# Notes for Math 450
# Lecture Notes 2

Renato Feres

## 1 Probability Spaces

We first explain the basic concept of a *probability space*, $(\Omega, \mathcal{F}, P)$. This may be interpreted as an *experiment* with random outcomes. The set $\Omega$ is the collection of all possible *outcomes* of the experiment; $\mathcal{F}$ is a family of subsets of $\Omega$ called *events*; and $P$ is a function that associates to an event its probability. These objects must satisfy certain logical requirements, which are detailed below. A random variable is a function $X : \Omega \to S$ of the output of the random system. We explore some of the general implications of these abstract concepts.

### 1.1 Events and the basic set operations on them

Any situation where the outcome is regarded as random will be referred to as an *experiment*, and the set of all possible outcomes of the experiment comprises its *sample space*, denoted by $S$ or, at times, $\Omega$. Each possible outcome of the experiment corresponds to a single element of $S$. For example, rolling a die is an experiment whose sample space is the finite set $\{1, 2, 3, 4, 5, 6\}$. The sample space for the experiment of tossing three (distinguishable) coins is

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

where $HTH$ indicates the ordered triple $(H, T, H)$ in the product set $\{H, T\}^3$. The delay in departure of a flight scheduled for 10:00 AM every day can be regarded as the outcome of an experiment in this abstract sense, where the sample space now may be taken to be the interval $[0, \infty)$ of the real line.

Sample spaces may be finite, countably infinite, or uncountable. By definition, a set $A$ is said to be *countable* if it is either finite or has the form $A = \{a_1, a_2, a_3, \cdots\}$. For example, the set of natural numbers, $\mathbb{N} = \{1, 2, 3, \cdots\}$ and the integers, $\mathbb{Z} = \{0, -1, 1, -2, 2, \cdots\}$, are countable sets. So is the set of rational numbers, $\mathbb{Q} = \{m/n : m, n \in \mathbb{Z}, n \neq 0\}$. The latter can be enumerated by listing the pairs $(m, n) \in \mathbb{Z}^2$ in some fashion, for example, as described in figure 1. The set $\mathbb{R}$ of all real numbers is uncountable, i.e., it cannot be enumerated in the form $r_1, r_2, r_3, \ldots$ In fact, if $a$ is any positive number, then the union of all intervals $(r_n - a/2^n, r_n + a/2^n)$ is a set of total length no greater than $2a$, as you can check by adding a geometric series. Since $a$ can be arbitrarily small,

$\{r_1, r_2, r_3, \dots\}$ cannot contain any set of non-zero length. Uncountable sample spaces such as the interval $[0, 1]$ will appear often.

A subset of the sample space is called an *event*. For example, $E = \{1, 3, 5\}$ is the event that a die returns an odd number of pips, and $[0, \pi]$ is the event that the ideal fortune wheel of the first lecture settles on a point of its upper-half part. We say that an event *occurs* if the outcome $x$ of the experiment belongs to $E$. This is denoted $x \in E$.

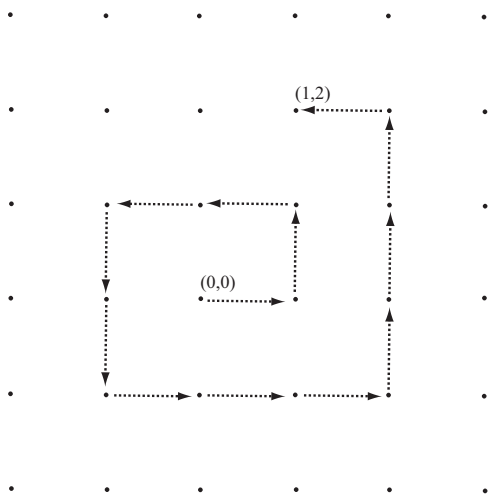


Figure 1: Enumerating pairs of integers

We write $E \subseteq S$ if $E$ is a subset of $S$. For reasons that will be further explained later, we may not want to regard arbitrary subsets of the sample space as events. The subsets to be regarded as events comprise the family $\mathcal{F}$. If $S$ is finite or countably infinite, we often assume that $\mathcal{F}$ contains all the subsets of $S$, but if $S$ is uncountable it will generally be necessary and desirable to restrict the sets that are included in $\mathcal{F}$. (It turns out that any reasonable definition of probability of events that assigns to intervals in $[0, 1]$ a probability equal to its length cannot be defined on all subsets of $[0, 1]$, only on the so-called *measurable sets*. We will see later an example of a non-measurable set.)

The events of $\mathcal{F}$ can be combined using the Boolean operations on sets: unions, intersections, complementation. Recall that the complement of a set $E \subseteq S$, written $E^c$, is the subset of all points in $S$ not contained in $E$. The empty set $\emptyset$ (also denoted $\{\}$) and the full set $S$ are always assumed to be elements of $\mathcal{F}$. The former is the *null* (or *impossible*) event and the latter the *sure* event. The following terms and notations will be used often:

1. The *union* of events $E_1$ and $E_2$, written $E_1 \cup E_2$, is the event that at least one of $E_1$ and $E_2$ occurs. In other words, the outcome of the experiment

lies in at least one of the two sets.

2. The *intersection* of the same two events, written $E_1 \cap E_2$ is the event that both $E_1$ and $E_2$ occur. That is, the outcome of the experiment lies in each of the two sets.

3. The *complement* of an event $E$, denoted $E^c$ or $S \setminus E$, or sometimes $\overline{E}$, is the event that $E$ does not occur. In other words, the outcome of the experiment does not lie in $E$.

4. Two events $E_1$ and $E_2$ are *disjoint*, or *mutually exclusive*, if they cannot both occur, that is, if $E_1 \cap E_2 = \emptyset$. Note that $E \cap E^c = \emptyset$ and $E \cup E^c = S$.

5. A set $E$ is said to be a *countable union*, or union of countably many sets, if there are sets $E_1, E_2, E_3, \cdots$ such that

$$E = E_1 \cup E_2 \cup E_3 \cup \cdots = \bigcup_{k=1}^{\infty} E_k.$$

6. A set $E$ is said to be a *countable intersection*, or the intersection of countably many sets, if there are sets $E_1, E_2, E_3, \cdots$ such that

$$E = E_1 \cap E_2 \cap E_3 \cap \cdots = \bigcap_{k=1}^{\infty} E_k.$$

7. A *partition* of a set $S$ is a collection of disjoint subsets of $S$ whose union is equal to $S$. It is said to be a finite (resp., countable) partition if the collection is finite (resp., countable). If these subsets belong to a family of events $\mathcal{F}$, we say that the partition forms a complete set of mutually exclusive events.

The set of events is required to be closed under (countable) unions and complementation. In other words, given a finite or countably infinite collection of events (in $\mathcal{F}$), their union, intersection and complements are also events (that is, also lie in $\mathcal{F}$). We will often refer to $\mathcal{F}$ as an *algebra of events* in the sample space $S$. The more precise term *$\sigma$-algebra* (sigma algebra), is often used. The next definition captures this in an efficient set of axioms.

**Definition 1.1 (Axioms for events)** The family $\mathcal{F}$ of events must be a *$\sigma$-algebra* on $S$, that is,

1. $S \in \mathcal{F}$

2. if $E \in \mathcal{F}$, then $E^c \in \mathcal{F}$

3. if each member of the sequence $E_1, E_2, E_3, \ldots$ belongs to $\mathcal{F}$, then the union $E_1 \cup E_2 \cup \ldots$ also belongs to $\mathcal{F}$.

When manipulating events abstractly, it is useful to keep in mind the elementary properties of set operations. We enumerate here some of them. The proofs are easy and they can be visualized using the so-called *Venn diagrams*.

**Proposition 1.1 (Basic operations on sets)** *Let $A, B, C$ be arbitrary subsets of a given set $S$. Then the following properties hold:*

$$
\begin{aligned}
A \cup B &= B \cup A && \textit{commutativity}\\
A \cap B &= B \cap A &&\\
(A \cup B) \cup C &= A \cup (B \cup C) && \textit{associativity}\\
(A \cap B) \cap C &= A \cap (B \cap C) &&\\
(A \cup B) \cap C &= (A \cap C) \cup (B \cap C) && \textit{distributive law}\\
(A \cap B) \cup C &= (A \cup C) \cap (B \cup C) &&\\
(A \cup B)^c &= A^c \cap B^c && \textit{DeMorgan's law}\\
(A \cap B)^c &= A^c \cup B^c &&\\
A \cup B &= (A \cap B^c) \cup (A^c \cap B) \cup (A \cap B) && \textit{disjoint union}
\end{aligned}
$$

## 1.2 Probability measures

Abstractly (that is, independently of any interpretation given to the notion of probability), probability is defined as an assignment of a number to each event, satisfying the conditions of the next definition.

**Definition 1.2 (Axioms for probability)** Given a set $S$ and an algebra of events $\mathcal{F}$, a *probability measure* $P$ is a function $P(\cdot)$ that associates to each event $E \in \mathcal{F}$ a real number $P(E)$ having the following properties:

1. $P(S) = 1$;
2. If $E \in \mathcal{F}$, then $P(E) \geq 0$;
3. If $E_1, E_2, E_3, \ldots$ are a finite or countably infinite collection of disjoint events in $\mathcal{F}$, then $P(E_1 \cup E_2 \cup \ldots) = P(E_1) + P(E_2) + P(E_3) + \ldots$.

These axioms can be interpreted as: (1) some outcome in $S$ must occur, (2) probabilities are non-negative quantities, and (3) the probability that at least one of a number of mutually exclusive events will occur is the sum of the probabilities of the individual events. The following properties are easily obtained from the axioms and the set operations of Proposition 1.1.

**Proposition 1.2** *Let $A, B$ be events. Then*

1. $P(A^c) = 1 - P(A)$;
2. $0 \leq P(E) \leq 1$, *for all* $E \in \mathcal{F}$;
3. $P(\emptyset) = 0$;

4. $P(A \cup B) = P(A) + P(B) - P(A \cup B)$;

5. *If* $A \subseteq B$, *then* $P(A) \leq P(B)$.

