

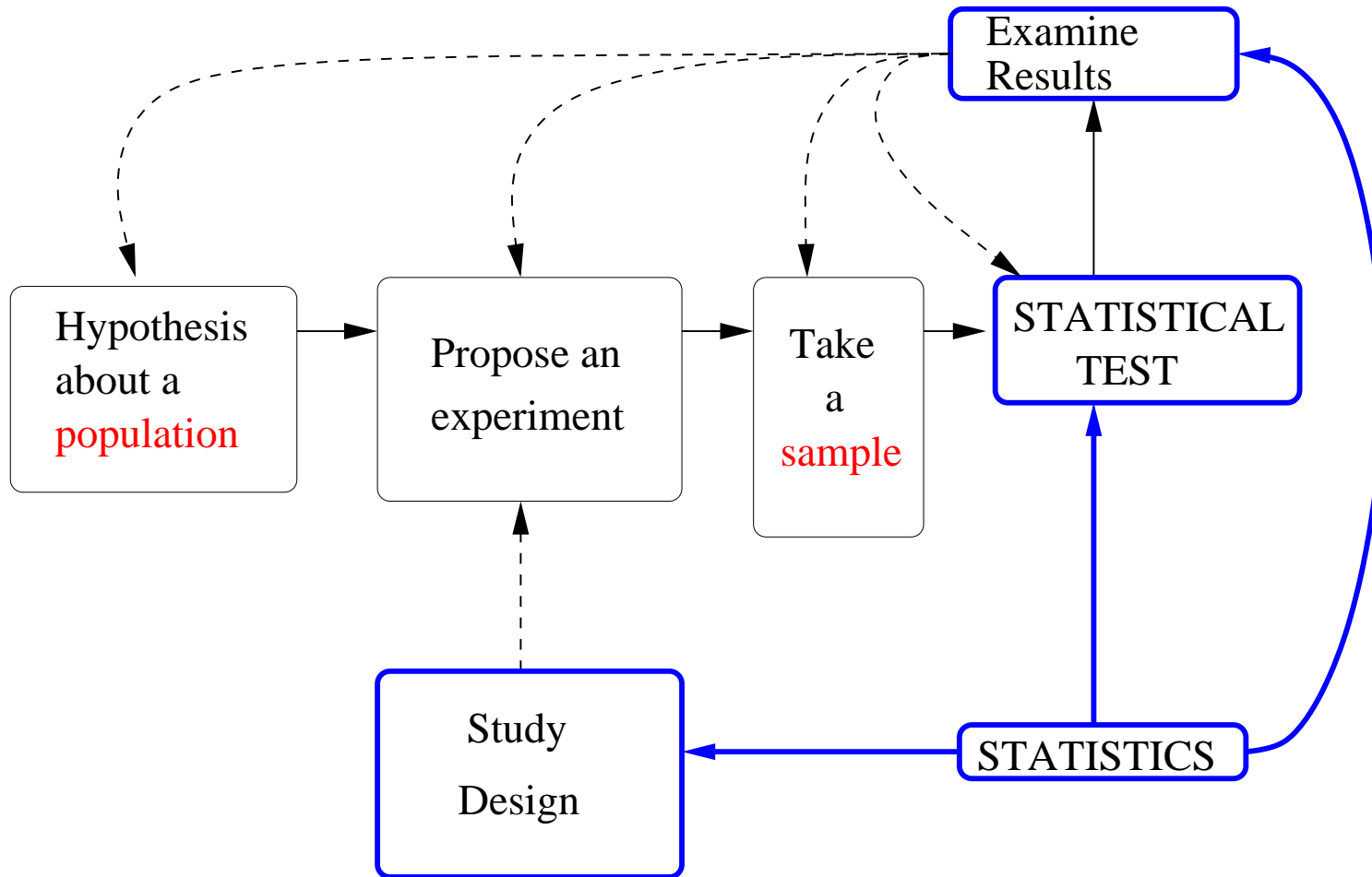
Lectures 2 and 3: Probability

Jonathan Marchini

Outline

- Why do we need to learn about probability?
- What is probability?
- How to assign probabilities
- How to manipulate probabilities and calculate probabilities of complex events

Why do we need probability?



Why do we need probability?

- The conclusions we make about a hypothesis depend upon the sample we take.
- The sample we take may lead us to the wrong conclusion.
- We need to know what the chances are of this happening.
- Probability is the study of chance.
- That's why we need probability!

The Baby-boom dataset

Hypothesis: Boys weigh more than girls at birth.

