# Probability Theory and Statistical Inference:
## Empirical Modeling with Observational Data

---

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

# Chapter 2: Probability Theory as a Modeling Framework

## 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 before the long journey to explore the theory of probability is: ► **Why do we need probability theory**?

The brief answer is that it frames both the foundation and 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.

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).

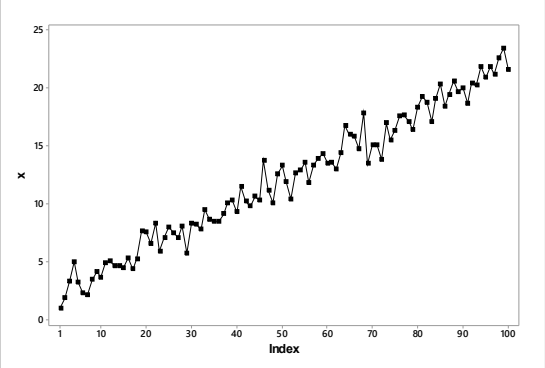


Fig. 2.1: t-plot of data $\mathbf{x}_0$

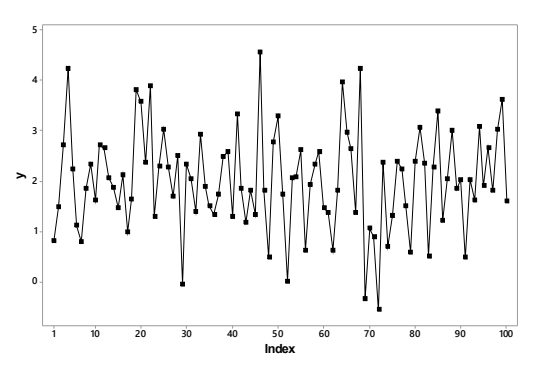


Fig. 2.2: Typical realization of NIID data

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.

**A bird's eye view of the chapter**. In section 2 we introduce the notion of a simple statistical model at an informal and intuitive level. Section 3 introduces the reader to probability theory from the viewpoint of statistical modeling. In section 4 we sidestep the axiomatic approach to probability in an attempt to motivate the required mathematical concepts by formalizing a simple generic stochastic phenomenon we call a *Random Experiment (RE)* defined by three conditions in plain English. In sections 5-6 we proceed to formalize the first two of these conditions in the form of (i) the outcomes set, (ii) the event space and (iii) the probability set function, together with the associated Kolmogorov axioms. Section 7 discusses the notion of conditioning needed to formalize the third condition in section 8.

# 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:

$$\text{[i] Probability model:} \quad \Phi=\{f(x;\boldsymbol{\theta}),\ \boldsymbol{\theta}\in\Theta,\ x\in\mathbb{R}_X\},$$

$$\text{[ii] Sampling model:} \quad \mathbf{X}:=(X_1,X_2,...,X_n) \text{ is a random sample.}$$

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* $\theta$, as the latter varies over its range of values $\Theta$: *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 $\Gamma[\alpha]=\int_0^\infty \exp(-u)\cdot u^{\alpha-1}du$.

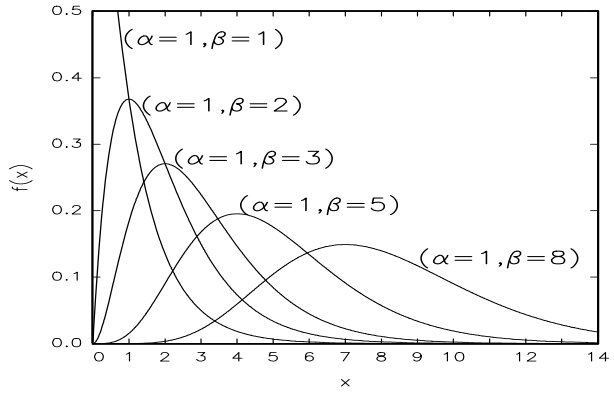


Fig. 2.3: The Gamma probability model

NOTE that the particular formula is of no intrinsic interest at this stage. What is important for the discussion in this section is to use this example in order to get some idea as to what lies behind the various symbols used in the generic case. For instance, the parameter space $\Theta$ and the range of values $\mathbb{R}_X$ of the random variable $X$, are the positive real line $\mathbb{R}_+:=(0,\infty)$, i.e., $\Theta:=\mathbb{R}_+$ and $\mathbb{R}_X:=\mathbb{R}_+$. Each curve in figure 2.3 represents the graph of one density function (varying over a subset of the range of values of the random variable $X:(0,14]\subset\mathbb{R}_+$) for a specific value of the parameter $\theta$. In figure 2.3 we can see five such curves for the values: $\theta=\ 1,2,3,5,8$; the latter being a small subset of the parameter space $\mathbb{R}_+$. In other words, the graphs of the density functions shown in figure 2.3 represent a small subset of the set of densities in (3). Figure 2.3 illustrates the notion of a probability model by helping one visualize the family of densities indexed by the parameter $\theta$.

Let us now briefly discuss the various concepts invoked in the above illustration.

## 2.2 The notion of a random variable: a naive view

The notion of a random variable constitutes one of the most important concepts in the theory of probability. For a proper understanding of the concept the reader is required to read through to chapter 3. In order to come to grips with the notion at an intuitive level, however, let us consider the naive view first introduced by Chebyshev (1821-1884) in the middle of the 19th century who defined a random variable as:

"a real variable that assumes different values with different probabilities.".

This definition comes close to the spirit of the modern concept but it leaves a lot to be desired from the mathematical viewpoint.

As shown in chapter 3, a random variable is a *function* from a set of outcomes to the real line; attaching numbers to outcomes! The need to define such a function arises because the outcomes of certain stochastic phenomena do not always come in the form of numbers but the data often do. The naive view of a random variable suppresses the set of outcomes and identifies the notion of a random variable with its range of values $\mathbb{R}_X$; hence the term *variable*.

**Example 2.3**. In the case of the experiment of casting two dice and looking at the uppermost faces, discussed in chapter 1, the outcomes come in the form of combinations of die faces (not numbers!), all 36 such combinations, denoted by, say, $\{s_1, s_2, ..., s_{36}\}$. Let us assume that we are interested in the sum of dots appearing on the two faces. This amounts to defining a random variable:

$$X(.):\ \{s_1, s_2, ..., s_{36}\} \rightarrow \mathbb{R}_X{:=}\{2, 3, ..., 12\}.$$

However, this is not the only random variable we could have defined. Another one might be:

$$Y(.):\ \{s_1, s_2, ..., s_{36}\} \rightarrow \{0, 1\},$$

if we want to define the outcomes: even ($Y{=}0$) and odd ($Y{=}1$). This example suggests that ignoring the outcomes set and identifying the random variable with its range of values can be misleading. Be that as it may, let us take this interpretation at face value and proceed to consider the other important dimension of the naive view of a random variable: its randomness. The simplest way to explain this dimension is to return to the above example.

**Example 2.4**. In the case of the experiment of casting two dice and adding the dots of the uppermost faces, we defined two random variables, which the naive view identifies with their respective range of values:

$$X \text{ with } \{2, 3, ..., 12\} \text{ and } Y \text{ with } \{0, 1\}.$$

In the case of the random variable $X$, the association of its values and the probabilities (density function), (chapter 1), takes the form:

| $x$ | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $f(x)$ | $\frac{1}{36}$ | $\frac{2}{36}$ | $\frac{3}{36}$ | $\frac{4}{36}$ | $\frac{5}{36}$ | $\frac{6}{36}$ | $\frac{5}{36}$ | $\frac{4}{36}$ | $\frac{3}{36}$ | $\frac{2}{36}$ | $\frac{1}{36}$ |

(4)

Similarly, the density function of the random variable $Y$ is:

| $y$ | 0 | 1 |
|---|---|---|
| $f(y)$ | $\frac{1}{2}$ | $\frac{1}{2}$ |

(5)

More generally, the density function is defined by:

$$\mathbb{P}(X{=}x){=}f(x),\ \text{for all } x{\in}\mathbb{R}_X, \tag{6}$$

and satisfies the properties:

$$\textbf{(a)}\ f_x(x){\geq}0, \text{ for all } x{\in}\mathbb{R}_X, \quad \textbf{(b)}\ \textstyle\sum_{x_i \in \mathbb{R}_X} f_x(x_i){=}1.$$

The last property just says that adding up the probabilities for all values of the random variable will give us one. The density function can be visualized as distributing a unit of mass (probability) over the range of values of $X$.

### 2.2.1 Continuous random variables

The above example involves two random variables which comply perfectly with Chebyshev's naive definition. With each value of the random variable we associate a probability. This is because both random variables are discrete: their range of values is *countable*. On the other hand, when a random variable takes values over an interval, i.e., its range of values is *uncountable*, things are not as simple. Attaching probabilities to particular values does not work (see chapter 3) and instead, we associate probabilities with intervals which belong to this range of values. Instead of (6), the density function for continuous random variables is defined over intervals as follows:

$$\mathbb{P}(x \leq X < x+dx)=f(x)dx, \text{ for all } x\in\mathbb{R}_X,$$

and satisfies the properties: **(a)** $f_x(x) \geq 0$, for all $x\in\mathbb{R}_X$, **(b)** $\int_{x\in\mathbb{R}_X} f_x(x)dx=1$.

It is important to note that the density function for continuous random variables takes values in the interval $[0,\infty)$; its values cannot be interpreted as probabilities In contrast, the density function for discrete random variables takes values in the interval $[0,1]$.

## 2.3 Density functions

The densities of the random variables $X$ and $Y$ associated with the casting of the two dice experiment, introduced above, involve no unknown parameters because the probabilities are known. This has been the result of implicitly assuming that the dice are symmetric and each side arises with the same probability. In the case where it is known that the dice are loaded, the above densities will change in the sense that they will now involve some unknown parameters.

**Example 2.5**. In the case of two outcomes $\{0,1\}$, assuming that $\mathbb{P}(Y=1)=\theta$ (an unknown parameter) $0 \leq \theta \leq 1$, the density function of $Y$ now takes the form:

| $y$ | 0 | 1 |
|---|---|---|
| $f(y;\theta)$ | $1-\theta$ | $\theta$ |

(7)

This can be expressed in the more compact form of the formula:

$$f(y;\theta)=\theta^y(1-\theta)^{1-y},\ \theta\in[0,1],\ y=0,1,$$

known as the **Bernoulli** density, with $\Theta:=[0,1]$ and $\mathbb{R}_Y:=\{0,1\}$.