The sample space $S$, together with a choice of $\mathcal{F}$ and a probability measure $P$, will often be referred to as a *probability space*. When it is necessary to be explicit about all the entities involved, we indicate the probability space by the triple $(S, \mathcal{F}, P)$. Here are a few simple examples.

**Example 1.1 (Tossing a coin once)** Take $S = \{0, 1\}$ and $\mathcal{F}$ the sets of all subsets of $S$:

$$\mathcal{F} = \{\emptyset, \{0\}, \{1\}, S\}\,.$$

The associated probabilities can be set as $P(\{0\}) = p$, $P(\{1\}) = 1 - p$. The probabilities of the other sets are fixed by the axioms: $P(S) = 1$ and $P(\emptyset) = 0$. For a *fair* coin $p = 1/2$.

**Example 1.2 (Spinning a fair fortune wheel once)** Let the state of the idealized fortune wheel of Lecture Notes 1 be represented by the variable $x = \theta/2\pi \in [0, 1]$, keeping in mind that $x = 0$ and $x = 1$ correspond to the same outcome. In this case, we take $S = [0, 1]$ and define $\mathcal{F}$ as the set of all subsets of $[0, 1]$ which are either intervals (of any type: $[a, b], [a, b), (a, b]$, or $(a, b)$ ), or can be obtained from intervals by applying the operations of unions, intersections, and complementation, at most countably many times. To define a probability measure $P$ on $\mathcal{F}$ it is enough to describe what values $P$ assumes on intervals since the other sets in $\mathcal{F}$ are generated by intervals using the set operations. We define $P([a, b]) = b - a$. (This is the so-called *Lebesgue measure* on the $\sigma$-algebra of *Borel subsets* of the interval $[0, 1]$.) This is the most fundamental example of probability space. Other examples can be derived from this one by an appropriate choice of random variable $X : [0, 1] \to S$.

**Example 1.3 (Three-dice game)** A roll of three dice can be described by the probability space $(S, \mathcal{F}, P)$ where $S$ is the set of triples $(i, j, k)$ where $i, j, k$ belong to $\{1, 2, \ldots, 6\}$, that is $(i, j, k) \in \{1, 2, \ldots, 6\}^3$; $\mathcal{F}$ is the family of all subsets of $S$, and $P$ is the probability measure $P(E) = \#(E)/6^3$, where $\#(E)$ denotes the number of elements in $E$.

**Example 1.4 (Random quadratic equations)** If we pick a quadratic equation at random, what is the probability that it will have real roots? To make sense of this question, we first need a model for a random equation. First, label the equation $ax^2 + bx + c = 0$ by its coefficients $(a, b, c)$. I will assume that the coefficients are statistically independent and uniformly distributed on different intervals. Say, $a$ and $c$ are drawn from $[0, 1]$ and $b$ from $[0, 2]$. So to pick a polynomial at random for this choice of probability measure amounts to choosing a point in the parallelepiped $S = [0, 1] \times [0, 2] \times [0, 1]$ with probability function given by the normalized volume

$$P(E) = \frac{1}{2} \int_0^1 \int_0^2 \int_0^1 I_E(a, b, c) da\, db\, dc.$$

The function $I_E$ is the *indicator function* of the set $E$, which takes value 1 if $(a, b, c)$ is a point in $E$ and 0 otherwise. The equations with real roots are those for which the discriminant $b^2 - 4ac$ is positive. The set whose probability we wish to calculate is

$$E = \{(a, b, c) \in S : b \leq 2\sqrt{ac}\}$$

and $P(E) = \text{Vol}(E)/2$. This volume can be calculated analytically by a simple iterated integral and the result can be confirmed by simulation. You will do both things in a later exercise. In this example, $\mathcal{F}$ is generated by parallelepipeds.

## 1.3 Conditional probabilities and independence

Fix a probability space $(S, \mathcal{F}, P)$. We would like to compute the probability that an event $A$ occurs given the knowledge that event $E$ occurs. We denote this probability by $P(A|E)$, the *probability of $A$ given $E$*. Before stating the definition, consider some of the properties that a conditional probability should satisfy: (i) $P(\cdot|E)$ should behave like a probability function on the restricted sample space $E$, i.e., it is non-negative, additive, and $P(E|E) = 1$; (ii) if $A$ and $E$ are mutually exclusive, then $P(A|E) = 0$; in particular, for a general $A$ we must have $P(A|E) = P(A \cap E|E)$; (iii) the knowledge that $E$ has occurred should not affect the relative probabilities of events that are already contained in $E$. In other words, if $A \subseteq E$, then $P(A|E) = cP(A)$ for some constant $c$.

Under the classical interpretation of probability, assuming all outcomes equally likely, (iii) implies that, by conditioning on $E$, individual outcomes in $E$ are equally likely while outcomes not in $E$ have zero probability.

Granted these properties we immediately have: $1 = P(E|E) = cP(E)$ so $c = 1/P(E)$. This leads to the following definition:

**Definition 1.3 (Conditional probability)** Let $A, E \in \mathcal{F}$ be two events such that $P(E) \neq 0$. The *conditional probability* of $A$ given $E$ is defined by

$$P(A|E) = \frac{P(A \cap E)}{P(E)}.$$

Conditional probabilities, given an event $E \subseteq S$, define a probability measure on the new sample space $E$, so the following properties clearly hold:

$$\begin{aligned} P(E|E) &= 1 \\ P(\emptyset|E) &= 0 \\ P(A \cup B|E) &= P(A|E) + P(B|E), \text{ if } A \cap B = \emptyset. \end{aligned}$$

**Example 1.5 (An urn problem)** We have two urns, labeled 1 and 2. Urn 1 contains 2 black balls and 3 red balls. Urn 2 contains 1 black ball and 1 red ball. An urn is chosen at random then a ball is chosen at random from it. Let us represent the sample space by

$$S = \{(1, B), (1, R), (2, B), (2, R)\}.$$

The information given suggests the following conditional probabilities:

$$P(B|1) = 2/5,\ P(R|1) = 3/5,\ P(B|2) = 1/2,\ P(R|2) = 1/2.$$

The conditional probabilities $P(1|R)$, $P(2|B)$, etc., can also be calculated. A simple way to do it is by using Bayes formula given below. We return to this point shortly.

**Example 1.6 (Three candidates running for office)** Let $A, B, C$ be three candidates running for office. As the election is still to happen the winner, $X$, is a random variable. Current opinion assigns probabilities

$$P(X = A) = P(X = B) = 2/5 \text{ and } P(X = C) = 1/5.$$

At some point candidate $A$ drops out of the race. Let $E$ be the event that either $B$ or $C$ wins. Then, $P(A|E) = 0$, and under no further knowledge of the political situation we have

$$P(B|E) = \frac{P(B \cap E)}{P(E)} = \frac{P(B)}{P(B) + P(C)} = \frac{2/5}{2/5 + 1/5} = 2/3.$$

Similarly, we calculate $P(C|E) = 1/3$. Notice that candidate $B$ is still twice as likely to win as $C$, even after we learned of $A$'s withdrawal. (Often in such cases we do have more information than is reflected in the simple event $E$; for example, candidate $A$ could be drawing votes from $C$, so if $A$ withdraws from the race, the relative chances of $C$ increase.)

One often wants to calculate $P(A \cap B)$ from knowledge of conditional probabilities. It follows immediately from definition 1.3 that

$$P(A|B)P(B) = P(A \cap B) = P(B|A)P(A).$$

This implies the following simple but important result.

**Theorem 1.1 (Bayes Theorem)** *For all events $A, B$ such that $P(A), P(B) \neq 0$, we have*

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

Bayes theorem is often used in combination with the next theorem. First, recall that a *partition* of a sample space $S$ is a collection of mutually exclusive events $E_1, E_2, E_3, \ldots$ whose union is $S$. We discard elements of zero probability, so that $P(E_i) > 0$ for each $i$ and the union of the $E_i$ still has probability measure equal to 1. Such "partitions up to a set of zero probability" will be called simply *partitions.*

**Theorem 1.2 (Total probability)** *Suppose that $E_1, E_2, E_3, \ldots$ is a partition of a sample space $S$ and let $E$ be an event. Then*

$$P(E) = \sum_{i=1}^{\infty} P(E|E_i)P(E_i).$$

*Proof.* Since $E = (E \cap E_1) \cup (E \cap E_2) \cup \ldots$, by the additivity axiom of a probability measure it follows that

$$\begin{aligned} P(E) &= P((E \cap E_1) \cup (E \cap E_2) \cup \ldots) \\ &= P(E \cap E_1) + P(E \cap E_2) + \ldots \\ &= P(E|E_1)P(E_1) + P(E|E_2)P(E_2) + \ldots \end{aligned}$$

which is what we wished to show. □

**Example 1.7 (The biology of twins)** There are two types of twins: monozygotic (developed from a single egg), and dizygotic. Monozygotic twins (M) often look very similar, but not always, while dizygotic twins (D) can sometimes show marked resemblance. Therefore, whether twins are monozygotic or dizygotic cannot be settled simply by inspection. However, it is always the case that monozygotic twins are of the same sex, whereas dizygotic twins can have opposite sexes. Denote the sexes of twins by $GG, BB$, or $GB$ (the order is immaterial.) Then, $P(GG|M) = P(BB|M) = P(GB|D) = 1/2$, $P(GG|D) = P(BB|D) = 1/4$, and $P(GB|M) = 0$. It follows that

$$\begin{aligned} P(GG) &= P(GG|M)P(M) + P(GG|D)P(D) \\ &= \frac{1}{2}P(M) + \frac{1}{4}(1 - P(M)) \end{aligned}$$

from which we conclude that

$$P(M) = 4P(GG) - 1.$$

Although it is not easy to be certain whether a particular pair are monozygotic or not, it is easy to discover the *proportion* of monozygotic twins in the whole population of twins simply by observing the sex distribution among *all* twins.

Combining Bayes formula with the theorem on total probability, we obtain the following.