Sample mean of boys weights = $\bar{x}_{\text{boys}} = 3375.308$

Sample mean of girls weights = $\bar{x}_{\text{girls}} = 3132.444$

$$\Rightarrow D = \bar{x}_{\text{boys}} - \bar{x}_{\text{girls}} = 3375.30 - 3132.444 = 242.8632$$

What are the chances of obtaining a value of D this big?

If the chances are small then we can be confident in our hypothesis.

Puzzle 1

Pick any two types of card that can occur in a normal pack of shuffled playing cards e.g. Queen and 6.

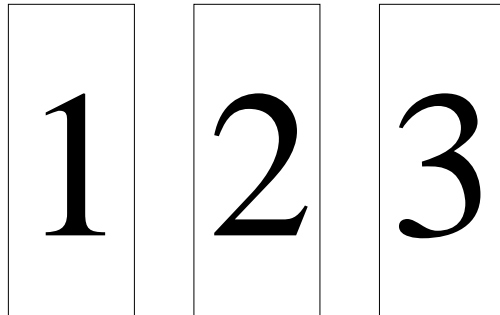
What do you think is the probability that somewhere in the pack a Queen is next to a 6?

Write down what you think the probability is.

We'll conduct an experiment in lecture to collect data and in future lectures test to see if the data is consistent with your hypothesis.

Puzzle 2

You have reached the final of a game show. The host shows you 3 doors and tells you that there is a prize behind one of the doors. You pick a door. The host then opens one of the doors you didn't pick that contains no prize and asks you if you want to change from the door you chose to the other remaining door. Should you change?



Example 2 : Throwing a die

Consider the **experiment** in which we throw a die once.

The **sample space** is

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

The outcome “the top face shows a three” is the **sample point** 3.

The **event** A_1 , that the die shows an even number is the subset $A_1 = \{2, 4, 6\}$ of the sample space.

The **event** A_2 that the die shows a number larger than 4 is the subset $A_2 = \{5, 6\}$ of S_2 .

Definitions

When we talk about probabilities we talk about the probability of events that might occur in some experiment.

An **experiment** is some activity with an observable outcome.

The set of all possible outcomes of the experiment is called the **sample space**.

A particular outcome is called a **sample point**.

A collection of possible outcomes is called an **event**.

Calculating simple probabilities

Simply speaking, the probability of an event is a number between 0 and 1, inclusive, that indicates how likely the event is to occur.

In some settings, it is natural to assume that all the sample points are equally likely.

In this case, we can calculate the probability of an event A as

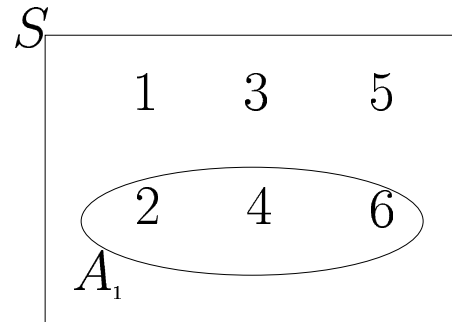
$$P(A) = \frac{|A|}{|S|},$$

where $|A|$ denotes the number of sample points in the event A .

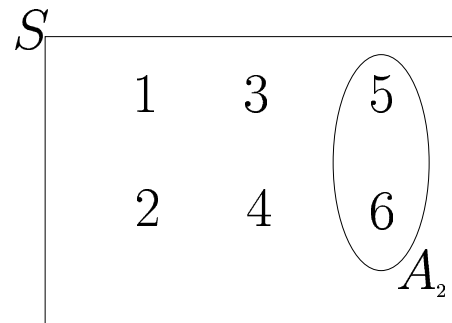
Example 2 (Continued)

$$S = \{1, 2, 3, 4, 5, 6\} \quad A_1 = \{2, 4, 6\} \quad A_2 = \{5, 6\}$$

$$P(A_1) = \frac{|A_1|}{|S|} = \frac{3}{6} = \frac{1}{2}$$



$$P(A_2) = \frac{|A_2|}{|S|} = \frac{2}{6} = \frac{1}{3}$$



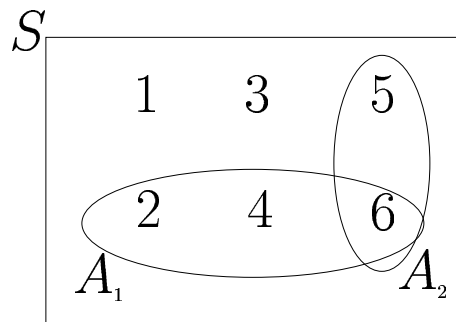
Intersection

What about $P(\text{face is even, and larger than 4})$?

We can write this event in set notation as $A_1 \cap A_2$.