**Example 2.6**. The notion of a parametric distribution (density) goes back to the 18th century with Bernoulli proposing the **Binomial** distribution with density function:

$$f(x;\theta)=\binom{n}{x}\theta^x(1-\theta)^{n-x},\ \theta\in[0,1],\ x=0,1,\ n=1,2,...,$$

where $\binom{n}{x}=\frac{n!}{(n-x)!x!}$, $n!=n\cdot(n-1)\cdot(n-2)\cdots(3)\cdot(2)\cdot(1)$.

**Example 2.7**. In the early 19th century De Moivre and Laplace introduced the **Normal** distribution whose density is:

$$f(x;\boldsymbol{\theta})=\frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\},\ \boldsymbol{\theta}:=(\mu,\sigma^2)\in\mathbb{R}\times\mathbb{R}_+,\ x\in\mathbb{R}.$$

The real interest in parametric densities for modeling purposes began with Pearson (1895) who proposed a family of distributions known today as the *Pearson family* which includes the Normal, the Student's t, the Laplace, the Pareto, the Gamma, and the Beta, as well as discrete distributions such as the Binomial, the Negative Binomial, the Hypergeometric and the Poisson (see Appendix 3.D).

### 2.3.1 The parameter(s) $\boldsymbol{\theta}$

As can be seen in figure 2.3, the parameters $\boldsymbol{\theta}$ are related to distinctive features of the density function such as the shape and the location. As the values of the parameters $\boldsymbol{\theta}$ change over their range of values $\Theta$, the parameter space. Hence, the notion of a parametric family of densities indexed by $\boldsymbol{\theta}\in\Theta$. The notion of a simple statistical model and its first component, a parametric family of densities will be discussed at length in chapter 3 and thus no further discussion will be given in this section; see Appendix A for a more complete list of parametric densities.

## 2.4 A random sample: a preliminary view

### 2.4.1 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.

### 2.4.2 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 $C_k^n = \frac{n!}{r!(n-r)!}$ |

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:

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

**Example 2.9**. From a deck of 52 cards, five cards are drawn without replacement. What is the probability that at least one card is an ace? The five cards can be selected in $\binom{52}{5}$ different ways. In $\binom{48}{5}$ of these, none of the cards selected is an ace. Hence:

$$\mathbb{P}(\text{at least one ace}) = 1 - \frac{\binom{48}{5}}{\binom{52}{5}} = 1 - \frac{\frac{48!}{(5!)(43!)}}{\frac{52!}{(5!)(47!)}} = .341$$

When the *order* of the elements is important, e.g. the sets $\{a, b\}$ and $\{b, a\}$ are considered different, an alternative counting result applies to account for that.

**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:

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

**Example 2.10**. How many ordered subsets with 3 letters can one form from the set of 4 letters $\{a, b, c, d\}$? When the order is irrelevant the answer is $C_3^4 := \binom{4}{3} = \frac{4!}{3!(4-3)!} = 4$, and they are $\{a, b, c\},\ \{a, b, d\},\ \{a, c, d\},\ \{b, c, d\}$. When the subsets are ordered, so that $\{a, b, c\} \neq \{b, a, c\}$ then the answer is: $P_3^4 := \frac{4!}{(4-3)!} = 24$.

**Example 2.11 - Galileo's three dice puzzle**. During the early 17th century a puzzle relating to casting three identical dice and adding up the number of dots was confusing the gamblers in Italy, including the Grand Duke Cosimo II of Tuscany. The puzzle was why the numbers 10 and 11 seem to occur empirically more often than 9 and 12 even though all four numbers occur in six combinations. For instance 10 occurs when elementary outcomes $\{(6, 3, 1), (6, 2, 2), (5, 4, 1), (5, 3, 2), (4, 4, 2), (4, 3, 3)\}$

occur, and 12 occurs when $\{(6,5,1),(6,4,2),(6,3,3),(5,5,2),(5,4,3),(4,4,4)\}$ occur. Duke Cosimo II asked Galileo Galilei (1564-1642), the pre-eminent astronomer, physicist, engineer, philosopher, and mathematician of the 17th century to elucidate the puzzle. In a pamphlet published in 1620 Galileo explained away the puzzle by pointing out that the apparent equality of the number of elementary outcomes associated with the events $(9,10,11,12)$ hides the different *permutations* of occurrence relating to each of the elementary outcomes. For instance, there are 6=3! permutations associated with $(6,3,1)$ but only 3=(3!/2!) permutations associated with $(6,2,2)$ because it includes the same number twice and $(4,4,4)$ has only one permutation: 1=(3!/3!). By extending table 1.3 to three dice, yielding $6^3$=216 elementary outcomes, Galileo explained the puzzle away by showing that the total number of permutations associated with the six outcomes for $(9,10,11,12)$ are $(25,27,27,25)$, respectively, which implies (Gorroochurn, 2012): $\mathbb{P}(9)=\mathbb{P}(12)=\frac{25}{216}=.116<.125=\frac{27}{216}=\mathbb{P}(10)=\mathbb{P}(11)$.

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. It is important to emphasize that the appropriateness of the IID assumptions when the data result from particular sampling procedures is often is a matter of *good design*.

**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:

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

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.

### 2.4.3 Sample survey procedures

**Simple random sampling.** Theoretically, the best way to design a random sample for the voting intentions of the USA electorate is to use *simple random sampling*: assign a number to every voter, irrespective of location, and then let a computer draw at random 1000 numbers; each voter has the same probability of being picked. Then proceed to ask the selected voters to register their voting intentions. This is often an impossible task, which raises the question

▶ what other sampling procedures are being used to collect a representative sample?

For the sample survey to be reliable, the modeler has to choose the subset of the voting population carefully so as to be representative of the population. The survey designer should account for the different factors which influence voting intentions by carefully selecting a subset of voters to be asked and tailoring the questionnaire to elicit the information needed.

**Stratified sampling**. This method of sampling is useful in cases where there is information at the outset relating to the heterogeneity of the target population from one group to another; the elements of each group are roughly homogeneous. In order to utilize this heterogeneity information to improve the accuracy of the results, the modeler divides the population into these heterogeneous groups (strata) and proceeds to collect data from each stratum using random sampling; hence the name stratified sampling. It can be shown that the accuracy of the estimated mean for the population, as estimated by its variance, increases with the difference in the means between strata.

**Cluster sampling**. This method of sampling is useful in cases where the target population is naturally divided into clusters and we need to economize on the cost of sampling. A way to do that is to draw a random sample of clusters first and then proceed to collect the data using random samples whose size reflects the proportion of the population represented by the cluster in question. For example, for a household consumption survey of the USA instead of drawing a random sample of, say, 5000 households from the whole of the USA by random sampling, one might draw a random sample of 100 counties first and then proceed to sample these proportionately to their population using random sampling.

**Quota sampling**. This is a popular method for public-opinion polls in which the interviewer is instructed to ask a pre-specified quota of people with certain specific characteristics such as sex, age, income, etc. The aim in this case is to try to account for the factors influencing the decision, ignoring the randomness of the sample. This method can introduce all kind of unknown biases into the analysis of the data.

Terminology. The terminology developed for sample survey data analysis in the 1930s and 1940s was bequeathed to statistical inference more broadly and has led to several confusions. The term *population* was first introduced for sample surveys to mean "a set of units such as people, states, households, and government agencies, about which the modeler wants some information". The term *sampling* was first introduced in the same context to mean 'selecting a subset of the target population'.

The term *random sample* was developed to mean 'selecting a subset of the target population in such a way so that every unit in the population has the same probability of being selected'. Unfortunately, this terminology can be very misleading in the context of observational data and should used with caution. For instance, the notion of a 'population' is completely inappropriate for observational data on macroeconomics variables, such as Gross Domestic Product over time.

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 Outlining the early milestones of probability theory

In an attempt to give the reader some idea as to the origins and the development of probability theory, we present the milestones over the last four centuries; for a more detailed account see Stigler (1986), Porter (1986), Hacking (2006), Hald (1990), Maistrov (1974) and Gorroochurn (2012).

Glimpses of probabilistic ideas relating to odds of winning or losing in dice and card games can be traced back to **Gerolamo Cardano** (1501-1576) in his book "The book on dice games", published posthumously in 1663. Cardano calculated the odds in dice and card games in the context of discussing fair bets and introduced the idea of the *number of equally possible outcomes* and the proportion relating to an event. Apart from certain isolated instances of combinatorial calculations, nothing very significant happened for the next century or so until the well-known series of letters between **Pierre de Fermat** (1601-1665) and **Blaise Pascal** (1623-1662) in relation to probabilities associated with games of chance. The origins of probability theory as providing systematic ways for solving problems in games of chance appeared in these letters. Pascal and Fermat are credited with the first correct answer to an old problem of *dividing the stakes when a fair game is stopped before either player wins*. The next important milestone was the book "How to reason in dice games" by **Christiaan Huyghens** (1629-1695) which proved to be the first widely read textbook on probability pertaining to games of chance. Huyghens introduced the fundamental notion of *mathematical expectation* and the basic rules of addition and multiplication of probabilities. The next influential book on probability entitled "The Art of Conjecturing" was written by **James Bernoulli** (1654-1705) and published posthumously in 1713 by his nephew Nicholas. This was a turning point for probability theory because it went beyond the probabilities associated with games of chance and proved the first of the so-called *limit theorems* known today as the Law of Large Numbers as a justification for using observed frequencies as probabilities. This thread was taken up by **Abraham de Moivre** (1667-1754) who proved the

second limit theorem, known today as the Central Limit theorem, in his book "The doctrine of chances" published in 1718. Important notions such as *independence* and *conditional probabilities* are formalized for the first time by de Moivre.

