Chapter 1: Probability Essentials

In this chapter we review the essential concepts of probability that will be needed as building blocks for the rest of the course.

1.1 Sample Space, Events, Probabilities, and Random Variables

First of all, **everything about probability starts with a sample space.** Probabilities have no meaning without reference to a sample space, and the values of probabilities change according to which sample space they relate to. Understanding the role and importance of the sample space is one of the most important steps in mastering probability and statistical theory.

Definition: A **random experiment** is an experiment whose outcome is not known until it is observed.

- *A random experiment describes a situation with an unpredictable, or random, outcome.*

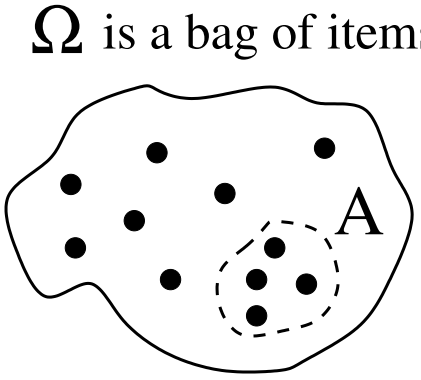
Definition: A **sample space**, Ω, is a set of outcomes of a random experiment. Every possible outcome is included in one, and only one, element of Ω.

- Ω *is a collection of all the things that could happen.*
- Ω *is a set. This means we can use the language of set theory, e.g.* $\cap$ *and* $\cup$*.*

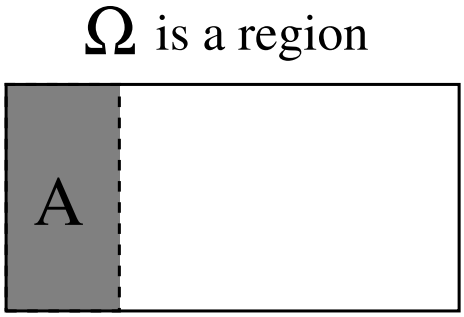
Definition: An **event**, A, is also a collection of outcomes. It is a subset of Ω.

- *An event* A *is*
- *An event* A *is a set of specific outcomes we are interested in.*
- *The formal definition of an event* A *is a subset of the sample space:* $A \subseteq \Omega$*.*
- *Just like* Ω*,* A *is also a set. This means we can use the language of set theory, e.g. for two events* A *and* B *we talk about* $A \cap B$*,* $A \cup B$*,* $\overline{A}$*, and so on.*
- *It makes no sense to talk about events unless we have first defined the random experiment and the sample space. This is not always as easy as it sounds!*

It is helpful to conceptualise sample spaces and events in pictures.



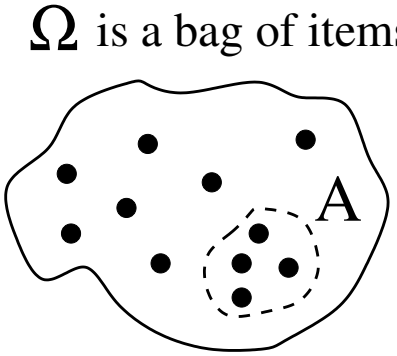
Event A is a smaller bag of items



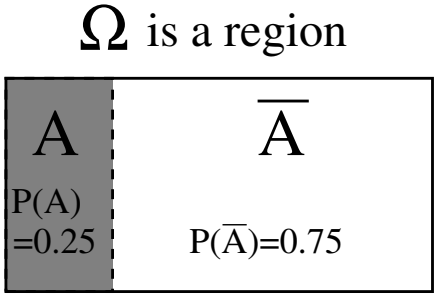
Event A is a subregion

Probability

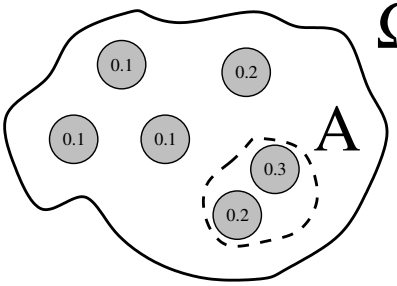
The idea of probability is to *attach a number to every item or event in* Ω *that reflects*



P(A)=4/11 if all items are equally likely



Probability is represented by AREA



P(A)=0.5 even though bag A contains only 2 out of 6 items

Question: What random experiments are we implicitly assuming here?

- Ω as a bag of items?
- Ω as a region?

Formal probability definition: the three axioms

As the pictures imply, the idea of probability is to allocate a number to every subset of Ω that reflects how likely we are to obtain an outcome in this subset. Imagine that all of Ω is given a cake: the idea of probability is to ***distribute*** a piece of cake to each item in Ω. Some items might get more cake than others, reflecting that the corresponding events are more likely to occur. This is why we talk about ***probability distributions:***

Definition: A **probability distribution** allocates an amount of probability to every possible subset of Ω.

This idea is formalized in the following three ***Axioms,*** which constitute the *definition* of a probability distribution. A rule for allocating probability to subsets of Ω is a valid probability distribution if and only if it satisfies the following three axioms or conditions.

Axiom 1: $\mathbb{P}(\Omega) = 1$.

- *This means that the total amount of 'cake' available is 1.*
- *It also makes it clear that*

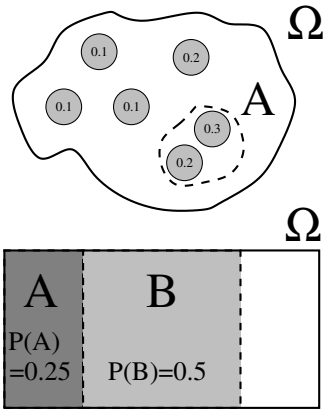
Axiom 2: $0 \leq \mathbb{P}(A) \leq 1$ for all events A.

- *This says that probability is always a number between 0 and 1.*

Axiom 3: If $A_1, A_2, \ldots, A_n$ are **mutually exclusive** events, (no overlap), then

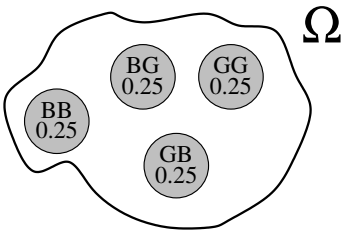
$$\mathbb{P}(A_1 \cup A_2 \cup \ldots \cup A_n) = \mathbb{P}(A_1) + \mathbb{P}(A_2) + \ldots + \mathbb{P}(A_n).$$

- *This says that if you have non-overlapping sets, the amount of cake they have in total is the sum of their individual amounts.*
- *This axiom is the reason why we can say that* $\mathbb{P}(A) = 0.3 + 0.2 = 0.5$ *in the bag diagram.*
- *In the region diagram,* $\mathbb{P}(A \cup B) =$



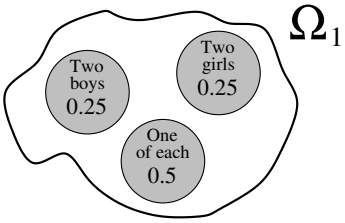
Examples of probability distributions

Suppose we are interested in the composition of a two-child family in terms of number of girls and boys. Assuming each child is equally likely to be a boy or a girl, there are

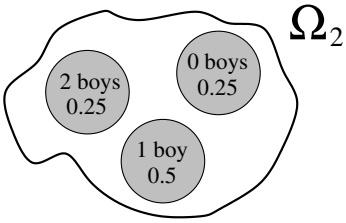


So if we pick a two-child family at random, we have probability $1/4 = 0.25$ of getting each of the outcomes BB, BG, GB, and GG.

