Probability: a short introduction.

Neil Walton

November 18, 2021

Contents

1 Preliminaries

Before we get going with these notes

General Advice

Do the following:

1. Make your own notes, particularly while watching lectures.
2. Read your notes after lectures (and before the next lecture).
3. Try exercises (to test understanding, as well as, for exams).
4. If you are ahead, read around (books, lectures, videos...).[1]
5. If you are behind, prioritize the most recent material.

Repeat each week.

The above advice applies to every university course you take.

Notation.

I assume that you know about:

- sets and special sets of numbers, e.g. $\mathbb{N} = \{n : n \in \mathbb{Z}, n > 0\}$.
- sets, $\{1,2,3,4\} = \{4,3,2,1\}$ and ordered lists $(1,2,3,4) \neq (4,3,2,1)$.
- functions, E.g. $f : \mathbb{R} \longrightarrow [0,\infty)$, $f(x) = x^2$.
- products, $\prod$, and sums, $\sum$.
- differentiation and integration of standard functions like x^n, e^x.
- you know roughly what is meant by $1 - \frac{1}{n} \longrightarrow 1$ as $n \to \infty$, i.e. that $1 - \frac{1}{n}$ gets closer and closer to 1 as n gets larger and larger.

Although each of these are introduced in other courses, all of these concepts are discussed in the appendix. I assume that you have read the appendix or are willing to refer to the appendix if and when required.

[1] This should be used to expand your own notes.

2 What is Probability?

I throw a coin 100 times. I got 52 heads.

Question. How many heads should I expect?

Answer. 50 heads should be expected.

Another experiment, I throw a dice.

Question. If I throw the dice forever, then what proportion of the throws are a 5.

Answer. The probability of this outcome is 1/6.

A slightly tougher one: I stop and start a stopwatch and look at the last digit.

Question. What is the probability that this digit is even?

Answer. It's 1/2. Because half the numbers are even. More precisely there are 5 even numbers each having probability 1/10 (as there are 10 digits). So

$$\begin{aligned}\mathbb{P}(\text{even}) =& \mathbb{P}(0) + \mathbb{P}(2) + \mathbb{P}(4) + \mathbb{P}(6) + \mathbb{P}(8)\\ =& \frac{1}{10} + \frac{1}{10} + \frac{1}{10} + \frac{1}{10} + \frac{1}{10}\\ =& \frac{5}{10}\\ =& \frac{1}{2}\end{aligned}$$

Discussion

Above we introduced various pieces of terminology:

experiment, *outcome*, *probability*, *expectation*...

We will define these more precisely soon.

In the dice question, we see that probabilities can be thought of as an idealized proportion, when we repeat an experiment infinitely many times. Since it is a proportion, notice the probabilities are less than 1.

In both the dice and stopwatch question, notice that counting was useful to us. E.g. In the stopwatch question, we counted the number of outcomes of interest (the 5 even digits) and the total number of outcome (the 10 possible digits). The probability of an even number was the ratio of these 5/10. In general, counting is an important starting point in probability.[2]

Notice that in the stopwatch question, the stopwatch is deterministic. However, our interaction with the stopwatch introduces randomness. Analogously, the particles of air in room can be argued to move deterministically, but small perturbations of the system move towards a state were the particles are uniformly random in the room.

When we discuss randomness colloquially, often we think of it as something than cannot be known. However, it is important to note that, when studying probability (and randomness), our uncertainty is quantifiable. Thus we can reason about randomness mathematically. The point of this course is to introduce initial concepts and principles in probability.

Beyond this course, it is worth noting that probability has many applications in statistics, finance and gambling, game theory, algorithm design, operational logistics, physics, machine learning...

[2] Although we initially spend a fair bit of time on counting, it must be added that there is much more to probability than counting.

3 Probability Terminology and Definitions

In probability, we consider an *experiment*. E.g. we throw two dice and add the total.

An *outcome* is the result of the experiment. E.g. if the first dice is a 5 and the 2nd is 2 then the outcome is $7(= 5 + 2)$.

The *sample space* is the set of possible outcomes, e.g. $\Omega = \{2, 3, 4, ..., 12\}$.

An *event* is a subset of outcomes from the sample space. E.g. $E = \{7\}$, $E = \{4, 7\}$, $E = \{\text{Even number}\}$.

We can define an event by explicitly listing the outcomes, e.g. $E = \{2, 4, 6, 8, 10, 12\}$, or by implicitly stating the outcomes, e.g. we can also write $E = \{ \text{ Even number } \}$.)

For a given set of events, there may be more than one way to define the sample space of an experiment. E.g. if I want to know the sum of two dice, we could consider the set of outcomes for the first and second dice throw. (See table below)

Examples

Experiment	Sample Space
Throw two coins	$\Omega = \{HH, HT, TH, TT\}$
Throw two dice	$\Omega = \{(1,1),(1,2),\cdots,(1,6),$ $(2,1),(2,2),\cdots,(2,6),$ $\vdots$ $(6,1),(6,2),\cdots(6,6)\}$
Number of clicks on a Gieger counter	$\{0, 1, 2, 3...\}$
Heights of people	$\{x : x \geq 0\}$ or $\mathbb{R}$

From the above table, note that sample spaces can be finite, (countably) infinite, or a continuum.

Definition of Discrete Probability. For finite of countably infinite sample spaces, we can define probabilities as follows.

Definition 1 (Probability – Discrete)**.** *For a sample space* $\Omega = \{\omega_1, \omega_2, \omega_3, \ldots\}$*, probabilities are numbers* $\mathbb{P}(\omega)$ *for each* $\omega \in \Omega$ *such that*

- *(Positive) For* $\omega \in \Omega$*,*
$$\mathbb{P}(\omega) \geq 0,$$
- *(Sums to one)*
$$\sum_{\omega \in \Omega} \mathbb{P}(\omega) = 1.$$

For events, $E \subseteq \Omega$*, we get the probability of the event by summing*

$$\mathbb{P}(E) = \sum_{\omega \in E} \mathbb{P}(\omega)\,.$$

The above is a good definition of for finite (or countably infinite) sample spaces. When we consider probabilities for continuous sample spaces definitions need to be modified.

An informal definition. The above definition gives us a working mathematical definition for probability. That said it is worth noting that intuitively we consider probabilities to represent the long-run proportion of time an event (or outcome) has occurred in an experiment. So informally if we repeat a number of experiments, which we denote by #{experiment}, and for those we count the number of times an event occurs #{event E occurs}, and if we let the number of experiments get large, that is #{experiment} $\longrightarrow \infty$, then it should hold that

$$\frac{\#\{\text{event } E \text{ occurs}\}}{\#\{\text{experiment}\}} \longrightarrow \mathbb{P}(E) \tag{1}$$

Later when we are a bit more precise about what we mean to "repeat a number of experiments", the above statement (1) will more formally be called the Law of Large Numbers.

Examples.

Example 1. *For the experiment where we throw two coins, calculate*

$$\mathbb{P}(\textit{at least one heads}).$$

Answer 1. *For the sample space $\Omega = \{HH, HT, TH, TT\}$, each probability is equally likely, i.e.*

$$p = \mathbb{P}(HH) = \mathbb{P}(HT) = \mathbb{P}(TH) = \mathbb{P}(TT)\,.$$

Also probabilities sum to one so

$$1 = \mathbb{P}(HH) + \mathbb{P}(HT) + \mathbb{P}(TH) + \mathbb{P}(TT)$$

This implies $4p = 1$ and so $p = \frac{1}{4}$.

From this point there are two ways to solve the question:

1. *Since $\{\textit{at least one head}\} = \{HH, HT, TH\}$, we can directly sum over the outcomes in the event*

$$\begin{aligned}\mathbb{P}(\textit{at least one heads}) =&\mathbb{P}(HH) + \mathbb{P}(HT) + \mathbb{P}(TH)\\ =&\frac{1}{4} + \frac{1}{4} + \frac{1}{4}\\ =&\frac{3}{4}\end{aligned}$$

2. *Since probabilities sum to one*

$$1 = \sum_{\omega\in\Omega} \mathbb{P}(\omega) = \mathbb{P}(\textit{at least one heads}) + \mathbb{P}(TT)\,,$$

Thus

$$\mathbb{P}(\textit{at least one heads}) = 1 - \mathbb{P}(TT) = 1 - \frac{1}{4} = \frac{3}{4}.$$

Example 2. *A bag contains three green balls and a red ball. Two balls are taken out at random what is the probability that both are green?*

Answer 2. *Here are three ways to answer this question:*
1. We can explicitly list by taking the balls out one at a time and count. Here we label the three green balls G_1, G_2, G_3 and the red ball R. The probability space is

$$\begin{array}{lllll} \Omega = \{ & & (R,G_1), & (R,G_2), & (R,G_3) \\ & (G_1,R) & & ,(G_1,G_2), & (G_1,G_3) \\ & (G_2,R), & (G_2,G_1) & & ,(G_2,G_3) \\ & (G_3,R), & (G_3,G_1), & (G_3,G_2) & \} \end{array}$$

There are 12 equally likely outcomes and 6 outcomes with both green so

$$\mathbb{P}(\textit{ both green }) = \frac{6}{12} = \frac{1}{2}.$$

(Notice we had to label the three balls G_1, G_2, G_3 because if we did not, then $\mathbb{P}((G,R)) \neq \mathbb{P}((R,G))$. So we could not count up events with equal probability.)

2. We can imagine we take out the balls simultaneously. Again we label the three green balls G_1, G_2, G_3 and the red ball R. Recall we use curly brackets sets where the order does not matter, and we use round brackets when the order matters. (E.g. $(G_1, G_2) \neq (G_2, G_1)$ but $\{G_1, G_2\} = \{G_2, G_1\}$). The probability space in this case is

$$\begin{array}{l} \Omega = \{\{R,G_1\}, \{R,G_2\}, \{R,G_3\} \\ \qquad\qquad ,\{G_1,G_2\}, \{G_1,G_3\} \\ \qquad\qquad\qquad\qquad ,\{G_2,G_3\}\} \end{array}$$

There are 6 equally likely outcomes and 3 outcomes with both green so

$$\mathbb{P}(\textit{ both green }) = \frac{3}{6} = \frac{1}{2}.$$

3. We can reason as follows. The probability the first ball removed is green is $3/4$, as three out of four balls are green. Given the first ball is green, the probability the 2nd ball is green is $2/3$, as now two out of three balls are green. So out of the three quarters of the time where the first ball is green, two thirds of the time the 2nd ball is green. Two thirds of three quarters is a half. So

$$\mathbb{P}(\textit{ both green }) = \frac{3}{4} \times \frac{2}{3} = \frac{1}{2}.$$

NSW

(This third argument might feel a little vague at first. We go into this in more detail when we discuss conditional probability, a bit later.)

4 Probability and Set Operations

We want to calculate probabilities for different events. Events are sets of outcomes, and we recall that there are various ways of combining sets. The current section is a bit abstract but will become more useful for concrete calculations later.

Operations on Events.

Definition 2 (Union). *For events $A, B \subseteq \Omega$, the union of A and B is the set of outcomes that are in A or B. The union is written*

$$A \cup B\,.$$

We can say "A or B" as well as "A union B".

Definition 3 (Intersection). *For events $A, B \subseteq \Omega$, the intersections of A and B is the set of outcomes that are in both A and B. The union is written*

$$A \cap B\,.$$

We can say "A and B" as well as "A intersection B".

Definition 4 (Complement). *For event $A \subseteq \Omega$, the complement of A is the set of outcomes in the sample space Ω that are not in A. We denote the complement of A with*[3]

$$A^c\,.$$

We can say "not A".

Definition 5 (Relative Complement). *For events $A, B \subseteq \Omega$, the relative complement of B relative to A is the set of outcomes not in B that are in A. The relative complement of B relative to AA is written*

$$A \backslash B\,.$$

We can say "A not B".

Warning! Later we will define $\mathbb{P}(A|B)$ to denote the probability of A given B. The line "|" in $\mathbb{P}(A|B)$ is straight and whereas line "\" in $A\backslash B$ in the relative complement is not a straight. So be warned that, in general,

$$\mathbb{P}(A|B) \neq \mathbb{P}(A\backslash B)\,.$$

[3] Other notations exist for the complement such as $\bar{A}$.