**Pierre Simon Laplace** (1749-1827) in his 1812 book "The Analytical theory of probability", drew together and extended the previous results on the probabilities associated with games of chance and the limit theorems and related these results to the development of methods for reconciling observations. Laplace and **Carl Frederic Gauss** (1777-1855) founded the tradition known as the *theory of errors* which linked probability theory to the modeling of observed data by operationalizing the central limit theorem and introducing the method of *least squares*. This was achieved by viewing errors of observations as the cumulative effect of numerous independent errors. The reign of the Normal distribution began with Laplace and Gauss (hence Gaussian distribution) and continues unabated to this day. Laplace's synthesis of probability theory and the reconciliation of observations provided the foundation of mathematical statistics: analysis of data by fitting models to observations.

The foundations of probability suggested by games of chance proved too limiting, and the search for new foundations began with **Lvovich Pafnufty Chebyshev** (1821-1884) and extended by his students **Andrei Andreiwich Markov** (1856-1922) and **Alexander Michailovich Lyapunov** (1857-1918). Chebyshev introduced the notion of a random variable and opened several new research paths with just four publications. His students Markov and Lyapunov met the challenge admirably and all three had a profound effect on probability theory. Their lasting effect is better seen in the limit theorems where they developed revolutionary new methods for studying the asymptotic behavior of sums of independent random variables. The modern mathematical foundations of probability theory were provided by **Andrei Nikolaevich Kolmogorov** (1903-1989) in his book "Foundations of probability theory" first published in 1933. This book established probability theory as part of mathematics proper and provided the foundation for modern statistical inference pioneered a decade earlier by **Ronald A. Fisher** (1890-1963).

## 3.2 Probability theory: a modeling perspective

Intuitively we can think of probability as an attempt to tame *chance regularity*. The failure to provide a satisfactory intrinsic definition of probability is mainly due to our failure to come to grips with the notion of chance regularity in a generally acceptable way. However, for most purposes the axiomatic (mathematical) definition, as given in section 5 below, is adequate. This definition amounts to saying that probability is what we define it to be via the chosen properties (axioms)!

The well-known axiomatic approach to a branch of mathematics, going back to Euclid, specifies the basic axioms and primitive objects and then develops the theory (theorems, lemmas, etc.) using deductive logic. The approach adopted in this chapter (see also Spanos, 1986) is pragmatic in the sense that the axioms and basic concepts will be motivated by striving to formalize the regularity patterns exhibited

by observable chance mechanisms of the type we seek to model in the context of probability theory. In particular, the basic concepts will be introduced initially as a formalization of a simple chance mechanism we call a *random experiment*, such as the examples of casting dice in chapter 1. This approach has certain advantages for non-mathematicians over the usual axiomatic approach.

*First*, it enables the reader to keep an eye on the forest and not get distracted by the beauty or the ugliness (in the eye of the beholder) of the trees. It is imperative for the reader not to lose sight of the main objective of probability theory, which is to provide a framework in the context of which stochastic phenomena can be modeled.

*Second*, motivating the mathematical concepts using a particular chance mechanism enables us to provide a manifest direct link between observable phenomena and abstract mathematical concepts. This enhances the intuition for the mathematical concepts and gives some idea as to why we need these concepts.

*Third,* historically the development of many branches of mathematics follows the pragmatic approach and the axiomatization follows after the area in question has reached a certain maturity. Probability theory existed for many centuries before it was axiomatized in 1933.

*Fourth*, it enables us to begin with a somewhat simplified mathematical structure, by formalizing a simple chance mechanism. We can then proceed to extend the mathematical apparatus to broaden its intended scope and encompass more realistic chance mechanisms of the type we encounter in several fields, including economics, ecology, biology and geoscience.

# 4 A simple generic stochastic mechanism

## 4.1 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:

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

[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:

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

[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.

## 4.2 A bird's eye view of the unfolding story

The formalization of the notion of a random experiment will occupy us for the next two chapters. In the process of formalization several new concepts and ideas will be introduced. The aim is to set up a mathematical framework for modeling economic data which exhibit chance regularity. In the discussion that follows, we will often find ourselves digressing from the main story line in an effort to do justice to the concepts introduced. Hence, it is of paramount importance for the reader to keep one eye firmly on the forest and not be distracted by the trees. With that in mind let us summarize the proposed formalization.

The *first step* will be to formalize condition [a], by defining the set of all possible distinct outcomes $(S)$ (see section 3). In section 4 we take the *second step* which is

concerned with the formalization of condition [b], relating to the uncertainty of the particular outcome in each trial. Even though at each trial the particular outcome is not known in advance, we often have information as to which outcomes are more probable (they are likely to occur more often) than others. This information will be formalized by attaching *probabilities* to the set of outcomes defined in the first step. In these two steps we construct what we call a *probability space*. It's worth summarizing the construction of a probability space to help the reader keep his/her eyes on the forest. We begin with a collection (a set) $S$ of what we call *elementary events* and then proceed to define another collection $\Im$, made up of subsets of $S$ we call *events,* so that $\Im$ is closed under set union, intersection and complementation. Probability is then defined as a non-negative function $\mathbb{P}(.)$ from $\Im$ to the subset of the real line $[0,1]$; assumed to satisfy $\mathbb{P}(S)=1$ and the additivity property:

$$\text{for } A\in\Im,\ B\in\Im \text{ and } A\cap B=\varnothing,\ \text{then } \mathbb{P}(A\cup B)=\mathbb{P}(A)+\mathbb{P}(B).$$

In the *third step*, taken in section 5, we will formalize condition [c]. The notion of repeating the experiment under identical conditions will be formalized in the form of *random trials*: a set of independent and identical trials.

**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. In the present chapter we present the first more abstract form of the formalization, known as the *statistical space*. The discussion that follows reverses the order followed by Kolmogorov (1933) in the sense that we begin with the phenomena of interest and proceed to motivate the adoption of the particular axioms. The abstract notion of a statistical space discussed in this chapter provides the mathematical foundations for probability theory. For modeling purposes, however, the statistical space needs to be transformed into a less abstract form, that of a statistical model; the subject matter of chapters 3 and 4.

**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.

# 5 Formalizing condition [a]: the outcomes set

## 5.1 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.

## 5.2 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**:

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

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:

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

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:

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

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

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

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

## 5.3 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) | Rational numbers: | $\mathbb{Q}=\{\frac{n}{m}: n\in\mathbb{Z} \text{ and } m\in\mathbb{N}\}$, |
| (iv) | Real numbers | $\mathbb{R}=\{x: x\in\mathbb{R}, -\infty < x < \infty\}$, |
| (v) | Positive real numbers: | $\mathbb{R}_+=\{x: x\in\mathbb{R}, 0 < x < \infty\}$. |

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

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

(b) The set of *integers* is countable because we can define the one-to-one correspondence:

$$\begin{array}{rccccccccc} \mathbb{Z}:=\{ & \cdots & -3 & -2 & -1 & 0 & 1 & 2 & 3 & \cdots & \} \\ & & \updownarrow & \updownarrow & \updownarrow & \updownarrow & \updownarrow & \updownarrow & \updownarrow & & \\ \mathbb{N}:=\{ & \cdots & 7 & 5 & 3 & 1 & 2 & 4 & 6, & \cdots & \}. \end{array}$$

(c) The set $\mathbb{Q}$ of *rational numbers* is a countable set. The one-to-one mapping is more complicated in this case and beyond the scope of this book; see Binmore (1980).

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).

# 6 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 its *occurrence value,* i.e., whether in a particular trial it 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.

## 6.1 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: $A \cup B{:=}\{x\colon\ x{\in}A \text{ or } x{\in}B\}$; 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: $A \cap B{:=}\{x\colon\ x{\in}A \text{ and } x{\in}B\}$; 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: $\bar{A}{:=}\{x\colon\ x{\in}S \text{ and } x{\notin}A\}$; 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 $\overline{\varnothing}=S$.

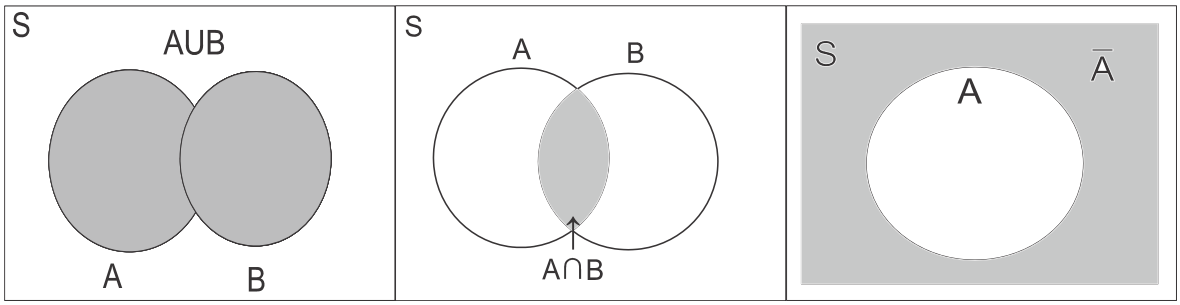


Figure 2.4: Venn diagrams depicting the basic set theoretic operations

(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*.

(c) The complement of the set of rational numbers $\mathbb{Q}$ with respect to the set of real numbers $\mathbb{R}$:

$$\overline{\mathbb{Q}}:=\{x:\ x\in\mathbb{R} \text{ and } x\notin\mathbb{Q}\}\,,$$

is known as the infinite set of irrational numbers.

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

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

that can be extended to: $[1]^*\ \overline{\bigcup_{i=1}^{\infty}A_i}=\bigcap_{i=1}^{\infty}\overline{A_i}$, $[2]^*\ \overline{\bigcap_{i=1}^{\infty}A_i}=\bigcup_{i=1}^{\infty}\overline{A_i}$.

**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 $\bar{A}\cup\bar{C}=\{(HT),(TH),(TT)\}=\bar{C}$.

For completeness we note that by combining the above basic operations with sets we define two other operations often encountered in books on probability.

By combining the set operations of intersection and complementation we define the *difference* between two sets as follows:

$$A-B=A\cap\overline{B}:=\{x:\ x\in A \text{ and } x\notin B\}\,.$$

By combining all three set operations we can define the *symmetric difference* between two sets as follows: $A\triangle B=(A\cap\overline{B})\cup(\overline{A}\cap B):=\{x:\ x\in A \text{ or } x\in B \text{ and } x\notin(A\cap B)\}\,.$

**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.

## 6.2 Events vs. outcomes

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

$$\text{If } A \subset S,\ A \text{ is an } \textit{event.}$$

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

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

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.