Now suppose we don't care *what order* the children are in: we only care *how many of each sex* are in the family. We could choose to represent this by a second sample space, Ω_1, in which the outcomes are no longer equally likely:

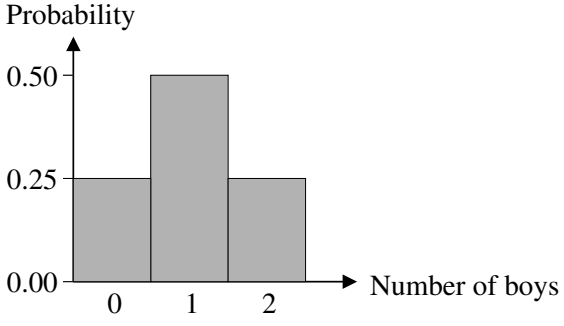


This is cumbersome to write down — especially if we consider listing the options for families of more than two children. Instead, we can be more efficient if we describe the outcomes by

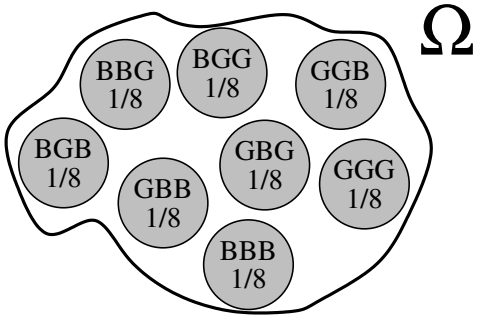


Now think of a different way of picturing Ω_2 that is easier to extend:

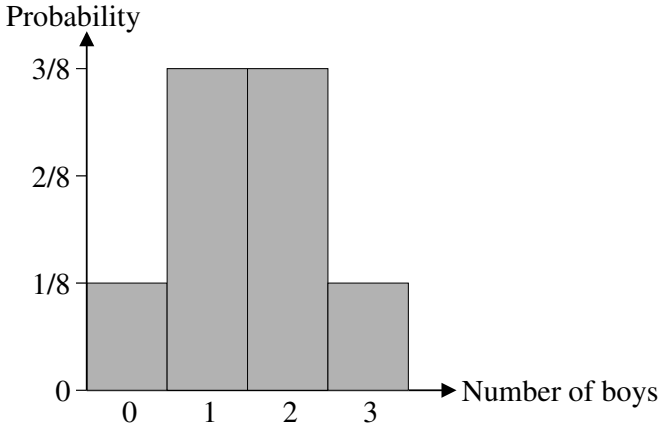
This representation is more like the 'region' image of Ω we used earlier, where probabilities are represented by
It also has the huge advantage of being a flexible, graphical display.
Question: where would you draw Ω_2?



If we move to three-child families, we quickly see the advantage of our graphical depiction of Ω:



(a) Representation of the probability distribution if child-order is of interest.



(b) Representation of the probability distribution if we count the number of boys.

We can see that the numerical expression of outcomes (0, 1, 2, or 3 boys) is much more succinct than describing all combinations, BBB, BBG, BGB, ..., GGG, as long as we do not care about the order that children occur in the family. However,

Random variables

The idea above of converting an outcome described in words (e.g. BBG) into a numeric summary of the outcome (e.g. 2 boys) is the definition of a ***random variable.*** Instead of writing $\mathbb{P}(2 \text{ boys})$ above, we give the unknown numerical outcome a capital letter, say X.

X is called a ***variable*** because it is a and it is called ***random*** because we don't know what value it will take until we make an observation of our random experiment. For example, if I pick a three-child family at random, I might observe $X = 0$, $X = 1$, $X = 2$, or $X = 3$ boys.

In essence,

In formal language, a random variable is a mapping from Ω to the real numbers: $X : \Omega \to \mathbb{R}$. For example, for outcome BBG (a member of Ω), the number of boys is 2, so we can write $X(BBG) = 2$. However, we usually use a more succinct notation and just say that our outcome is $X = 2$.

Everything you need to know about random variables

- Random variables always have

  Understand the capital letter to mean that X denotes a quantity that will take on values at random.

- The term 'random variable' is often abbreviated to

- You can think of a random variable simply as

  In the example above, X generates random numbers 0, 1, 2, or 3 by picking a 3-child family at random and counting how many boys are in it.

- Each possible ***value*** of a random variable has a probability associated with it. In the example above, where X is the number of boys in a three-child family:

- If we want to refer to a generic, unspecified, value of a random variable, we use a

  For example, we might be interested in finding a formula for $\mathbb{P}(X = x)$ for all values $x = 0, 1, 2, 3$.

Differences between X, x, $\{X = x\}$, and $\mathbb{P}(X = x)$

It is very important to understand this standard notation and how it is used.

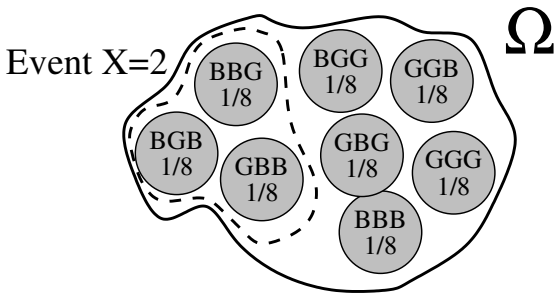
- X (capital letter) is a ***random variable:*** a mechanism for generating random real numbers. It is mainly used as a

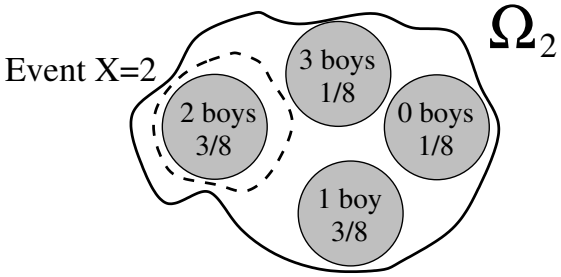
- x (lower-case letter) is a ***real number*** like 2 or 3. It is used to indicate an unspecified value that X might take.

