Chapter 1

Basic Definitions

This chapter covers the most basic definitions of probability theory and explores some fundamental properties of the probability function.

1.1 Sample Space and Events

Our starting point is the concept of an abstract random experiment. This is an experiment whose outcome is not necessarily determined before it is conducted. Examples include flipping a coin, the outcome of a soccer match, and the weather. The set of all possible outcomes associated with the random experiment is called the sample space. Events are subsets of the sample space, or in other words sets of possible outcomes. The probability function assigns real values to events in a way that is consistent with our intuitive understanding of probability. Formal definitions appear below.

Definition 1.1.1. A sample space Ω associated with a random experiment is the set of all possible outcomes of the experiment.

A sample space can be finite, for example

$$\Omega = \{1, \ldots, 10\}$$

in the experiment of observing a number from 1 to 10. Or Ω can be countably infinite, for example

$$\Omega = \{0, 1, 2, 3, \ldots\}$$

in the experiment of counting the number of phone calls made on a specific day. A sample space may also be uncountably infinite, for example

$$\Omega = \{x \,:\, x \in \mathbb{R},\, x \geq 0\}$$

in the experiment of measuring the height of a passer-by.

The notation $\mathbb{N}$ corresponds to the natural numbers $\{1, 2, 3, \ldots\}$, and the notation $\mathbb{N} \cup \{0\}$ corresponds to the set $\{0, 1, 2, 3, \ldots\}$. The notation $\mathbb{R}$ corresponds to the real numbers and the notation $\{x : x \in \mathbb{R}, x \geq 0\}$ corresponds to the non-negative real numbers. See Chapter A in the appendix for an overview of set theory, including the notions of a power set and countably infinite and unconuntably infinite sets.

In the examples above, the sample space contained unachievable values (number of people and height are bounded numbers). A more careful definition could have been used, taking into account bounds on the number of potential phone calls or height. For the sake of simplicity, we often use simpler sample spaces containing some unachievable outcomes.

Definition 1.1.2. An event E is a subset of the sample space Ω, or in other words a set of possible outcomes.

In particular, the empty set $\emptyset$ and the sample space Ω are events. Figure 1.1 shows an example of a sample space Ω and two events $A, B \subset \Omega$ that are neither $\emptyset$ nor Ω. The R code below shows all possible events of an experiment with $\Omega = \{a, b, c\}$. There are $2^{|\Omega|}$ such sets, assuming Ω is finite (see Chapter A for more information on the power set).

```
# bring the sets package into scope (install
# it first using install.packages('sets') if
# needed)
library(sets)
Omega = set("a", "b", "c")
# display a set containing all possible
# events of an experiment with a sample
# space Omega
2^Omega
```

```
## {{}, {"a"}, {"b"}, {"c"}, {"a", "b"},
##  {"a", "c"}, {"b", "c"}, {"a", "b",
##  "c"}}
```

Example 1.1.1. *In the random experiment of tossing a coin three times and observing the results (heads or tails), with ordering, the sample space is the set*

$$\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}.$$

The event

$$E = \{HHH, HHT, HTT, HTH\} \subset \Omega$$

describes "a head was obtained in the first coin toss." In this case both the sample space Ω and the event E are finite sets.

Example 1.1.2. *Consider a random experiment of throwing a dart at a round board without missing the board. Assuming the radius of the board is 1, the sample space is the set of all two dimensional vectors inside the unit circle*

$$\Omega = \left\{(x, y) \,:\, x, y \in \mathbb{R},\, \sqrt{x^2 + y^2} < 1\right\}.$$

An event describing a bullseye hit may be

$$E = \left\{(x, y) \,:\, x, y \in \mathbb{R},\, \sqrt{x^2 + y^2} < 0.1\right\} \subset \Omega.$$

In this case both the sample space Ω and the event E are uncountably infinite.

For an event E, the outcome of the random experiment $\omega \in \Omega$ is either in E ($\omega \in E$) or not in E ($\omega \notin E$). In the first case, we say that the event E occurred, and in the second case we say that the event E did not occur. $A \cup B$ is the event of either A or B occurring and $A \cap B$ is the event of both A and B occurring. A^c (in the complement, the universal set is taken to be Ω: $A^c = \Omega \setminus A$) is the event that A did not occur. If the events A, B are disjoint ($A \cap B = \emptyset$), the two events cannot happen at the same time, since no outcome of the random experiment belongs to both A and B. If $A \subset B$, then B occurring implies that A occurs as well.

1.2 The Probability Function

Definition 1.2.1. Let Ω be a sample space associated with a random experiment. A probability function P is a function that assigns real numbers to events $E \subset \Omega$ satisfying the following three axioms.

1. $$\mathsf{P}(E) \geq 0 \qquad \text{for all } E.$$

2. $$\mathsf{P}(\Omega) = 1$$

3. If $E_n, n \in \mathbb{N}$, is a sequence of pairwise disjoint events ($E_i \cap E_j = \emptyset$ whenever $i \neq j$), then
$$\mathsf{P}\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} \mathsf{P}(E_i).$$

Some basic properties of the probability function appear below.

Proposition 1.2.1.

$$\mathsf{P}(\emptyset) = 0.$$

Proof. Using the second and third axioms of probability,

$$\begin{aligned} 1 = \mathsf{P}(\Omega) &= \mathsf{P}(\Omega \cup \emptyset \cup \emptyset \cup \cdots) = \mathsf{P}(\Omega) + \mathsf{P}(\emptyset) + \mathsf{P}(\emptyset) + \cdots \\ &= 1 + \mathsf{P}(\emptyset) + \mathsf{P}(\emptyset) + \cdots, \end{aligned}$$

implying that $\mathsf{P}(\emptyset) = 0$ (since $P(E) \geq 0$ for all E). ∎

Proposition 1.2.2 (Finite Additivity of Probability)**.** *For every finite sequence $E_1, \ldots, E_N$ of pairwise disjoint events ($E_i \cap E_j = \emptyset$ whenever $i \neq j$),*

$$\mathsf{P}(E_1 \cup \cdots \cup E_N) = \mathsf{P}(E_1) + \cdots + \mathsf{P}(E_N).$$

Proof. Setting $E_k = \emptyset$ for $k > N$ in the third axiom of probability, we have

$$\mathsf{P}(E_1 \cup \cdots \cup E_N) = \mathsf{P}\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} \mathsf{P}(E_i) = \mathsf{P}(E_1) + \cdots + \mathsf{P}(E_N) + 0.$$

The last equality above follows from the previous proposition. ∎

Proposition 1.2.3.

