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: Ω =?

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: Ω ={head, tail}

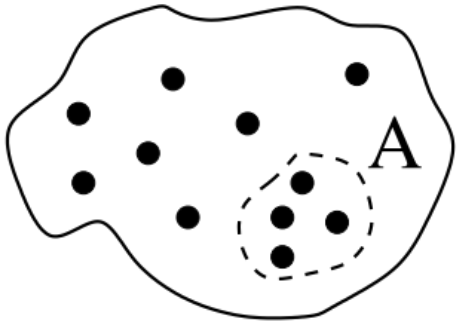
Events vs Sample Space

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 $A \cap B$, $A \cup B$, $\overline{A}$, and so on.*
- *It makes no sense to talk about events unless we have first defined the random experiment and the sample space. This is not always as easy as it sounds!*

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

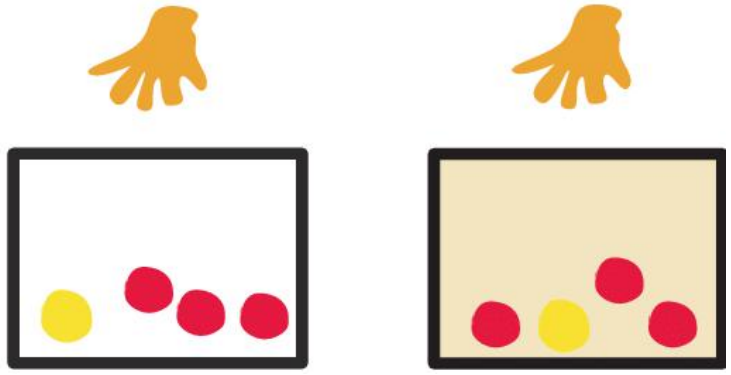
Ω is a bag of items



Event A is a smaller bag of items

Ω is a region

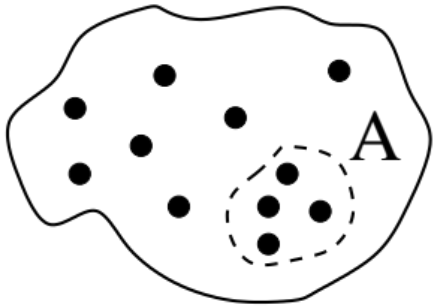
Event A is a subregion



Probabilities

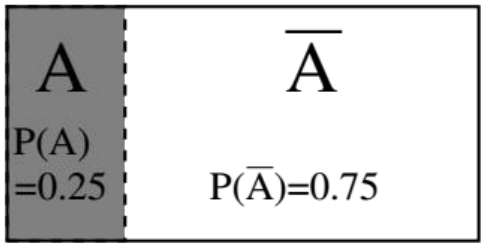
The idea of probability is to *attach a number to every item or event in* Ω *that reflects how likely the event is to occur.*

Ω is a bag of items

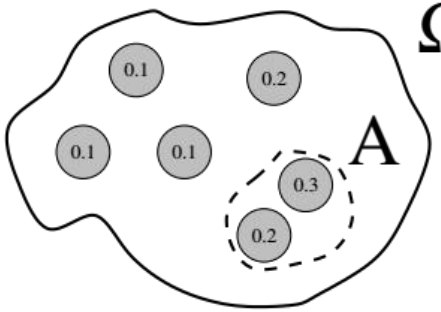


P(A)=4/11 if all items are equally likely

Ω is a region



Probability is represented by AREA

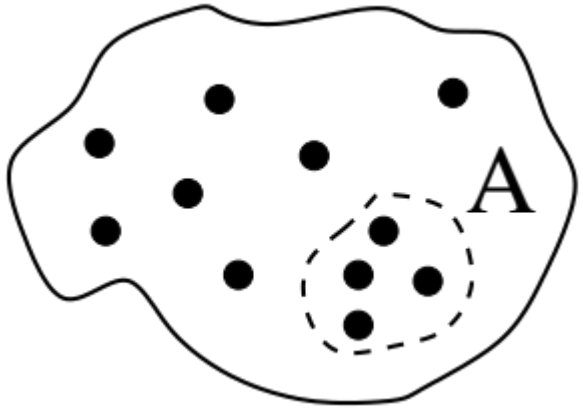


P(A)=0.5 even though bag A contains only 2 out of 6 items

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.*

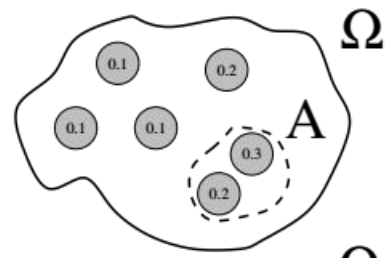
Axiom 2: $0 \leq \mathbb{P}(A) \leq 1$ for all events A.

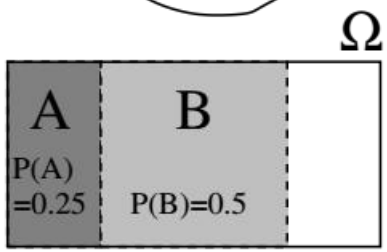
- *This says that probability is always a number between 0 and 1.*

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

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

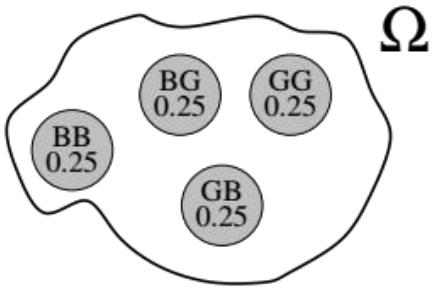
- *This says that if you have non-overlapping sets, the amount of cake they have in total is the sum of their individual amounts.*
- *This axiom is the reason why we can say that $\mathbb{P}(A) = 0.3 + 0.2 = 0.5$ in the bag diagram.*
- *In the region diagram, $\mathbb{P}(A \cup B) = 0.25 + 0.5 = 0.75$.*





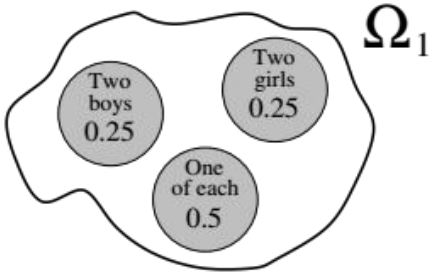
Examples of probability distributions

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:***



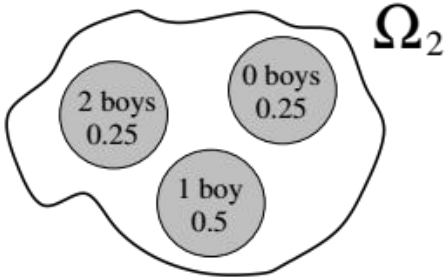
So if we pick a two-child family at random, we have probability $1/4 = 0.25$ of getting each of the outcomes BB, BG, GB, and GG.

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:



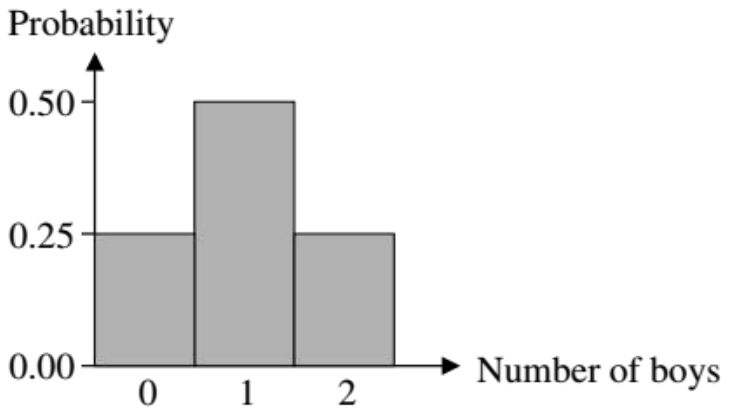
Examples of probability distributions

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* $\Omega_2 = \{0, 1, 2\}$*:*



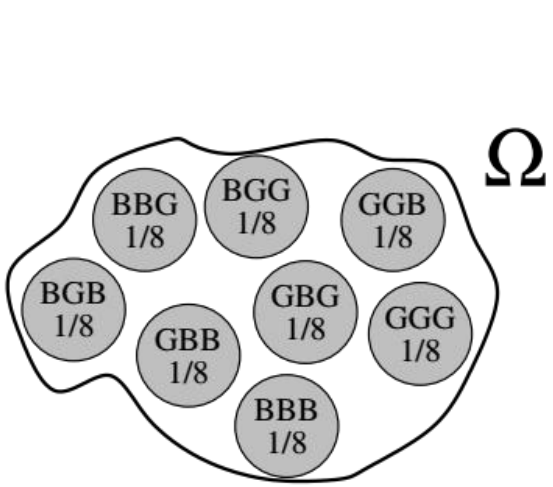
Now think of a different way of picturing Ω_2 that is easier to extend:

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?

