

Introduction to Machine Learning

A quick refresher course in probability theory

Second lecture, 12.01.2022

Phuc Loi Luu, PhD
p.luu@garvan.org.au
luu.p.loi@googlemail.com

Roadmap for today

- Discrete probability
- Random variables
- Joint density, marginal density and the transformation law
- Expectation, variance, covariance, correlation, quantiles
- Independence, conditional probability, conditional independence
- Basic notions from statistics

Random experiments and Sample Space (1)

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 .

Example:

Random experiment: Toss a coin once.

Sample space: $\Omega = ?$

Random experiments and Sample Space (2)

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 .

Example:

Random experiment: Toss a coin once.

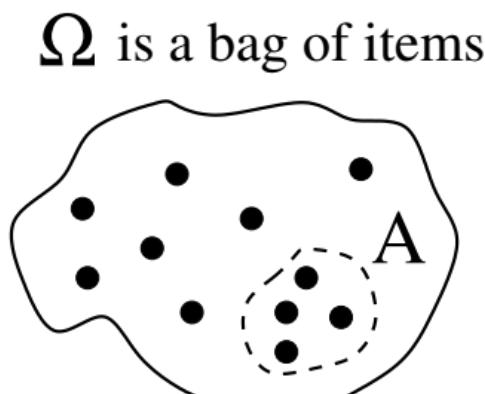
Sample space: $\Omega = \{\text{head, tail}\}$

Events vs Sample Space

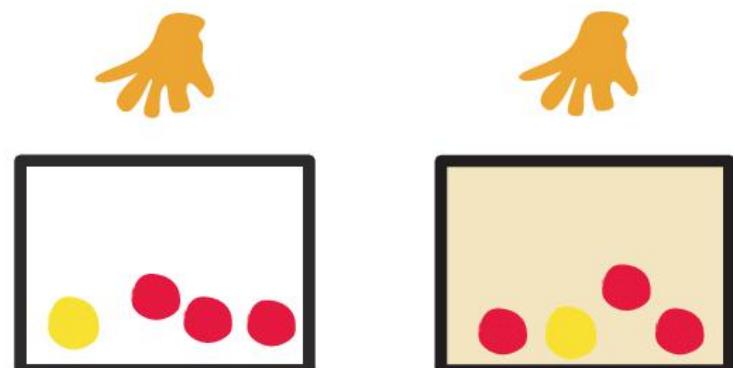
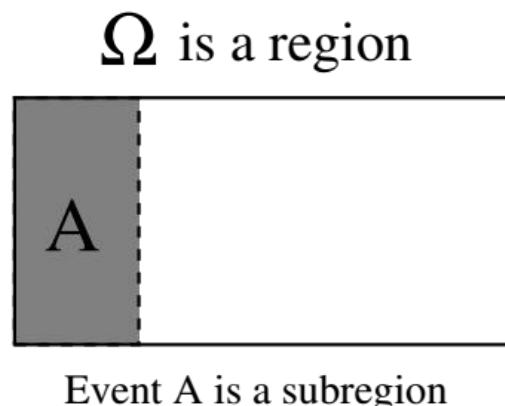
Definition: An event, A , is also a collection of outcomes. It is a subset of Ω .

- An event A is ‘something that could happen’.
- 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 $A \cap B$, $A \cup B$, \bar{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.



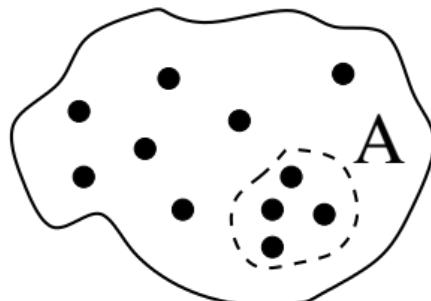
Event A is a smaller bag of items



Probabilities

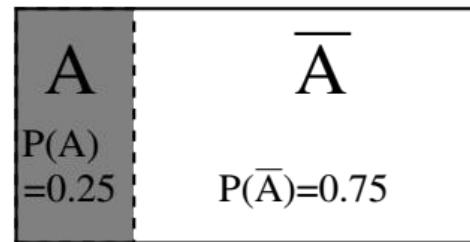
The idea of probability is to *attach a number to every item or event in Ω that reflects how likely the event is to occur.*

Ω is a bag of items

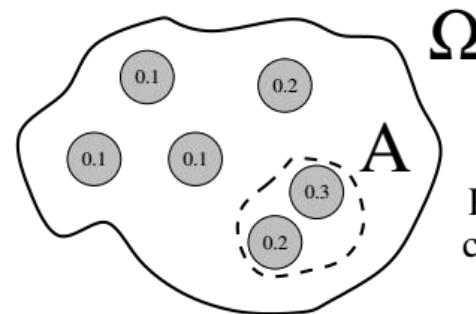


$P(A)=4/11$ if all items
are equally likely

Ω is a region



Probability is represented by AREA

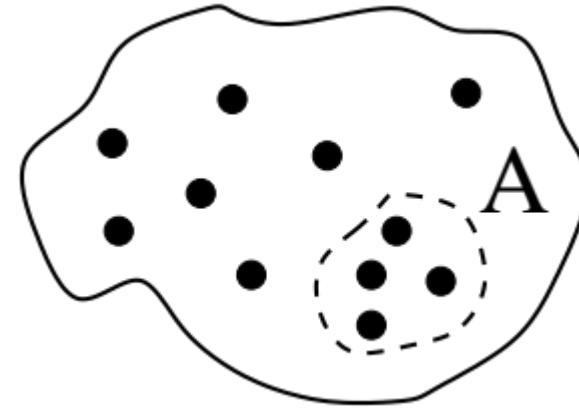


$P(A)=0.5$ even though bag A
contains only 2 out of 6 items

Question: What random experiments are we implicitly assuming here?

- Ω as a bag of items? *Pick an item at random from the bag.*
- Ω as a region? *Select a point at random from the region.*

Probability distribution



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**: a probability distribution describes how much ‘cake’ is given to each subset of Ω .

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

Formal probability definition: the three axioms

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

- This means that the total amount of ‘cake’ available is 1.
- It also makes it clear that probability depends on the sample space, Ω , and has no meaning unless Ω is defined first.

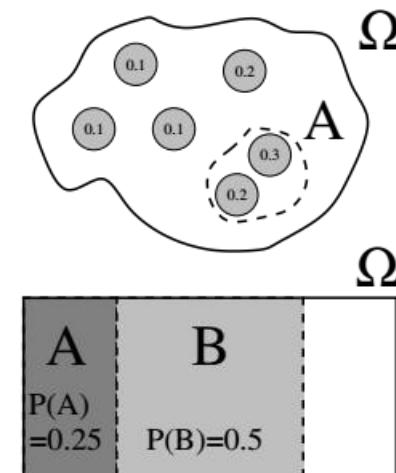
Axiom 2: $0 \leq \mathbb{P}(A) \leq 1$ for all events A .

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

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

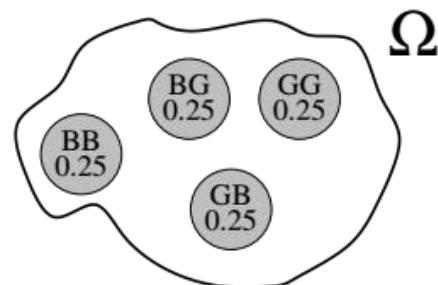
$$\mathbb{P}(A_1 \cup A_2 \cup \dots \cup A_n) = \mathbb{P}(A_1) + \mathbb{P}(A_2) + \dots + \mathbb{P}(A_n).$$

- 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, $\mathbb{P}(A \cup B) = 0.25 + 0.5 = 0.75$.



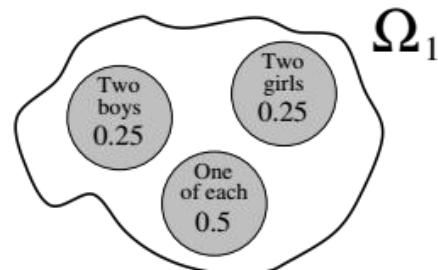
Examples of probability distributions

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 *four equally-likely outcomes*:



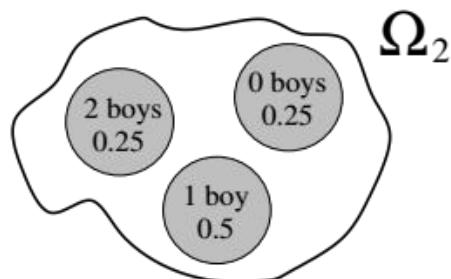
So if we pick a two-child family at random, we have probability $1/4 = 0.25$ of getting each of the outcomes BB, BG, GB, and GG.

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:



Examples of probability distributions

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 *counting the number of boys: 0, 1, or 2, giving $\Omega_2 = \{0, 1, 2\}$:*

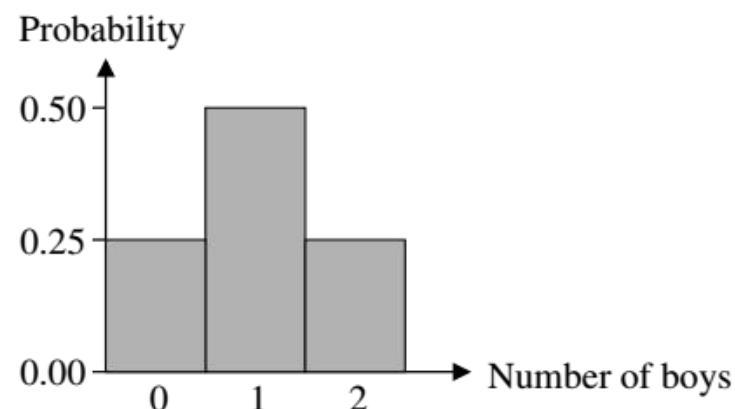


Now think of a different way of picturing Ω_2 that is easier to extend:

This representation is more like the ‘region’ image of Ω we used earlier, where probabilities are represented by *areas*.