- $\{X = x\}$, often written just as $X = x$, is an ***event:*** it is a

  For example, $X = 2$ is the event that we count 2 boys when we pick a three-child family at random.

Because $X = 2$ is an event, it is a





Crucially,

- $\mathbb{P}(X = x)$ is a ***real number:*** it is a number between 0 and 1.

When talking about ***events,*** like $\{X = x\}$, use set notation like $\cap$ and $\cup$.

When talking about ***probabilities,*** like $\mathbb{P}(X = x)$, use ordinary addition and multiplication $+$ and $\times$, just as you would for any other real numbers.

Right	**Wrong**
$X = 2 \cup X = 3$	$X = 2 + X = 3$
$\mathbb{P}(X = 2 \cup X = 3)$	$\mathbb{P}(X = 2) \cup \mathbb{P}(X = 3)$
$\mathbb{P}(X = 2) + \mathbb{P}(X = 3)$ *Probability of the event that X is 2 OR 3*	$\mathbb{P}(X = 2 + X = 3)$
$X \leq 2 \cap X > 1$	$X \leq 2 \times X > 1$
$\mathbb{P}(X \leq 2 \cap X > 1)$ *Probability that X is BOTH ≤ 2 AND > 1*	$\mathbb{P}(X \leq 2) \cap \mathbb{P}(X > 1)$

1.2 Bernoulli trials and the Binomial Distribution

The Binomial distribution is one of the simplest probability distributions. We shall use it extensively throughout the course for illustrating statistical concepts.

The Binomial distribution

Such trials are called ***Bernoulli trials,*** named after the 17th century Swiss mathematician Jacques Bernoulli.

Definition: A sequence of **Bernoulli trials** is a sequence of independent trials where each trial has two possible outcomes, denoted Success and Failure, and the probability of Success stays constant at p for all trials.

Examples: (1) Repeated tossing of a fair coin:

(2) Repeated rolls of a fair die:

Jacques Bernoulli, and his brother Jean, were bitter rivals. They both studied mathematics secretly, against their father's will. Their father wanted Jacques to be a clergyman and Jean to be a merchant.

Note: Saying the trials are ***independent*** means that

Thus, whether the current trial yields a Success or a Failure is not influenced by the outcomes of any previous trials. For example, you are ***not*** more likely to have a win after a run of losses: the previous outcomes simply have no influence.

Definition: The random variable Y is called a **Bernoulli random variable** if

Definition: For any random variable Y, we define the **probability function** of Y to be the function $f_Y(y) = \mathbb{P}(Y = y)$.

The probability function of the Bernoulli random variable is:

We often write the probability function in table format:

Binomial distribution

The Binomial distribution describes the outcome from a fixed number, n, of Bernoulli trials. For example:

- X is the number of boys in a 3-child family:

- X is the number of 6's obtained in 10 rolls of a die:

Definition: Let X be the number of successes obtained in n independent Bernoulli trials, each of which has probability of success p.

Then X has the **Binomial distribution with parameters n and p.**

We write

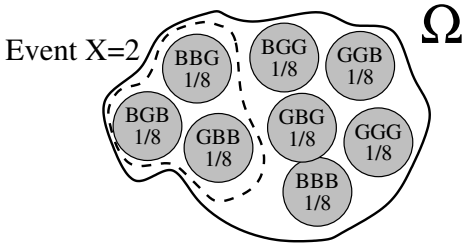
> The Binomial distribution counts the number of **successes** in a **fixed number** of Bernoulli trials.

If $X \sim \text{Binomial}(n, p)$, then $X = x$ if there are x successes in the n trials. We don't care what order the successes occur in — in other words, we don't care *which* of the trials are successes and which are failures. However, we do have to bear in mind all the different orderings when we calculate the probabilities of the distribution.

Take the example of $X =$ number of boys in a 3-child family, so

If we want to calculate $\mathbb{P}(X = 2)$, we have to take account of all the different ways that we can achieve $X = 2$:

In this case, there are 3 ways of getting the outcome we are interested in: 2 boys and 1 girl. How would we calculate the number of ways in general?



Question: How many ways are there of achieving 6 boys in a 10-child family?
Answer:

Question: How many ways are there of achieving x successes in n trials?
Answer:

Question: If each trial has probability p of being a success, what is the probability of getting the precise outcome $SFFSF$ from $n = 5$ trials?
Answer:

Question: What is the probability of **one** ordering that contains x successes and $n - x$ failures?
Answer:

Question: So what is the overall probability of achieving x successes in n trials: $\mathbb{P}(X = x)$ when $X \sim \text{Binomial}(n, p)$?
Answer:

This gives the **probability function for the Binomial distribution:**

Let $X \sim \text{Binomial}(n, p)$. The probability function for X is:

$$f_X(x) = \mathbb{P}(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \quad \text{for} \quad x = 0, 1, \ldots, n\,.$$

Note: 1. Importantly,
The correct way to write the range of values is
— Writing $x \in [0, n]$ is **wrong,** because this includes decimals like 0.4.
— Writing $x = 0, 1, \ldots$ is **wrong,** because the range of values must stop at n: you can't have more than n successes in n trials.

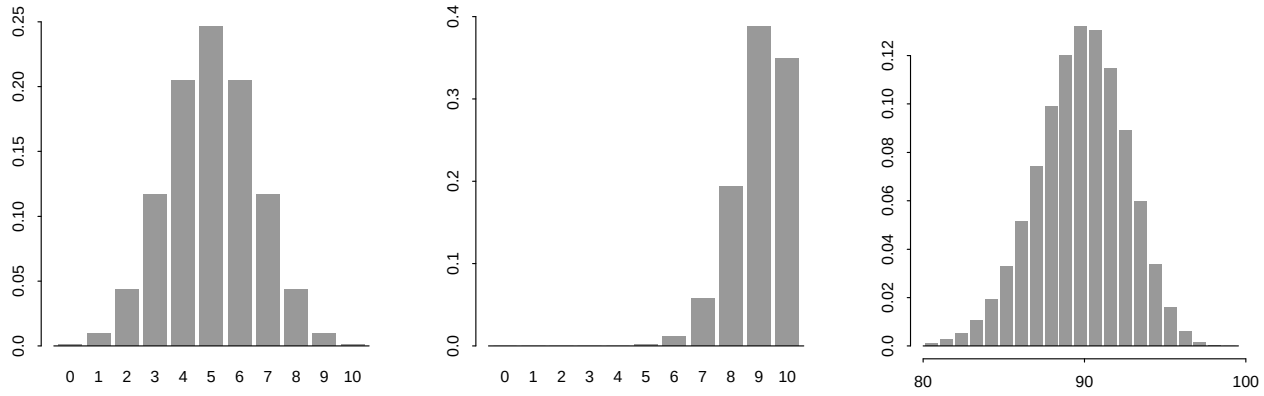
2. $f_X(x)$ means, *'the probability function belonging to the r.v. I've named X'.*
Use a capital X in the subscript and a lower-case x as the argument.