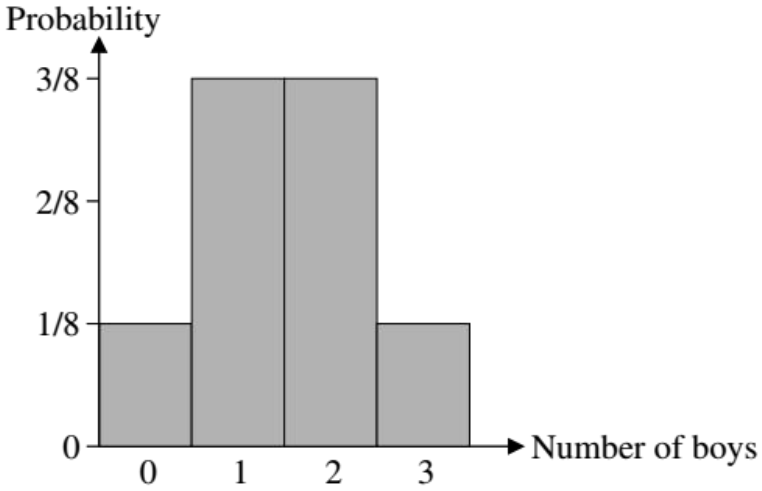


Examples of probability distributions

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:*

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

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

Random variables

Example:

Random experiment: Toss a coin once.
Sample space: $\Omega = \{\text{head}, \text{tail}\}$.
An example of a random variable: $X : \Omega \to \mathbb{R}$ maps "head" $\to 1$, "tail" $\to 0$.

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.

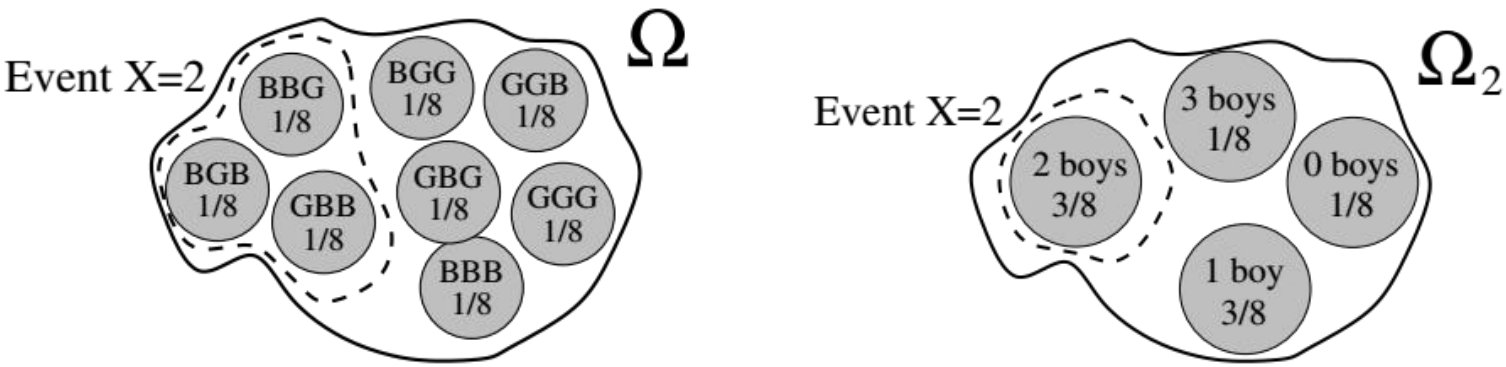
- x (lower-case letter) is a ***real number*** like 2 or 3. It is used to indicate an unspecified value that X might take.

- $\{X = x\}$, 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	**Wrong**
$X = 2 \cup X = 3$ *Event that X takes the value 2 OR 3*	$X = 2 + X = 3$
$\mathbb{P}(X = 2 \cup X = 3)$ *Probability of the event that X is 2 OR 3*	$\mathbb{P}(X = 2) \cup \mathbb{P}(X = 3)$
$\mathbb{P}(X = 2) + \mathbb{P}(X = 3)$ *Probability of the event that X is 2 OR 3*	$\mathbb{P}(X = 2 + X = 3)$
$X \leq 2 \cap X > 1$ *Event that X takes a value that is BOTH ≤ 2 AND > 1* *(the value $X = 2$ is the only possibility)*	$X \leq 2 \times X > 1$
$\mathbb{P}(X \leq 2 \cap X > 1)$ *Probability that X is BOTH ≤ 2 AND > 1*	$\mathbb{P}(X \leq 2) \cap \mathbb{P}(X > 1)$

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

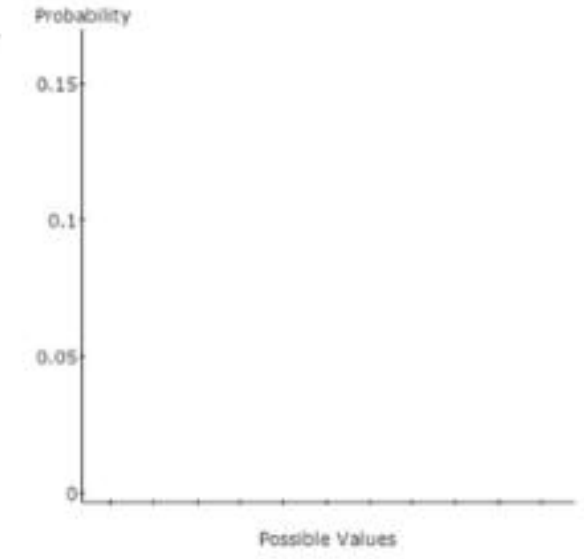
X	1	2	3	4	5	6
P(X=x)	1/6	1/6	1/6	1/6	1/6	1/6

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)

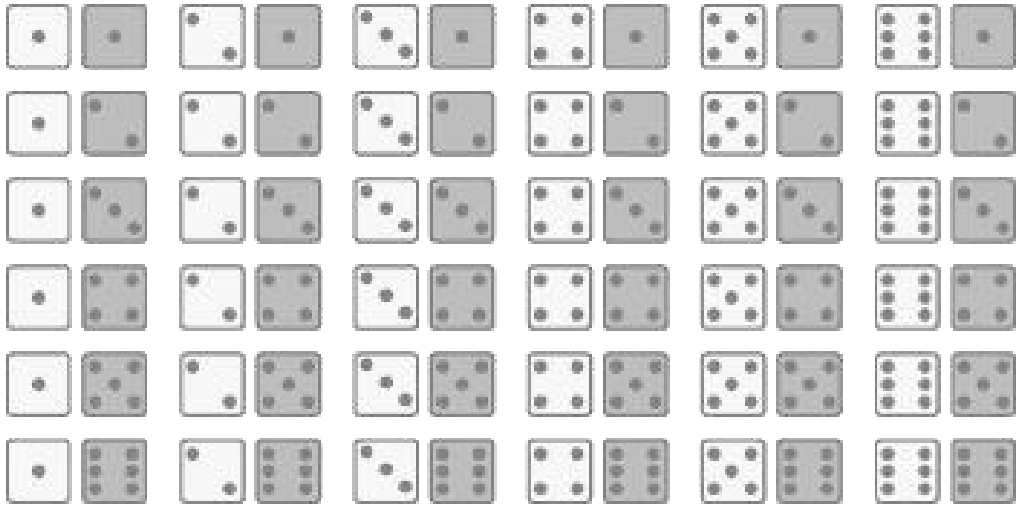
a. Let X = the sum of the two rolls. Fill in the table below with the possible values of the random variable and the probability of each possible value. Then use this information to construct a graph of the probability distribution where the x-axis represents the possible values of X and the y-axis represents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

Possible values of X	$P(x)$



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'; $p = \mathbb{P}(\textit{Head}) = 0.5$.
(2) Repeated rolls of a fair die: $p = \mathbb{P}(\textit{Get a 6}) = 1/6$.

Bernoulli trials and the Binomial Distribution

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

Note: Saying the trials are ***independent*** means that *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* $Y \sim$ *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:

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

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);* $p = \mathbb{P}(\textit{Boy}) = 0.5$ *for each child.*
- X is the number of 6's obtained in 10 rolls of a die: $n = 10$ *trials (die rolls);* $p = \mathbb{P}(\textit{Get a 6}) = 1/6$ *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 $X \sim \textit{Binomial}(n, p)$, *or* $X \sim \textit{Bin}(n, p)$.