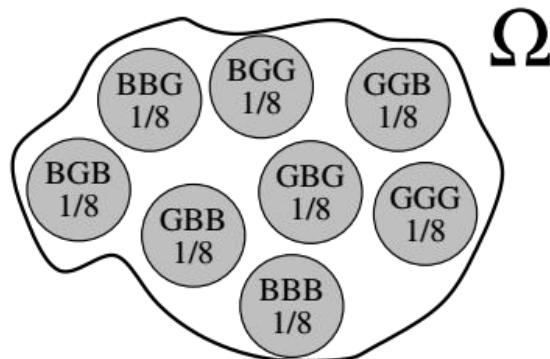
It also has the huge advantage of being a flexible, graphical display.

Question: where would you draw Ω_2 ?

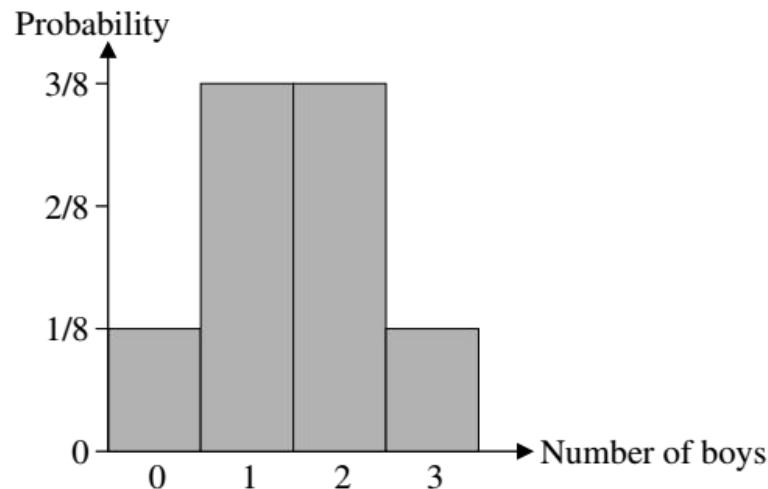


Examples of probability distributions

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, *we have to take account of all the different orderings when we calculate probabilities:*

$$\mathbb{P}(2 \text{ boys}) = \mathbb{P}(BBG) + \mathbb{P}(BGB) + \mathbb{P}(GBB) = 3 \times \frac{1}{8} = \frac{3}{8}.$$

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 **variable number**, 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, *a random variable is a numeric summary of the outcome of a random experiment.*

In formal language, a random variable is a mapping from Ω to the real numbers: $X : \Omega \rightarrow \mathbb{R}$. 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 $X = 2$.

Random variables

Example:

Random experiment: Toss a coin once.

Sample space: $\Omega = \{\text{head, tail}\}$.

An example of a random variable: $X : \Omega \rightarrow \mathbb{R}$ maps “head” $\rightarrow 1$, “tail” $\rightarrow 0$.

Essential point: A random variable is a way of producing random real numbers.

	HH	HT	TT
X	0	1	2

Everything you need to know about random variables

- Random variables always have **CAPITAL LETTERS**, e.g. X or Y .

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 **rv**.
- You can think of a random variable simply as *the name of a mechanism for generating random real numbers*.

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:

$$\mathbb{P}(X = 0) = \frac{1}{8}; \quad \mathbb{P}(X = 1) = \frac{3}{8}; \quad \mathbb{P}(X = 2) = \frac{3}{8}; \quad \mathbb{P}(X = 3) = \frac{1}{8}.$$

- If we want to refer to a generic, unspecified, value of a random variable, we use a **lower-case letter, such as x or y** .

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 $P(X = x)$

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 ***name — just like your own name***.

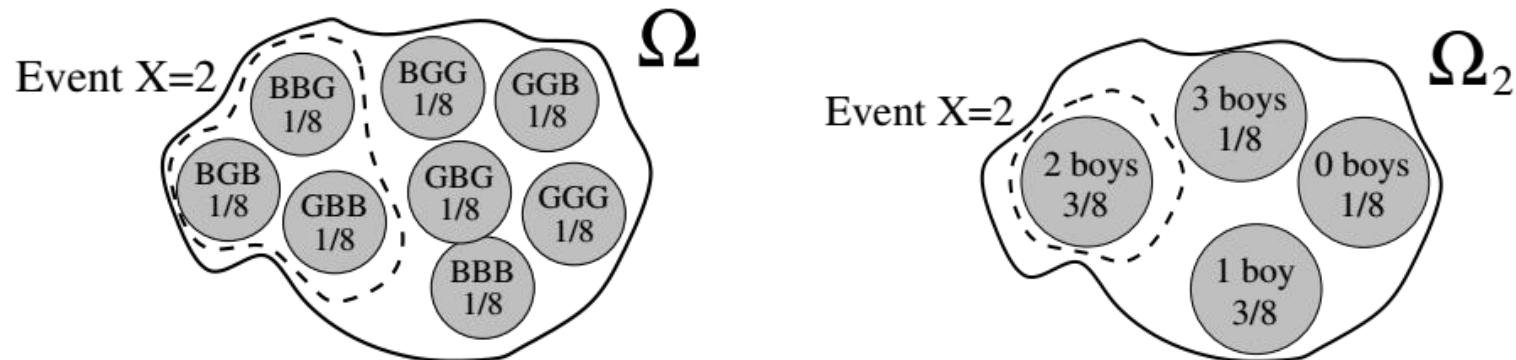
If we say $X = 2$, it is a bit like saying, ‘Susan is in the kitchen.’ It tells us a current observation of the random variable that we have called X .

- 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 ***thing that happens***.

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

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

Because $X = 2$ is an event, it is a *set: a subset of the sample space.*



Crucially, *use set notation to combine expressions like $X = 2$.*

- $\mathbb{P}(X = x)$ is a **real number**: it is a number between 0 and 1. *Use operations like + and * to combine probabilities: for example, $\mathbb{P}(X \leq 1) = \mathbb{P}(X = 0) + \mathbb{P}(X = 1)$.*

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.

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

Right

$$X = 2 \cup X = 3$$

Event that X takes the value 2 OR 3

$$\mathbb{P}(X = 2 \cup X = 3)$$

Probability of the event that X is 2 OR 3

$$\mathbb{P}(X = 2) + \mathbb{P}(X = 3)$$

Probability of the event that X is 2 OR 3

$$X \leq 2 \cap X > 1$$

Event that X takes a value that is

BOTH ≤ 2 AND > 1

(the value $X = 2$ is the only possibility)

$$\mathbb{P}(X \leq 2 \cap X > 1)$$

Probability that X is BOTH ≤ 2 AND > 1

Wrong

$$X = 2 + X = 3$$

$$\mathbb{P}(X = 2) \cup \mathbb{P}(X = 3)$$

$$\mathbb{P}(X = 2 + X = 3)$$

$$X \leq 2 \times X > 1$$

$$\mathbb{P}(X \leq 2) \cap \mathbb{P}(X > 1)$$

Class task for differences between X , x , $\{X = x\}$, and $P(X = x)$

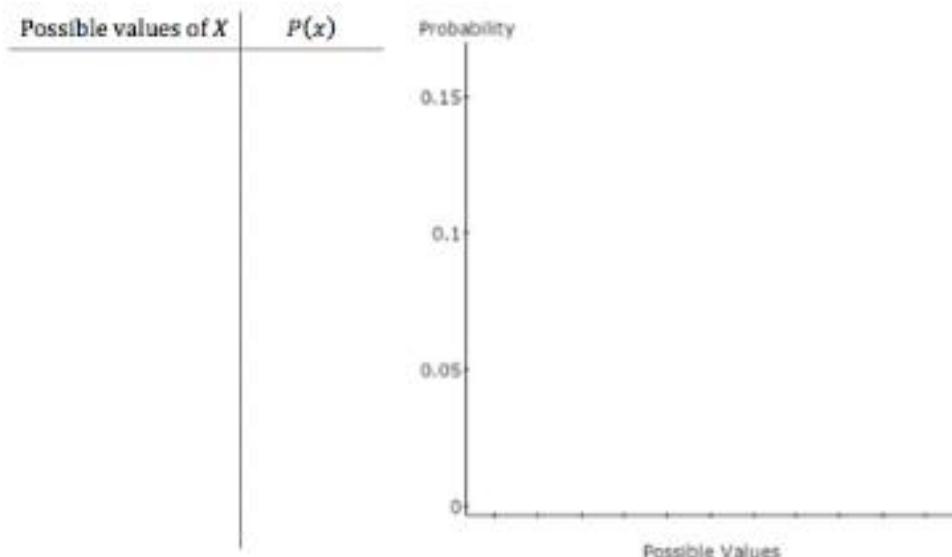


X	1	2	3	4	5	6
$P(X=x)$	1/6	1/6	1/6	1/6	1/6	1/6

2. When you roll a fair, 6-sided die twice, there are 36 possible outcomes. The possible outcomes are listed below where the first number is the result of the first roll and the second number is the result of the second roll.