This is the **intersection** of the two events, A_1 and A_2
i.e the set of elements which belong to both A_1 and A_2 .

$$A_1 \cap A_2 = \{6\} \quad \Rightarrow \quad P(A_1 \cap A_2) = \frac{|A_1 \cap A_2|}{|S|} = \frac{1}{6}$$



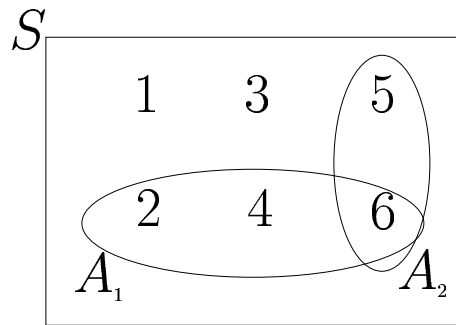
Union

What about $P(\text{face is even, or larger than 4})$?

We can write this event in set notation as $A_1 \cup A_2$.

This is the **union** of the two events, A_1 and A_2
i.e the set of elements which belong either A_1 and A_2
or both.

$$A_1 \cup A_2 = \{2, 4, 5, 6\} \quad \Rightarrow \quad P(A_1 \cup A_2) = \frac{|A_1 \cup A_2|}{|S|} = \frac{4}{6} = \frac{2}{3}$$



Complement

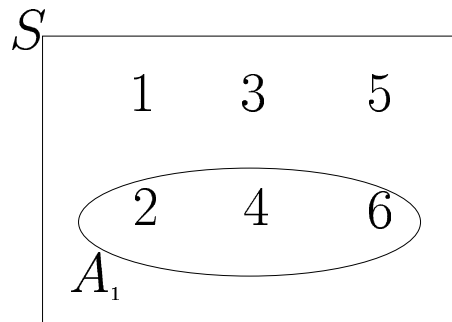
What about $P(\text{face is **not** even})$?

We can write this event in set notation as A_1^c .

This is the **complement** of the event, A_1

i.e the set of elements which do not belong to A_1 .

$$A_1^c = \{1, 3, 5\} \quad \Rightarrow \quad P(A_1^c) = \frac{|A_1^c|}{|S|} = \frac{3}{6} = \frac{1}{2}$$



Exercise 1

Throw a fair die twice.

1. Write down the sample space for this experiment.
2. Calculate the probability of the event E_1 that the sum of the faces showing is less than 9?
3. Calculate the probability of the event E_2 that the sum of the faces showing is even?
4. Calculate the probability of the event $E_1 \cap E_2$?

Probability in more general settings

In many settings, either the sample space is infinite or all possible outcomes of the experiment are not equally likely. We still wish to associate probabilities with events of interest.

Luckily, there are some rules/laws that allow us to calculate and manipulate such probabilities with ease.

Probability Axioms (Building Blocks)

There are three axioms which we need in order to develop our laws

- I. $0 \leq P(A) \leq 1$ for any event A .
- II. $P(S) = 1$.
- III. If A_1, \dots, A_n are **mutually exclusive** events, then

$$P(A_1 \cup \dots \cup A_n) = P(A_1) + \dots + P(A_n).$$

A set of events are **mutually exclusive** if at most one of the events can occur in a given experiment.

Complement Law

Suppose

- A = The event that a randomly selected student from a class has a bike

What is the probability that a student does not have a bike?

This is the **complement** of the event A , i.e. A^c

$$P(A^c) = 1 - P(A)$$

Eg If $P(A) = 0.36$ then $P(A^c) = 1 - 0.36 = 0.64$

Addition Law (Union)

Suppose

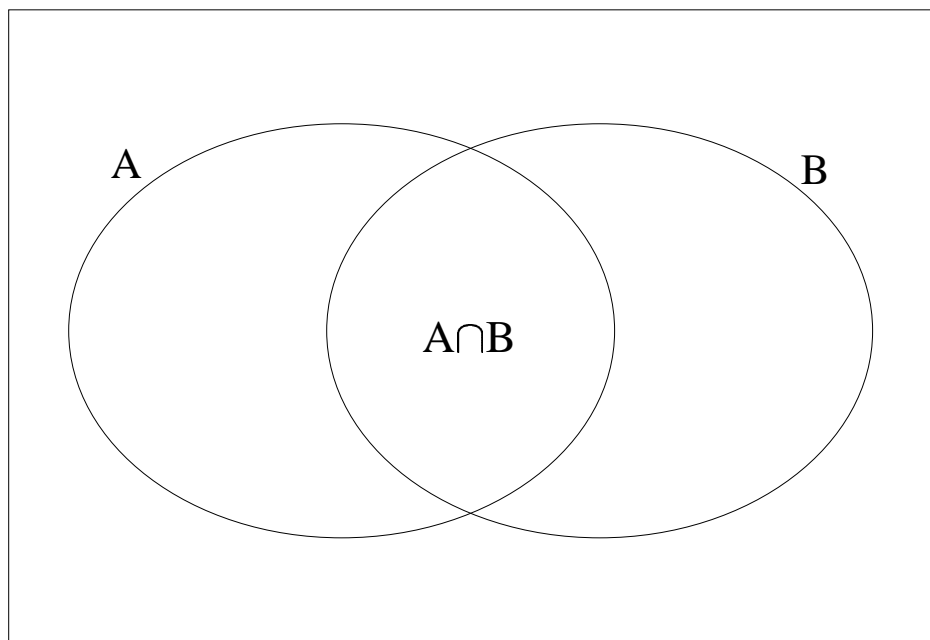
- A = a randomly selected student from a class has brown eyes
- B = a randomly selected student from a class has blue eyes

