomes set S,
[A2]	$\mathbb{P}(A)\geq 0$, for any event $A{\in}\Im$,
[A3]	*Countable Additivity.* For a countable sequence of mutually exclusive events, i.e., $A_i{\in}\Im$, $i{=}1,2,...,n,..$ such that $A_i\cap A_j{=}\varnothing$, for all $i{\neq}j$, $i,j{=}1,2,...,n,...,$ then $\mathbb{P}(\bigcup_{i=1}^{\infty}A_i){=}\sum_{i=1}^{\infty}\mathbb{P}(A_i)$.

In an attempt to understand the role of axiom **[A3],** let us consider the question of assigning probabilities to different events in $\Im$ beginning the simplest to more complicated examples.

(a) Finite outcomes set: $S{=}\{s_1,s_2,...,s_n\}$

In this case one can assign probabilities to the elementary outcomes $s_1,s_2,...,s_n$ without worrying about any inconsistencies or technical difficulties because one can always consider the relevant event space $\Im$ to be $\mathcal{P}(S)$, the set of all subsets of

S. Moreover, since the elementary events $s_1, s_2, ..., s_n$ constitute a *partition* of S (mutually exclusive and $\bigcup_{i=1}^{n} s_i{=}S$), axiom [**A3**] implies that (by axiom [**A1**]):

$$\mathbb{P}(\bigcup_{i=1}^{n} s_i){=}\sum_{i=1}^{n}\mathbb{P}(s_i){=}1,$$

and suggests that by assigning probabilities to the outcomes yields the *simple probability distribution* on S:

$$[p(s_1), p(s_2), ..., p(s_n)], \text{ and } \sum_{i=1}^{n} p(s_i){=}1.$$

The probability of event A in $\Im$ is then defined as follows. First we express event A in terms of the elementary outcomes, say $A{=}\{s_1, s_2, ..., s_k\}$. Then we derive its probability by adding the probabilities of the outcomes $s_1, s_2, ..., s_k$, i.e.

$$\mathbb{P}(A){=}p(s_1)+p(s_2)+...+p(s_k){=}\sum_{i=1}^{k} p(s_i).$$

Example 2.40. (a) Consider the case of the random experiment of "tossing a coin three times," and the event space is the power set:

$$S_3{=}\{(HHH), (HHT), (HTT), (HTH), (TTT), (TTH), (THT), (THH)\}.$$

Let $A_1{=}\{(HHH)\}$ and $A_2{=}\{(TTT)\}$, and derive the probabilities of the events $A_3{:=}(A_1 \cup A_2)$, $A_4{:=}\bar{A}_1$, $A_5{:=}\bar{A}_2$ and $A_6{:=}(\bar{A}_1 \cap \bar{A}_2)$:

$$\mathbb{P}(A_3){=}\mathbb{P}(A_1)+\mathbb{P}(A_2){=}\tfrac{1}{8}{+}\tfrac{1}{8}{=}\tfrac{1}{4},\quad \mathbb{P}(A_4){=}\mathbb{P}(S_3)-\mathbb{P}(A_1){=}1-\tfrac{1}{8}{=}\tfrac{7}{8},$$
$$\mathbb{P}(A_5){=}\mathbb{P}(S_3){-}\mathbb{P}(A_2){=}1{-}\tfrac{1}{8}{=}\tfrac{7}{8},\quad \mathbb{P}(A_6){=}\mathbb{P}(\bar{A}_1{\cap}\bar{A}_2){=}1{-}\mathbb{P}(A_1{\cup}A_2){=}\tfrac{3}{4}.$$

If we go back to the previous section we can see that these are the probabilities we attached using common sense. More often than not, the elementary events $s_1, s_2, ..., s_n$ are equiprobable.

(b) Consider the assignment of probability to the event $A{=}\{(HH), (HT), (TH)\}$, in the case of the random experiment [**ii**] "tossing a fair coin twice". The probability distribution in this case takes the form:

$$\{\mathbb{P}(HH){=}\tfrac{1}{4},\ \mathbb{P}(HT){=}\tfrac{1}{4},\ \mathbb{P}(TH){=}\tfrac{1}{4},\ \mathbb{P}(TT){=}\tfrac{1}{4}\}.$$

This suggests that $\mathbb{P}(A){=}\mathbb{P}(HH)+\mathbb{P}(HT)+\mathbb{P}(TH){=}\tfrac{3}{4}$.

(b) Countable outcomes set: $S{=}\{s_1, s_2, ..., s_n, ...\}$

This case is a simple extension of the finite case where the elementary outcomes $s_1, s_2, ..., s_n, ...$ are again mutually exclusive and they constitute a partition of S, i.e., $\bigcup_{i=1}^{\infty} s_i{=}S$. Axiom [**A3**] implies that:

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} s_i\right)=\sum_{i=1}^{\infty}\mathbb{P}(s_i){=}1.$$

(by axiom [**A1**]) and suggests that by assigning probabilities to the outcomes yields the *probability distribution* on S:

$$[p(s_1), p(s_2), ..., p(s_n), ...], \text{ such that } \sum_{i=1}^{\infty} p(s_i){=}1.$$

As in case (a), the probability of event A in $\Im$ (which might coincide with the power set of S) is defined similarly by:

$$\mathbb{P}(A){=}\textstyle\sum_{[i:s_i\in A]} p(s_i). \tag{4}$$

In contrast to the finite S case, the probabilities $\{p(s_1), p(s_2), ..., p(s_n), ...,\}$ can easily give rise to inconsistencies, such as the case $p(s_n)$, $n{=}1,2,...$, are constant and non-negative. For instance, if we assume that $p(s_n){=}p{>}0$, for all $n{=}1,2,3...$, this gives rise to an inconsistency arises because, however tiny p is,

$$\sum\nolimits_{n=1}^{\infty} p{=}\infty.$$

The only way to render this summation bounded is to make p a decreasing function of n.
For example, assuming $p_n{=}\frac{1}{n^2}$ implies that $(\frac{1}{1.6449})\sum_{n=1}^{\infty} n^{-2}{=}1$,
which is consistent with axioms [**A1**]-[**A3**]; NOTE that for any $k>1$: $\sum_{n=1}^{\infty} n^{-k}{<}\infty$.