(1, 1) (1, 2) (1, 3) (1, 4) (1, 5) (1, 6)
(2, 1) (2, 2) (2, 3) (2, 4) (2, 5) (2, 6)
(3, 1) (3, 2) (3, 3) (3, 4) (3, 5) (3, 6)
(4, 1) (4, 2) (4, 3) (4, 4) (4, 5) (4, 6)
(5, 1) (5, 2) (5, 3) (5, 4) (5, 5) (5, 6)
(6, 1) (6, 2) (6, 3) (6, 4) (6, 5) (6, 6)

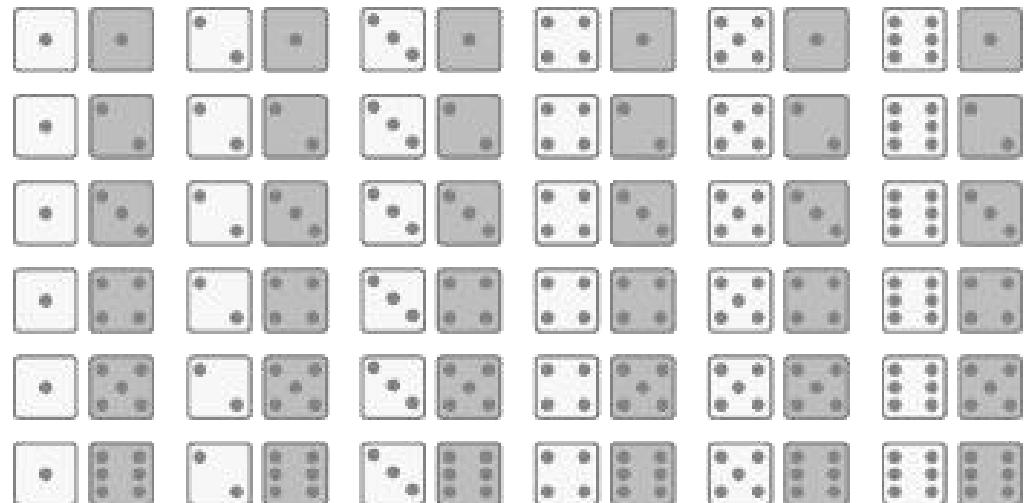
- a. Let X = the sum of the two rolls. Fill in the table below with the possible values of the random variable and the probability of each possible value. Then use this information to construct a graph of the probability distribution where the x-axis represents the possible values of X and the y-axis represents the probability of each possible value.



Class task for differences between X , x , $\{X = x\}$, and $P(X = x)$



X	$p(X)$	Dice combinations
2	1/36	(1,1)
3	2/36	(1,2); (2,1)
4	3/36	(1,3); (3,1); (2,2)
5	4/36	(1,4); (4,1); (2,3); (3,2)
6	5/36	(1,5); (5,1); (2,4); (4,2); (3,3)
7	6/36	(1,6); (6,1); (2,5); (5,2); (3,4); (4,3)
8	5/36	(2,6); (6,2); (3,5); (5,3); (4,4)
9	4/36	(3,6); (6,3); (4,5); (5,4)
10	3/36	(4,6); (6,4); (5,5)
11	2/36	(5,6); (6,5)
12	1/36	(6,6)



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 *counts the number of successes in a fixed number n of independent trials, where each trial has two possible outcomes: Success with probability p , and Failure with probability $1 - p$.*

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:

‘Success’ = ‘Head’; $p = \mathbb{P}(\text{Head}) = 0.5$.

(2) Repeated rolls of a fair die: $p = \mathbb{P}(\text{Get a } 6) = 1/6$.



Bernoulli trials and the Binomial Distribution



Note: Saying the trials are *independent* means that *they do not influence each other*.

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 *it takes only two values, 0 and 1. We write $Y \sim \text{Bernoulli}(p)$, where $p = \mathbb{P}(Y = 1)$.*

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:

$$f_Y(y) = \mathbb{P}(Y = y) = \begin{cases} p & \text{if } y = 1 & (\text{Success}) \\ 1 - p & \text{if } y = 0 & (\text{Failure}) \end{cases}$$

We often write the probability function in table format:

y	0	1
$\mathbb{P}(Y = y)$	$1 - p$	p

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: $n = 3$ *trials (children)*; $p = \mathbb{P}(\text{Boy}) = 0.5$ *for each child*.
- X is the number of 6's obtained in 10 rolls of a die: $n = 10$ *trials (die rolls)*; $p = \mathbb{P}(\text{Get a 6}) = 1/6$ *for each roll*.

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 $X \sim \text{Binomial}(n, p)$, or $X \sim \text{Bin}(n, p)$.

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.

Binomial Distribution

Take the example of $X = \text{number of boys in a 3-child family}$, so $X \sim \text{Binomial}(n = 3, p = 0.5)$.

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

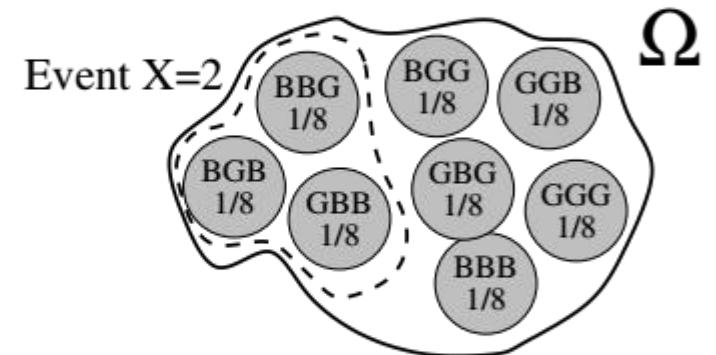
$$\mathbb{P}(X = 2) = \mathbb{P}(BBG) + \mathbb{P}(BGB) + \mathbb{P}(GBB).$$

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?

There are 3 trials (children), and we need to choose 2 of them to be boys. The number of ways of choosing 2 trials from 3 is:

$${}^3C_2 = \binom{3}{2} = \frac{3!}{(3-2)! 2!} = \frac{3 \times 2 \times 1}{1 \times (2 \times 1)} = 3.$$



Binomial Distribution

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

Answer:

$${}^{10}C_6 = \binom{10}{6} = \frac{10!}{(10-6)!6!} = 210 \quad \text{--- use calculator button } {}^nC_r .$$

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

Answer:

$${}^nC_x = \binom{n}{x} = \frac{n!}{(n-x)!x!} .$$

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: $p \times (1-p) \times (1-p) \times p \times (1-p) = p^2(1-p)^3$. This will be the same whatever order the successes and failures are in. But it only gives the probability for one ordering.

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

Answer: $p^x(1-p)^{n-x}$.

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: (Number of orderings) \times (probability of each ordering) $= \binom{n}{x} p^x(1-p)^{n-x}$.

Binomial Distribution

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

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

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

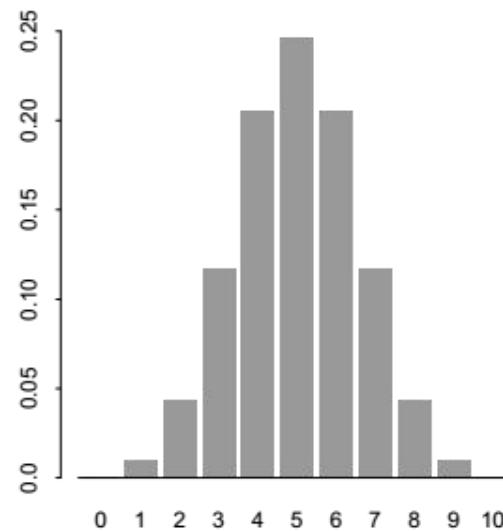
- Note:**
1. Importantly, $f_X(x) = 0$ if x is not one of the values $0, 1, \dots, n$.
The correct way to write the range of values is $x = 0, \dots, n$.
 - Writing $x \in [0, n]$ is **wrong**, because this includes decimals like 0.4.
 - Writing $x = 0, 1, \dots$ is **wrong**, because the range of values 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

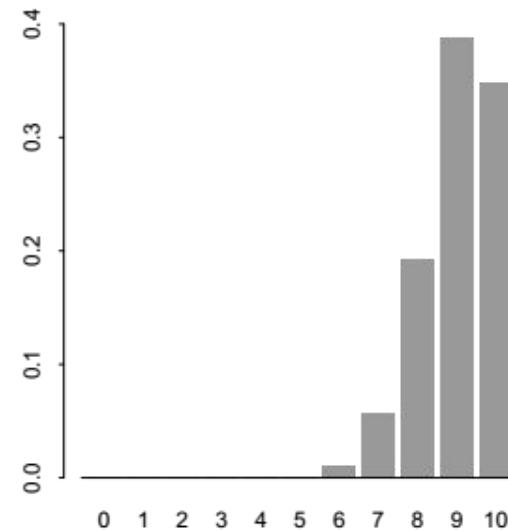
The shape of the Binomial distribution depends upon the values of n and p . For small n , the distribution is almost symmetrical for values of p close to 0.5, but highly skewed for values of p close to 0 or 1. As n increases, the distribution becomes more and more symmetrical, and there is noticeable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

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

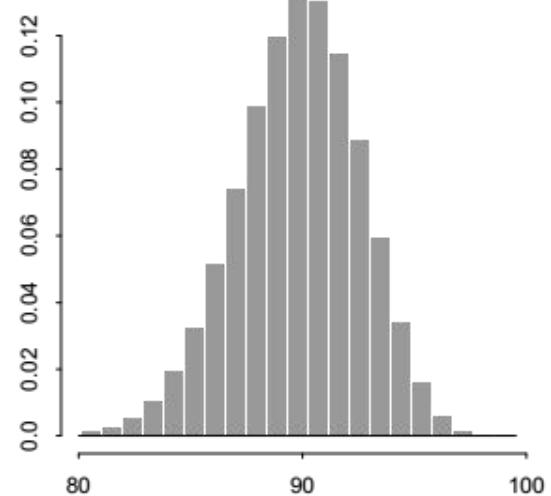
$n = 10, p = 0.5$



$n = 10, p = 0.9$



$n = 100, p = 0.9$



```
barplot(dbinom(x=0:10, size=10, prob=0.5))
```

<https://statisticsglobe.com/binomial-distribution-in-r-dbinom-pbinom-qbinom-rbinom>

Shape of the Binomial distribution

Probabilities for $X \sim \text{Binomial}(n = 25, p = 0.25)$

```
barplot(dbinom(x=0:25, size=25, prob=0.25))
```

<https://statisticsglobe.com/binomial-distribution-in-r-dbinom-pbinom-qbinom-rbinom>

Sum of independent Binomial random variables

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

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

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 $n + m$ *trials*.

Note: X and Y must both share *the same value of p* .