Here are a few facts about sets

Lemma 1. *For events $A, B, C \subseteq \Omega$*
i)

$$A \cap B = B \cap A, \quad A \cup B = B \cup A, \quad (A^c)^c = A\,.$$

ii)

$$\begin{aligned} A \cup (B \cup C) &= (A \cup B) \cup C \\ A \cap (B \cap C) &= (A \cap B) \cap C\,. \end{aligned}$$

iii)

$$\begin{aligned} A \cap (B \cup C) &= (A \cap B) \cup (A \cap C) \\ A \cup (B \cap C) &= (A \cup B) \cap (A \cup C)\,. \end{aligned}$$

iv) (De Morgan's Laws)

$$\begin{aligned} (A \cup B)^c &= A^c \cap B^c \\ (A \cap B)^c &= A^c \cup B^c\,. \end{aligned}$$

v) If $A \subseteq B$ then $B^c \subseteq A^c$.
vi) $A \backslash B = A \cap B^c$.

Remarks.

- The proof of each statement above is quite straightforward. For each statement, we can draw a Venn diagram (i.e. we draw 1,2, or 3 intersecting circles and shade the appropriate areas to check the result.) For instance, the first statement in Lemma 1iii) can be seen in Figure 1, below.
- From i) and ii) we see that it does not matter what order we apply unions and intersections. For that reason, when we apply a sequence of unions or a sequence of intersections we can unambiguously write

$$\bigcup_{i=1}^{\infty} A_i \qquad \text{and} \qquad \bigcap_{n=1}^{\infty} A_n\,.$$

- The statement iii) is analogous to the statement that $a \times (b + c) = a \times b + a \times c$.

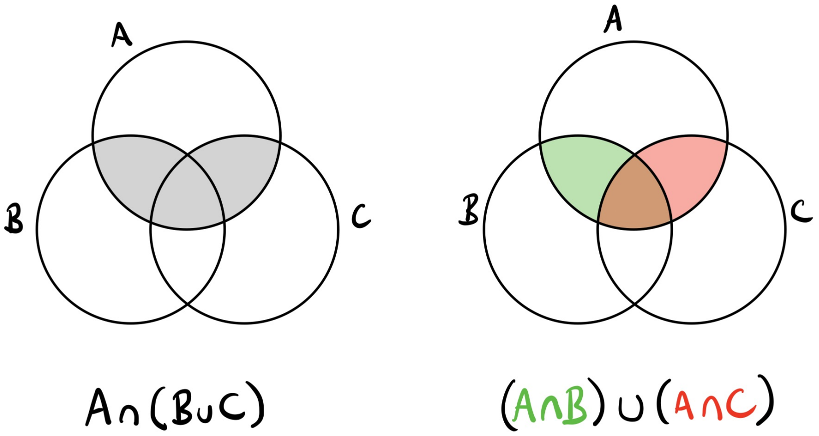


Figure 1: A Venn Diagram verifying that $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$

• In De Morgan's Laws, iv), we can think of the complement of union being intersection. I.e. $\cup^c = \cap$ and $\cap^c = \cup$. [4] Thus

$$(A \cup B)^c = A^c \cup^c B^c = A^c \cap B^c\,.$$

[4] This is not a mathematically precise statement but is useful for remembering the rule.

Probability Rules for Operations on Sets

We can now start to think about what these operations on events imply for the probabilities of those events. Here are a sequence of lemma to this effect. (You can skip the proofs on first reading.)

Lemma 2. *If A and B have not outcome in common, i.e. $A \cap B = \emptyset$, then*[5]

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)\,.$$

Proof.

$$\mathbb{P}(A \cup B) = \sum_{\omega \in A \cup B} \mathbb{P}(\omega) = \sum_{\omega \in A} \mathbb{P}(\omega) + \sum_{\omega \in B} \mathbb{P}(\omega) = \mathbb{P}(A) + \mathbb{P}(B)\,.$$

□

Lemma 2 can be extended to a countable set of sets $A_1, A_2, \ldots$ with $A_n \cap A_m = \emptyset$. So

$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mathbb{P}(A_n)\,.$$

The above equality is actually an axiom of probability.

Lemma 3. *If $A \subseteq B$ then $\mathbb{P}(A) \leq \mathbb{P}(B)$*

Proof. Using Lemma 1ii) and v),

$$B = (A \cap B) \cup (A^c \cap B) = A \cup (B \backslash A)\,.$$

Using Lemma 2 and the fact probabilities are non-negative

$$\mathbb{P}(B) = \mathbb{P}(A) + \mathbb{P}(B \backslash A) \geq \mathbb{P}(A)\,.$$

□

Suppose I have a statement like "I have a red car" this implies "I have a car". So the probability "I have a car" is more than the probability "I have a red car". The above shows that this inequality holds for any two statements where one implies the other.

---

[5] Here recall that $\emptyset$ is the empty set, which is the set containing no outcomes.

Lemma 4.
$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)\,.$$

(An intuitive proof of this lemma: draw a Venn diagram with events A and B. Notice if we colour in A and then colour in B then we end up colouring $B \cap A$ twice. So if we want to count that region once, which we do for the event $A \cup B$, we need to subtract $B \cap A$ once.)

Proof. Note that

$$\mathbb{P}(A \cup B) = \mathbb{P}(A \backslash B) + \mathbb{P}(B \backslash A) + \mathbb{P}(A \cap B) \tag{2}$$

where as

$$\begin{aligned}\mathbb{P}(A) + \mathbb{P}(B) =& [\mathbb{P}(A \backslash B) + \mathbb{P}(A \cap B)] + [\mathbb{P}(B \backslash A) + \mathbb{P}(A \cap B)]\\ =& \mathbb{P}(A \backslash B) + \mathbb{P}(B \backslash A) + 2\mathbb{P}(A \cap B)\,. \end{aligned} \tag{3}$$

Subtracting (2) from (3) gives

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

Rearranging the above expression gives the result. □

It is often easy to calculate $\mathbb{P}(A \cap B)$ from $\mathbb{P}(A)$ and $\mathbb{P}(B)$ (see later discussion on independence). The above lemma gives a way to access $\mathbb{P}(A \cup B)$ from $\mathbb{P}(A \cap B)$.

Lemma 5.

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

Proof. $A \cup A^c = \Omega$ while $A \cap A^c = \emptyset$. Also $\mathbb{P}(\Omega) = 1$ since probabilities sum to one (see Definition 1). So

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

Rearranging gives the required result. □

Sometimes the number of outcomes in A is large. So we may need to sum the probability of a lot of different outcomes to get $\mathbb{P}(A)$. However, it may be that A^c only has a few outcomes. So it may be simple to calculate $\mathbb{P}(A^c)$. The above lemma then gives us a way to get $\mathbb{P}(A)$.

---

5 Counting Principles

Counting in Probability. If each outcome is equally likely, i.e. $\mathbb{P}(\omega) = p$ for all $\omega \in \Omega$, then since

$$1 = \sum_{\omega\in\Omega} \mathbb{P}(\omega) = \sum_{\omega\in\Omega} p = |\Omega| p$$

(where $|\Omega|$ is the number of outcomes in the set Ω) it must be that

$$\mathbb{P}(\omega) = \frac{1}{|\Omega|} \qquad \text{for all } \omega \in \Omega. \tag{4}$$

Further, since $\mathbb{P}(E) = \sum_{\omega\in E} \mathbb{P}(\omega)$ then

$$\mathbb{P}(E) = \frac{|E|}{|\Omega|}. \tag{5}$$

So from both (4) and (5), we see that, when outcomes are equally likely, then to calculate probabilities we need to be able to count the number of outcomes in different sets.

Some Counting Rules. 1,2,3,4,... aside, we cover the following counting methods

1. Multiplication
2. Factorials
3. Permutations
4. Combinations

Card Game. We consider the probability of winning a card game in the following setting:

> *A card dealer has* 5 *cards* $A, K, Q, 4, 2$.
> *You are dealt* 3 *cards.*
> *You win if you get* A, K, Q.

The probability of winning depends on what we mean by "the cards being dealt" and what is means by to "get A,K,Q".

Multiplication.

Rules of the the card game. Suppose we play the card game in the following way

> The dealer shuffles the deck. Shows a random card. Replaces it in the deck. And then repeats.
> You win if you get A then K and then Q, in that order.

The probability of winning. We can calculate the probability of winning as follows. The set of outcomes can be written as

$$\Omega = \{(C_1, C_2, C_3) : C_1, C_2, C_3 \in \{A, K, Q, 4, 2\}\}\,.$$

I.e. it is the set of triplets (C_1, C_2, C_3) where the first, second and third card are in $\{A, K, Q, 4, 2\}$. If holds that

$$|\Omega| = 5^3 = 125\,.$$

Since there is only one winning hand among these 125 equally likely possibilities, we have

$$\mathbb{P}(\text{ winning }) = \frac{1}{125}\,.$$

Note $|\Omega| = 5^3$ holds because.. Notice once we fix the first two cards (C_1, C_2) there are 5 different values for the third card. So the number of triplets (C_1, C_2, C_3) is five times the number of pairs (C_1, C_2). By the same logic, once we fix the first card C_1 there are five values that C_2 can take. So the number of pairs (C_1, C_2) is 5 times the number of values C_1 can take. Finally C_1 can take 5 values. So we see the number of triples is $5 \times 5 \times 5$. See Figure 2.

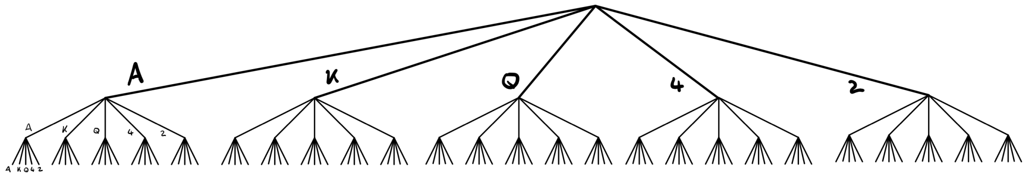


Figure 2: This card game has 5 possibilities at each stage.

In General. For sets $\Omega_1,\dots.,\Omega_k$, we define the product set by

$$\prod_{i=1}^{K} \Omega_i := \Big\{(\omega_1,\dots,\omega_k) : \omega_1 \in \Omega_1,\dots,\omega_k \in \Omega_k)\Big\}.$$

Following the argument as above we can see that

$$\left|\prod_{i=1}^{k} \Omega_i\right| = \prod_{i=1}^{k} |\Omega_i|.$$

For our card game, we had $k = 3$ and $|\Omega_1| = |\Omega_2| = |\Omega_3| = 5$. Later this principle will generalize into something that we call independence.

Factorials

Rules of the the card game. Suppose we play the card game in the following way

> The dealer shuffles the deck. Shows a random card. Replaces it in the deck. And then repeats.
> You win if you get A then K and then Q, in <u>any order</u>.

So now (Q, K, A) is a now a winning hand as well as (A, K, Q).

The probability of winning. Since the way that the dealer deals has not changed we know the probability of any hand, that is,

$$\mathbb{P}((A, K, Q)) = \cdots = \mathbb{P}((Q, K, A)) = \frac{1}{125}.$$

So if we want to calculate the probability of winning we need to calculate the number of events with a winning hand. It is not hard to check that

$$\{\text{ winning }\} = \{(A, K, Q), (A, Q, K), (K, A, Q), (K, Q, A), (Q, A, K), (Q, K, A)\}$$

That is there are 6 winning hands. So the probability of winning is

$$\mathbb{P}(\text{ winning }) = \frac{6}{125}.$$

Six winning hands holds because.. Notice there are 3 possibilities for the first card, where we are still on for a win, namely, A, K, Q. We can suppose without any lost generality that the first card was Q, then for the 2nd card there are 2 possibilities that can be dealt where we can still win namely A and K. Again we can suppose that the card was K, then for the 3rd card there is 1 possibility where we win, namely, A. So there if we look across the three stages of this card game there are 3 possibilities for the 1st card then for each of these there are 2 possibilities for the 2nd card and then 1 for the final card, which gives $3 \times 2 \times 1 = 6$ winning hands. See Figure 3

In general. Given a set C with $|C| = n$, which we can think of as a deck of cards. We are interested in counting the different of ways of dealing out the deck:

$$F := \{(c_1, c_2, ..., c_n) : c_1, ..., c_n \in C \text{ where } c_i \neq c_j \text{ for all } i \neq j\}$$

In general, following the argument as above it holds that

$$|F| = n \times (n-1) \times (n-2) \times .... \times 2 \times 1$$

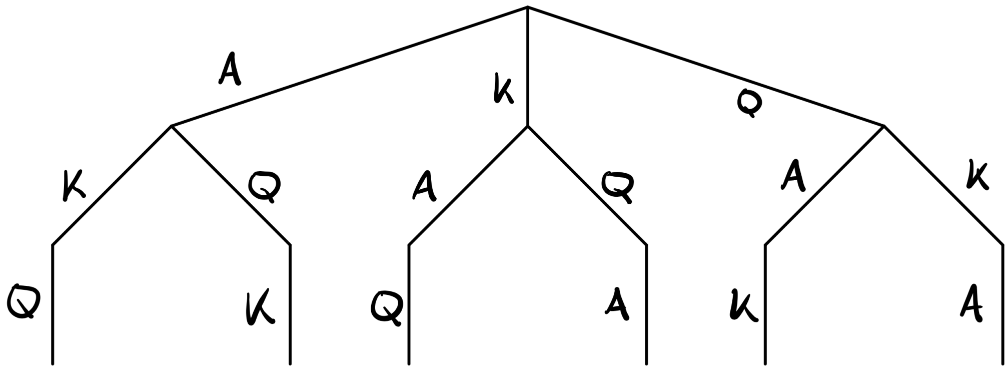


Figure 3: There are initially 3 potentially winning first cards, and the one less at each stage.

