Chapter 2

Probability Theory

2.1 Random Experiments

The basic notion in probability is that of a **random experiment**: an experiment whose outcome cannot be determined in advance, but is nevertheless still subject to analysis.

Examples of random experiments are:

1. tossing a die,
2. measuring the amount of rainfall in Brisbane in January,
3. counting the number of calls arriving at a telephone exchange during a fixed time period,
4. selecting a random sample of fifty people and observing the number of left-handers,
5. choosing at random ten people and measuring their height.

Example 2.1 (Coin Tossing) The most *fundamental* stochastic experiment is the experiment where a coin is tossed a number of times, say n times. Indeed, much of probability theory can be based on this simple experiment, as we shall see in subsequent chapters. To better understand how this experiment behaves, we can carry it out on a digital computer, for example in Matlab. The following simple Matlab program, simulates a sequence of 100 tosses with a fair coin(that is, heads and tails are equally likely), and plots the results in a bar chart.

```
x = (rand(1,100) < 1/2)
bar(x)
```

Here `x` is a vector with 1s and 0s, indicating Heads and Tails, say. Typical outcomes for three such experiments are given in Figure 1.1.

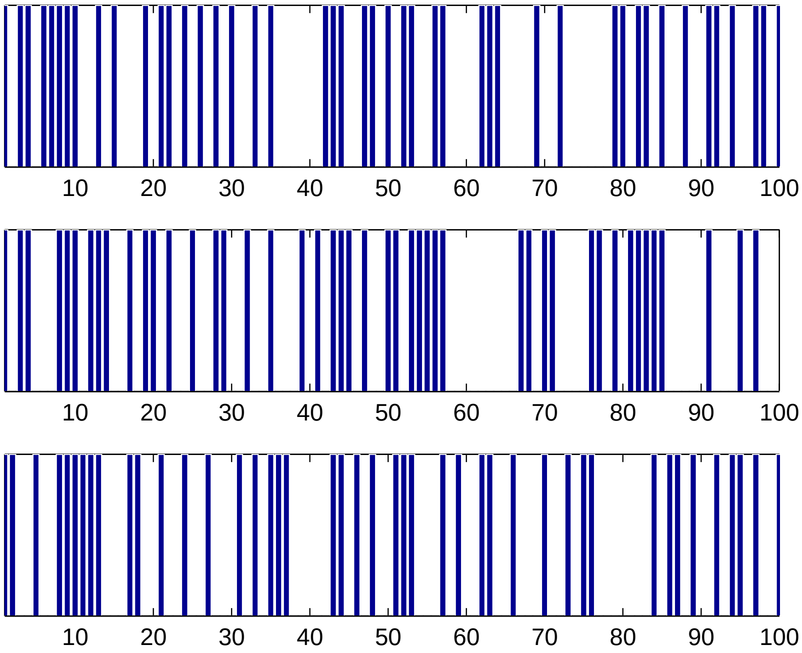


Figure 2.1: Three experiments where a fair coin is tossed 100 times. The dark bars indicate when "Heads" (=1) appears.

We can also plot the average number of "Heads" against the number of tosses. In the same Matlab program, this is done in two extra lines of code:

```
y = cumsum(x)./[1:100]
plot(y)
```

The result of three such experiments is depicted in Figure 1.2. Notice that the average number of Heads seems to converge to 1/2, but there is a lot of random fluctuation.

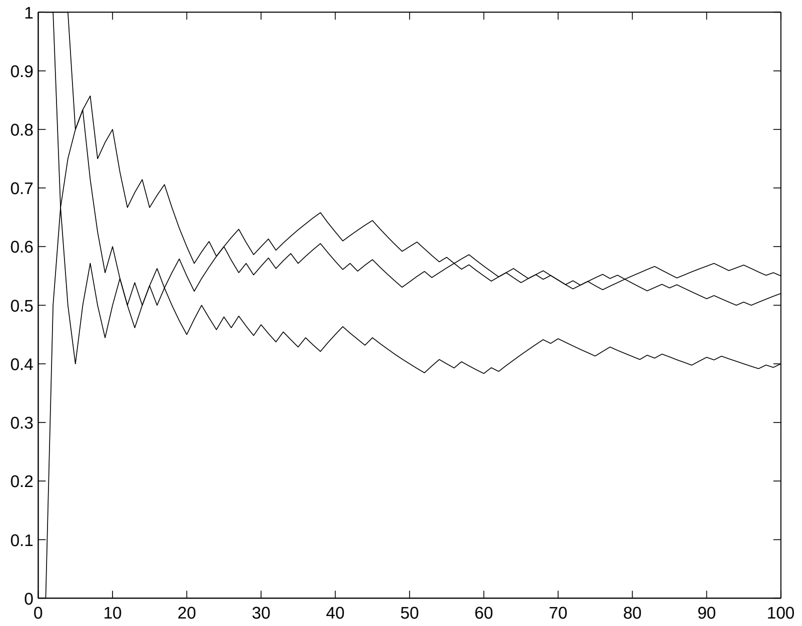


Figure 2.2: The average number of heads in n tosses, where $n = 1, \ldots, 100$.

Example 2.2 (Control Chart) Control charts, see Figure 1.3, are frequently used in manufacturing as a method for *quality control.* Each hour the average output of the process is measured — for example, the average weight of 10 bags of sugar — to assess if the process is still "in control", for example, if the machine still puts on average the correct amount of sugar in the bags. When the process $>$ *Upper Control Limit* or $<$ *Lower Control Limit* and an alarm is raised that the process is out of control, e.g., the machine needs to be adjusted, because it either puts too much or not enough sugar in the bags. The question is how to set the control limits, since the random process naturally uctuatesaround its "centre" or "target" line.

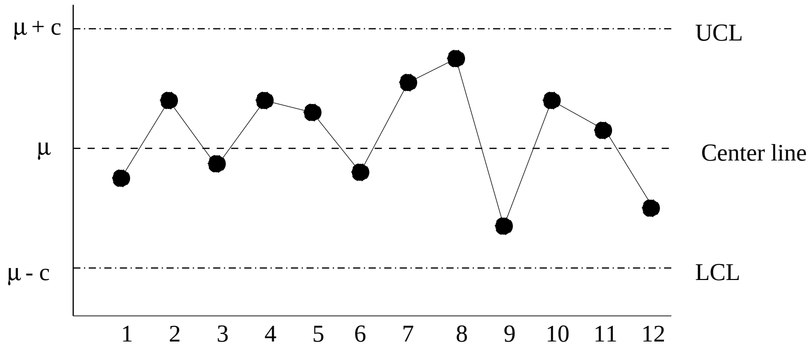


Figure 2.3: Control Chart

Example 2.3 (Machine Lifetime) Suppose 1000 identical components are monitored for failure, up to 50,000 hours. The outcome of such a random experiment is typically summarised via the cumulative lifetime table and plot, as given in Table 1.1 and Figure 1.3, respectively. Here $\hat{F}(t)$ denotes the proportion of components that have failed at time t. One question is how $\hat{F}(t)$ can be modelled via a continuous function F, representing the lifetime distribution of a typical component.