Binomial random variable as a sum of Bernoulli random variables

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

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

This is because X and $Y_1 + \dots + Y_n$ both represent *the number of successes in n independent trials, where each trial has success probability p .*

Cumulative distribution function, $F_X(x)$

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

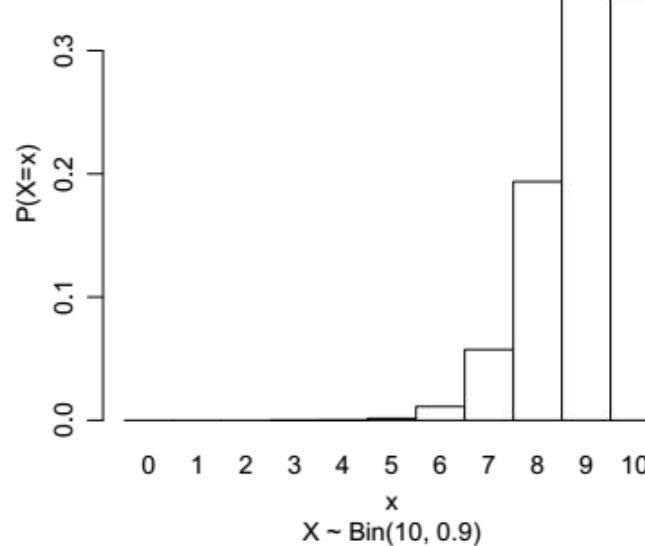
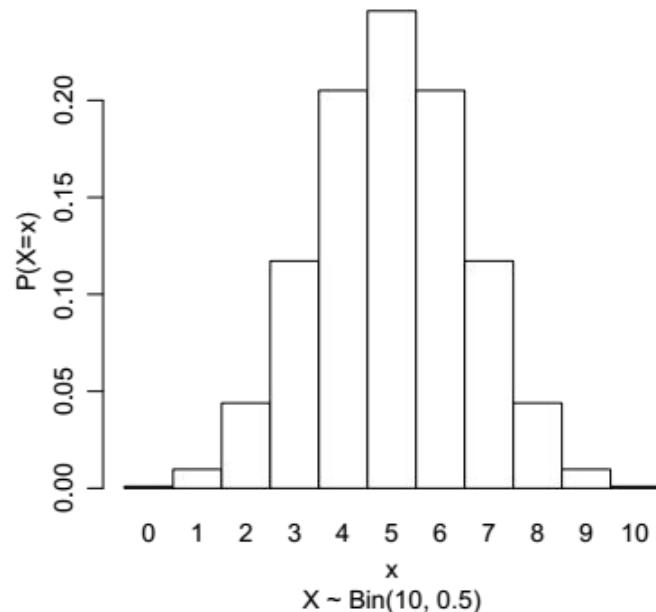
Another function that is widely used is the *cumulative distribution function*, or CDF, written as $F_X(x)$.

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

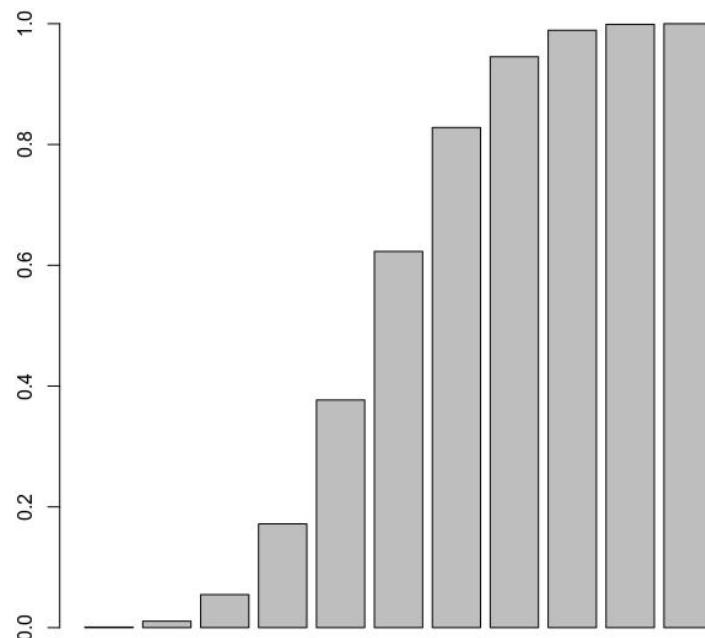
The cumulative distribution function, $F_X(x)$, *sweeps up all the probability up to and including the point x .*



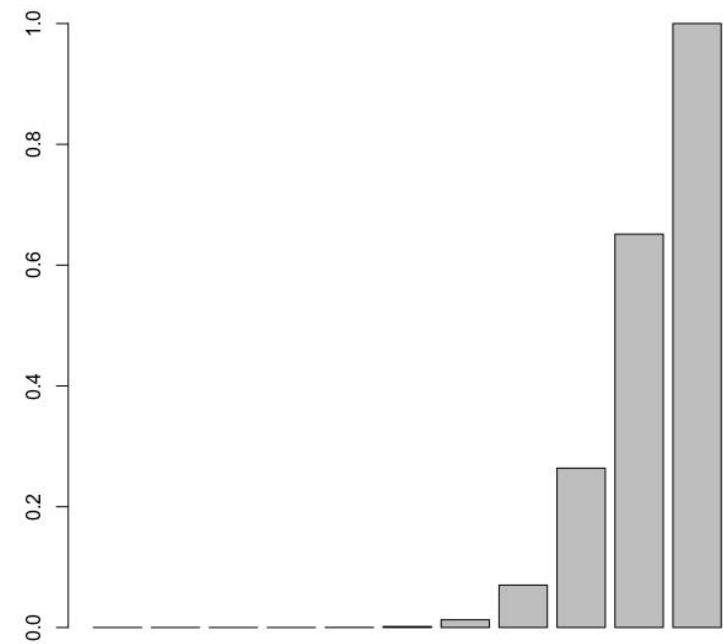
Clearly,

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

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



```
barplot(pbinom(0:10, size=10, prob=0.5))
```



```
barplot(pbinom(0:10, size=10, prob=0.9))
```

Using the cumulative distribution function to find probabilities

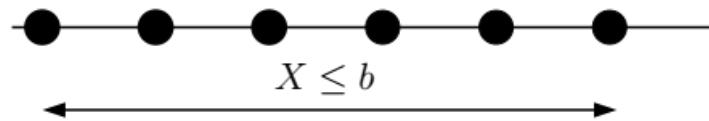
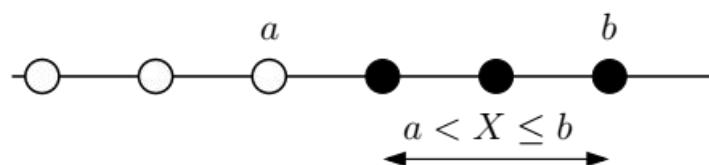
$$\mathbb{P}(a < X \leq b) = F_X(b) - F_X(a) \quad \text{if } b > a.$$

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

$$\mathbb{P}(X \leq b) = \mathbb{P}(X \leq a) + \mathbb{P}(a < X \leq b)$$

So $F_X(b) = F_X(a) + \mathbb{P}(a < X \leq b)$

$$\Rightarrow F_X(b) - F_X(a) = \mathbb{P}(a < X \leq b).$$



Using the cumulative distribution function to find probabilities

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. $\mathbb{P}(X \leq 30)? \quad F_X(30).$

2. $\mathbb{P}(X < 30)? \quad F_X(29).$

3. $\mathbb{P}(X \geq 56)?$

$$1 - \mathbb{P}(X < 56) = 1 - \mathbb{P}(X \leq 55) = 1 - F_X(55).$$

4. $\mathbb{P}(X > 42)?$

$$1 - \mathbb{P}(X \leq 42) = 1 - F_X(42).$$

5. $\mathbb{P}(50 \leq X \leq 60)?$

$$\mathbb{P}(X \leq 60) - \mathbb{P}(X \leq 49) = F_X(60) - F_X(49).$$

Conditional probability

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

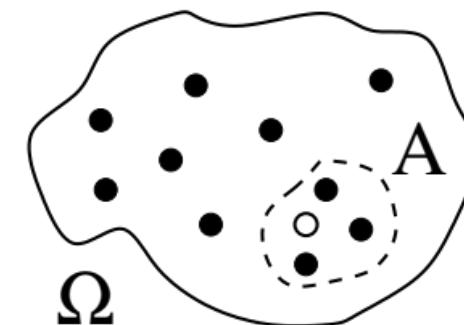
$\mathbb{P}(\Omega) = 1$, so *the symbol \mathbb{P} is only defined relative to a particular sample space Ω .*

Conditional probability is about *changing the sample space*.

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?

$$\mathbb{P}(\text{white ball}) = \frac{1}{11}.$$



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 **event**.

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

$$\mathbb{P}(\text{white ball if we select only from } A) = \frac{1}{4}.$$

We use a shorthand notation to write this down:

$$\mathbb{P}(\text{white ball if we select only from } A) = \mathbb{P}(\text{white ball} | A) = \frac{1}{4}.$$

Conditional probability

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

$\mathbb{P}(\text{white ball} | A)$ is called a *conditional probability*, and we say we have *conditioned on event A* .

Note: The vertical bar in $\mathbb{P}(\text{white ball} | A)$ is vertical: |.

Do not write a conditional probability as $\mathbb{P}(W/A)$ or $\mathbb{P}(W \setminus A)$: it is $\mathbb{P}(W | A)$.

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 $\mathbb{P}(W | A)$, we have *changed the sample space* from Ω to A .

Conditional probability, an example

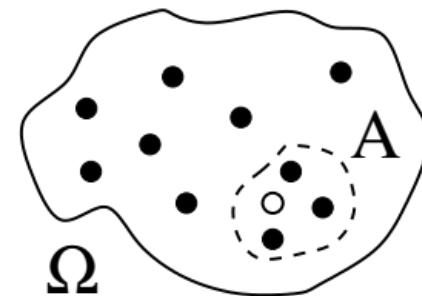
Define event $W = \{\text{pick white ball}\}$.