Example 2.41. Consider the case of the random experiment of "tossing a coin until the first H appears", where the relevant event space is $\mathcal{P}(S_4)$ (the power set of S_4):

$$S_4{=}\{(H),(TH),(TTH),(TTTH),(TTTTH),...\}.$$

For $\mathbb{P}(H){=}\theta,\ 0{<}\theta{<}1$, and $\mathbb{P}(T){=}1{-}\theta$, the assignment of probabilities takes the form:

$$\mathbb{P}(TH){=}(1{-}\theta)\theta,\ \mathbb{P}(TTH){=}(1{-}\theta)^2\theta,\ \cdots,\ \mathbb{P}(\underbrace{TT...TH}_{n \text{ times}}){=}(1-\theta)^{n-1}\theta, \cdots,$$

where $\sum_{n=1}^{\infty}\theta(1{-}\theta)^{n-1}{=}1$ for any $\theta{\in}(0,1)$.

(c) Uncountable outcomes set S

Without any loss of generality let us consider the case where:

$$S_{[0,1]}{=}\{x\text{: } 0\leq x\leq 1,\ x{\in}\mathbb{R}\}.$$

We can utilize axiom [**A3**] if we can express the interval $[0,1]$ as a countable union of disjoint sets $A_i, i{=}1,2,3,...$ It turns out that with the use of some sophisticated mathematical arguments (axiom of choice, etc.) we can express this interval in the form of:

$$[0,1]{=}\textstyle\bigcup_{i=1}^{\infty} A_i,$$

where $A_i{\cap}A_j{=}\varnothing,\ i{\neq}j,\ i,j{=}1,2,...$, and $\mathbb{P}(A_i)$ is the *same for all* $A_i,\ i{=}1,2,3,...$
This, however, leads to inconsistencies because by axiom [**A3**]:

$$\mathbb{P}([0,1])\ \ {=}\mathbb{P}\left(\textstyle\bigcup_{i=1}^{\infty} A_i\right)\ \ {=}\textstyle\sum_{i=1}^{\infty}\mathbb{P}(A_i),$$

and thus $\mathbb{P}([0,1]){=}0$, if $\mathbb{P}(A_i){=}0$, or $\mathbb{P}([0,1]){=}\infty$, if $\mathbb{P}(A_i)>0$.

The reason why the above attempt failed lies with the nature of the disjoint sets $A_i,\ i{=}1,2,3,...$ They are members of the power set $\mathcal{P}([0,1])$ but not necessarily elements of a σ-field associated with $[0,1]$, needed for a consistent assignment of probabilities.

▶ **How can one circumvent this technical problem**?

In the case where we can define a countable *partition* of S one can generate a σ-field associated with S, and obtain the probability of any event A via (4); see Williams (1991).

Example 2.42. This procedure is best illustrated in the case where the outcomes set is the real line $\mathbb{R}$ and the appropriate σ-field is the Borel field $\mathcal{B}(\mathbb{R})$ which is generated by subsets of the form: $B_x=\{(-\infty, x]:\ x\in\mathbb{R}\}$. We can define $\mathbb{P}(.)$ on B_x first and then proceed to extend it to all subsets (a, ∞), $(a, b]$, $\{a\}$, (a, b), for any real numbers $a < b$, using Caratheodory's extension theorem.

Finally, it is important to emphasize that the combination of axiom [**A3**]-countable additivity and concept of the σ-field for $\Im$ provided the key to Kolmogorov's axiomatization because it ensured the *continuity* of $\mathbb{P}(.)$; see theorem 6 below. Previous attempts to axiomatize probability failed primarily because they could not secure the continuity of $\mathbb{P}(.)$.

5.5 Probability space $(\mathbf{S},\Im, \mathbb{P}(.))$

From the mathematical viewpoint this completes the formalization of conditions [a]−[b] defining a random experiment ($\mathcal{E}$). Condition [a] has become a set S called an outcomes set (with elements the elementary outcomes) and condition [b] has taken the form of $(\Im, \mathbb{P}(.))$ where $\Im$ is a σ-field of subsets of S called an event space and $\mathbb{P}(.)$ is a probability set function which satisfies axioms [**A1**]-[**A3**] (table 2.9).

From a mathematical perspective the next step is to use the above mathematical set up, in conjunction with mathematical logic, to derive a number of conclusions making up probability theory. The approach adopted in this book takes a different route by emphasizing probability theory as providing the foundation of empirical modeling. It is instructive, however, to get a taste of what the mathematical approach entails before we proceed with the modeling perspective.

6 Conditional probability and Independence

6.1 Conditional probability and its properties

As a prelude to formalizing condition [c] of a Random Experiment $\mathcal{E}$, we need to take a digression to discuss a very important notion in probability theory, that of *conditioning.* This notion arises naturally when one has certain additional information relating to the experiment in question that might affect the relevant probabilities.

Example 2.46. In the case of tossing a coin twice, if we (somehow) know that the actual outcome has at least one T, this information will affect the probabilities of certain events. For instance, the outcome (HH) now has zero probability, and thus the outcomes $(HT), (TH)$ and (TT) have probabilities equal to $\frac{1}{3}$, not $\frac{1}{4}$ as before. Let us formalize this argument in a more systematic fashion by defining the event B "at least one T": $B=\{(HT), (TH), (TT)\}$.