### 6.2.1 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**: | $\bigcup_{i=1}^{m} A_i{=}S.$ |

## 6.3 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:

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

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: $\overbrace{2 \times 2 \times \cdots \times 2}^{n \text{ times}} = 2^n$.

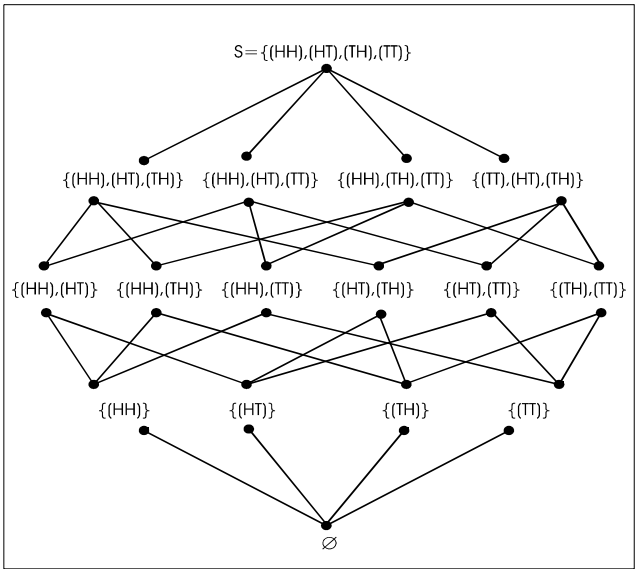


Fig. 2.3: Constructing a power set

**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 $\sigma$-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 $S$. $\Im_0$ is a field because:

$$S \in \Im_0,\ S \cup \varnothing = S \in \Im_0,\ S \cap \varnothing = \varnothing \in \Im_0 \text{ and } S - \varnothing = S \in \Im_0.$$

(c) $\Im(A)=\{S, \varnothing, A, \bar{A}\}$ is the field generated by event $A$. $\Im(A)$ is a field because:

$$\begin{array}{llll} S \in \Im(A), & S \cup \varnothing = S \in \Im(A), & S \cap \varnothing = \varnothing \in \Im(A), & S - \varnothing = S \in \Im(A), \\ \bar{A} \in \Im(A), & A \cup \bar{A} = S \in \Im(A), & A \cap \bar{A} = \varnothing \in \Im(A), & A \cup S = S \in \Im(A), \\ & A \cap S = A \in \Im(A), & \bar{A} \cup S = S \in \Im(A), & \bar{A} \cap S = \bar{A} \in \Im(A). \end{array}$$

**Example 2.32** (counter-examples). (a) $\{S, \varnothing, A, B\}$ cannot be a field because $(A \cup B)$ is not an element of the set, unless $B=\bar{A}$.

(b) $\{S, \varnothing, A, B, (A \cup B)\}$ cannot be a field because $(A \cap B)$ is not an element, unless $A \cap B = \varnothing$.

(c) $\{S, A, \bar{A}\}$ cannot be a field because $\varnothing$ is not an element of this set even though $\varnothing \subset \{S, A, \bar{A}\}$ since 'belongs to' and being 'a subset of' are different concepts.

► **How does one generate a field?**

**Example 2.33.** To illustrate how a field is generated from a set of events of interest, let us consider the case where the set is $D_1=\{A, B\}$ and consider generating the corresponding field. In an effort to avoid getting lost in abstractions we will discuss the generation of a field in relation to our favorite example of tossing a coin twice, where $S:=\{(HH), (HT), (TH), (TT)\}$, $A=\{(HH), (HT)\}$, $B=\{(HT), (TH)\}$ and the field is the power set $\mathcal{P}(S_2)$ as defined above.

**Step 1.** Form the set $D_2=\{S, \varnothing, A, B, \bar{A}, \bar{B}\}$ which includes the complements of events $A$ and $B$. In relation to the above example:

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

**Step 2.** Form the set which also includes all intersections of the elements of $D_2$ i.e., form: $D_3=\{S,\varnothing,A,B,\bar{A},\bar{B},(A\cap B),(\bar{A}\cap B),(A\cap\bar{B}),(\bar{A}\cap\bar{B})\}$.

In relation to our example:

$$A\cap B=\{(HT)\},\ (A\cap\bar{B})=\{(HH)\},\ (\bar{A}\cap B)=\{(TH)\},\ (\bar{A}\cap\bar{B})=\{(TT)\}.$$

Notice that these intersections generate all the events with one outcome.

**Step 3.** Form the set which also includes all unions of the elements of $D_3$:

$$\mathcal{D}=\{D_3,(A\cup B),(A\cup\bar{B}),(\bar{A}\cup B),(\bar{A}\cup\overline{B})\text{ etc.}\}.$$

In relation to example 2.33:

$$[A\cup B]=\{(HH),(HT),(TH)\},\ \left[A\cup\bar{B}\right]=\{(HH),(HT),(TT)\},$$

$$\left[\bar{A}\cup B\right]=\{(HT),(TH),(TT)\},\ \left[\bar{A}\cup\overline{B}\right]=\{(HH),(TH),(TT)\},$$

$$\left[(A\cap\bar{B})\cup(\bar{A}\cap B)\right]=\{(HH),(TH)\},\ \left[(A\cap B)\cup(\bar{A}\cap\overline{B})\right]=\{(HT),(TT)\}.$$

The reader is encouraged to check that the power set of $S$ has indeed been generated!

NOTE that $D_1\subset D_2\subset D_3\subset\mathcal{D}$ and $\mathcal{D}$ is a field. Indeed, $\mathcal{D}$ is the smallest field containing $D_1$, referred to as *the field generated by* $D_1$, and denoted by $\Im(D_1)=\mathcal{D}$.

**Example 2.34.** In the case of tossing a coin three times:

$$S_3=\{(HHH),(HHT),(HTT),(HTH),(TTT),(TTH),(THT),(THH)\}. \tag{8}$$

If the events of interest are, say $A_1=\{(HHH)\}$ and $A_2=\{(TTT)\}$, the set $\{A_1,A_2\}$ is clearly not a field but we can always generate such a field with these two events. Add $(\overline{A_1\cup A_2})$:$=A_3=\{(HHT),(HTT),(HTH),(TTH),(THT),(THH)\}$ to create a partition of $S_3=A_1\cup A_2\cup A_3$ and then create the field of events of interest as:

$$\Im_3=\{S_3,\varnothing,A_1,A_2,A_3,(A_1\cup A_2),(A_1\cup A_3),(A_2\cup A_3)\}. \tag{9}$$

It should be clear from the above examples that generating a field using set theoretic operations, starting from a set of events of interest, is a non-trivial exercise in cases where the number of initial events of interest is greater than 2. The exception to this is the case where the initial events form a *partition* of $S$; hence the addition of $A_3$ to create such a partition.

Consider the events $\{A_1,A_2,...,A_m\}$ that constitute a *partition* of $S$, then the set of all possible unions of the elements of $\mathcal{A}$:$=\{\varnothing,A_1,A_2,...,A_m\}$ forms a field:

$$\Im(\mathcal{A})=\left\{\mathcal{B}\text{: }\mathcal{B}=\textstyle\bigcup_{i\in I}A_i,\ I\subseteq\{1,2,...,n\}\right\}.$$

**Example 2.35.** In the case of tossing a coin three times with outcomes set $S_3$ let the events of interest be: $A_1=\{(HHH),(HHT),(HTT)\}$, $A_2=\{(HTH),(TTT),(TTH)\}$ and $A_3=\{(THT),(THH)\}$. The set $\{A_1,A_2,A_3\}$ is clearly a partition of $S_3$, and thus the relevant field can be generated as:

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

This event space has the mathematical structure of being closed under the set-theoretic operations ($\cup,\cap$ and $^-$), i.e., if we perform any of these operations on any elements of $\Im_3$ the end result will be elements of $\Im_3$ (verify).

The above method can be extended to the case where $S$ is infinite by defining a *countable partition* of it, say $\{A_1, A_2, ..., A_n, ...\}=\{A_i,\ i\in\mathbb{N}\}$. The set of subsets generated by $\mathcal{A}:=\{\varnothing, A_1, A_2, ...A_n, ...\}$ takes the form:

$$\Im(\mathcal{A})=\left\{\mathcal{B}\colon\ \mathcal{B}=\textstyle\bigcup_{i\in I} A_i,\ I\subseteq\mathbb{N}\right\},$$

and constitutes an extension of the notion of a field, known as a $\sigma$-field (sigma-field). The extension amounts to the $\sigma$-field being closed under countable unions and intersections of events.

TERMINOLOGY: in the literature the terms *algebra* and *$\sigma$-algebra* are more widely used than that of *field* and $\sigma$-field. The latter terminology was introduced by Kolmogorov (1933).

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

**Table 2.7: Definition of a sigma-field ($\sigma$-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 $\sigma$-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. The properties (i)-(iv) provide the mathematical structure needed to formalize the notion of an event space. The intuition behind this result stems from the fact that the key characteristic of events $A$ and $B$ is that at a particular trial they might or might not occur. The same characteristic, however, is inherited by the related events $\overline{A},\ \overline{B},\ A\cup B,\ A\cap B$, etc.

**Borel** $\sigma$-field. The most important $\sigma$-field in probability theory is the one defined on the real line $\mathbb{R}$, known as a Borel $\sigma$-field, or *Borel-field* for short, and denoted by $\mathcal{B}(\mathbb{R})$. So far we considered $\sigma$-fields generated by arbitrary sets of outcomes $S$ which were endowed with no other mathematical structure than the set theoretic. The real line $\mathbb{R}$ is obviously not just a set in the same sense of the set of outcomes of the experiment of tossing a coin twice, but it enjoys a rich mathematical structure which enables us to define order among its elements, define distance between any two elements, define convergence in relation to a sequence of its elements etc. The structure that is of particular interest in the present context is the one that enables us to define convergence, known to the mathematical connoisseurs as *topological structure*. Naturally, the Borel field $\mathcal{B}(\mathbb{R})$ enjoys a certain additional mathematical structure generated by that of the real line.

