

## Chapter 2

# Introduction to Probability

# Probability: big idea

---

In statistical studies, we collect data to learn about real-world phenomena.

However, we recognize that as we only collect a finite sample of data, this cannot tell us everything about the system being studied.

Therefore, we need to develop tools to allow us to talk about the “uncertainty” associated with this lack of complete information.

# Probability: big idea

---

Also, we recognize that when we collect data, they are subject to variability

- measurement variability
- variability due to small changes in experimental conditions.

We want to describe the nature of this variability.

# Probability: big idea

---

## Example: Repeated coin tosses

Five sequences of 10 tosses of a fair coin

|              | 1   | 2   | 3   | 4   | 5   | 6   | 7   | 8   | 9   | 10  |
|--------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| + Sequence 1 | "T" | "T" | "T" | "T" | "H" | "H" | "H" | "H" | "H" | "H" |
| + Sequence 2 | "H" | "T" | "T" | "T" | "H" | "T" | "H" | "H" | "H" | "H" |
| + Sequence 3 | "T" | "T" | "T" | "T" | "H" | "H" | "H" | "T" | "T" | "T" |
| + Sequence 4 | "T" | "H" | "T" | "H" | "T" | "H" | "H" | "T" | "T" | "T" |
| + Sequence 5 | "T" | "H" | "T" | "T" | "H" | "T" | "T" | "T" | "T" | "H" |

The number of Heads recorded is different for each sequence; this 'outcome' is variable.

# Uncertainty

---

**Uncertainty** corresponds to a lack of perfect knowledge; it could be a consequence of

- incomplete observation of a system;
- unpredictable variation;
- simple lack of knowledge of the “state of nature”, such as
  - ▶ the **current** state that is imperfectly observed, or
  - ▶ the **future** state, that is, the result of a study yet to be carried out.

# Uncertainty

---

## Example: Uncertain states of nature

- the outcome of a single coin toss;
- the millionth digit of  $\pi$ ;

3.1415926535897932384626433832795028841971....

- the height of the building I am in;
- the number of people on campus now;
- the temperature at noon tomorrow.

# Uncertainty

---

## Note

The millionth digit of  $\pi$  is a fixed number: however, unless we know its value, we are still in a state of uncertainty when asked to assess it, as we have a lack of perfect knowledge.

# Uncertainty

---

## Example: Coin tossing

If I take a coin and decide to carry out an experiment where I toss the coin once and see which face is turned upward when the coin comes to rest, then I can assess that there are two possible results of the toss:



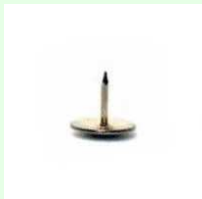
The outcome of the toss is uncertain **before** I toss the coin, and **after** I toss the coin **until I see the result**.

# Uncertainty

---

## Example: Thumbtack tossing

If I take a thumbtack and decide to carry out an experiment where I toss the thumbtack once, then I can assess that there are two possible results of the toss:



The outcome of the toss is uncertain **before** I toss the thumbtack, and **after** I toss the thumbtack **until I see the result**.

# What is probability ?

---

By **probability**, we generally mean a numerical assessment of  
*the chance of a particular event occurring, given a particular set of circumstances.*

- *the chance of a particular event occurring ...*
- *... given a particular set of circumstances.*

# Constructing the mathematical framework

---

We will use the term **experiment** to mean some situation that generates multiple possible outcomes

- we carry out an operation, make a measurement etc.

For a given experiment, we aim to

- identify the possible outcomes;
- assign numerical values to collections of possible values to represent the chance that they will coincide with the actual outcome;
- lay out the rules for how the numerical values can be manipulated.

# Basic concepts in set theory

---

A set  $S$  is a collection of individual **elements**, such as  $s$ .

We write

$$s \in S$$

to mean “ $s$  is one of the elements of  $S$ ”.

- $S = \{0, 1\}$ ;  $0 \in S, 1 \in S$ .
- $S = \{\text{“dog”}, \text{“cat”}, \text{“mouse”}, \text{“rat”}\}$ ;  $\text{“rat”} \in S$
- $S = [0, 1]$  (the **closed** interval from zero to one); then

$$0.3428 \in S$$

# Basic concepts in set theory

---

$S$  may be

- **finite** if it contains a finite number of elements;
  - ▶ e.g.  $\{1, 2, 3, 4, 5\}$ ,  $\{\text{Science, Arts, Engineering}\}$ .
- **countable** if it contains an infinite number of elements that can be listed;
  - ▶ e.g.  $\{0, 1, 2, 3, 4, 5, \dots\}$ .
- **uncountable** if it contains an infinite number of elements that cannot be listed;
  - ▶ e.g.  $\{\text{All values on a continuum between 0 and 100}\}$ .

## Basic concepts in set theory

---

A is a **subset** of  $S$  if it contains some, all, or none of the elements of  $S$ . We write

some:  $A \subset S$       **some or all**:  $A \subseteq S$ .

- $S = \{0, 1\}$ ;  $A = \{0\}$ , or  $A = \{1\}$ , or  $A = \{0, 1\}$ .
- $S = \{\text{"dog"}, \text{"cat"}, \text{"mouse"}, \text{"rat"}\}$ ;  $A = \{\text{"cat"}, \text{"mouse"}\}$
- $S = [0, 1]$ ;  $A = (0.25, 0.3)$ .

The **empty** set,  $\emptyset$ , is a subset that contains no elements.

# Set operations

---

To manipulate sets, we use three basic operations:

- intersection  $\cap$
- union  $\cup$
- complement  $A^c$

# Set operations

---

- **intersection**: the **intersection** of two sets  $A$  and  $B$  is the collection of elements that are elements of **both**  $A$  and  $B$ .

$$s \in A \cap B \quad \text{means} \quad s \in A \text{ and } s \in B.$$

We have for any  $A, B \subseteq S$  that

$$A \cap \emptyset = \emptyset$$

$$A \cap S = A$$

$$A \cap B \subseteq A$$

$$A \cap B \subseteq B.$$

# Set operations

---

- **union**: the **union** of two sets  $A$  and  $B$  is the set of **distinct** elements that are either in  $A$ , or in  $B$ , or in both  $A$  and  $B$ .

$$s \in A \cup B \quad \text{means} \quad s \in A, \text{ or } s \in B, \text{ or } s \in A \cap B.$$

We have for any  $A, B \subseteq S$  that

$$A \cup \emptyset = A$$

$$A \cup S = S$$

$$A \subseteq A \cup B$$

$$B \subseteq A \cup B$$

# Set operations

---

- **complement**: the **complement** of set  $A$  is the collection of elements of  $S$  that are **not** elements of  $A$ .

$$s \in A^c \quad \text{means} \quad s \in S \text{ but } s \notin A.$$

We have that

$$A \cap A^c = \emptyset \quad \text{and} \quad A \cup A^c = S.$$

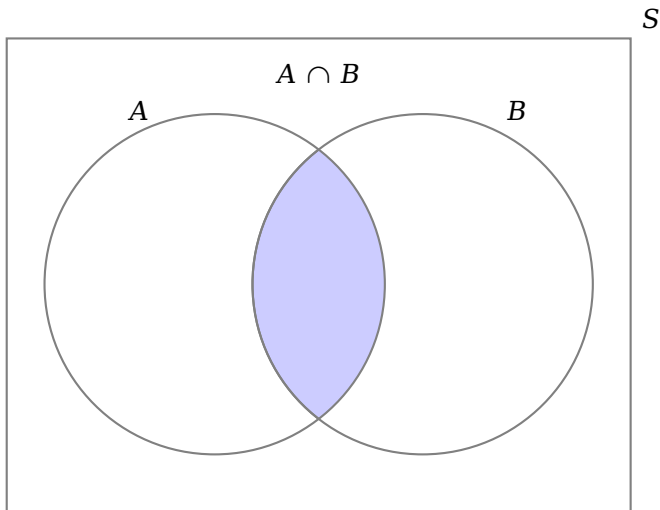
We have for any  $A$  that

$$(A^c)^c = A.$$

# Set operations: in pictures

---

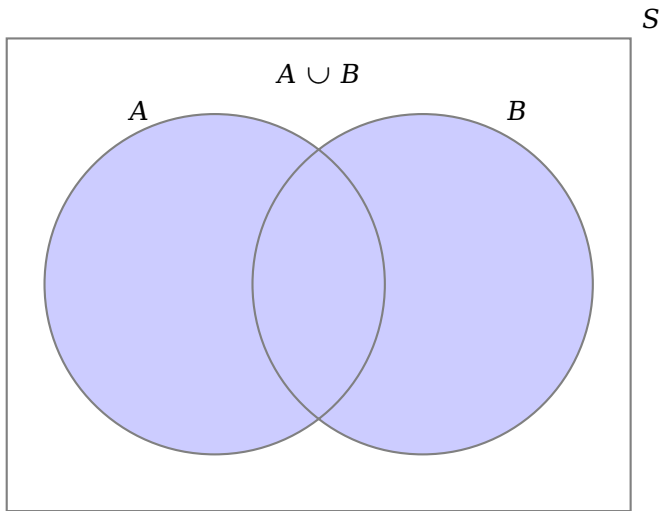
Intersection: blue shaded area is  $A \cap B$ .