We have said: $\mathbb{P}(W) = \frac{1}{11}$.

This means $\mathbb{P}(W \text{ from within } \Omega) = \frac{1}{11}$,

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: $\mathbb{P}(W | A) = \frac{1}{4}$.



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

Answer: $\mathbb{P}(A | A) = 1$, because if we select items from within A , we are definitely going to select something in A .

The conditional probability $\mathbb{P}(W | A)$ 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 $\mathbb{P}(W | A)$ 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.

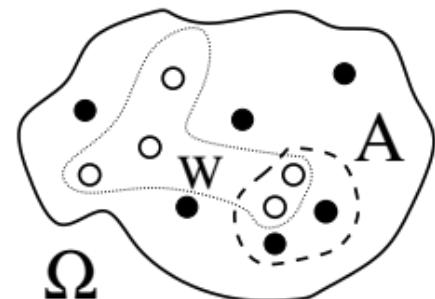
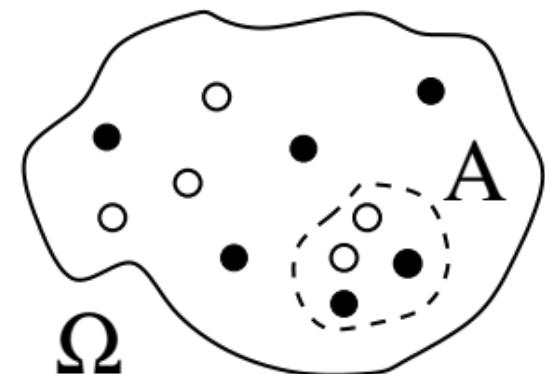
Question: What is $\mathbb{P}(W)$?

Answer: $\mathbb{P}(W)$ refers to the probability within the whole sample space Ω , so $\mathbb{P}(W) = \frac{5}{11}$.

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

Answer: $\mathbb{P}(W | A)$ refers to the probability within bag A only, so $\mathbb{P}(W | A) = \frac{2}{4}$.

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



Formula for conditional probability

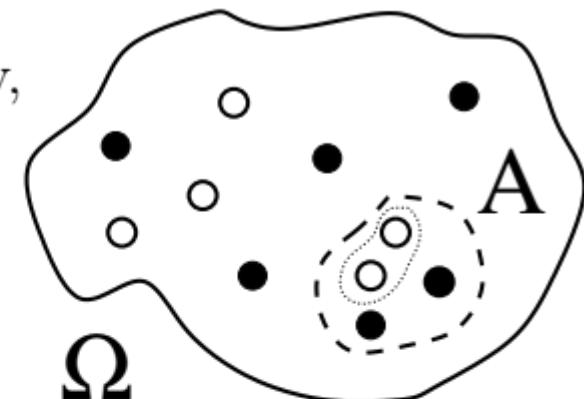
It is obvious from the diagram that $\mathbb{P}(W | A) = \frac{2}{4}$.

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 $W \cap A$.

The dashed bag represents the set A .

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



$$\mathbb{P}(W | A) = \frac{\text{probability in the dotted bag}}{\text{probability in the dashed bag}} = \frac{\mathbb{P}(W \cap A)}{\mathbb{P}(A)}.$$

This reasoning gives us our formal definition of conditional probability.

Formula for 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 $\mathbb{P}(B | A)$, and defined as

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

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

Note: $\mathbb{P}(B | A)$ gives $\mathbb{P}(B \text{ and } A, \text{ from within the set of } A\text{'s only})$.

$\mathbb{P}(B \cap A)$ gives $\mathbb{P}(B \text{ and } A, \text{ from the whole sample space } \Omega)$.

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

The notation $\mathbb{P}(B | A)$ is good because it emphasises that the *denominator* of the proportion is A . In a sense, $\mathbb{P}(B | A)$ 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 *when they are picked from the whole sample space, Ω* .
- $\mathbb{P}(B | A)$ means the probability of picking someone who satisfies B *when they are picked only from set A (dark hair people): it is the probability of B WITHIN A* .
- $\mathbb{P}(B \cap A)$ means the probability of picking someone who satisfies **BOTH B AND A** , *when they are picked from the whole sample space, Ω* .

This means you can easily identify which probabilities are conditional and which are intersections by looking to see *who we are picking FROM*. 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: $F = \{\text{first years}\}$; $S = \{\text{second years}\}$; $T = \{\text{third years}\}$; $O = \{\text{other students, e.g. exchange students, COPs, ...}\}$.

Language of conditional probability, exercise

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: $\mathbb{P}(A \cap B)$.
- Probability a dark-haired person has blue eyes: $\mathbb{P}(B | A)$.
- Probability a person has passed two GenEd papers: $\mathbb{P}(X = 2)$.
- Probability a second-year has passed two GenEd papers: $\mathbb{P}(X = 2 | S)$.
- Probability a first-year has passed two GenEd papers: $\mathbb{P}(X = 2 | F)$.
- Probability a dark-haired first-year has passed one or two GenEd papers:
$$\mathbb{P}(X = 1 \cup X = 2 | F \cap A) = \mathbb{P}(X = 1 | F \cap A) + \mathbb{P}(X = 2 | F \cap A).$$

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 $\mathbb{P}(\cdot | A)$ should behave exactly like ordinary probabilities $\mathbb{P}(\cdot)$, as long as *all* the probabilities are conditioned on the same event A .

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

Answer: Replace A by Ω : this gives

$$\mathbb{P}(B | \Omega) + \mathbb{P}(\overline{B} | \Omega) = \mathbb{P}(B) + \mathbb{P}(\overline{B}) = 1.$$

So, yes, $\mathbb{P}(B | A) + \mathbb{P}(\overline{B} | A) = 1$ for any other sample space A too.

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

Answer: Try to replace the conditioning set by Ω : we can't! There are two conditioning sets: A and \overline{A} .

The expression is NOT true. 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 ,

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

Proof: *Immediate from the definitions:*

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

and $\mathbb{P}(B | A) = \frac{\mathbb{P}(B \cap A)}{\mathbb{P}(A)} \Rightarrow \mathbb{P}(B \cap A) = \mathbb{P}(A \cap B) = \mathbb{P}(B | A)\mathbb{P}(A).$ □

Statistical independence

Events A and B are said to be ***independent*** if they *have no influence on each other*.

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 $X = \text{number of GenEd papers passed}$.

- A and B ? *Probably not: dark-haired people in this class might be more likely to have brown eyes, so less likely to have blue eyes?*
- A and F ? *Yes, these are probably independent.*
- F and S ? *Definitely not independent. No-one can be in BOTH first year AND second year, so each event STOPS the other one from happening. This is a very strong dependence.*
- Events $X = 2$ and F ? *Probably not independent: first-years are less likely to have passed two GenEd papers than second-years.*

Statistical independence

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

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

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 $\mathbb{P}(A | B)$ and $\mathbb{P}(B | A)$?

Statistical independence

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

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

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

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

So the ***single*** statement $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$, implies ***both*** statements $\mathbb{P}(A | B) = \mathbb{P}(A)$ ***and*** $\mathbb{P}(B | A) = \mathbb{P}(B)$. 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, \dots, A_n are **mutually independent** if

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

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

Independence for random variables

Random variables are independent if *they have no influence on each other.*

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

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

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

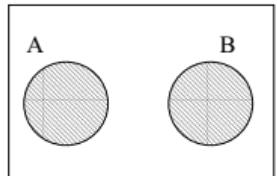
Thus X and Y are *independent if and only if*

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

Independence in pictures

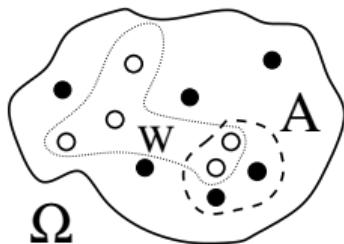
It is very difficult to draw a picture of statistical independence.

Are events A and B statistically independent?



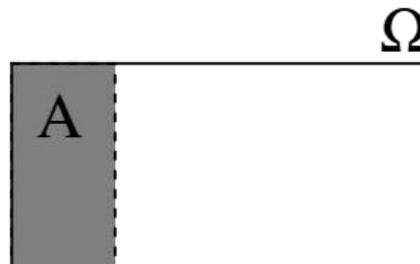
Ω *No, they are NOT independent.
Events A and B can't happen together.
They STOP each other from happening.
This is STRONG dependence — high influence.*

Are events W and A statistically independent?



*No, they are NOT independent.
 $\mathbb{P}(W | A) = 2/4$, but $\mathbb{P}(W) = 5/11$.
So $\mathbb{P}(W | A) \neq \mathbb{P}(W)$, so they are
NOT 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 $\mathbb{P}(B | A) = \mathbb{P}(B)$,
and what this means if we represent probabilities by **areas**.]

Bayes' Theorem

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

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

Apply the multiplication rule to each side:

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

Thus

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

Example 1

- Given:
 - A doctor knows that meningitis causes stiff neck 50% of the time
 - Prior probability of any patient having meningitis is 1/50,000
 - Prior probability of any patient having stiff neck is 1/20
- If a patient has stiff neck, what's the probability he/she has meningitis?

$$P(M | S) = \frac{P(S | M)P(M)}{P(S)} = \frac{0.5 \times 1/50000}{1/20} = 0.0002$$



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

Bayes' Theorem

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

- A, B = events
 $P(A|B)$ = probability of A given B is true
 $P(B|A)$ = probability of B given A is true
 $P(A), P(B)$ = the independent probabilities of A and B



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

Example 2

A doctor is called to see a sick child. The doctor has prior information that 90% of sick children in that neighborhood have the flu, while the other 10% are sick with measles. Let F stand for an event of a child being sick with flu and M stand for an event of a child being sick with measles. Assume for simplicity that $F \cup M = \Omega$, i.e., that there no other maladies in that neighborhood. A well-known symptom of measles is a rash (the event of having which we denote R). Assume that the probability of having a rash if one has measles is $P(R | M) = 0.95$. However, occasionally children with flu also develop rash, and the probability of having a rash if one has flu is $P(R | F) = 0.08$. Upon examining the child, the doctor finds a rash. What is the probability that the child has measles?