t (h)	failed	$\widehat{F(t)}$	t (h)	failed	$\widehat{F(t)}$
0	0	0.000	3000	140	0.140
750	22	0.020	5000	200	0.200
800	30	0.030	6000	290	0.290
900	36	0.036	8000	350	0.350
1400	42	0.042	11000	540	0.540
1500	58	0.058	15000	570	0.570
2000	74	0.074	19000	770	0.770
2300	105	0.105	37000	920	0.920

Table 2.1: The cumulative lifetime table

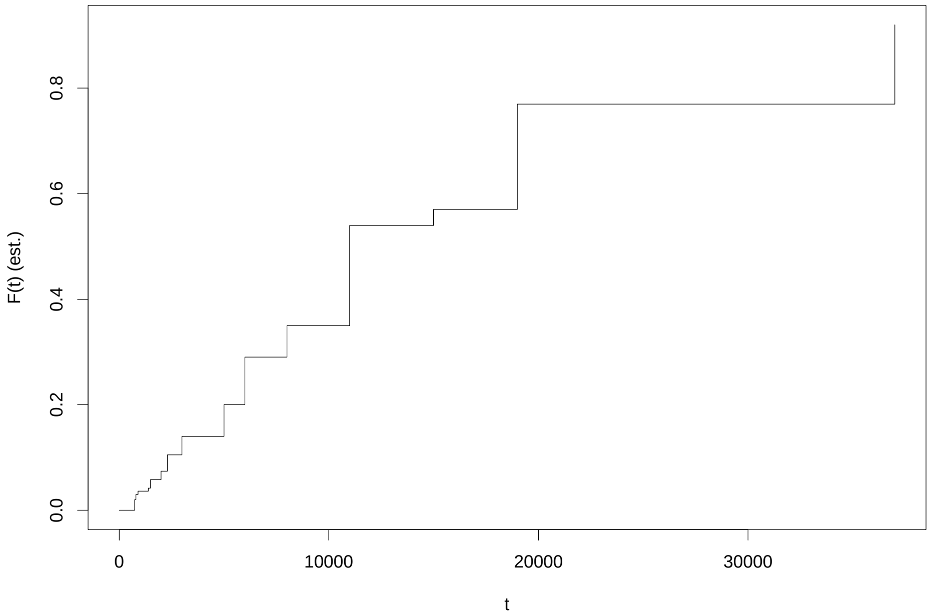


Figure 2.4: The cumulative lifetime table

Example 2.1 A 4-engine aeroplane is able to fly on just one engine on each wing. All engines are unreliable.

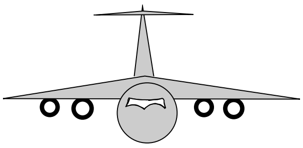


Figure 2.5: A aeroplane with 4 unreliable engines

Number the engines: 1,2 (left wing) and 3,4 (right wing). Observe which engine works properly during a specified period of time. There are $2^4 = 16$ possible outcomes of the experiment. Which outcomes lead to "system failure"? Moreover, if the probability of failure within some time period is known for each of the engines, what is the probability of failure for the entire system? Again this can be viewed as a random experiment.

Below are two more pictures of randomness. The first is a computer-generated "plant", which looks remarkably like a real plant. The second is real data depicting the number of bytes that are transmitted over some communications link. An interesting feature is that the data can be shown to exhibit "fractal" behaviour, that is, if the data is aggregated into smaller or larger time intervals, a similar picture will appear.

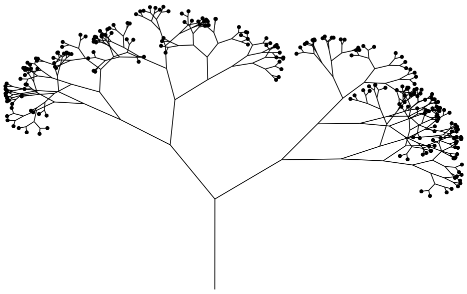


Figure 2.6: Plant growth

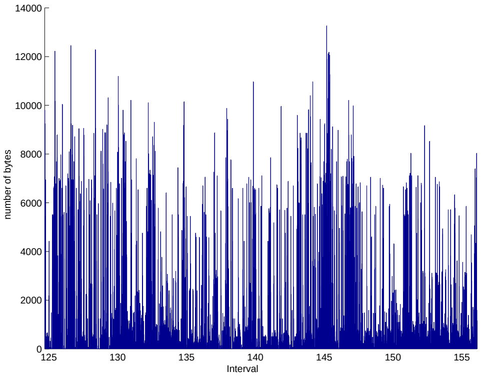


Figure 2.7: Telecommunications data

We wish to describe these experiments via appropriate mathematical models. These models consist of three building blocks: a *sample space*, a set of *events* and a *probability*. We will now describe each of these objects.

2.2 Sample Space

Although we cannot predict the outcome of a random experiment with certainty we usually can specify a set of possible outcomes. This gives the first ingredient in our model for a random experiment.

Definition 2.1 The **sample space** of a random experiment is the set of all possible outcomes of the experiment.

Examples of random experiments with their sample spaces are:

1. Cast two dice consecutively,

$$\Omega = \{(1,1),(1,2),\ldots,(1,6),(2,1),\ldots,(6,6)\}.$$

2. The lifetime of a machine (in days),

$$\Omega = \mathbb{R}_+ = \{\text{ positive real numbers }\}.$$

3. The number of arriving calls at an exchange during a specified time interval,

$$\Omega = \{0,1,\cdots\} = \mathbb{Z}_+ .$$

4. The heights of 10 selected people.

$$\Omega = \{(x_1,\ldots,x_{10}), x_i \geq 0, i = 1,\ldots,10\} = \mathbb{R}_+^{10} .$$

   Here $(x_1,\ldots,x_{10})$ represents the outcome that the length of the first selected person is x_1, the length of the second person is x_2, et cetera.

Notice that for modelling purposes it is often easier to take the sample space larger than necessary. For example the actual lifetime of a machine would certainly not span the entire positive real axis. And the heights of the 10 selected people would not exceed 3 metres.

2.3 Events

Often we are not interested in a single outcome but in whether or not one of a *group* of outcomes occurs. Such subsets of the sample space are called **events**. Events will be denoted by capital letters $A, B, C, \ldots$. We say that event A **occurs** if the outcome of the experiment is one of the elements in A.

Examples of events are:

1. The event that the sum of two dice is 10 or more,
$$A = \{(4,6),(5,5),(5,6),(6,4),(6,5),(6,6)\}.$$
2. The event that a machine lives less than 1000 days,
$$A = [0,1000)\ .$$
3. The event that out of fifty selected people, five are left-handed,
$$A = \{5\}\,.$$

Example 2.5 (Coin Tossing) Suppose that a coin is tossed 3 times, and that we "record" every head and tail (not only the number of heads or tails). The sample space can then be written as

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

where, for example, HTH means that the first toss is heads, the second tails, and the third heads. An alternative sample space is the set $\{0,1\}^3$ of binary vectors of length 3, e.g., HTH corresponds to (1,0,1), and THH to (0,1,1).