Without knowing B the outcomes set and the probability distribution are:

$$\begin{aligned} S_2= & \ \{(HH),(HT),(TH),(TT)\}, \\ \mathbf{P}= & \ \left\{\mathbb{P}(HH)=\tfrac{1}{4}, \mathbb{P}(HT)=\tfrac{1}{4}, \mathbb{P}(TH)=\tfrac{1}{4}, \mathbb{P}(TT)=\tfrac{1}{4}\right\}. \end{aligned}$$

With the knowledge provided by B these become:

$$S_B=\{(HT),(TH),(TT)\}, \ \ \mathbf{P}_B=\left\{P_B(HT)=\tfrac{1}{3}, \ P_B(TH)=\tfrac{1}{3}, \ P_B(TT)=\tfrac{1}{3}\right\}.$$

In a sense the event B has become the new outcomes set and the probabilities are now conditional on B in the sense that:

$$P_B(HT)=\mathbb{P}((HT)|B)=\tfrac{1}{3}, \ \ P_B(TH)=\mathbb{P}((TH)|B)=\tfrac{1}{3}, \ \ P_B(TT)=\mathbb{P}((TT)|B)=\tfrac{1}{3}.$$

A general way to derive these conditional probabilities is the conditional rule:

$$\boxed{\mathbb{P}(A|B)=\tfrac{\mathbb{P}(A\cap B)}{\mathbb{P}(B)}, \text{ for } \mathbb{P}(B)>0,} \tag{5}$$

for any event $A\in\Im$, where $\mathbb{P}(.)$ is the original probability set function defined on $\Im$.

Example 2.47. For $A=\{(TH)\}$ and $A\cap B=\{(TH)\}$ (5) implies: $\mathbb{P}(A|B)=\frac{(1/4)}{(3/4)}=\frac{1}{3}$.

Example 2.48. Boy-girl problem. Consider a family with two children. We learn that one of the two is a girl (G), what is the probability that the other is a boy (B)? First, we need to avoid the crucial mistake made by Leibniz (section 2.2) by listing all possible distinct outcomes:

$$S=\{(BG), \ (GB), \ (BB), \ (GG)\},$$

assumed to be equally likely. The obvious answer that the probability of the event of interest, $A_2=\{(BG), \ (GB)\}$ is $\frac{1}{2}$ is clearly wrong, because it ignores the information available that the event 'at least one of the two is a girl', i.e. $A_1=\{(BG), \ (GB), \ (GG)\}$ has occurred. The proper way to account for that is to evaluate the conditional probability:

$$\mathbb{P}(A_2|A_1)=\frac{\mathbb{P}(A_1\cap A_2)}{\mathbb{P}(A_1)}=\frac{(1/2)}{(3/4)}=\frac{2}{3}.$$

Properties. The conditional probability in (5) enjoys a number of properties.

(CP1) The product rule for conditional probability. Using the conditional probability formula (5) we can deduce the *product rule*:

$$\mathbb{P}(A\cap B)=\mathbb{P}(A|B)\cdot\mathbb{P}(B)=\mathbb{P}(B|A)\cdot\mathbb{P}(A). \tag{6}$$

This formula can be easily extended to an ordered sequence of three events $\{A_1, A_2, A_2\}$:

$$\mathbb{P}(A_1\cap A_2\cap A_3)=\mathbb{P}(A_3|A_2, A_1)\cdot\mathbb{P}(A_2|A_1)\cdot\mathbb{P}(A_1). \tag{7}$$

(CP2). The total probability rule. This results from combining theorem 4 [For events $A\in\Im$, $B\in\Im$: $\mathbb{P}(B)=\mathbb{P}(A\cap B)+\mathbb{P}(\overline{A}\cap B)$] and (18):

$$\boxed{\mathbb{P}(B)=\mathbb{P}(A)\cdot\mathbb{P}(B|A)+\mathbb{P}(\overline{A})\cdot\mathbb{P}(B|\overline{A}).} \tag{8}$$

This formula can be extended to a finite partition $\{A_1, A_2, \ldots, A_n\}$ of S:

$$\mathbb{P}(B)=\textstyle\sum_{i=1}^{n}\mathbb{P}(A_i)\cdot\mathbb{P}(B|A_i). \tag{9}$$

(CP3). Bayes rule. Combining (5), (18) and (9), we derive *Bayes' rule*:

$$\boxed{\mathbb{P}(A_i|B)=\tfrac{\mathbb{P}(A_i)\cdot\mathbb{P}(B|A_i)}{\sum_{i=1}^{n}\mathbb{P}(A_i)\cdot\mathbb{P}(B|A_i)},\text{ for }\mathbb{P}(B)>0.} \tag{10}$$

Terminology. It is important to bring out the fact that the attribution of formula in (10) to Bayes (1764) is a classic example of Stigler's (1980) "Law of Eponymy" stating that no scientific discovery is named after its original discoverer. The formula in (5) pertains to conditional probability between events which was used in the early 16th century by Cardano, and both formulae (5) and (18) are clearly stated in de Moivre (1718/1738); see Hald (1998). Moreover, the claim that (10) provides the foundation of Bayesian statistics is also misleading because in Bayesian inference A_i, i=1,...,n, are not *observable events,* as in the above context, but unobservable parameters $\boldsymbol{\theta}:=(\theta_1, \theta_2, \ldots, \theta_n)$; see chapter 10.

Example 2.50. False positive/negative. Consider the case of a medical test to detect a particular disease. It is well-known that such tests are almost never 100% accurate. Let us assume that for this particular test it has been established that:

(a) If a patient has the disease, the test will likely detect it (give a positive result) with .95 probability, i.e. its *false negative* probability is .05.