Definition 6 (Factorial)**.** *Given a positive integer n, n factorial is defined to be*

$$n! := n \times (n-1) \times (n-2) \times .... \times 2 \times 1$$

(By convention we define $0! := 1$*)*

So $n!$ counts the number of ways we can order a set of n elements.

Permutations

Rules of the the card game. Suppose we play the card game in the following way

> The dealer shuffles the deck. Takes out a random card, but does not put it back in the deck. And then repeats. You win if you get A then K and then Q, in that order.

The probability of winning. In this case, there is only one winning hand (A, K, Q). However, the set of hands has now changed as we can not longer repeat any card. This is similar to factorials, however, as you may recall, for a factorial we deal out all the cards, whereas here we only deal three out of the five cards. So here the sample space is

$$\Omega = \{(C_1, C_2, C_3) : C_1, C_2, C_3 \in \{A, K, Q, 4, 2\} \text{ where } C_i \neq C_j \text{ for all } i \neq j\}.$$

As we will argue shortly $|\Omega| = 60$. Thus

$$\mathbb{P}(\text{ winning }) = \frac{1}{60}.$$

It holds that $|\Omega| = 60$ because.. Similar our discussion on factorials, there are 5 possibilities for the first card dealt. Suppose for instance the first card is A, although the same argument would hold equally for any other card. Now once the first card is dealt there are 4 remaining possibilities for the second card, namely, $\{K, Q, 4, 2\}$. Suppose for instance the second card is a K. Then there are 3 possibilities for the third, namely, $\{Q, 4, 2\}$. In total this leads to $5 \times 4 \times 3 = 60$ possible ways of dealing 3 cards from a pack of 5 cards. See Figure 4

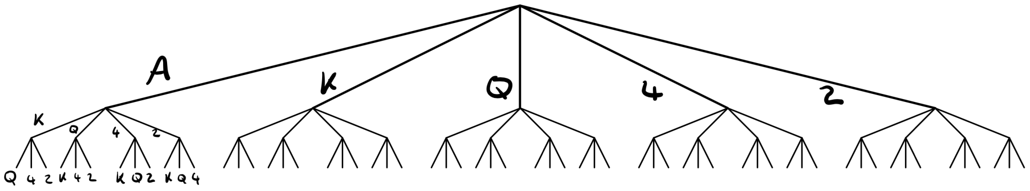


Figure 4: There are initially 5 potentially winning first cards, and the one less at each stage of the 3 stages.

In general. Given a set C with $|C| = n$, which we can think of as a deck of cards. We are interested in counting the different of ways of dealing out k cards without replacement:

$$P := \{(c_1, c_2, ..., c_k) : c_1, ..., c_k \in C \text{ where } c_i \neq c_j \text{ for all } i \neq j\}$$

In general, following the argument as above it holds that

$$|P| = n \times (n-1) \times (n-2) \times .... \times (n-k+2) \times (n-k+1)$$

Definition 7 (Permutation)**.** *Given non-negative integers n and k with $k \leq n$ the permuation is defined to be*

$$P^n_k := n \times (n-1) \times (n-2) \times .... \times (n-k+2) \times (n-k+1)$$

or more compactly using factorials:

$$P^n_k := \frac{n!}{(n-k)!}$$

So P^n_k counts the number of ways we can order k elements from a set of n .

Combinations.

Rules of the the card game. Suppose we play the card game in the following way

> The dealer shuffles the deck. Takes out a random card,
> but does not put it back in the deck. And then repeats.
> You win if you get A then K and then Q, in any order.

The probability of winning. There are two ways to think about the above game. We focus first on the way that is most similar to out previous argument on permutations.

As before the ways of dealing the cards is exactly the same as our discussion on permutations.

$$\Omega_1 = \{(C_1, C_2, C_3) : C_1, C_2, C_3 \in \{A, K, Q, 4, 2\} \text{ where } C_i \neq C_j \text{ for all } i \neq j\}.$$

Thus $|\Omega_1| = 5!/2! = 5 \times 4 \times 3 = 60$ for the same reason as previously. Further since the order of the cards $\{A, K, Q\}$ does not matter. For the same reason as our discussion on factorials there are $3! = 3{\times}2{\times}1 = 6$ winning hands. Thus

$$\mathbb{P}(\text{ winning }) = \frac{3! \cdot 2!}{5!} = \frac{6}{60} = \frac{1}{10}. \tag{6}$$

A further way of thinking: removing the order... Rather than thinking of removing cards one at a time, we could think of taking out three cards at the same time from the deck of five cards, without paying attention to the order that they came out. Here we are interest in the set of three cards removed, e.g. $\{A, K, Q\} = \{Q, K, A\}$, rather than the order that the cards where taken out, e.g. $(A, K, Q) \neq (Q, K, A)$. Here we could then think of a slightly different sample space:

$$\Omega_2 = \{\{C_1, C_2, C_3\} : C_1, C_2, C_3 \in \{A, K, Q, 4, 2\} \text{ where } C_i \neq C_j \text{ for all } i \neq j\}$$

We can think of this as dividing the deck of cards $\{A, K, Q, 4, 2\}$ into two decks: the cards dealt out, and the cards still in the deck, E.g.

$$\text{(the deck of cards)} \qquad \overset{|}{A}, \overset{\circ}{K}, \overset{|}{Q}, \overset{\circ}{4}, \overset{|}{2} \rightarrow \{\overset{|}{A}, \overset{|}{Q}, \overset{|}{2}\} \qquad \text{(cards dealt out)}$$
$$\searrow \{\overset{\circ}{K}, \overset{\circ}{2}\} \qquad \text{(cards in the deck)}$$

Here "|" indicates the card is dealt and "∘" indicates that the card stays in the pack.

In addition to the above argument that gave in (6) above. A way to think about counting the number of outcomes in Ω_2. Is to imagine we deal out the whole deck. The first three cards are given to the players and the last two cards remain with the dealer. The number of ways of dealing out the deck we know is 5! from our discussion on factorials. This over-counts as we don't care what the order is for the first 3 cards dealt or for the last 2 cards that remain in the deck. For example suppose we are dealt A, K, Q, there are 3! ways of dealing out A, K, Q as the first three cards, and, for each way the A, K, Q is dealt, for the cards $4, 2$ still in the deck, there are 2! ways of ordering these cards. In other words 5! over counts by a factor of $3! \times 2!$. From this we can see that

$$|\Omega_2| = \frac{5!}{3!2!} = 10\,.$$

There is only one outcome in Ω_2 where we win so

$$\mathbb{P}(\text{ winning }) = \frac{1}{10}\,.$$

In General. Given a set C with $|C| = n$, which we can think of as a deck of cards. We are interested in the number of ways of dealing k cards:

$$D := \{\{c_1, c_2, ..., c_k\} : c_1, ..., c_k \in C \text{ where } c_i \neq c_j \text{ for all } i \neq j\}$$

In general the argument above gives that

$$|D| = \frac{n!}{k!(n-k)!}\,.$$

Definition 8 (Combination). *Given a non-negative integers n and k with $k \leq n$, the combination of n choose k is defined to be*

$$C^n_k := \binom{n}{k} := \frac{n!}{k!(n-k)!}\,.$$

So C^n_k counts the number of ways we can chose k elements from a set of n .

Multinomials. In combinations the cards were dealt to one person (and the dealer) what if we wanted to consider dealing more hands. It is not hard to check that if there are n cards and $k_1, ..., k_m$ are the number of cards dealt to each player. (here we think of k_m as the number of cards that the dealer has left.) Note $k_1 + ... + k_m = n$. Then the natural extension of the combination, which gives the number of ways of dealing cards, as above is a multinomial which is given by

$$\binom{n}{k_1, ..., k_m} = \frac{n!}{k_1!k_2!...k_m!}\,.$$

Additional Remarks.*

Here are a few additional remarks.
(These remarks are not examinable. So you may choose to skip, but it might yet be helpful).

A bit more interpretation for combinations. Notice earlier we thought of a combination as dividing up the deck of five cards:

$$\overset{1}{A}, \overset{0}{K}, \overset{1}{Q}, \overset{0}{4}, \overset{1}{2}$$

Here "1" indicates the card is dealt and "0" indicates that the card stays in the pack. So we could think of C^n_k as counting the number of strings of n digits with exactly k ones. E.g. For $C^5_3 = 10$ because there are 10 such strings:

$$11100, 11010, 11001, 10110, 10101, 10011, 01110, 01101, 01011, 00111.$$

(For this reason you can see that $\sum_{k=0}^n C^n_k = 2^n$.)

Polynomials, Pascals Triangle. Consider the expansion of $(x+1)^n$. Notice that

$$(1 + x)^n = \binom{n}{0} + \binom{n}{1}x + \binom{n}{2}x^2 + ... + \binom{n}{n}x^n\,. \tag{7}$$

Why? Well for instance if we look at the x^3 term in $(1 + x)^5$ then to get one x^3 term we need to include three x terms and two 1 terms from the expansion of $(1 + x)^5$. For instance here is one example:

$$(1 + \overset{|}{x})(\overset{\circ}{1} + x)(1 + \overset{|}{x})(\overset{\circ}{1} + x)(1 + \overset{|}{x})\,.$$

This is the same as $\overset{|}{A}, \overset{\circ}{K}, \overset{|}{Q}, \overset{\circ}{4}, \overset{|}{2}$ from earlier so the number of such terms is C^5_3. This why the coefficients are combinations.

The Binomial Theorem. Notice from the above result we can see that

$$\binom{n+1}{k} = \binom{n}{k} + \binom{n}{k-1}$$

Hint: $(1+x)^{n+1} = (1+x)^n(1+x)$ now apply (7). This is called the Binomial Theorem.

Pascal's Triangle. If we repeatedly apply the above formula we get a nice shape for listing out combinations:

$$\begin{array}{c} 1 \\ 1 \quad 1 \\ 1 \quad 2 \quad 1 \\ 1 \quad 3 \quad 3 \quad 1 \\ 1 \quad 4 \quad 6 \quad 4 \quad 1 \\ 1 \quad 5 \quad 10 \quad 10 \quad 5 \quad 1 \\ \vdots \end{array}$$

This is called Pascal's Triangle.

Stirlings Formula. When n is large, $n!$ can be a really big number. (A good reason to have the exclamation mark.) We can quantify how much with the following formula

$$n! \sim \sqrt{(2\pi n)}e^{-n}n^n$$

Here the $\sim$ means that the ratio between the two numbers approaches 1 as n gets large.

Examples

Example 3. *What is the probability of winning the jackpot in the national lottery (lotto)?*
(Recall the national lottery 6 *random balls from a set of* 59 *balls. You win if you get all* 6 *correct, regardless of order.)*

Answer 3. *The number of draws in* $C_6^{59} = 45057474$ *and only one wins so* $\mathbb{P}(win) = 1/45057474$.