$$\mathsf{P}(A^c) = 1 - \mathsf{P}(A).$$

Proof. By finite additivity,

$$1 = \mathsf{P}(\Omega) = \mathsf{P}(A \cup A^c) = \mathsf{P}(A) + \mathsf{P}(A^c).$$

∎

Proposition 1.2.4.

$$\mathsf{P}(A) \leq 1.$$

Proof. The previous proposition implies that $\mathsf{P}(A^c) = 1 - \mathsf{P}(A)$. Since all probabilities are non-negative $\mathsf{P}(A^c) = 1 - \mathsf{P}(A) \geq 0$, proving that $\mathsf{P}(A) \leq 1$. ∎

Proposition 1.2.5. *If $A \subset B$ then*

$$\begin{aligned} &\mathsf{P}(B) = \mathsf{P}(A) + \mathsf{P}(B \setminus A) \\ &\mathsf{P}(B) \geq \mathsf{P}(A). \end{aligned}$$

Proof. The first statement follows from finite additivity:

$$\mathsf{P}(B) = \mathsf{P}(A \cup (B \setminus A)) = \mathsf{P}(A) + \mathsf{P}(B \setminus A).$$

The second statement follows from the first statement and the non-negativity of the probability function. ∎

Proposition 1.2.6 (Principle of Inclusion-Exclusion)**.**

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

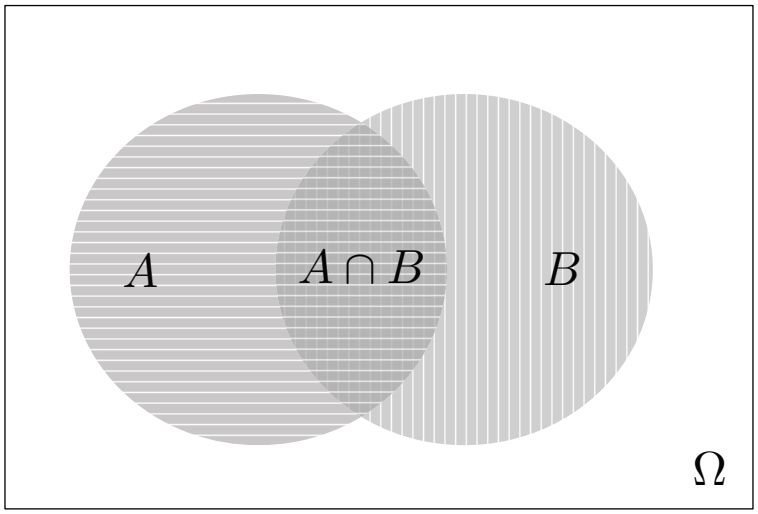


Figure 1.1: Two circular sets A, B, their intersection $A \cap B$ (gray area with horizontal and vertical lines), and their union $A \cup B$ (gray area with either horizontal or vertical lines or both). The set $\Omega \setminus (A \cup B) = (A \cup B)^c = A^c \cap B^c$ is represented by white color.

Proof. Using the previous proposition, we have

$$\begin{aligned}
\mathsf{P}(A \cup B) &= \mathsf{P}((A \setminus (A \cap B)) \cup (B \setminus (A \cap B)) \cup (A \cap B)) \\
&= \mathsf{P}((A \setminus (A \cap B)) + \mathsf{P}(B \setminus (A \cap B))) + \mathsf{P}(A \cap B) \\
&= \mathsf{P}(A) - \mathsf{P}(A \cap B) + \mathsf{P}(B) - \mathsf{P}(A \cap B) + \mathsf{P}(A \cap B) \\
&= \mathsf{P}(A) + \mathsf{P}(B) - \mathsf{P}(A \cap B).
\end{aligned}$$

∎

Figure 1.1 illustrates the Principle of Inclusion-Exclusion. Intuitively, the probability function $\mathsf{P}(A)$ measures the size of the set A (assuming a suitable definition of size). The size of the set A plus the size of the set B equals the size of the union $A \cup B$ plus the size of the intersection $A \cap B$: $\mathsf{P}(A) + \mathsf{P}(B) = \mathsf{P}(A \cup B) + \mathsf{P}(A \cap B)$ (since the intersection $A \cap B$ is counted twice in $\mathsf{P}(A) + \mathsf{P}(B)$).

Definition 1.2.2. For a finite sample space Ω, an event containing a single element $E = \{\omega\}, \omega \in \Omega$ is called an elementary event.

If the sample space is finite $\Omega = \{\omega_1, \ldots, \omega_n\}$, it is relatively straightforward to define probability functions by defining the n probabilities of the elementary events. More specifically, for a sample space with n elements, suppose that we are given a set of n non-negative numbers $\{p_\omega : \omega \in \Omega\}$ that sum to one. There exists

then a unique probability function P over events such that $\mathsf{P}(\{\omega\}) = p_\omega$. This probability is defined for arbitrary events through the finite additivity property

$$\mathsf{P}(E) = \sum_{\omega \in E} \mathsf{P}(\{\omega\}) = \sum_{\omega \in E} p_\omega.$$

A similar argument holds for sample spaces that are countably infinite.

The R code below demonstrates such a probability function, defined on $\Omega = \{1, 2, 3, 4\}$ using $p_1 = 1/2$, $p_2 = 1/4$, $p_3 = p_4 = 1/8$.

```
# sample space
Omega = c(1, 2, 3, 4)
# probabilities of 4 elementary events
p = c(1/2, 1/4, 1/8, 1/8)
# make sure they sum to 1
sum(p)
```

```
## [1] 1
```

```
# define an event 1,4 using a binary
# representation
A = c(1, 0, 0, 1)
# compute probability of A using
# probabilities of elementary events
sum(p[A == 1])
```

```
## [1] 0.625
```

1.3 The Classical Probability Model on Finite Spaces

In the classical interpretation of probability on finite sample spaces, the probabilities of all elementary events $\{\omega\}, \omega \in \Omega$, are equal. Since the probability function must satisfy $\mathsf{P}(\Omega) = 1$ we have

$$\mathsf{P}(\{\omega\}) = |\Omega|^{-1}, \qquad \text{for all} \quad \omega \in \Omega.$$

This implies that under the classical model on a finite Ω, we have

$$\mathsf{P}(E) = \frac{|E|}{|\Omega|}.$$

Example 1.3.1. *Consider the experiment of throwing two distinct dice and observing the two faces with order. The sample space is*

$$\Omega = \{1, \ldots, 6\} \times \{1, \ldots, 6\} = \{(x, y) : x, y \in \{1, 2, \ldots, 6\}\}$$