(b) If a patient does *not* have the disease, the test will likely incorrectly give a positive result with .1 probability (*false positive*).

Let us also assume that a person randomly selected from the relevant population will have the disease with probability of .03. The question of interest is: ▶ when a person from that population tests positive, what is the probability that he/she actually has the disease? To answer that question we need to define the events of interest in terms of the two primary events, a patient:

$$A \text{ - has the disease, } B\text{- tests positive.}$$

(a)-(b) suggest that the relevant probabilities are:

$$\mathbb{P}(A)=.03,\ \mathbb{P}(B|A)=.95,\ \mathbb{P}(B|\overline{A})=.1.$$

Applying Bayes' formula (10) yields:

$$\mathbb{P}(A|B)=\tfrac{\mathbb{P}(A)\cdot\mathbb{P}(B|A)}{\mathbb{P}(A)\cdot\mathbb{P}(B|A)+\mathbb{P}(\overline{A})\cdot\mathbb{P}(B|\overline{A})}=\tfrac{(.03)(.95)}{(.03)(.95)+(.97)(.1)}=.227$$

At first sight this probability might appear rather small since the test has a 95% accuracy, but that ignores the fact that the incidence of the disease in this particular population is small, 3%.

CAUTION: the above example is a purely probabilistic example (deductive in nature) and has nothing to do with data statistical models or any form of inference. Any allusions to populations and sampling are coincidental and have connection to real data and statistical inference! This is very important because there is a major confusion between **false positive/negative probabilities** and **legitimate type I and II error probabilities**; the relationship is more apparent than real!

Example 2.51. Monty Hall Puzzle

A contestant on a TV game 'Let's Make a Deal', is presented with three doors numbered $\boxed{\mathbf{1}}$, $\boxed{\mathbf{2}}$, $\boxed{\mathbf{3}}$. One of doors has a *car* behind it and the other two have goats. The contestant will be asked to choose one of the doors and then the game host will open one of the other doors and give the contestant a chance to switch.

Stage 1. The contestant is asked to pick one door, and he chooses door $\boxed{\mathbf{1}}$**.**

Stage 2. The game host opens door $\boxed{\mathbf{3}}$ to reveal a goat.

Stage 3. The game host asks the contestant: do you want to switch to door $\boxed{\mathbf{2}}$?

▶ Will switching to door **2** improve the contestant's chances of winning the car?

A professor of mathematics claimed in print that the answer is a definite No!:

"If one door is shown to be a loser, that information changes the probability of either remaining choice, neither of which has any reason to be more likely, to 1/2."
It turns out that this claim is wrong! Why? The probability 1/2 for doors $\boxed{\mathbf{1}}$ and $\boxed{\mathbf{2}}$ hiding the car ignores one important but subtle piece of information: the game host opened door 3 knowing where the car is! This information is relevant for evaluating the pertinent probabilities, not only of the primary event of interest, which is C_k-door k hides the car, but also of the related event, D_k-the host opened door k.

Probability theory can frame the relationship between these events in terms of their joint, marginal and conditional probabilities. Initially, the car could have been behind any one of the doors, and thus the *marginal* probabilities for events C_k are:

$$\mathbb{P}(C_1)=\mathbb{P}(C_2)=\mathbb{P}(C_3)=\tfrac{1}{3}.$$

After the contestant selected door $\boxed{\mathbf{1}}$, the game host has only two doors to choose from, and thus the *marginal* probabilities for events D_k are:

$$\mathbb{P}(D_1)=0,\ \mathbb{P}(D_2)=\mathbb{P}(D_3)=\tfrac{1}{2}.$$

The professor of mathematics was wrong because he attempted to account for the occurrence of the *event* D_3: the game host opened door $\boxed{\mathbf{3}}$, by erroneously changing the original marginal probabilities from:

$$\mathbb{P}(C_1){=}\mathbb{P}(C_2){=}\mathbb{P}(C_3){=}\tfrac{1}{3} \text{ to } \mathbb{P}(C_1){=}\mathbb{P}(C_2){=}\tfrac{1}{2}.$$

Probabilistic reasoning teaches us that the proper way to take into account the information that event D_3 *has occurred* is to condition on it. Using the conditional probability formula (5), one can elicit the probabilities the contestant needs:

$$\mathbb{P}(C_1|D_3){=}\tfrac{\mathbb{P}(C_1\cap D_3)}{\mathbb{P}(D_3)} \text{ and } \mathbb{P}(C_2|D_3){=}\tfrac{\mathbb{P}(C_2\cap D_3)}{\mathbb{P}(D_3)}. \tag{11}$$

To evaluate (11), however, one requires the probabilities: $\mathbb{P}(C_1 \cap D_3)$ and $\mathbb{P}(C_2 \cap D_3)$. The joint probability rule in (18) suggests that one can evaluate these joint probabilities via:

$$\mathbb{P}(C_1{\cap}D_3){=}\mathbb{P}(D_3|C_1){\cdot}\mathbb{P}(C_1),\ \ \mathbb{P}(C_2{\cap}D_3){=}\mathbb{P}(D_3|C_2){\cdot}\mathbb{P}(C_2).$$

But how can one retrieve $\mathbb{P}(D_3|C_1)$ and $\mathbb{P}(D_3|C_2)$?

The *game host's reasoning,* based on his information, that led him to open door 3, can be used to elicit these conditional probabilities. Pondering on the game host's reasoning in opening door 3: If the car is behind door 1, the game host is free to pick between doors $\boxed{\mathbf{2}}$ and $\boxed{\mathbf{3}}$ at random, hence: $\mathbb{P}(D_3|C_1){=}\frac{1}{2}$.
If the car is behind door $\boxed{\mathbf{2}}$, he has no option but open door $\boxed{\mathbf{3}}$, hence: $\mathbb{P}(D_3|C_2){=}1$.
The car could not have been behind door $\boxed{\mathbf{3}}$, and thus: $\mathbb{P}(D_3|C_3){=}0$.