Shape of the Binomial distribution

The shape of the Binomial distribution depends upon the values of n and p. For small n, the distribution is almost symmetrical for values of p close to 0.5, but highly skewed for values of p close to 0 or 1. As n increases, the distribution becomes more and more symmetrical, and there is noticeable skew only if p is very close to 0 or 1.

The probability functions for various values of n and p are shown below.

$n = 10$, $p = 0.5$ $n = 10$, $p = 0.9$ $n = 100$, $p = 0.9$



Sum of independent Binomial random variables:

If X and Y are , and $X \sim \text{Binomial}(n, p)$, $Y \sim \text{Binomial}(m, p)$, then

This is because X counts the number of successes out of n trials, and Y counts the number of successes out of m trials: so overall, $X + Y$ counts the total number of successes out of

Note: X and Y must both share

Binomial random variable as a sum of Bernoulli random variables

It is often useful to express a Binomial(n, p) random variable as the sum of n Bernoulli(p) random variables. If $Y_i \sim \text{Bernoulli}(p)$ for $i = 1, 2, \ldots, n$, and if $Y_1, Y_2, \ldots, Y_n$ are independent, then:

$$X = Y_1 + Y_2 + \ldots + Y_n \sim \text{Binomial}(n, p).$$

This is because X and $Y_1 + \ldots + Y_n$ both represent

Cumulative distribution function, $F_X(x)$

We have defined the *probability function*, $f_X(x)$, as $f_X(x) = \mathbb{P}(X = x)$.

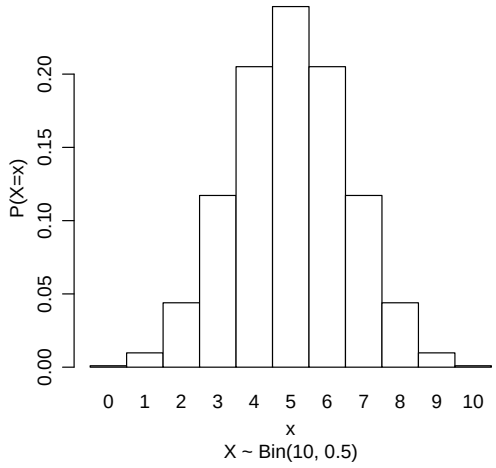
Another function that is widely used is the *cumulative distribution function*, or CDF, written as $F_X(x)$.

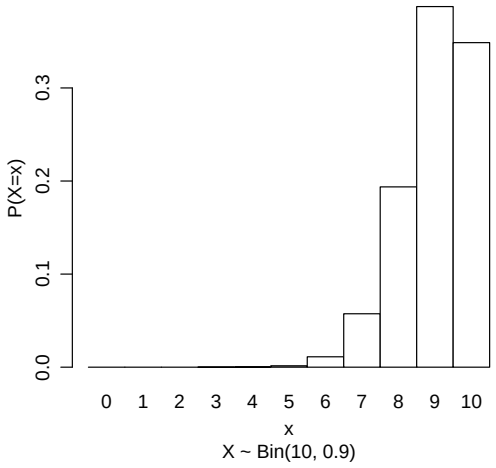
Definition: *The cumulative distribution function, or CDF, is*

$$F_X(x) = \mathbb{P}(X \leq x) \text{ for } -\infty < x < \infty$$

The cumulative distribution function $F_X(x)$ as a probability sweeper

The cumulative distribution function, $F_X(x)$,





Clearly,
$$F_X(x) = \sum_{y \leq x} f_X(y)\,.$$

Using the cumulative distribution function to find probabilities

$$\mathbb{P}(a < X \leq b) = F_X(b) - F_X(a) \quad \text{if } b > a.$$

Proof that $\mathbb{P}(a < X \leq b) = F_X(b) - F_X(a)$:

$$\begin{aligned} & \mathbb{P}(X \leq b) &=& \mathbb{P}(X \leq a) + \mathbb{P}(a < X \leq b) \\ \textit{So} \qquad & F_X(b) &=& F_X(a) + \mathbb{P}(a < X \leq b) \\ \Rightarrow \qquad & F_X(b) - F_X(a) &=& \mathbb{P}(a < X \leq b). \end{aligned}$$

Warning: endpoints

Be careful of endpoints and the difference between $\leq$ and $<$.
For example,

$$\mathbb{P}(X < 10) = \mathbb{P}(X \leq 9) = F_X(9).$$

Examples: Let $X \sim \text{Binomial}(100, 0.4)$. In terms of $F_X(x)$, what is:

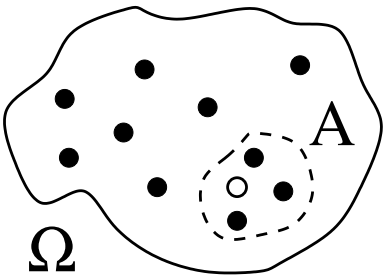
1. $\mathbb{P}(X \leq 30)$?
2. $\mathbb{P}(X < 30)$?
3. $\mathbb{P}(X \geq 56)$?
4. $\mathbb{P}(X > 42)$?
5. $\mathbb{P}(50 \leq X \leq 60)$?

1.3 Conditional probability

We have mentioned that ***probability*** depends upon the sample space, Ω:

Conditional probability is about
In particular, conditional probability is about ***reducing the sample space to a smaller one.***

Look at Ω on the right. Pick a ball at random. All 11 balls are equally likely to be picked. What is the probability of selecting the white ball?



Now suppose we select a ball only from within the smaller bag A. Recall that A is a subset of Ω, so in probability language, A is an

What is the probability of selecting the white ball, if we pick only from the balls in bag A?

We use a shorthand notation to write this down:

We read this as, 'probability of the white ball ***given*** A', or 'probability of selecting the white ball from ***within*** A'.

$\mathbb{P}(\text{white ball} \mid A)$ is called a ***conditional probability,*** and we say we have

Note: The vertical bar in $\mathbb{P}(\text{white ball} \mid A)$ is vertical: $|$.
Do not write a conditional probability as $\mathbb{P}(W/A)$ *or* $\mathbb{P}(W \backslash A)$*: it is* $\mathbb{P}(W \mid A)$.