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

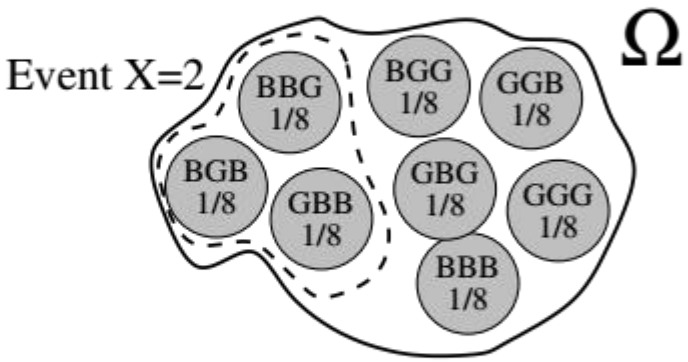
If $X \sim \text{Binomial}(n, p)$, then $X = x$ if there are x successes in the n trials. We don't care what order the successes occur in — in other words, we don't care *which* of the trials are successes and which are failures. However, we do have to bear in mind all the different orderings when we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

Binomial Distribution

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

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:

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

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* <u>one</u> *ordering.*

Question: What is the probability of **<u>one</u>** 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:

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

Note:

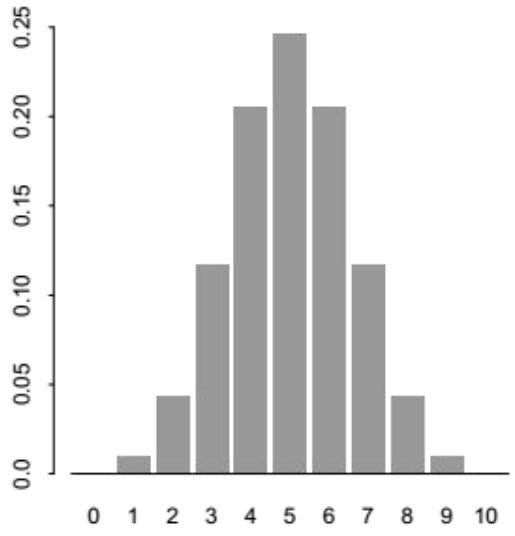
1. Importantly, $f_X(x) = 0$ *if* x *is not one of the values* $0, 1, \ldots, n$.
   The correct way to write the range of values is $x = 0, \ldots, n$.
   — Writing $x \in [0, n]$ is **wrong,** because this includes decimals like 0.4.
   — Writing $x = 0, 1, \ldots$ is **wrong,** because the range of values must stop at n: you can't have more than n successes in n trials.
2. $f_X(x)$ means, *'the probability function belonging to the r.v. I've named* X*'.*
   Use a capital X in the subscript and a lower-case x as the argument.

Shape of the Binomial distribution

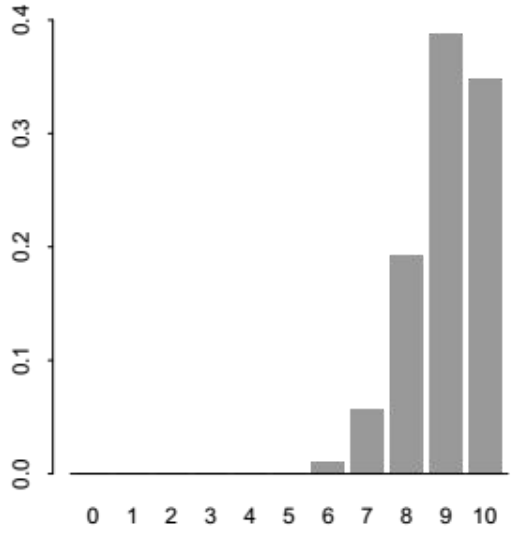
The shape of the Binomial distribution depends upon the values of n and p. For small n, the distribution is almost symmetrical for values of p close to 0.5, but highly skewed for values of p close to 0 or 1. As n increases, the distribution becomes more and more symmetrical, and there is noticeable skew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

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

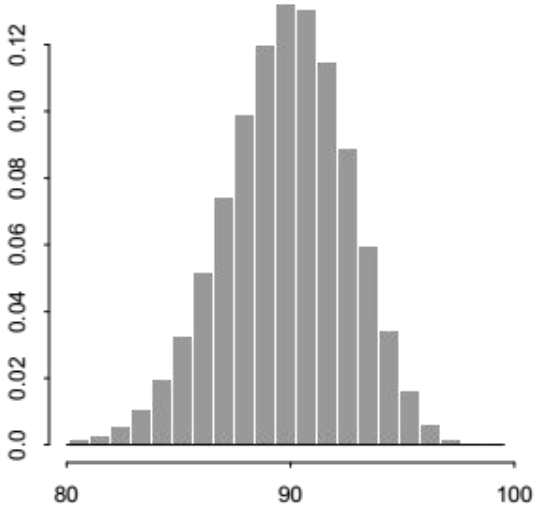
$n = 10, p = 0.5$



$n = 10, p = 0.9$



$n = 100, p = 0.9$



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

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

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

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

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 Binomial(n, p) random variable as the sum of n Bernoulli(p) random variables. If $Y_i \sim$ Bernoulli(p) for $i = 1, 2, \ldots, n$, and if $Y_1, Y_2, \ldots, Y_n$ are independent, then:

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

This is because X and $Y_1 + \ldots + Y_n$ both represent *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*** $f_X(x) = \mathbb{P}(X = x)$.

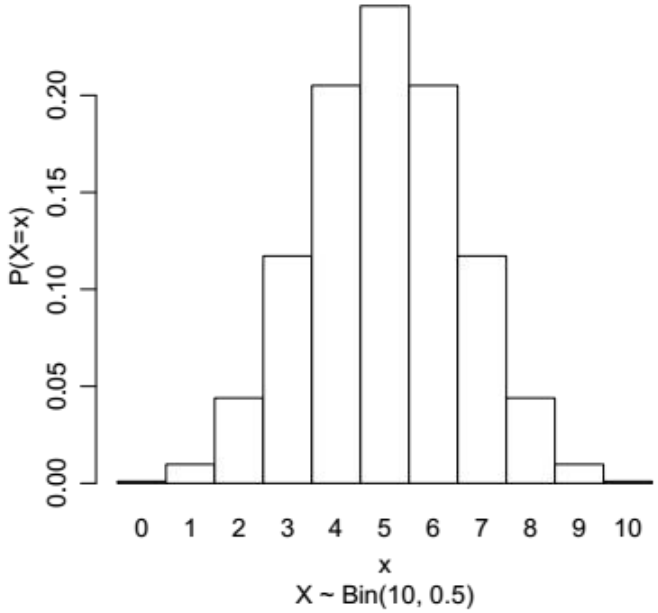
Another function that is widely used is the ***cumulative distribution function***, or CDF, written as $F_X(x)$.

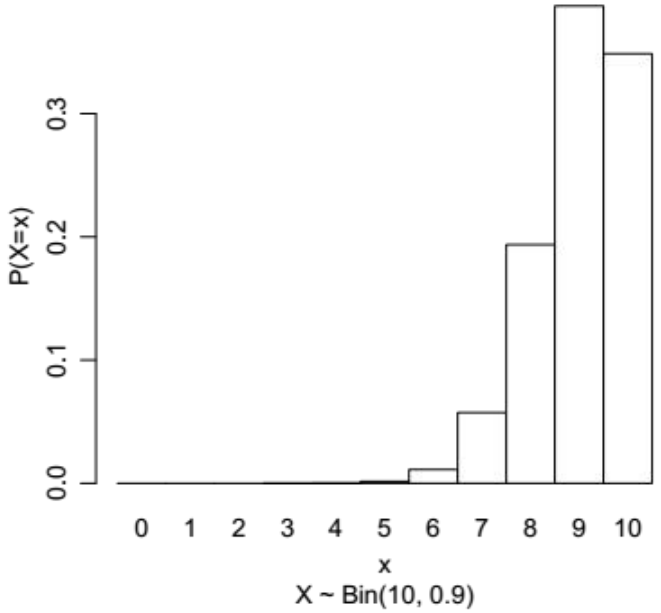
Definition: ***The cumulative distribution function, or CDF, is***

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

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.