Example 4. *What is the probability of winning the jackpot in Euromillions?*
(Recall that Euromillions has 5 *main numbers between* 1 *and* 50 *and* 2 *"lucky stars" between* 1 *and* 12*.*

Answer 4. *The total number of combinations for the main numbers is* $C^{50}_{5} = 2118,760$ *and for the lucky stars* $C^{12}_{2} = 66$*. So in total* $C^{50}_{5} \times C^{12}_{2} = 139,838,160$*. So the probability is* $1/139838160$*.*

Example 5. *How many distinct words can be made out of* A, B, B, A*.*

Answer 5. $C^{4}_{2} = 6$ *. (Like above think of the positions* $(1,2,3,4)$ *of letters as cards and a card being dealt being the same as being assigned an* A *in that position.)*

Example 6. *How many distinct words can be made out of* B, A, N, A, N, A*.*

Answer 6. $6!/(1!3!2!) = 60$*. (Same principle as previous question but now is a multinomial.)*

Example 7 (Birthday Paradox)**.** *In a football game there are 23 players (including the referee). What is the probability that two or more players have the same birthday.*[6]

Answer 7. *Note that*

$$\mathbb{P}(\textit{two or more same birthday}\ ) = 1 - \mathbb{P}(\textit{all different})\,.$$

So we need to count the number of ways the birthdays can be organized and the number of ways that all birthdays can be different. Notice that the sample space has size

$$|\Omega| = (365)^{23}\,.$$

(I.e. Notice that the number of birthday for the 1st player is 365, and for the 2nd player it is 365 and so forth), whereas the number of ways the players can have different birthdays is a permutation:

$$|\{\ \textit{all different}\,\}| = P^{365}_{23}\,.$$

(I.e. Notice that the number of birthday for the 1st player is 365, and for the 2nd player it is 364 – because player 1*'s birthday is taken– and so forth).*

[6] You can ignore Feb 29th on leap years.

Thus

$$\mathbb{P}(\textit{all different}) = \frac{P^{365}_{23}}{365^{23}} = 0.4927$$

and

$$\mathbb{P}(\textit{two or more same birthday}\ ) = 0.5073\,.$$

The result may be quite surprising as you would think it unlikely to share a birthday with any individual. However, there are many pairs of birthdays between the players. These balance out to give 50% chance of two players having the same birthday.

Example 8. *A parking inspector randomly inspects* 7 *different cars in a car park with* 35 *cars where* 5 *cars have not bought a ticket. If the inspector inspects a car without a ticket then the car receives a fine. Find the probability that exactly two cars are fined.*

Answer 8. *The number of ways of inspecting* 7 *cars out of* 35 *is* C^{35}_7*. I.e.*

$$|\Omega| = \binom{35}{7} = 6724520$$

We can suppose, without loss of generality, that the first 5 *cars have not bought a ticket. If* 2 *cars are fined, the inspector must choose to inspect* 2 *among these first* 5 *cars and must choose to inspect* 5 *the remaining* 30 *cars: E.g.*

$$\underbrace{\overset{I}{1},\overset{I}{2},3,4,5,}_{\textit{no ticket}}\underbrace{\overset{I}{6},7,8,\overset{I}{9},\overset{I}{10},11,12,\overset{I}{13},14,15,16,17,\ldots,31,32,\overset{I}{33},34,35}_{\textit{ticket}}\ .$$

Here "I" means inspected. The are C^5_2 *ways of choosing the two cars without a ticket, and* C^{30}_5 *ways of choosing the five cars with a ticket. Since any way of choosing the first two cars does not influence our choices for remaining five. For this reason we multiply the both combinations to give:*

$$|\{\ \textit{two cars fined}\ \}| = \binom{5}{2} \times \binom{30}{5} = 1425060\,.$$

Thus

$$\mathbb{P}(\ \textit{two cars fined}\ ) = \frac{\binom{5}{2} \times \binom{30}{5}}{\binom{35}{7}} = 0.212\,.$$

6 Conditional Probability

Conditional probabilities are probabilities when we have assumed that another event has occurred.

An example: Two aces. Suppose we have a deck of cards and we consider the probability of getting an ace. If we take one card out, then the probability of an ace is $4/52 = 1/13$. (There are 4 aces in a pack of 52 cards.) Now, given that the first card was an ace, we take a second card out the pack, notice now the probability of the second card being an ace has changed. Specifically there are now 3 aces and 51 cards in the pack. So the probability is $3/51 = 1/17$. (Also observe that this probability is different than if we had assumed that the first card was not an ace, which in that case the probability would be $4/51$.) Thus our condition on the first card has effected the probability for the second card. This is an example of conditional probability, which we now develop more generally.

Motivating a definition of conditional probability. Recall that in Section 3, we thought of probability as

$$\frac{\#\{\text{event } E \text{ occurs}\}}{\#\{\text{experiment}\}} \longrightarrow \mathbb{P}(E)\,. \tag{8}$$

Similarly, as we think of conditional probability as the proportion of time that an event occurs knowing that an other event has occurred. In this case, an analogous statement would be that

$$\frac{\#\{\text{event B and event A occur}\}}{\#\{\text{event A occurs}\}} \longrightarrow \mathbb{P}(B|A)\,. \tag{9}$$

where here we use $\mathbb{P}(B|A)$ to denote the probability of event B conditional on event A.

Observer, that we can express (9) in terms of (8). In particular,

$$\frac{\#\{\text{event B and A occur}\}}{\#\{\text{event A occurs}\}} = \frac{\dfrac{\#\{\text{event B and A occur}\}}{\#\{\text{experiment}\}}}{\dfrac{\#\{\text{event A occurs}\}}{\#\{\text{experiment}\}}} \longrightarrow \frac{\mathbb{P}(A\cap B)}{\mathbb{P}(A)}$$

Here we divide the denominator and numerator from the left hand side of (9) by the number of experiments and then we apply (8) to both expressions.

This motivates the following

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

Definition of conditional probability. Given the above we have:

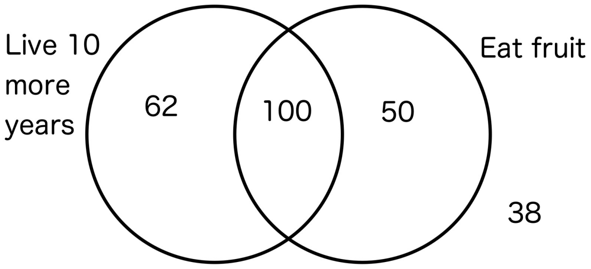
Definition 9 (Conditional Probability)**.** *For events A and B, the conditional probability of B given A is*

$$\mathbb{P}(B|A) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)}$$

when $\mathbb{P}(A) > 0$. By convention, if $\mathbb{P}(A) = 0$ then we define $\mathbb{P}(B|A) = 0$.

A couple of examples.

Example 9. *The results of a survey of 250 different 75 year-olds is given in the following Venn diagram.*



For a randomly selected participant in the survey:

1) Calculate the probability living 10 more years given that they eat fruit.

2) Calculate the probability living 10 more years given that they don't eat fruit.

Answer 9. *1)*

$$\mathbb{P}(\textit{ live 10 years} \mid \textit{eat fruit}\,) = \frac{100}{100+50} = 0.667$$

2)

$$\mathbb{P}(\textit{ live 10 years} \mid \textit{don't eat fruit}\,) = \frac{62}{62+38} = 0.62$$

So you should eat fruit...

Example 10. *I have two siblings, given I have a brother, what is the probability that I also have a sister.*

Answer 10. *Note that our sample space is $\Omega = \{BB, BS, SB, SS\}$. E.g. here BS means the eldest sibling is a brother and the youngest is a sister. Each outcome has equal probability of $1/4$. So*

$$\mathbb{P}(\text{ sister} \mid \text{brother }) = \frac{\mathbb{P}(BS) + \mathbb{P}(SB)}{\mathbb{P}(BB) + \mathbb{P}(BS) + \mathbb{P}(SB)} = \frac{2/4}{3/4} = \frac{2}{3}$$

For some this is an example of conditional probability being a bit counter-intuitive, as ones gut reaction is that the answer is a half. Notice if I specified that my eldest sibling was a brother then the answer would indeed be a half. This is just one small example of the subtle art of manipulating conditional probabilities.

Independence

We are interested in the setting where knowing that an event has occurred does not affect another event. E.g. if I roll a dice twice, in principle the outcome of the first roll should not influence the second roll. If event A does not influence event B then it should be that

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

I.e. conditioning on A having happened does not change the probability of B occuring. Since $\mathbb{P}(B|A) = \mathbb{P}(A \cap B)/\mathbb{P}(A)$, we can slightly more symmetrically express the above equality as

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

This is what we call independence of two events.

Definition 10 (Independence)**.** *We say that events A and B are independent if*

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

So if knowing that A has happened does not affect the probability of B then we multiply the probabilities together.

Warning! The following is sometimes confused by students. We say that two events are "mutually exclusive" if $A \cap B = \emptyset$. In that case we know from Lemma 2 that we *add* the probabilities together. This not the same as independence where we *multiply* the probabilities together.

As discussed, independence says that knowing an outcome is in A does not effect the probability of B. However, if events A and B are mutually exclusive, then knowing the outcome is in A effects the probability of B. Specifically, if we know an outcome is in A then it definitely is not in B.

Example 11. *What is the probability of getting* 10 *heads in a row from an unbiased coin?*

Answer 11. *Since with multiple probabilities together*

$$\begin{aligned}\mathbb{P}(HHHHHHHHHH) &= \mathbb{P}(H)\mathbb{P}(H)\mathbb{P}(H)\mathbb{P}(H)\mathbb{P}(H)\mathbb{P}(H)\mathbb{P}(H)\mathbb{P}(H)\mathbb{P}(H)\mathbb{P}(H)\\ &= \left(\frac{1}{2}\right)^{10} = \frac{1}{1024}.\end{aligned}$$

This is actually a "magic trick". Notice there are 1440 *minutes in a day. So enough time to have a reasonable chance of getting* 10 *heads in a row over the course of a day. There are TV magicians that had performed this as a trick (by cutting out the roughly* 1023 *previous camera takes).*

Rules for Conditional Probabilities

Here are few useful formulas for Conditional Probabilities. (Like with operations on sets the proofs are not entirely necessary to know for exams.)

Lemma 6.

$$\mathbb{P}(A \cap B) = \mathbb{P}(B|A)\mathbb{P}(A)$$

Proof. Follows immediately from the definition of $\mathbb{P}(B|A)$ □

This is useful as it can be easier to find $\mathbb{P}(B|A)$. E.g. like with the earlier two aces example, we can easily find the probability of drawing a ace from a deck given the previous card was an ace, and from that calculate the probability of two aces.

Lemma 7.

$$\mathbb{P}(B) = \mathbb{P}(B|A)\mathbb{P}(A) + \mathbb{P}(B|A^c)\mathbb{P}(A^c)$$

Proof. Since $B = (A \cap B) \cup (A^c \cap B)$, then Lemma 6 gives

$$\mathbb{P}(B) = \mathbb{P}(B \cap A) + \mathbb{P}(B \cap A^c)$$

and then applying Lemma 2 gives

$$\mathbb{P}(B) = \mathbb{P}(B|A)\mathbb{P}(A) + \mathbb{P}(B|A^c)\mathbb{P}(A^c)$$

as required. □

Note that the above result can be applied to any number of sets $A_1, \ldots, A_n$. So long as $\cup_i A_i = \Omega$ and $A_i \cap A_j = \emptyset$ for $i \neq j$, it holds that

$$\mathbb{P}(B) = \sum_{i=1}^{n} \mathbb{P}(B|A_i)\mathbb{P}(A_i)\,.$$

Lemma 8 (Bayes' Rule)**.**

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

The result is sometimes called Bayes' Theorem, as well.

Proof.

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

□

Bayes' Rule reverses the order of the conditional probability. There is a whole branch of statistics developed to this which we will very briefly touch upon shortly.