What we have done is to ***reduce the sample space*** from Ω, which was a bag containing 11 equally-likely items, to a smaller bag A which contains only 4 equally-likely items.

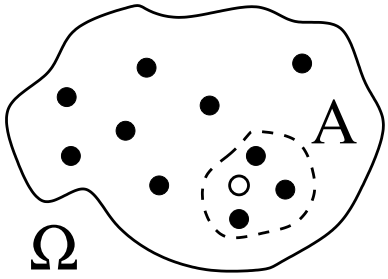
But A is still a bag of items — so A is a valid sample space in its own right. When we write $\mathbb{P}(W \mid A)$, we have ***changed the sample space*** from Ω to A.

Define event

We have said:

This means

where we recall that the symbol $\mathbb{P}$ is defined relative to Ω because $\mathbb{P}(\Omega) = 1$.



Now if we reduce to selecting only from the balls in A, we write:

Question: What is $\mathbb{P}(A \mid A)$?

Answer:

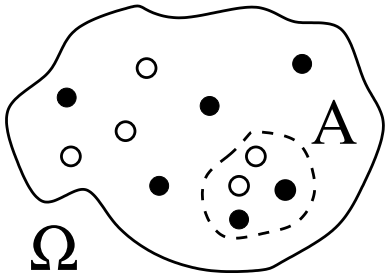
> The conditional probability $\mathbb{P}(W \mid A)$ means ***the probability of event W, when selecting only from within set A.***
>
> Read it as 'probability of event W, given event A',
> or 'probability of event W ***from within the set A***.'
>
> It is equivalent to ***changing the sample space from Ω to A.***
>
> The notation $\mathbb{P}(W \mid A)$ is like saying,
> '$\mathbb{P}(W)$ when my symbol $\mathbb{P}$ is defined relative to A instead of to Ω.'

Formula for conditional probability

Suppose we have several white balls in Ω, instead of just one. As before, we pick a ball at random and event W is the event that we select a white ball.

Question: What is $\mathbb{P}(W)$?

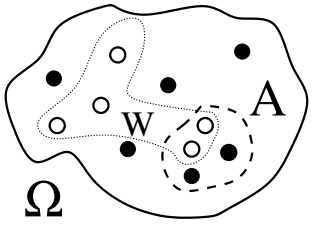
Answer: $\mathbb{P}(W)$ refers to the probability within the whole sample space Ω, so



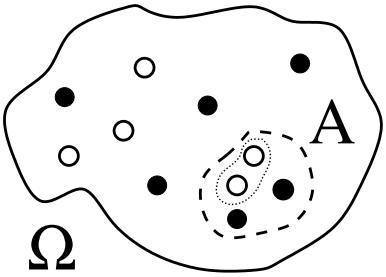
Question: What is $\mathbb{P}(W \mid A)$?

Answer: $\mathbb{P}(W \mid A)$ refers to the probability within bag A only, so

Question: Can you see why $\mathbb{P}(W \mid A) = \dfrac{\mathbb{P}(W \cap A)}{\mathbb{P}(A)}$?



It is obvious from the diagram that $\mathbb{P}(W \mid A) = \frac{2}{4}$.



The probability of W, when selecting from bag A only, is the probability contained in the small dotted bag as a fraction of the probability in the dashed bag.

The small dotted bag represents the set

The dashed bag represents the set

Thus, the probability of W when selecting from within A is:

This reasoning gives us our formal definition of conditional probability.

Definition: Let A and B be two events on a sample space Ω. The **conditional probability of event B, given event A**, is written $\mathbb{P}(B \mid A)$, and defined as

Read $\mathbb{P}(B \mid A)$ as *"probability of B, given A"*, or *"probability of B within A"*.

Note:

Follow this reasoning carefully. It is important to understand why conditional probability is the probability of the intersection within the new sample space.

Conditioning on event A means ***changing the sample space*** to A.

Think of $\mathbb{P}(B \mid A)$ as the chance of getting a B, from the set of A's only.

The notation $\mathbb{P}(B \mid A)$ is good because it emphasises that the ***denominator*** of the proportion is A. In a sense, $\mathbb{P}(B \mid A)$ is asking for event B as a ***fraction*** of event A.

Language of conditional probability

Conditional probability corresponds to ***changing the sample space.*** This means it affects *the set we are picking FROM, when we calculate the probability that its members satisfy a certain event.*

Suppose we are picking a person at random from this class (Ω). Event A is that the person has dark hair, and event B is that the person has blue eyes.

- $\mathbb{P}(B)$ means we want the probability of picking someone who satisfies B
- $\mathbb{P}(B \mid A)$ means the probability of picking someone who satisfies B
- $\mathbb{P}(B \cap A)$ means the probability of picking someone who satisfies

This means you can easily identify which probabilities are conditional and which are intersections by looking to see Recall:

$\Omega = \{\text{people in this class}\}$; $A = \{\text{dark-haired people}\}$; $B = \{\text{blue-eyed people}\}$

Define also a random variable X = number of GenEd papers a person has passed. At the University of Auckland, most students have to complete two GenEd (General Education) papers as part of their undergraduate degree. The GenEd papers can be completed at any time during the degree. Nearly everyone in this class will satisfy one of the events $X = 0$, $X = 1$, or $X = 2$.

Define further events: $F = \{\text{first years}\}$; $S = \{\text{second years}\}$; $T = \{\text{third years}\}$; $O = \{\text{other students, e.g. exchange students, COPs, } \ldots\}$.

Exercise: Translate the following statements into probability notation. Assume in all cases we are picking a person at random from this class.

- Probability a person has dark hair and blue eyes:
- Probability a dark-haired person has blue eyes:
- Probability a person has passed two GenEd papers:
- Probability a second-year has passed two GenEd papers:
- Probability a first-year has passed two GenEd papers:
- Probability a dark-haired first-year has passed one or two GenEd papers:

Trick for checking conditional probability calculations:

A useful trick for checking a conditional probability expression is to *replace the conditioned set by Ω, and see whether the expression is still true.*

The conditioned set is just another sample space, so probabilities $\mathbb{P}(\cdot \mid A)$ should behave exactly like ordinary probabilities $\mathbb{P}(\cdot)$, as long as ***all*** the probabilities are conditioned on the same event A.

Question: Is $\mathbb{P}(B \mid A) + \mathbb{P}(\overline{B} \mid A) = 1$?

Answer:

Question: Is $\mathbb{P}(B \mid A) + \mathbb{P}(B \mid \overline{A}) = 1$?

Answer: *Try to replace the conditioning set by Ω:*

The expression is *It doesn't make sense to try to add together probabilities from two different sample spaces.*

The Multiplication Rule

For any events A and B,

Proof: *Immediate from the definitions:*

1.4 Statistical independence

Events A and B are said to be ***independent*** if they

For example, in the previous section, would you expect the following pairs of events to be statistically independent?