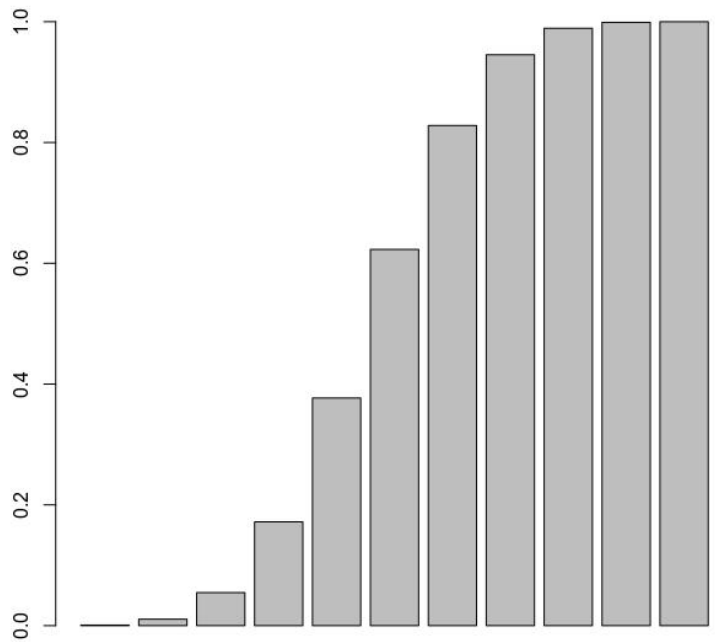




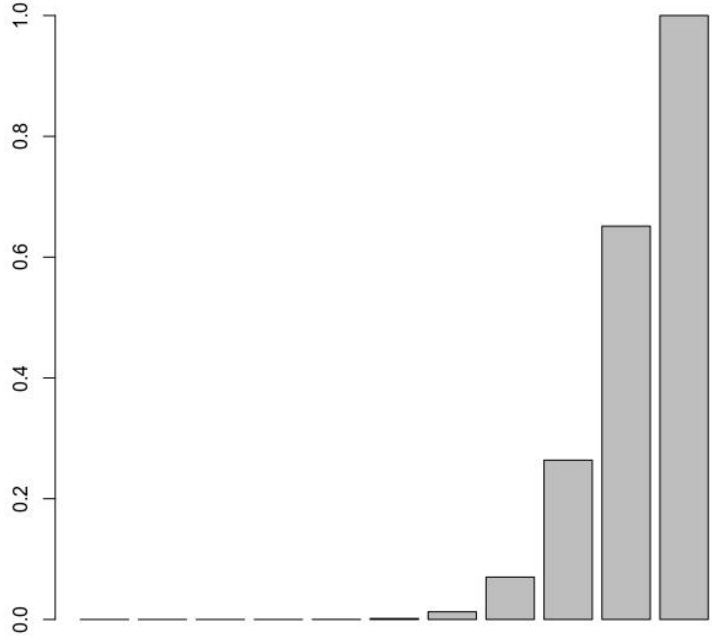
Clearly,

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

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



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



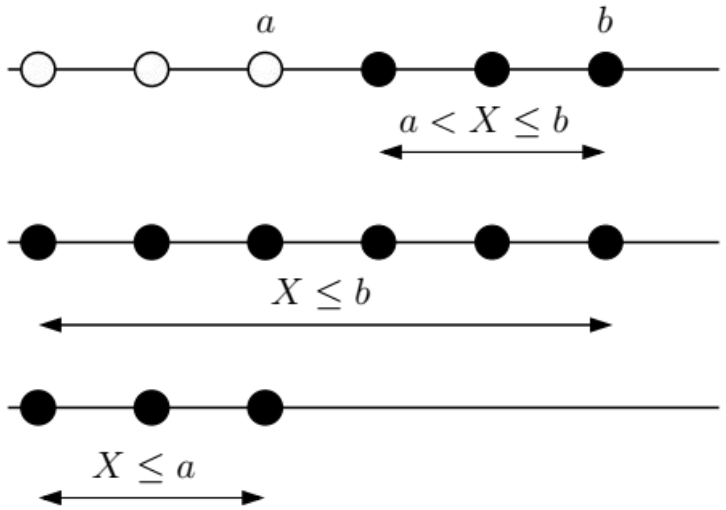
barplot(pbinom(0:10, size=10, prob=0.9))

Using the cumulative distribution function to find probabilities

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

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

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



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. $\mathbb{P}(X \leq 30)$? $F_X(30)$.
2. $\mathbb{P}(X < 30)$? $F_X(29)$.
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. $\mathbb{P}(50 \leq X \leq 60)$?

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

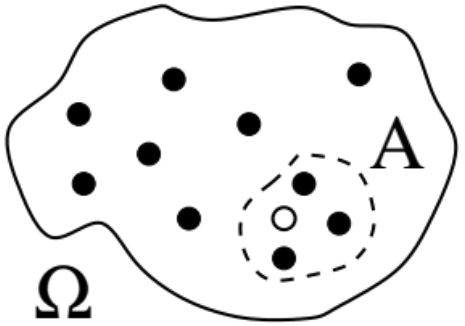
Conditional probability

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?



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

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?

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

We use a shorthand notation to write this down:

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

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* $\mathbb{P}(W \backslash A)$*: 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.

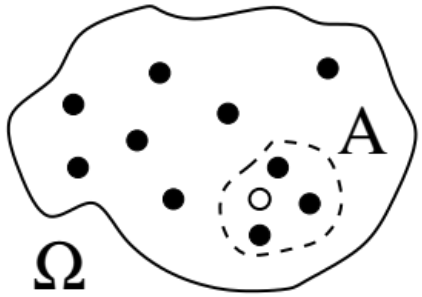
Conditional probability, an example

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

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

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

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

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

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

The conditional probability $\mathbb{P}(W \mid A)$ means ***the probability of event W, when selecting only from within set A.***

Read it as 'probability of event W, given event A',
or 'probability of event W ***from within the set A***.'

It is equivalent to ***changing the sample space from Ω to A.***

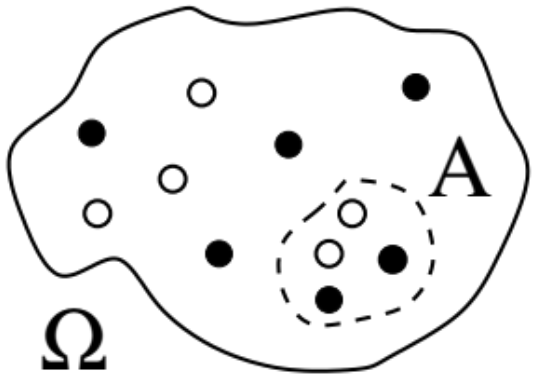
The notation $\mathbb{P}(W \mid A)$ is like saying,
'$\mathbb{P}(W)$ when my symbol $\mathbb{P}$ is defined relative to A instead of to Ω.'

Formula for conditional probability

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

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

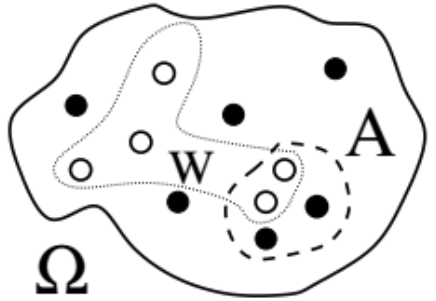
Answer: $\mathbb{P}(W)$ refers to the probability within the whole sample space Ω, so $\mathbb{P}(W) = \frac{5}{11}$.



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

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

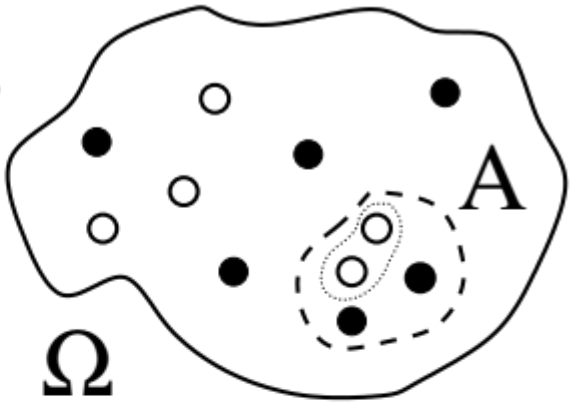
Question: Can you see why $\mathbb{P}(W \mid A) = \dfrac{\mathbb{P}(W \cap A)}{\mathbb{P}(A)}$?



Formula for conditional probability

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

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



The small dotted bag represents the set $W \cap A$.

The dashed bag represents the set A.

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

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

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 $\mathbb{P}(B \mid A)$, and defined as

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

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

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

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

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

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

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

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

Language of conditional probability

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

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

- $\mathbb{P}(B)$ means we want the probability of picking someone who satisfies B *when they are picked from the whole sample space, Ω.*
- $\mathbb{P}(B \mid A)$ 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: $F = \{\text{first years}\}$; $S = \{\text{second years}\}$; $T = \{\text{third years}\}$; $O = \{\text{other students, e.g. exchange students, COPs, } \ldots\}$.

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

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

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

Answer: *Replace A by Ω: this gives*

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

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

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

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,

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

Proof: *Immediate from the definitions:*

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

and $\mathbb{P}(B \mid A) = \dfrac{\mathbb{P}(B \cap A)}{\mathbb{P}(A)} \quad \Rightarrow \quad \mathbb{P}(B \cap A) = \mathbb{P}(A \cap B) = \mathbb{P}(B \mid A)\mathbb{P}(A).$ $\square$

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

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

Using the multiplication rule:

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

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

Statistical independence

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

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

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

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

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

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

Independence for random variables

Random variables are independent if *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:

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

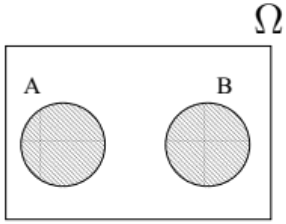
Thus X *and* Y *are independent if and only if*

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

Independence in pictures

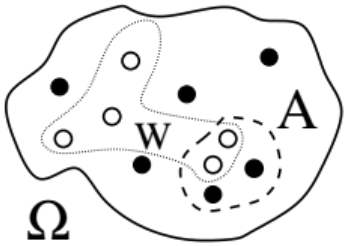
It is very difficult to draw a picture of statistical independence.

Are events A and B statistically independent?



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.
$\mathbb{P}(W \mid A) = 2/4$, but $\mathbb{P}(W) = 5/11$.
So $\mathbb{P}(W \mid A) \neq \mathbb{P}(W)$, 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 $\mathbb{P}(B \mid A) = \mathbb{P}(B)$,
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 $\mathbb{P}(B \mid A)$ in terms of $\mathbb{P}(A \mid B)$.