What is the probability that a student has brown eyes **OR** blue eyes?

This is the **union** of the two events A and B, i.e. $A \cup B$

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

We have to subtract $P(A \cap B)$ because when we add $P(A)$ and $P(B)$ we count $P(A \cap B)$ twice.



Example 3 : SNPs

Single nucleotide polymorphisms (SNPs) are nucleotide positions in a genome which exhibit variation amongst individuals in a species. In some studies in humans, SNPs are discovered in European populations. Suppose that of such SNPs, 70% also show variation in an African population, 80% show variation in an Asian population and 60% exhibit variation in both the African and Asian population.

Suppose one such SNP is chosen at random, what is the probability that it is variable in either the African or the Asian population?

Write A for the event that the SNP is variable in Africa, and B for the event that it is variable in Asia. We are told

$$P(A) = 0.7$$

$$P(B) = 0.8$$

$$P(A \cap B) = 0.6.$$

We require $P(A \cup B)$. From the addition rule:

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

$$= 0.7 + 0.8 - 0.6$$

$$= 0.9.$$

Conditional Probability Laws

Suppose

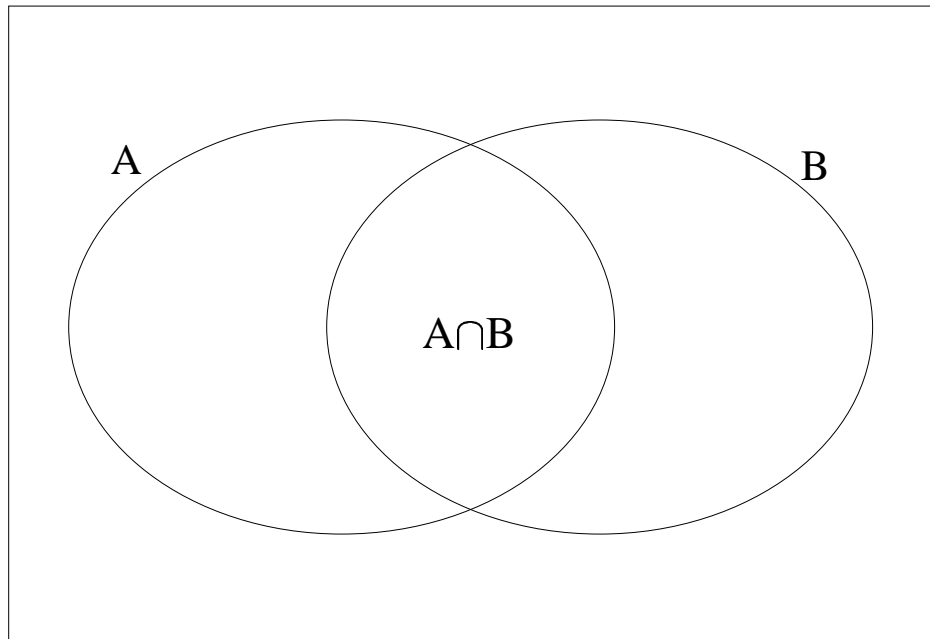
- A = a randomly selected student from the class has a bike
- B = a randomly selected student from the class has blue eyes

What is the probability that a student has blue eyes **GIVEN** that the student has a bike?

This is a **conditional** probability.

We write this probability as $P(B|A)$ (pronounced 'probability of B given A')

Think of $P(B|A)$ as ‘how much of A is taken up by B ’.



Then we see that

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

Example 4

For the SNP example what is the probability that a SNP is variable in the African population given that it is variable in the Asian population?