$\Omega = \{\text{people in this class}\}$; $A = \{\text{dark-haired people}\}$; $B = \{\text{blue-eyed people}\}$ $F = \{\text{first years}\}$; $S = \{\text{second years}\}$; $T = \{\text{third years}\}$; $O = \{\text{other students}\}$; and random variable $X =$ number of GenEd papers passed.

- A and B?
- A and F?
- F and S?
- Events $X = 2$ and F?

To give a formal definition of statistical independence, we need a notion of what it means for two events to have ***no influence*** on each other:

- A has no influence on B if
- B has no influence on A if
- So A and B have ***no influence on each other if both*** $\mathbb{P}(B \mid A) = \mathbb{P}(B)$ ***and*** $\mathbb{P}(A \mid B) = \mathbb{P}(A)$.

However, it is untidy to have a definition with two statements to check. It would be better to have a definition with just one statement.

Using the multiplication rule:

- If $\mathbb{P}(B \mid A) = \mathbb{P}(B)$, then $\mathbb{P}(A \cap B) =$
- If $\mathbb{P}(A \mid B) = \mathbb{P}(A)$, then $\mathbb{P}(A \cap B) =$

So both statements imply that $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$. What about the other way around? Suppose that $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$. What does that imply about $\mathbb{P}(A \mid B)$ and $\mathbb{P}(B \mid A)$?

- If $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$, then

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

- Similarly, if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$, then

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

So the ***single*** statement $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$, implies ***both*** statements $\mathbb{P}(A \mid B) = \mathbb{P}(A)$ ***and*** $\mathbb{P}(B \mid A) = \mathbb{P}(A)$. Likewise, either of these two statements implies $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$. We can therefore use this single statement as our definition of statistical independence.

Definition: Events A and B are **statistically independent** if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$.

Definition: If there are more than two events, we say events $A_1, A_2, \ldots, A_n$ are **mutually independent** if

$$\mathbb{P}(A_1 \cap A_2 \cap \ldots \cap A_n) = \mathbb{P}(A_1)\mathbb{P}(A_2)\ldots\mathbb{P}(A_n), \quad \text{AND}$$

the same multiplication rule holds for every subcollection of the events too.

Independence for random variables

Random variables are independent if

That is, random variables X and Y are independent if, whatever the outcome of X, it has no influence on the outcome of Y.

Definition: Random variables X and Y are **statistically independent** if

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

for all *possible values* x *and* y.

We usually replace the cumbersome notation $\mathbb{P}(\{X = x\} \cap \{Y = y\})$ by the simpler notation

From now on, we will use the following notations interchangeably:

$$\mathbb{P}(\{X = x\} \cap \{Y = y\}) = \mathbb{P}(X = x \textit{ AND } Y = y) = \mathbb{P}(X = x, Y = y).$$

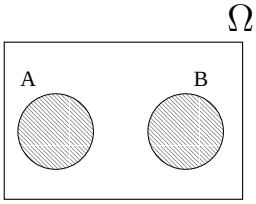
Thus X *and* Y *are independent if and only if*

$$\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x)\mathbb{P}(Y = y) \quad \textit{for ALL possible values } x, y.$$

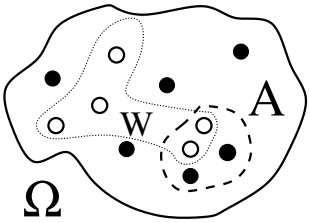
Independence in pictures

It is very difficult to draw a picture of statistical independence.

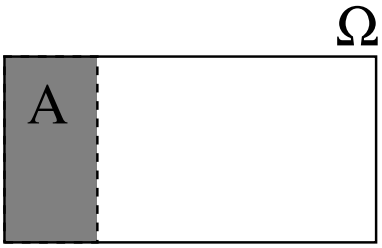
Are events A and B statistically independent?



Are events W and A statistically independent?



Question: How ***would*** you convey independence between events A and B on a diagram? Where would you draw event B?



[Hint: think of the formula $\mathbb{P}(B \mid A) = \mathbb{P}(B)$,
and what this means if we represent probabilities by ***areas.***]

1.5 Bayes' Theorem

Rev. Thomas Bayes (1702–1761), English clergyman and founder of Bayesian Statistics.

Bayes' Theorem follows directly from the multiplication rule. It shows how to invert the conditioning in conditional probabilities, i.e. how to express $\mathbb{P}(B \mid A)$ in terms of $\mathbb{P}(A \mid B)$.

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

Apply the multiplication rule to each side:

Thus $$\boxed{\mathbb{P}(B \mid A) = \frac{\mathbb{P}(A \mid B)\mathbb{P}(B)}{\mathbb{P}(A)}\,.}$$

1.6 The Partition Theorem (Law of Total Probability)

Definition: Events A and B are **mutually exclusive**, or **disjoint**, if

This means events A and B cannot happen together. If A happens, it excludes B from happening, and vice-versa.

If A and B are mutually exclusive,
For all other A and B,

Definition: Any number of events $B_1, B_2, \ldots, B_k$ are **mutually exclusive** if every pair of the events is mutually exclusive: ie. $B_i \cap B_j = \emptyset$ *for all* i, j *with* $i \neq j$.

Definition: A **partition** of Ω is a

That is, sets $B_1, B_2, \ldots, B_k$ form a partition of Ω if

$$B_i \cap B_j = \emptyset \text{ for all } i, j \text{ with } i \neq j,$$

$$\textbf{and} \quad \bigcup_{i=1}^{k} B_i = B_1 \cup B_2 \cup \ldots \cup B_k = \Omega.$$

$B_1, \ldots, B_k$ form a partition of Ω if they

Examples:

Partitioning an event *A*

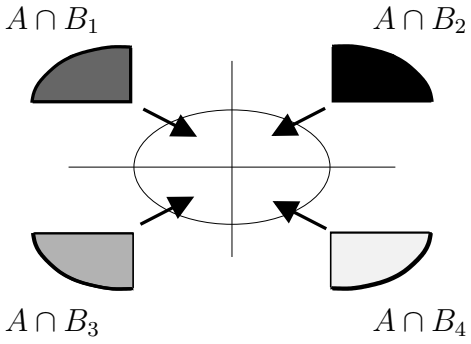
Any set A can be partitioned: it doesn't have to be Ω.
In particular, if $B_1, \ldots, B_k$ form a partition of Ω, then $(A \cap B_1), \ldots, (A \cap B_k)$ form a partition of A.