**Corollary 1.1** *Let $E_1, E_2, E_3, \ldots$ be a partition of a sample space $S$ and $E$ an event such that $P(E) > 0$. Then for each $E_j$, the probability of $E_j$ given $E$ is given by*

$$P(E_j|E) = \frac{P(E|E_j)P(E_j)}{P(E|E_1)P(E_1) + P(E|E_2)P(E_2) + \ldots}.$$

*In particular, given any event $A$,*

$$P(A|E) = \frac{P(E|A)P(A)}{P(E|A)P(A) + P(E|A^c)P(A^c)}.$$

**Example 1.8 (The urn problem, part II)** Consider the same situation described in example 1.5. We wish to calculate the probability $P(1|R)$ that urn 1

was selected given that a red ball was drawn at the end of the two stage process. By Bayes theorem,

$$\begin{aligned} P(1|R) &= \frac{P(R|1)P(1)}{P(R|1)P(1) + P(R|2)P(2)} \\ &= \frac{(3/5)(1/2)}{(3/5)(1/2) + (1/2)(1/2)} \\ &= 6/11. \end{aligned}$$

Bayes theorem is usually applied to situations in which we want to obtain the probability of some "hidden" (cause) event given the occurrence of some "surface manifestation" or signal (effect) that is correlated with the event. Here is a classical example.

**Example 1.9 (Clinical test)** Let $A$ be the event that a given person has a disease for which a clinical test is available. Let $B$ be the event that the test gives a positive reading. We may have prior information as to how reliable the test is, so we may already know the conditional probability $P(B|A)$ that if the test is given to a person who is known to have the disease, the reading will be positive, and similarly $P(B|A^c)$, of a positive reading if the person is healthy. We may also know how common or rare the disease is in the population at large, so the "prior" probability, $P(A)$, that a person has the disease may be known. Bayes theorem then gives the probability of having the disease given a positive test reading. You will compute a numerical example later in an exercise.

**Definition 1.4 (Independence of events)** Events $E_1, E_2, \ldots, E_k \in \mathcal{F}$, are said to be *independent* if

$$P(E_1 \cap \cdots \cap E_k) = P(E_1) \cdots P(E_k).$$

Events $E_1, E_2, \ldots$ in an infinite sequence are independent if $E_1, E_2, \ldots, E_k$ are independent for each $k$. In particular, if $P(B) \neq 0$, $A$ and $B$ are independent if and only if $P(A|B) = P(A)$. This means that knowledge of occurrence of $B$ does not affect the probability of occurrence of $A$.

**Example 1.10 (Coin tossing)** A coin is tossed twice. Let $F$ be the event that the first toss is a head and $E$ the event that the two outcomes are the same. Taking $S = \{HH, HT, TH, TT\}$, then $F = \{HH, HT\}$ and $E = \{HH, TT\}$. The probability of $F$ given $E$ is

$$P(F|E) = \frac{P(HH)}{P(HH, TT)} = \frac{1/4}{1/2} = 1/2 = P(E).$$

Therefore, $F$ and $E$ are independent events.

The proof of the following proposition is left as exercise. It uses the notations

$$\begin{aligned} P(E_1, \ldots, E_n) &= P(E_1 \cap \cdots \cap E_n) \\ P(A|E_1, \ldots, E_n) &= P(A|E_1 \cap \cdots \cap E_n). \end{aligned}$$

**Proposition 1.3 (Bayes' sequential formula)** *Let $E_1, E_2, \ldots, E_n$ be events such that $P(E_1 \cap E_2 \cap \cdots \cap E_n) \neq 0$. Then*

$$P(E_1, E_2, \ldots, E_n) = P(E_1)P(E_2|E_1)P(E_3|E_1, E_2)\ldots P(E_n|E_1, \ldots, E_{n-1}).$$

Note that the sequential formula reduces to definition 1.4 when the events are independent.

# 2 Random variables

A random variable is a quantity (typically in $\mathbb{R}$, or $\mathbb{R}^k$) that depends on some random factor. That is, a random variable is a function $X$ from a probability space $S$ to some set of possible values that $X$ can attain. We also require that subsets of $S$ defined in terms of $X$ be events in $\mathcal{F}$. The precise definition is given next.

**Definition 2.1 (Random variables)** Let $(S, \mathcal{F}, P)$ be a probability space. We say that a function $X : S \to \mathbb{R}$ is a random variable if it is *measurable* relative to $\mathcal{F}$ (or, simply, *$\mathcal{F}$-measurable*). This means that for all $a \in \mathbb{R}$,

$$\{s \in S : X(s) \leq a\} \in \mathcal{F}.$$

In other words, the condition $X \leq a$ defines an event. The probability of this event is usually written $P(X \leq a) = P(\{s \in S : X(s) \leq a\})$.

The probability of other events, such as $a \leq X \leq b$, are obtained from the values of $P(X \leq c)$, $c \in \mathbb{R}$, by taking intersections, unions, complements, and employing the axioms of probability measures. For example,

$$P(a < X \leq b) = P(X \leq b) - P(x \leq a).$$

Also note that if an increasing sequence $a_n$ converges to $a$, then

$$P(X < a) = \lim_{n \to \infty} P(X \leq a_n).$$

This follows from the countable additivity of the probability measure (see definition 1.2) and the fact that the union of the sets $\{X \leq a_n\}$ equals $\{X < a\}$. The probability of other events defined by conditions on $X$ can be obtained in a similar way. This is an indication that the function $F(x) = P(X \leq x)$ determines the probability measure $P$ on events defined in terms of $X$.

**Definition 2.2 (Cumulative distribution function)** The function

$$F_X(x) := P(X \leq x)$$

is called the *cumulative distribution function* of the random variable $X$.

The random variable $X : S \to \mathbb{R}$ gives rise to a probability space $(\mathbb{R}, \mathcal{B}, P_X)$, where $\mathcal{B}$ is the $\sigma$-algebra of subsets of $\mathbb{R}$ generated by the intervals $(-\infty, a]$, called the *Borel $\sigma$-algebra*, and $P_X$ is defined as follows: for any $A \in \mathcal{B}$,

$$P_X(A) = P(X \in A).$$

In particular, the probability $P_X$ of an interval such as $(a, b]$ is

$$P_X((a, b]) = F_X(b) - F_X(a).$$

$P_X$ is sometimes called the *probability law* of $X$. Because of this remark, we often do not need to make explicit the probability space $(S, \mathcal{F}, P)$; we only use the set of values of $X$ in $\mathbb{R}$ and the probability law $P_X$ (on Borel sets).

**Definition 2.3 (Independent random variables)** Two random variables $X$ and $Y$ are independent if any two events, $A$ and $B$, *defined in terms of* $X$ and $Y$, respectively, are independent.

Notice that we have not given a formal definition of "being defined in terms of." We will do it shortly. The intuitive meaning of the term suffices for the moment.

**Example 2.1 (Three-dice game)** To illustrate these ideas let us return to the game of rolling three dice. Set $S = \{1, 2, \ldots, 6\}^3$, $\mathcal{F}$ the collection of all subsets of $S$, and for $A \subset S$ define

$$P(A) := \frac{\#A}{\#S},$$

where $\#A$ denotes the number of elements of $A$. (Note: $\#S = 216$.) Thus an elementary outcome is a triple $\omega = (i, j, k)$, where $i, j, k$ are integers from 1 to 6. For example, $(1, 4, 3)$ represents the outcome shown in the next figure.

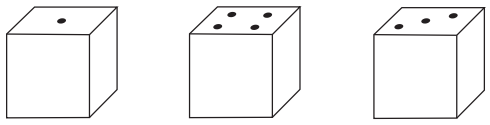

Figure 2: Roll of three dice

Define random variables $X_1, X_2, X_3$ by

$$X_1(i, j, k) = i, \quad X_2(i, j, k) = j, \quad \text{and} \quad X_3(i, j, k) = k.$$

An event $A \in \mathcal{F}$ is defined in terms of $X_2$ if and only if it is the union of sets of the form $X_2 = j$. Note that $P(X_l = u) = 1/6$ for $l = 1, 2, 3$, and $u = 1, 2, \ldots, 6$.

For example: if $A$ is the event $X_1 = i$ and $B$ is the event $X_2 = j$, then $A$ and $B$ are independent since

$$\begin{aligned} P(A \cap B) &= \#\{(i,j,1),\ldots,(i,j,6)\}/6^3 \\ &= 6/6^3 \\ &= (1/6)(1/6) \\ &= P(A)P(B). \end{aligned}$$

A very convenient way to express the idea of events defined in terms of a random variable $X$ is through the $\sigma$-algebra $\mathcal{F}_X$ of the next definition.

**Example 2.2 (Buffon's needle problem)** Buffon was a French naturalist who proposed the famous problem of computing the probability that a needle intersects a regular grid of parallel lines. We suppose that the needle has length $l$ and the lines are separated by a distance $a$, where $a > l$.

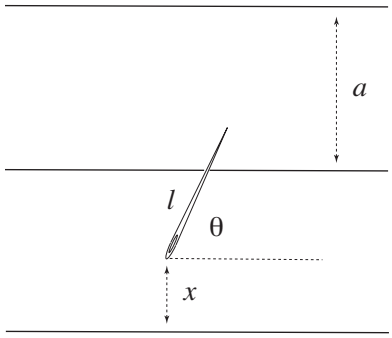


Figure 3: Buffon's needle problem.

To solve the problem we first need a mathematical model of it as a function $X$ from some appropriate probability space into $\{0,1\}$, where 1 represents 'intersect' and 0 represents 'not intersect.' We regard the grid as perfectly periodic and infinite. The position of the needle in it can be specified by the vertical distance, $x$, of the end closest to the needle's eye to the first line below it, and the angle, $\theta$, between 0 and $2\pi$ that the needle makes with a line parallel to the grid. It seems reasonable to regard these two parameters as independent and to assume that $\theta$ is uniformly distributed over $[0, 2\pi]$ and $x$ is uniformly distributed over $[0, a]$. So our mathematical definition of throwing the needle over the grid is to pick a pair $(x, \theta)$ from the rectangle $S = [0, a] \times [0, 2\pi]$ with the uniform distribution, i.e., so that the probability of an event $E \subset S$ is

$$P(E) = \frac{1}{2\pi a} \int_0^a \int_0^{2\pi} I_E(x,\theta) dx d\theta$$

where $I_E$ is the indicator function of $E$.