## Set operations: in pictures

---

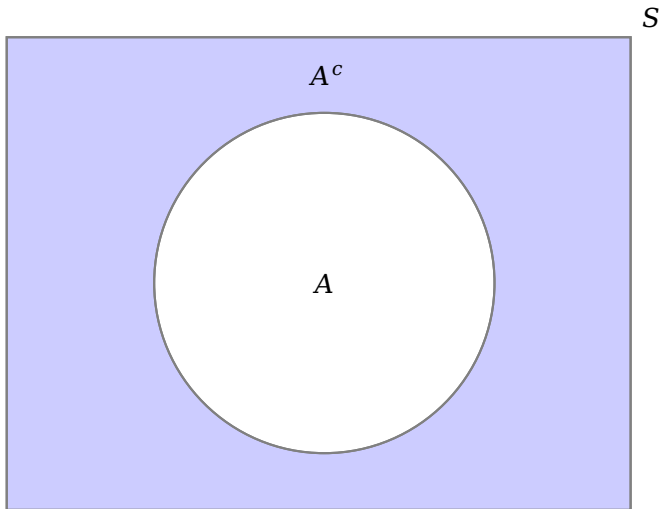
Union: blue shaded area is  $A \cup B$ .



## Set operations: in pictures

---

Complement: blue shaded area is  $A^c$



# Example

---

## Example: Finite set

- $S = \{1, 2, 3, \dots, 9, 10\}$
- $A = \{2, 4, 6\}$
- $B = \{1, 2, 5, 7, 9\}$

Then

- $A \cap B = \{2\}$
- $A \cup B = \{1, 2, 4, 5, 6, 7, 9\}$
- $A^c = \{1, 3, 5, 7, 8, 9, 10\}$

# Extensions

---

- Intersections:

$$(A \cap B) \cap C = A \cap B \cap C$$

- Unions:

$$(A \cup B) \cup C = A \cup B \cup C$$

- Intersections and Unions:

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$$

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$$

These can be extended to any number of events.

# Partitions

---

We call  $A_1, A_2, \dots, A_K$  a **partition** of  $S$  if these sets are

(i) pairwise **disjoint** (or **mutually exclusive**):

$$A_j \cap A_k = \emptyset \quad \text{for all } j \neq k$$

(ii) **exhaustive**:

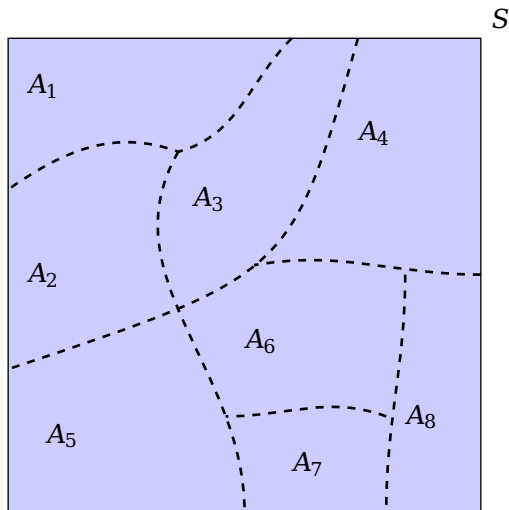
$$\bigcup_{k=1}^K A_k = S.$$

(i.e. the  $A$ s cover the **whole** of  $S$ , but **do not overlap**).

Every  $s \in S$  is an element of **precisely one** of the  $A_k$ .

# Partitions

---



## Some simple results

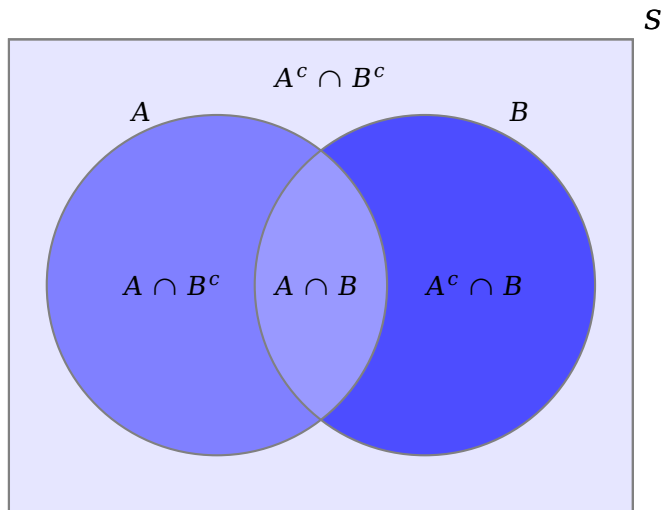
---

For any two subsets  $A$  and  $B$  of  $S$ , we have the partition of  $S$  via the four sets

$$A \cap B \quad A \cap B^c \quad A^c \cap B \quad A^c \cap B^c$$

## Some simple results

---



## Some simple results

---

This picture tells us that

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

- $(A \cup B)$ : three darker blue regions
- $A^c \cap B^c$ : lightest blue region

Similarly

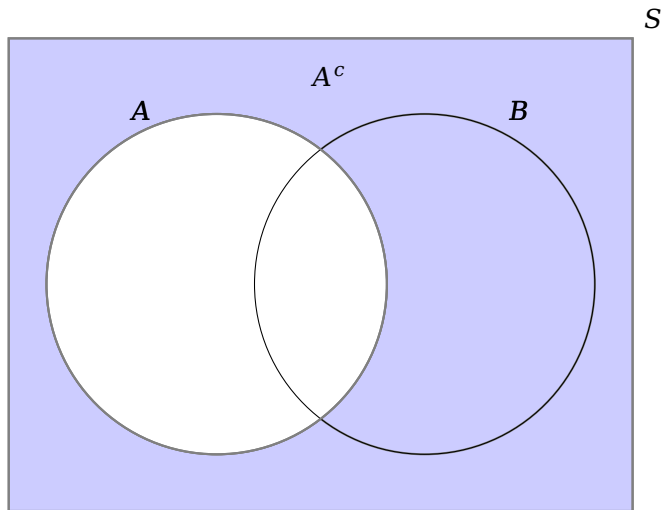
$$A^c \cup B^c = (A \cap B)^c$$

so that

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

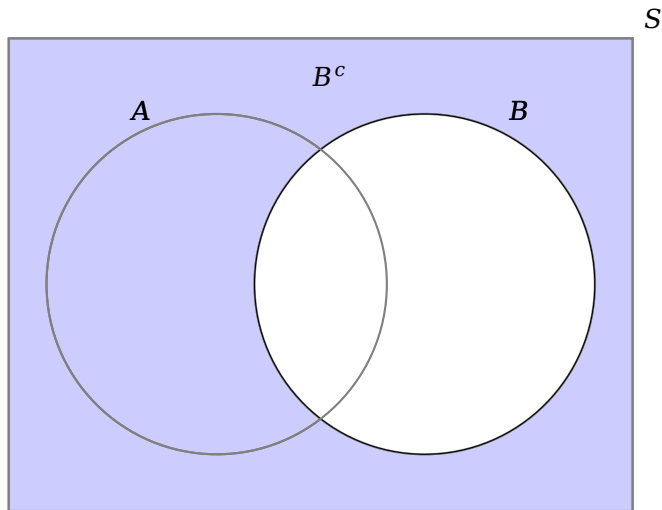
## Some simple results

---



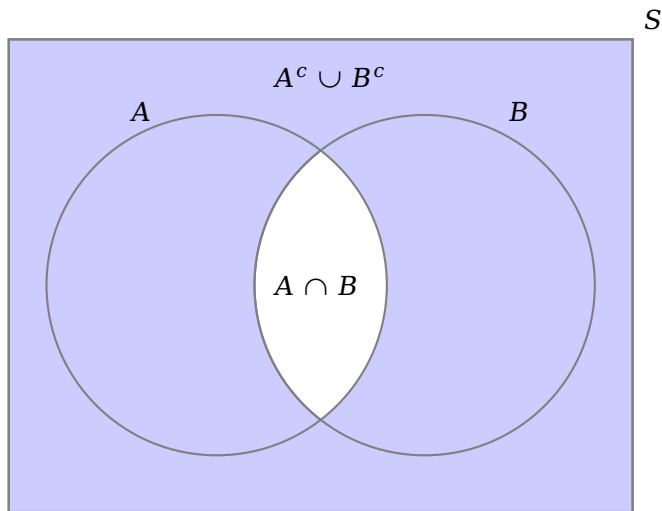
## Some simple results

---



## Some simple results

---



# Some simple results

---

These two results

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

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

are sometimes known as **de Morgan's Laws**.