(see Chapter A in the appendix for the notation of a Cartesian product of two sets). Since Ω *has 36 elements, the probability of the elementary event* $E = \{(4,4)\}$ *is* $\mathsf{P}(E) = 1/|\Omega| = 1/36$*. The probability of getting a sum of 9 in both dice is*

$$\begin{aligned}\mathsf{P}(sum = 9) = \mathsf{P}(\{(6,3),(3,6),(4,5),(5,4)\}) &= \frac{|\{(6,3),(3,6),(4,5),(5,4)\}|}{36} \\ &= \frac{4}{36}.\end{aligned}$$

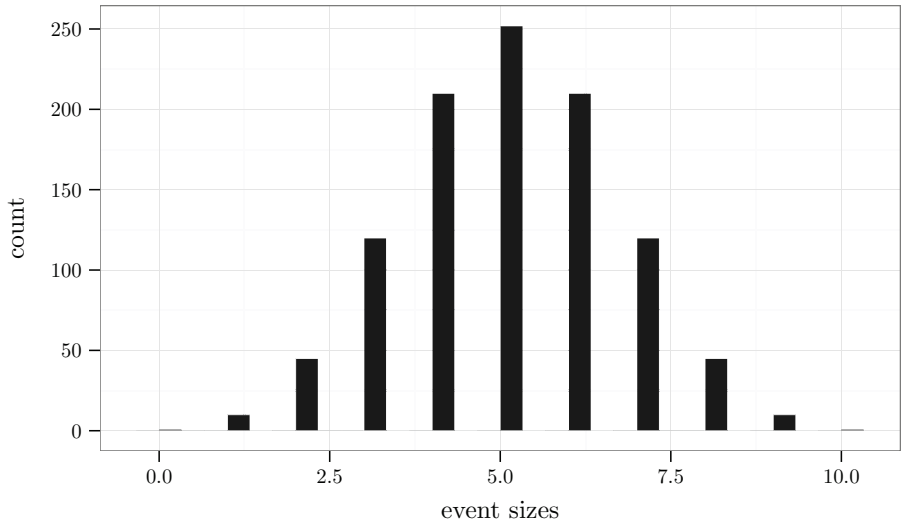
The classical model in this case is reasonable, assuming the dice are thrown independently and are fair.

The R code below demonstrates the classical model and the resulting probabilities on a small Ω.

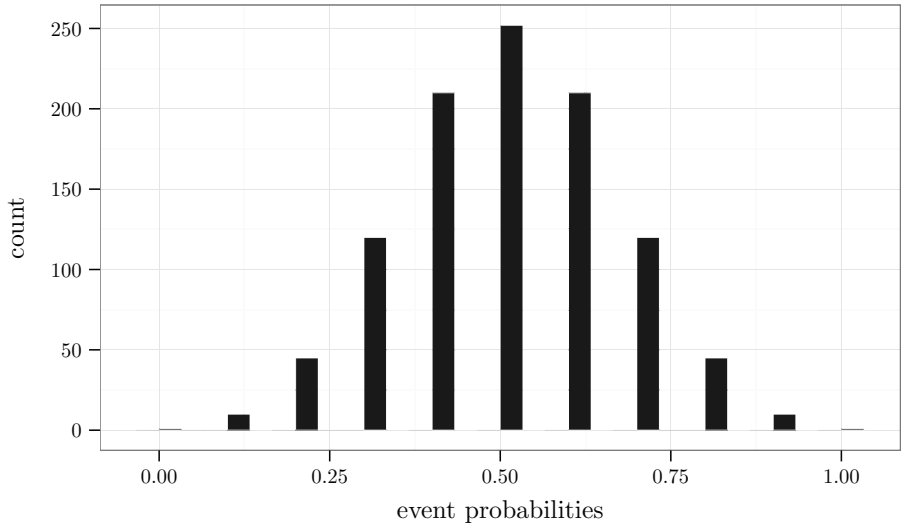
```
Omega = set(1, 2, 3)
# all possible events
2^Omega

## {{}, {1}, {2}, {3}, {1, 2}, {1, 3}, {2,
##  3}, {1, 2, 3}}

# size of all possible events
sapply(2^Omega, length)

## [1] 0 1 1 1 2 2 2 3

# probabilities of all possible events under
# the classical model
sapply(2^Omega, length)/length(Omega)

## [1] 0.0000 0.3333 0.3333 0.3333 0.6667 0.6667
## [7] 0.6667 1.0000
```

Note that the sequence of probabilities above does not sum to one since it contains probabilities of non-disjoint events. The R code below demonstrates this below for a larger set using by graphing the histogram of sizes and probabilities.

```
library(ggplot2)
Omega = set(1, 2, 3, 4, 5, 6, 7, 8, 9, 10)
# histogram of sizes of all possible events
qplot(sapply(2^Omega, length), xlab = "event sizes")
```




```
# histogram of probabilities of all possible
# events under classical model
probs = sapply(2^Omega, length)/length(Omega)
qplot(probs, xlab = "event probabilities")
```



The left-most and right-most bars represent two sets with probabilities 0 and 1, respectively. These sets are obviously $\emptyset$ and Ω.

1.4 The Classical Probability Model on Continuous Spaces

For a continuous sample space of dimension n (for example $\Omega = \mathbb{R}^n$), we define the classical probability function as

$$\mathsf{P}(A) = \frac{\text{vol}_n(A)}{\text{vol}_n(\Omega)},$$

where $\text{vol}_n(S)$ is the n-dimensional volume[1] of the set S.

Example 1.4.1. *In an experiment measuring the weight of residents in a particular geographical region, the sample space could be* $\Omega = (0, 1000) \subset \mathbb{R}^1$ *(assuming our measurement units are pounds and people weigh less than 1000 pounds). The probability of getting a measurement between 150 and 250 (in the classical model) is the ratio of the 1-dimensional volumes or lengths:*

$$\mathsf{P}((150, 250)) = \frac{|250 - 150|}{|1000 - 0|} = 0.1.$$

The classical model in this case is highly inaccurate and not likely to be useful.

Example 1.4.2. *Assuming the classical model on the sample space of Example 1.1.2, the probability of hitting the bullseye is*

$$\mathsf{P}\left(\left\{(x, y) \,:\, \sqrt{x^2 + y^2} < 0.1\right\}\right) = \frac{\pi\, 0.1^2}{\pi\, 1^2} = 0.01$$

(since the area of a circle of radius r is $\pi \cdot r^2$). The classical model in this case assumes that the person throwing the darts does not make any attempt to hit the center. For most dart throwers this model is inaccurate.