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

Apply the multiplication rule to each side:

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

Thus

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

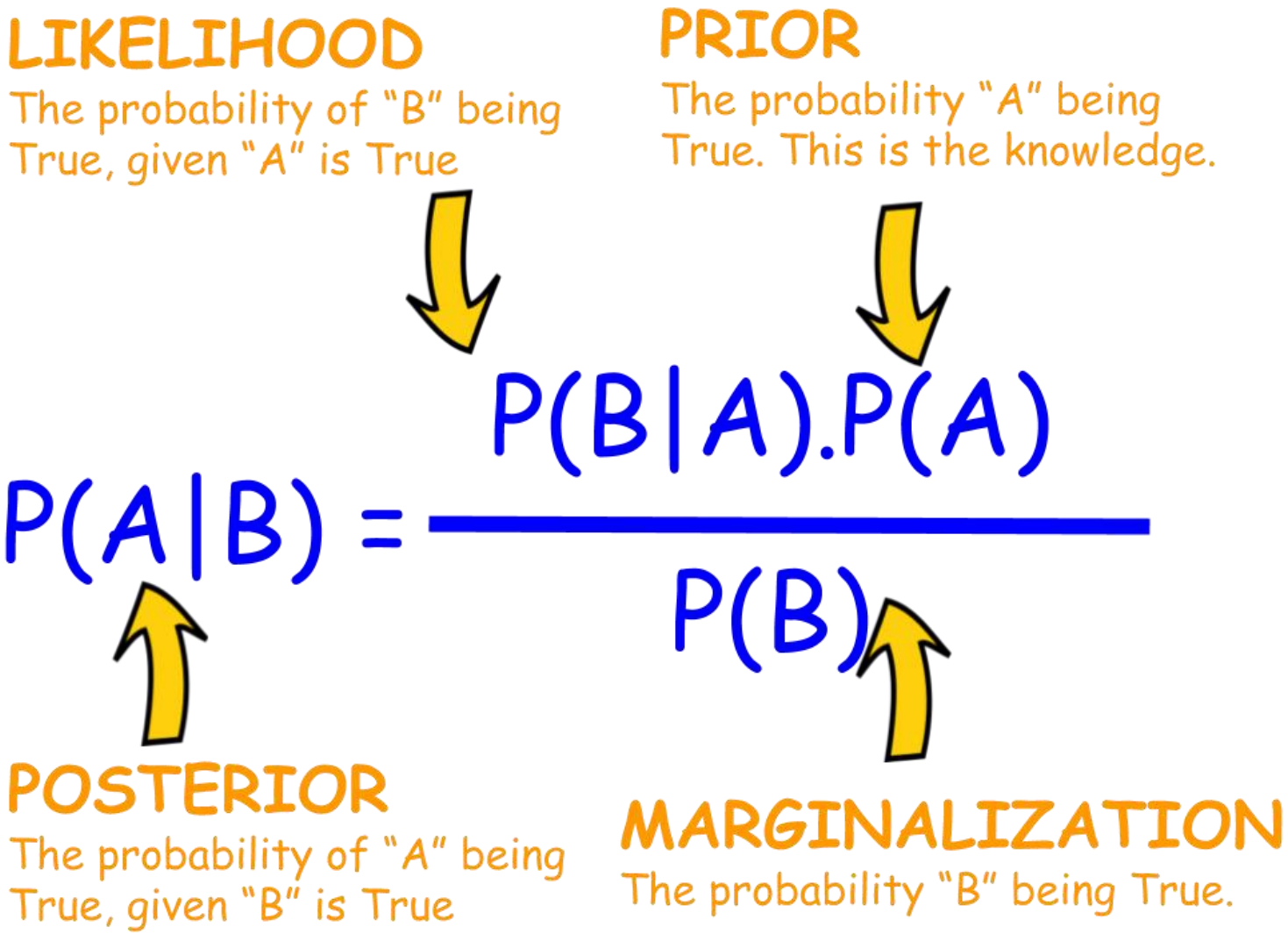
Rev. Thomas Bayes (1702–1761), English clergyman and founder of Bayesian Statistics.

Example

- 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?

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

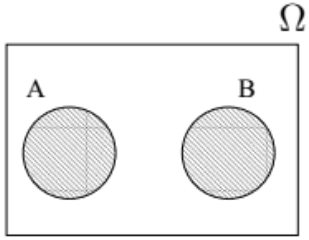
Bayes' Theorem



The Partition Theorem (Law of Total Probability)

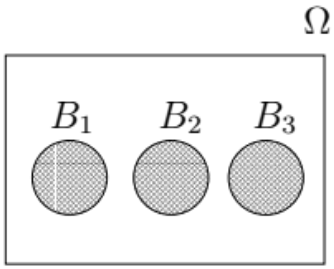
Definition: Events A and B are **mutually exclusive**, or **disjoint**, if $A \cap B = \emptyset$.

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



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

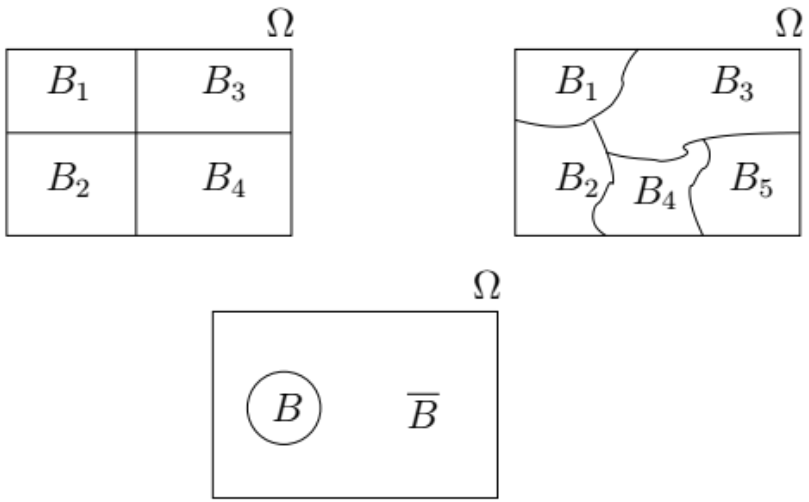
$$B_i \cap B_j = \emptyset \text{ for all } i, j \text{ with } i \neq j,$$

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

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

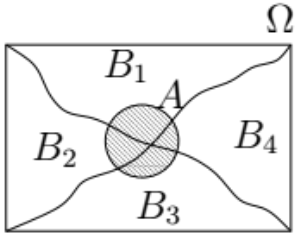
The Partition Theorem (Law of Total Probability)

Examples:



Partitioning an event A

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



The Partition Theorem (Law of Total Probability)

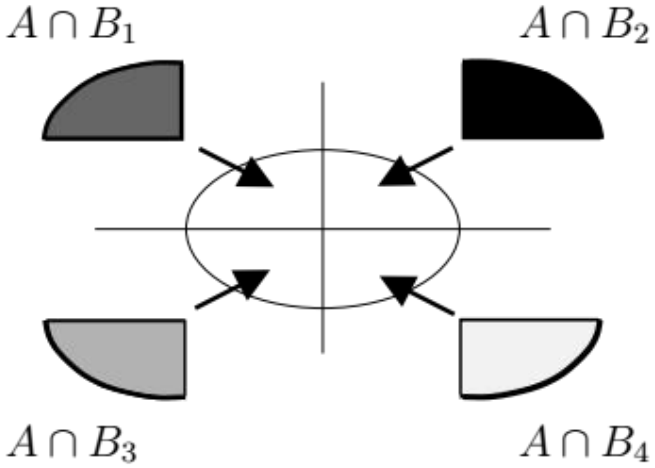
Let $B_1, \ldots, B_m$ form a partition of Ω. Then for any event A,

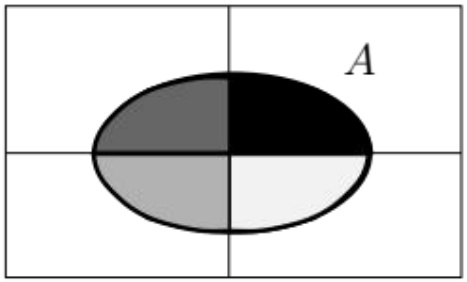
$$\mathbb{P}(A) = \sum_{i=1}^{m} \mathbb{P}(A \cap B_i) = \sum_{i=1}^{m} \mathbb{P}(A \mid B_i)\mathbb{P}(B_i)$$

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

The Partition Theorem in pictures

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





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

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:

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

From the information given, the events have probabilities:

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

(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 $C = \{$ *crowded* $\}$, $N = \{$ *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.

$$\mathbb{P}(C \mid T) = 0.5; \qquad \mathbb{P}(C \mid L) = 0.7.$$
$$\mathbb{P}(N \mid C) = 0.8; \qquad \mathbb{P}(N \mid \overline{C}) = 0.4.$$

Examples of conditional probability and partitions

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

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

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

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

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 $\mathbb{P}(B \mid I) = 0.4$. Each sentence should have a different wording or sentence structure, but each sentence should unambiguously convey that $\mathbb{P}(B \mid I) = 0.4$. 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 $\mathbb{P}(B \mid I) = 0.4$. 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 $\mathbb{P}(B \mid I) = 0.4$.

(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 $I_1, \ldots, I_n$ are mutually exclusive and $\mathbb{P}(B \mid I_r) = k$ (a constant) for all $r = 1, \ldots, n$, then $\mathbb{P}(B \mid I_1 \cup I_2 \cup \ldots \cup I_n) = k$ 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.