**Example 2.36**. Consider the distance function (metric): $\rho(x,y)=\frac{|x-y|}{|x-y|+1},\ (x,y)\in\mathbb{R}^2$, and define $\mathcal{B}_\rho(\mathbb{R})$ to be the smallest $\sigma$-field generated by the open sets:

$$D_\rho(x_0)=\{x\in\mathbb{R}\colon\ \rho(x,x_0)<r,\ r>0,\ x_0\in\mathbb{R}\}.$$

It can be shown that $\mathcal{B}_\rho(\mathbb{R})=\mathcal{B}(\mathbb{R})$; see Shiryaev (1996), p. 144.

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 $\sigma$-field over an infinite set is to define it via the elements that can generate this set. In the case of the real line a number of different intervals such as $(a,\infty)$, $(a,b]$ $(a,b)$, $(-\infty,b)$, can be used to generate the Borel field. However, 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 $\sigma(B_x)$=$\mathcal{B}(\mathbb{R})$, i.e. these intervals can be created using the set theoretic operations of union, intersection and complementation using $B_x$, $x\in\mathbb{R}$; 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) $[a,b]\ =\cap_{n=1}^{\infty}(a-\frac{1}{n},b]\ \Rightarrow\ [a,b]\in\mathcal{B}(\mathbb{R})$.

It is important to note that the Borel-field $\mathcal{B}(\mathbb{R})$ includes just about all subsets of the real line $\mathbb{R}$, but not quite all! That is, there are subsets of $\mathbb{R}$ which belong to the power set but not to $\mathcal{B}(\mathbb{R})$, i.e. $\mathcal{B}(\mathbb{R})\subset\mathcal{P}(\mathbb{R})$ but $\mathcal{B}(\mathbb{R})\supsetneq\mathcal{P}(\mathbb{R})$.

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 |
| $\sigma$-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 $\Im$.

## 6.4 A digression: what is a function?

Before we proceed to complete the second component in formalizing condition [b] defining a random experiment, we need to take a digression in order to define the concept of a *function* because the type of functions we will need in this and the next chapter go beyond the usual point to point numerical functions. The naive notion of a function as a *formula* enabling $f(x)$ to be calculated in terms of $x$, is embarrassingly inadequate for our purposes.

It is no exaggeration to claim that the notion of a function is perhaps the most important concept in mathematics. However, the concept of a function has caused problems in several areas of mathematics since the time of Euclid because it has changed numerous times. The definitions adopted at different times during the $18^{\text{th}}$ and $19^{\text{th}}$ centuries ranged from "a closed (finite analytical) expression" to "every quantity whose value depends on one or several others" (see Klein, 1972, for a fascinating discussion). The problems caused by the absence of a precise notion of a function were particularly acute during the 19th century when several attempts were made by famous mathematicians, such as Cauchy, Riemann and Weirestrass, to provide more rigorous foundations for Calculus; see Gray (2015). One can go as far as to claim that the requirements of 'Analysis' forced mathematicians to invent more and more general categories of functions which were instrumental in the development of many areas of modern mathematics such as set theory, the modern theory of integration and the theory of topological spaces. In turn, the axiomatization of set theory provided the first general and precise definition of a function in the early $20^{\text{th}}$ century. Intuitively, a function is a special type of "marriage" between two sets.

A **function** $f(.)$: $A\rightarrow B$ is a *relation* between sets $A$ and $B$ satisfying the restriction that for each $x{\in}A$, there exists a *unique* element $y{\in}B$ such that $(x,y){\in}f$. The sets $A$ and $B$ are said to be the *domain* and the *co-domain* of the function $f(.)$.

Intuitively, a relation $R$ connects elements of $A$ to elements of $B$, to define *pairs* $(x,y)$ where $x{\in}A$ and $y{\in}B$, and denoted by $xRy$ or $(x,y){\in}R$. To distinguish between the elements of the input set $A$ and output set $B$ we treat $(x,y)$ as an ordered pair; to indicate the order we use the set theoretic notation: $(x,y){:=}\{x,\{x,y\}\}$.

Formally, a **relation** $R$ is defined to be any *subset* of the Cartesian product $(A{\times}B)$-the set of all *ordered pairs* $(x,y)$ where $x{\in}A$ and $y{\in}B$. That is, a function is a special kind of a relation $(x,y){\in}f$ such that:

(i) every element $x$ of the domain $A$ has an image $y$ in $B$,

(ii) for each $x{\in}A$, there exists a *unique* element $y{\in}B$, i.e. if $y_1{=}f(x)$ and $y_2{=}f(x)$, then $y_1{=}y_2$.

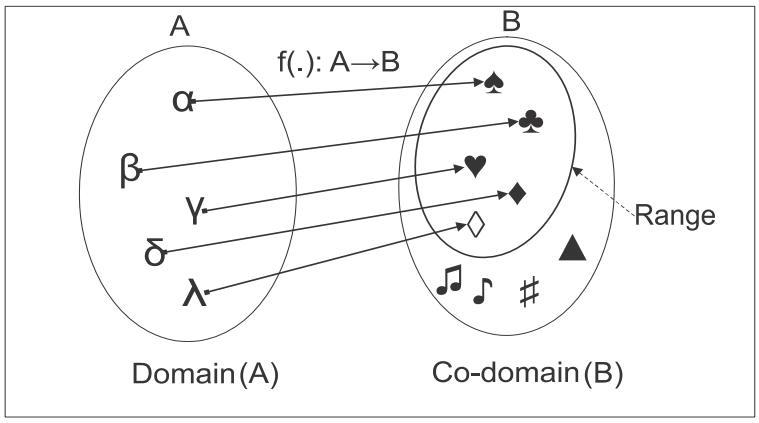


Fig. 2.4: Defining a function $f(.)$: $A \to B$

Looking at figure 2.4 brings out three important features of a function.

(i) The underlying intuition is that two arrows from two different elements in $A$ can connect to the same element in $B$, but no two arrows can emanate from the same element in $A$.

(ii) The nature of the two sets and their elements is arbitrary; it does not have to be a formula relating numbers to numbers.

(iii) The uniqueness restriction concerns the elements of the co-domain which are paired with elements of the domain ($B_A$:=$f(A) \subset B$, and $B_A$ is called the *range* of $f$) and intuitively means that only one arrow emanates from each element $x \in A$, but more than one arrows can end up with any one element $y \in B_A$.

NOTE that we distinguish between the *co-domain* and the *range* of $f$ (figure 2.4). In the case where $B_A$:=$f(A)$=$B$ the function is called *surjective* (onto). Also, in the case where for each $y \in B_A$ there corresponds a unique $x \in A$, the function is said to be *injective* (one-to-one). If the function is both one-to-one and onto it is called a *bijection*.

**Example 2.38.** Let the domain and co-domain of a numerical relation be $A$=$\{1,2,3,4\}$ and $B$=$\{2,3,5,7,11,13\}$, respectively. The set of ordered pairs $f$=$\{(1,13),(2,11),(3,7),(4,5)\}$ constitute a function with range $R_f$=$\{4,5,7,11,13\}$. In contrast, the set of ordered pairs $h$=$\{(1,13),(2,11),(2,3),(3,7),(4,5)\}$ does *not* constitute a function because an element of $A$, is paired with two different numbers in $B$; the number 2 in $A$ takes the values 3 and 11 in $B$.

## 6.5 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.

**Example 2.39.** For example 3.34, with $S_3$ and $\Im_3$ as defined in (8)-(9), common sense suggests that the following assignment of probabilities seems appropriate:

$$\mathbb{P}(A_1)=\tfrac{1}{8},\ \mathbb{P}(A_2)=\tfrac{1}{8},\ \mathbb{P}(\bar{A}_1)=\tfrac{7}{8},\ \mathbb{P}(\bar{A}_2)=\tfrac{7}{8},\ \mathbb{P}(A_1 \cup A_2)=\tfrac{1}{4},\ \mathbb{P}(\bar{A}_1 \cap \bar{A}_2)=\tfrac{3}{4}.$$

In calculating the above probabilities we assumed that the coin is fair and used common sense to argue that for an event such as $A_1 \cup A_2$ we find its probability

by adding that of $A_1$ and $A_2$ together since the two are mutually exclusive. In mathematics, however, we cannot rely exclusively on such things as common sense and intuition when framing a mathematical set up. We need to formalize the common sense arguments by giving a mathematical definition for $\mathbb{P}(.)$.

### 6.5.1 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 theory of probability theory, as a mathematical discipline, can and should be developed from axiom in exactly the same way as Geometry and Algebra. This means that after we have defined the elements to be studied and their basic relations, and have stated the axioms by which these relations are to be governed, all further exposition must be based exclusively on these axioms, independent of the usual concrete meaning of these elements and their relations."; see Kolmogorov (1933), p. 1.

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: $\mathbb{P}(.)$: $\Im \rightarrow [0,1]$,
is said to be *a probability set function* if it satisfies the axioms in table 2.9. Looking at the axioms in table 2.9, it is apparent that the concept of an 'event' is as fundamental in the axiomatic framing of probability theory as the concept of a straight line for Euclidean geometry. Moreover, relations among events pertain exclusively to their occurrence. Axioms [**A1**] and [**A2**] are self-evident but [**A3**] requires some explanation because it is not self-evident and it largely determines the mathematical structure of the probability set function. The countable additivity axiom provides a way to attach probabilities to events by utilizing mutually exclusive events.