The event A that the third toss is heads is

$$A = \{HHH, HTH, THH, TTH\}\ .$$

Since events are sets, we can apply the usual set operations to them:

1. the set $A \cup B$ (A **union** B) is the event that A *or* B *or* both occur,
2. the set $A \cap B$ (A **intersection** B) is the event that A *and* B both occur,
3. the event A^c (A **complement**) is the event that A does *not* occur,
4. if $A \subset B$ (A is a **subset** of B) then event A is said to *imply* event B.

Two events A and B which have no outcomes in common, that is, $A \cap B = \emptyset$, are called **disjoint** events.

Example 2.6 Suppose we cast two dice consecutively. The sample space is $= \{(1,1),(1,2),\ldots,(1,6),(2,1),\ldots,(6,6)\}$. Let $A = \{(6,1),\ldots,(6,6)\}$ be the event that the first die is 6, and let $B = \{(1,6),\ldots,(1,6)\}$ be the event that the second dice is 6. Then $A \cap B = \{(6,1),\ldots,(6,6)\} \cap \{(1,6),\ldots,(6,6)\} = \{(6,6)\}$ is the event that both die are 6.

It is often useful to depict events in a **Venn diagram**, such as in Figure 1.8

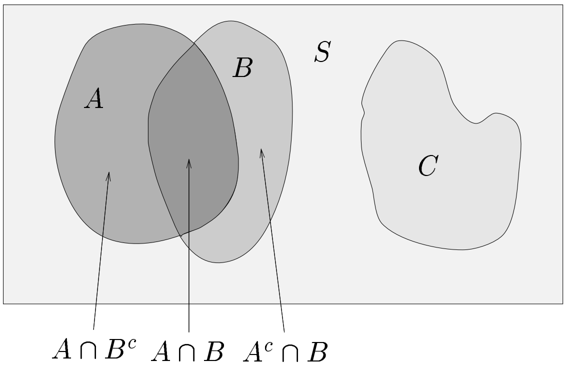


Figure 1.8: A Venn diagram

In this Venn diagram we see

(i) $A \cap C = \emptyset$ and therefore events A and C are disjoint.

(ii) $(A \cap B^c) \cap (A^c \cap B) = \emptyset$ and hence events $A \cap B^c$ and $A^c \cap B$ are disjoint.

Example 2.7 (System Reliability) In Figure 1.9 three systems are depicted, each consisting of 3 unreliable components. The *series* system works if and only if (abbreviated as iff) *all* components work; the *parallel* system works iff *at least one* of the components works; and the 2-out-of-3 system works iff at least 2 out of 3 components work.

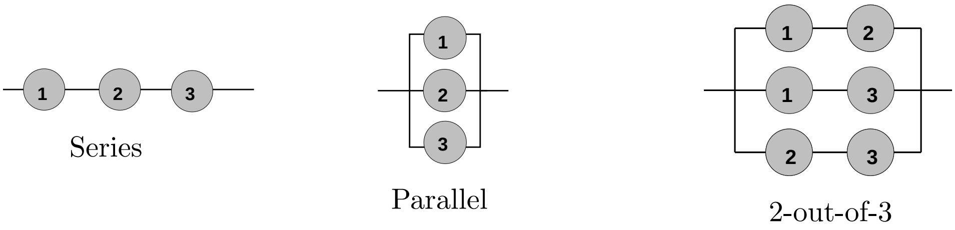


Figure 1.9: Three unreliable systems

Let A_i be the event that the ith component is functioning, $i = 1, 2, 3$; and let D_a, D_b, D_c be the events that respectively the series, parallel and 2-out-of-3 system is functioning. Then,

$$D_a = A_1 \cap A_2 \cap A_3 \ ,$$

and

$$D_b = A_1 \cup A_2 \cup A_3 \ .$$

Also,

$$\begin{aligned} D_c &= (A_1 \cap A_2 \cap A_3) \cup (A_1^c \cap A_2 \cap A_3) \cup (A_1 \cap A_2^c \cap A_3) \cup (A_1 \cap A_2 \cap A_3^c) \\ &= (A_1 \cap A_2) \cup (A_1 \cap A_3) \cup (A_2 \cap A_3) \ . \end{aligned}$$

Two useful results in the theory of sets are the following, due to **De Morgan**: If $\{A_i\}$ is a collection of events (sets) then

$$\left(\bigcup_i A_i\right)^c = \bigcap_i A_i^c \tag{2.1}$$

and

$$\left(\bigcap_i A_i\right)^c = \bigcup_i A_i^c \ . \tag{2.2}$$

This is easily proved via Venn diagrams. Note that if we interpret A_i as the event that a component works, then the left-hand side of (1.1) is the event that the corresponding parallel system is not working. The right hand is the event that at all components are not working. Clearly these two events are the same.

2.4 Probability

The third ingredient in the model for a random experiment is the specification of the probability of the events. It tells us how *likely* it is that a particular event will occur.

Definition 2.2 A probability $\mathbb{P}$ is a rule (function) which assigns a positive number to each event, and which satisfies the following **axioms**:

Axiom 1: $\mathbb{P}(A) \geq 0$.
Axiom 2: $\mathbb{P}(\Omega) = 1$.
Axiom 3: For any sequence $A_1, A_2, \ldots$ of *disjoint* events we have

$$\boxed{\mathbb{P}(\bigcup_i A_i) = \sum_i \mathbb{P}(A_i)} \ . \tag{2.3}$$

Axiom 2 just states that the probability of the “certain” event Ω is 1. Property (1.3) is the *crucial* property of a probability, and is sometimes referred to as the **sum rule**. It just states that if an event can happen in a number of different ways *that cannot happen at the same time*, then the probability of this event is simply the sum of the probabilities of the composing events.

Note that a probability rule $\mathbb{P}$ has exactly the same properties as the common “area measure”. For example, the total area of the union of the triangles in Figure 2.10 is equal to the sum of the areas of the individual triangles. This

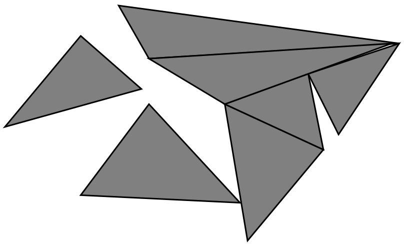


Figure 2.10: The probability measure has the same properties as the “area” measure: the total area of the triangles is the sum of the areas of the idividual triangles.

is how you should interpret property (1.3). But instead of measuring areas, $\mathbb{P}$ measures probabilities.

As a direct consequence of the axioms we have the following properties for $\mathbb{P}$.

Theorem 2.1 Let A and B be events. Then,

1. $\mathbb{P}(\emptyset) = 0$.
2. $A \subset B \Longrightarrow \mathbb{P}(A) \leq \mathbb{P}(B)$.
3. $\mathbb{P}(A) \leq 1$.
4. $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.
5. $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.

Proof.