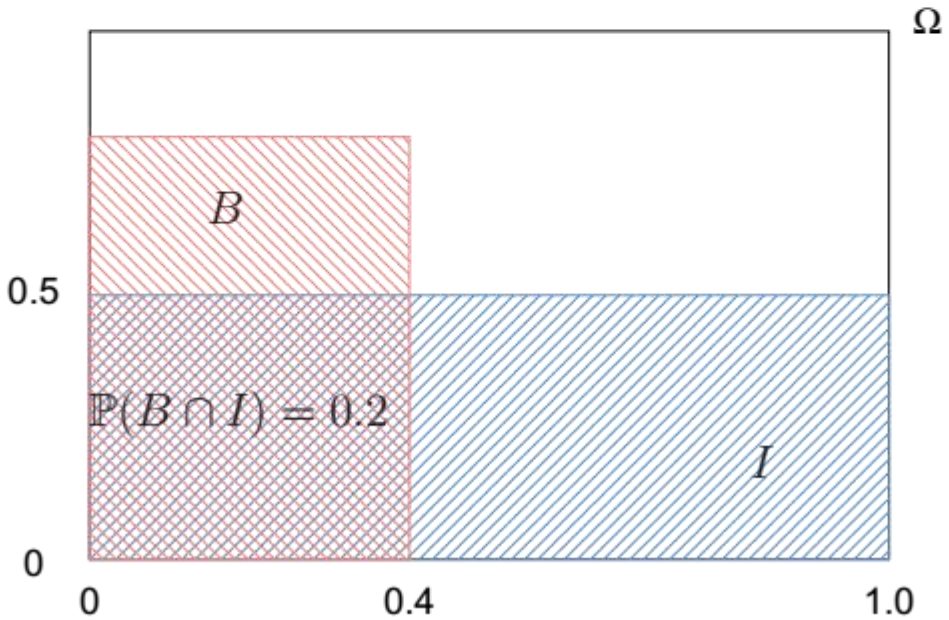
Class task conditional probability and partitions

(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

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

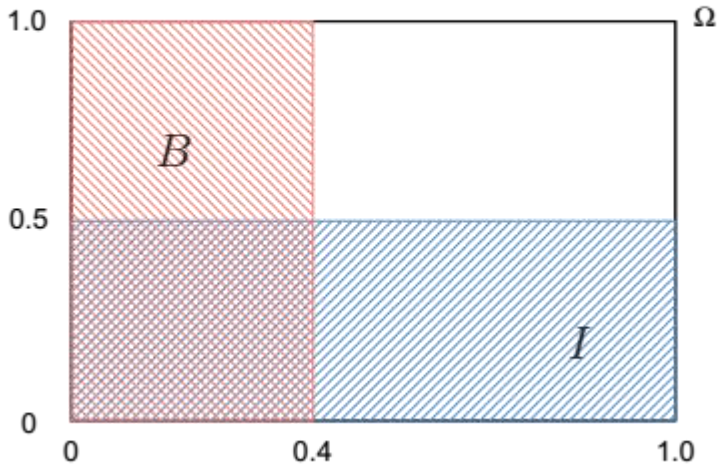
However, we cannot calculate $\mathbb{P}(B)$ from the information given, so B can be any height in the diagram below.



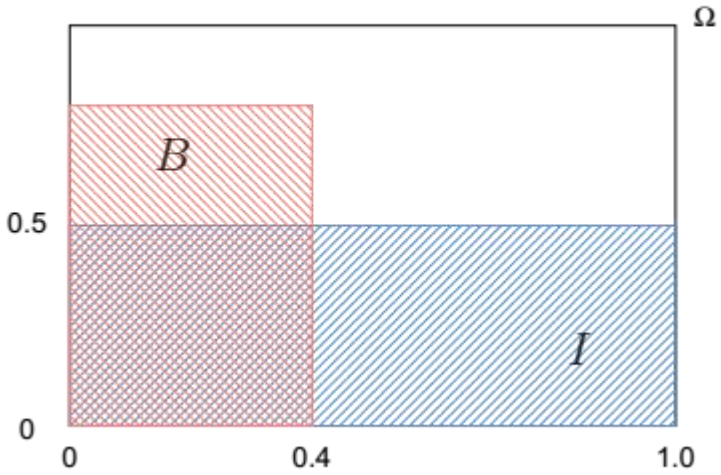
$\mathbb{P}(B \mid I) = 0.4$

Class task conditional probability and partitions

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



(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:

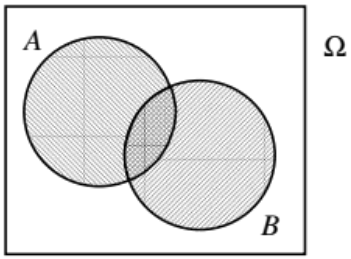
$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B).$$

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

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

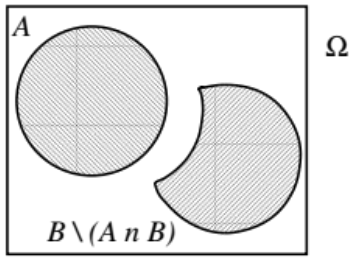
Explanation

To understand the formula, think of the Venn diagrams:



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

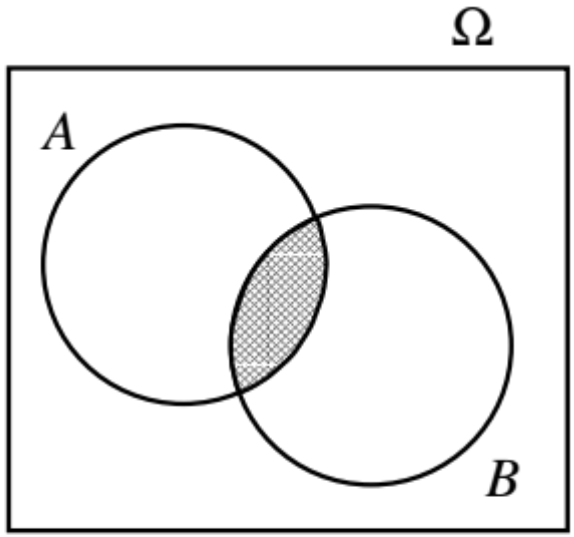
So we have to *subtract the intersection once to get* $\mathbb{P}(A \cup B)$*:* $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.



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

Probability of an intersection

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



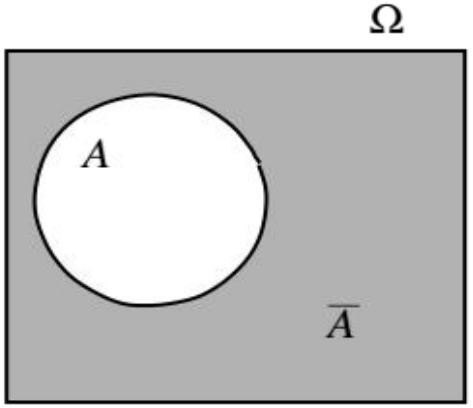
We might be able to use ***statistical independence:***
if A and B are independent, then
$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$.

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

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

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

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

Examples of basic probability calculations, example 1

Next write down all the information given:

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

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

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

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

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

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

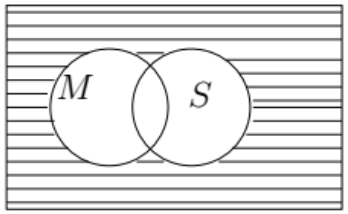
(a)

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

(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.

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

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

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

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

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

Probability Reference List (Keep in mind!!!)

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

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

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

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

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

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

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

- **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 $A_1, \ldots, A_m$ partition the sample space Ω, then

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

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

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

Exercises

(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:

$$\mathbb{P}(\text{patient tests positive} \,|\, \text{patient has HIV}) = 0.98 \tag{1}$$

$$\mathbb{P}(\text{patient tests positive} \,|\, \text{patient does not have HIV}) = 0.07 \tag{2}$$

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

$$\mathbb{P}(\text{patient has HIV} \,|\, \text{patient tests positive}). \tag{3}$$

(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)).

Exercises

(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 $\mathbb{P}(\text{patient has HIV} \,|\, \text{patient tests positive})$. 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.

Expectation and variance of a random variable

- The ***expectation*** of a random variable is the value it takes ***on average.***
- The ***variance*** of a random variable measures how much the random variable ***varies about its average.***

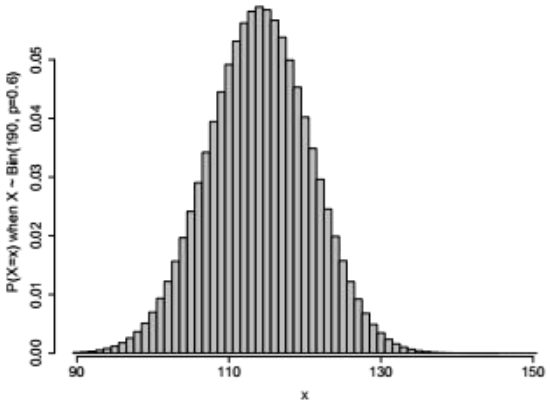
Expectation

Given a random variable X that measures something, we often want to know *what is the* **average** *value of* X*?*

For example, here are 30 random observations taken from the distribution $X \sim \text{Binomial}(n = 190, p = 0.6)$:

***R* command:** `rbinom(30, 190, 0.6)`

```
116 116 117 122 111 112 114 120 112 102
125 116  97 105 108 117 118 111 116 121
107 113 120 114 114 124 116 118 119 120
```



The average, or *mean*, of the ***first ten*** values is:

$$\frac{116 + 116 + \ldots + 112 + 102}{10} = 114.2.$$

Expectation of a random variable

The mean of the ***first twenty*** values is:

$$\frac{116 + 116 + \ldots + 116 + 121}{20} = 113.8.$$

The mean of the ***first thirty*** values is:

$$\frac{116 + 116 + \ldots + 119 + 120}{30} = 114.7.$$

The answers all seem to be close to *114*. What would happen if we took the average of hundreds of values?