1. For the classical model to apply, the sample space Ω must by finite or be continuous with a finite non-zero volume.

2. The classical model (on both finite and continuous spaces) satisfies the three axioms defining a probability function.

3. A consequence of the classical model on continuous spaces is that the probability of an elementary event is zero (the volume of a single element is 0).

4. In the next two chapters we will explore a number of alternative probability models that may be more accurate than the classical model.

[1] The 1-dimensional volume of a set $S \subset \mathbb{R}$ is its length. The 2-dimensional volume of a set $S \subset \mathbb{R}^2$ is its area. The 3-dimensional volume of a set $S \subset \mathbb{R}^3$ is its volume. In general, the n-dimensional volume of A is the n-dimensional integral of the constant function 1 over the set A.

1.5 Conditional Probability and Independence

Definition 1.5.1. The conditional probability of an event A given an event B with $\mathsf{P}(B) > 0$ is

$$\mathsf{P}(A \,|\, B) = \frac{\mathsf{P}(A \cap B)}{\mathsf{P}(B)}.$$

If $\mathsf{P}(A) > 0$ and $\mathsf{P}(B) > 0$ we have

$$\mathsf{P}(A \cap B) = \mathsf{P}(A \,|\, B)\,\mathsf{P}(B) = \mathsf{P}(B \,|\, A)\,\mathsf{P}(A).$$

Intuitively, $\mathsf{P}(A \,|\, B)$ is the probability of A occurring assuming that the event B occurred. In accordance with that intuition, the conditional probability has the following properties.

1. If $B \subset A$, then $\mathsf{P}(A \,|\, B) = \mathsf{P}(B)/\,\mathsf{P}(B) = 1$.
2. If $A \cap B = \emptyset$, then $\mathsf{P}(A \,|\, B) = 0/\,\mathsf{P}(B) = 0$.
3. If $A \subset B$ then $\mathsf{P}(A \,|\, B) = \mathsf{P}(A)/\,\mathsf{P}(B)$.
4. The conditional probability may be viewed as a probability functions

   $$\mathsf{P}_A(E) \stackrel{\text{def}}{=} \mathsf{P}(E \,|\, A)$$

   satisfying Definition 1.2.1 (Exercise 1.7). In addition, all the properties and intuitions that apply to probability functions apply to P_A as well.
5. Assuming the event A occurred, P_A generally has better forecasting abilities than P.

As mentioned above, conditional probabilities are usually intuitive. The following example from [16], however, shows a counter-intuitive situation involving conditional probabilities. This demonstrates that intuition should not be a substitute for rigorous computation.

Example 1.5.1. *Consider families with two children where the gender probability of each child is symmetric (1/2). We select a family at random and consider the sample space describing the gender of the children* $\Omega = \{MM, MF, FM, FF\}$. *We assume a classical model, implying that the probabilities of all 4 elementary events are 1/4.*

We define the event that both children in the family are boys as $A = \{MM\}$, *the event that a family has a boy as* $B = \{MF, FM, MM\}$, *and the event that the first child is a boy as* $C = \{MF, MM\}$.

Given that the first child is a boy, the probability that both children are boys is

$$\mathsf{P}(A \,|\, C) = \mathsf{P}(A \cap C)/\,\mathsf{P}(C) = \mathsf{P}(A)/\,\mathsf{P}(C) = (1/4)/(1/2) = 1/2.$$

This matches our intuition. Given that the family has a boy, the probability that both children are boys is the counterintuitive

$$\mathsf{P}(A \,|\, B) = \frac{\mathsf{P}(A \cap B)}{\mathsf{P}(B)} = \frac{\mathsf{P}(A)}{\mathsf{P}(B)} = (1/4)/(3/4) = 1/3.$$

Definition 1.5.2. Two events A, B are independent if $\mathsf{P}(A \cap B) = \mathsf{P}(A)\,\mathsf{P}(B)$. A finite number of events $A_1, \ldots, A_n$ are independent if

$$\mathsf{P}(A_1 \cap \cdots A_n) = \mathsf{P}(A_1) \cdots \mathsf{P}(A_n)$$

and are pairwise independent if every pair $A_i, A_j, i \neq j$ are independent.

The following definition generalizes independence to an arbitrary collection of events, indexed by a (potentially infinite) set Θ.

Definition 1.5.3. Multiple events $A_\theta, \theta \in \Theta$ are pairwise independent if every pair of events is independent. Multiple events $A_\theta, \theta \in \Theta$ are independent if for every $k > 0$ and for every size k-subset of distinct events $A_{\theta_1}, \ldots, A_{\theta_k}$, we have

$$\mathsf{P}(A_{\theta_1} \cap \ldots \cap A_{\theta_k}) = \mathsf{P}(A_{\theta_1}) \cdots \mathsf{P}(A_{\theta_k}).$$

Note that pairwise independence is a strictly weaker condition than independence.

In agreement with our intuition, conditioning on an event that is independent of A does not modify the probability of A:

$$\mathsf{P}(A \,|\, B) = \mathsf{P}(A)\,\mathsf{P}(B)/\,\mathsf{P}(B) = \mathsf{P}(A).$$

On the other hand, two disjoint events cannot occur simultaneously and should therefore be dependent. Indeed, in this case $\mathsf{P}(A \,|\, B) = 0 \neq \mathsf{P}(A)$ (assuming that $\mathsf{P}(A)$ and $\mathsf{P}(B)$ are non-zero).

Example 1.5.2. *We consider a random experiment of throwing two dice independently and denote by A the event that the first throw resulted in 1, by B the event that the sum in both throws is 3, and by C the event that the second throw was even. Assuming the classical model, the events A, B are dependent*

$$\mathsf{P}(A \cap B) = \mathsf{P}(B \,|\, A)\,\mathsf{P}(A) = (1/6)(1/6) \neq (1/6)(2/36) = \mathsf{P}(A)\,\mathsf{P}(B),$$

while A and C are independent

$$\mathsf{P}(A \cap C) = \mathsf{P}(C \,|\, A)\,\mathsf{P}(A) = (1/2)(1/6) = \mathsf{P}(A)\,\mathsf{P}(C).$$

Proposition 1.5.1. *If A, B are independent, then so are the events A^c, B, the events A, B^c, and the events A^c, B^c.*

Proof. For example,

$$\begin{aligned}\mathsf{P}(A^c \cap B) &= \mathsf{P}(B \setminus A) = \mathsf{P}(B) - \mathsf{P}(A \cap B) = \mathsf{P}(B) - \mathsf{P}(A)\,\mathsf{P}(B) \\ &= (1 - \mathsf{P}(A))\,\mathsf{P}(B) = \mathsf{P}(A^c)\,\mathsf{P}(B).\end{aligned}$$