The coherence of these conditional probabilities is confirmed by the total probability rule in (8):

$$\mathbb{P}(D_3){=}\textstyle\sum_{k=1}^{3}\mathbb{P}(D_3|C_k){\cdot}\mathbb{P}(C_k){=}\tfrac{1}{2}(\tfrac{1}{3}){+}1(\tfrac{1}{3}){+}0(\tfrac{1}{3}){=}\tfrac{1}{2}.$$

Collecting all the relevant probabilities derived above:

$$\mathbb{P}(C_1){=}\mathbb{P}(C_2){=}\mathbb{P}(C_3){=}\tfrac{1}{3},\ \ \mathbb{P}(D_3){=}\mathbb{P}(D_2){=}\tfrac{1}{2}. \qquad \mathbb{P}(D_3|C_1){=}\tfrac{1}{2},\ \ \mathbb{P}(D_3|C_2){=}1$$

and evaluating the relevant conditional probabilities yields:

$$\mathbb{P}(C_1|D_3){=}\tfrac{\mathbb{P}(D_3|C_1)\cdot\mathbb{P}(C_1)}{\mathbb{P}(D_3)}{=}\tfrac{\left(\frac{1}{2}\right)\left(\frac{1}{3}\right)}{\left(\frac{1}{2}\right)}{=}\tfrac{1}{3} < \mathbb{P}(C_2|D_3){=}\tfrac{\mathbb{P}(D_3|C_2)\cdot\mathbb{P}(C_2)}{\mathbb{P}(D_3)}{=}\tfrac{(1)\left(\frac{1}{3}\right)}{\left(\frac{1}{2}\right)}{=}\tfrac{2}{3}.$$

This shows that switching doors doubles the chances of winning the car from $\frac{1}{3}$ to $\frac{2}{3}$. The moral of this true story is that being a professor of mathematics does not necessarily mean that you can reason systematically with probabilities. That takes more than just sound common sense and good mathematical background! It requires mastering the mathematical structure of probability theory and its rules of reasoning.

6.2 The concept of independence among events

The notion of conditioning can be used to determine whether two events A and B are related in the sense that information about the occurrence of one, say B, alters the probability of occurrence of A. Recall that:

$$\boxed{\mathbb{P}(A|B)=\tfrac{\mathbb{P}(A\cap B)}{\mathbb{P}(B)},\text{ for }\mathbb{P}(B)>0,} \tag{12}$$

If knowledge of the occurrence of B does not alter the probability of event A, it is natural to say that A and B are independent.

More formally A and B are *independent* if:

$$\mathbb{P}(A|B)=\mathbb{P}(A) \Leftrightarrow \mathbb{P}(B|A)=\mathbb{P}(B) \tag{13}$$

Using the conditional probability formula (12), we can deduce that two events A and B are *independent* if:

$$\mathbb{P}(A\cap B)=\mathbb{P}(A)\cdot\mathbb{P}(B). \tag{14}$$

NOTE that this notion of independence can be traced back to Cardano in the 1550s.

Example 2.52. For $A=\{(HH),(TT)\}$ and $B=\{(TT),(HT)\}$, $A\cap B=\{(TT)\}$ and thus:

$$\mathbb{P}(A\cap B)=\tfrac{1}{4}=\mathbb{P}(A)\cdot\mathbb{P}(B),$$

implying that A and B are independent.

It is very important to distinguish between *independent* and *mutually exclusive* events; the definition of the latter does not involve probability. Indeed, two independent events with positive probability cannot be mutually exclusive. This is because if $\mathbb{P}(A)>0$ and $\mathbb{P}(B)>0$ and they are independent then $\mathbb{P}(A\cap B)=\mathbb{P}(A)\cdot\mathbb{P}(B)>0$, but mutual exclusiveness implies that $\mathbb{P}(A\cap B)=0$ since $A\cap B=\varnothing$. The intuition behind this result is that mutually exclusive events are informative about each other because the occurrence of one precludes the occurrence of the other.

Example 2.53. For $A=\{(HH),(TT)\}$ and $B=\{(HT),(TH)\}$, $A\cap B=\varnothing$ but:

$$\mathbb{P}(A\cap B)=0\neq\tfrac{1}{4}=\mathbb{P}(A)\cdot\mathbb{P}(B).$$

Joint independence. Independence can be generalized to more than two events but in the latter case we need to distinguish between pair wise, joint and mutual independence. For example in the case of three events A, B and C we say that they are *jointly independent* if:

$$\mathbb{P}(A\cap B\cap C)=\mathbb{P}(A)\cdot\mathbb{P}(B)\cdot\mathbb{P}(C). \tag{15}$$

Pairwise independence. The notion of joint independence, however, is not equivalent to *pairwise independence* defined by the conditions:

$$\mathbb{P}(A\cap B)=\mathbb{P}(A)\cdot\mathbb{P}(B),\quad \mathbb{P}(A\cap C)=\mathbb{P}(A)\cdot\mathbb{P}(C),\quad \mathbb{P}(B\cap C)=\mathbb{P}(B)\cdot\mathbb{P}(C).$$