In an exercise you will show how to express the outcome of a throw of the needle by a function $X$ and will calculate the probability of $X = 1$. This turns out to be

$$P(X = 1) = \frac{2}{\pi}\frac{l}{a}.$$

If $a = 2l$ we have the probability $1/\pi$ of intersection, providing an interesting (if not very efficient) way to compute approximations to $\pi$.

**Example 2.3 (A sequence of dice rolls)** We show here how to represent the experiment of rolling a die infinitely many times as a random variable $X$ from $[0, 1]$ into the space $S$ of all sequences of integers from 1 to 6. (If you find the discussion a bit confusing, simply skip it. It won't be needed later.) We first make a few remarks about representing numbers between 0 and 1 in base 6. Let $K = \{0, 1, 2, 3, 4, 5\}$ and $I_k = (k/6, (k+1)/6]$, if $k \neq 0$, and $I_0 = [0, 1/6]$. These are the six intervals on the top of Figure 4.

A number $x \in [0, 1]$ has expansion in base 6 of the form $0.\omega_1\omega_2\omega_3\ldots$, where $\omega_i \in K$, if we can write

$$x = \frac{\omega_1}{6} + \frac{\omega_2}{6^2} + \frac{\omega_3}{6^3} + \ldots.$$

The representation is in general not unique; for example, $0.1000\cdots = 0.0555\ldots$, since

$$\frac{1}{6} = \frac{0}{6} + \frac{5}{6^2} + \frac{5}{6^3} + \frac{5}{6^4} + \ldots.$$

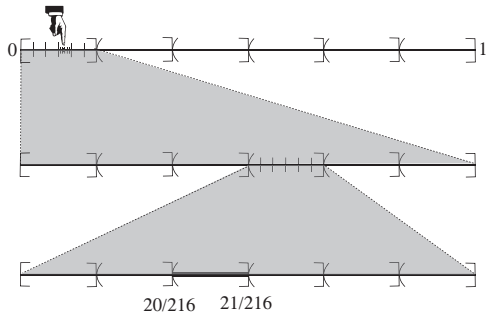


Figure 4: Six-adic interval associated to the dice roll event $(1, 4, 3)$.

To make the choice of $\omega_i$ for a given $x$ unique, we can proceed as follows. First, for any $x > 0$ (not necessarily less than 1) define $\eta(x)$ as the non-negative integer (possibly 0) such that $\eta(x) < x \leq \eta(x)+1$. Now, define a transformation $T : [0, 1] \to [0, 1]$ by

$$T(x) = \begin{cases} 6x - \eta(6x) & \text{if } x > 0 \\ 0 & \text{if } x = 0. \end{cases}$$

A function from $[0,1]$ to the set of sequences $(\omega_1, \omega_2, \dots)$ can now be defined as follows. For each $n = 1, 2, \dots$, write:

$$\omega_n = X_n(x) = k+1 \Leftrightarrow T^{n-1}(x) \in I_k.$$

In other words, the value of, say, $\omega_9$ is 4, if and only if the 8th iterate of $T$ applied to $x$, $T^8(x)$ falls into the fifth interval, $I_4 = (4/6, 5/6]$. It is easy to see why this works. Observe that:

$$\begin{aligned} x = T^0(x) &= 0.\omega_1\omega_2\omega_3\dots && \in I_{\omega_1} \\ T^1(x) &= 0.\omega_2\omega_3\omega_4\dots && \in I_{\omega_2} \\ T^2(x) &= 0.\omega_3\omega_4\omega_5\dots && \in I_{\omega_3} \\ &\vdots && \vdots \end{aligned}$$

We can now define our mathematical model of infinitely many dice rollings as follows. Choose a point $x \in [0,1]$ at random with uniform probability distribution. Write its six-adic expansion, $x = 0.\omega_1\omega_2\omega_3\dots$, where $\omega_n$ is chosen by the rule: $T^{n-1}(x) \in I_{\omega_n}$. Then the infinite vector

$$X(x) = (\omega_1 + 1, \omega_2 + 1, \omega_3 + 1, \dots)$$

can be regarded as the outcome of rolling a die infinitely many times. One still needs to justify that the random variables $X_1, X_2, \dots$ are all independent and that $P(X_n = j) = 1/6$ for each $j = 1, 2, \dots, 6$. This will be checked in one of the exercises.

**Definition 2.4 (Events defined in terms of a random variable)** If $X$ is a random variable on a probability space $(S, \mathcal{F}, P)$, let $\mathcal{F}_X$ be the smallest subset of $\mathcal{F}$ that contains the events $X \le a$, $a \in \mathbb{R}$, and satisfies the axioms for a $\sigma$-algebra of events, namely: (i) $S \in \mathcal{F}_X$, (ii) the complement of any event in $\mathcal{F}_X$ is also in $\mathcal{F}_X$, and (iii) the union of countably many events in $\mathcal{F}_X$ is also in $\mathcal{F}_X$. We call $\mathcal{F}_X$ the *$\sigma$-algebra generated by $X$*. An event in $\mathcal{F}$ is said to be *defined in terms of $X$* if it belongs to $\mathcal{F}_X$.

More generally, we say that random variables $X_1, X_2, \dots$ are independent if sets $A_1, A_2, \dots$ defined in terms of the respective $X_i$ are independent.

**Definition 2.5 (I.i.d.)** Random variables $X_1, X_2, X_3, \cdots$ are said to be i.i.d. (independent, identically distributed) if they are independent and, for each $a \in \mathbb{R}$, the probabilities $P(X_i \le a)$ are the same for all $i$, that is, they have the same cumulative distribution functions $F_{X_i}(x)$.

As indicated above, the $X_1, X_2, \dots$, in the sequence of die-rollings example are i.i.d., random variables.

# 3 Continuous random variables

The previous discussion about probability spaces, conditional probability, independence, etc., are completely general, although we have for the most part applied the concepts to discrete random variables and finite sample spaces. A random variable $X$ on a probability space $(\Omega, \mathcal{F}, P)$ is said to be *discrete* if with probability one it can take only a finite or countably infinite number of values. For discrete random variables the probability measure $P_X$ is concentrated on a set $\{a_1, a_2, \dots\}$ in the following sense: if $A$ is a subset of $\mathbb{R}$ that does not contain any of the $a_i$, then $P_X(A) = 0$. The cumulative distribution function $F_X(x)$ is then a step-wise increasing, discontinuous function:

$$F(x) = \sum_{n \le x} P_X(a_n).$$

We consider now, more specifically, the continuous case. The term "continuous" is not used here in the strict sense of calculus. It is meant in the sense that the probability of single outcomes is zero. For example, the probability of the event $\{a\}$ in $[0, 1]$ with the uniform probability measure is zero.

A more restrictive definition will actually be used. We call $X : S \to \mathbb{R}$ a *continuous random variable* if the probability $P_X$ is derived from a density function: this is a function $f_X : \mathbb{R} \to [0, \infty)$ (not required to be continuous) for which

$$F_X(x) = \int_{-\infty}^{x} f_X(s) ds,$$

where $F_X$ is the cumulative distribution function of $X$. (Our continuous random variables should more properly be called *absolutely continuous*, in the language of measure theory.)

**Example 3.1 (Normal random variables)** An important class of a continuous random variables are the *normal* (or *Gaussian*) random variables. A random variable $X$ is normal if there are constants $\sigma$ and $\mu$ such that

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}.$$

We will have much to say about normal random variables later. The same is true about the next example.

**Example 3.2 (Exponential random variables)** A random variable $X$ is said to have an *exponential distribution with parameter* $\lambda$ if its probability density function $f_X(x)$ has the form

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \ge 0 \\ 0 & \text{if } x < 0. \end{cases}$$

It is also possible to have mixed-type random variables with both continuous and discrete parts. The abstract definitions given above apply to all cases.

## 3.1 Two or more continuous random variables

We may also consider random variables taking values in $\mathbb{R}^k$, for $k > 1$. We discuss the case $k = 2$ to be concrete, but higher dimensions can be studied in essentially the same way.

A random variable $X = (X_1, X_2) : S \to \mathbb{R}^2$, on a probability space $(S, \mathcal{F}, P)$, is a *continuous* random variable if there is a function $p(x_1, x_2)$ such that the probability of the event $X \in E$, where $E$ is a subset of $\mathbb{R}^2$, is given by the double integral of $p$ over $E$:

$$P_X(E) = \iint_E p(x_1, x_2) dx_1 dx_2.$$

Strictly speaking, $E$ should be a *measurable* element in the $\sigma$-algebra $\mathcal{B}$ of Borel sets generated by parallelepipeds, which are the higher dimensional counterpart to intervals. We will not worry about such details since the sets we are likely to encounter in this course or elsewhere are of this type. Also, the continuous random variables we will often encounter have differentiable density function $p(x_1, x_2)$.