The other parts of the proof are similar. ∎

Proposition 1.5.2 (Bayes Theorem)**.** *If* $\mathsf{P}(B) \neq 0$ *and* $\mathsf{P}(A) \neq 0$*, then*

$$\mathsf{P}(A \mid B) = \frac{\mathsf{P}(B \mid A)\,\mathsf{P}(A)}{\mathsf{P}(B)}.$$

Proof.

$$\mathsf{P}(A \mid B)\,\mathsf{P}(B) = \mathsf{P}(A \cap B) = \mathsf{P}(B \cap A) = \mathsf{P}(B \mid A)\,\mathsf{P}(A).$$

∎

Example 1.5.3. *We consider the following imaginary voting pattern of a group of 100 Americans, classified according to their party and whether they live in a city or a small town. The last row and last column capture the sum of the columns and the sum of the rows, respectively.*

	City	*Small Town*	*Total*
Democrats	*30*	*15*	*45*
Republicans	*20*	*35*	*55*
Total	*50*	*50*	*100*

We consider the experiment of drawing a person at random and observing the vote. The sample space contains 100 elementary events and we assume a classical model, implying that each person may be selected with equal $1/100$ *probability.*

Defining A *as the event that a person selected at random lives in the city, and* B *as the event that a person selected at random is a democrat, we have*

$$\begin{aligned}
\mathsf{P}(A \cap B) &= 30/100 \\
\mathsf{P}(A^c \cap B) &= 15/100 \\
\mathsf{P}(A \cap B^c) &= 20/100 \\
\mathsf{P}(A^c \cap B^c) &= 35/100 \\
\mathsf{P}(A) &= 50/100 \\
\mathsf{P}(B) &= 45/100 \\
\mathsf{P}(A \mid B) &= 0.3/0.45 \\
\mathsf{P}(A \mid B^c) &= 0.2/0.55 \\
\mathsf{P}(B \mid A) &= 0.3/0.5 \\
\mathsf{P}(B \mid A^c) &= 0.15/0.5.
\end{aligned}$$

Since A, B *are dependent, conditioning on city dwelling raises the probability that a randomly drawn person is democrat from* $\mathsf{P}(B) = 0.45$ *to* $\mathsf{P}(B \mid A) = 0.6$*.*

Proposition 1.5.3 (General Multiplication Rule)**.**

$$\mathsf{P}(A_1 \cap \cdots \cap A_n) = \mathsf{P}(A_1)\,\mathsf{P}(A_2 \mid A_1)\,\mathsf{P}(A_3 \mid A_2 \cap A_1) \cdots \mathsf{P}(A_n \mid A_1 \cap \cdots \cap A_{n-1}).$$

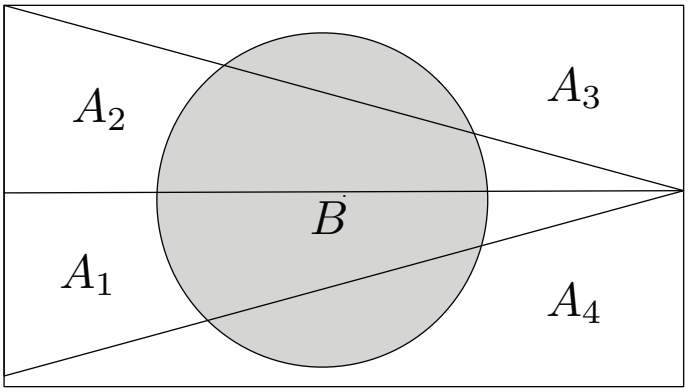


Figure 1.2: The partition $A_1, \ldots, A_4$ of Ω induces a partition $B \cap A_i$, $i = 1, \ldots, 4$ of B (see Proposition 1.5.4).

Proof. Using induction and $\mathsf{P}(A \cap B) = \mathsf{P}(A \,|\, B)\,\mathsf{P}(B)$, we get

$$\begin{aligned}\mathsf{P}(A_1 \cap \cdots \cap A_n) &= \mathsf{P}(A_n \,|\, A_1 \cap \cdots \cap A_{n-1})\,\mathsf{P}(A_1 \cap \cdots \cap A_{n-1}) \\ &= \cdots \\ &= \mathsf{P}(A_1)\,\mathsf{P}(A_2 \,|\, A_1)\,\mathsf{P}(A_3 \,|\, A_2 \cap A_1) \cdots \mathsf{P}(A_n \,|\, A_1 \cap \cdots \cap A_{n-1}).\end{aligned}$$

∎

Proposition 1.5.4 (The Law of Total Probability)**.** *If $A_i, i \in S$, form a finite or countably infinite partition of Ω (see Definition A.1.12)*

$$\mathsf{P}(B) = \sum_{i \in S} \mathsf{P}(A_i)\,\mathsf{P}(B \,|\, A_i).$$

Proof. The partition $A_i, i \in S$, of Ω induces a partition $B \cap A_i$, $i \in S$, of B. The result follows from countable additivity (third probability axiom) or finite additivity applied to that partition

$$\mathsf{P}(B) = \mathsf{P}\left(\bigcup_{i \in S} (B \cap A_i)\right) = \sum_{i \in S} \mathsf{P}(A_i \cap B) = \sum_{i \in S} \mathsf{P}(A_i)\,\mathsf{P}(B \,|\, A_i).$$

∎

Figure 1.2 illustrates the above proposition and its proof.

The definition below extends the notion of independence to multiple experiments.

Definition 1.5.4. Consider n random experiments with sample spaces $\Omega_1, \ldots, \Omega_n$. The set $\Omega = \Omega_1 \times \cdots \times \Omega_n$ (see Chapter A for a definition of the cartesian product

$\times$) is the sample space expressing all possible results of the experiments. The experiments are independent if for all sets $A_1 \times \cdots \times A_n$ with $A_i \subset \Omega_i$,

$$\mathsf{P}(A_1 \times \cdots \times A_n) = \mathsf{P}(A_1) \cdots \mathsf{P}(A_n).$$

In the equation above, the probability function on the left hand side is defined on $\Omega_1 \times \cdots \times \Omega_n$ and the probability functions on the right hand side are defined on Ω_i, $i = 1, \ldots, n$.

Example 1.5.4. *In two independent die throwing experiments* $\Omega = \{1, \ldots, 6\} \times \{1, \ldots, 6\}$ *and*

$$\mathsf{P}(\textit{first die is 3, second die is 4}) = \mathsf{P}(\textit{first die is 3})\,\mathsf{P}(\textit{second die is 4}) = \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36}.$$

Chapter 4 contains an extended discussion of probabilities associated with multiple experiments.