**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 $S$: $[p(s_1), p(s_2), ..., p(s_n)]$, 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:

$$\mathbb{P}(A)=\sum_{[i:s_i \in A]} p(s_i). \tag{10}$$

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, $\sum_{n=1}^{\infty} p=\infty$. 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:

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

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:

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

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**]:

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

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 $\sigma$-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 $\sigma$-field associated with $S$, and obtain the probability of any event $A$ via (10). The question which naturally arises is whether we can start with an arbitrary class of subsets of $S$, say $\mathcal{D}$, with $\mathbb{P}(.)$ defined for every element of $\mathcal{D}$ and then proceed to extend it to a $\sigma$-field generated by $\mathcal{D}$. This strategy will work only if $\mathcal{D}$ is a *field*. This is because axiom [**A3**] restricts the assignment of probabilities to countable unions of disjoint sets, and thus one needs a set $\mathcal{D}$ that is closed under the usual set theoretic operations, i.e. a field. $\mathcal{D}$ is then used to generate the $\sigma$-field $\Im=\sigma(\mathcal{D})$ by applying the three set theoretic operations. Having defined the probability set function $\mathbb{P}(.)$ on $\mathcal{D}$, Caratheodory's extension theorem can be used to stretch $\mathbb{P}(.)$ to assign probabilities to all the elements of $\sigma(\mathcal{D})$; 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 $\sigma$-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 $(a,\infty)$, $(a,b]$, $\{a\}$, $(a,b)$, 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 $\sigma$-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 $\mathbb{P}(.)$.

## 6.6 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 $\sigma$-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).

**Example 2.43*. Lebesgue measure on** $S=(0,1]$. Let $\Im=\mathcal{B}((0,1])$ and $\mathfrak{S}=\{(a,b],\ 0\leq a\leq b\leq 1\}$, then the mapping (Shiryayev, 1996):

$$\lambda(.):\ \mathfrak{S}\to[0,1] \text{ such that } \lambda(\varnothing)=0 \text{ and } \lambda((a,b])=b-a,$$

is the *Lebesgue* measure that reduces to a probability assignment $\mathbb{P}((a,b])=\lambda((a,b])$.

Collecting all these components together we can define what we call a *probability space* $(S,\Im,\mathbb{P}(.))$ where $S$ is an outcomes set, $\Im$ is an *event space* associated with $S$, and $\mathbb{P}(.)$ a *probability function* that satisfies axioms [**A1**]-[**A3**] (table 2.9); see Pfeiffer (1978) and Khazanie (1976) for further details. The probability space has all the necessary mathematical structure to be used as the foundation on which one can build the whole edifice we call probability theory.

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.

## 6.7 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. To get some idea of mathematical deduction, let us derive a few such theorems pertaining to probability as specified above.

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**]:

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

**Example 2.45.** For $A=\{(HH),(HT),(TH)\}$ and $B=\{(HH),(TT)\}$, theorem 3 yields: $\mathbb{P}(A\cup B)=\frac{3}{4}+\frac{1}{2}-\frac{1}{4}=1$.

**Boole's inequality**. An obvious extension of theorem 3 is:

$$\mathbb{P}(\cup_{i=1}^{n}A_i)\leq\sum_{i=1}^{n}\mathbb{P}(A_i)$$

**Theorem 4**. For events $A\in\Im,\ B\in\Im$: $\mathbb{P}(B)=\mathbb{P}(A\cap B)+\mathbb{P}(\overline{A}\cap B)$.

This follows from axiom [**A3**] and the fact that events $(A\cap B)$ and $(\overline{A}\cap B)$ are mutually exclusive.

**Theorem 5**. For any event $B\in\Im$ and a set of events $A_i\in\Im,\ i=1,2,,...,n$ defining a partition of $S$, i.e. $A_i\cap A_j=\varnothing$, for all $i\neq j,\ i,j=1,2,...,n$, and $S=\cup_{k=1}^{n}A_i$:

$$\mathbb{P}(B)=(\textstyle\sum_{i=1}^{n}\mathbb{P}(B\cap A_i)) \tag{11}$$

First $S=\cup_{k=1}^{n}A_i$ implies that $B=B\cap S=B\cap(\cup_{k=1}^{n}A_i)=\cup_{k=1}^{n}(B\cap A_i)$. Since the events $(B\cap A_i),\ i=1,2,...,n$ are mutually exclusive, (11) follows from axiom [**A3**].

**Theorem 6**. (a) For $\{A_n\}_{n=1}^{\infty}\in\Im$ an increasing sequence, i.e. $A_n\supseteq A_{n-1},\ n=1,2,...,$:

$$\lim_{n\to\infty}\mathbb{P}(A_n)=\mathbb{P}(A),\ \text{ where } A=\lim_{n\to\infty}A_n=\textstyle\bigcup_{n=1}^{\infty}A_n.$$

(b) For $\{A_n\}_{n=1}^{\infty}\in\Im$ a decreasing sequence, i.e. $A_n\subseteq A_{n-1},\ n=1,2,...,$:

$$\lim_{n\to\infty}\mathbb{P}(A_n)=\mathbb{P}(A),\ \text{ where } A=\lim_{n\to\infty}A_n=\textstyle\bigcap_{n=1}^{\infty}A_n.$$

The proof is based on defining a new *disjoint* sequence $\{B_n\}_{n=1}^{\infty}\in\Im$, where $B_n=A_n-A_{n-1},\ n=1,2,...,$ with $A_0=\varnothing$ and $\bigcup_{n=1}^{\infty}B_n=A_n$, and then using axiom [**A3**]; see Shiryayev (1996). This theorem known as the *continuity property* of $\mathbb{P}(A_n)$ since in both cases (a) and (b) it ensures that:

$$\lim_{n\to\infty}\mathbb{P}(A_n)=\mathbb{P}(\lim_{n\to\infty}A_n),$$

i.e. the limit of the probability of $\{A_n\}_{n=1}^{\infty}$ coincides with the probability of the limit.

In conclusion, let us state a theorem known as the *Bonferroni inequality*:

**Theorem 7**. $\mathbb{P}(\bigcap_{k=1}^{n}A_k)\geq 1-\sum_{k=1}^{n}\mathbb{P}(\overline{A_k}),\ A_k\in\Im,\ k=1,2,...,n.$

For a formal proof, see Williams (1991).

# 7 Conditional probability and Independence

## 7.1 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:

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

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

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

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

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

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 $A\cap B=\{(TH)\}$ (12) 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 (12) enjoys a number of properties.

**(CP1) The product rule for conditional probability**. Using the conditional probability formula (12) we can deduce the *product rule*:

$$\mathbb{P}(A\cap B)=\mathbb{P}(A|B)\cdot\mathbb{P}(B)=\mathbb{P}(B|A)\cdot\mathbb{P}(A). \tag{13}$$

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

$$\mathbb{P}(A_1\cap A_2\cap A_3)=\mathbb{P}(A_3|A_2,A_1)\cdot\mathbb{P}(A_2|A_1)\cdot\mathbb{P}(A_1). \tag{14}$$

**Example 2.49**. Consider an urn containing 3 red, 4 white and 5 blue balls. The probabilities of drawing 'at random' a red ($R$), a white ($W$) or a blue ($B$) ball, separately, are:

$$\mathbb{P}(R)=\tfrac{3}{3+4+5}=\tfrac{1}{4},\ \mathbb{P}(W)=\tfrac{4}{3+4+5}=\tfrac{1}{3},\ \mathbb{P}(B)=\tfrac{5}{3+4+5}=\tfrac{5}{12}.$$

The probability of the event $A$-'draw three balls, a red, a white and a blue, in that sequence' is: $\mathbb{P}(A)=\mathbb{P}(R_1\cap W_2\cap B_3)$. This probability depends on whether we draw *with* or *without replacement*. When drawing with replacement the three events are independent and thus:

$$\mathbb{P}(R_1\cap W_2\cap B_3)=\mathbb{P}(R_1)\cdot\mathbb{P}(W_2)\cdot\mathbb{P}(B_3)=(\tfrac{1}{4})(\tfrac{1}{3})(\tfrac{5}{12})=\tfrac{5}{144}.$$

When drawing without replacement these events are *not* independent. Using (14):

$$\mathbb{P}(R_1 \cap W_2 \cap B_3) = \mathbb{P}(R_1) \cdot \mathbb{P}(W_2|R_1) \cdot \mathbb{P}(B_3|W_2 \cap R_1) = (\tfrac{3}{3+4+5})(\tfrac{4}{2+4+5})(\tfrac{5}{2+3+5}) = \tfrac{1}{22}.$$

Note that: $\mathbb{P}(R_1) \cdot \mathbb{P}(W_2) \cdot \mathbb{P}(B_3) = \frac{55}{1584} < \frac{72}{1584} = \mathbb{P}(R_1) \cdot \mathbb{P}(W_2|R_1) \cdot \mathbb{P}(B_3|W_2 \cap R_1)$.

The formula in (14) can be extended to $n$ events $\{A_1, A_2, ..., A_n\}$ yielding the *sequential* conditioning rule:

$$\boxed{\mathbb{P}\left(\bigcap_{k=1}^{n} A_k\right) = \mathbb{P}(A_n|A_{n-1}, \ldots, A_1) \cdot \mathbb{P}(A_{n-1}|A_{n-2}, \ldots, A_1) \cdot \cdots \cdot \mathbb{P}(A_2|A_1) \cdot \mathbb{P}(A_1).} \tag{15}$$

**(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 (13):

$$\boxed{\mathbb{P}(B) = \mathbb{P}(A) \cdot \mathbb{P}(B|A) + \mathbb{P}(\overline{A}) \cdot \mathbb{P}(B|\overline{A}).} \tag{16}$$

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

$$\mathbb{P}(B) = \textstyle\sum_{i=1}^{n} \mathbb{P}(A_i) \cdot \mathbb{P}(B|A_i). \tag{17}$$

**(CP3). Bayes rule**. Combining (12), (13) and (17), we derive *Bayes' rule*:

$$\boxed{\mathbb{P}(A_i|B) = \frac{\mathbb{P}(A_i) \cdot \mathbb{P}(B|A_i)}{\sum_{i=1}^{n} \mathbb{P}(A_i) \cdot \mathbb{P}(B|A_i)}, \text{ for } \mathbb{P}(B) > 0.} \tag{18}$$

TERMINOLOGY. It is important to bring out the fact that the attribution of formula in (18) 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 (12) pertains to conditional probability between events which was used in the early 16th century by Cardano, and both formulae (12) and (13) are clearly stated in de Moivre (1718/1738); see Hald (1998). Moreover, the claim that (18) 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: $A$ - has the disease, $B$- 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 (18) yields:

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

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%.

**Example 2.51. Monty Hall Puzzle**[1]

A contestant on a TV game 'Let's Make a Deal', is presented with three doors numbered **1**, **2**, **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 **1**.