Example 2.54. Consider the outcomes set $S{=}\{(HH),(HT),(TH),(TT)\}$ and the events:

$$A{=}\{(TT),(TH)\},\ B{=}\{(TT),(HT)\},\text{ and }C{=}\{(TH),(HT)\}.$$

Given that $A\cap B{=}\{(TT)\}$, $A\cap C{=}\{(TH)\}$, $B\cap C{=}\{(HT)\}$ and $A\cap B\cap C{=}\varnothing$ we can deduce:

$$\begin{aligned}&\mathbb{P}(A\cap B){=}\mathbb{P}(A)\cdot\mathbb{P}(B){=}\tfrac{1}{4}, && \mathbb{P}(B\cap C){=}\mathbb{P}(B)\cdot\mathbb{P}(C){=}\tfrac{1}{4},\\ &\mathbb{P}(A\cap C){=}\mathbb{P}(A)\cdot\mathbb{P}(C){=}\tfrac{1}{4},\text{ but} && \mathbb{P}(A\cap B\cap C){=}0{\neq}\mathbb{P}(A)\cdot\mathbb{P}(B)\cdot\mathbb{P}(C){=}\tfrac{1}{8}.\end{aligned}$$

Similarly, joint independence does not imply pairwise independence. Moreover, both of these forms of independence are weaker than independence which involves joint independence for all sub-collections of the events in question.

Independence. The events $A_1, A_2, ..., A_n$ are said to be *independent* iff:

$$\mathbb{P}(A_1\cap A_2\cap\cdots\cap A_k){=}\mathbb{P}(A_1)\cdot\mathbb{P}(A_2)\cdot\cdots\cdot\mathbb{P}(A_k),\ \text{ for each } k{=}2,3,...,n.$$

That is, this holds for *any sub-collection* $A_1, A_2, ..., A_k$ $(k\leq n)$ of $A_1, A_2, ..., A_n$.

In the case of three events A, B and C pairwise and joint independence together imply independence and conversely.

7 A close look at Bayes' rule

Let us take a closer lool at Bayes's rule in the context of the probabilty space $(S, \Im, \mathbb{P}(.))$ knowing the mathematical structure of $\Im$ and the distinct feature that makes an event an element of $\Im$.

7.1 Bayes' rule in terms of specific 'events'

The conditional probability formula in (5) is transformed into an *updating rule* by interpreting the two events A and B as a *hypothesis* H (e.g. $\theta = .5$) and *evidence* E, (e.g. data $\mathbf{x}_0$), respectively, to yield:

$$P(H|E) {=} \frac{P(E|H) {\cdot} P(H)}{P(E)}, \ \mathbb{P}(E) > 0, \tag{16}$$

attributed to Bayes' (1763). Its components are interpreted as follows:

(i) $P(H|E)$ is the *posterior probability* of H given E,

(ii) $P(E|H)$ is the *likelihood* of E given H,

(iii) $P(H)$ is the *prior probability* of H, and

(iv) $P(E)$ is the initial *probability of evidence* E.

(1) For Bayes' rule in (16) to represent an instantiation of the conditional probability formula:

$$\mathbb{P}(A|B) {=} \frac{\mathbb{P}(B|A) {\cdot} \mathbb{P}(A)}{\mathbb{P}(B)}, \text{ for } \mathbb{P}(B) > 0. \tag{17}$$

the events H and E are required to be:

(a) defined on the same event space $\Im$,

(b) potentially *observable,* and

(c) related in the sense that the events $H \cap E$, $H \cup E$, H^c, E^c belong to $\Im$, where H^c denotes the complement of H with respect to S).

Conditions (a)-(c) are potentially problematic for Bayes' rule since E is, in principle, observable and lies in the real world, but H is usually *unobservable* and belongs to the world of mathematics. Ignoring the gap between these two worlds by assuming they are interrelated via $\Im$ in the above simplistic way raises fundational issues in empirical modeling.

(2) It is not obvious how the likelihood function $\mathbb{P}(E|H)$ assigns a probability to E by conditioning on an *unobservable* event H. How does one 'condition on the occurrence of an unobservable event H' without running into an oxymoron? A generous possible interpretation might be that the conditioning is only *notional* in the sense that the hypothesis H relates to a particular instance of the mechanism that gave rise to E. A generous interpretation might be that $\mathbb{P}(E|H)$ refers to the 'objective probability of the occurrence of E presuming that H is true'. In practice, however, 'presuming that H is true' could not represent the occurrence of an unobservable event as such.

(3) Despite the bold move of merging hypotheses (mathematical world) and evidence (real world) into the same S, when it comes to acknowledging this overlap

$(E{\cap}H)$, Bayesians sidestep the issue by using the identity $\mathbb{P}(E{\cap}H){=}\mathbb{P}(E|H){\cdot}\mathbb{P}(H)$ in:

$$\mathbb{P}(H|E) = \tfrac{\mathbb{P}(E\cap H)}{\mathbb{P}(E)},\ \mathbb{P}(E) > 0. \tag{18}$$

It is presumed that the assignments $\mathbb{P}(E|H){\cdot}\mathbb{P}(H)$ or $\mathbb{P}(H|E){\cdot}\mathbb{P}(E)$ are easier to justify!

(4) The most problematic of the probabilistic assignments (i)-(iv) is the assingment of probability to evidence E, $\mathbb{P}(E)$, because it's not obvious where the probability could come from; see Earman (1992), p. 172. The Bayesian attempt to address this conundrum by defining (iv) in terms of (ii)-(iii) which they deem less questionable. In particular, they use H and $\overline{H}$ denoted by ($\overline{H}$, the complement of H with respect to S), to define a *partition* of S: $S{=}H \cup \overline{H}$, and then use:

$$\mathbb{P}(H \cup \overline{H}){=}\mathbb{P}(H){+}\mathbb{P}(\overline{H}){=}\mathbb{P}(S){=}1,$$

to deduce the *total probability formula*:

$$\mathbb{P}(E){=}\mathbb{P}(H){\cdot}\mathbb{P}(E|H) + \mathbb{P}(\overline{H}){\cdot}\mathbb{P}(E|\overline{H}). \tag{19}$$

The formula in (19) is used to rewrite *Bayes' rule* as:

$$\mathbb{P}(H|E){=}\frac{\mathbb{P}(E|H){\cdot}\mathbb{P}(H)}{\mathbb{P}(H){\cdot}\mathbb{P}(E|H){+}\mathbb{P}(\overline{H}){\cdot}\mathbb{P}(E|\overline{H})},\ \mathbb{P}(E) > 0. \tag{20}$$

The so-called *Bayesian catchall factor* $\mathbb{P}(E|\overline{H})$ in (20) has been severely criticized by Mayo (1996), pp. 116-118, as highly misleading in practice.