1.6 Basic Combinatorics for Probability

Some knowledge of combinatorics is essential for probability. For example, computing the probability $\mathsf{P}(E)$ in the classical model on finite sample spaces $\mathsf{P}(E) = |E|/|\Omega|$ is equivalent to the combinatorial problem of enumerating the elements in E and Ω.

In the case of a two-stage experiment where stage 1 has k outcomes and stage 2 has l outcomes and every combination of results in the two stages is possible, the total number of combinations of results of stage 1 and 2 is $k \cdot l$. A formal generalization appears below.

Definition 1.6.1. A k-tuple over the sets $S_1, \ldots, S_k$ is a finite ordered sequence $(s_1, \ldots, s_k)$ such that $s_i \in S_i$.

Proposition 1.6.1. *There are* $\prod_{j=1}^{k} |S_j|$ *ways to form* k*-tuples over the finite sets* $S_1, \ldots, S_k$*. In particular if* $S_1 = \cdots = S_k = S$ *there are* $|S|^k$ *possible* k*-tuples.*

Proof. A k-tuple is characterized by picking one element from each group. There are n_1 elements in group 1, n_2 in group 2, and so on. Since choices in one group do not constrain the choices in other groups, the number of possible choices is $\prod_{j=1}^{k} |S_j|$. ∎

Example 1.6.1. *The R code below generates all possible 3-tuples over* $S_1 = \{1, 2\}$*,* $S_2 = \{1, 2, 3\}$*, and* $S_3 = \{1, 2\}$*. There are* $2 \cdot 3 \cdot 2 = 12$ *such possibilities.*

```
expand.grid(S1 = 1:2, S2 = 1:3, S3 = 1:2)
```

```
##    S1 S2 S3
## 1   1  1  1
```

```
## 2   2  1  1
## 3   1  2  1
## 4   2  2  1
## 5   1  3  1
## 6   2  3  1
## 7   1  1  2
## 8   2  1  2
## 9   1  2  2
## 10  2  2  2
## 11  1  3  2
## 12  2  3  2
```

Definition 1.6.2. Assuming that n is a positive integer and $r \leq n$ is another positive integer, we use the following notation:

$$\begin{aligned} n! &\stackrel{\text{def}}{=} n \cdot (n-1) \cdot (n-2) \cdots 2 \cdot 1 \\ (n)_r &\stackrel{\text{def}}{=} \frac{n!}{(n-r)!} \\ \binom{n}{r} &\stackrel{\text{def}}{=} \frac{(n)_r}{r!} = \frac{n!}{r!(n-r)!}. \end{aligned}$$

We refer to the function $f(n) = n!$ as the factorial function and to $\binom{n}{r}$ as n-choose-r.

The factorial function grows very rapidly as $n \to \infty$. The R code below shows the considerable magnitude of $n!$ even for small n.

```
factorial(1:8)
```

```
## [1]     1     2     6    24   120   720  5040
## [8] 40320
```

The following proposition shows that the growth rate of $n!$ is similar to the growth rate of $(n/e)^n$ (see Definition B.2.1 for a definition of the limit notation below).

Proposition 1.6.2 (Stirling's Formula)**.**

$$\lim_{n\to\infty} \frac{n!}{n^{n+1/2}e^{-n}} = \sqrt{2\pi}.$$

Proofs are available in [16] and [37].

Proposition 1.6.3. *The factorial function grows faster than any exponential:*

$$\lim_{n\to\infty} \frac{a^n}{n!} = 0 \qquad \textit{for all } a > 0.$$

Proof. Using

$$\lim_{n\to\infty} \frac{a^n}{n^{n+1/2}e^{-n}} = \lim_{n\to\infty} \frac{1}{\sqrt{n}} \left(\frac{ae}{n}\right)^n = 0, \qquad a > 0,$$

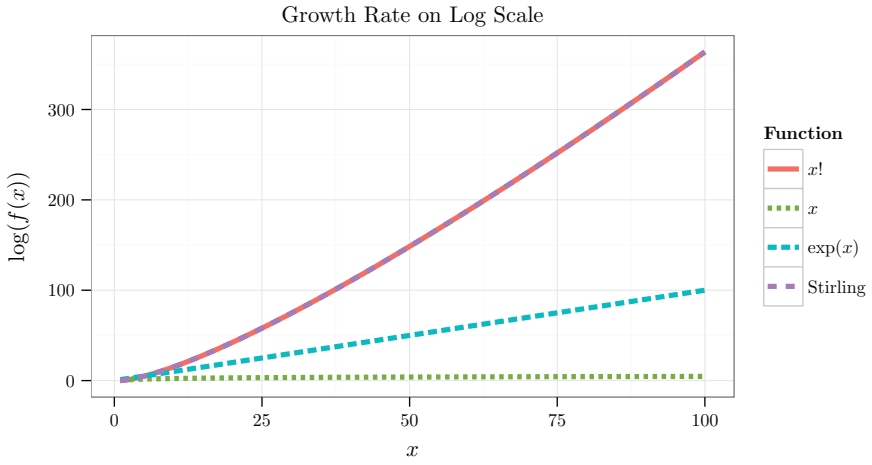
and Proposition 1.6.2, we get

$$\lim_{n\to\infty} \frac{a^n}{n!} = 0, \qquad a > 0. \tag{1.1}$$

∎

The proposition above is illustrated by the following graph comparing the growth rate on a log scale of the factorial with Stirling's approximation, an exponential function, and a linear function x. Stirling's approximation overlaps the factorial line indicating extremely good approximation.

```
x = 1:100
R = stack(list(`$x!$` = lfactorial(x), Stirling = log(2 *
    pi)/2 + (x + 1/2) * log(x) - x, `exp($x$)` = x,
    `$x$` = log(x)))
names(R) = c("lf", "Function")
R$x = x
qplot(x, lf, color = Function, lty = Function,
    geom = "line", xlab = "$x$", ylab = "$\\log(f(x))$",
    data = R, size = I(2), main = "Growth Rate on Log Scale")
```



Proposition 1.6.4. *The number of r-tuples over a finite set S in which no element appears twice is $(|S|)_r$ and the number of different orderings of n elements is $n!$.*

Proof. The first statement is a direct corollary of Proposition 1.6.1, where each value in the r tuple is selected from the population of remaining or unselected items. The second statement follows from the first ($n = r = |S|$). ∎