Example 12 (Two aces)**.** *Taking out two cards from a well-shuffled deck. What is the probability that both cards are aces? What is the probability that the 2nd card is an ace?*

Answer 12. *We know that* $\mathbb{P}(A_1) = 4/52$. *Also we know that* $\mathbb{P}(A_2|A_1) = 3/51$, *because after one ace is dealt then there are* 3 *aces and* 51 *cards. Thus using Lemma 6*

$$\mathbb{P}(A_1 \cap A_2) = \mathbb{P}(A_1)\mathbb{P}(A_2|A_1) = \frac{4}{52}\cdot\frac{3}{51} = \frac{1}{221} = 0.00452\,.$$

For the 2nd part, we can apply Lemma 7. Here

$$\begin{aligned}\mathbb{P}(A_2) &= \mathbb{P}(A_2|A_1)\mathbb{P}(A_1) + \mathbb{P}(A_2|A_1^c)\mathbb{P}(A_2^c)\\ &= \frac{4}{52}\cdot\frac{3}{51} + \frac{48}{52}\cdot\frac{4}{51} = \frac{1}{13}.\end{aligned}$$

This should come as no surprise, as the probability that the 2nd card is an ace should be the same as the probability that the 1st card is an ace. (Imagine taking the first card out the pack and putting it to the back, and then taking the 2nd card out and looking at it.)

Example 13 (Frequentist vs Bayesian Statistics)**.** *I have a biased coin, it is biased so the probability of heads,* θ, *is either* $3/4$ *or* $1/4$, *but you don't know which. So you throw the coin three times and get three heads. From this data determine if* $\theta = 3/4$ *or* $\theta = 1/4$.

Answer 13. *This is clearly not a well-defined question, because we cannot determine with certainty the value of* θ. *Given it's subjective nature (where there is no right answers). We give two approaches: a* *frequentist approach, which is a more classical statistical approach, and a bayesian statistical approach.*

The Frequentist Answer. *The likelihood of three heads for both choices of θ is*

$$\mathbb{P}(HHH|\theta = 1/4) = \left(\frac{1}{4}\right)^3 \quad \textit{and} \quad \mathbb{P}(HHH|\theta = 3/4) = \left(\frac{3}{4}\right)^3 .$$

The parameter that gives the highest probability is $\hat{\theta} = 3/4$. So our answer for this problem is the estimator $\hat{\theta} = 3/4$.

The Bayesian Answer. *Since we don't know in prior to throwing the coin, which of the two possibilities hold. We could give each possibility equal likelihood that is*

$$\mathbb{P}(\theta = 1/4) = \mathbb{P}(\theta = 3/4) = \frac{1}{2}$$

This is called the prior distribution. After throwing the coin and getting three heads then we want to update our estimate to find

$$\mathbb{P}(\theta = 1/4|HHH) \quad \textit{and} \quad \mathbb{P}(\theta = 3/4|HHH)$$

This called the posterior distribution. We can use Bayes' Rule to calculate the posterior:

$$\begin{aligned} \mathbb{P}(\theta = 3/4|HHH) &= \mathbb{P}(HHH|\theta = 3/4)\frac{\mathbb{P}(\theta = 3/4)}{\mathbb{P}(HHH)} \\ &= \left(\frac{3}{4}\right)^3 \frac{\frac{1}{2}}{\mathbb{P}(HHH)} \end{aligned} \tag{10}$$

To calculate $\mathbb{P}(HHH)$ *we can apply Lemma 7*

$$\begin{aligned} \mathbb{P}(HHH) =& \mathbb{P}(HHH|\theta = 3/4)\mathbb{P}(\theta = 3/4) + \mathbb{P}(HHH|\theta = 1/4)\mathbb{P}(\theta = 1/4) \\ =& \left(\frac{3}{4}\right)^3 \frac{1}{2} + \left(\frac{1}{4}\right)^3 \frac{1}{2} = \frac{28}{128} \end{aligned}$$

which then gives (10) *that*

$$\mathbb{P}(\theta = 3/4|HHH) = \frac{27}{28}$$

and thus $\mathbb{P}(\theta = 1/4|HHH) = \frac{1}{28}$. Thus in the Bayesian approach we says that we think $\theta = 3/4$ with probability $27/28$.

Under reasonable assumptions and enough data both the Bayesian and Frequentist approaches will converge on the correct parameter. The choice of the prior in the Bayesian approach is quite subjective. When the range of parameters gets large (or continuous) then we need to solve an optimization problem in the frequentist approach, while in the Bayesian approach we need to sum over a large number of terms to find the normalizing constant, which was the $\mathbb{P}(HHH)$ term in the above example.

7 Random Variables

Often we are interested in the magnitude of an outcome as well as its probability. E.g. in a gambling game amount you win or loss is as important as the probability each outcome.

Definition 11 (Random Variable)**.** *A random variable is a function $X : \Omega \to \mathbb{R}$ that gives a value for every outcome in the sample space.*

For example, if we roll two dice then $\Omega = \{(d_1, d_2) : d_1, d_2 \in \{1,2,3,4,5,6\}\}$ is our sample space and the sum of the two dice can be given by $X((d_1, d_2)) = d_1 + d_2$.

Random variables (RVs) can be discrete, e.g. taking values $1, 2, 3, \ldots$, or they can be continuous, e.g. taking any value in $\mathbb{R}$.

Typically we use capital letters to denote random variables X, Y, Z. Often we suppress the underlying sample space when writing probabilities. E.g. for our sum of two dice examples, we might write

$$\mathbb{P}(X = 2) = \frac{1}{36},$$

where here by $\{X = 2\}$ we really mean the event $\{\omega \in \Omega : X(\omega) = 2\}$.

Discrete Probability Distributions. A key way to characterize a random variable is with its distribution.

Definition 12 (Probability Mass Function)**.** *The probability distribution of a discrete random variable X with values in the finite or countable set $\mathcal{X} \subseteq \mathbb{R}$ is given by*

$$p(x) = \mathbb{P}(X = x), \qquad x \in \mathcal{X}.$$

This is some times called the probability mass function (PMF). Notice it satisfies the properties

- *(Positive) For all $x \in \mathcal{X}$,*
$$p(x) \geq 0$$
- *(Sums to one)*
$$\sum_{x \in \mathcal{X}} p(x) = 1.$$

Another way to characterize the distribution of the a random variable is through its cumulative distribution function, which is simply gives the probability that the random variable is below a given value.

Definition 13 (Cumulative Distribution Functiion). *The cumulative distribution function (CDF) of a random variable X is*

$$F_X(x) := \mathbb{P}(X \leq x)\,,$$

for $x \in \mathbb{R}$.

Add a picture

We can define more than one random variable on the same probability space. E.g. from our two dice throws, X could be the value of first dice, Y could be the value of the second dice, and Z could be the sum of the two dice.

Definition 14 (Joint Probability Distribution). *Suppose there are two random variables defined on the same probability space, $X : \Omega \to \mathcal{X} \subseteq \mathbb{R}$ and $Y : \Omega \to \mathcal{Y} \subseteq \mathbb{R}$, then the joint probability mass function is given by*

$$p(x, y) := \mathbb{P}(X = x, Y = y),$$

for $x \in \mathcal{X}$ and $y \in \mathcal{Y}$.

Definition 15 (Independent Random Variables). *We say that X and Y are independent random variables if*

$$\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x) \cdot \mathbb{P}(Y = y)$$

for all $x \in \mathcal{X}$, $y \in \mathcal{Y}$ (I.e. $p(x, y) = p_X(x)p_Y(y)$).

We can extend the about definition to any number of random variables. E.g. a set of random variable $X_1, X_2, X_3, ..., X_n$ are independent if

$$\mathbb{P}(X_1 = x_1, ..., X_n = x_n) = \prod_{i=1}^{n} \mathbb{P}(X_i = x_i)\,.$$

A common situtation that we are interested in is where $X_1, ..., X_n$ are independent identically distributed random variables or IIDRVs, for short. Here in addition to being independent, the random variables each have the same CDF. Here we think of IIDRVs repeating the same random experiment n times and recording the answer to each.

8 Expectation and Variance

Definition 16. *The expectation of a discrete random variable X is*

$$\mathbb{E}[X] = \sum_{x \in \mathcal{X}} x \cdot \mathbb{P}(X = x)\,.$$

The expectation gives an average value of the random variable. We could think of placing one unit of mass along the number line, where at point x we place a weight of $\mathbb{P}(X = x)$. The expectation, $\mathbb{E}[X]$, is then the point of the number line that balances the weights on the left with the right.

Example 14. *Calculate the expectation for the following random variable*

$$X = \begin{cases} -8 & \textit{with probability } \frac{1}{2} \\ 2 & \textit{with probability } \frac{1}{4} \\ 4 & \textit{with probability } \frac{1}{4} \end{cases}$$

Answer 14.

$$\mathbb{E}[X] = -8 \times \frac{1}{2} + 2 \times \frac{1}{4} + 4 \times \frac{1}{4} = -2\tfrac{1}{2}\,.$$

Properties of the expectation

Here are various properties of the expectation. Proofs are included and are good to know but are not essential reading for exams. (For the most part the following lemmas are really properties of summation.)

Lemma 9. *For a function $g : \mathcal{X} \to \mathbb{R}$,*

$$\mathbb{E}[g(X)] = \sum_{x \in \mathcal{X}} g(x)\mathbb{P}(X = x)\,.$$

Proof.

$$\begin{aligned}\mathbb{E}[g(X)] &= \sum_{y} y\mathbb{P}(g(X) = y)\\ &= \sum_{y} y \sum_{x:g(x)=y} \mathbb{P}(X = x)\\ &= \sum_{y} \sum_{x:g(x)=y} g(x)\mathbb{P}(X = x)\\ &= \sum_{x} g(x)\mathbb{P}(X = x)\,.\end{aligned}$$

□

Lemma 10. *For constants a and b*

$$\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$$

Proof. Applying Lemma 9,

$$\begin{aligned}\mathbb{E}[aX + b] =& \sum_{x\in\mathcal{X}} (ax + b)\mathbb{P}(X = x)\\ =& a\sum_{x\in\mathcal{X}} x\mathbb{P}(X = x) + b\sum_{x\in\mathcal{X}} \mathbb{P}(X = x)\\ =& a\mathbb{E}[X] + b\,.\end{aligned}$$

□

Lemma 11. *For two random variables X and Y*

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]\,.$$

Proof.

$$\begin{aligned}\mathbb{E}[X + Y] =& \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} (x + y)\mathbb{P}(X = x, Y = y)\\ =& \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} x\mathbb{P}(X = x, Y = y) + \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} y\mathbb{P}(X = x, Y = y)\\ =& \sum_{x\in\mathcal{X}} x\mathbb{P}(X = x) + \sum_{y\in\mathcal{Y}} y\mathbb{P}(Y = y)\\ =& \mathbb{E}[X] + \mathbb{E}[Y]\,.\end{aligned}$$

In the above we use the fact that, since $\{X = x\} = \cup_{y\in\mathcal{Y}}\{X = x, Y = y\}$, it holds that

$$\mathbb{P}(X = x) = \sum_{y\in\mathcal{Y}} \mathbb{P}(X = x, Y = y)\,.$$

□

Lemma 12. *For two independent random variables X and Y*

$$\mathbb{E}[X\cdot Y] = \mathbb{E}[X]\cdot\mathbb{E}[Y]\,.$$

Proof.

$$\begin{aligned}\mathbb{E}[X]\mathbb{E}[Y] =& \left(\sum_{x\in\mathcal{X}} x\mathbb{P}(X = x)\right)\left(\sum_{y\in\mathcal{Y}} y\mathbb{P}(Y = y)\right)\\ =& \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} xy\mathbb{P}(X = x)\mathbb{P}(Y = y)\\ =& \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} xy\mathbb{P}(X = x, Y = y) && \text{(by independence)}\\ =&\mathbb{E}[XY]\end{aligned}$$

□

Warning! In general, we cannot multiply expectations in this way. We need independence to hold. For instance if $X = Y = 1$ with probability 1/2 and $X = Y = 0$ otherwise a quick check shows that $1/2 = \mathbb{E}[XY] \neq \mathbb{E}[X]\mathbb{E}[Y] = 1/4$.

Example 15. *Let Y the number of heads from* 100 *coin tosses then* $\mathbb{E}[Y] = 50$.

Answer 15. *The exact distribution of Y is not so straight-forward to explicitly calculate. However, we can use the above rules.*

Note that $Y = \sum_{i=1}^{100} X_i$ where

$$X_i = \begin{cases} 1, & \text{if } i\text{th coin is heads,}\\ 0, & \text{otherwise.}\end{cases}$$

It is easy to see that $\mathbb{E}[X_i] = 1\mathbb{P}(X_i = 1) + 0\mathbb{P}(X_i = 0) = 1/2$. Thus by Lemma 11

$$\mathbb{E}[Y] = \sum_{i=1}^{100} \mathbb{E}[X_i] = \frac{100}{2} = 50\,.$$

Variance

While the expectation gives an average value for a random variable. The variance determines how spread out a probability distribution is.