7.2 Bayesian confirmation: a very brief introduction

The Bayesian confirmation theory relies on comparing the prior with the posterior probability of a particular hypothesis H:

$$\begin{array}{ll}\text{[i] Confirmation:} & P(H|E) > P(H)\\ \text{[ii] Disconfirmation:} & P(H|E) < P(H)\end{array}$$

The *degree of confirmation* is evaluated using some measure $\mathfrak{c}(H,E)$ of the 'degree to which E raises the probability of H'. Examples of such Bayesian measures are (Fitelson, 1999):

$$d(H,E){=}P(H|E){-}P(H),$$
$$m(H,E){=}P(E|H){-}P(E),$$
$$r(H,E){=}\tfrac{P(H|E)}{P(H)}.$$

Using $\mathfrak{c}(H,E)$ one can define *relative evidence* as:

The measure $\mathfrak{c}(H,E)$ indicates that evidence E favors hypothesis H_1 over H_0, iff:

$$\mathfrak{c}(H_1,E) > \mathfrak{c}(H_0,E).$$

For instance using the measure $r(H, E)$ in the case of two competing hypotheses H_0 and H_1 :

$$\frac{P(H_1|E)}{P(H_1)} > \frac{P(H_0|E)}{P(H_0)} \overset{\text{Bayes}}{\Leftrightarrow} \frac{P(E|H_1)}{P(E)} > \frac{P(E|H_0)}{P(E)} \Leftrightarrow \frac{P(E|H_1)}{P(E|H_0)} > 1$$

where $\frac{P(E|H_1)}{P(E|H_0)}$ is the (Bayesian) likelihood ratio.

For comparison purposes let us contrast this to the *ratio of posteriors*:

$$\frac{P(H_1|E)}{P(H_0|E)} = \frac{\frac{P(E|H_1)\cdot\mathbb{P}(H_1)}{P(E)}}{\frac{P(E|H_0)\cdot\mathbb{P}(H_0)}{P(E)}} = \left(\frac{P(E|H_1)}{P(E|H_0)}\right)\left(\frac{P(H_1)}{P(H_0)}\right) > 1, \tag{21}$$

which is the product of the ‘likelihood ratio’ $\frac{P(E|H_1)}{P(E|H_0)}$ *and* the ratio of the prior probabilities $\frac{P(H_1)}{P(H_0)}$.

A critic can make a strong argument that in light of the fact that the above Bayesian confirmation theory inherits all the weaknesses (1)-(4) of Bayes’ rule, the confirmation endeavor feels like **playing war games on an imaginary map** without any actual connection to the reality one is aiming to understand .

8 Formalizing condition [c]: sampling space

8.1 The concept of random trials

The last condition defining the notion of a *random experiment* is:

[c] The experiment can be repeated under identical conditions.

This is interpreted to mean that the circumstances and conditions from one trial to the next remain the same. This entails two interrelated but different components:

(i) the *set up* of the experiment remains the same for all trials and

(ii) the *outcome* in one trial does *not* modify the probability of the outcome in another trial.

▶ **How do we formalize these conditions?**

The first notion we need to formalize is that of a finite sequence of trials. Let us denote the n trials by $\{\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_3, ..., \mathcal{A}_n\}$ and associate each trial with a probability space $(S_i, \Im_i, \mathbb{P}_i(.))$, $i{=}1, 2, \ldots, n$, respectively. In order to be able to discuss any relationship between trials we need to encompass them in an overall probability space; without it we cannot formalize condition (ii) above.

The first component of condition [c] can be easily formalized by ensuring that the probability space $(S, \Im, \mathbb{P}(.))$ remains the same from trial to trial in the sense:

$$\text{[i] } (S_i, \Im_i, \mathbb{P}_i(.)){=}(S, \Im, \mathbb{P}(.)), \text{ for all } i{=}1, 2, \ldots, n, \tag{22}$$

and we refer to this as the *Identical Distribution* (ID) condition.

The second component [c]-(ii) is formalized by postulating that the trials are *independent*:

$$\begin{array}{ll} \text{[ii]} & \mathbb{P}_{(n)}(\mathcal{A}_1 \cap \mathcal{A}_2 \cap \cdots \cap \mathcal{A}_k){=}\mathbb{P}_1(\mathcal{A}_1)\cdot\mathbb{P}_2(\mathcal{A}_2)\cdot\cdots\cdot\mathbb{P}_k(\mathcal{A}_k), \text{ for } k{=}2, 3, ..., n, \\ \text{or} & \\ \text{[ii]*} & \mathbb{P}_{(n)}(\mathcal{A}_k|\mathcal{A}_1, \mathcal{A}_2, ...\mathcal{A}_{k-1}, \mathcal{A}_{k+1}, ..., \mathcal{A}_n){=}\mathbb{P}_k(\mathcal{A}_k), \text{ for } k{=}1, 2, ..., n. \end{array} \tag{23}$$

Taking the conditions of *Independence* (23) and *Identical Distribution* (22) we define what we call *a sequence of Random trials.*