1. $\Omega = \Omega \cap \emptyset \cap \emptyset \cap \cdots$, therefore, by the sum rule, $\mathbb{P}(\Omega) = \mathbb{P}(\Omega) + \mathbb{P}(\emptyset) + \mathbb{P}(\emptyset) + \cdots$, and therefore, by the second axiom, $1 = 1 + \mathbb{P}(\emptyset) + \mathbb{P}(\emptyset) + \cdots$, from which it follows that $\mathbb{P}(\emptyset) = 0$.
2. If $A \subset B$, then $B = A \cup (B \cap A^c)$, where A and $B \cap A^c$ are disjoint. Hence, by the sum rule, $\mathbb{P}(B) = \mathbb{P}(A) + \mathbb{P}(B \cap A^c)$, which is (by the first axiom) greater than or equal to $\mathbb{P}(A)$.
3. This follows directly from property 2 and axiom 2, since $A \subset \Omega$.
4. $\Omega = A \cup A^c$, where A and A^c are disjoint. Hence, by the sum rule and axiom 2: $1 = \mathbb{P}(\Omega) = \mathbb{P}(A) + \mathbb{P}(A^c)$, and thus $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.
5. Write $A \cup B$ as the disjoint union of A and $B \cap A^c$. Then, $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B \cap A^c)$. Also, $B = (A \cap B) \cup (B \cap A^c)$, so that $\mathbb{P}(B) = \mathbb{P}(A \cap B) + \mathbb{P}(B \cap A^c)$. Combining these two equations gives $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.

We have now completed our model for a random experiment. It is up to the modeller to specify the sample space Ω and probability measure $\mathbb{P}$ which most closely describes the actual experiment. This is not always as straightforward as it looks, and sometimes it is useful to model only certain *observations* in the experiment. This is where *random variables* come into play, and we will discuss these in the next chapter.

Example 2.8 Consider the experiment where we throw a fair die. How should we define and $\mathbb{P}$?

Obviously, $\Omega = \{1, 2, \ldots, 6\}$; and some common sense shows that we should define $\mathbb{P}$ by

$$\mathbb{P}(A) = \frac{|A|}{6}, \quad A \subset \Omega,$$

where $|A|$ denotes the number of elements in set A. For example, the probability of getting an even number is $\mathbb{P}(\{2, 4, 6\}) = 3/6 = 1/2$.

In many applications the sample space is *countable*, i.e. $\Omega = \{a_1, a_2, \ldots, a_n\}$ or $\Omega = \{a_1, a_2, \ldots\}$. Such a sample space is called **discrete**.

The easiest way to specify a probability $\mathbb{P}$ on a discrete sample space is to specify first the probability p_i of each **elementary event** $\{a_i\}$ and then to define

$$\mathbb{P}(A) = \sum_{i:a_i \in A} p_i \ , \ \text{ for all } \ A \subset \Omega.$$

This idea is graphically represented in Figure 1.11. Each element a_i in the sample is assigned a probability weight p_i represented by a black dot. To find the probability of the set A we have to sum up the weights of all the elements in A.

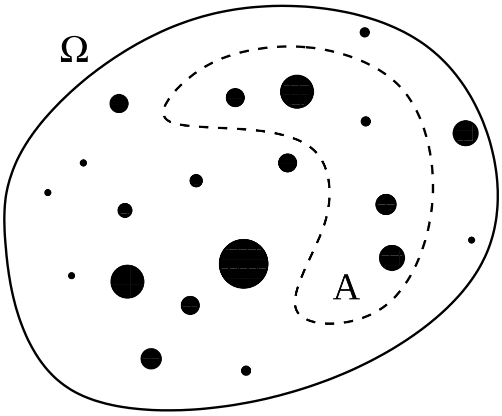


Figure 2.11: A discrete sample space

Again, it is up to the modeller to properly specify these probabilities. Fortunately, in many applications all elementary events are *equally likely*, and thus the probability of each elementary event is equal to 1 divided by the total number of elements in Ω. E.g., in Example 1.8 each elementary event has probability 1/6.

Because the "equally likely" principle is so important, we formulate it as a theorem.

Theorem 2.2 (Equilikely Principle) If has a finite number of outcomes, and all are equally likely, then the probability of each event A is defined as

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

Thus for such sample spaces the calculation of probabilities reduces to *counting* the number of outcomes (in A and Ω).

When the sample space is not countable, for example $\Omega = \mathbb{R}_+$, it is said to be **continuous**.

Example 2.9 We draw at random a point in the interval $[0, 1]$. Each point is equally likely to be drawn. How do we specify the model for this experiment?

The sample space is obviously $\Omega = [0, 1]$, which is a continuous sample space. We cannot define $\mathbb{P}$ via the elementary events $\{x\}$, $x \in [0, 1]$ because each of these events must have probability 0 (!). However we can define $\mathbb{P}$ as follows: For each $0 \leq a \leq b \leq 1$, let

$$\mathbb{P}([a, b]) = b - a\ .$$

This completely specifies $\mathbb{P}$. In particular, we can find the probability that the point falls into any (sufficiently nice) set A as the *length* of that set.

2.5 Counting

Counting is not always easy. Let us first look at some examples:

1. A multiple choice form has 20 questions; each question has 3 choices. In how many possible ways can the exam be completed?
2. Consider a horse race with 8 horses. How many ways are there to gamble on the placings (1st, 2nd, 3rd).
3. Jessica has a collection of 20 CDs, she wants to take 3 of them to work. How many possibilities does she have?

4. How many different throws are possible with 3 dice?

To be able to comfortably solve a multitude of counting problems requires a lot of experience and *practice*, and even then, some counting problems remain exceedingly hard. Fortunately, many counting problems can be cast into the simple framework of drawing balls from an urn, see Figure 2.12.

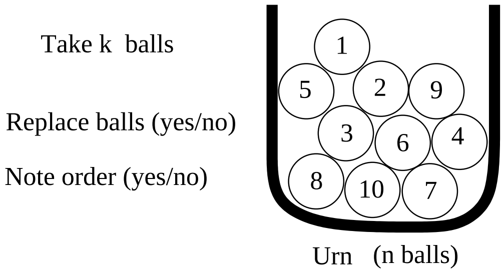


Figure 2.12: An urn with n balls

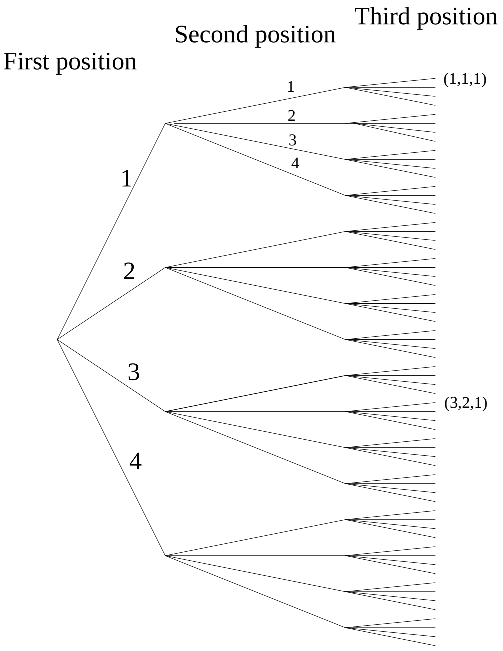
Consider an urn with n different balls, numbered $1, \ldots, n$ from which k balls are drawn. This can be done in a number of different ways. First, the balls can be drawn one-by-one, or one could draw all the k balls at the same time. In the first case the **order** in which the balls are drawn can be noted, in the second case that is not possible. In the latter case we can (and will) still assume the balls are drawn one-by-one, but that the order is not noted. Second, once a ball is drawn, it can either be put back into the urn (after the number is recorded), or left out. This is called, respectively, drawing with and without **replacement**. All in all there are 4 possible experiments: (ordered, with replacement), (ordered, without replacement), (unordered, without replacement) and (ordered, with replacement). The art is to recognise a seemingly unrelated counting problem as one of these four urn problems. For the 4 examples above we have the following