The function $p(x_1, x_2)$ is called the *(joint) probability density function*, abbreviated p.d.f., of the vector-valued random variable $X$. The p.d.f. must satisfy the conditions $p(x_1, x_2) \geq 0$ and

$$\iint_{\mathbb{R}^2} p(x_1, x_2) dx_1 dx_2 = 1.$$

The components $X_1$ and $X_2$ of $X$ will also be continuous random variables, if $X$ is. Their probability density functions will be written $p_1(x_1)$ and $p_2(x_2)$. Given a subset $E$ on the real line, the event $X_1 \in E$ has probability

$$P_{X_1}(E) = P((X_1, X_2) \in E \times \mathbb{R}) = \int_{-\infty}^{\infty} \int_E p(x_1, x_2) dx_1 dx_2.$$

Therefore,

$$p_1(x_1) = \int_{-\infty}^{\infty} p(x_1, x_2) dx_2.$$

Similarly,

$$p_2(x_2) = \int_{-\infty}^{\infty} p(x_1, x_2) dx_1.$$

**Definition 3.1 (Conditional density)** The *conditional density function* of $x_2$ given $x_1$ is defined as

$$p(x_2|x_1) = p(x_1, x_2)/p_1(x_1).$$

The *conditional probability* of $X_2 \in E$, where $E$ is a subset of $\mathbb{R}$, *given* $X_1 = a$ is defined by the limit:

$$P(X_2 \in E | X_1 = a) = \lim_{h \to 0} \frac{P((X_1, X_2) \in [a, a + h] \times E)}{P(X_1 \in [a, a + h])}.$$

**Proposition 3.1** *Using the same notation as in definition 3.1, we have*

$$P(X_2 \in E | X_1 = a) = \int_E p(y|a)dy$$

*A similar formula holds for conditioning $X_1$ on an outcome of $X_2$.*

*Proof.* Using the fundamental theorem of calculus at the second step and the above definitions we obtain:

$$\begin{aligned} P(X_2 \in E | X_1 = a) &= \lim_{h \to 0} \frac{\int_E \int_a^{a+h} p(x_1, x_2) dx_1 dx_2}{\int_a^{a+h} p_1(x_1) dx_1} \\ &= \lim_{h \to 0} \frac{\int_E \frac{1}{h} \int_a^{a+h} p(x_1, x_2) dx_1 dx_2}{\frac{1}{h} \int_a^{a+h} p_1(x_1) dx_1} \\ &= \frac{\int_E p(a, x_2) dx_2}{p_1(a)} \\ &= \int_E p(a, x_2)/p_1(a) dx_2 \\ &= \int_E p(x_2|a) dx_2. \end{aligned}$$

□

The next proposition is a simple consequence of the definitions.

**Proposition 3.2 (Independence)** *Two continuous random variables $X_1$, $X_2$ are independent if and only if the probability density of $X = (X_1, X_2)$ decomposes into a product:*

$$p(x_1, x_2) = p_1(x_1) p_2(x_2).$$

## 3.2 Transformations of continuous random variables

It is useful to know how the probability density functions transform if the random variables change in simple ways. Let $X = (X_1, \ldots, X_n)$ be a random vector with the probability density function $f_X(x)$, $x \in \mathbb{R}^n$. Define the random variable $Y = g(X)$, where $g : \mathbb{R}^n \to \mathbb{R}^n$. More explicitly,

$$\begin{aligned} Y_1 &= g_1(X_1, \ldots, X_n) \\ Y_2 &= g_2(X_1, \ldots, X_n) \\ &\vdots \\ Y_n &= g_n(X_1, \ldots, X_n). \end{aligned}$$

Assume that $g$ is a differentiable coordinate change: it is invertible, and its Jacobian determinant $Jg(x) = \det Dg(x)$ of $g$ is non-zero at all points $x \in \mathbb{R}^n$. Here $Dg(x)$ denotes the matrix

$$Dg(x)_{ij} = \frac{\partial g_i}{\partial x_j}(x).$$

**Theorem 3.1 (Coordinate change)** *Let $X$ be a continuous random variable with probability density function $f_X(x)$. Let $g : \mathbb{R}^n \to \mathbb{R}^n$ be a coordinate change, then $Y = g(X)$ has probability density function*

$$f_Y(y) = f_X(g^{-1}(y))|\det(Dg(g^{-1}(y)))|^{-1}$$

The theorem can be proved by multivariable calculus methods. We omit the proof. An important special case is when $g$ is an invertible affine transformation. This means that for some invertible matrix $A$ and constant vector $b \in \mathbb{R}^n$,

$$Y = AX + b,$$

assuming here that $X$, $b$, and $Y$ are column vectors and $AX$ denotes matrix multiplication. In this case, $Dg(x) = A$.

**Corollary 3.1 (Invertible affine transformations)** *Suppose that*

$$Y = AX + b,$$

*where $A$ is an invertible matrix. If $f_X(x)$ is the probability density function of $X$, and $f_Y(y)$ is the probability density function of $Y = g(X)$, then*

$$f_Y(y) = f_X(A^{-1}(y - b))/|\det A|.$$

*If $n = 1$, then the random variable $Y = aX + b$ has probability density function*

$$f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right).$$

**Example 3.3** We say that $Z$ is a *standard normal* random variable if its probability density function is

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}.$$

If $X = \sigma Z + \mu$, then $X$ is a normal random variable with probability density function

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}.$$

A different kind of transformation is a coordinate projection, $\Pi : \mathbb{R}^n \to \mathbb{R}^m$, such that $\Pi(x_1, \ldots, x_n) = (x_1, \ldots, x_m)$, where $m \leq n$. We have already considered this earlier when deriving the probability density $p_1(x_1)$ from $p(x_1, x_2)$. The same argument applies here to show the following.

**Proposition 3.3 (Projection transformation)** *Let $m < n$ and $Y = \Pi(X)$ where $X$ is a random variable in $\mathbb{R}^n$ with probability density function $f_X(x)$ and $\Pi : \mathbb{R}^n \to \mathbb{R}^m$ is the coordinate projection onto the first $m$ dimensions. If $f_Y(y)$ is the probability density function of $Y$, then*

$$f_Y(y_1, \ldots, y_m) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_X(y_1, \ldots, y_m, x_{m+1}, \ldots, x_n) dx_{m+1} \ldots dx_n.$$

**Example 3.4 (Linear combinations of random variables)** Consider independent random variables $X_1, X_2 : S \to \mathbb{R}$ with probability density functions $f_1(x)$ and $f_2(x)$. Given constants $a_1$ and $a_2$, we wish to find the probability density function of the linear combination

$$Y = a_1 X_1 + a_2 X_2.$$

We may as well assume that $a_1$ is different from 0 since, otherwise, this problem would be a special case of corollary 3.1, which shows what to do when we multiply a random variable by a nonzero number. By the same token, we may take $a_1 = 1$. So suppose that $Y = X_1 + aX_2$, where $a$ is arbitrary. Let $A : \mathbb{R}^2 \to \mathbb{R}^2$ be the invertible linear transformation with matrix

$$A = \begin{pmatrix} 1 & a \\ 0 & 1 \end{pmatrix}$$

and $\Pi : \mathbb{R}^2 \to \mathbb{R}$ the linear projection to the first coordinate. Also introduce the column vector $X = (X_1, X_2)^t$, where the upper-script indicates 'transpose.' Then $Y$ is the image of $X$ under the linear map $\Pi A$:

$$Y = \Pi A X.$$

We can thus solve the problem in two steps using corollary 3.1 and proposition 3.3. The joint density function for $X$ is

$$f_X(x_1, x_2) = f_1(x_1) f_2(x_2)$$

since we are assuming that the random variables are independent. Then

$$\begin{aligned} f_Y(y) &= f_{\Pi AX}(y) \\ &= \int_{-\infty}^{\infty} f_{AX}(y, s) ds \\ &= \int_{-\infty}^{\infty} f_X(y - as, s) ds \\ &= \int_{-\infty}^{\infty} f_1(y - as) f_2(s) ds. \end{aligned}$$

Therefore, the probability density function of $Y = X_1 + aX_2$ is

$$f_Y(y) = \int_{-\infty}^{\infty} f_1(y - as) f_2(s) ds.$$

A case of special interest is when $a = 1$. In this case $f_Y(y)$ is the so-called *convolution* of $f_1(x)$ and $f_2(x)$. We state this special case as a separate proposition.

**Definition 3.2 (Convolution)** The *convolution* of two functions $f(x)$ and $g(x)$ is the function $f * g$ given by

$$(f * g)(x) = \int_{-\infty}^{\infty} f(x - s) g(s) ds.$$

It is not difficult to show that if $f_1(x)$ and $f_2(x)$ are probability density functions, then the convolution integral exists (as an improper integral) and the result is a non-negative function with total integral equal to 1. Therefore, it is also a probability density function.

**Proposition 3.4 (Sum of two independent random variables)** *The sum of independent random variables $X_1$ and $X_2$ having probability density functions $f_1(x)$, $f_2(x)$ is a random variable with probability function $f(x) = (f_1 * f_2)(x)$.*

**Example 3.5** Define for a positive number $\delta$ the function

$$g_\delta(x) = \frac{1}{\delta} I_{[0,\delta]}(x),$$

where $I_{[0,\delta]}(x)$ is the *indicator function* of the interval $[0, \delta]$. I.e., $g_\delta(x)$ is $1/\delta$ if $x$ lies in the interval and 0 if not. Then $g_\delta(x)$ is a non-negative function with total integral 1, so we can think of it as the probability density function of a continuous random variable $X$ (a random point in $[0, \delta]$ with uniform probability distribution). Suppose we pick two random points, $X_1, X_2$, in $[0, \delta]$ independently with the same distribution $g_\delta(x)$. We wish to compute the probability distribution of the difference $Y = X_1 - X_2$. Note that $Y$ has probability 0 of falling outside the interval $[-\delta, \delta]$, so $f_Y(y)$ must be 0 outside this symmetric interval. Explicitly,

$$\begin{aligned} f_Y(y) &= \int_{-\infty}^{\infty} g_\delta(y+s) g_\delta(s) ds \\ &= \frac{1}{\delta^2} \int_{-\infty}^{\infty} I_{[0,\delta]}(y+s) I_{[0,\delta]}(s) ds \\ &= \frac{1}{\delta^2} \int_{-\infty}^{\infty} I_{[-y,-y+\delta]}(s) I_{[0,\delta]}(s) ds \\ &= \frac{1}{\delta^2} \int_{-\infty}^{\infty} I_{[-y,-y+\delta]\cap[0,\delta]}(s) ds \\ &= \frac{1}{\delta^2} \text{ length of the interval } [-y, -y+\delta] \cap [0, \delta]. \end{aligned}$$

A little thought gives the following expression for the length of the intersection:

$$f_Y(y) = \begin{cases} 0 & \text{if } |y| \geq \delta \\ (\delta - |y|)/\delta^2 & \text{if } |y| < \delta. \end{cases}$$

The graph of this function is given below.