We have that

$$P(A) = 0.7$$

$$P(B) = 0.8$$

$$P(A \cap B) = 0.6.$$

We want

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{0.6}{0.8} = 0.75$$

We can rearrange the conditional probability laws to obtain a general Multiplication Law

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

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

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

Independence of Events

Definition Two events A and B are said to be *independent* if $P(A \cap B) = P(A)P(B)$.

Note that in this case (provided $P(B) > 0$), if A and B are independent

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

and similarly $P(B|A) = P(B)$ (provided $P(A) > 0$).

So for independent events, knowledge that one of the events has occurred does not change our assessment of the probability that the other event has occur.

Example 5 : Snails

In a population of a particular species of snail, individuals exhibit different forms. It is known that 45% have a pink background colouring, while 55% have a yellow background colouring. In addition, 30% of individuals are striped, and 20% of the population are pink and striped.

1. Is the presence or absence of striping independent of background colour?
2. Given that a snail is pink, what is the probability that it will have stripes.

Define the events: A , B , that a snail has a pink, respectively yellow, background colouring, and S for the event that it has stripes.

Then we are told $P(A) = 0.45$, $P(B) = 0.55$, $P(S) = 0.3$, and $P(A \cap S) = 0.2$.

For part (1), note that

$$0.2 = P(A \cap S) \neq 0.135 = P(A)P(S),$$

so the events A and S are not independent.

For part (2),

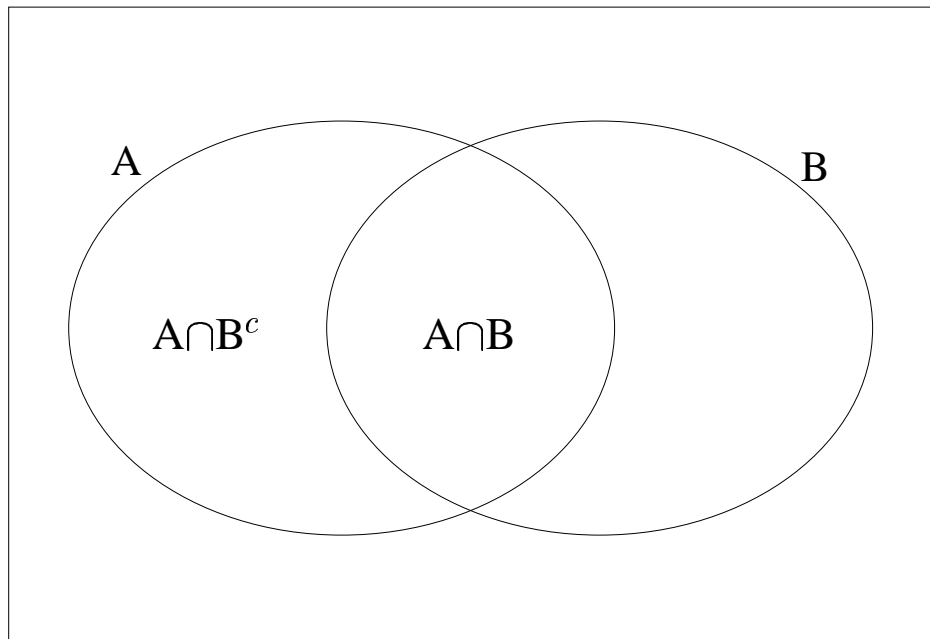
$$P(S|A) = \frac{P(S \cap A)}{P(A)} = \frac{0.2}{0.45} = 0.44.$$

Thus, knowledge that a snail has a pink background colouring increases the probability that it is striped. (That $P(S|A) \neq P(S)$ also establishes that background colouring and the presence of stripes, are not independent.)

The Partition Rule

If $P(A \cap B) = 0.52$ and $P(A \cap B^c) = 0.14$ what is $p(A)$?

$P(A)$ is made up of two parts (i) the part of A contained in B (ii) the part of A contained in B^c .



So we have the rule

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

and $P(A) = P(A \cap B) + P(A \cap B^c) = 0.52 + 0.14 = 0.66$

More generally, if E_1, \dots, E_n are a set of *mutually exclusive* events then

$$\boxed{P(A) = \sum_{i=1}^n P(A \cap E_i) = \sum_{i=1}^n P(A | E_i)P(E_i)}$$

A set of events are *mutually exclusive* if at most one of the events can occur in a given experiment.

Example 6 : Mendelian segregation

At a particular locus in humans, there are two alleles A and B , and it is known that the population frequencies of the genotypes AA , AB , and BB , are 0.49, 0.42, and 0.09, respectively. An AA man has a child with a woman whose genotype is unknown.

What is the probability that the child will have genotype AB ?