Bayes' Theorem

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

A, B = events

$P(A|B)$ = probability of A given B is true

$P(B|A)$ = probability of B given A is true

$P(A), P(B)$ = the independent probabilities of A and B



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

We use Bayes's formula.

$$\begin{aligned} P(M | R) &= \frac{P(R | M)P(M)}{(P(R | M)P(M) + P(R | F)P(F))} \\ &= \frac{0.95 \times 0.10}{(0.95 \times 0.10 + 0.08 \times 0.90)} \simeq 0.57. \end{aligned}$$

Which is nowhere close to 95% of $P(R|M)$.

Bayes' Theorem example 3

In a study, physicians were asked what the odds of breast cancer would be in a woman who was initially thought to have a 1% risk of cancer but who ended up with a positive mammogram result (a mammogram accurately classifies about 80% of cancerous tumors and 90% of benign tumors.)

95 out of a hundred physicians estimated the probability of cancer to be about 75%. Do you agree?



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

Bayes' Theorem example 3

In a study, physicians were asked what the odds of breast cancer would be in a woman who was initially thought to have a 1% risk of cancer but who ended up with a positive mammogram result (a mammogram accurately classifies about 80% of cancerous tumors and 90% of benign tumors.)

95 out of a hundred physicians estimated the probability of cancer to be about 75%. Do you agree?

Introduce the events:

$+$ = mammogram result is positive,

B = tumor is benign,

M = tumor is malignant.

Note that $B^c = M$. We are given $P(M) = .01$, so $P(B) = 1 - P(M) = .99$. We are also given the conditional probabilities $P(+ | M) = .80$ and $P(- | B) = .90$, where the event $-$ is the complement of $+$, thus $P(+ | B) = .10$

Bayes' formula in this case is

$$\begin{aligned}P(M | +) &= \frac{P(+ | M)P(M)}{(P(+ | M)P(M) + P(+ | B)P(B))} \\&= \frac{0.80 \times 0.01}{(0.80 \times 0.01 + 0.10 \times 0.99)} \\&\simeq 0.075\end{aligned}$$

So the chance would be 7.5%. A far cry from a common estimate of 75



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

Bayes' Theorem homework

1. Consider a test to detect a disease that 0.1 % of the population have. The test is 99 % effective in detecting an infected person. However, the test gives a false positive result in 0.5 % of cases. If a person tests positive for the disease what is the probability that they actually have it?
2. It is estimated that 50% of emails are spam emails. Some software has been applied to filter these spam emails before they reach your inbox. A certain brand of software claims that it can detect 99% of spam emails, and the probability for a false positive (a non-spam email detected as spam) is 5%.

Now if an email is detected as spam, then what is the probability that it is in fact a non-spam email?



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

Bayes' Theorem

LIKELIHOOD

The probability of "B" being True, given "A" is True

PRIOR

The probability "A" being True. This is the knowledge.

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

POSTERIOR

The probability of "A" being True, given "B" is True

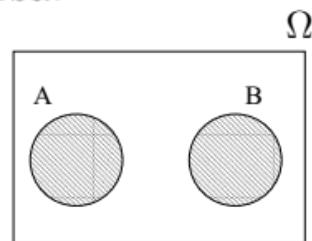
MARGINALIZATION

The probability "B" being True.

The Partition Theorem (Law of Total Probability)

Definition: Events A and B are mutually exclusive, or disjoint, if $A \cap B = \emptyset$.

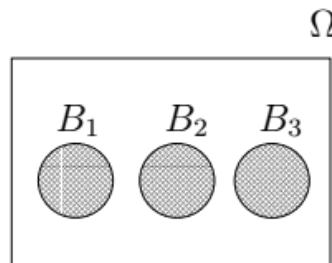
This means events A and B cannot happen together. If A happens, it excludes B from happening, and vice-versa.



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

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

Definition: Any number of events B_1, B_2, \dots, 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 $i \neq j$.



Mutually exclusive vs Statistical Independent

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

AND goes with multiplying

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

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

OR goes with adding

The Partition Theorem (Law of Total Probability)

Definition: A partition of Ω is a collection of mutually exclusive events whose union is Ω .

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

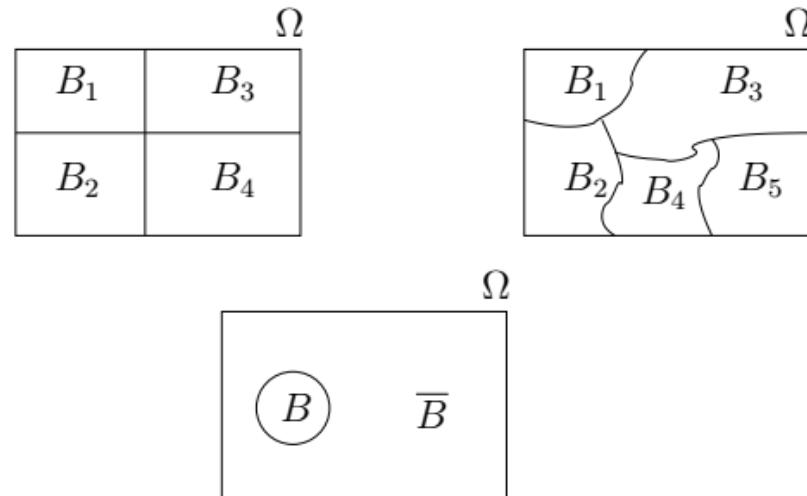
$$B_i \cap B_j = \emptyset \text{ for all } i, j \text{ with } i \neq j,$$

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

B_1, \dots, B_k form a partition of Ω if they *have no overlap*
and collectively cover all possible outcomes.

The Partition Theorem (Law of Total Probability)

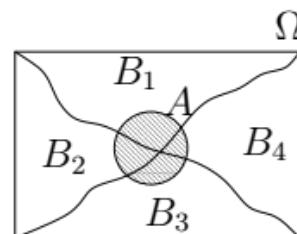
Examples:



Partitioning an event A

Any set A can be partitioned: it doesn't have to be Ω .

In particular, if B_1, \dots, B_k form a partition of Ω , then $(A \cap B_1), \dots, (A \cap B_k)$ form a partition of A .



The Partition Theorem (Law of Total Probability)

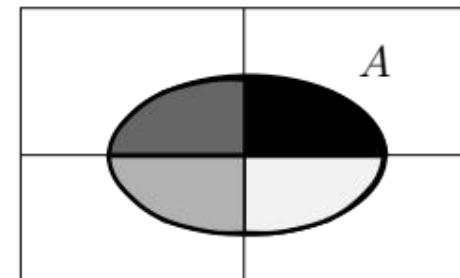
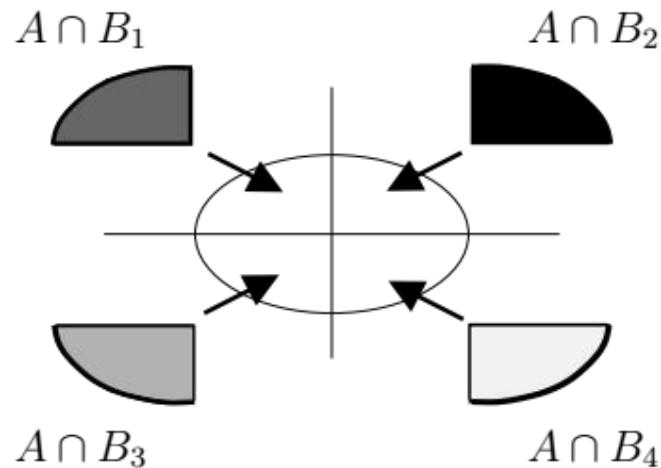
Let B_1, \dots, B_m form a partition of Ω . Then for any event A ,

$$\mathbb{P}(A) = \sum_{i=1}^m \mathbb{P}(A \cap B_i) = \sum_{i=1}^m \mathbb{P}(A | B_i) \mathbb{P}(B_i)$$

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

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.”



$$\mathbb{P}(A) = \mathbb{P}(A \cap B_1) + \mathbb{P}(A \cap B_2) + \mathbb{P}(A \cap B_3) + \mathbb{P}(A \cap B_4).$$

So:

$$\mathbb{P}(A) = \mathbb{P}(A | B_1)\mathbb{P}(B_1) + \mathbb{P}(A | B_2)\mathbb{P}(B_2) + \mathbb{P}(A | B_3)\mathbb{P}(B_3) + \mathbb{P}(A | B_4)\mathbb{P}(B_4).$$

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.

Yes: they cover all possible journeys (probabilities sum to 1), and there is no overlap in the events by definition.

Examples of conditional probability and partitions

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.

Let $C = \{ \text{crowded} \}$, $N = \{ \text{noisy} \}$.

C and N do NOT form a partition of Ω . It is possible for the bus to be noisy when it is crowded, so there must be some overlap between C and N .

- (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 .

$$\begin{aligned}\mathbb{P}(C | T) &= 0.5; & \mathbb{P}(C | L) &= 0.7. \\ \mathbb{P}(N | C) &= 0.8; & \mathbb{P}(N | \bar{C}) &= 0.4.\end{aligned}$$

Examples of conditional probability and partitions

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

$$\begin{aligned}\mathbb{P}(C) &= \mathbb{P}(C | T)\mathbb{P}(T) + \mathbb{P}(C | L)\mathbb{P}(L) && (\textit{Partition Theorem}) \\ &= 0.5 \times 0.6 + 0.7 \times 0.4 \\ &= 0.58.\end{aligned}$$

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

$$\begin{aligned}\mathbb{P}(N) &= \mathbb{P}(N | C)\mathbb{P}(C) + \mathbb{P}(N | \bar{C})\mathbb{P}(\bar{C}) && (\textit{Partition Theorem}) \\ &= 0.8 \times 0.58 + 0.4 \times (1 - 0.58) \\ &= 0.632.\end{aligned}$$

Class task conditional probability and partitions

Mr Tambourine runs a musical coffee shop called Cafe Swan Cake, that attracts musical customers from all over the world. The sample space is $\Omega = \{\text{customers}\}$.