100 values from Binomial(190, 0.6):

R command: `mean(rbinom(100, 190, 0.6))`
Result: `114.86`

Note: You will get a different result every time you run this command.

Expectation of a random variable

1000 values from Binomial(190, 0.6):

R command: `mean(rbinom(1000, 190, 0.6))`
Result: `114.02`

1 million values from Binomial(190, 0.6):

R command: `mean(rbinom(1000000, 190, 0.6))`
Result: `114.0001`

The average seems to be *converging to the value 114.*

The larger the sample size, *the closer the average seems to get to 114.*

If we kept going for larger and larger sample sizes, we would keep getting answers closer and closer to 114. This is because *114 is the DISTRIBUTION MEAN: the mean value that we would get if we were able to draw an infinite sample from the Binomial(190, 0.6) distribution.*

This distribution mean is called the *expectation, or expected value, of the Binomial(190, 0.6) distribution.*

It is a *FIXED property of the Binomial(190, 0.6) distribution.* This means it is a *fixed constant: there is nothing random about it.*

Expectation of a random variable

Definition: The **expected value**, also called the **expectation** or **mean**, of a discrete random variable X, *can be written as either* $\mathbb{E}(X)$, *or E(X), or* μ_X, *and is given by*

$$\mu_X = \mathbb{E}(X) = \sum_x x f_X(x) = \sum_x x \mathbb{P}(X = x)\,.$$

The expected value is a measure of the centre, or average, of the set of values that X can take, weighted according to the probability of each value.

If we took a very large sample of random numbers from the distribution of X, *their average would be approximately equal to* μ_X.

Expectation of a random variable

Example: Let $X \sim \text{Binomial}(n = 190, p = 0.6)$. What is $\mathbb{E}(X)$?

$$\begin{aligned}\mathbb{E}(X) &= \sum_x x\mathbb{P}(X = x) \\ &= \sum_{x=0}^{190} x\binom{190}{x}(0.6)^x(0.4)^{190-x}.\end{aligned}$$

Although it is not obvious, the answer to this sum is $n \times p = 190 \times 0.6 = 114$. We will see why in Section 2.10.

Explanation of the formula for expectation

We will move away from the Binomial distribution for a moment, and use a simpler example.

Let the random variable X be defined as $X = \begin{cases} 1 & \text{with probability } 0.9, \\ -1 & \text{with probability } 0.1. \end{cases}$

X takes only the values 1 and -1. What is the 'average' value of X?

Expectation of a random variable

Using $\frac{1+(-1)}{2} = 0$ *would not be useful, because it ignores the fact that usually* $X = 1$, *and only occasionally is* $X = -1$.

Instead, think of observing X many times, say 100 times.

Roughly *90* of these 100 times will have $X = 1$.
Roughly *10* of these 100 times will have $X = -1$

The average of the 100 values will be roughly

$$\frac{90 \times 1 + 10 \times (-1)}{100},$$

$$= 0.9 \times 1 + 0.1 \times (-1)$$

$$(= 0.8.)$$

We could repeat this for any sample size.

Expectation of a random variable

As the sample gets large, the average of the sample will get ever closer to

$$0.9 \times 1 + 0.1 \times (-1).$$

This is why the distribution mean is given by

$$\mathbb{E}(X) = \mathbb{P}(X = 1) \times 1 + \mathbb{P}(X = -1) \times (-1),$$

or in general,

$$\mathbb{E}(X) = \sum_x \mathbb{P}(X = x) \times x.$$

> $\mathbb{E}(X)$ *is a fixed constant giving the average value we would get from a large sample of* X.

Linear property of expectation

Expectation is a ***linear*** operator:

Theorem 2.7: *Let a and b be constants. Then*

$$\mathbb{E}(aX + b) = a\mathbb{E}(X) + b.$$

Proof:

Immediate from the definition of expectation.

$$\begin{aligned}
\mathbb{E}(aX + b) &= \sum_x (ax + b) f_X(x) \\
&= a \sum_x x f_X(x) + b \sum_x f_X(x) \\
&= a\,\mathbb{E}(X) + b \times 1. \qquad \square
\end{aligned}$$

Example 1: finding expectation from the probability function

Example 1: Let $X \sim \text{Binomial}(3, 0.2)$. Write down the probability function of X and find $\mathbb{E}(X)$.

We have:

$$\mathbb{P}(X = x) = \binom{3}{x}(0.2)^x(0.8)^{3-x} \text{ for } x = 0, 1, 2, 3.$$

x	*0*	*1*	*2*	*3*
$f_X(x) = \mathbb{P}(X = x)$	0.512	0.384	0.096	0.008

Then

$$\begin{aligned}\mathbb{E}(X) = \sum_{x=0}^{3} x f_X(x) &= 0 \times 0.512 + 1 \times 0.384 + 2 \times 0.096 + 3 \times 0.008 \\ &= 0.6.\end{aligned}$$

Note: We have: $\mathbb{E}(X) = 0.6 = 3 \times 0.2$ *for* $X \sim$ ***Binomial***$(3, 0.2)$.
We will prove in Section 2.10 that whenever $X \sim \text{Binomial}(n, p)$, then $\mathbb{E}(X) = np$.

Example 2: finding expectation from the probability function

Example 2: Let Y be Bernoulli(p) (Section 1.2). That is,

$$Y = \begin{cases} 1 & \textit{with probability } p, \\ 0 & \textit{with probability } 1-p. \end{cases}$$

Find $\mathbb{E}(Y)$.

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

$$\mathbb{E}(Y) = 0 \times (1-p) + 1 \times p = p.$$

Expectation of a sum of random variables: E(X + Y)

For ANY random variables $X_1, X_2, \ldots, X_n$*,*

$$\boxed{\mathbb{E}\,(X_1 + X_2 + \ldots + X_n) = \mathbb{E}(X_1) + \mathbb{E}(X_2) + \ldots + \mathbb{E}(X_n).}$$

In particular, $\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$ *for ANY* X *and* Y*.*

This result holds for ***any*** random variables $X_1, \ldots, X_n$. *It does NOT require* $X_1, \ldots, X_n$ *to be independent.*

We can summarize this important result by saying:

The expectation of a sum
is the sum of the expectations – ALWAYS.

The proof requires multivariate methods, to be studied in later courses.

Note: We can combine the result above with the linear property of expectation. For any constants $a_1, \ldots, a_n$, we have:

$$\mathbb{E}\,(a_1X_1 + a_2X_2 + \ldots + a_nX_n) = a_1\mathbb{E}(X_1) + a_2\mathbb{E}(X_2) + \ldots + a_n\mathbb{E}(X_n).$$

Expectation of a product of random variables: E(XY)

There are two cases when finding the expectation of a product:

1. **General case:**

For general X and Y, $\quad \mathbb{E}(XY)$ *is NOT equal to* $\mathbb{E}(X)\mathbb{E}(Y)$.

   We have to find $\mathbb{E}(XY)$ either using their joint probability function (see later), or using their covariance (see later).

2. **Special case:** when X and Y are *INDEPENDENT:*

When X and Y are INDEPENDENT, $\quad \mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$.

Variable transformations

We often wish to *transform* random variables through a function. For example, given the random variable X, possible transformations of X include:

$$X^2, \qquad \sqrt{X}, \qquad 4X^3, \qquad \ldots$$

We often summarize all possible variable transformations by referring to $Y = g(X)$ *for some function* g.

For discrete random variables, it is very easy to find the probability function for $Y = g(X)$, given that the probability function for X is known. Simply ***change all the values and keep the probabilities the same.***

Example 1: Let $X \sim \text{Binomial}(3, 0.2)$, and let $Y = X^2$. Find the probability function of Y.

The probability function for X is:

x	0	1	2	3
$\mathbb{P}(X = x)$	0.512	0.384	0.096	0.008

Thus *the probability function for* $Y = X^2$ *is:*

y	0^2	1^2	2^2	3^2
$\mathbb{P}(Y = y)$	0.512	0.384	0.096	0.008

This is because Y *takes the value* 0^2 *whenever* X *takes the value* 0, *and so on.*

Variable transformations

Thus the probability that $Y = 0^2$ is *the same as the probability that* $X = 0$.