1. Example 1 above can be viewed as drawing 20 balls from an urn containing 3 balls, noting the order, and with replacement.
2. Example 2 is equivalent to drawing 3 balls from an urn containing 8 balls, noting the order, and without replacement.
3. In Example 3 we take 3 balls from an urn containing 20 balls, not noting the order, and without replacement
4. Finally, Example 4 is a case of drawing 3 balls from an urn containing 6 balls, not noting the order, and with replacement.

Before we proceed it is important to introduce a notation that reflects whether the outcomes/arrangements are ordered or not. In particular, we denote ordered arrangements by *vectors*, e.g., $(1, 2, 3) \neq (3, 2, 1)$, and unordered arrangements by *sets*, e.g., $\{1, 2, 3\} = \{3, 2, 1\}$. We now consider for each of the four cases how to count the number of arrangements. For simplicity we consider for each case how the counting works for $n = 4$ and $k = 3$, and then state the general situation.

Drawing with Replacement, Ordered

Here, after we draw each ball, note the number on the ball, and put the ball back. Let $n = 4, k = 3$. Some possible outcomes are $(1, 1, 1), (4, 1, 2)$, $(2, 3, 2)$, $(4, 2, 1), \ldots$ To count how many such arrangements there are, we can reason as follows: we have three positions $(\cdot, \cdot, \cdot)$ to fill in. Each position can have the numbers 1,2,3 or 4, so the total number of possibilities is $4 \times 4 \times 4 = 4^3 = 64$. This is illustrated via the following tree diagram:

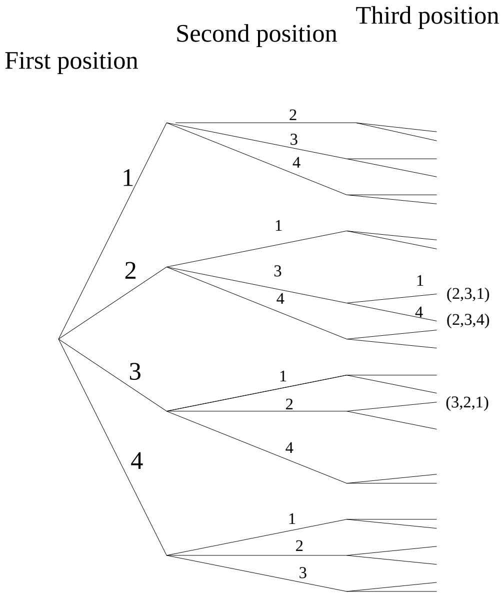


For general n and k we can reason analogously to find:

> The number of ordered arrangements of k numbers chosen from $\{1, \ldots, n\}$, with replacement (repetition) is n^k.

Drawing Without Replacement, Ordered

Here we draw again k numbers (balls) from the set $\{1, 2, \ldots, n\}$, and note the order, but now do not replace them. Let $n = 4$ and $k = 3$. Again there are 3 positions to fill $(\cdot, \cdot, \cdot)$, but now the numbers cannot be the same, e.g., (1,4,2),(3,2,1), etc. Such an ordered arrangements called a **permutation** of size k from set $\{1, \ldots, n\}$. (A permutation of $\{1, \ldots, n\}$ of size n is simply called a permutation of $\{1, \ldots, n\}$ (leaving out "of size n"). For the 1st position we have 4 possibilities. Once the first position has been chosen, we have only 3 possibilities left for the second position. And after the first two positions have been chosen there are 2 positions left. So the number of arrangements is $4 \times 3 \times 2 = 24$ as illustrated in Figure 1.5, which is the same tree as in Figure 1.5, but with all "duplicate" branches removed.



For general n and k we have:

> The number of permutations of size k from $\{1, \ldots, n\}$ is ${}^nP_k = n(n-1)\cdots(n-k+1)$.

In particular, when $k = n$, we have that the number of ordered arrangements of n items is $n! = n(n-1)(n-2)\cdots 1$, where $n!$ is called n-**factorial**. Note that

$${}^nP_k = \frac{n!}{(n-k)!}.$$

Drawing Without Replacement, Unordered

This time we draw k numbers from $\{1, \ldots, n\}$ but do not replace them (no replication), and do not note the order (so we could draw them in one grab). Taking again $n = 4$ and $k = 3$, a possible outcome is $\{1, 2, 4\}$, $\{1, 2, 3\}$, etc. If we noted the order, there would be ${}^nP_k$ outcomes, amongst which would be (1,2,4),(1,4,2),(2,1,4),(2,4,1),(4,1,2) and (4,2,1). Notice that these 6 permutations correspond to the single unordered arrangement $\{1, 2, 4\}$. Such unordered arrangements without replications are called **combinations** of size k from the set $\{1,\ldots,n\}$.

To determine the number of combinations of size k simply need to divide ${}^nP_k$ be the number of permutations of k items, which is $k!$. Thus, in our example $(n = 4, k = 3)$ there are $24/6 = 4$ possible combinations of size 3. In general we have:

The number of combinations of size k from the set $\{1,\ldots n\}$ is

$$ {}^nC_k = \binom{n}{k} = \frac{{}^nP_k}{k!} = \frac{n!}{(n-k)!\,k!} \, . $$

Note the two different notations for this number. We will use the second one.

Drawing With Replacement, Unordered

Taking $n = 4, k = 3$, possible outcomes are $\{3,3,4\}$, $\{1,2,4\}, \{2,2,2\}$, etc. The trick to solve this counting problem is to represent the outcomes in a different way, via an ordered vector $(x_1,\ldots,x_n)$ representing how many times an element in $\{1,\ldots,4\}$ occurs. For example, $\{3,3,4\}$ corresponds to $(0,0,2,1)$ and $\{1,2,4\}$ corresponds to $(1,1,0,1)$. Thus, we can count how many distinct vectors $(x_1,\ldots,x_n)$ there are such that the sum of the components is 3, and each x_i can take value 0,1,2 or 3. Another way of looking at this is to consider placing $k = 3$ balls into $n = 4$ urns, numbered 1,. . . ,4. Then $(0,0,2,1)$ means that the third urn has 2 balls and the fourth urn has 1 ball. One way to distribute the balls over the urns is to distribute $n - 1 = 3$ "separators" and $k = 3$ balls over $n - 1 + k = 6$ positions, as indicated in Figure 1.13.