# Sample spaces and events

---

Recall the earlier informal definition:

*By probability, we generally mean the chance of a particular **event** occurring, given a **particular set of circumstances**. The probability of an event is generally expressed as a quantitative measurement.*

We need to carefully define what an '**event**' is, and what constitutes a '**particular set of circumstances**'.

# Sample spaces and events

---

We will consider the general setting of an **experiment**:

- this can be interpreted as any setting in which an uncertain consequence is to arise;
- could involve observing an outcome, taking a measurement etc.

# Constructing the framework

---

## 1. Consider the possible outcomes of the experiment:

Identify all the outcomes that can arise, and denote the corresponding set by  $S$ .

- $S = \{0, 1\}$ ;
- $S = \{\text{"head"}, \text{"tail"}\} \equiv \{H, T\}$ ;
- $S = \{\text{"cured"}, \text{"not cured"}\}$ ;
- $S = \{\text{"Arts"}, \text{"Engineering"}, \text{"Medicine"}, \text{"Science"}\}$ ;
- $S = \{1, 2, 3, 4, 5, 6\}$ ;
- $S = \mathbb{R}^+$ .

The set  $S$  is termed the **sample space** of the experiment. The individual elements of  $S$  are termed **sample points** (or **sample outcomes**).

## Constructing the framework

---

2. **Events:** An **event**  $A$  is a collection of sample outcomes. That is,  $A$  is a subset of  $S$ ,  $A \subseteq S$ .

For example

- $A = \{0\}$ ;
- $A = \{\text{"tail"}\} \equiv \{T\}$ ;
- $A = \{\text{"cured"}\}$ ;
- $A = \{\text{"Arts"}, \text{"Engineering"}\}$ ;
- $A = \{1, 3, 5\}$ ;
- $A = [2, 3)$ .

The individual sample outcomes are termed **simple** (or **elementary**) events, and may be denoted

$$E_1, E_2, \dots, E_K, \dots$$

## Constructing the framework

---

3. **Terminology:** We say that event  $A$  **occurs** if the actual outcome,  $s$ , is an element of  $A$ .

For two events  $A$  and  $B$

- $A \cap B$  occurs if and only if  $A$  occurs **and**  $B$  occurs, that is

$$s \in A \cap B.$$

- $A \cup B$  occurs if  $A$  occurs **or** if  $B$  occurs, **or** if **both**  $A$  and  $B$  occur, that is

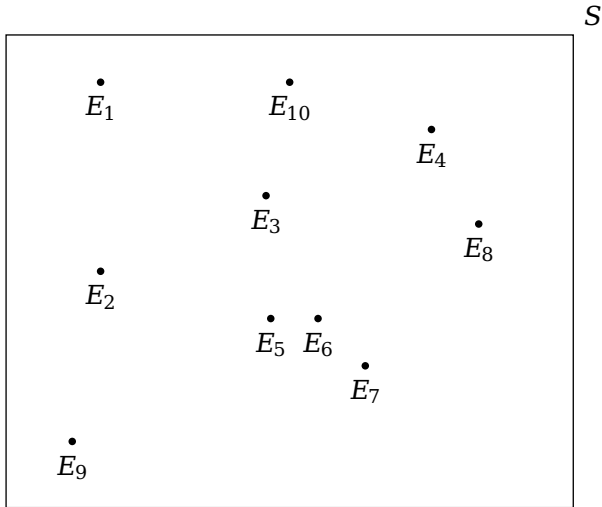
$$s \in A \cup B$$

- if  $A$  occurs, then  $A^c$  **does not** occur.

# Constructing the framework

---

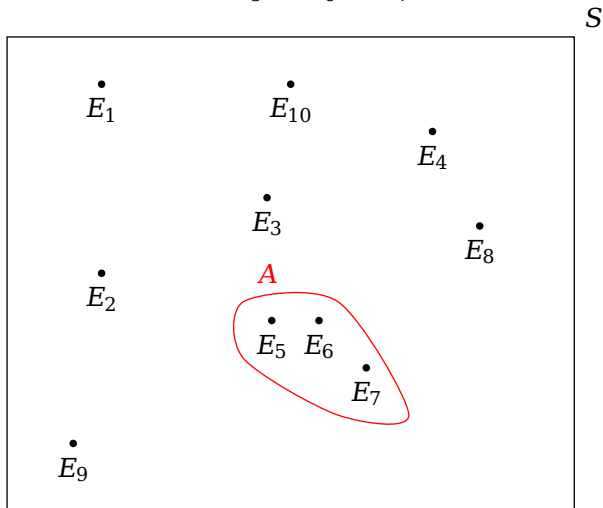
$$S = E_1 \cup E_2 \cup \cdots \cup E_{10}$$



# Constructing the framework

---

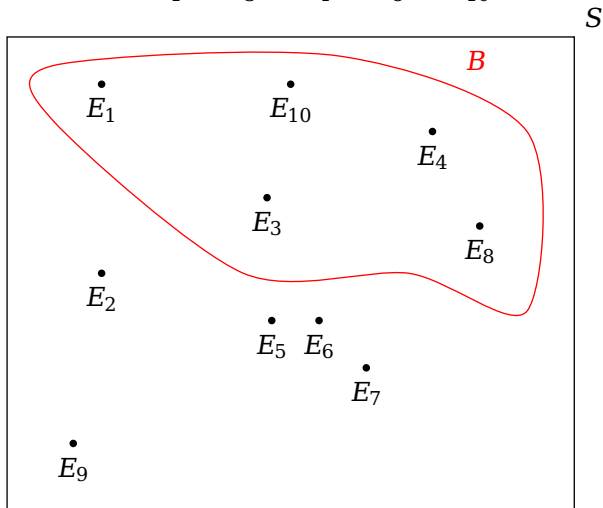
$$A = E_5 \cup E_6 \cup E_7$$



# Constructing the framework

---

$$\mathbf{B} = E_1 \cup E_3 \cup E_4 \cup E_8 \cup E_{10}$$



# Constructing the framework

---

## Note

- The event  $S$  is termed the **certain event**;
- The event  $\emptyset$  is termed the **impossible event**.

# Mathematical definition of probability

---

For event  $A$ ,  $P$  is the function that assigns

$$P(A) = p$$

where  $p$  is a numerical value.

However, this definition is too general:

- what properties does  $P(\cdot)$  have ?
- how does  $P(\cdot)$  assign numerical values; that is, how do we compute

$$P(A)$$

for a given event  $A$  in sample space  $S$  ?

# The rules of probability

---

Suppose  $S$  is a sample space for an experiment,  $A$  and  $B$  are events (i.e. subsets of  $S$ ).

Then  $P(A)$  obeys the following rules:

(I)  $P(A) \geq 0$ .

(II)  $P(S) = 1$ .

(III) If  $A \cap B = \emptyset$ , then

$$P(A \cup B) = P(A) + P(B).$$

# The rules of probability

---

Informally,  $P$  measures the 'content' of  $A$ . The rules say

- (I) The content cannot be negative.
- (II) The total content of  $S$  is standardized to be 1.
- (III) If  $A$  and  $B$  do not overlap, then their total content is the sum of the their individual contents.

# The rules of probability

---

Rule (III) extends to **any** number of events:

If  $A_1, A_2, \dots, A_K$  are a collection of events such that

$$A_j \cap A_k = \emptyset \quad \text{for all } j \neq k$$

then

$$P(A_1 \cup A_2 \cup \dots \cup A_K) = P(A_1) + P(A_2) + \dots + P(A_K).$$

# The rules of probability

---

## Example: Partitions

If  $A_1, A_2, \dots, A_K$  form a partition of  $S$ , and

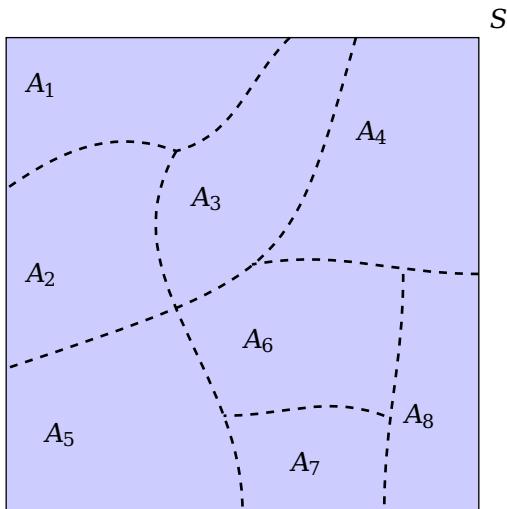
$$P(A_k) = p_k \quad k = 2, \dots, K$$

say, then

$$\sum_{k=1}^K P(A_k) = \sum_{k=1}^K p_k = 1.$$

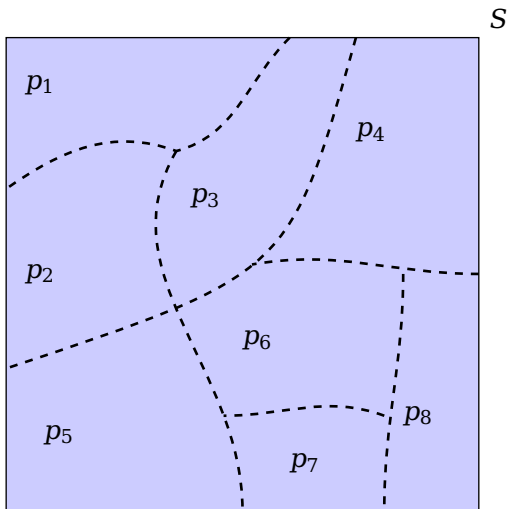
# The rules of probability

---



# The rules of probability

---



# The rules of probability

---

Some immediate consequences of the rules:

- (i) For any  $A$ ,  $P(A^c) = 1 - P(A)$ .
- (ii)  $P(\emptyset) = 0$ .
- (iii) For any  $A$ ,  $P(A) \leq 1$ .
- (iv) For two events  $A$  and  $B$ , if  $A \subseteq B$ , then

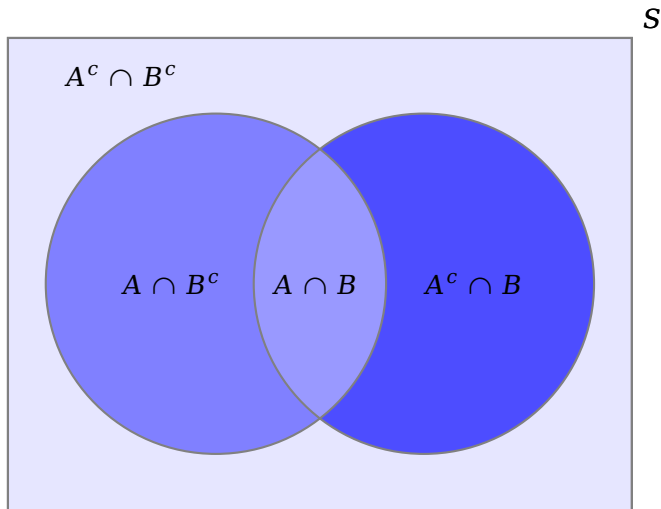
$$P(A) \leq P(B)$$

- (v) General Addition Rule: For two events  $A$  and  $B$ ,

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

# The rules of probability

---



# The rules of probability

---

We can see from the diagram that if we simply compute

$$P(A) + P(B)$$

we are **double counting** the intersection, so we need to subtract  $P(A \cap B)$  to obtain that

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

We may deduce from this that

$$P(A \cup B) \leq P(A) + P(B)$$

as  $P(A \cap B) \geq 0$ .

# The rules of probability

---

For events  $A_1, A_2, \dots, A_K$ , we can construct a formula for

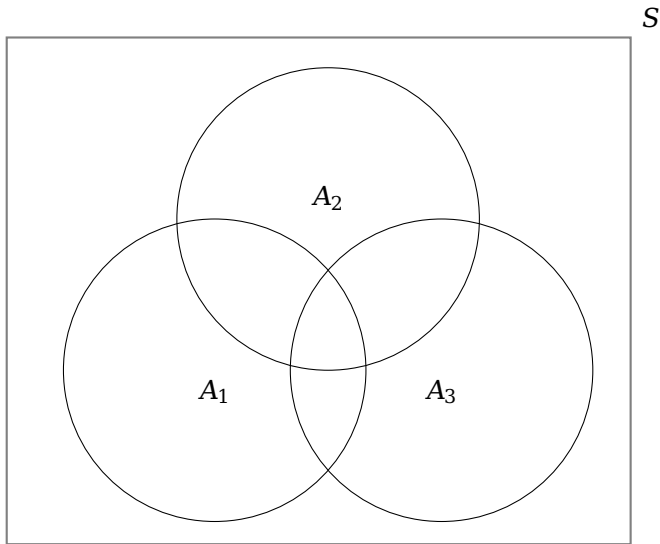
$$P\left(\bigcup_{k=1}^K A_k\right)$$

For example, with  $K = 3$ , we have that

$$\begin{aligned} P(A_1 \cup A_2 \cup A_3) \\ &= P(A_1) + P(A_2) + P(A_3) \\ &\quad - P(A_1 \cap A_2) - P(A_1 \cap A_3) - P(A_2 \cap A_3) \\ &\quad + P(A_1 \cap A_2 \cap A_3). \end{aligned}$$

# The rules of probability

---



# The rules of probability

---

A more straightforward general result is **Boole's Inequality**

$$P \left( \bigcup_{k=1}^K A_k \right) \leq \sum_{k=1}^K P(A_k) .$$

# Specifying probabilities

---

We have defined how the probability function must behave **mathematically**.

We now consider **two** ways via which we could specify the **numerical** probability of an event.

## Equally likely sample outcomes

---

Suppose that  $S$  is finite, with  $N$  sample outcomes in total that are considered to be **equally likely** to occur.

Then for the elementary events  $E_1, E_2, \dots, E_N$ , we have

$$P(E_i) = \frac{1}{N} \quad i = 1, \dots, N.$$

For  $A \subseteq S$

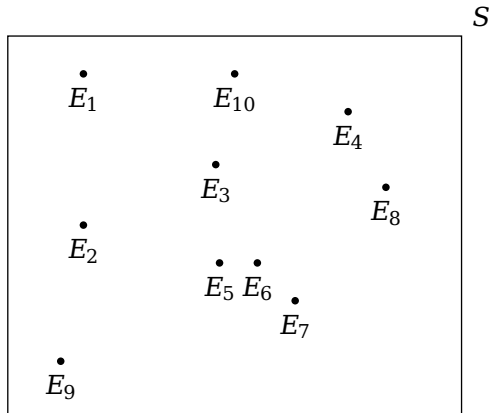
$$P(A) = \frac{\text{Number of sample outcomes in } A}{\text{Number of sample outcomes in } S} = \frac{n}{N}$$

say.

# Equally likely sample outcomes

---

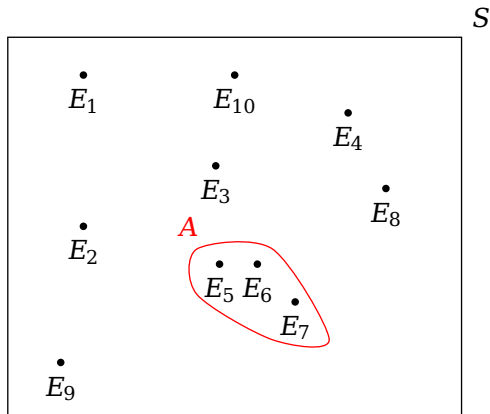
Suppose we have  $N = 10$ .



# Equally likely sample outcomes

---

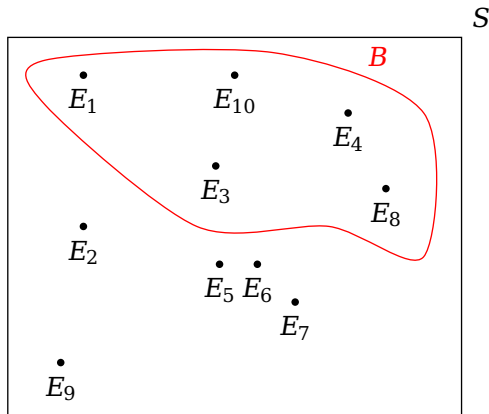
$$P(\textcolor{red}{A}) = 3/10.$$



# Equally likely sample outcomes

---

$$P(\textcolor{red}{B}) = 5/10.$$



## Equally likely sample outcomes

---

We have “equally likely outcomes” where we can appeal to some form of **symmetry** argument.