## 3.3 Bayes theorem for continuous random variables

Bayes theorem 1.1 holds in full generality. In the continuous case it can be expressed in terms of the density functions:

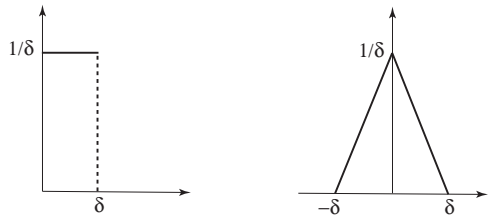


Figure 5: The graph on the left-hand side represents $g_\delta(x)$. On the right is the graph of the probability density function $f_Y(x)$.

**Theorem 3.2 (Bayes theorem for continuous random variables)** *Using the same notations as above in this section, we have*

$$p(x_2|x_1) = p(x_1|x_2)p_2(x_2)/p_1(x_1).$$

As in the discrete random variable examples, Bayes theorem is often used in combination with the total probability formula 1.2. In terms of the densities of continuous random variables, this formula is expressed as an integral:

**Proposition 3.5 (Total probability)** *The following equality, and a similar one for $p_2(x_2)$, hold:*

$$p_1(x_1) = \int_{-\infty}^{\infty} p(x_1|x_2)p_2(x_2)dx_2.$$

Note that $p(y|x)$ is proportional to $p(x|y)p_2(y)$ as we vary $y$ for a fixed $x$. This means that, to find the most likely $y$ given the knowledge of $x$, we need only maximize the expression $p(x|y)p_2(y)$ without taking into account the denominator in Bayes formula.

**Example 3.6** An experimental set-up emits a sequence of blips at random times. The experimenter registers the occurrence of each blip by pressing a button. Suppose that the time $T$ between two consecutive blips is an exponentially distributed random variable with parameter $\lambda$. This means that it has probability density function

$$p_T(t) = \lambda e^{-\lambda t}.$$

(We will study exponentially distributed random variables at great length later in the course. For now, it is useful to keep in mind that the mean value of $T$ is $1/\lambda$, so $\lambda$ is the overall "frequency" at which the blips are emitted.) The time it takes the experimenter to respond to the emission of a blip is also random. A crude model of the experimenter's delayed reaction is this: if a blip occurs at time $s$, then the time it is registered is $s + \tau$, where $\tau$ is a random variable with probability density function:

$$g_\delta(u) = \begin{cases} 1/\delta & \text{if } u \in [0, \delta] \\ 0 & \text{otherwise.} \end{cases}$$

We also assume independent delay times. Let $S_1, S_2$ be the times of two consecutive blips, $\tau_1$, $\tau_2$ the respective response delays, $T = S_2 - S_1$, and $\tau = \tau_2 - \tau_1$. Then the registered time difference between the two events is

$$R = S_2 + \tau_2 - (S_1 + \tau_1) = T + \tau.$$

Notice that, if we knew that the time between two consecutive blips was $t$, then the registered time difference, $R = t + \tau$, would have probability density

$$p(r|t) = f_\tau(r - t)$$

according to corollary 3.1, where $f_\tau$ is obtained from $g_\delta$ as in example 3.5. The problem is to learn about the inter-emission intervals, $T$, given the times the blips are actually registered. A reasonable estimate of $T$ is the value $t$ that maximizes $p(t|r)$. Using the result of example 3.5 and Bayes theorem in the form $p(t|r) \propto p(r|t)p(t)$ we obtain:

$$p(t|r) \propto \begin{cases} 0 & \text{if } |r - t| \geq \delta \\ \lambda e^{-\lambda t}(\delta - |r - t|)/\delta^2 & \text{if } |r - t| < \delta. \end{cases}$$

($a \propto b$ means that $a$ is *proportional to* $B$.) The maximum depends on the relative sizes of the parameters. For concreteness assume that $\delta$ is less than $r$. (The measurement error is small compared to the measured value.) Then

$$t = \begin{cases} r & \text{if } 1/\lambda > \delta \\ r - \delta + 1/\lambda & \text{otherwise.} \end{cases}$$

I leave the details of the calculation as an exercise for you.

# 4 Exercises and Computer Experiments

Familiarize yourself with the Matlab programs of Lecture Notes 1. They may be useful in some of the problems below. All simulations for now will be limited to discrete random variables. We will deal with simulating continuous random variables in the next set of notes.

## 4.1 General properties of probability spaces

Whenever solving a problem about finite probability spaces, be explicit about the sample space $S$ and the subsets of $S$ corresponding to each event. "Plausible" chatty arguments can be misleading. Solving a problem in finite probability often reduces to counting the number of elements in a set.

**Exercise 4.1 (Galileo's problem)** A game consists of rolling three dice and counting the total number of pips. A gambler famously asked Galileo to solve the following problem: Is it more likely to obtain a 9 or a 10? Solve this problem by direct counting. Confirm your answer by carrying out this experiment 50000 times and counting the relative frequencies of 9 and 10. (Suggestion: use our `samplefromp` function.)

**Exercise 4.2** A coin is tossed three times. What is the probability that exactly two heads occur, given that :

1. The first outcome was a head.
2. The first outcome was a tail.
3. The first two outcomes were heads.
4. The first two outcomes were tails.
5. The first outcome was a head and the third outcome was a head.

In each case, be explicit about the subsets of the sample space $S$ involved in the problem.

**Exercise 4.3** What is the probability that a family of two children has

1. Two boys given that it has at least one boy.
2. Two boys given that the first child was a boy.

As always, be explicit about the sample space and the sets involved.

**Exercise 4.4** Check that the random variables $X_1, X_2, X_3, \ldots$ defined in the three-dice example are i.i.d.

The following program simulates the experiment of tossing a coin $k$ times using the main idea behind example 2.3. A similar program could be written for the die-rolling experiment.