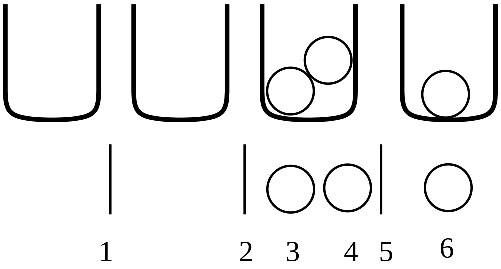


Figure 1.13: distributing k balls over n urns

The number of ways this can be done is the equal to the number of ways k positions for the balls can be chosen out of $n-1+k$ positions, that is, $\binom{n+k-1}{k}$. We thus have:

The number of different sets $\{x_1,\ldots,x_k\}$ with $x_i \in \{1,\ldots,n\}$, $i = 1,\ldots,k$ is

$$ \binom{n+k-1}{k} \, . $$

Returning to our original four problems, we can now solve them easily:

1. The total number of ways the exam can be completed is $3^{20} = 3,486,784,401$.
2. The number of placings is ${}^{8}P_3 = 336$.
3. The number of possible combinations of CDs is $\binom{20}{3} = 1140$.
4. The number of different throws with three dice is $\binom{8}{3} = 56$.

More examples

Here are some more examples. Not all problems can be directly related to the 4 problems above. Some require additional reasoning. However, the counting principles remain the same.

1. In how many ways can the numbers 1,. . . ,5 be arranged, such as 13524, 25134, etc?

   Answer: $5! = 120$.

2. How many different arrangements are there of the numbers 1,2,. . . ,7, such that the first 3 numbers are 1,2,3 (in any order) and the last 4 numbers are 4,5,6,7 (in any order)?

   Answer: $3! \times 4!$.

3. How many different arrangements are there of the word “arrange”, such as “aarrnge”, “arrngea”, etc?

   Answer: Convert this into a ball drawing problem with 7 balls, numbered 1,. . . ,7. Balls 1 and 2 correspond to ’a’, balls 3 and 4 to ’r’, ball 5 to ’n’, ball 6 to ’g’ and ball 7 to ’e’. The total number of permutations of the numbers is 7!. However, since, for example, (1,2,3,4,5,6,7) is identical to (2,1,3,4,5,6,7) (when substituting the letters back), we must divide 7! by $2! \times 2!$ to account for the 4 ways the two ’a’s and ’r’s can be arranged. So the answer is $7!/4 = 1260$.

4. An urn has 1000 balls, labelled 000, 001, . . . , 999. How many balls are there that have all number in ascending order (for example 047 and 489, but not 033 or 321)?

   Answer: There are $10 \times 9 \times 8 = 720$ balls with different numbers. Each triple of numbers can be arranged in $3! = 6$ ways, and only one of these is in ascending order. So the total number of balls in ascending order is $720/6 = 120$.

5. In a group of 20 people each person has a different birthday. How many different arrangements of these birthdays are there (assuming each year has 365 days)?

   Answer: ${}^{365}P_{20}$.

Once we've learned how to count, we can apply the equilikely principle to calculate probabilities:

1. What is the probability that out of a group of 40 people all have different birthdays?

   Answer: Choosing the birthdays is like choosing 40 balls with replacement from an urn containing the balls 1,...,365. Thus, our sample space Ω consists of vectors of length 40, whose components are chosen from $\{1, \dots, 365\}$. There are $|\Omega| = 365^{40}$ such vectors possible, and all are *equally likely.* Let A be the event that all 40 people have different birthdays. Then, $|A| = {}^{365}P_{40} = 365!/325!$ It follows that $\mathbb{P}(A) = |A|/|\Omega| \approx 0.109$, so not very big!

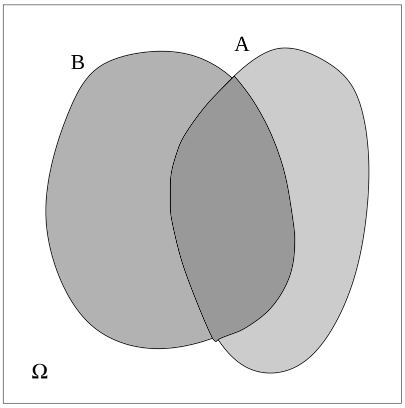
2. What is the probability that in 10 tosses with a fair coin we get exactly 5 Heads and 5 Tails?

   Answer: Here Ω consists of vectors of length 10 consisting of 1s (Heads) and 0s (Tails), so there are 2^{10} of them, and all are *equally likely.* Let A be the event of exactly 5 heads. We must count how many binary vectors there are with exactly 5 1s. This is equivalent to determining in how many ways the positions of the 5 1s can be chosen out of 10 positions, that is, $\binom{10}{5}$. Consequently, $\mathbb{P}(A) = \binom{10}{5}/2^{10} = 252/1024 \approx 0.25$.

3. We draw at random 13 cards from a full deck of cards. What is the probability that we draw 4 Hearts and 3 Diamonds?

   Answer: Give the cards a number from 1 to 52. Suppose 1–13 is Hearts, 14–26 is Diamonds, etc. Ω consists of unordered sets of size 13, without repetition, e.g., $\{1, 2, \dots, 13\}$. There are $|\Omega| = \binom{52}{13}$ of these sets, and they are all equally likely. Let A be the event of 4 Hearts and 3 Diamonds. To form A we have to choose 4 Hearts out of 13, and 3 Diamonds out of 13, followed by 6 cards out of 26 Spade and Clubs. Thus, $|A| = \binom{13}{4} \times \binom{13}{3} \times \binom{26}{6}$. So that $\mathbb{P}(A) = |A|/|\Omega| \approx 0.074$.

2.6 Conditional probability and independence



How do probabilities change when we know some event $B \subset \Omega$ has occurred? Suppose B has occurred. Thus, we know that the outcome lies in B. Then A will occur if and only if $A \cap B$ occurs, and the relative chance of A occurring is therefore

$$\mathbb{P}(A \cap B)/\mathbb{P}(B).$$

This leads to the definition of the **conditional probability** of A given B:

$$\boxed{\mathbb{P}(A\,|\,B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}} \qquad 2.4$$

Example 2.10 We throw two dice. Given that the sum of the eyes is 10, what is the probability that one 6 is cast?