(a) **Coins:** For a single coin flip, the sample space is

$$S = \{\text{Head}, \text{Tail}\}$$

with elementary events  $E_1 = \{\text{Head}\}$  and  $E_2 = \{\text{Tail}\}$ , and it is usual to assume that

$$P(E_1) = P(E_2) = \frac{1}{2}.$$

## Equally likely sample outcomes

---

(b) **Dice:** For six-sided die, the sample space is

$$S = \{1, 2, 3, 4, 5, 6\}$$

with elementary events  $E_i = \{i\}$  for  $i = 1, 2, 3, 4, 5, 6$ , and it is usual to assume that

$$P(E_i) = \frac{1}{6}.$$

If  $A$  is the event that the outcome is an **even** number,

$$A = E_2 \cup E_4 \cup E_6$$

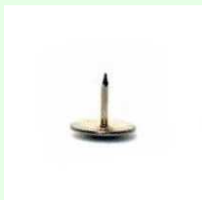
and

$$P(A) = \frac{3}{6} = \frac{1}{2}.$$

# Relative frequency

---

## Example: Thumbtack tossing



- $S = \{\text{Up}, \text{Down}\}.$
- $A = \{\text{Up}\}.$

What is  $P(A)$  ?

## Relative frequency

---

We may define  $P(A)$  by considering the **relative frequency** with which event  $A$  occurs in a long sequence of repeated identical experiments.

- Consider a finite sequence of  $N$  repeat experiments.
- Let  $n$  be the number of times (out of  $N$ ) that  $A$  occurs.
- Define

$$P(A) = \lim_{N \rightarrow \infty} \frac{n}{N}.$$

This is the **frequentist** definition of probability;

# Relative frequency

---

The frequentist definition covers the thumbtack example:



Is it always possible to consider an infinite sequence of repeats ?

## Rules for counting outcomes

---

A sequence of  $k$  operations, in which operation  $i$  can result in  $n_i$  possible outcomes, can result in

$$n_1 \times n_2 \times \cdots \times n_k$$

possible sequences of outcomes.

# Rules for counting outcomes

---

## Example: Two dice

Two dice are rolled: there are  $6 \times 6 = 36$  possible outcomes.

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

## Rules for counting outcomes

---

### Example:

We are to pick  $k = 5$  numbers sequentially from the set

$$\{1, 2, 3, \dots, 100\},$$

where every number is available on every pick. The number of ways of doing that is

$$100 \times 100 \times 100 \times 100 \times 100 = 100^5.$$

## Rules for counting outcomes

---

### Example:

We are to pick  $k = 11$  players sequentially from a squad of 25 players. The number of ways of doing that is

$$25 \times 24 \times 23 \times \cdots \times 15.$$

as players cannot be picked twice.

## Rules for counting outcomes

---

When selecting repeatedly from a finite set  $\{1, 2, \dots, N\}$  we may select

- **with** replacement, that is,
  - ▶ each successive selection can be one of the original set, irrespective of previous selections,or
- **without** replacement, that is,
  - ▶ the set is **depleted** by each successive selection,

# Permutations

---

An ordered arrangement of  $r$  objects is called a **permutation**.

The number of ways of ordering  $r$  objects selected, without replacement, from  $n$  is denoted  $P_r^n$ , and we have that

$$P_r^n = n \times (n-1) \times (n-2) \times \cdots (n-r+2) \times (n-r+1) = \frac{n!}{(n-r)!}$$

where

$$n! = n \times (n-1) \times \cdots 3 \times 2 \times 1.$$

# Combinations

---

The number of **combinations** of  $n$  objects taken  $r$  at a time is the number of subsets of size  $r$  that can be formed.

This number is denoted

$$C_r^n = \binom{n}{r}$$

where

$$\binom{n}{r} = \frac{n!}{r!(n-r)!}$$

i.e. we “choose  $r$  from  $n$ ”.

# Combinations

---

The number of possible selections of  $r$  objects (in sequence, without replacement) is

$$n \times (n - 1) \times \dots \times (n - r + 1) = \frac{n!}{(n - r)!} = P_r^n$$

leaving  $(n - r)$  objects unselected.

If the order of selected objects is not important in identifying a combination, we must have that

$$P_r^n = r! \times C_r^n.$$

as there are  $r!$  equivalent combinations that yield the same permutation.

# Binomial coefficients

---

## Note

Binomial theorem:

$$(a + b)^n = \sum_{j=0}^n \binom{n}{j} a^j b^{n-j}$$

where

$$\binom{n}{j} = \frac{n!}{j!(n-j)!}$$

# Binary sequences

---

## Binary sequences

101000100101010

arise in many probability settings; if we take '1' to indicate inclusion and '0' to indicate exclusion, then we can identify combinations with binary sequences.

- the number of binary sequences of length  $n$  containing  $r$  1s is

$$\binom{n}{r}$$

# Binary sequences

---

The above results are used to compute probabilities in the case of equally likely outcomes.

- $S$ : complete list of possible sequences of selections.
- $A$ : sequences having property of interest.
- We have

$$P(A) = \frac{\text{Number of elements in } A}{\text{Number of elements in } S} = \frac{n_A}{n_S}$$

# Conditional probability

---

In the case of ‘ordinary’ probability, the outcome **must** be an element in the set  $S$ .

- the only ‘certain’ knowledge we have is that the outcome,  $s$ , is in  $S$ .
- we can then assess  $P(A)$  for event  $A$ .

# Conditional probability

---

Suppose we know that, in fact,

$$s \in B \subseteq S$$

for some  $B$ ; that is, we **know** that event  $B$  occurs.

In light of the information that  $B$  occurs, what do we **now** think about the probability that  $A$  occurs also ?

# Conditional probability

---

With the knowledge that event  $B$  occurs, we restrict the parts of the sample space that should be considered;

- we are **certain** that the sample outcome must lie in  $B$ ,
- we can **forget** about those outcomes that are not in  $B$ .

## Conditional probability

---

For two events  $A$  and  $B$ , the **conditional probability** of  $A$  **given**  $B$  is denoted  $P(A|B)$ , and is defined by

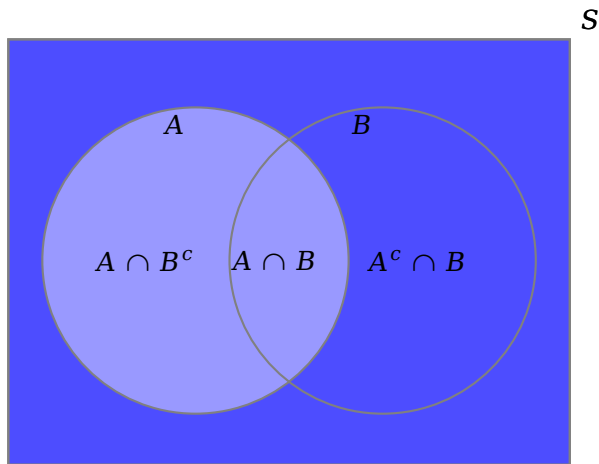
$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Note that we consider this **only** in cases where  $P(B) > 0$ .

# Conditional probability

---

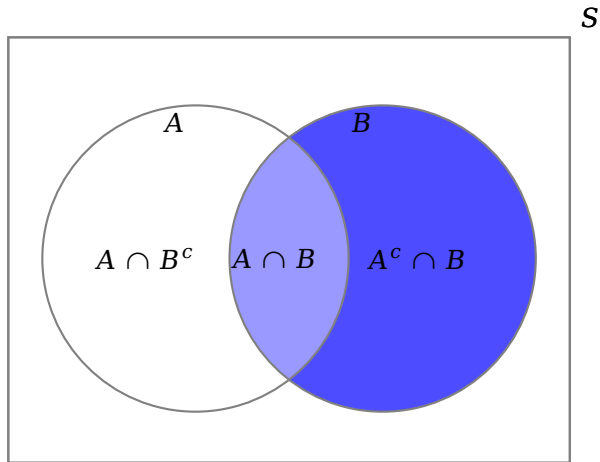
Ordinary Probability:  $P(A)$ .



# Conditional probability

---

Conditional Probability:  $P(A|B) = P(A \cap B)/P(B)$ .



# Conditional probability

---

## Example:

Experiment: roll a fair die, record the score.

- $S = \{1, 2, 3, 4, 5, 6\}$ , all outcomes equally likely.
- $A = \{1, 3, 5\}$  (score is odd).
- $B = \{4, 5, 6\}$  (score is more than 3).
- $A \cap B = \{5\}$ .

$$P(A) = \frac{3}{6} = \frac{1}{2} \quad P(B) = \frac{3}{6} = \frac{1}{2} \quad P(A \cap B) = \frac{1}{6}$$

so

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{1/6}{1/2} = \frac{1}{3}.$$

# Conditional probability

---

## Example:

In a class of 100 students:

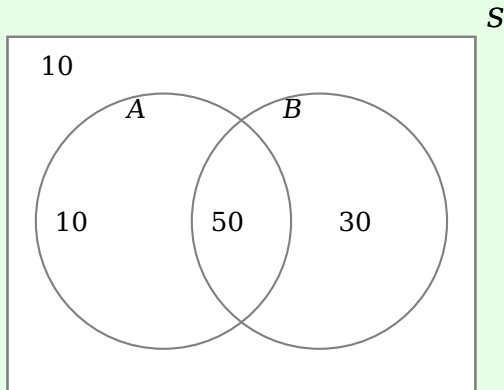
- 60 are from Science, 40 are from Arts.
- 80 are in a Major program, 20 are in another program.
- 50 of the Science students are in a Major program.