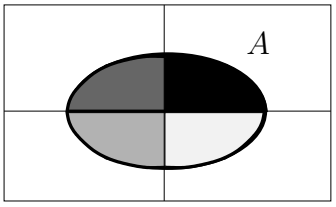
Theorem 1.6: The Partition Theorem (Law of Total Probability)

Both formulations of the Partition Theorem are very widely used, but especially the conditional formulation $\sum_{i=1}^{m} \mathbb{P}(A \mid B_i)\mathbb{P}(B_i)$.

The Partition Theorem in pictures

The Partition Theorem is easy to understand because it simply states that "the whole is the sum of its parts."





So:

Examples of conditional probability and partitions

Tom gets the bus to campus every day. The bus is on time with probability 0.6, and late with probability 0.4.

The sample space can be written as $\Omega = \{\text{bus journeys}\}$. We can formulate events as follows:

$$T = \{\text{on time}\} \qquad\qquad L = \{\text{late}\}$$

From the information given, the events have probabilities:

$$\mathbb{P}(T) = 0.6\,; \qquad\qquad \mathbb{P}(L) = 0.4.$$

(a) Do the events T and L form a partition of the sample space Ω? Explain why or why not.

The buses are sometimes crowded and sometimes noisy, both of which are problems for Tom as he likes to use the bus journeys to do his Stats assignments. When the bus is on time, it is crowded with probability 0.5. When it is late, it is crowded with probability 0.7. The bus is noisy with probability 0.8 when it is crowded, and with probability 0.4 when it is not crowded.

(b) Formulate events C and N corresponding to the bus being crowded and noisy. Do the events C and N form a partition of the sample space? Explain why or why not.

(c) Write down probability statements corresponding to the information given above. Your answer should involve two statements linking C with T and L, and two statements linking N with C.

(d) Find the probability that the bus is crowded.

(e) Find the probability that the bus is noisy.

1.7 Extra practice and reference

The following sections include some extra reading and examples taken from the old Stats 210 notes (pre-2015) before this material became a prerequisite for taking the course.

1. Probability of a union

The union operator, $A \cup B$, means A ***OR*** B ***OR both.*** For any events A and B on a sample space Ω:

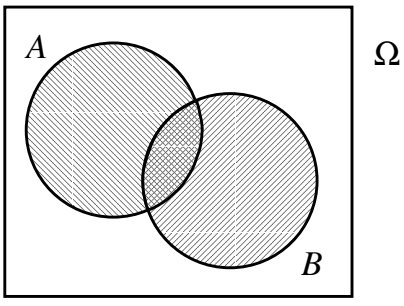
$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B).$$

For three or more events: e.g. for any events A, B, and C on Ω:

$$\begin{aligned} \mathbb{P}(A \cup B \cup C) \;=\; & \mathbb{P}(A) + \mathbb{P}(B) + \mathbb{P}(C) \\ & - \mathbb{P}(A \cap B) - \mathbb{P}(A \cap C) - \mathbb{P}(B \cap C) \\ & + \mathbb{P}(A \cap B \cap C)\,. \end{aligned}$$

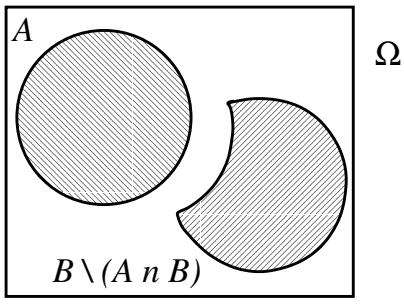
Explanation

To understand the formula, think of the Venn diagrams:



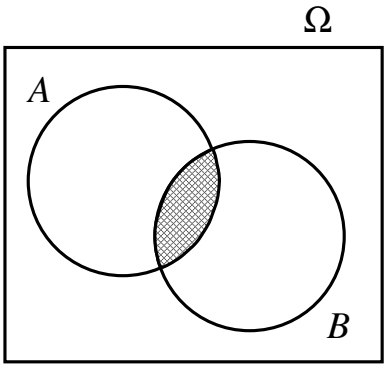
When we add $\mathbb{P}(A) + \mathbb{P}(B)$, we *add the intersection twice.*

So we have to *subtract the intersection once to get* $\mathbb{P}(A \cup B)$*:* $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.



Alternatively, think of $A \cup B$ as *two disjoint sets: all of* A*, and the bits of* B *without the intersection. So* $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \Big\{ \mathbb{P}(B) - \mathbb{P}(A \cap B) \Big\}$.

2. Probability of an intersection

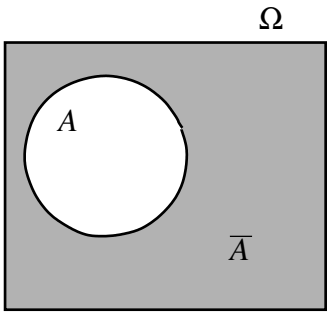


The intersection operator, $A \cap B$, means
both A AND B together.
There is no easy formula for $\mathbb{P}(A \cap B)$.

We might be able to use *statistical independence:*
if A and B are independent, then
$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$.

If A and B are not statistically independent,
we usually use *conditional probability:* $\mathbb{P}(A \cap B) = \mathbb{P}(A \mid B)\mathbb{P}(B)$ *for any events A and B. It is usually easier to find a conditional probability than an intersection.*

3. Probability of a complement



The complement of A is written $\overline{A}$ and denotes
everything in Ω that is not in A.

Clearly, $\mathbb{P}(\overline{A}) = 1 - \mathbb{P}(A)$.

Examples of basic probability calculations

An Australian survey asked people what sort of car they would like if they could choose any car at all. 13% of respondents had children and chose a large car. 12% of respondents did not have children and chose a large car. 33% of respondents had children.

Find the probability that a respondent:
(a) chose a large car;
(b) either had children or chose a large car (or both).

First define the sample space: $\Omega =$ { *respondents* }. *Formulate events:*

Let $C =$ { *has children* } $\overline{C} =$ { *no children* }

$L =$ { *chooses large car* }.

Next write down <u>all</u> the information given:

$$\begin{aligned} \mathbb{P}(C) &= 0.33 \\ \mathbb{P}(C \cap L) &= 0.13 \\ \mathbb{P}(\overline{C} \cap L) &= 0.12. \end{aligned}$$

(a) Asked for $\mathbb{P}(L)$.

$$\begin{aligned} \mathbb{P}(L) &= \mathbb{P}(L \cap C) + \mathbb{P}(L \cap \overline{C}) \qquad \textit{(Partition Theorem)} \\ &= \mathbb{P}(C \cap L) + \mathbb{P}(\overline{C} \cap L) \\ &= 0.13 + 0.12 \\ &= 0.25. \qquad\qquad \mathbb{P}\textit{(chooses large car)} = 0.25. \end{aligned}$$

(b) Asked for $\mathbb{P}(L \cup C)$.

$$\begin{aligned} \mathbb{P}(L \cup C) &= \mathbb{P}(L) + \mathbb{P}(C) - \mathbb{P}(L \cap C) \qquad \textit{(formula for probability of a union)} \\ &= 0.25 + 0.33 - 0.13 \\ &= 0.45. \end{aligned}$$

Example 2: Facebook statistics for New Zealand university students aged between 18 and 24 suggest that 22% are interested in music, while 34% are interested in sport. Define the sample space $\Omega = \{\text{NZ university students aged 18 to 24}\}$. Formulate events: $M = \{\text{interested in music}\}$, $S = \{\text{interested in sport}\}$.