We assume that as far as her genotype at this locus is concerned the woman is chosen randomly from the population.

Use the partition rule, where the partition corresponds to the three possible genotypes for the woman. Then

$$\begin{aligned} P(\text{child } AB) &= P(\text{child } AB \text{ and mother } AA) \\ &\quad + P(\text{child } AB \text{ and mother } AB) \\ &\quad + P(\text{child } AB \text{ and mother } BB) \\ &= P(\text{mother } AA)P(\text{child } AB|\text{mother } AA) \\ &\quad + P(\text{mother } AB)P(\text{child } AB|\text{mother } AB) \\ &\quad + P(\text{mother } BB)P(\text{child } AB|\text{mother } BB) \\ &= 0.49 \times 0 + 0.42 \times 0.5 + 0.09 \times 1 \\ &= 0.3, \end{aligned}$$

Bayes Rule

Bayes Rule is a *very* powerful probability law.

An example from medicine

Let D be a disease and S a symptom.

A doctor may be interested in $P(D|S)$.

This is a hard probability to assign.

A probability that is much easier to calculate is $P(S|D)$,
i.e. from patient records..

The power of Bayes Rule is its ability to take $P(S|D)$
and calculate $P(D|S)$.

We have actually already seen a version of Bayes Rule before

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

Using the Multiplication Law we can re-write this as

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

(Bayes Rule)

Example 7 : Disease genes

A gene has two possible types A_1 and A_2 . 75% of the population have A_1 . B is a disease that has 3 forms B_1 (mild), B_2 (severe) and B_3 (lethal). A_1 is a protective gene, with the probabilities of having the three forms given A_1 as 0.9, 0.1 and 0 respectively. People with A_2 are unprotected and have the three forms with probabilities 0, 0.5 and 0.5 respectively.

What is the probability that a person has gene A_1 given they have the severe disease?

The first thing to do with such a question is ‘decode’ the information, i.e. write it down in a compact form we can work with.

$$P(A_1) = 0.75 \quad P(A_2) = 0.25$$

$$P(B_1|A_1) = 0.9 \quad P(B_2|A_1) = 0.1 \quad P(B_3|A_1) = 0$$

$$P(B_1|A_2) = 0 \quad P(B_2|A_2) = 0.5 \quad P(B_3|A_2) = 0.5$$

We want $P(A_1|B_2)$?

From Bayes Rule we know that

$$\mathbf{P}(A_1|B_2) = \frac{\mathbf{P}(B_2|A_1)\mathbf{P}(A_1)}{\mathbf{P}(B_2)}$$

We know $\mathbf{P}(B_2|A_1)$ and $\mathbf{P}(A_1)$ but what is $\mathbf{P}(B_2)$?

$$\begin{aligned}\mathbf{P}(B_2) &= \mathbf{P}(B_2 \cap A_1) + \mathbf{P}(B_2 \cap A_2) \\ &= \mathbf{P}(B_2|A_1)\mathbf{P}(A_1) + \mathbf{P}(B_2|A_2)\mathbf{P}(A_2) \\ &= 0.1 \times 0.75 + 0.5 \times 0.25 \\ &= 0.2\end{aligned}$$