Random trials. A sequence of trials $\mathcal{G}_n^{\text{IID}}{:=}\{\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_3, ..., \mathcal{A}_n\}$, which is both *independent* and *identically distributed*:

$$\boxed{\begin{array}{l} \text{[i] } (S_i, \Im_i, \mathbb{P}_i(.)){=}(S, \Im, \mathbb{P}(.)), \text{ for all } i{=}1, 2, \ldots, n, \\ \\ \text{[ii] } \mathbb{P}_{(n)}(\mathcal{A}_1 \cap \mathcal{A}_2 \cap ... \cap \mathcal{A}_k){=}\mathbb{P}(\mathcal{A}_1) \cdot \mathbb{P}(\mathcal{A}_2) \cdots \mathbb{P}(\mathcal{A}_k), \text{ for } k{=}2, 3, ..., n. \end{array}}$$

The story so far in symbols

$$\mathcal{E}{:=}\left[\begin{array}{c} \text{[a]} \\ \text{[b]} \\ \text{[c]} \end{array}\right] \begin{array}{c} \Rightarrow \\ \Rightarrow \\ \Rightarrow \end{array} \left(\begin{array}{l} S \\ (\Im, \mathbb{P}(.)) \\ \mathcal{G}_n \end{array}\right) \Longrightarrow \left[(S, \Im, \mathbb{P}(.))^n, \mathcal{G}_n^{\text{IID}}\right], \ \mathcal{G}_n^{\text{IID}}{=}\{\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_3, ..., \mathcal{A}_n\}$$

The purpose of this chapter has been to provide an introduction to probability theory using the formalization of a simple chance mechanism we called a random experiment ($\mathcal{E}$) defined by conditions [a]-[c].

The primary objective of this formalization is to motivate some of the most important concepts of probability theory and define them in a precise mathematical way in the form of a statistical space.

The questions addressed along the way include the following:

▶ Why these particular primitive notions $(S, \Im, \mathbb{P}(.))$?

The probability space $(S, \Im, \ \mathbb{P}(.))$ provides an idealized mathematical description of the stochastic mechanism that gives rise to the events in $\Im$.

▶ Why is the set of events of interest $\Im$ a sigma-field?

Mathematically $\Im$ has the structure of a $\sigma-$field, because of the nature of the basic concept of probability we call an *event*: a subset of S, which is an element of $\Im$, that might or might not occur at any particular trial. If A and B are events so are $A \cup B, \ A \cap B, \ \overline{A}, \ \overline{B}$, etc. The notion of an *event* (an element of $\Im$) in probability plays an analogous role to the notion of a *point* in geometry. $\Im$ is a set of subsets of S that is closed under the set theoretic operations $\cup$,$\cap$,$^{-}$.

▶ Why choose the particular axioms **[A1]-[A3]** in table 2.9?

This formalization places probability squarely into the mathematical field of *measure theory* concerned more broadly with assigning size, length, content, area, volume, etc. to sets; see Billinglsley (1995). The axioms [**A1**]-[**A3**] ensure that $\mathbb{P}(.)$ assigns probabilities to events in $\Im$ in a consistent and coherent way.

▶ What is the scope of the IID trials in $\mathcal{G}_n^{\text{IID}}=\{\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_3, ..., \mathcal{A}_n\}$?

The notion of a set of IID trials in $\mathcal{G}_n^{\text{IID}}$ formalizes two vague notions often invoked in descriptive statistics: (a) the 'uniformity' of the target *population (or nature)* and (b) the 'representativeness' of the *sample.*

9 In lieu of a summary and conclusions

1. Introduction: descriptive statistics vs. statistical inference proper
2. A simple statistical model: a preliminary view
3. Probability theory as a modeling framework: a probability space $(S, \Im,\ \mathbb{P}(.))$
4. Formalizing the notion of a Random Experiment: conditions [a]-[c]
5. Conditional probability and independence among events
6. A closer look at Bayes' rule

Important concepts

Random experiment, outcomes set (sample space), elementary outcomes, events, sure event, impossible event, set theoretic union, intersection, complementation, partition of a set, empty set, finite set, infinite set, countable set, uncountable set, Venn diagrams, de Morgan's law, mutually exclusive events, event space, power set, field of events, sigma field of events, Borel-field, function, domain and co-domain of a function, range of a function, probability set function, countable additivity, probability space, mathematical deduction, conditional probability, total probability rule, Bayes rule, independent events, pairwise independent events, sampling space, Independent trials, Identically Distributed trials, statistical space.

Crucial distinctions

Descriptive vs. inferential statistics, elementary outcomes vs. events, countable vs. uncountable sets, power set vs. a sigma-field, independent events vs. mutual exclusive events, co-domain vs. range of a function, probabilistic vs. set theoretic terminology, independence vs. joint independence vs. pairwise independence among events.

Essential ideas

- There is no such thing as 'descriptive measures for particular data' that do not invoke probabilistic assumptions.
- Probability theory as the foundation and overarching framework for empirical modeling is crucial for defining the premises of statistical induction as well as calibrating the capacity of the inference procedures stemming from this premises.
- In statistics one aims to model the stochastic mechanism that gave rise to the data, and not to summarize the particular data. Indeed, the inference pertains to this mechanism, even though it is framed in terms of the parameters of the model.
- The most effective way to transform an uncountable set in probability theory into a countable one is to use partitioning.
- The concept of a σ-field played a crucial role in Kolmogorov's framing of the axiomatic approach to probability theory because it captures the key features of the concept of an event for all outcomes sets, including uncountable ones.

- The axiomatization of probability revolves around the concept of an 'event' and its occurrence.
- The concept of a σ-field provides the key to understanding the concept of conditioning in its various forms, e.g. $E(Y|\mathbf{X}=\mathbf{x})$ vs. $E(Y|\sigma(\mathbf{X}))$. Kolmogorov (1933) was the first to properly formalize conditional probability using the concept of a σ-field.