# Conditional probability

---

## Example:

$A$  – Science;  $B$  – Major program.



# Conditional probability

---

## Example:

A student is selected from the class, with all students equally likely to be selected.

- What is the probability that the selected student is from Science,  $P(A)$  ?:

$$P(A) = \frac{60}{100} = \frac{3}{5}$$

# Conditional probability

---

## Example:

- What is the probability that the selected student is from Science **and** in a Major program,  $P(A \cap B)$  ?:

$$P(A \cap B) = \frac{50}{100} = \frac{1}{2}$$

# Conditional probability

---

## Example:

- Given that the selected student is **known** to be in a Major program, what is the probability that the student is from Science,  $P(A|B)$  ?

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{50/100}{80/100} = \frac{50}{80} = \frac{5}{8}.$$

# Conditional probability

---

## Note

- If  $B \equiv S$ , then

$$P(A|S) = \frac{P(A \cap S)}{P(S)} = \frac{P(A)}{1} = P(A).$$

- If  $B \equiv A$ , then

$$P(A|A) = \frac{P(A \cap A)}{P(A)} = \frac{P(A)}{P(A)} = 1.$$

- We have

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \leq \frac{P(B)}{P(B)} = 1.$$

## Conditional probability

---

Direct from the definition, we have that

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

(recall  $P(B) > 0$ ).

# Conditional probability

---

## Note

It is important to understand the distinction between

$$P(A \cap B) \quad \text{and} \quad P(A|B).$$

- $P(A \cap B)$  is the chance of A **and** B occurring **relative to S**. A and B are treated symmetrically in the calculation.
- $P(A|B)$  records the chance of A **and** B occurring **relative to B**. A and B are **not** treated symmetrically; from the definition, we see that in general

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

# Conditional probability

---

## Note

From the above calculation, we see that we could have

$$P(A|B) \leq P(A) \quad \text{or} \quad P(A|B) \geq P(A) \quad \text{or} \quad P(A|B) = P(A).$$

That is, knowledge that  $B$  occurs could **decrease**, **increase** or **leave unchanged** the probability that  $A$  occurs.

Recall: Simpson's Paradox

# Conditional probability

---

## Note

If  $P(B) > 0$ ,

(I) Non-negativity:

$$P(A|B) = P(A \cap B)/P(B) \geq 0.$$

(II) For the event  $S$ :

$$P(S|B) = P(S \cap B)/P(B) = P(B)/P(B) = 1$$

(III) Additivity: If  $A_1$  and  $A_2$  are mutually exclusive, then

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

i.e. conditional probability obeys the Probability Rules !

# Conditional probability

---

## Example:

A box of 500 light bulbs is purchased from each of two factories (labelled 1 and 2).

- In the box from Factory 1, there are 25 defective bulbs.
- In the box from Factory 2, there are 10 defective bulbs.

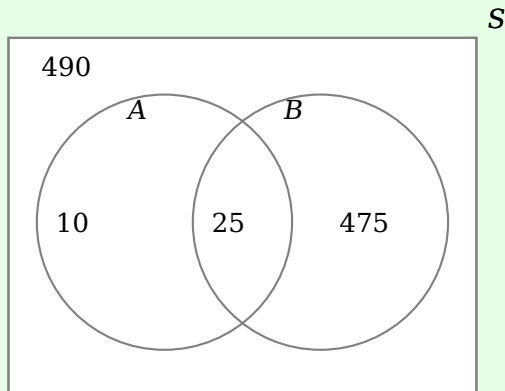
The bulbs are unpacked and placed in storage.

# Conditional probability

---

## Example:

$A$  – Defective;  $B$  – Factory 1.



# Conditional probability

---

## Example:

At the next installation, a bulb is selected from the store.

- What is the probability that the bulb **is defective**,  $P(A)$  ?:

$$P(A) = \frac{35}{1000} = 0.035.$$

# Conditional probability

---

## Example:

- What is the probability that the bulb came from Factory 1,  $P(B)$  ?:

$$P(B) = \frac{500}{1000} = 0.5.$$

# Conditional probability

---

## Example:

- Given that the selected bulb **came from Factory 1**, what is the probability it **is defective**  $P(A|B)$  ?:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{25/1000}{500/1000} = \frac{25}{500} = 0.05.$$

# Conditional probability

---

## Example:

- Given that the selected bulb **is defective**, what is the probability it **came from Factory 1**,  $P(B|A)$  ?:

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{25/1000}{35/1000} = \frac{25}{35} = \frac{5}{7}.$$

# Conditional probability

---

## Example:

We measure the failure time of an electrical component.

- $S = \mathbb{R}^+$ .
- $A = [10, \infty)$  (fails after 10 hours or more).
- $B = [5, \infty)$  (fails after 5 hours or more).

Suppose we assess

$$P(A) = 0.25 \quad P(B) = 0.4.$$

Now here  $A \subset B \subset S$ , so therefore  $P(A \cap B) \equiv P(A)$ , and

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)}{P(B)} = \frac{0.25}{0.4} = 0.625.$$

# Conditional probability

---

## Example: Drug testing

A drug testing authority tests 100000 athletes to assess whether they are using performance enhancing drugs (PEDs).

The drug test is not perfect as it produces

- **False positives:** declares an athlete to be using PEDs, when in reality they are not.
- **False negatives:** declares an athlete not to be using PEDs, when in reality they are.

# Conditional probability

---

## Example: Drug testing

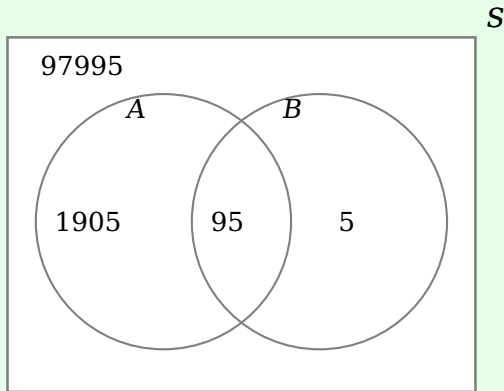
Suppose that after detailed investigation it is discovered that

- 2000 athletes gave positive test results.
- 100 athletes were using PEDs.
- 95 athletes who were using PEDs tested positive.

# Conditional probability

## Example: Drug testing

$A$  – Drug test is positive;  $B$  – Athlete is using PED.



# Conditional probability

---

## Example: Drug testing

An athlete is selected at random from the 100000, and is observed to test positive.

What is the probability that they actually were using PEDs ?

# Conditional probability

---

## Example: Drug testing

The required conditional probability is

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{95}{2000} = 0.0475.$$

Note that is very different from

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{95}{100} = 0.95.$$

# Independence

---

Two events  $A$  and  $B$  in sample space  $S$  are **independent** if

$$P(A|B) = P(A)$$

equivalently, they are independent if

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

Note that if  $P(A|B) = P(A)$ , then

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

# Independence

---

## Note

- Independence is not a property of the **events**  $A$  and  $B$  themselves, it is a property of the **probabilities** assigned to them.
- Independence is **not** the same as **mutual exclusivity**.

$$\text{Independent: } P(A \cap B) = P(A)P(B)$$

$$\text{Mutually exclusive: } P(A \cap B) = 0.$$

## Independence for multiple events

---

For events  $A_1, A_2, A_3$  in sample space  $S$ , we can consider independence **pairwise**: we might assume that

$$P(A_1 \cap A_2) = P(A_1)P(A_2)$$

$$P(A_1 \cap A_3) = P(A_1)P(A_3)$$

$$P(A_2 \cap A_3) = P(A_2)P(A_3).$$

Can we deduce from this

$$P(A_1 \cap A_2 \cap A_3) = P(A_1)P(A_2)P(A_3)?$$

In general, **no**.

# Independence for multiple events

---

## Example: Two dice

Suppose we roll two dice with outcomes independent.

- $A_1$  – first roll outcome is odd.
- $A_2$  – second roll outcome is odd.
- $A_3$  – total score is odd.

We have, for instance,

$$P(A_1) = P(A_2) = P(A_1|A_3) = P(A_2|A_3) = \frac{1}{2}$$

(identify all 36 pairs of equally likely scores and compute).

Thus  $A_1, A_2$  and  $A_3$  are **pairwise** independent.

# Independence for multiple events

---

## Example: Two dice

However,

$$P(A_1 \cap A_2 \cap A_3) = 0$$

and

$$P(A_1 | A_2 \cap A_3) = 0$$

etc.