(a) What is $\mathbb{P}(\overline{M})$?

(b) What is $\mathbb{P}(M \cap S)$?

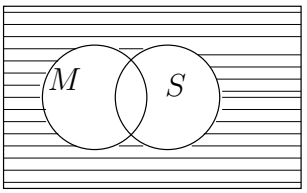
Information given: $\mathbb{P}(M) = 0.22$ $\mathbb{P}(S) = 0.34$.

(a)

$$\begin{aligned} \mathbb{P}(\overline{M}) &= 1 - \mathbb{P}(M) \\ &= 1 - 0.22 \\ &= 0.78. \end{aligned}$$

(b) We can not calculate $\mathbb{P}(M \cap S)$ *from the information given.*

(c) Given the further information that 48% of the students are interested in neither music nor sport, find $\mathbb{P}(M \cup S)$ and $\mathbb{P}(M \cap S)$.



Information given: $\mathbb{P}(\overline{M \cup S}) = 0.48$.

Thus
$$\begin{aligned} \mathbb{P}(M \cup S) &= 1 - \mathbb{P}(\overline{M \cup S}) \\ &= 1 - 0.48 \\ &= 0.52. \end{aligned}$$

Probability that a student is interested in music, or sport, or both.

$$\begin{aligned} \mathbb{P}(M \cap S) &= \mathbb{P}(M) + \mathbb{P}(S) - \mathbb{P}(M \cup S) \qquad \textit{(probability of a union)} \\ &= 0.22 + 0.34 - 0.52 \\ &= 0.04. \end{aligned}$$

Only 4% of students are interested in both *music and sport.*

(d) Find the probability that a student is interested in music, but *not* sport.

$$\begin{aligned} \mathbb{P}(M \cap \overline{S}) &= \mathbb{P}(M) - \mathbb{P}(M \cap S) \qquad \textit{(Partition Theorem)} \\ &= 0.22 - 0.04 \\ &= 0.18. \end{aligned}$$

1.8 Probability Reference List

The following properties hold for all events A, B, and C on a sample space Ω.

- $\mathbb{P}(\emptyset) = 0$ and $\mathbb{P}(\Omega) = 1$. *$\emptyset$ is the 'empty set': the event with no outcomes.*
- $0 \leq \mathbb{P}(A) \leq 1$: *probabilities are always between 0 and 1.*
- **Complement:** $\mathbb{P}(\overline{A}) = 1 - \mathbb{P}(A)$.
- **Probability of a union:** $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.
  For three events A, B, C:

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

  If A and B are **mutually exclusive**, then $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$.

- **Conditional probability:** $\mathbb{P}(A \mid B) = \dfrac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$.

- **Multiplication rule:** $\mathbb{P}(A \cap B) = \mathbb{P}(A \mid B)\mathbb{P}(B) = \mathbb{P}(B \mid A)\mathbb{P}(A)$.

- **The Partition Theorem:** if $B_1, B_2, \ldots, B_m$ form a partition of Ω, then

$$\mathbb{P}(A) = \sum_{i=1}^{m} \mathbb{P}(A \cap B_i) = \sum_{i=1}^{m} \mathbb{P}(A \mid B_i)\mathbb{P}(B_i) \quad \text{for any event } A.$$

  As a special case, B and $\overline{B}$ partition Ω, so:

$$\begin{aligned} \mathbb{P}(A) &= \mathbb{P}(A \cap B) + \mathbb{P}(A \cap \overline{B}) \\ &= \mathbb{P}(A \mid B)\mathbb{P}(B) + \mathbb{P}(A \mid \overline{B})\mathbb{P}(\overline{B}) \quad \text{for any } A,\ B. \end{aligned}$$

- **Bayes' Theorem:** $\mathbb{P}(B \mid A) = \dfrac{\mathbb{P}(A \mid B)\mathbb{P}(B)}{\mathbb{P}(A)}$.

  More generally, if $B_1, B_2, \ldots, B_m$ form a partition of Ω, then

$$\mathbb{P}(B_j \mid A) = \frac{\mathbb{P}(A \mid B_j)\mathbb{P}(B_j)}{\sum_{i=1}^{m} \mathbb{P}(A \mid B_i)\mathbb{P}(B_i)} \quad \text{for any } j.$$

- **Chains of events:** for any events A_1, A_2, A_3,

$$\mathbb{P}(A_1 \cap A_2 \cap A_3) = \mathbb{P}(A_1)\,\mathbb{P}(A_2 \mid A_1)\,\mathbb{P}(A_3 \mid A_2 \cap A_1)\,.$$

- **Statistical independence:** events A and B are **independent** if and only if

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

  Alternatively, either of the following statements is necessary and sufficient for A and B to be independent: $\mathbb{P}(A \mid B) = \mathbb{P}(A)$ and $\mathbb{P}(B \mid A) = \mathbb{P}(B)$.

- **Manipulating conditional probabilities:**

  If $\mathbb{P}(B) > 0$, then we can treat $\mathbb{P}(\cdot \mid B)$ just like $\mathbb{P}$: for example,

  ★ if A_1 and A_2 are mutually exclusive, then

$$\begin{aligned} \mathbb{P}(A_1 \cup A_2 \mid B) &= \mathbb{P}(A_1 \mid B) + \mathbb{P}(A_2 \mid B) \\ \text{compare with the usual formula,} \quad \mathbb{P}(A_1 \cup A_2) &= \mathbb{P}(A_1) + \mathbb{P}(A_2). \end{aligned}$$

  ★ if $A_1, \ldots, A_m$ partition the sample space Ω, then

$$\mathbb{P}(A_1 \mid B) + \mathbb{P}(A_2 \mid B) + \ldots + \mathbb{P}(A_m \mid B) = 1;$$

  ★ $\mathbb{P}(A \mid B) = 1 - \mathbb{P}(\overline{A} \mid B)$ for any A.

Note: it is ***not*** generally true that $\mathbb{P}(A \mid B) = 1 - \mathbb{P}(A \mid \overline{B})$.