Let B be the event that the sum is 10,

$$B = \{(4,6),(5,5),(6,4)\}.$$

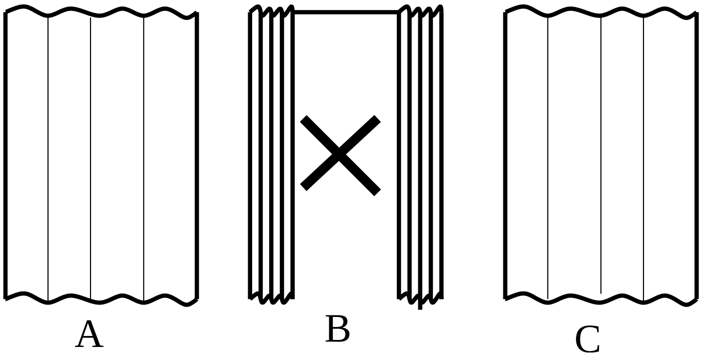
Let A be the event that one 6 is cast,

$$A = \{(1,6),\dots,(5,6),(6,1),\dots,(6,5)\}.$$

Then, $A \cap B = \{(4,6),(6,4)\}$. And, since all elementary events are equally likely, we have

$$\mathbb{P}(A\,|\,B) = \frac{2/36}{3/36} = \frac{2}{3}.$$

Example 2.11 (Monte Hall Problem) This is a nice application of conditional probability. Consider a quiz in which the final contestant is to choose a prize which is hidden behind one three curtains (A, B or C). Suppose without loss of generality that the contestant chooses curtain A. Now the quiz master (Monte Hall) always opens one of the other curtains: if the prize is behind B, Monte opens C, if the prize is behind C, Monte opens B, and if the prize is behind A, Monte opens B or C with equal probability, e.g., by tossing a coin (of course the contestant does not see Monte tossing the coin!).



Suppose, again without loss of generality that Monte opens curtain B. The contestant is now offered the opportunity to switch to curtain C. Should the contestant stay with his/her original choice (A) or switch to the other unopened curtain (C)?

Notice that the sample space consists here of 4 possible outcomes: Ac: The prize is behind A, and Monte opens C; Ab: The prize is behind A, and Monte opens B; Bc: The prize is behind B, and Monte opens C; and Cb: The prize is behind C, and Monte opens B. Let A, B, C be the events that the prize is behind A, B and C, respectively. Note that $A = \{Ac, Ab\}$, $B = \{Bc\}$ and $C = \{Cb\}$, see Figure 1.14.

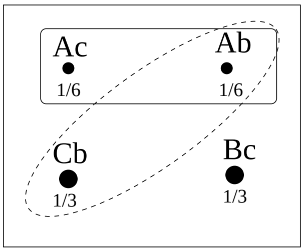


Figure 2.14: The sample space for the Monte Hall problem.

Now, obviously $\mathbb{P}(A) = \mathbb{P}(B) = \mathbb{P}(C)$, and since Ac and Ab are equally likely, we have $\mathbb{P}(\{Ab\}) = \mathbb{P}(\{Ac\}) = 1/6$. Monte opening curtain B means that we have information that event $\{Ab, Cb\}$ has occurred. The probability that the prize is under A given this event, is therefore

$$\mathbb{P}(A \,|\, \text{B is opened}) = \frac{\mathbb{P}(\{Ac, Ab\} \cap \{Ab, Cb\})}{\mathbb{P}(\{Ab, Cb\})} = \frac{\mathbb{P}(\{Ab\})}{\mathbb{P}(\{Ab, Cb\})} = \frac{1/6}{1/6 + 1/3} = \frac{1}{3}.$$

This is what we expected: the fact that Monte opens a curtain does not give us any extra information that the prize is behind A. So one could think that it doesn't matter to switch or not. But wait a minute! What about $\mathbb{P}(B \,|\, \text{B is opened})$? Obviously this is 0 — opening curtain B means that we know that event B cannot occur. It follows then that $\mathbb{P}(C \,|\, \text{B is opened})$ must be 2/3, since a conditional probability behaves like any other probability and must satisfy axiom 2 (sum up to 1). Indeed,

$$\mathbb{P}(C \,|\, \text{B is opened}) = \frac{\mathbb{P}(\{Cb\} \cap \{Ab, Cb\})}{\mathbb{P}(\{Ab, Cb\})} = \frac{\mathbb{P}(\{Cb\})}{\mathbb{P}(\{Ab, Cb\})} = \frac{1/3}{1/6 + 1/3} = \frac{2}{3}.$$

Hence, given the information that B is opened, it is twice as likely that the prize is under C than under A. Thus, the contestant should switch!

2.6.1 Product Rule

By the definition of conditional probability we have

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B \,|\, A). \tag{1.5}$$

We can generalise this to n intersections $A_1 \cap A_2 \cap \cdots \cap A_n$, which we abbreviate as $A_1 A_2 \cdots A_n$. This gives the **product rule** of probability (also called *chain rule*).

Theorem 2.3 (Product rule) Let $A_1, \ldots, A_n$ be a sequence of events with $\mathbb{P}(A_1 \ldots A_{n-1}) > 0$. Then,

$$\boxed{\mathbb{P}(A_1 \cdots A_n) = \mathbb{P}(A_1)\,\mathbb{P}(A_2 \,|\, A_1)\,\mathbb{P}(A_3 \,|\, A_1 A_2) \cdots \mathbb{P}(A_n \,|\, A_1 \cdots A_{n-1}).} \tag{1.6}$$

Proof. We only show the proof for 3 events, since the $n > 3$ event case follows similarly. By applying (1.4) to $\mathbb{P}(B \mid A)$ and $\mathbb{P}(C \mid A \cap B)$, the left-hand side of (1.6) is we have,

$$\mathbb{P}(A)\,\mathbb{P}(B \mid A)\,\mathbb{P}(C \mid A \cap B) = \mathbb{P}(A)\frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)}\frac{\mathbb{P}(A \cap B \cap C)}{\mathbb{P}(A \cap B)} = \mathbb{P}(A \cap B \cap C)\ ,$$

which is equal to the left-hand size of (1.6). ■

Example 2.12 We draw consecutively 3 balls from a bowl with 5 white and 5 black balls, without putting them back. What is the probability that all balls will be black?

Solution: Let A_i be the event that the ith ball is black. We wish to find the probability of $A_1A_2A_3$, which by the product rule (1.6) is

$$\mathbb{P}(A_1)\,\mathbb{P}(A_2 \mid A_1)\,\mathbb{P}(A_3 \mid A_1A_2) = \frac{5}{10}\,\frac{4}{9}\,\frac{3}{8} = 0.083.$$

Note that this problem can also be easily solved by counting arguments, as in the previous section.

Example 2.13 (Birthday Problem) In Section 1.5 we derived by counting arguments that the probability that all people in a group of 40 have different birthdays is

$$\frac{365 \times 364 \times \cdots \times 326}{365 \times 365 \times \cdots \times 365} \approx 0.109. \tag{1.7}$$

We can derive this also via the product rule. Namely, let A_i be the event that the first i people have different birthdays, $i = 1, 2, \ldots$. Note that $A_1 \supset A_2 \supset A_3 \supset \cdots$. Therefore $A_n = A_1 \cap A_2 \cap \cdots \cap A_n$, and thus by the product rule

$$\mathbb{P}(A_{40}) = \mathbb{P}(A_1)\mathbb{P}(A_2 \mid A_1)\mathbb{P}(A_3 \mid A_2) \cdots \mathbb{P}(A_{40} \mid A_{39})\ .$$

Now $\mathbb{P}(A_k \mid A_{k-1} = (365 - k + 1)/365$ because given that the first $k-1$ people have different birthdays, there are no duplicate birthdays if and only if the birthday of the k-th is chosen from the $365 - (k-1)$ remaining birthdays. Thus, we obtain (1.7). More generally, the probability that n randomly selected people have different birthdays is

$$\mathbb{P}(A_n) = \frac{365}{365} \times \frac{364}{365} \times \frac{363}{365} \times \cdots \times \frac{365 - n + 1}{365}, \quad n \geq 1\ .$$

A graph of $\mathbb{P}(A_n)$ against n is given in Figure 1.15. Note that the probability $\mathbb{P}(A_n)$ rapidly decreases to zero. Indeed, for $n = 23$ the probability of having no duplicate birthdays is already less than 1/2.