Definition 17 (Variance / Standard Deviation)**.** *The variance of a random variable X is defined to be*

$$\mathbb{V}(X) := \mathbb{E}[(X - \mathbb{E}[X])^2]$$

Further the standard deviation is the square-root of the variance. That is

$$std(X) := \sqrt{\mathbb{V}(X)}$$

It is common to use μ to denote the expectation of a random variable and for σ to denote the variance of a random variable. So

$$\sigma^2 = \mathbb{E}[(X - \mu)^2] .$$

Here we use $(X-\mu)^2$ to determine a square distance about the mean. We then take the square root to give the rescale the distance. This is very similar to the way we think of (Euclidean) distances for vectors, E.g. for $\boldsymbol{x} = (x_1, x_2)$, $|\boldsymbol{x}| = \sqrt{x_1^2 + x_2^2}$.

Lemma 13.

$$\mathbb{V}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 .$$

So the variance is the mean square minus the square of the mean.

Proof.

$$\begin{aligned}\mathbb{V}(X) =&\mathbb{E}[(X - \mathbb{E}[X])^2]\\ =&\mathbb{E}[X^2 - 2X\mathbb{E}[X] + \mathbb{E}[X]^2]\\ =&\mathbb{E}[X^2] - 2\mathbb{E}[X]\mathbb{E}[X] + \mathbb{E}[X]^2\\ =&\mathbb{E}[X^2] - \mathbb{E}[X]^2 .\end{aligned}$$

□

Lemma 14. *For $a, b \in \mathbb{R}$,*

$$\mathbb{V}(aX + b) = a^2\mathbb{V}(X)\,.$$

Proof.

$$\begin{aligned}\mathbb{V}(aX + b) =& \mathbb{E}[(aX + b - \mathbb{E}[aX + b])^2\\ =& \mathbb{E}[(aX - a\mathbb{E}[X])^2]\\ =& a^2\mathbb{E}[(X - \mathbb{E}[X])^2]\\ =& a^2\mathbb{V}(X)\,.\end{aligned}$$

□

Lemma 15. *If X and Y are independent random variables then*

$$\mathbb{V}(X + Y) = \mathbb{V}(X) + \mathbb{V}(Y)\,.$$

Proof.

$$\begin{aligned}\mathbb{V}(X + Y) =& \mathbb{E}[(X + Y)^2] - \mathbb{E}[X + Y]^2\\ =& \mathbb{E}[X^2 + 2XY + Y^2] - (\mathbb{E}[X]^2 + 2\mathbb{E}[X]\mathbb{E}[Y] + \mathbb{E}[Y]^2)\\ =& \mathbb{E}[X^2] - \mathbb{E}[X]^2 + \mathbb{E}[Y^2] - \mathbb{E}[Y]^2 + \underbrace{2\mathbb{E}[XY] - 2\mathbb{E}[X]\mathbb{E}[Y]}_{\text{=0, by independence.}}\\ =& \mathbb{V}(X) + \mathbb{V}(Y)\,.\end{aligned}$$

□

Notice if we have a sequence of independent identically distributed random variables (IIDRVs) $X_1, ..., X_n$ where $\mathbb{V}(X_1) = \sigma^2$ then we can see from the above lemma that

$$\text{std}\Big(\sum_{i=1}^{n} X_i\Big) = \sqrt{\mathbb{V}\Big(\sum_{i=1}^{n} X_i\Big)} = \sqrt{\sum_{i=1}^{n} \mathbb{V}\big(X_i\big)} = \sigma\sqrt{n}\,.$$

So we see that as we add up more and more IIDRVs the distance from the mean is about $\sqrt{n}$. This is an important observation that will be refined later on where we discuss the normal distribution.

9 Discrete Probability Distributions

There are some probability distributions that occur frequently. This is because they either have a particularly natural or simple construction. Or they arise as the limit of some simpler distribution. Here we cover

- Bernoulli random variables
- Binomial distribution
- Geometric distribution
- Poisson distribution.

Of course there are many other important distributions. Like with counting, it is often easy when first learning about probability to think of different probability distributions as being a main destination for probability. However, it is perhaps better to think of the probability distributions that we cover now as simple building blocks that we can then be later used to construct more expressive probabilistic and statistical models.

Again we focus on probability distributions that take discrete values though shortly we will begin to discuss their continuous counterpart.

Notation. There are many different distributions which often have different parameters. For instance, shortly we will define the Binomial Distribution, which has two parameters n and p and we denote by $\text{Bin}(n, p)$. If a random variable X has a specific distribution then we use "$\sim$" to denote this. E.g. if X is a random variable with parameters $n = 4$ and $p = 0.2$ then we write

$$X \sim \text{Bin}(4, 0.2).$$

Bernoulli Distribution.

We start with the simplest discrete probability distribution.

Definition 18 (Bernoulli Distribution)**.** *A random variable that is either zero or one is a Bernoulli random variable. That is we we write $X \sim Bern(p)$ if*

$$\mathbb{P}(X = 1) = p \qquad \textit{and} \qquad \mathbb{P}(X = 0) = 1 - p.$$

It is a straight-forward calculation to show that for $X \sim \mathrm{Bern}(p)$

$$\mathbb{E}[X] = p, \qquad \text{and} \qquad \mathbb{V}(X) = p(1-p)\,.$$

Binomial Distribution. If we take $X_1, X_2, ..., X_n$ to be independent Bernoulli random variables with parameter p, and we add then together

$$X = \sum_{i=1}^{n} X_i\,,$$

then we get a Binomial distribution with parameters n and p.

So if we consider an experiment with probability of success p and we repeat an experiment n times and count up the number of successes, then the resulting probability distribution is a Binomial distribution.

Let's briefly consider the probability that $X = k$. One event where $\{X = k\}$ occurs is when the first k experiments end in success and the rest fail, $\{X_1 = 1, ..., X_k = 1, X_{k+1} = 0, ..., X_n = 0\}$. Note that by independence

$$\begin{aligned}
&\mathbb{P}(X_1 = 1, ..., X_k = 1, X_{k+1} = 0, ..., X_n = 0)\\
&= \mathbb{P}(X_1 = 1)...\mathbb{P}(X_k = 1)\mathbb{P}(X_{k+1} = 0)...\mathbb{P}(X_n = 0)\\
&= p^k(1-p)^{n-k}
\end{aligned}$$

Indeed the probability of any individual sequence $X_1, ..., X_n$ where $\sum_i X_i = k$. So how may such sequences are there? Well we have seen this before. It is the number of ways we can label n points with k ones (see additional remarks on combinations in Section 5), that is the combination C^n_k. Thus we have

$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}\,.$$

This motivates the following definition.

Definition 19 (Binomial Distribution). *A random variable X has a binomial distribution with parameters n and p, if it has probability mass function*

$$p(k) = \binom{n}{k} p^k (1-p)^{n-k}\,,$$

for $k = 0, 1, ...n$, and we write $X \sim Bin(n, p)$.

Here are some results on Binomial distributions that might be handy.

Lemma 16. *If $X \sim Bin(n,p)$, then*

$$\mathbb{E}[X] = np, \quad \textit{and} \quad \mathbb{V}(X) = np(1-p)\,.$$

Proof. Using the fact that $X = \sum_{i=1}^{n} X_i$ for X_i independent and $X_i \sim \text{Bern}(p)$. Then

$$\mathbb{E}[X] = \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{E}[X_i] = np$$

and

$$\mathbb{V}(X) = \mathbb{V}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathbb{V}(X_i) = np(1-p)\,.$$

□

Lemma 17. *If $X \sim Bin(n,p)$ and $Y \sim Bin(m,p)$ and are independent then*

$$X + Y \sim Bin(n+m,p)\,.$$

Proof. $X = \sum_{i=1}^{n} X_i$ for X_i independent and $X_i \sim \text{Bern}(p)$, and $X = \sum_{i=n+1}^{m} X_i$ for X_i independent and $X_i \sim \text{Bern}(p)$. So $X + Y = \sum_{i=1}^{n+m} X_i$ thus is Bernoulli with parameters $n + m$ and p. □

Lemma 18. *If $X \sim Bin(n,p)$ and $Y_1, \ldots, Y_n$ are independent Bernoulli random variables with parameter q then*

$$\sum_{i=1}^{X} Y_i \sim Bin(n,pq)\,.$$

Proof. Since we can write $X = \sum_{i=1}^{n} X_i$ for $X_i \sim \text{Bern}(p)$. Note that an equivalent way to represent the above random variable is

$$\sum_{i=1}^{n} X_i Y_i$$

Since $X_i Y_i \sim \text{Bern}(pq)$, then the above random variable must be $\text{Bin}(n,pq)$. □

Geometric Distribution

Suppose we throw a biased coin until the first time that it lands on heads. The distribution of the number of throws is a geometric distribution. For instance, the probability that it takes $X = 5$ coin throws is the same as the probability of 4 tails in a row and then one heads which is

$$\mathbb{P}(X = 5) = \mathbb{P}(TTTTH) = (1-p)^4 p$$

where p is the probability of heads. In general, the probability we need k throws is

$$\mathbb{P}(X = k) = (1-p)^k p\,.$$

This gives the geometric distribution.

Definition 20 (Geometric distribution)**.** *The geometric distribution with success probability p is the distribution with probability mass function*

$$p(k) = (1-p)^{k-1} p$$

for $k = 1, 2, \ldots,$ and we write $X \sim Geo(p)$.

The following lemma is useful for geometrics distributions but also various forms of compound interest and other applications.

Lemma 19 (Geometric Series)**.** *For $|x| < 1$,*

$$\sum_{n=0}^{\infty} x^n = \frac{1}{1-x}, \quad \sum_{n=0}^{\infty} n x^{n-1} = \frac{1}{(1-x)^2} \quad \textit{and} \quad \sum_{n=0}^{\infty} n(n-1) x^{n-2} = \frac{2}{(1-x)^3}$$

Proof.

$$\begin{aligned} \sum_{n=0}^{\infty} x^n &= 1 + x + x^2 + x^3 + \ldots \\ x \times \sum_{n=0}^{\infty} x^n &= \phantom{1+{}} x + x^2 + x^3 + \ldots \end{aligned}$$

Now subtracting gives

$$(1-x)\sum_{n=0}^{\infty} x^n = 1$$

Thus

$$\sum_{n=0}^{\infty} x^n = \frac{1}{(1-x)}\,.$$

Differentiating the above with respect to x gives

$$\sum_{n=0}^{\infty} nx^{n-1} = \sum_{n=0}^{\infty} \frac{d}{dx}x^n = \frac{d}{dx}\sum_{n=0}^{\infty} x^n = \frac{d}{dx}\frac{1}{(1-x)} = \frac{1}{(1-x)^2}\,.$$

Differentiating again gives

$$\sum_{n=0}^{\infty} n(n-1)x^{n-2} = \frac{2}{(1-x)^3}\,.$$

□

Lemma 20. *If $X \sim$ Geo(p) then*

$$\mathbb{P}(X > k) = (1-p)^k, \quad \mathbb{E}[X] = \frac{1}{p}, \quad \textit{and} \quad \mathbb{V}(X) = \frac{1-p}{p^2}\,.$$

Proof. The event $\{X > k\}$ is the event where we get k tails in a row, which has probability $(1-p)^k$. So

$$\mathbb{P}(X > k) = (1-p)^k\,.$$

Next,

$$\mathbb{E}[X] = \sum_{k=0}^{\infty} k(1-p)^{k-1}p = p\sum_{k=0}^{\infty} k(1-p)^{k-1} = \frac{p}{(1-(1-p))^2} = \frac{1}{p}\,.$$

Similarly

$$\begin{aligned}\mathbb{E}[X(X-1)] =& \sum_{k=0}^{\infty} k(k-1)(1-p)^{k-1}p \\ =&(1-p)p\sum_{k=0}^{\infty} k(k-1)(1-p)^{k-2} \\ =&\frac{2(1-p)p}{(1-(1-p))^3} = \frac{2(1-p)}{p^2}\end{aligned}$$

So

$$\begin{aligned}\mathbb{V}(X) =&\mathbb{E}[X^2] - \mathbb{E}[X]^2\\ =&\mathbb{E}[X(X-1)] + \mathbb{E}[X] - \mathbb{E}[X]^2\\ =&\frac{2(1-p)}{p^2} + \frac{1}{p} - \frac{1}{p^2} = \frac{1-p}{p^2}\,.\end{aligned}$$

□

If we throw a coin and get 8 tails in a row, and we ask how long should we wait until we next get a heads, then (even though it might feel like we are now due a heads) it is the same as the time we would have expected when we first started throwing the coin. This is key property of the geometric distribution and its called memoryless property.