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
function y=kcoins(k)
%Input   - positive integer k
%
%Output  - vector of 0s and 1s of legnth k
%          This simulates flipping k unbiased coins
%          independently by using the binary digits
%          of a single U(0,1) random number.
y=[];
x=rand;
for i=1:k
    y=[y (i*x - floor(i*x)<0.5)];
end
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
```

**Exercise 4.5 (Random quadratic equations)** We will say that a quadratic equation $ax^2 + bx + c = 0$ is a random equation if the real coefficients $a, b, c$ are random variables. Suppose that the coefficients are independent, that $a, c$ are uniformly distributed in $[0, 1]$ and $b$ is uniformly distributed in $[0, 2]$. What

is the probability that a random quadratic equation has real roots? Confirm your answer by simulation. (Pick a sample of 50000 random polynomials and determine the relative sample frequency of polynomials with non-negative discriminant.) Show your program.

**Exercise 4.6 (Crossing the street)** Suppose that, on a certain street, with probability $p = 1/3$ a car will pass at each second $k = 0, 1, 2, \ldots$. (That is, the probability that a car will pass at second $k$ is $1/3$, for all $k$.) A pedestrian who starts waiting at time 0, will need 5 seconds without a car passing in order to cross the street. Write a Matlab program to simulate this process 10000 times, calculating the proportion of times a pedestrian will have to wait before being able to cross. Can you guess the exact value?

**Exercise 4.7** Prove proposition 1.2

**Exercise 4.8** Let $E_n$, $n = 1, 2, \ldots$, be a sequence of events. Prove:

1. $P(E_1 \cup E_2 \cup \ldots) \leq P(E_1) + P(E_2) + \ldots$
2. if $E_n \subseteq E_{n+1}$ for all $n$, then

$$P\left(\bigcup_{i=1}^{\infty} E_i\right) = \lim_{n\to\infty} P(E_n).$$

**Exercise 4.9** Let $(\Omega, \mathcal{F}, P)$ be a probability space. Show that:

1. $P(\emptyset) = 0$;
2. Given $A_1, A_2, A_3, \cdots$ in $\mathcal{F}$, then $\bigcap_k A_k$ is also in $\mathcal{F}$;
3. If $A, B \in \mathcal{F}$ satisfy $A \subset B$, then $P(A) \leq P(B)$;
4. If $A_1, A_2, A_3, \cdots$ are elements of $\mathcal{F}$ (not necessarily disjoint), then

$$P\left(\bigcup_k A_k\right) \leq \sum_k P(A_k).$$

A hint: for part 3 of the exercise, write $B$ as a disjoint union: $B = A \cup (B \setminus A)$. For part 4, note that $A_1 \cup A_2 \cup \cdots$ equals the union $A_1' \cup A_2' \cup \cdots$, where

$$A_1' = A_1,\ A_2' = A_2 \setminus A_1,\ A_3' = A_3 \setminus (A_1 \cup A_2),\ \ldots.$$

**Exercise 4.10** Prove that the set of all rational numbers in $[0, 1]$ is an event in the $\sigma$-algebra of Borel sets $\mathcal{B}$ and that it has probability is 0.

## 4.2 Bayes theorem

**Exercise 4.11 (Clinical test)** In a large city, a certain disease is present in about one out of 10000 persons. A program of testing is to be carried out using a clinical test which gives a positive reading with probability 98% for a diseased person, and with probability 1% for a healthy person. What is the probability that a person who has a positive reading actually has the disease?

**Exercise 4.12 (The biology of twins)** Example 1.7 was based on the assumption that births of boys and girls occur equally frequently, and yet it is known that fewer girls are born than boys. Suppose that the probability of a girl is $p$, so that

$$\begin{array}{lll} P(GG|M) = p & P(BB|M) = 1-p & P(GB|M) = 0 \\ P(GG|D) = p^2 & P(BB|D) = (1-p)^2 & P(GB|D) = 2p(1-p). \end{array}$$

Find the proportion of monozygotic twins in the whole population of twins in terms of $p$ and the sex distribution among all twins.

For the next problem, consider the following situation. John claims that he can predict whether the stock market will end the day up or down with 80% accuracy. Gary is skeptical of John's claim and wants to put his powers of prediction to a test. Denote by $H_q$ the hypothesis that John can determine the daily up and down movement of the stock market with $q \times 100\%$ accuracy. For simplicity, we consider a finite set of values $q = k/10$, $k = 0, 1, 2, \ldots, 10$. For example, $H_{0.5}$ is the hypothesis that John is making entirely random guesses, and $H_{0.8}$ is the hypothesis that John is correct in his claim. Gary's skepticism is quantified by his assigning a high likelihood to hypothesis $H_{0.5}$. Let $P(q)$ denote the probability, in Gary's estimation, of hypothesis $H_q$. This is a probability measure on the set of hypotheses, so

$$P(0.0) + P(0.1) + \cdots + P(0.9) + P(1.0) = 1.$$

Now suppose that John is asked to make a forecast. Call $X = C$ the event that he makes a correct one-day forecast, and $X = \overline{C}$ that he gets it wrong. The problem is to know how Gary should modify his initial estimation $P(q)$ of the likelihood of each $H_q$ given the outcome of the random variable $X$. Let $P(q|C)$ be Gary's modified estimation of the likelihood of $H_q$, given that John gets the one-day prediction right, and $P(q|\overline{C})$ the corresponding probability if he gets it wrong. By definition, the probability that $C$ will occur given that hypothesis $H_q$ is true is $P(C|q) = q$. Then by Bayes theorem (and denoting $q_k = k/10$),

$$P(q|C) = \frac{qP(q)}{\sum_{k=0}^{10} q_k P(q_k)}, \quad \text{and} \quad P(q|\overline{C}) = \frac{(1-q)P(q)}{\sum_{k=0}^{10}(1-q_k)P(q_k)}.$$

(Notice that $P(0|C) = 0 = P(1|\overline{C})$, as it should be.) The $P(q_k)$ represent Gary's *prior* probabilities, and $P(q_k|X)$ his *posterior* probabilities after some experimental evidence is offered. These equations can be iterated: given a

sequence of observations, $X_1, X_2, X_3, \ldots$, each giving $C$ or $\overline{C}$, we produce a sequence of probability distributions $P_1, P_2, \ldots$ on the set of hypothesis $H_q$, describing the evolution of Gary's belief in the predictive power of John. In the following exercise you will produce a simulation of this situation.

**Exercise 4.13** Suppose that hypothesis $H_{0.6}$ is correct. (John has some ability to predict market movement, but not as great as he claims.) Simulate a sample sequence $X_1, X_2, X_3, \ldots, X_{100}$ of outcomes $\{C, \overline{C}\}$ then use it to obtain the probability vectors $P_1, P_2, \ldots$. Assume $P_0(q) = a[0.5^{1/2} - (q - 0.5)^{1/2}]$ for the initial probabilities, where $a$ is a normalizing constant. On the same coordinate system, plot the graphs of $P_0$ , $P_{100}$, and $P_{1000}$, as a functions of $q$.

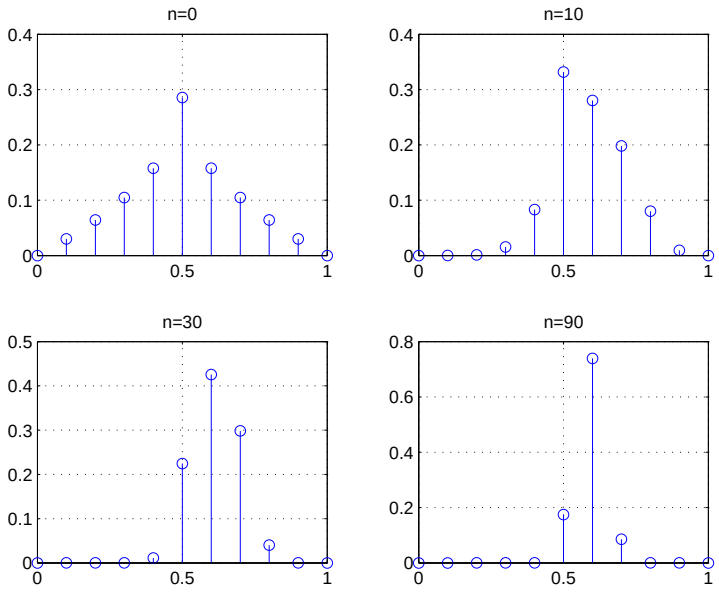


Figure 6: The greater the number of trials, the more concentrated is $P$ around $q = 0.6$, which means that Gary is more and more confident that the hypothesis $H_{0.6}$ is correct. The top left graph represents Gary's initial beliefs. The other graphs correspond to his re-evalutation after observing John's predictions over $n$ days.

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
rand('seed',121);
s=0.6;                          %Actual predictive power of John.
u=1/2;                          %Can be modified to represent different
```

```
                          %models of Gary's initial belief.
n=90;                     %number of days tested
q=0:0.1:1;
p=0.5^u-(abs(q-0.5)).^u;
p=p/sum(p);
R=[p];
for i=1:n
    x=(rand<=s);
    if x==1
        p=q.*p/sum(q.*p);
    else
        p=(1-q).*p/sum((1-q).*p);
    end
    R=[R;p];
end

stem(q,R(n,:))
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
```

## 4.3 Continuous random variables

**Exercise 4.14 (Convolution of exponential random variables)** Suppose that the lifetime of a component is exponentially distributed and that an identical and independent backup component is available. The system operates as long as one of the components is functional; therefore, the distribution of the life of the system is that of the sum of two independent exponential random variables. Let $T_1$ and $T_2$ be independent exponential random variables with parameter $\lambda$, and let $S = T_1 + T_2$. Find the probability density function of $S$. Note: the probability density function of the exponential random variable with parameter $\lambda$ is $\lambda \exp(-\lambda x)$ if $x \geq 0$ and 0 if $x < 0$. (Be careful with the limits of integration when calculating the convolution integral.) Answer: $f_S(s) = \lambda^2 s \exp(-\lambda s)$.

**Exercise 4.15 (Convolution of normal distributions)** Show that a linear combination $Y = a_1 X_1 + a_2 X_2$ of independent normal random variables $X_1$, $X_2$ is also a normal random variable.

## 4.4 The Monty Hall problem

The *Monty Hall problem* in based on a TV game show from the 1960s called "Let's Make a Deal." The show host, Monty Hall, would ask a contestant to choose one of three doors. Behind one of the doors was a valuable prize, say a car, and each of the other two concealed a goat. Instead of opening the door picked by the contestant, Monty would open one of the remaining two that did not conceal the car, and then give the contestant a choice whether to switch or to stick with the first selection. Should the contestant switch or not?

To avoid issues of interpretation, we assume that the same contestant has the opportunity to play the game many times, and has to decide on a strategy to follow. Should she always switch doors? Should she never switch doors? Should she follow a mixed strategy? Here is a more precise formulation of the game, from the point of view of the contestant:

step 1. Monty hides the car behind one of the three doors. No other information being available from the outset, the contestant assumes that the door concealing a car was chosen at random with equal probabilities.

step 2. The contestant picks a door at random with probabilities $(1/3, 1/3, 1/3)$.

step 3. Monty selects one of the other two doors to open. If the contestant's initial choice did not hid the car, then Monty necessarily opens the only remaining door that conceals a goat. If the contestant's first pick actually hid the car, Monty chooses from the other two at random with probabilities $(1/2, 1/2)$.

step 4. The contestant now applies one of several strategies chosen beforehand: (1) on every game, she chooses to stick with her initial choice; (2) on every game, she chooses to flip; (3) she picks from the set {flip, not flip} with probabilities $P(\text{flip}) = p$ and $P(\text{not flip}) = 1 - p$, independently, every game. (One can imagine other strategies, but these seem the simplest and most natural.)

In fact, these constitute a single family of strategies parametrized by $p$. If $p = 0$ we have strategy (1) and if $p = 1$ we have strategy (2). The problem is to find $p$ that maximizes the expected payoff, which we may agree to define as 1 for the car and 0 for a goat. We represent by $\mathcal{S}_p$ the strategy parametrized by $p$.

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
function y=montyhall(p,k)
%Input  - p probability of switching initial choice of door
%       - k number of games played
%
%Output - string of 0s (lose) and 1s (win) of length k

y=[];
for i=1:k
    %The doors are numbered 1, 2, 3.
    %Monty selects one door at random with probabilities
    %(1/3, 1/3, 1/3) and places the car behind it.
    %The winning door with the car is representeed by w.
    x=rand;
    w=1*(x<=1/3) + 2*(x>1/3 & x<= 2/3) + 3*(x>2/3);

    %The doors with goats will be called g1 and g2
```

```
    if w==1
        g1=2;
        g2=3;
    elseif w==2
        g1=1;
        g2=3;
    else
        g1=1;
        g2=2;
    end

    %Contestant makes initial choice of door
    %with probabilities (1/3, 1/3, 1/3). Call the choice c1.
    x=rand;
    c1=1*(x<=1/3) + 2*(x>1/3 & x<= 2/3) + 3*(x>2/3);

    %Now, Monty picks a door different from c1 and opens it.
    %Call this door m.
    if c1 == w
        x=rand;
        m=g1*(x<=1/2)+g2*(x>1/2);
    elseif c1==g1
        m=g2;
    else
        m=g1;
    end

    %The two remaining closed doors are c1 and another
    %which we call r, determined by:
    r=6-(c1+m);

    %The contestant now chooses a door c2 between c1 and r.
    %With probability p, she chooses to switch to r.
    x=rand;
    c2=r*(x<=p)+c1*(x>p);

    %The payoff, a, is either 0 or 1:
    if c2==w
        a=1;
    else
        a=0;
    end
    y=[y a];
end
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
```

**Exercise 4.16** For each $p$ from 0 to 1 in steps of 0.2, find the relative frequency of wins for 5000 trials of the game, then plot the result as a function of $p$.

**Exercise 4.17** Derive analytically the exact value of the expected payoff for strategy $\mathcal{S}_p$ as a function of $p$.

To argue exercise 4.17 we need to know that once the contestant has decided on the strategy of not switching doors, her estimated probability, 1/3, of winning a car is not affected by the extra information gained when Monty opened one of the other two doors. This can be justified as follows. Let $D_1$, $D_2$, $D_3$ be the events that the car is behind doors $1, 2, 3$, respectively. Say that the contestant has chosen door $D_1$. Let $M_2, M_3$ be the events that Monty opens doors 2 and 3, respectively (and contestant already selected door 1). Then

$$\begin{aligned} P(D_1|M_2) &= \frac{P(M_2|D_1)P(D_1)}{P(M_2)} \text{ (by Bayes theorem)} \\ &= \frac{P(M_2|D_1)P(D_1)}{P(M_2|D_1)P(D_1) + P(M_2|D_2)P(D_2) + P(M_2|D_3)P(D_3)} \\ &= \frac{(1/2)(1/3)}{(1/2)(1/3) + 0(1/3) + 1(1/3)} \\ &= \frac{1}{3} \\ &= P(D_1). \end{aligned}$$

## 4.5 Random permutations

**Exercise 4.18 (Lotto simulation)** Write a program to simulate a lotto machine. It should pick five balls at random from a set numbered 1 to 50, one by one without replacement. The following line of Matlab code may be helpful: we can obtain a permutation of the numbers $1, 2, \ldots, n$ using

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
[ignore,p] = sort(rand(1,n));
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
```

The permutation itself is the output `p`, which is a vector of length $n$ with entries from 1 to $n$. The variable `ignore` (which is a vector of random numbers between 0 and 1 sorted in increasing order) is not used. Why does this command actually produce a random permutation? Use the Matlab help facility to learn about the `sort` function.