Define events $I = \{\text{customer is Irish}\}$; $B = \{\text{customer plays the Banjo}\}$.

- (a) Write down **four different sentences** that express the information $\mathbb{P}(B | I) = 0.4$. Each sentence should have a different wording or sentence structure, but each sentence should unambiguously convey that $\mathbb{P}(B | I) = 0.4$. The sentences should be written in natural English, using the words ‘Irish’ and ‘banjo’ rather than the letters I and B . (4)
- (b) Write down **two** different sentences that express the information $\mathbb{P}(B \cap I) = 0.2$. (2)
- (c) Draw a diagram, showing the sample space Ω , and two possible events B and I , that clearly conveys the two pieces of information that $\mathbb{P}(B \cap I) = 0.2$ and $\mathbb{P}(B | I) = 0.4$. Use **areas** on the diagram to convey probabilities, and label your diagram so that it clearly depicts both the required quantities and the correct areas. [Hint: rectangular regions on a diagram will convey areas better than circular regions.] (4)
- (d) Draw two further diagrams showing Ω , B , and I , where $\mathbb{P}(B \cap I) = 0.2$ and $\mathbb{P}(B | I) = 0.4$.
 - (i) The first diagram should show B and I as **independent events**. Include any brief pieces of working or calculation that have helped you to create this diagram.
 - (ii) The second diagram should show B and I as **non-independent events**. (6)
- (e) Suppose that B and I are **independent** events. You want to explain to a high-school student, in words, why independence means that $\mathbb{P}(B \cap I) = \mathbb{P}(B)\mathbb{P}(I)$. Write down your **brief explanation**, in **at most two sentences**. If you wish, you may use numeric values calculated earlier in the question as part of your explanation. (4)

If you have extra time: Draw a diagram to explain why, if events I_1, \dots, I_n are mutually exclusive and $\mathbb{P}(B | I_r) = k$ (a constant) for all $r = 1, \dots, n$, then $\mathbb{P}(B | I_1 \cup I_2 \cup \dots \cup I_n) = k$ also. *This is an example of a question that is easily solved visually, once you understand visual representations of conditional probability, but is tedious to solve algebraically.*

Class task conditional probability and partitions

- (a) The key is to bring out that the 40% is selecting from the **Irish customers**. Some possible sentences:
- (i) 40% of the Irish customers play the banjo.
 - (ii) Of the Irish customers, 40% play the banjo.
 - (iii) The probability that an Irish customer plays the banjo is 0.4.
 - (iv) The probability of playing the banjo is 0.4 for the Irish customers.
 - (v) Out of the Irish customers, the proportion of banjo-players is 40%.
 - (vi) Given that a customer is Irish, the probability he or she plays the banjo is 0.4.
- (b) The key is to bring out that the 20% is selecting from **all customers**. Some possible sentences:
- (i) The probability that a customer is an Irish banjo-player is 0.2.
 - (ii) 20% of the customers are Irish and play the banjo.
 - (iii) A customer is Irish and plays the banjo with probability 0.2.

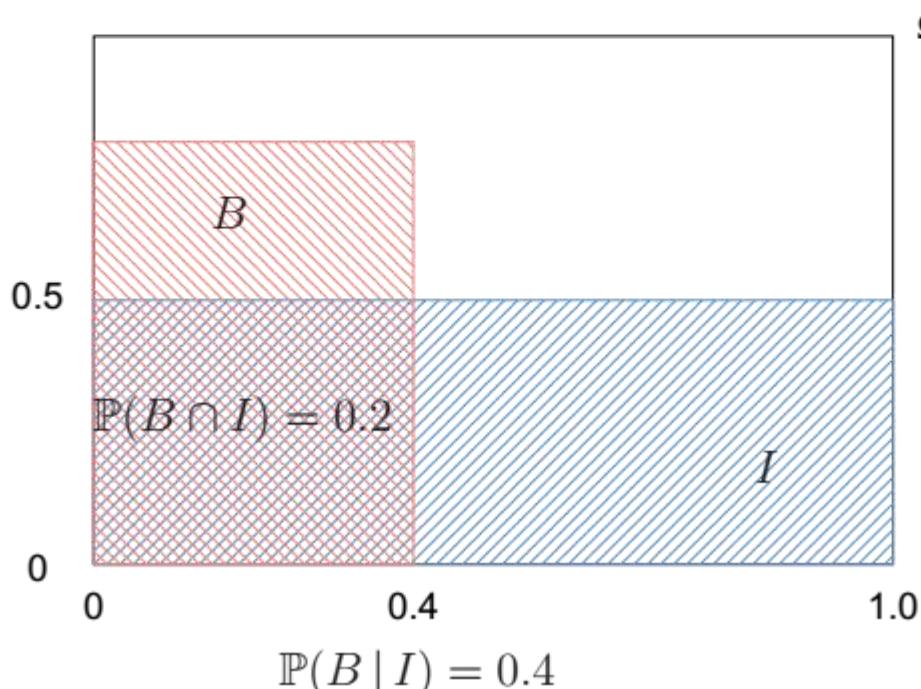
Class task conditional probability and partitions

- (c) It's easiest to convey areas precisely if you use rectangles instead of circles on the Venn diagram. The diagram below shows the intersection area as being exactly fraction 0.2 of the whole, and exactly fraction 0.4 of the blue area (event I). A scale is marked on the diagram.

Note that

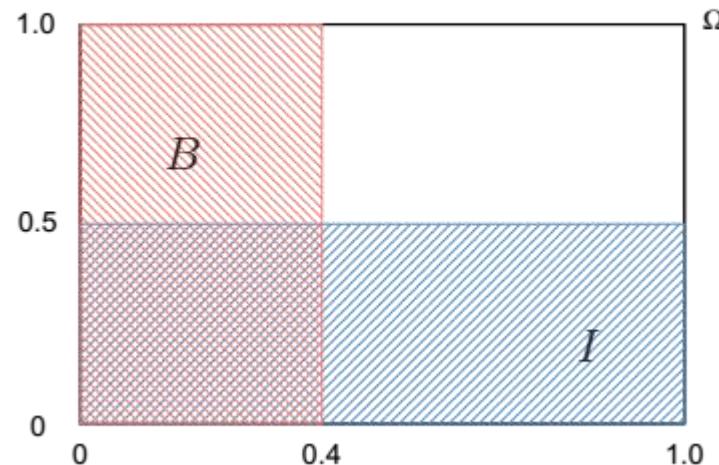
$$0.4 = \mathbb{P}(B | I) = \frac{\mathbb{P}(B \cap I)}{\mathbb{P}(I)} = \frac{0.2}{\mathbb{P}(I)} \Rightarrow \mathbb{P}(I) = 0.5.$$

However, we cannot calculate $\mathbb{P}(B)$ from the information given, so B can be any height in the diagram below.

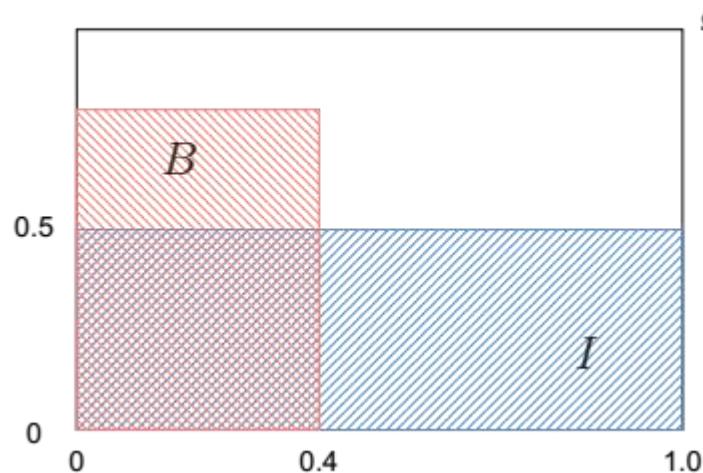


Class task conditional probability and partitions

(d)(i) If B and I are independent, then $\mathbb{P}(B) = \mathbb{P}(B | I) = 0.4$, and $\mathbb{P}(I | B) = \mathbb{P}(I) = 0.5$.



(ii) Any other diagram (like in part (c)) will show the events being non-independent:



Class task conditional probability and partitions

- (e) If B and I are independent, they occur in the same proportions regardless of whether they happen together or separately. I occurs half the time overall, so it will also occur on half the times that B happens, so $P(I \cap B) = 0.5\mathbb{P}(B) = \mathbb{P}(I)\mathbb{P}(B)$.

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 :

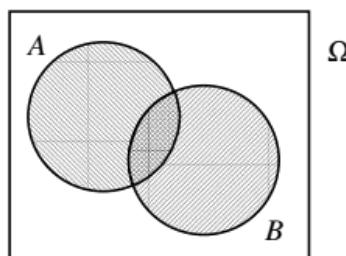
$$\boxed{\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B).}$$

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

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

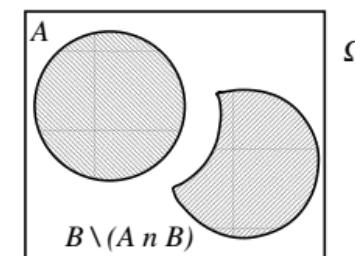
Explanation

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



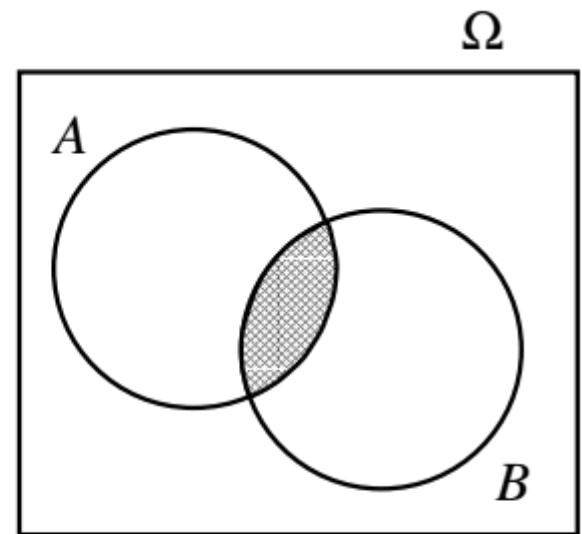
Alternatively, think of $A \cup B$ as *two disjoint sets: all of A, and the bits of B without the intersection*. So $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \left\{ \mathbb{P}(B) - \mathbb{P}(A \cap B) \right\}$.