# Conditional independence

---

Events  $A_1, A_2$  and  $B$  in sample space  $S$ , with  $P(B) > 0$ , are **conditionally independent** given  $B$  if

$$P(A_1|A_2 \cap B) = P(A_1|B)$$

or equivalently

$$P(A_1 \cap A_2|B) = P(A_1|B)P(A_2|B).$$

# General Multiplication Rule

---

For events  $A_1, A_2, \dots, A_K$ , we have the general result that

$$\begin{aligned} P(A_1 \cap A_2 \cap \dots \cap A_K) &= P(A_1) \\ &\quad \times P(A_2|A_1) \\ &\quad \times P(A_3|A_1 \cap A_2) \\ &\quad \vdots \\ &\quad \times P(A_K|A_1 \cap A_2 \cap \dots \cap A_{K-1}) \end{aligned}$$

Also known as the **Chain Rule** for probabilities.

# General Multiplication Rule

---

If the events are **mutually independent**, then we have that

$$P(A_1 \cap A_2 \cap \cdots \cap A_K) = \prod_{k=1}^K P(A_k)$$

# Probability Trees

---

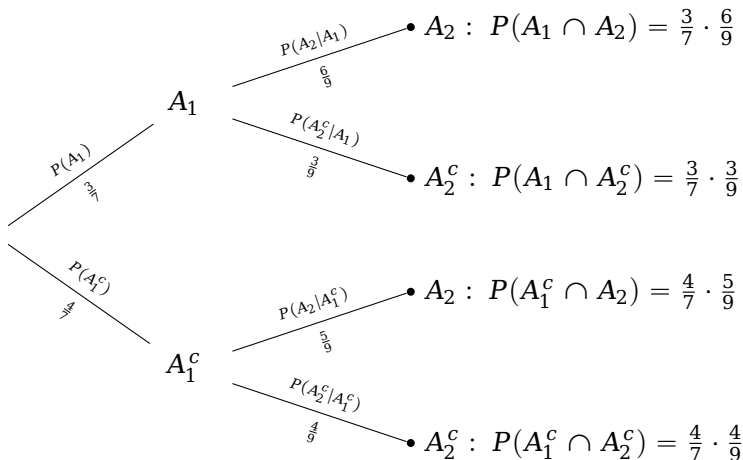
**Probability trees** are simple ways to display joint probabilities of multiple events. They comprise

- **Junctions:** correspond to the multiple events
- **Branches:** correspond to the sequence of (conditional) choices of events, given the previous choices.

We multiply along the branches to get the joint probabilities.

# Probability Trees

---



# Probability Trees

---

Note that each junction has two possible branches coming from it, corresponding to

$$A_k \quad \text{and} \quad A_k^c$$

respectively.

Such a tree extends to as many events as we need.

# The Theorem of Total Probability

---

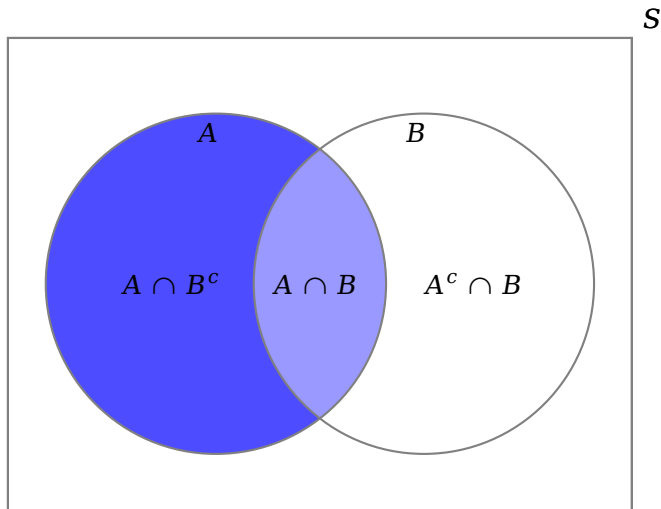
For two events  $A$  and  $B$  in sample space  $S$ , we have the **partition** of  $A$  as

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

- $(A \cap B)$ : “in  $A$  and in  $B$ ”
- $(A \cap B^c)$ : “in  $A$  and **not** in  $B$ ”

# The Theorem of Total Probability

---



# The Theorem of Total Probability

---

Therefore

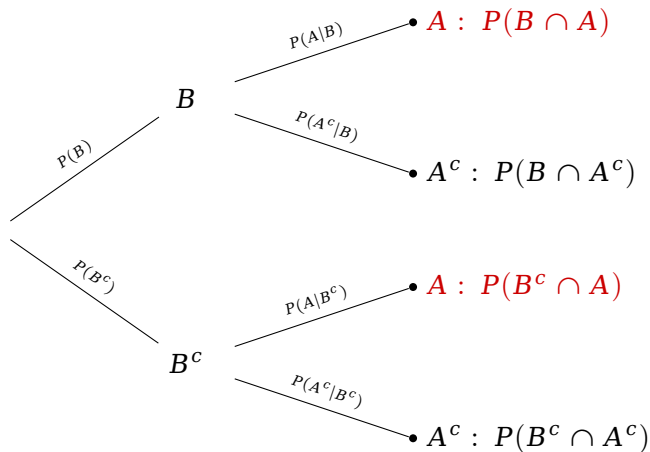
$$P(A) = P(A \cap B) + P(A \cap B^c)$$

and using the multiplication rule, we may rewrite this as

$$P(A) = P(A|B)P(B) + P(A|B^c)P(B^c).$$

# The Theorem of Total Probability

---



# The Theorem of Total Probability

---

That is, there are two ways to get to  $A$ :

- via  $B$ ;
- via  $B^c$ .

To compute  $P(A)$  we add up all the probabilities on paths that end up at  $A$ .

Note that  $B$  and  $B^c$  together form a partition of  $S$ .

# The Theorem of Total Probability

---

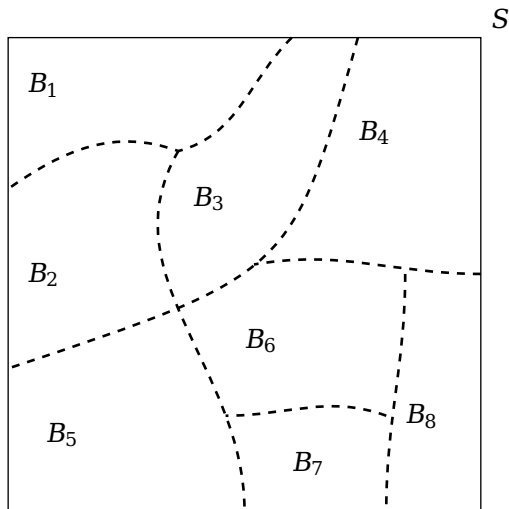
Now suppose we have a partition into  $K$  subsets

$$B_1, B_2, \dots, B_K.$$

Again, these events also partition  $A$ .

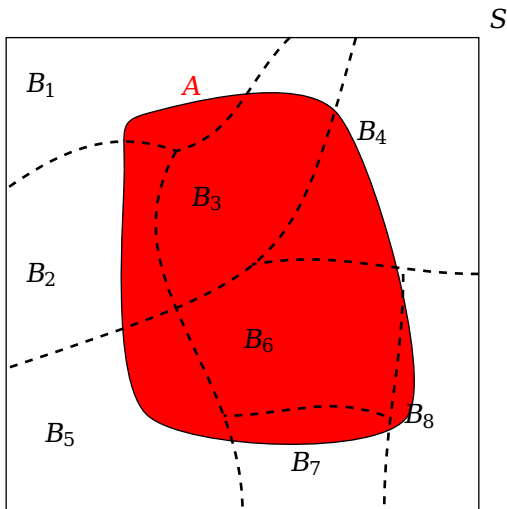
# The Theorem of Total Probability

---



# The Theorem of Total Probability

---



# The Theorem of Total Probability

---

That is, we have that

$$A = (A \cap B_1) \cup (A \cap B_2) \cup \cdots \cup (A \cap B_K)$$

and

$$P(A) = P(A \cap B_1) + P(A \cap B_2) + \cdots + P(A \cap B_K).$$

# The Theorem of Total Probability

---

Using the definition of conditional probability, we have

$$P(A) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \cdots + P(A|B_K)P(B_K).$$

that is

$$P(A) = \sum_{k=1}^K P(A|B_k)P(B_k).$$

This result is known as the **Theorem of Total Probability**.

# The Theorem of Total Probability

---

## Note

- The formula assumes that  $P(B_k) > 0$  for  $k = 1, \dots, K$ .
- It might be that

$$P(A \cap B_k) = P(A|B_k) = 0.$$

for some  $k$ .

- This formula represents how probability is used to compute  $P(A)$ .

# The Theorem of Total Probability

---

## Example: Three bags

Suppose that an experiment involves selecting a ball from one of three bags. Let

- Bag 1: 4 red and 4 white balls.
- Bag 2: 1 red and 10 white balls.
- Bag 3: 7 red and 11 white balls.