**Stage 2**. The game host opens door **3** to reveal a goat.

**Stage 3**. The game host asks the contestant: do you want to switch to door **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 **1** and **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: $\mathbb{P}(C_1)$=$\mathbb{P}(C_2)$=$\mathbb{P}(C_3)$=$\frac{1}{3}$.

After the contestant selected door **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, $\mathbb{P}(D_2)$=$\mathbb{P}(D_3)$=$\frac{1}{2}$.

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

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

[1] https://www.facebook.com/virginiatecheconomics/videos/1371481736198880/

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 (12), one can elicit the probabilities the contestant needs:

$$\mathbb{P}(C_1|D_3)=\tfrac{\mathbb{P}(C_1\cap D_3)}{\mathbb{P}(D_3)} \text{ and } \mathbb{P}(C_2|D_3)=\tfrac{\mathbb{P}(C_2\cap D_3)}{\mathbb{P}(D_3)}. \tag{19}$$

To evaluate (19), 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 (13) 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 **2** and **3** at random, hence: $\mathbb{P}(D_3|C_1)=\frac{1}{2}$.
If the car is behind door **2**, he has no option but open door **3**, hence: $\mathbb{P}(D_3|C_2)=1$.
The car could not have been behind door **3**, and thus: $\mathbb{P}(D_3|C_3)=0$.

The coherence of these conditional probabilities is confirmed by the total probability rule in (16): $\mathbb{P}(D_3)=\sum_{k=1}^{3}\mathbb{P}(D_3|C_k)\cdot\mathbb{P}(C_k)=\frac{1}{2}(\frac{1}{3})+1(\frac{1}{3})+0(\frac{1}{3})=\frac{1}{2}$.

Collecting all the relevant probabilities derived above:

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

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.

## 7.2 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 $A$. 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:

$$\mathbb{P}(A|B)=\mathbb{P}(A) \Leftrightarrow \mathbb{P}(B|A)=\mathbb{P}(B) \tag{20}$$

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

$$\mathbb{P}(A\cap B)=\mathbb{P}(A)\cdot\mathbb{P}(B). \tag{21}$$

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

$$\mathbb{P}(A\cap B\cap C)=\mathbb{P}(A)\cdot\mathbb{P}(B)\cdot\mathbb{P}(C). \tag{22}$$

**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: $A=\{(TT),(TH)\}$, $B=\{(TT),(HT)\}$, 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.

# 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* affect that of another.

▶ **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 overall probability space that suggests itself is the *product probability space*:

$$(S_1, \Im_1, \mathbb{P}_1(.)) \times (S_2, \Im_2, \mathbb{P}_2(.)) \times \cdots \times (S_n, \Im_n, \mathbb{P}_n(.)),$$

which can be thought of as a triple of the form:

$$([S_1 \times S_2 \times \cdots \times S_n], [\Im_1 \times \Im_2 \times \cdots \times \Im_n], [\mathbb{P}_1 \times \mathbb{P}_2 \times \cdots \times \mathbb{P}_n]) := \left(\mathbf{S}_{(n)}, \Im_{(n)}, \mathbb{P}_{(n)}\right),$$

in an obvious notation. The technical question that arises is whether $\left(\mathbf{S}_{(n)}, \Im_{(n)}, \mathbb{P}_{(n)}\right)$ is a proper probability space. To be more precise, the problem is whether $\mathbf{S}_{(n)}$ is a proper outcomes set, $\Im_{(n)}$ has the needed structure of a $\sigma$-field and $\mathbb{P}_{(n)}$ defines a set function which satisfies the three axioms. The answer to the first scale of the question is in the affirmative since the outcomes set can be defined by:

$$\mathbf{S}_{(n)} = \left\{\mathbf{s}_{(n)}\colon\ \mathbf{s}_{(n)} := (s_1, s_2, ..., s_n),\ s_i \in S_i,\quad i=1, 2, ..., n\right\}.$$

It turns out that indeed $\Im_{(n)}$ has the needed structure of a $\sigma$-field (for a finite $n$) and $\mathbb{P}_{(n)}$ defines a set function which satisfies the three axioms; the technical arguments needed to prove these claims are beyond the scope of the present book; see Billingsley (1995).

Having established that the product probability space is a proper probability space, we can proceed to view the sequence of trials $\{\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_3, ..., \mathcal{A}_n\}$ as an *event* in $\left(\mathbf{S}_{(n)}, \Im_{(n)}, \mathbb{P}_{(n)}\right)$. An event to which we can attach probabilities.

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:

$$\text{[i] } (S_i, \Im_i, \mathbb{P}_i(.)) = (S, \Im, \mathbb{P}(.)), \text{ for all } i=1, 2, \ldots, n, \tag{23}$$

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

**Example 2.55**. Let $S=\{s_1, s_2, ..., s_k\}$ be a generic outcomes set with:

$$\mathbf{P}=[p(s_1), p(s_2), ..., p(s_k)] \text{ such that } \sum_{i=1}^{k} p(s_i)=1,$$

the associated probability distribution. Then condition [i] amounts to saying that:

[i] **P** is the same for all $n$ trials $\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_3, ..., \mathcal{A}_n$.

Formally, the ID condition reduces $\left(\mathbf{S}_{(n)}, \Im_{(n)}, \mathbb{P}_{(n)}\right)$ into something simpler:

$$\left(\mathbf{S}_{(n)}, \Im_{(n)}, \mathbb{P}_{(n)}\right) \overset{\text{ID}}{\to} (S, \Im, \mathbb{P}(.)) \times (S, \Im, \mathbb{P}(.)) \times ... \times (S, \Im, \mathbb{P}(.)) := [(S, \Im, \mathbb{P}(.))]^n,$$

with the same probability space $(S, \Im, \mathbb{P}(.))$ associated with each trial $k=1, 2, ..., n$.

The second component is more difficult to formalize because it involves ensuring that the outcome in the $i$-th trial does not affect and is not affected by the outcome in the $j$-th trial for $i \neq j$, $i, j=1, 2, ..., n$. Viewing the $n$ trials $(\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_3, ..., \mathcal{A}_n)$ as an event in the context of the product probability space $\left(\mathbf{S}_{(n)}, \Im_{(n)}, \mathbb{P}_{(n)}\right)$, we can formalize this in the form of *independence* among the trials. Intuitively, trial $i$, does not affect and is not affected by the outcome of trial $j$. That is, given the outcome in trial $j$ the probabilities associated with the various outcomes in trial $i$ are unchanged and vice versa. The idea that "given the outcome of trial $j$ the outcome of trial $i$ is unaffected" can be formalized using the notion of *conditioning,* discussed in the previous section.

Let us return to the formalization of the notion of a random experiment $(\mathcal{E})$ by proceeding to formalize condition [c]-(ii): the outcome in one trial does not affect and is not affected by that of another.

**Sampling space.** A sequence of $n$ trials, denoted by $\mathcal{G}_n = \{\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_3, ..., \mathcal{A}_n\}$ where $\mathcal{A}_i$, represents the $i-th$ trial of the experiment, associated with the product probability space $\left(\mathbf{S}_{(n)}, \Im_{(n)}, \mathbb{P}_{(n)}\right)$ is said to be *a sampling space.*

As argued above, we view the $n$ trials $\mathcal{G}_n := \{\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_3, ..., \mathcal{A}_n\}$, as an event in the context of the product probability space $\left(\mathbf{S}_{(n)}, \Im_{(n)}, \mathbb{P}_{(n)}\right)$. As such we can attach a probability to this event using the set function $\mathbb{P}_{(n)}$. Hence, we formalize [c]-(ii) by postulating that the trials are *independent*:

$$\begin{array}{ll} \text{[ii]} & \mathbb{P}_{(n)}(\mathcal{A}_1 \cap \mathcal{A}_2 \cap \cdots \cap \mathcal{A}_k) = \mathbb{P}_1(\mathcal{A}_1) \cdot \mathbb{P}_2(\mathcal{A}_2) \cdot \cdots \cdot \mathbb{P}_k(\mathcal{A}_k), \text{ for } k=2, 3, ..., n, \\ \text{or} & \\ \text{[ii]}^* & \mathbb{P}_{(n)}(\mathcal{A}_k | \mathcal{A}_1, \mathcal{A}_2, ...\mathcal{A}_{k-1}, \mathcal{A}_{k+1}, ..., \mathcal{A}_n) = \mathbb{P}_k(\mathcal{A}_k), \text{ for } k=1, 2, ..., n. \end{array} \tag{24}$$

NOTE that $\mathbb{P}_{(n)}(.)$ and $\mathbb{P}_k(.)$ are different probability set function which belong to the probability spaces $\left(\mathbf{S}_{(n)}, \Im_{(n)}, \mathbb{P}_{(n)}\right)$ and $(\mathbf{S}_k, \Im_k, \mathbb{P}_k)$, respectively; see Pfeiffer (1978).

Taking the conditions of *Independence* (24) and *Identical Distribution* (23) 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*, i.e.

$$\boxed{\mathbb{P}_{(n)}(\mathcal{A}_1 \cap \mathcal{A}_2 \cap ... \cap \mathcal{A}_k) = \mathbb{P}(\mathcal{A}_1) \cdot \mathbb{P}(\mathcal{A}_2) \cdots \mathbb{P}(\mathcal{A}_k), \text{ for } k=2, 3, ..., n,}$$

is referred to as a sequence of *Random trials.*

NOTE that $\mathcal{G}_n^{\text{IID}}$ is a special case of a sampling space $\mathcal{G}_n$ associated with $\left(\mathbf{S}_{(n)}, \Im_{(n)}, \mathbb{P}_{(n)}\right)$, defined above, in the sense that $\mathcal{G}_n^{\text{IID}}$ is associated with $(S, \Im, \mathbb{P}(.))^n$, a sequence of Random trials. In general, the components of $\left(\mathbf{S}_{(n)}, \Im_{(n)}, \mathbb{P}_{(n)}\right)$ can be both *non-Identically Distributed* and *non-Independent.*

## 8.2 The concept of a statistical space

Combining a simple product probability space and a sequence of Random trials we define a *simple statistical space*, denoted by:

$$[(S,\Im,\mathbb{P}(.))^n,\mathcal{G}_n^{\text{IID}}].$$

The term *simple* stems from the fact that this represents a particular case of the more general formulation of a *statistical space* $\left[\left(\mathbf{S}_{(n)},\Im_{(n)},\mathbb{P}_{(n)}\right),\mathcal{G}_n\right]$, where each trial, say $\mathcal{A}_i$, is associated with a different probability space $(S_i,\Im_i,\mathbb{P}_i(.))$ (i.e., non-ID) and the trials are not necessarily independent. As argued in chapters 5-8, in many disciplines the IID formulation is inadequate.