Lemma 21 (Memoryless Property)**.** *If $T \sim Geo(p)$ then, conditional on $\{T > t\}$, the distribution of $T - t$ is geometrically distributed with parameter p. In otherwords $(T - t|T \geq t) \sim Geo(p)$.*

Proof.

$$\begin{aligned}\mathbb{P}(T-t > k|T > t) =&\frac{\mathbb{P}(T > t+k, T > t)}{\mathbb{P}(T > t)}\\ =&\frac{\mathbb{P}(T > t+k)}{\mathbb{P}(T > t)}\\ =&\frac{(1-p)^{t+k}}{(1-p)^t} = (1-p)^k\end{aligned}$$

□

Example 16 (Waiting for a bus)**.** *At a bus stop, the probability that a bus arrives at any given minute is p and is independent from one minute to the next.*

1. *What is the expect gap in the time between any two busses?*
2. *You arrive at the bus stop and there is no bus there. What is the expected gap between the last time a bus arrived and the next bus to arrive?*

Answer 16. *1. The time from one bus to the next is geometric p, so the expected wait is $1/p$.*

2. Given you at a time with no bus the time until the last bus too arrive is geometrically distributed with parameter p and so is the time until the next bus to arrive. The time between this bus arrivals is thus the sum of these geometeric distributions, and so the expected time is $2/p$.

This is sometimes called the waiting time paradox. Here we see that when we turn up at the bus station the gap between the buses is twice as long as the mean time between the buses. This is because when we turn up and there is no bus there then we are more likely to have chosen a time with a bigger gap between the buses.

Poisson Distribution.

The Poisson distribution arises when we count the number of successes of an unlikely event over a large population. This occurs in all manner of settings from nuclear decay, to insurance, to call over a telephone line.

We present a definition first and then we will motivate the Poisson distribution.

Definition 21 (Poisson distribution)**.** *For a parameter $\lambda > 0$, the Poisson distribution has probability mass function*

$$p(k) = \frac{\lambda^k}{k!} e^{\lambda}$$

for $k = 0, 1, 2, \ldots$ and we write $X \sim Po(\lambda)$

Motivation for Poisson Distribution. If we take a Binomial distribution where the number of trails n is large but the probability of success in each trail is small, specifically $p = \lambda / n$, then the Binomial distribution is well approximated by a Poisson distribution.

This is the reason the Poisson distribution is a reasonable distribution to represent pheonomena like nuclear decay. In nuclear decay, there are a large number of atoms in a radio-active substance, and, in any given time interval, there is a very small probability of one of these atom undergoing nuclear decay and the emitting a particle (e.g. a gamma-ray). For this reason the distribution of the number of observed gamma-rays over a time interval is well approximated by a Poisson distribution.

The following lemma sets out how the Poisson distribution approximates the Binomial distribution (again students primarily interested in assessment can skip with argument).

Theorem 1 (Binomial to Poisson Limit)**.** *Consider a sequence of Binomial random variables $X^{(n)} \sim Bin(n, \frac{\lambda}{n})$ for $n \in \mathbb{N}$, and let $X \sim Po(\lambda)$. Then*

$$\mathbb{P}(X^{(n)} = k) \xrightarrow[n \to \infty]{} \mathbb{P}(X = k)$$

That is as n gets large the probability of $X^{(n)} = k$ approaches the probability that $X = k$ for each k.

Proof.

$$
\begin{aligned}
\mathbb{P}(X^{(n)} = k) =& \binom{n}{k}\left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} \\
=& \frac{n}{n} \cdot \frac{n-1}{n} \cdot \ldots \cdot \frac{n-k+1}{n} \cdot \frac{\lambda^k}{k!} \cdot \left(1 - \frac{\lambda}{n}\right)^{-k} \cdot \left(1 - \frac{\lambda}{n}\right)^n \qquad (11)\\
\to& \frac{\lambda^k}{k!} e^{-\lambda} = \mathbb{P}(X = k)\,.
\end{aligned}
$$

Above we notice each term like $(n-1)/n = 1 - \frac{1}{n} \to 1$ as $n \to \infty$. So most of the terms above in (11) converge to 1. Also we note that

$$\left(1 - \frac{\lambda}{n}\right)^n \to e^{-\lambda}\,.$$

This is shown in Lemma 22 below. □

Lemma 22. *For any $x \in \mathbb{R}$*

$$\left(1 - \frac{x}{n}\right)^n \to e^{-x}, \qquad \textit{as} \qquad n \to \infty\,.$$

Proof.

$$
\begin{aligned}
\left(1 - \frac{x}{n}\right)^n =& \exp\left\{n\left[\log\left(1 - \frac{x}{n}\right) - \log(1)\right]\right\} \\
=& \exp\left\{-x\left[\frac{\log(1 - x/n) - \log(1)}{-x/n}\right]\right\}
\end{aligned}
$$

Notice that

$$\frac{\log(1 - x/n) - \log(1)}{-x/n}$$

is the slope of the line between $(1 - x/n)$ and 1 for the function $f(z) = \log z$. So as n get large this will converge to the gradient of $f(z) = \log z$ at $z = 1$. Note $f'(1) = 1$. So applying this to the above

$$
\begin{aligned}
\left(1 - \frac{x}{n}\right)^n =& \exp\left\{-x\left[\frac{\log(1 - x/n) - \log(1)}{-x/n}\right]\right\} \\
\xrightarrow[n\to\infty]{}& \exp\left\{-x \cdot \left.\frac{d\log z}{dz}\right|_{z=1}\right\} = e^{-x}\,.
\end{aligned}
$$

□

Now for some more standard facts about the Poisson distribution.

Lemma 23. *If $X \sim Po(\lambda)$ then*

$$\mathbb{E}[X] = \lambda, \qquad \textit{and} \qquad \mathbb{V}(X) = \lambda\,.$$

Proof.

$$\mathbb{E}[X] = \sum_{k=0}^{\infty} k \cdot \frac{\lambda^k}{k!} e^{-\lambda} = \sum_{k=0}^{\infty} \lambda \frac{\lambda^{k-1}}{(k-1)!} e^{-\lambda} = \lambda e^{\lambda} e^{-\lambda} = \lambda\,.$$

Above we recall the Taylor expansion of the exponential

$$e^{\lambda} = 1 + \lambda + \lambda^2/2! + \lambda^3/3! + \ldots$$

$$\mathbb{E}[X(X-1)] = \sum_{k=2}^{\infty} k(k-1)\frac{\lambda^k}{k!} e^{-\lambda} = \lambda^2 \sum_{k=2}^{\infty} \frac{\lambda^{k-2}}{(k-2)!} e^{-\lambda} = \lambda^2 e^{\lambda} e^{-\lambda} = \lambda^2\,.$$

Thus

$$\mathbb{V}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \mathbb{E}[X(X-1)] + \mathbb{E}[X] - \mathbb{E}[X]^2 = \lambda^2 + \lambda - \lambda^2 = \lambda\,.$$

□

Here are two lemmas on Poisson random variables (we omit their proofs for now).

Lemma 24 (Poisson Summation Property)**.** *If $X \sim Po(\lambda)$, $Y \sim Po(\mu)$, and X and Y are independent then*

$$X + Y \sim Po(\lambda + \mu).$$

Lemma 25 (Poisson Thinning Property)**.** *If $N \sim Po(\lambda)$ and, independent of N, we let $X_1, X_2, \ldots$ be independent Bernoulli random variables with parameter p then*

$$\sum_{n=1}^{N} X_i \sim Po(p\lambda)\,.$$

In Lemma 24, we can begin to see how we can think of a Poisson distribution as part of a process that evolves in time. For instance we might say that the number of calls on a set of telephone lines in each minute is Poisson distributed with mean 4, then the number of calls per hour is Poisson mean $4 \times 60 = 240$.

In Lemma 25, we can see that if we exclude points according to an independent random variable then the resulting random variable is still Poisson. This is useful for instance in insurance. Here the number of claims an insurance company receives in a given day might be Poisson with mean 20. The company might split the claims into big and small claims (say on average half the claims are big and half small). Since there is some fixed probability that each claim is, say, big then the resulting number of big claims is Poisson mean 10. This is useful for an insurance company as they can divide up, reinsure or resell some of their risk.

Both lemmas can be proved directly by summing things but is a bit of a messy calculation. Intuitively the above lemmas holds because an equivalent results, Lemma 17 and Lemma 18, hold for Binomial random variables. So the both properties persists when we take the limit to a Poisson random variable (like in Theorem 1). The cleanest proof (using moment generating functions) is beyond the scope of this course, so we omit the proof for now.

10 Continuous Probability Distributions

We consider distributions that have a continuous range of values. Discrete probability distributions where defined by a probability mass function. Analogously continuous probability distributions are defined by a probability density function.

Definition 22 (Probability Density Function)**.** *A probability density function (pdf) is a function $f : \mathbb{R} \to \mathbb{R}_+$ that has two properties*

- *(Positive) For $x \in \mathbb{R}$*
$$f(x) \geq 0.$$
- *(Integrates to one)*
$$\int_{-\infty}^{\infty} f(x)dx = 1\,.$$

From this we can define the following.

Definition 23 (Continuous Probability Distribution)**.** *A random variable X with values in $\mathbb{R}$ has a continuous probability distribution with pdf $f(x)$ if*

$$F(x) := \mathbb{P}(X \leq x) = \int_{-\infty}^{x} f(y)dy\,.$$

As before, $F(x)$ is called the cumulative distribution function (CDF). As before it satisfies

- $0 \leq F(x) \leq 1$
- *$F(x)$ is non-decreasing*

and also it satisfies

$$F'(x) = f(x). \tag{12}$$

A key observation is that when making the conceptual switch from (discrete) probability mass functions to (continuous) probability density distributions, we have replaced summations with integration. This is the main difference, and since most properties of sums apply to integrals[7] many properties follow over for continuous random variables.

A few observations. Notice that the Equation (12) is a consequence of the Fundamental Theorem of Calculus. Note while a pmf must be bounded above by 1, in principle, a pdf can be unbounded.[8] Notice we may want to restrict a continuous random variable to a range of values. For example, we may want to assume that our random variable is positive, in this case the pdf will satisfy $f(x) = 0$ for $x < 0$. Notice it does not make sense to think of a continuous random variable as taking any specific value since the integral of a point is zero. Instead we often think of the random variable belonging to some range of values. For instance, for $a < b$, we have

$$\mathbb{P}(a < X \leq b) = \mathbb{P}(X \leq b) - \mathbb{P}(X \leq a) = F(b) - F(a) = \int_a^b f(x)dx\,.$$

Joint distributions. We can consider the pdf for two random variables (or more). If X, Y are continuous random variables (defined on the same probability space) then their joint pdf is a function $f(x, y)$ such that

- For $x, y \geq 0$,
$$f(x, y) \geq 0\,.$$
- $$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x, y)dxdy = 1$$

and from this

$$\mathbb{P}(X \leq a, Y \leq b) = \int_{-\infty}^{a}\int_{-\infty}^{b} f(x, y)dxdy\,.^9$$

If X and Y are independent then the joint pdf is the product of the pdfs

$$f(x, y) = f_X(x)f_Y(y)\,.$$

All other the above extends out to more than two random variables $X_1, ..., X_n$ in the way you might naturally expect. E.g. the pdf is a function of the form $f(x_1, ..., x_n)$.

[7] After all integrals are just fancy sums.

[8] It is possible to let $f(x) = \infty$ for some values but we ignore such cases for now.

Expectations

Analogous to the expectation in discrete random variables we have the following definition.

Definition 24 (Expectation, continuous case)**.** *The expectation of a continuous random variable X is given by*

$$\mathbb{E}[X] := \int_{-\infty}^{\infty} x f(x) dx \,.$$

Similarly the variance is defined much as before

$$\mathbb{V}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \int_{-\infty}^{\infty} x^2 f(x) dx - \left(\int_{-\infty}^{\infty} x f(x) dx\right)^2 .$$

The following proposition is an amalgamation of the lemmas that we had for discrete random variables.