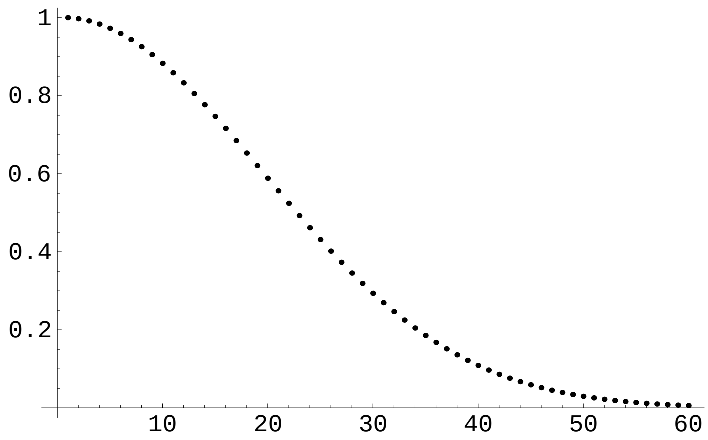


Figure 2.15: The probability of having no duplicate birthday in a group of n people, against n.

2.6.2 Law of Total Probability and Bayes' Rule

Suppose $B_1, B_2, \ldots, B_n$ is a **partition** of Ω. That is, $B_1, B_2, \ldots, B_n$ are disjoint and their union is Ω, see Figure 1.16

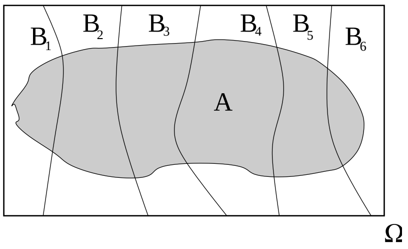


Figure 2.16: A partition of the sample space

Then, by the sum rule, $\mathbb{P}(A) = \sum_{i=1}^{n} \mathbb{P}(A \cap B_i)$ and hence, by the definition of conditional probability we have

$$\boxed{\mathbb{P}(A) = \sum_{i=1}^{n} \mathbb{P}(A|B_i)\, \mathbb{P}(B_i)}$$

This is called the **law of total probability**.

Combining the Law of Total Probability with the definition of conditional probability gives **Bayes' Rule**:

$$\boxed{\mathbb{P}(B_j|A) = \frac{\mathbb{P}(A|B_j)\, \mathbb{P}(B_j)}{\sum_{i=1}^{n} \mathbb{P}(A|B_i)\mathbb{P}(B_i)}}$$

Example 2.14 A company has three factories (1, 2 and 3) that produce the same chip, each producing 15%, 35% and 50% of the total production. The probability of a defective chip at 1, 2, 3 is 0.01, 0.05, 0.02, respectively. Suppose someone shows us a defective chip. What is the probability that this chip comes from factory 1?

Let B_i denote the event that the chip is produced by factory i. The $\{\mathcal{B}_{\rangle}\}$ form a partition of Ω. Let A denote the event that the chip is faulty. By Bayes' rule,

$$\mathbb{P}(B_1 \mid A) = \frac{0.15 \times 0.01}{0.15 \times 0.01 + 0.35 \times 0.05 + 0.5 \times 0.02} = 0.052 \ .$$

2.6.3 Independence

Independence is a very important concept in probability and statistics. Loosely speaking it models the *lack of information* between events. We say A and B are *independent* if the knowledge that A has occurred does not change the *probability* that B occurs. That is

$$A,\, B \text{ independent} \Leftrightarrow \mathbb{P}(A|B) = \mathbb{P}(A)$$

Since $\mathbb{P}(A|B) = \mathbb{P}(A \cap B)/\mathbb{P}(B)$ an alternative definition of independence is

$$\boxed{A,\, B \text{ independent} \Leftrightarrow \mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)}$$

This definition covers the case $B = \emptyset$ (empty set). We can extend the definition to arbitrarily many events:

Definition 2.3 The events $A_1, A_2, \ldots,$ are said to be **(mutually) independent** if for any n and any choice of distinct indices $i_1, \ldots, i_k$,

$$\mathbb{P}(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}) = \mathbb{P}(A_{i_1})\mathbb{P}(A_{i_2}) \cdots \mathbb{P}(A_{i_k}) \ .$$

Remark 2.1 In most cases independence of events is a **model assumption**. That is, we assume that there exists a $\mathbb{P}$ such that certain events are independent.

Example 2.15 (A Coin Toss Experiment and the Binomial Law) We flip a coin n times. We can write the sample space as the set of binary n-tuples:

$$\Omega = \{(0, \ldots, 0), \ldots, (1, \ldots, 1)\} \ .$$

Here 0 represent Tails and 1 represents Heads. For example, the outcome $(0, 1, 0, 1, \ldots)$ means that the first time Tails is thrown, the second time Heads, the third times Tails, the fourth time Heads, etc.

How should we define $\mathbb{P}$? Let A_i denote the event of Heads during the ith throw, $i = 1, \ldots, n$. Then, $\mathbb{P}$ should be such that the events $A_1, \ldots, A_n$ are *independent*. And, moreover, $\mathbb{P}(A_i)$ should be the same for all i. We don't know whether the coin is fair or not, but we can call this probability p $(0 \leq p \leq 1)$.

These two rules completely specify $\mathbb{P}$. For example, the probability that the first k throws are Heads and the last $n-k$ are Tails is

$$\begin{aligned} \mathbb{P}(\{(1,1,\dots,1,0,0,\dots,0)\}) &= \mathbb{P}(A_1)\cdots\mathbb{P}(A_k)\cdots\mathbb{P}(A_{k+1}^c)\cdots\mathbb{P}(A_n^c) \\ &= p^k(1-p)^{n-k}. \end{aligned}$$

Also, let B_k be the event that there are k Heads in total. The probability of this event is the sum the probabilities of elementary events $\{(x_1,\dots,x_n)\}$ such that $x_1+\cdots+x_n=k$. Each of these events has probability $p^k(1-p)^{n-k}$, and there are $\binom{n}{k}$ of these. Thus,

$$\mathbb{P}(B_k) = \binom{n}{k} p^k(1-p)^{n-k}, \quad k=0,1,\dots,n\ .$$

We have thus discovered the **binomial distribution**.