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



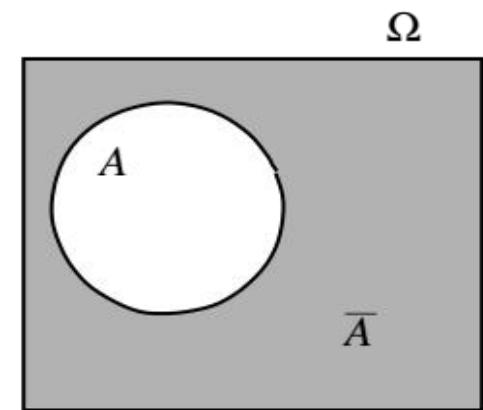
If A and B are not statistically independent, we usually use *conditional probability*: $\mathbb{P}(A \cap B) = \mathbb{P}(A | B)\mathbb{P}(B)$ for any events A and B . *It is usually easier to find a conditional probability than an intersection.*

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, example 1

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: $\Omega = \{ \text{respondents} \}$. Formulate events:

Let $C = \{ \text{has children} \}$ $\bar{C} = \{ \text{no children} \}$

$L = \{ \text{chooses large car} \}$.



Examples of basic probability calculations, example 1

Next write down all the information given:

$$\mathbb{P}(C) = 0.33$$

$$\mathbb{P}(C \cap L) = 0.13$$

$$\mathbb{P}(\bar{C} \cap L) = 0.12.$$

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

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

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

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

Examples of basic probability calculations, example 2

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.

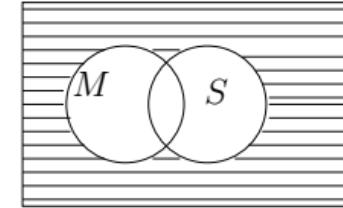
Examples of basic probability calculations, example 2

(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.

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

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.

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

Probability Reference List (Keep in mind!!!)

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

Probability Reference List (Keep in mind!!!)

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

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

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

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

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

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

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

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

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

- **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: $\mathbb{P}(A | B) = \mathbb{P}(A)$ and $\mathbb{P}(B | A) = \mathbb{P}(B)$.

Probability Reference List (Keep in mind!!!)

- Manipulating conditional probabilities:

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

- ★ if A_1 and A_2 are mutually exclusive, then

$$\mathbb{P}(A_1 \cup A_2 | B) = \mathbb{P}(A_1 | B) + \mathbb{P}(A_2 | B)$$

compare with the usual formula, $\mathbb{P}(A_1 \cup A_2) = \mathbb{P}(A_1) + \mathbb{P}(A_2)$.

- ★ if A_1, \dots, A_m partition the sample space Ω , then

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

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

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

Homework from 1 to 10

1. One ball is to be selected from a box containing red, white, blue, yellow, and green balls. If the probability that the selected ball will be red is $1/5$ and the probability that it will be white is $2/5$, what is the probability that it will be blue, yellow, or green?
2. A student selected from a class will be either a boy or a girl. If the probability that a boy will be selected is 0.3 , what is the probability that a girl will be selected?
3. Consider two events A and B such that $\Pr(A) = 1/3$ and $\Pr(B) = 1/2$. Determine the value of $\Pr(B \cap A^c)$ for each of the following conditions: **(a)** A and B are disjoint; **(b)** $A \subset B$; **(c)** $\Pr(A \cap B) = 1/8$.
4. If the probability that student A will fail a certain statistics examination is 0.5 , the probability that student B will fail the examination is 0.2 , and the probability that both student A and student B will fail the examination is 0.1 , what is the probability that at least one of these two students will fail the examination?
5. For the conditions of Exercise 4, what is the probability that neither student A nor student B will fail the examination?
6. For the conditions of Exercise 4, what is the probability that exactly one of the two students will fail the examination?
7. Consider two events A and B with $\Pr(A) = 0.4$ and $\Pr(B) = 0.7$. Determine the maximum and minimum possible values of $\Pr(A \cap B)$ and the conditions under which each of these values is attained.
8. If 50 percent of the families in a certain city subscribe to the morning newspaper, 65 percent of the families subscribe to the afternoon newspaper, and 85 percent of the families subscribe to at least one of the two newspapers, what percentage of the families subscribe to both newspapers?
9. Prove that for every two events A and B , the probability that exactly one of the two events will occur is given by the expression

$$\Pr(A) + \Pr(B) - 2 \Pr(A \cap B).$$

10. For two arbitrary events A and B , prove that

$$\Pr(A) = \Pr(A \cap B) + \Pr(A \cap B^c).$$

11. A point (x, y) is to be selected from the square S containing all points (x, y) such that $0 \leq x \leq 1$ and $0 \leq y \leq 1$. Suppose that the probability that the selected point will belong to each specified subset of S is equal to the area of that subset. Find the probability of each of the following subsets: **(a)** the subset of points such that $(x - \frac{1}{2})^2 + (y - \frac{1}{2})^2 \geq \frac{1}{4}$; **(b)** the subset of points such that $\frac{1}{2} < x + y < \frac{3}{2}$; **(c)** the subset of points such that $y \leq 1 - x^2$; **(d)** the subset of points such that $x = y$.

12. Let A_1, A_2, \dots be an arbitrary infinite sequence of events, and let B_1, B_2, \dots be another infinite sequence of events defined as follows: $B_1 = A_1$, $B_2 = A_1^c \cap A_2$, $B_3 = A_1^c \cap A_2^c \cap A_3$, $B_4 = A_1^c \cap A_2^c \cap A_3^c \cap A_4, \dots$. Prove that

$$\Pr\left(\bigcup_{i=1}^n A_i\right) = \sum_{i=1}^n \Pr(B_i) \text{ for } n = 1, 2, \dots,$$

and that

$$\Pr\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \Pr(B_i).$$

13. Prove Theorem 1.5.8. *Hint:* Use Exercise 12.

14. Consider, once again, the four blood types A, B, AB, and O described in Exercise 8 in Sec. 1.4 together with the two antigens anti-A and anti-B. Suppose that, for a given person, the probability of type O blood is 0.5 , the probability of type A blood is 0.34 , and the probability of type B blood is 0.12 .

- a. Find the probability that each of the antigens will react with this person's blood.
- b. Find the probability that both antigens will react with this person's blood.

Homework

- 1.** Consider the unfavorable game in Example 2.4.2. This time, suppose that the initial fortune of gambler A is i dollars with $i \leq 98$. Suppose that the initial fortune of gambler B is $100 - i$ dollars. Show that the probability is greater than $1/2$ that gambler A losses i dollars before winning $100 - i$ dollars.

- 2.** Consider the following three different possible conditions in the gambler's ruin problem:
 - a.** The initial fortune of gambler A is two dollars, and the initial fortune of gambler B is one dollar.
 - b.** The initial fortune of gambler A is 20 dollars, and the initial fortune of gambler B is 10 dollars.
 - c.** The initial fortune of gambler A is 200 dollars, and the initial fortune of gambler B is 100 dollars.

Suppose that $p = 1/2$. For which of these three conditions is there the greatest probability that gambler A will win the initial fortune of gambler B before he loses his own initial fortune?

- 3.** Consider again the three different conditions (a), (b), and (c) given in Exercise 2, but suppose now that $p < 1/2$. For which of these three conditions is there the greatest probability that gambler A will win the initial fortune of gambler B before he loses his own initial fortune?

- 4.** Consider again the three different conditions (a), (b), and (c) given in Exercise 2, but suppose now that $p > 1/2$. For which of these three conditions is there the greatest probability that gambler A will win the initial fortune of gambler B before he loses his own initial fortune?

Homework

- (1) Early HIV tests relied on an absorbance ratio for antibodies. Given a threshold, all patients below the threshold were given a negative result (no HIV), while all above the threshold got a positive result (HIV infected). However, there is always some overlap between healthy and infected patients.

Using figures from 1987/1988:

$$\mathbb{P}(\text{patient tests positive} \mid \text{patient has HIV}) = 0.98 \quad (1)$$

$$\mathbb{P}(\text{patient tests positive} \mid \text{patient does not have HIV}) = 0.07 \quad (2)$$

These numbers make it look as if the test is reasonably accurate, both for people with and without HIV. We will use Bayes' Theorem to calculate

$$\mathbb{P}(\text{patient has HIV} \mid \text{patient tests positive}). \quad (3)$$

- (a) Write down events H and P for patients with HIV and patients testing positive. Express (1) and (2) in terms of these events and their complements, and write down the probability that we want to calculate (expression (3)).

Homework

- (b) In NZ in 1987, roughly 4650 people out of a population of 3.2 million were infected with HIV. Find $\mathbb{P}(H)$, and hence deduce $\mathbb{P}(\overline{H})$.
- (c) Use the Partition Theorem to find $\mathbb{P}(P)$. [Hint: use H and \overline{H} as your partition.]
- (d) Use Bayes' Theorem to find $\mathbb{P}(\text{patient has HIV} \mid \text{patient tests positive})$. Comment on the size of your result, and what it means for people who obtained a positive test for HIV in 1987/1988.

Note: the reason for the small size of the answer is that HIV is very rare in the overall population. Even when we restrict to the set of people with positive test results, this set is still dominated largely by people without HIV.