Proposition 1. *For continuous random variables X, Y,*
1) For an integrable function $g : \mathbb{R} \to \mathbb{R}$,

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f(x) dx \,.$$

2) For constants $a, b \in \mathbb{R}$,

$$\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$$

3) For two random variables,

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y] \,.$$

4) If X and Y are independent, then

$$\mathbb{E}[XY] = \mathbb{E}[X] \cdot \mathbb{E}[Y] \,.$$

The proof of the above result really follow by an almost identical proof to the earlier discrete results. Just replace the summations with integrals. For that reason we omit the proof of this proposition.

The Normal Distribution

The normal distribution arrises in many situations involving measurement. E.g. the distributions of heights, the relative change in a stock index, the measurement of physical phenomena (e.g. a comet passing the sun), the result from an election poll, the distribution of heat.

The normal distribution is, perhaps, the most important probability distribution. Why is this? Well roughly because it is the distributions that arises when you add up lots of small independent errors. This is more formally states as a result called the *central limit theorem*, which we will discuss shortly.

Definition 25 (Standard Normal Distribution)**.** *The standard normal distribution has probability density function*

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left\{ -\frac{x^2}{2} \right\},$$

for $-\infty < x < \infty$*. If a random variable* Z *is a standard normal random variable we write* $Z \sim \mathcal{N}(0,1)$*. The cumulative distribution function is*

$$\Phi(z) = \mathbb{P}(Z \leq x) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$$

It can be shown that a standard normal random variable has mean 0 and variance 1. By shifting and scaling we can acheive other values for the mean and variance.

Definition 26 (Normal Distribution)**.** *The normal distribution with mean* μ *and variance* σ^2 *has probability density function*

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\}$$

for $-\infty < x < \infty$*. If* X *is a normally distributed random variable with mean and variance* σ^2 *then we write* $X \sim \mathcal{N}(\mu, \sigma^2)$ *.*

An useful point is that

$$\text{if} \quad X \sim \mathcal{N}(\mu,\sigma^2) \quad \text{then} \quad Z := \frac{X-\mu}{\sigma} \sim \mathcal{N}(0,1)\,.$$

Thus we see that a normal random variable is simply a standard normal random variable that has been rescaled (by σ) and shifted (by μ).

11 The Law of Large Numbers and Central Limit Theorem

Let's explain why the normal distribution is so important. Suppose that I throw a coin 100 times and count the number of heads

$$S_{100} = \sum_{i=1}^{100} X_i, \qquad \text{where} \qquad X_i = \begin{cases} 1 & \text{if } i\text{th throw is heads,} \\ 0 & \text{otherwise.} \end{cases}$$

The proportion of heads should be close to its mean

$$\frac{S_{100}}{100} \approx \frac{1}{2} = \mathbb{E}[X]$$

and for $10,000$ it should be even closer. This can be shown mathematically (not just for coin throws but for quite general random variable)s

Theorem 2. *For independent random variables X_i, $i = 1, \ldots, n$, with mean μ and variance bounded above by σ, if we define*

$$S_n := \sum_{i=1}^{n} X_i$$

then for all $\epsilon > 0$

$$\mathbb{P}\Big(\mu - \epsilon \leq \frac{S_n}{n} \leq \mu + \epsilon\Big) \xrightarrow[n \to \infty]{} 1\,.$$

We will prove this result a little later. But, continuing the discussion, suppose $X_1, \ldots, X_n$ are independent identically distributed random variables with mean μ and variance σ^2. We see from the above result that S_n/n is getting close to μ. Nonetheless, in general, there is going to be some error. So let's define

$$\epsilon_n := \frac{S_n}{n} - \mu = \frac{S_n - n\mu}{n}\,.$$

So what does ϵ_n look like? We know that, in some sense, $\epsilon_n \to 0$ as $n \to \infty$ but how fast?

For this we can analyze the variance of the random variable ϵ_n:

$$\begin{aligned}\mathbb{V}(\epsilon_n) = \mathbb{V}\left(\frac{S_n - n\mu}{n}\right) =& \frac{1}{n^2}\mathbb{V}\left(S_n - n\mu\right)\\ =& \frac{1}{n^2}\mathbb{V}\left(\sum_{i=1}^{n}(X_i - \mu)\right)\\ =& \frac{1}{n^2}\sum_{i=1}^{n}\underbrace{\mathbb{V}(X_i - \mu)}_{=\sigma^2} = \frac{\sigma^2}{n}\end{aligned}$$

Thus the standard deviation of ϵ_n decreases as $\sigma/\sqrt{n}$. Given this we can define

$$Z_n = \frac{\sqrt{n}}{\sigma}\epsilon_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}.$$

Notice that $\mathbb{E}[Z_n] = 0$ and

$$\mathbb{V}(Z_n) = \frac{n}{\sigma}\mathbb{V}(\epsilon_n) = 1.$$

So Z_n has mean zero and its variance is fixed. I.e. the error as measured by Z_n is not vanishing, but is staying roughly constant. So it seems like there is sometime happening for this random variable Z_n, a question is what happens to Z_n. The answer is that Z_n converges to a normal distribution.

This is a famous and fundamental result in probability and statistics called the central limit theorem.

Theorem 3 (Central Limit Theorem). *For independent random variables X_i with mean μ and variance σ^2, for $S_n = \sum_{i=1}^{n} X_i$ and*

$$Z_n := \frac{S_n - n\mu}{\sigma\sqrt{n}}$$

then

$$\mathbb{P}(Z_n \leq z) \xrightarrow[n\to\infty]{} \mathbb{P}(Z \leq z)$$

where Z is a standard normal random variable.

Given the discussion above the Central Limit Theorem, roughly says that

$$S_n \approx \mu n + \sqrt{n}\sigma Z$$

where Z is a standard normal random variable. So whenever we measure errors about some expected value we should start to consider normal random variables.

A Mathematical Notation

Lists and Sets.

Sets. A set is a list of numbers, letters, objects,.. what ever you want really. We contain these within curly brackets { and } . Eg.

$$\{1,2,3,4,5,6,7,8,9,10\}$$
$$\{a,b,c,...,z\}.$$

- The order does not matter in a set. For instance,

  $$\{1,2,3,4\} = \{4,3,2,1\} = \{2,4,1,3\}.$$

- We ofter refer to the items inside as “elements".
- Sometimes we use dots “. . . " when it is clear what is happening next:

  $$\{a,b,c,...,z\}$$
  $$\{1,2,3,...\}$$

- We can use a colon “:" to specify conditions on a set. We can read this as “such that". Eg. numbers such that x is positive

  $$\{x : x > 0\}.$$

  or numbers such that they are between 1 and 10 and even

  $$\{x : 0 < x \leq 10,\ \ x \text{ even}\}$$

  The set of numbers greater than zero less than or equal to ten and even. Notice the comma is like an "and".

Set Notation. A couple of pieces of notation.

- $\in$ – means “in" or “belongs to". E.g. two belongs to the numbers from 1 to 10:

  $$2 \in \{1,2,3,4,5,6,7,8,9,10\}.$$

- $\subseteq$ – means “subset". E.g. the set of number 2,4,6,8,10 is a subset of the numbers from 1 to 10:

  $$\{2,4,6,8,10\} \subseteq \{1,2,3,4,5,6,7,8,9,10\}.$$

There are various other notations that I will introduce shortly.

Special sets. There are some commonly occuring sets with a special notation:

- $\mathbb{N}$ – the natural numbers, $\mathbb{N} = \{1, 2, 3, ...\}$.
- $\mathbb{Z}$ – the integers, $\mathbb{Z} = \{..., -2, -1, 0, 1, 2, ...\}$.
- $\mathbb{Q}$ – the rational numbers (aka. fractions), $\mathbb{Q} = \left\{\frac{a}{b} : a \in \mathbb{Z}, b \in \mathbb{N}\right\}$.
- $\mathbb{R}$ – the real numbers, e.g. $\pi \in \mathbb{R}$
- $[a, b], (a, b)$ – numbers between a and b, inclusive and exclusive.

Note that $\mathbb{N} \subseteq \mathbb{Z} \subseteq \mathbb{Q} \subseteq \mathbb{R}$.

Ordered lists. Sometimes we want to list elements where the order matters. We contain these with round brackets (and). E.g.

$$(x, y, z),$$
$$(\text{heads}, \text{tails}, \text{heads}, \text{tails})$$

(Note this is useful for co-ordinates for geometry but also when we can in what order a sequence of events occur in probability.)

- Here the order of elements in these lists does matter:

$$(1, 2, 3, 4) \neq (4, 3, 2, 1) \neq (2, 4, 1, 3).$$

- Again we often use ":" to list the items in the list or specify the conditions. E.g. Here we list the probabilities for each outcome from two coin throws.

$$(p_\omega : \omega \in \{HH, HT, TH, HH\})$$

Cardinality of a set. The cardinality of a set is the number of elements in that set. We use brackets | and | to denote the cardinality. Eg.

$$|\{a, b, c, d\}| = 4.$$

Products and Sums.

Sums. We use the symbol $\sum$ for sums over a specified range:

$$1+2+3+4+5+6+7+8+9=\sum_{n=1}^{9} n,$$

$$\sum_{\omega\in\Omega} p_\omega = 1,$$

$$1+\frac{1}{2^2}+\frac{1}{3^2}+\frac{1}{4^2}+\ldots=\sum_{n=1}^{\infty}\frac{1}{n^2}\,.$$

Notice sums do not need to be finite. Notice we sum over a range of values in a set. (This is useful in probability.)

Products. Normally at school "×" is used to mean multiplication. However, people also often use "·". I.e.

$$1\cdot 2\cdot 3\cdot 4 = 24$$

We use the symbol $\prod$ for products of a range over values. E.g.

$$2^n = \prod_{i=1}^{n} 2,$$

$$1\times 4\times 6\times 10 = \prod_{n\in\{\frac{1}{2},2,3,5\}} (2n)$$

Notice that here do products over sets. (This is useful in probability.)

Recall that for mathematical symbols we often omit the product symbol altogether. E.g. for $x=3$ and $y=5$,

$$xy = 15\,.$$

Cartesian Products. We can do products for sets. That is where we create a set consisting of the order pairs from two or more sets.

$$\{H,T\}\times\{H,T\}=\{(H,H),(H,T),(T,H),(T,T)\}$$

$$\prod_{i=1}^{3}\mathbb{R}=\{(x,y,z): x\in\mathbb{R}, y\in\mathbb{R}, z\in\mathbb{R}\}$$

Notice the cardinality a product set is the product of the sizes of the sets:

$$\left|\prod_{i=1}^{n}\Omega_i\right| = \prod_{i=1}^{n}|\Omega_i|$$

This is why it makes sense to think of it as a product.

Functions.

A function is something that takes an element from one set and gives you an element from another. E.g. $f(x) = x^2$ or $f(\theta) = e^{i\theta}$. We write $f : \mathcal{D} \to \mathcal{R}$ where $\mathcal{D}$ is the domain, the set of elements to which we apply the function, and $\mathcal{R}$ is the range, the set where the function takes its values. In probability we work with the function $\mathbb{P} : \Omega \to [0,1]$, i.e. for each outcome in our probability space we assign a probability which is a number between zero and one.

Equals.

We use the equals sign when two things are the same. I.e. $2 = 6/3$. We also use add a colon ":=" when we are defining something. I.e. the natural numbers are defined by

$$\mathbb{N} := \{1, 2, 3, 4, \ldots\}\,.$$

(We could use = here, but, by using :=, it makes it more explicit that we are defining a new piece of notation.)

Limits.

We will occasionally write statements like

$$\left(1 - \frac{1}{n}\right) \longrightarrow 0, \quad \text{as} \quad n \to 1\,.$$

There is a formal mathematical definition for this, which we do not get into. But it should be reasonably clear that what the above statement is staying is that as n gets very close to 1 then $1 - \frac{1}{n}$ gets closer and closer to 0.

Further the following statement holds:

$$\left(1 - \frac{x}{n}\right)^n \longrightarrow e^{-x}, \quad \text{as} \quad n \to \infty\,.$$

The statement says that as n gets larger and larger, the expression $(1 - x/n)^n$ gets closer and closer to e^{-x}.