The following code generates all possible orderings of the letters a, b, and c. There are $3! = 6$ such orderings.

```
# generate all 6 permutations over three
# letters
library(gtools)
permutations(3, 3, letters[1:3])
```

```
##      [,1] [,2] [,3]
## [1,] "a"  "b"  "c"
## [2,] "a"  "c"  "b"
## [3,] "b"  "a"  "c"
## [4,] "b"  "c"  "a"
## [5,] "c"  "a"  "b"
## [6,] "c"  "b"  "a"
```

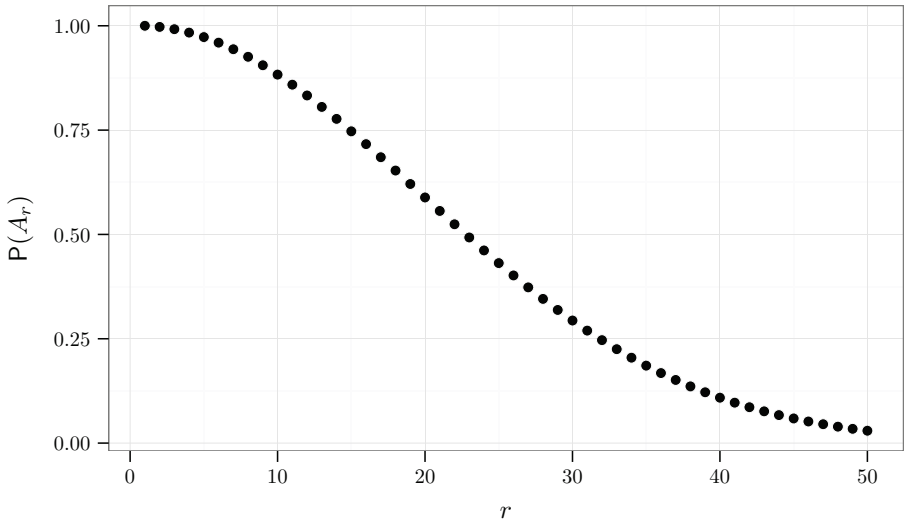
Example 1.6.2 (The Birthday Paradox)**.** *There are* 365^r *possible assignments of birthdays to* r *people. Using the previous proposition, the number of assignments of birthdays to* r *people, assuming that all birthdays are different, is* $(365)_r$*. Under the classical probability model, the probability* $\mathsf{P}(A_r)$ *that a group of* r *people will have all different birthdays is*

$$\mathsf{P}(A_r) = \frac{|A_r|}{|\Omega|} = \frac{(365)_r}{365^r}.$$

For example, $\mathsf{P}(A_{30}) \approx 0.294$*, implying that it is likely to find recurring birthdays in a group of 30 people. The graph below shows how the probability of having different birthdays decays to zero as* r *increases. The median (the value at which the probability is approximately 1/2) is* $r = 23$*. The name "The Birthday Paradox" is sometimes associated with this example, since it is intuitively likely that 23 people will all have different birthdays with high probability.*

The following R code graphs the probability of r *people having all different birthdays as a function of* r*.*

```
# perform calculation on log-scale to avoid
# overflow
r = 1:50
p = exp(lfactorial(365) - lfactorial(365 - r) -
    r * log(365))
qplot(x = r, y = p, size = I(2), xlab = "$r$",
    ylab = "$\\P(A_r)$")
```



Proposition 1.6.5. *A population of n elements has $\binom{n}{r}$ different subsets of size r. Equivalently there are $\binom{n}{r}$ ways to select r elements out of n distinct elements with no element appearing twice (selection without replacement) if order is neglected.*

Proof. There are $(n)_r$ ways to select r elements out of n elements if ordering matters (number of r-tuples over n elements). Since there are $r!$ possible orderings of the selected values, the number we are interested in times $r!$ equals $(n)_r$. Dividing $(n)_r$ by $r!$ completes the proof. ∎

Example 1.6.3. *We use R below to enumerate all possible subsets of size 3 out of a set of size 4. There are ten columns listing these subsets in accordance with $\binom{5}{3} = 20/2$.*

```
# list all possible combinations of 3 out of
# 5 letters
combn(letters[1:5], 3)
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8]
## [1,] "a"  "a"  "a"  "a"  "a"  "a"  "b"  "b"
## [2,] "b"  "b"  "b"  "c"  "c"  "d"  "c"  "c"
## [3,] "c"  "d"  "e"  "d"  "e"  "e"  "d"  "e"
##      [,9] [,10]
## [1,] "b"  "c"
## [2,] "d"  "d"
## [3,] "e"  "e"
```

Example 1.6.4. *In poker, a hand is a subset of 5 cards (order does not matter) out of 52 distinct cards. The cards have face values (1-13) and suits (clubs, spades, hearts, diamonds). There are* $|\Omega| = \binom{52}{5}$ *different hands at poker since this is the number of subsets of size 5 from the 52 distinct cards. The probability that a random hand has five different face values under the classical model is*

$$4^5 \binom{13}{5} / \binom{52}{5} \approx 0.507$$

as face values are chosen in $\binom{13}{5}$ *ways and there are four suits possible for each of the five face values.*

Example 1.6.5. *The number of sequences of length* $p+q$ *containing* p *zeros and* q *ones is* $\binom{p+q}{p}$ *(choosing* p *among* $p+q$ *sequence positions and assigning them to zero values causes the remaining positions to be automatically assigned to one values).*

Example 1.6.6. *Assuming that the U.S. Senate has 60 male senators and 40 female senators, the probability under the classical probability model of selecting an all-male committee of 3 senators is*

$$\mathsf{P}(E) = \frac{|E|}{|\Omega|} = \frac{\textit{number of samples of 3 out of 60 without order and replacement}}{\textit{number of samples of 3 out of 100 without order and replacement}}$$
$$= \frac{\binom{60}{3}}{\binom{100}{3}} = \frac{60 \cdot 59 \cdot 58}{3 \cdot 2} \frac{3 \cdot 2}{100 \cdot 99 \cdot 98} \approx 0.211.$$

Intuitively, if the frequency of all male committees is significantly larger than 21%, we may conclude that the classical model is inappropriate.

Proposition 1.6.6 (Binomial Theorem)**.**

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^{n-k} y^k.$$

Proof. Expanding the expression

$$(x+y)^n = (x+y)(x+y)\cdots(x+y)$$

we see that it contains many additive terms, each corresponding to a pick of x or y from each of the product terms above. Collecting equal additive terms $x^{n-k}y^k$ for $k = 0, \ldots, n$ (the sum of the two exponents must be n) we have $\binom{n}{k}$ possible selections of k choices of y out of n leading to the term $\binom{n}{k} x^{n-k} y^k$. Repeating this argument for all possible $k = 0, \ldots, n$ completes the proof. ∎

The description above corresponds to choosing r elements out of n distinct elements, or alternatively placing n distinct elements into two bins — one with r elements and one with $n-r$ elements. A useful generalization is placing n distinct elements in k bins with r_i elements being placed at bin i, for $i = 1, \ldots, k$.