If you play cards, you will likely know the meaning of terms such as 'riffle shuffles,' 'cuts,' etc. These are permutations done to a deck of cards to randomized them. You can find a discussion of this topic and about simulation of card shuffling in [CGI]. Reference [LC] has a more mathematical, but very brief, discussion of the problem of how many shuffles are needed to mixed the deck up. This is fundamentally a problem about random walks on groups, and there is a lot of interesting group theory involved.

**Exercise 4.19 (Riffles and cuts)** Describe the permutation of $1, \ldots, n$ ($n$ is the number of cards in a deck) that represents a riffle. Do the same for a cut. (You may need to look up for the meaning of these terms.)

## 4.6 Buffon's needle problem

**Exercise 4.20** Show how to construct the function $X$ described in the example 2.2, draw the region $E \subseteq [0, a] \times [0, 2\pi]$ corresponding to intersection and prove that $P(E) = 2l/\pi a$, as claimed there. Hint: For each fixed $\theta$ find the range of $x$ to have intersection.

You can do a simulation of the needle problem with the Matlab script below.

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
tic
rand('seed',121)
c=0;
N=1000000;
for i=1:N
    h=rand;
    u=2*pi*rand;
    y=h+0.5*sin(u);
    if (y>0 & y<1)
        c=c;
    else
        c=c+1;
    end
end
buffonpi=N/c
toc
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
```

## 4.7 A non-measurable set

It was pointed out earlier that one of the reasons for making explicit the $\sigma$-algebra of events $\mathcal{F}$ in the definition of probability space is that, for the most important example of the interval $[0, 1]$ with probability measure derived from the rule $P([a, b]) = b-a$, there are subsets for which it is not possible to associate a probability measure (or length) at all. Such subsets cannot be regarded as events of any probability experiment without incurring logical contradictions, and must be excluded from the outset. The purpose of this extended exercise is to have you construct one example of such a non-measurable set.

Rather than construct the set in $[0, 1]$ directly, it is simpler to construct it in the unit circle $S^1 = \{z \in \mathbb{C} : |z| = 1\}$ and note that we can map the interval onto the circle almost bijectively (except for the endpoints 0 and 1, which are sent to $1 \in S^1$) using the map

$$x \in [0, 1] \mapsto e^{2\pi i x} \in S^1.$$

Also note that the measure on the interval corresponds under this map to the length of segments of arc on the circle. Therefore, this measure is invariant under rotation. We denote by $R_\theta : S^1 \to S^1$ the rotation by angle $\theta$:

$$R_\theta(z) = e^{i\theta} z.$$

Fix $\alpha = 2\pi\sqrt{2}$. (There is nothing special about $\sqrt{2}$ other than that it is irrational.) Define for each $z \in S^1$ the set

$$I_z := \{R_\alpha^m(z) : m \in \mathbb{Z}\}.$$

In other words, $I_z$ consists of all the points on the circle obtained by rotating $z$ by an integer multiple of the angle $\alpha$.

If you took Math 310, you should not have much difficulty checking the claims of the following exercise.

**Exercise 4.21** Show that the following claims hold:

1. For any two points $z_1, z_2 \in S^1$, either $I_{z_1} = I_{z_2}$ or $I_{z_1} \cap I_{z_2} = \emptyset$. Note: check that the relation

   $$z_1 \sim z_2 \Leftrightarrow z_2 = R_\alpha^m(z_1), \text{ for some } m \in \mathbb{Z}$$

   is an equivalence relation.

2. By item one, argue that $S^1$ is an (uncountable) union of sets

   $$S^1 = \bigcup_{z \in A} I_z,$$

   where $A$ is a set that contains a single element from each distinct equivalence class. (Being able to do this requires that you have faith in the axiom of choice. In fact, the set $A$ is defined by choosing from each distinct set $I_z$ a single element.)

3. Let $A_m = R_\alpha^m(A)$ denote the sets obtained by rotating $A$ by the angle $m\alpha$. Using the fact that $\alpha/2\pi$ is irrational, show that

   $$S^1 = \bigcup_{m \in \mathbb{Z}} A_m,$$

   and that the union is disjoint.

4. Now argue that $A$ cannot be a measurable set. In fact, suppose, in order to arrive at a contradiction, that it makes sense to define the arc-length of $A$. We denote this arc-length by $2\pi P(A)$, where $P(A)$ is the measure the set would have if we view it as a subset of $[0, 1]$ through the identification with $S^1$ described above. Arc-lengths are not changed under rotations, so each set $A_m$ must have the same length as $A$. But now we have a dilemma: if $A$ has positive length, the length of $S^1$ would be infinite, by the axiom

of additivity of probability under countable disjoint unions and the claim of Part 3. On the other hand, if $A$ had arc-length 0, $S^1$ would be forced to have arc-length zero, being a countable union of sets of zero length. Either way we have a contradiction. The only escape is to conclude that $A$ cannot be assigned an arc-length measure consistent with the axioms of probability.

# 5 Appendix: The Lebesgue integral

The definition of measurable sets leads in a natural way to a very general and powerful notion of integration, known as the *Lebesgue integral*. We take here a very brief look at this concept. Our main motivation is to have a general notion of expected value of a random variable.

Suppose that $(S, \mathcal{F}, P)$ is a probability space and let $X : S \to \mathbb{R}$ be a random variable. For technical convenience we assume that $X$ is bounded by two numbers: $a \leq X \leq b$, although this is not esssential.

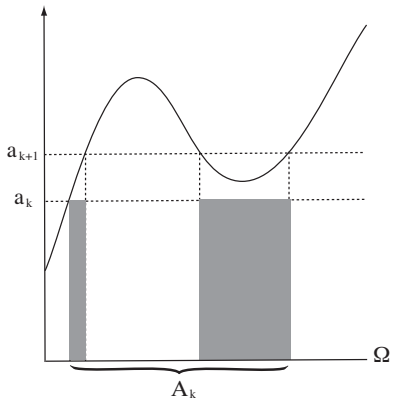


Figure 7: Pre-image of an interval

For a given positive integer $n$ subdivide the interval $[a, b]$ into $n$ equally spaced intervals of length $(b - a)/n$:

$$[a, b] = [a_0, a_1] \cup (a_1, a_2] \cup \cdots \cup (a_{n-1}, a_n],$$

where $a_0 = a$ and $a_n = b$. Let $A_k \subset S$ denote the event $a_k < X \leq a_{k+1}$. Now write

$$I_n(X) = \sum_{k=1}^{n-1} a_k P(A_k).$$

The limit of $I_n(X)$ as $n \to \infty$ is called the *Lebesgue integral* of $X$ and is denoted

$$\int_S X(s)dP(s) = \lim_{n\to\infty} I_n(X).$$

Sometimes the notation $P(ds)$ is used for $dP(s)$.

If $f : \mathbb{R} \to \mathbb{R}$ is a Borel measurable (bounded) function and $X : S \to \mathbb{R}$ is a random variable with probability law $P_X$, it is not too difficult to obtain from the definition that the expected value of $f \circ X$ is given by

$$E[f \circ X] := \int_S f(X(s))dP(s) = \int_{-\infty}^{\infty} f(x)dP_X(x),$$

where the second integral is the Lebesgue integral on the set of values of $X$ relative to the probability measure $P_X$.

We note the following remarks. First, what makes this definition much more general than Riemann's definition of integral is that we are freed from the limitation of using simple sets (intervals) to partition the domain of the function $X$. Here we partition $S$ using measurable sets, on each of which $X$ falls into a narrow interval in its image set.

Another feature of the Lebesgue integral is its notational convenience. Notice that $S$ could have been continuous or discrete. In the continuous case, the integral may reduce to ordinary Riemann integral, whereas in the discrete case it reduces to a discrete sum. For example, if $S = \{1, 2, \ldots, 6\}^3$ as in the rolling of three dice example, and $X(s) = (i, j, k)$ is the outcome of a roll, then it can be shown that

$$\int_S f(X(s))dP(s) = \frac{1}{216} \sum_{(i,j,k)\in S} f(i, j, k).$$

# References


[Brem] Pierre Brémaud. *An Introduction to Probabilistic Modeling*, Springer, 1988.

[CGI] K. Chen, P. Giblin, A. Irving. *Mathematical explorations with Matlab*, Cambridge University Press, 1999.

[Kay] Steven Kay. *Intuitive Probability and Random Processes using Matlab*, Springer, 2006.

[LC] G.F. Lawler and L.N. Coyle. *Lectures on Contemporary Probability*, AMS/IAS, Student Mathematical Library, Volume 2, 1999.

[Lee] Peter M. Lee. *Bayesian Statistics - an introduction*, Hodder Arnold, 2004.

[Snell] J. Laurie Snell. *Introduction to Probability Theory with Computing*, Prentice-Hall, 1975.