Finally

$$\mathbf{P}(A_1|B_2) = \frac{0.1 \times 0.75}{0.2} = 0.375$$

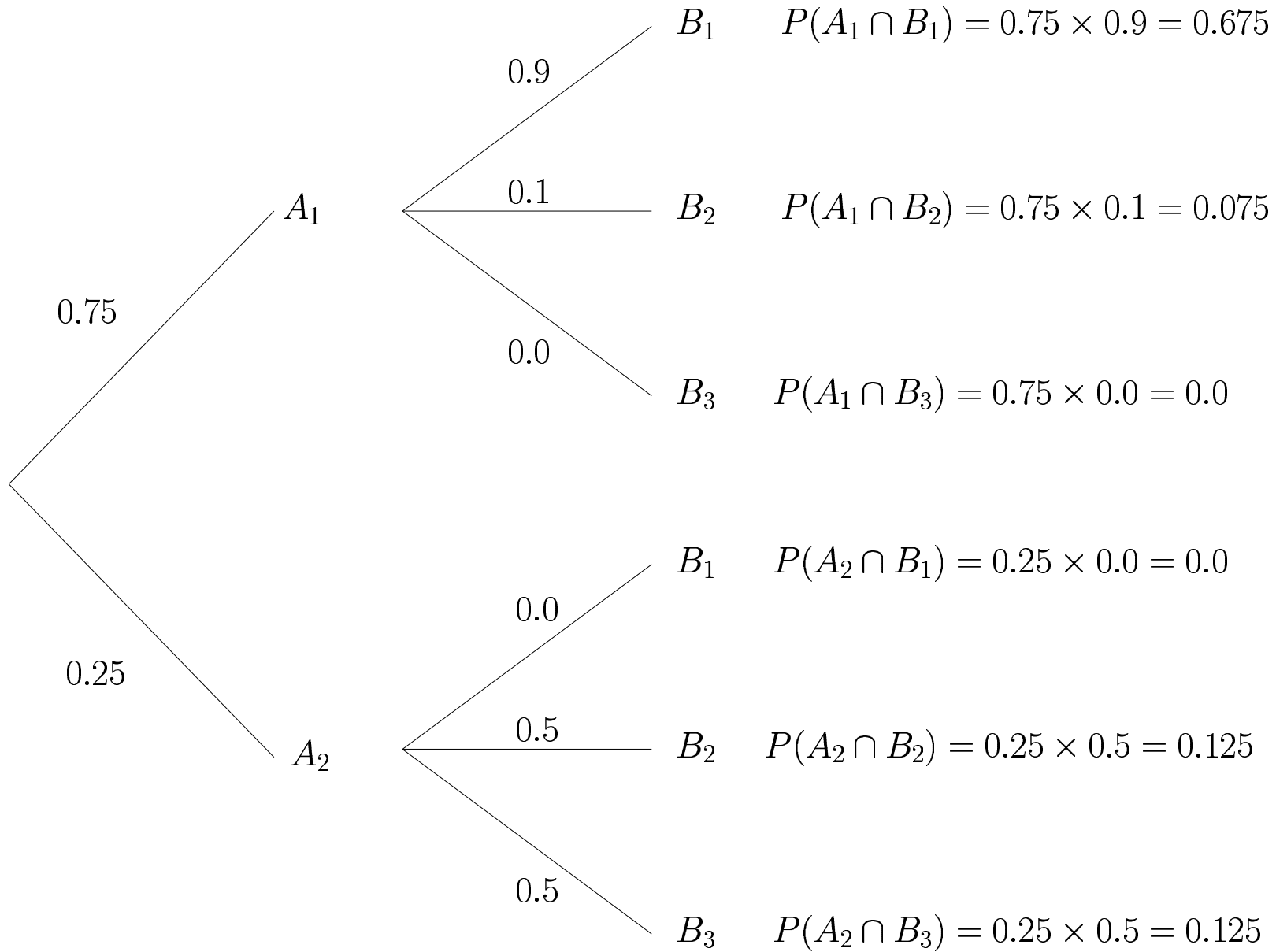
Probability Trees

A useful way of tackling many probability problems is to draw a ‘probability tree’.

The branches of the tree represent different possible events.

Each branch is labelled with the probability of choosing it given what has occurred before.

The probability of a given route through the tree can then be calculated by multiplying all the probabilities along that route (using the Multiplication Rule)



We can calculate $P(B_2)$ by adding up the probabilities of the paths that involve B_2 .

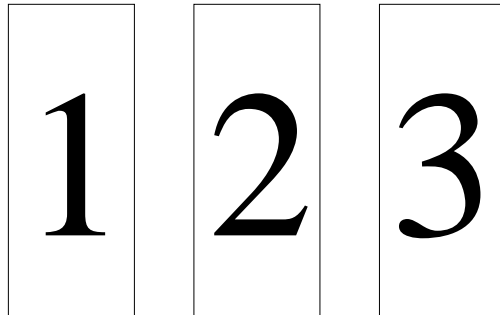
Thus, $P(B_2) = 0.075 + 0.125 = 0.2$.

Also

$$P(A_1|B_2) = \frac{P(A_1 \cap B_2)}{P(B_2)} = \frac{0.075}{0.2} = 0.375$$

Puzzle 2

You have reached the final of a game show. The host shows you 3 doors and tells you that there is a prize behind one of the doors. You pick a door. The host then opens one of the doors you didn't pick that contains no prize and asks you if you want to change from the door you chose to the other remaining door. Should you change?



Permutations and Combinations (Probabilities of patterns)

In some situations we observe a specific pattern from a large number of possible patterns.

To calculate the probability of the pattern we need to count the number of ways our pattern could have arisen.

This is why we need to learn about permutations and combinations.