Proposition 1.6.7. *The number of ways to deposit n distinct objects into k bins with r_i objects in bin i, $i = 1, \ldots, k$, is the multinomial coefficient $n!/(r_1! \cdots r_k!)$.*

Proof. Repeated use of Proposition 1.6.5 shows that the number is

$$\binom{n}{r_1}\binom{n-r_1}{r_2}\binom{n-r_1-r_2}{r_3}\cdots\binom{n-r_1-\cdots-r_{k-2}}{r_{k-1}}.$$

Canceling common factors in the numerators and denominators completes the proof. ∎

Example 1.6.7. *A throw of twelve dice can result in 6^{12} different outcomes. The event A that each face appears twice can occur in as many ways as twelve dice can be arranged in six groups of two each. Assuming the classical probability model, the above proposition implies that*

$$\mathsf{P}(A) = \frac{|A|}{|\Omega|} = \frac{12!/(2^6)}{6^{12}} \approx 0.0034.$$

The following inequality is useful in bounding the probability of complex events in terms of the probability of multiple simpler events.

Proposition 1.6.8 (Boole's Inequality)**.** *For a finite or countably infinite set of events A_i, $i \in C$*

$$\mathsf{P}\left(\bigcup_{i\in C} A_i\right) \leq \sum_{i\in C} \mathsf{P}(A_i).$$

Proof. For two events the proposition holds since $\mathsf{P}(A \cup B) = \mathsf{P}(A) + \mathsf{P}(B) - \mathsf{P}(A \cap B)$ (Principle of Inclusion-Exclusion). The case of a finite number of sets holds by induction. The case of a countably infinite number of sets follows from Proposition E.2.1 in the appendix. ∎

1.7 Probability and Measure Theory*

Definition 1.2.1 appears to be formal, and yet is not completely rigorous. It states that a probability function P assigns real values to events $E \subset \Omega$ in a manner consistent with the three axioms. The problem is that the domain of the probability function P is not clearly specified. In other words, if P is a function $\mathsf{P} : \mathcal{F} \to \mathbb{R}$ from a set $\mathcal{F}$ of subsets of Ω to $\mathbb{R}$, the set $\mathcal{F}$ is not specified. The importance of this issue stems from the fact that the three axioms need to hold for all sets in $\mathcal{F}$.

At first glance this appears to be a minor issue that can be solved by choosing $\mathcal{F}$ to be the power set of Ω: 2^Ω. This works nicely whenever Ω is finite or countably infinite. But selecting $\mathcal{F} = 2^\Omega$ does not work well for uncountably infinite Ω such as continuous spaces. It is hard to come up with useful functions $\mathsf{P} : 2^\Omega \to \mathbb{R}$ that satisfy the three axioms for all subsets of Ω.

A satisfactory solution that works for uncountably infinite Ω is to define $\mathcal{F}$ to be a σ-algebra of subsets of Ω (see Section E.1) that is smaller than 2^Ω. In particular, when $\Omega \subset \mathbb{R}^d$, the Borel σ-algebra (Definition E.5.1) is sufficiently large to include the "interesting" subsets of Ω and yet is small enough to not restrict P too much.

We also note that a probability function P is nothing but a measure μ on a measurable space $(\Omega, \mathcal{F})$ (Definition E.2.1) satisfying $\mu(\Omega) = 1$. In other words, the triplet $(\Omega, \mathcal{F}, \mathsf{P})$ is a measure space (see Definition E.2.1) where $\mathcal{F}$ is the σ-algebra of measurable sets and P is a measure satisfying $\mathsf{P}(\Omega) = 1$. Thus, the wide array of mathematical results from measure theory (Chapter E) and Lebesgue integration (Chapter F.3) are directly applicable to probability theory.

1.8 Notes

Our exposition follows the axiomatic view of probability, promoted by A. Kolmogorov. Alternative viewpoints are available, including the frequency viewpoint ($\mathsf{P}(A)$ is the frequency of A occurring in a long sequence of repetitive experiments) and the subjective viewpoint ($\mathsf{P}(A)$ measures the belief that A will occur).

More information on the basic concepts of probability is available in nearly any probability textbook. One example is Feller's first volume [16], which inspired a generation of probabilists as well as several of the examples in this chapter. Examples of books with rigorous coverage of probability theory are [17, 10, 5, 1, 33, 25]. Elementary exposition that avoids measure theory is available in most undergraduate probability textbooks, such as [48, 36, 14]. More information on combinatorics is available in combinatorics textbooks, for example [35] (undergraduate level) and [42] (graduate level).

1.9 Exercises

1. Extend the argument at the end of Section 1.2 and characterize probability functions on a countably infinite Ω using a sequence of non-negative numbers that sum to one. What is the problem with extending this argument further to uncountably infinite Ω?

2. Can there be a classical probability model on sample spaces that are countably infinite? Provide an example or prove that it is impossible.

3. Complete the proof of Proposition 1.5.1.

4. Describe a sample space consistent with the experiment of drawing a hand in poker. Write the events E corresponding to drawing three aces and drawing a full house (and their sizes $|E|$). What is the event corresponding to the intersection of the two events above, and what is its size and probability under the classical model?

5. Formulate a theory of probability that mirrors the standard theory, with the only difference that the second axiom would be $\mathsf{P}(\Omega) = 2$. How would the propositions throughout the chapter change (if at all)?

6. Show a situation where we have three events that are independent but not mutually independent. Hint: Look for a probability function satisfying

$$\begin{aligned} \mathsf{P}(A) = \mathsf{P}(B) = \mathsf{P}(C) &= 1/3 \\ \mathsf{P}(A \cap B) = \mathsf{P}(A \cap C) = \mathsf{P}(B \cap C) &= 1/9 = \mathsf{P}(A \cap B \cap C). \end{aligned}$$

7. Prove that $\mathsf{P}_E(A) = \mathsf{P}(A \,|\, E)$ is a probability function if $\mathsf{P}(E) \neq 0$.

8. Consider the experiment of throwing three fair six-sided dice independently and observing the results without order. Identify the sample space, and the most and least probable elements of it.

9. Repeat the previous exercise, if the results are observed with order.

10. Generalize Proposition 1.2.6 (Principle of Inclusion-Exclusion) to a union of three sets. Can you further generalize it to a union of an arbitrary number of sets?