A bag is selected (with all bags equally likely), and then a ball is selected from that bag (with all balls equally likely).

What is the probability that the ball selected is red ?

# The Theorem of Total Probability

---

## Example: Three bags

Let

- $S$ : all possible selections of balls.
- $B_1$ : bag 1 selected.
- $B_2$ : bag 2 selected.
- $B_3$ : bag 3 selected.
- $A$ : ball selected is red.

# The Theorem of Total Probability

---

## Example: Three bags

We have that

$$\begin{aligned}P(A) &= P(A \cap B_1) + P(A \cap B_2) + P(A \cap B_3) \\&= P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + P(A|B_3)P(B_3) \\&= \left(\frac{4}{8} \times \frac{1}{3}\right) + \left(\frac{1}{11} \times \frac{1}{3}\right) + \left(\frac{7}{18} \times \frac{1}{3}\right) \\&= \frac{97}{297}\end{aligned}$$

# The Theorem of Total Probability

---

## Example: Three bags

Note that this is **not** equal to

$$\frac{\text{Total number of red balls}}{\text{Total number of balls}}$$

that is

$$\frac{4 + 1 + 7}{8 + 11 + 18} = \frac{12}{37}$$

as the red balls are **not equally likely** to be selected – this is only true conditional on the bag selection.

# Bayes Theorem

---

The second great theorem of probability is **Bayes Theorem**.

- Named after Reverend Thomas Bayes (1701 – 1761)

[https://en.wikipedia.org/wiki/Thomas\\_Bayes](https://en.wikipedia.org/wiki/Thomas_Bayes)

- Could be written **Bayes's Theorem**.
- It is not really a theorem.

# Bayes Theorem

---

## Theorem: Bayes Theorem

For two events  $A$  and  $B$  with  $P(A) > 0$  and  $P(B) > 0$ ,

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

If  $0 < P(B) < 1$ , then we may expand the denominator

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

using the Theorem of Total Probability.

## Bayes Theorem

---

“Proof”: By the **definition** of conditional probability

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

As  $P(A) > 0$  and  $P(B) > 0$  we can cross multiply and write

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

If  $0 < P(B) < 1$ , then we can write

$$P(A) = P(A|B)P(B) + P(A|B^c)P(B^c)$$

and substitute this expression in the denominator.

## Bayes Theorem

---

If  $B_1, B_2, \dots, B_K$  form a partition of  $S$ , with  $P(B_k) > 0$  for  $k = 1, \dots, K$ , and  $A$  is an event with  $P(A) > 0$ , then for  $k = 1, \dots, K$

$$P(B_k|A) = \frac{P(A|B_k)P(B_k)}{P(A)} = \frac{P(A|B_k)P(B_k)}{\sum_{j=1}^K P(A|B_j)P(B_j)}.$$

Probability tree interpretation: **Given** that we end up at  $A$ , what is the probability that we got there via branch  $B_k$  ?

# Bayes Theorem

---

## Example: Three bags

Suppose that an experiment involves selecting a ball from one of three bags. Let

- Bag 1: 4 red and 4 white balls.
- Bag 2: 1 red and 10 white balls.
- Bag 3: 7 red and 11 white balls.

A bag is selected (with all bags equally likely), and then a ball is selected from that bag (with all balls equally likely).

The ball selected is red. What is the probability it came from Bag 2 ?

# Bayes Theorem

---

## Example: Three bags

Let

- $S$ : all possible selections of balls.
- $B_1$ : bag 1 selected.
- $B_2$ : bag 2 selected.
- $B_3$ : bag 3 selected.
- $A$ : ball selected is red.

# Bayes Theorem

---

## Example: Three bags

Recall that

$$\begin{aligned}P(A) &= P(A \cap B_1) + P(A \cap B_2) + P(A \cap B_3) \\&= P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + P(A|B_3)P(B_3) \\&= \left(\frac{4}{8} \times \frac{1}{3}\right) + \left(\frac{1}{11} \times \frac{1}{3}\right) + \left(\frac{7}{18} \times \frac{1}{3}\right) \\&= \frac{97}{297}\end{aligned}$$

# Bayes Theorem

---

## Example: Three bags

Then

$$\begin{aligned}P(B_2|A) &= \frac{P(A|B_2)P(B_2)}{P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + P(A|B_3)P(B_3)} \\&= \frac{\frac{1}{11} \times \frac{1}{3}}{\frac{4}{8} \times \frac{1}{3} + \frac{1}{11} \times \frac{1}{3} + \frac{7}{18} \times \frac{1}{3}} \\&= \frac{1/33}{97/297} \\&= \frac{9}{97} \simeq 0.09278.\end{aligned}$$

# Bayes Theorem

---

## Example: Three bags

We can also compute that

$$P(B_1|A) = \frac{4/24}{97/297} = \frac{297}{582} \simeq 0.51031$$

and

$$P(B_3|A) = \frac{21/54}{97/297} = \frac{1746}{2079} \simeq 0.39691.$$

# Bayes Theorem

---

## Note

We have that

$$P(B_1|A) + P(B_2|A) + P(B_3|A) = 1$$

but note that

$$P(A|B_1) + P(A|B_2) + P(A|B_3) \neq 1.$$

In the first formula, we are conditioning on  $A$  everywhere; in the second, we have different conditioning sets.

# Using Bayes Theorem

---

Bayes theorem is often used to make probability statements concerning an event  $B$  that **has not been observed**, given an event  $A$  that **has been observed**.

# Using Bayes Theorem

---

## Example: Medical screening

A health authority tests individuals from a population to assess whether they are sufferers from some disease. The screening test is not perfect as it produces

- **False positives:** declares someone to be a sufferer, when in reality they are a non-sufferer.
- **False negatives:** declares someone to be a non-sufferer, when in reality they are a sufferer.

# Using Bayes Theorem

---

## Example: Medical screening

Suppose that we denote by

- $A$  – Screening test is positive;
- $B$  – Person is actually a sufferer.

# Using Bayes Theorem

---

## Example: Medical screening

Suppose that

- $P(B) = p$ ;
- $P(A|B) = 1 - \alpha$  (true positive rate);
- $P(A|B^c) = \beta$  (false positive rate).

for probabilities  $p, \alpha, \beta$ .

# Using Bayes Theorem

---

## Example: Medical screening

We have for the rate of positive tests

$$P(A) = P(A|B)P(B) + P(A|B^c)P(B^c) = (1 - \alpha)p + \beta(1 - p)$$

by the Theorem of Total Probability, and by Bayes theorem

$$\begin{aligned} P(B|A) &= \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|B^c)P(B^c)} \\ &= \frac{(1 - \alpha)p}{(1 - \alpha)p + \beta(1 - p)}. \end{aligned}$$

# Using Bayes Theorem

---

## Example: Medical screening

Similarly, for the **false negative rate**, we have

$$\begin{aligned}P(B|A^c) &= \frac{P(A^c|B)P(B)}{P(A^c|B)P(B) + P(A^c|B^c)P(B^c)} \\&= \frac{\alpha p}{\alpha p + (1 - \beta)(1 - p)}.\end{aligned}$$

# Using Bayes Theorem

---

## Example: Medical screening

We have

$$P(A|B) = (1 - \alpha) \quad P(B|A) = (1 - \alpha) \frac{p}{(1 - \alpha)p + \beta(1 - p)}$$

and these two values are potentially very different.

# Using Bayes Theorem

---

## Example: Medical screening

That is, in general,

$$P(\text{"Spots"}|\text{"Measles"}) \neq P(\text{"Measles"}|\text{"Spots"})$$

# Using Bayes Theorem

---

## Note

This is sometimes known as the **Prosecutor's Fallacy**:

$$P(\text{"Evidence"}|\text{"Guilt"}) \neq P(\text{"Guilt"}|\text{"Evidence"})$$

## Bayes Theorem with multiple events

---

If  $A_1$  and  $A_2$  are two events in a sample space  $S$  so that

$$P(A_1 \cap A_2) > 0$$

and there is a partition  $B_1, \dots, B_K$ , with  $P(B_k) > 0$ , then

$$P(B_k|A_1 \cap A_2) = \frac{P(A_1 \cap A_2|B_k)P(B_k)}{\sum_{j=1}^K P(A_1 \cap A_2|B_j)P(B_j)}$$

## Bayes Theorem with multiple events

---

- we use the previous version of Bayes Theorem with

$$A \equiv A_1 \cap A_2.$$

- this extends to conditioning events  $A_1, A_2, \dots, A_K$ .

By earlier results

$$P(A_1 \cap A_2 | B_k) = P(A_1 | B_k)P(A_2 | A_1 \cap B_k)$$

that is, the chain rule applies to conditional probabilities also.