A simple statistical space $[(S,\Im,\mathbb{P}(.))^n,\mathcal{G}_n^{\text{IID}}]$ represents our first formalization of the notion of a random experiment $\mathcal{E}$. This formulation, however, is rather abstract because it involves arbitrary sets and set functions. The main aim of the next chapter is to reduce it to a more appropriate form by mapping this mathematical structure onto the real line where numerical data live.

**The story so far in symbols**

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

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 formalization had a primary objective: 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*.

## 8.3 The unfolding story ahead

In chapter 3 the probability space $(S, \Im, \mathbb{P}(.))$ is mapped onto the real line ($\mathbb{R}$) to define a *probability model* of the form: $\Phi = \{f(x;\boldsymbol{\theta}),\ \boldsymbol{\theta}\in\Theta,\ x\in\mathbb{R}\}$. In chapter 4 the sampling space is transformed into a special type of sampling model we call a *random sample*: a set of random variables $\mathbf{X} := (X_1, X_2, ..., X_n)$ which are Independent and Identically Distributed (IID). The unfolding story in symbols:

$$(S, \Im, \mathbb{P}(.)) \rightarrow \Phi = \{f(x;\boldsymbol{\theta}),\ \boldsymbol{\theta}\in\Theta,\ x\in\mathbb{R}\}, \quad \mathcal{G}_n^{\text{IID}} \rightarrow \mathbf{X} := (X_1, X_2, ..., X_n).$$

---

**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 $\sigma$-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 $\sigma$-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 $\sigma$-field.

# 9 Questions and Exercises

**1**. (a) Explain the main differences between *descriptive* and *inferential (proper) statistics* as they relate to their objectives and the role of probability.

(b) "There is no such a thing as descriptive statistics that summarizes the statistical information in the data in hand without invoking any probabilistic assumptions." Discuss.

(c) Explain the difference between modeling the particular data and modeling the underlying stochastic mechanism that gave rise to the data.

**2**. (a) Compare and contrast figures 2.1 and 2.2 in terms of the type of chance regularity pattern they exhibit.

(b) Explain intuitively why the descriptive statistics for the $\mathbf{x}_0$ data (fig. 2.1) based on $\overline{x}{=}\frac{1}{n}\sum_{k=1}^{n} x_k$, and $s_x^2{=}\frac{1}{n}\sum_{k=1}^{n}(x_k{-}\overline{x})^2$ are unreliable as measures of location and central tendency.

(c) Explain intuitively why the values based on $s_x^2{=}\frac{1}{n}\sum_{k=1}^{n}(x_k{-}\overline{x})^2$ misleadingly inflates the true variation around the mean.

(d) Explain why the descriptive statistics in (b) give rise to reliable measures when applied to the $\mathbf{y}_0$ data (fig. 2.2).

**3**. In example 2.11 on casting three dice and adding up the dots, explain the different permutations for the occurrence of the events $(11, 12)$ and evaluate their probabilities using Galileo's reasoning.

**4**. (a) Explain the difference between combinations and permutations.

(b) Compare and contrast the following sample survey procedures: (i) simple random sampling, (ii) stratified sampling, (iii) cluster sampling, (iv) quota sampling.

**5**. (a) Which of the following observable phenomena can be considered as *Random Experiments*, as defined by conditions [a]-[c] and explain your answer briefly.

(i) A die is cast and the number of dots facing up is counted.

(ii) For a period of a year observe the newborns in NYC as male or female.

(iii) Observe the daily price of a barrel of crude oil.

(b) For each of the experiments (i)-(iii), specify the set $S$ of all distinct outcomes.

(c) Contrast the notions of outcome vs. event; use experiment (i) to illustrate.

**6**. For the sets $A$=$\{2,4,6\}$and $B$=$\{4,8,12\}$ derive the following:

(a) $A \cup B$, (b) $A \cap B$, (c) $\overline{A \cup B}$ relative to $S$=$\{2,4,6,8,10,12\}$. Illustrate your answers using Venn diagrams.

**7**. A die is cast and the number of dots facing up is counted.

(a) Specify the set of all possible outcomes.

(b) Define the sets: $A$-the outcome is an odd number, $B$-the outcome is an even number, $C$-the outcome is less than or equal to 4.

(c) Using your answer in (b), derive: $A \cup B$, $A \cup C$, $B \cup C$, $A \cap B$, $A \cap C$, $B \cap C$.

(d) Derive the probabilities of $A, B, C$ and all the events in (c).

(e) Derive the probabilities: $\mathbb{P}(A|B)$, $\mathbb{P}(A|C)$, $\mathbb{P}(B|C)$.

**8**. (a) "Two mutually exclusive events $A$ and $B$ cannot be independent". Discuss.

(b) Explain the notions of mutually exclusive events and a partition of an outcomes set $S$. How is the latter useful in generating event spaces?

**9**. Define the concept of a $\sigma$-field and explain why we need such a concept for the set of all events of interest. Explain why we cannot use the power set as the event space in all cases.

**10**. Consider the outcomes set $S$=$\{2,4,6,8\}$ and let $A$= $\{2,4\}$ and $B$= $\{4,6\}$ be the events of interest. Show that the field generated by these two events coincides with the power set of $S$.

**11**. Explain how intervals of the form $(-\infty, x]$ can be used to define intervals such as $\{a\}$, $(a,b)$, $[a,b)$, $(a,b]$, $[a,\infty)$, using set theoretic operations.

**12**. (a) Explain the difference between a relation and a function.

(b) Let the domain and co-domain of a numerical relation be $A$=$\{1,2,3,4\}$ and $B$=$\{2,3,5,7,11,13\}$, respectively. Explain whether the set of ordered pairs, (i) $f$=$\{(1,13),(2,11),(3,7),(4,7)\}$, (ii) $h$=$\{(1,13),(2,11),(2,3),(3,7),(4,5)\}$ constitute a function or not.

**13**. Explain whether the probability functions defined below are proper ones:
(i) $\mathbb{P}(A)$=$\frac{2}{3}$, $\mathbb{P}(\bar{A})$=$\frac{1}{3}$, $\mathbb{P}(S)$=1, $\mathbb{P}(\varnothing)$=0,(ii) $\mathbb{P}(A)$=$\frac{1}{3}$, $\mathbb{P}(\bar{A})$=$\frac{1}{3}$, $\mathbb{P}(S)$=1, $\mathbb{P}(\varnothing)$=0,
(iii) $\mathbb{P}(A)$=$\frac{1}{4}$, $\mathbb{P}(\bar{A})$=$\frac{3}{4}$, $\mathbb{P}(S)$=0, $\mathbb{P}(\varnothing)$=1, (iv) $\mathbb{P}(A)$=-$\frac{1}{4}$, $\mathbb{P}(\bar{A})$=$\frac{5}{4}$, $\mathbb{P}(S)$=1, $\mathbb{P}(\varnothing)$=0.

**14**. (a) Explain how we can define a simple probability distribution in the case where the outcomes set is finite.

(b) Explain how we define the probability of an event $A$ in the case where the outcomes set has a finite number of elements, i.e., $S$=$\{s_1, s_2, ..., s_n\}$.

(c) How do we deal with the assignment of probabilities in the case of an uncountable outcomes set?

**15**. Describe briefly, using your own words, the formalization of conditions [a] and [b] of a random experiment into a probability space $(S, \Im, \mathbb{P}(.))$.

**16**. Explain how theorem 4: For events $A$ and $B$: $\mathbb{P}(B)$=$\mathbb{P}(A \cap B)$+$\mathbb{P}(\overline{A} \cap B)$, relates to the total probability formula: $\mathbb{P}(B)$=$\mathbb{P}(A)\cdot\mathbb{P}(B|A) + \mathbb{P}(\overline{A})\cdot\mathbb{P}(B|\overline{A})$.

**17**. Draw a ball from an urn containing 6 black, 8 white and 10 yellow balls.

(a) Evaluate the probabilities of drawing a black ($B$), a white ($W$) or a yellow ($Y$) ball, separately.

(b) Evaluate the probability of the event 'draw three balls, a red, a white and a blue in that sequence', when sampling with or without replacement.

**18**. Apply the reasoning in example 2.50 using the probabilities, $\mathbb{P}(A)$=.3, $\mathbb{P}(B|A)$=.95, $\mathbb{P}(B|\overline{A})$=.05, and contrast your results with those in that example.

**19**. In the Monty-Hall puzzle (example 2.51) explain why the reasoning used by the Math professor to reach the conclusion that keeping the original door or switching to the other will make no difference to the probability of finding the car is erroneous.

**20**. Describe briefly the formalization of condition [c] of a random experiment into a simple sampling space $\mathcal{G}_n^{\text{IID}}$.

**21**. Explain the notions of Independent events and Identically Distributed trials.

**22**. Explain how conditioning can be used to define independence; give examples.

**23**. Explain the difference between a sampling space in general and the simple sampling space $\mathcal{G}_n^{\text{IID}}$ in particular.

**24**. In the context of the random experiment of tossing a coin twice derive the probability of event $A$=$\{(HT),(TH)\}$ given event $B$=$\{(HH),(HT)\}$. Explain why event $A$ and $B$ are independent.

**25***. For two events $A$ and $B$ in $S$, use Venn diagrams to evaluate (a) the smallest possible value of $\mathbb{P}(A \cup B)$, and (b) the greatest possible value of $\mathbb{P}(A \cap B)$, under the following two scenarios: (i) $\mathbb{P}(A)$=.4, $\mathbb{P}(B)$=.6, (ii) $\mathbb{P}(A)$=.3, $\mathbb{P}(B)$=.5; hint: consider all possible relationships between $A$ and $B$ within $S$.