Permutations of n objects

Consider 2 objects A B

Q. How many ways can they be arranged?
i.e. how many **permutations** are there?

A. 2 ways AB BA

Consider 3 objects A B C

Q. How many ways can they be arranged (permuted)?

A. 6 ways ABC ACB BCA BAC CAB CBA

Consider 4 objects A B C D

Q. How many ways can they be arranged (permuted)?

A. 24 ways

ABCD ABDC ACBD ACDB ADBC ADCB
BACD BADC BCAD BCDA BDAC BDCA
CBAD CBDA CABD CADB CDBA CDAB
DBCA DBAC DCBA DCAB DABC DACB

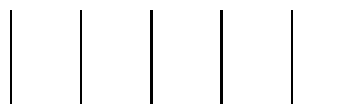
There is a pattern emerging here.

No. of objects	2	3	4	5	6	...
No. of permutations	2	6	24	120	720	...

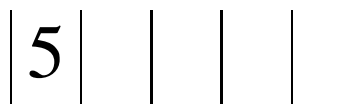
Can we find a formula for the number of permutations of n objects?

A good way to think about permutations is to think of putting objects into boxes.

Suppose we have 5 objects. How many different ways can we place them into 5 boxes?



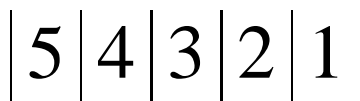
There are 5 choices of object for the first box.



There are now only 4 objects to choose from for the second box.



There are 3 choices for the 3rd box, 2 for the 4th and 1 for the 5th box.



Thus, the number of permutations of 5 objects is

$$5 \times 4 \times 3 \times 2 \times 1$$

In general, the number of permutations of n objects is

$$n(n - 1)(n - 2) \dots (3)(2)(1)$$

We write this as $n!$ (pronounced ‘n factorial’).

There should be a button on your calculator that calculates factorials.

Permutations of r objects from n

Now suppose we have 4 objects and only 2 boxes. How many permutations of 2 objects are there when we have 4 to choose from?

There are 4 choices for the first box and 3 choices for the second box

$$\boxed{4|3}$$

So there are 12 permutations of 2 objects from 4. We write this as

$${}^4P_2 = 12$$

We say there are ${}^n P_r$ permutations of r objects chosen from n .

The formula for ${}^n P_r$ is given by

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

To see why this works consider the example above ${}^4 P_2$.
Using the formula we get

$${}^4 P_2 = \frac{4!}{2!} = \frac{4 \times 3 \times 2 \times 1}{2 \times 1} = 4 \times 3$$

Combinations of r objects from n

Now consider the number of ways of choosing 2 objects from 4 when the order doesn't matter. We just want to count the number of possible **combinations**.

We know that there are 12 permutations when choosing 2 objects from 4. These are

AB AC AD BC BD CD
BA CA DA CB DB DC

Notice how the permutations are grouped in 2's which are the same combination of letters.

Thus there are $12/2 = 6$ possible combinations.

AB AC AD BC BD CD

We write this as

$${}^4C_2 = 6$$

We say there are nC_r combinations of r objects chosen from n .

The formula for nC_r is given by

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

Another way of writing this formula that makes it clearer is

$${}^n C_r = \frac{{}^n P_r}{r!}$$

Effectively this says we count the number of permutations of r objects from n and then divide by $r!$ because the ${}^n P_r$ permutations will occur in groups of $r!$ that are the same combination.

Example: The National Lottery

In the National Lottery you need to choose 6 balls from 49.

What is the probability that I choose all 6 balls correctly?

There are 2 ways of answering this question (i) using permutations and combinations (ii) using a tree diagram

Method 1 - using permutations and combinations

$$\begin{aligned} P(6 \text{ correct}) &= \frac{\text{No. of ways of choosing the 6 correct balls}}{\text{No. of ways of choosing 6 balls}} \\ &= \frac{{}^6P_6}{{}^{49}P_6} \\ &= \frac{6!}{\frac{49!}{43!}} \\ &= \frac{6 \times 5 \times 4 \times 3 \times 2 \times 1}{49 \times 48 \times 47 \times 46 \times 45 \times 44} \\ &= 0.0000000715112 \quad (1 \text{ in } 14 \text{ million}) \end{aligned}$$

Method 2 - using a tree diagram

The tree we would draw is big and at each stage has 2 branches, one for picking a correct ball and one for picking an incorrect ball. We need to calculate the product of probabilities along the branch where we always pick the right ball.

Consider the first ball I choose, the probability it is correct is

$$\frac{6}{49}$$

The second ball I choose is correct with probability

$$\frac{5}{48}$$

The third ball I choose is correct with probability

$$\frac{4}{47}$$

and so on.

Thus the probability that I get all 6 balls correct is

$$\frac{6}{49} \frac{5}{48} \frac{4}{47} \frac{3}{46} \frac{2}{45} \frac{1}{44} = 0.0000000715112 \quad (1 \text{ in } 14 \text{ million})$$