Overall, we would write the probability function of $Y = X^2$ as:

y	0	1	4	9
$\mathbb{P}(Y = y)$	0.512	0.384	0.096	0.008

To transform a discrete random variable, *transform the values and leave the probabilities alone.*

Variable transformations

Example 2: Mr Chance hires out giant helium balloons for advertising. His balloons come in three sizes: heights 2m, 3m, and 4m. 50% of Mr Chance's customers choose to hire the cheapest 2m balloon, while 30% hire the 3m balloon and 20% hire the 4m balloon.

The amount of helium gas in cubic metres required to fill the balloons is $h^3/2$, where h is the height of the balloon. Find the probability function of Y, the amount of helium gas required for a randomly chosen customer.

Let X be the height of balloon ordered by a random customer. The probability function of X is:

height, x (m)	*2*	*3*	*4*
$\mathbb{P}(X = x)$	*0.5*	*0.3*	*0.2*

Let Y be the amount of gas required: $Y = X^3/2$.
The probability function of Y is:

gas, y (m^3)	*4*	*13.5*	*32*
$\mathbb{P}(Y = y)$	*0.5*	*0.3*	*0.2*

Expected value of a transformed random variable

We can find the expectation of a transformed random variable just like any other random variable. For example, in Example 1 we had $X \sim \text{Binomial}(3, 0.2)$, and $Y = X^2$.

The probability function for X is:

x	0	1	2	3
$\mathbb{P}(X = x)$	0.512	0.384	0.096	0.008

and for $Y = X^2$:

y	0	1	4	9
$\mathbb{P}(Y = y)$	0.512	0.384	0.096	0.008

Thus the expectation of $Y = X^2$ is:

$$\begin{aligned} \mathbb{E}(Y) = \mathbb{E}(X^2) &= 0 \times 0.512 + 1 \times 0.384 + 4 \times 0.096 + 9 \times 0.008 \\ &= 0.84. \end{aligned}$$

Note: $\mathbb{E}(X^2)$ *is NOT the same as* $\{\mathbb{E}(X)\}^2$. *Check that* $\{\mathbb{E}(X)\}^2 = 0.36$.

Expected value of a transformed random variable

To make the calculation quicker, we could cut out the middle step of writing down the probability function of Y. Because we transform the values and keep the probabilities the same, we have:

$$\mathbb{E}(X^2) = 0^2 \times 0.512 + 1^2 \times 0.384 + 2^2 \times 0.096 + 3^2 \times 0.008.$$

If we write $g(X) = X^2$, this becomes:

$$\mathbb{E}\{g(X)\} = \mathbb{E}(X^2) = g(0) \times 0.512 + g(1) \times 0.384 + g(2) \times 0.096 + g(3) \times 0.008.$$

Clearly the same arguments can be extended to any function $g(X)$ and any discrete random variable X:

$$\boxed{\mathbb{E}\{g(X)\} = \sum_x g(x)\mathbb{P}(X = x).}$$

Expected value of a transformed random variable

Definition: For any function g and discrete random variable X, the expected value of $g(X)$ is given by

$$\mathbb{E}\{g(X)\} = \sum_x g(x)\mathbb{P}(X = x) = \sum_x g(x) f_X(x).$$

Example: Recall Mr Chance and his balloon-hire business from page 74. Let X be the height of balloon selected by a randomly chosen customer. The probability function of X is:

height, x (m)	2	3	4
$\mathbb{P}(X = x)$	0.5	0.3	0.2

(a) What is the average amount of gas required per customer?

Expected value of a transformed random variable

Gas required was $X^3/2$ from page 74.
Average gas per customer is $\mathbb{E}(X^3/2)$.

$$\begin{aligned}
\mathbb{E}\left(\frac{X^3}{2}\right) &= \sum_x \frac{x^3}{2} \times \mathbb{P}(X = x) \\
&= \frac{2^3}{2} \times 0.5 + \frac{3^3}{2} \times 0.3 + \frac{4^3}{2} \times 0.2 \\
&= 12.45 \ \textit{m}^3 \textit{ gas.}
\end{aligned}$$

(b) Mr Chance charges \$400$\times h$ to hire a balloon of height h. What is his expected earning per customer?

Expected value of a transformed random variable

Expected earning is $\mathbb{E}(400X)$.

$$\begin{aligned}
\mathbb{E}(400X) &= 400 \times \mathbb{E}(X) \qquad \textit{(expectation is linear)} \\
&= 400 \times (2 \times 0.5 + 3 \times 0.3 + 4 \times 0.2) \\
&= 400 \times 2.7 \\
&= \$1080 \textit{ per customer.}
\end{aligned}$$

(c) How much does Mr Chance expect to earn in total from his next 5 customers?

Expected value of a transformed random variable

Let $Z_1, \ldots, Z_5$ be the earnings from the next 5 customers. Each Z_i has $\mathbb{E}(Z_i) =$ 1080 by part (b). The total expected earning is

$$\begin{aligned} \mathbb{E}(Z_1 + Z_2 + \ldots + Z_5) &= \mathbb{E}(Z_1) + \mathbb{E}(Z_2) + \ldots + \mathbb{E}(Z_5) \\ &= 5 \times 1080 \\ &= \$5400. \end{aligned}$$

Expected value of a transformed random variable

Suppose $X = \begin{cases} 3 & \text{with probability } 3/4, \\ 8 & \text{with probability } 1/4. \end{cases}$

Then $3/4$ of the time, X takes value 3, and $1/4$ of the time, X takes value 8.

So $\mathbb{E}(X) = \frac{3}{4} \times 3 + \frac{1}{4} \times 8.$

ADD UP THE VALUES
TIMES HOW OFTEN THEY OCCUR

Common mistakes in calculate expected value of a transformed random variable

i) $\mathbb{E}(\sqrt{X}) = \sqrt{\mathbb{E}X} = \sqrt{\frac{3}{4} \times 3 \quad + \quad \frac{1}{4} \times 8}$

ii) $\mathbb{E}(\sqrt{X}) = \sqrt{\frac{3}{4}} \times 3 \quad + \quad \sqrt{\frac{1}{4}} \times 8$

iii) $\mathbb{E}(\sqrt{X}) = \sqrt{\frac{3}{4} \times 3} \quad + \quad \sqrt{\frac{1}{4} \times 8}$

$= \sqrt{\frac{3}{4}} \times \sqrt{3} \quad + \quad \sqrt{\frac{1}{4}} \times \sqrt{8}$

Common mistakes in calculate expected value of a transformed random variable

What about $\mathbb{E}(\sqrt{X})$?

$$\sqrt{X} = \begin{cases} \sqrt{3} & \text{with probability } 3/4, \\ \sqrt{8} & \text{with probability } 1/4. \end{cases}$$

ADD UP THE VALUES
TIMES HOW OFTEN THEY OCCUR

$$\mathbb{E}(\sqrt{X}) = \frac{3}{4} \times \sqrt{3} + \frac{1}{4} \times \sqrt{8}.$$

Properties of Expectation

i) Let g and h be functions, and let a and b be constants. For any random variable X (discrete or continuous),

$$\mathbb{E}\Big\{ag(X) + bh(X)\Big\} = a\mathbb{E}\Big\{g(X)\Big\} + b\mathbb{E}\Big\{h(X)\Big\}.$$

In particular,

$$\mathbb{E}(aX + b) = a\mathbb{E}(X) + b.$$

ii) Let X and Y be ANY random variables (discrete, continuous, independent, or non-independent). Then

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

More generally, for ANY random variables $X_1, \ldots, X_n$,

$$\mathbb{E}(X_1 + \ldots + X_n) = \mathbb{E}(X_1) + \ldots + \mathbb{E}(X_n).$$

iii) Let X and Y be **independent** random variables, and g, h be functions. Then

$$\begin{aligned} \mathbb{E}(XY) &= \mathbb{E}(X)\mathbb{E}(Y) \\ \mathbb{E}\Big(g(X)h(Y)\Big) &= \mathbb{E}\Big(g(X)\Big)\mathbb{E}\Big(h(Y)\Big). \end{aligned}$$

Notes: 1. $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$ is ONLY generally true if X and Y are *INDEPENDENT.*

2. If X and Y are independent, then $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$. However, the converse is not generally true: it is possible for $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$ even though X and Y are dependent.

Variance

Example: Mrs Tractor runs the Rational Bank of Remuera. Every day she hopes to fill her cash machine with enough cash to see the well-heeled citizens of Remuera through the day. She knows that the expected amount of money withdrawn each day is \$50,000. How much money should she load in the machine? \$50,000?

No: \$50,000 is the average, near the centre of the distribution. About half the time, the money required will be GREATER than the average.

How much money should Mrs Tractor put in the machine if she wants to be 99% certain that there will be enough for the day's transactions?

Answer: it depends how much the amount withdrawn *varies above and below its mean.*

For questions like this, we need the study of *variance.*

Variance is the *average squared distance of a random variable from its own mean.*

Variance

Definition: The **variance** of a random variable X *is written as either* $\mathrm{Var}(X)$ *or* σ_X^2, *and is given by*

$$\sigma_X^2 = \mathrm{Var}(X) = \mathbb{E}\left[(X - \mu_X)^2\right] = \mathbb{E}\left[(X - \mathbb{E}X)^2\right].$$

Similarly, the variance of a function of X is

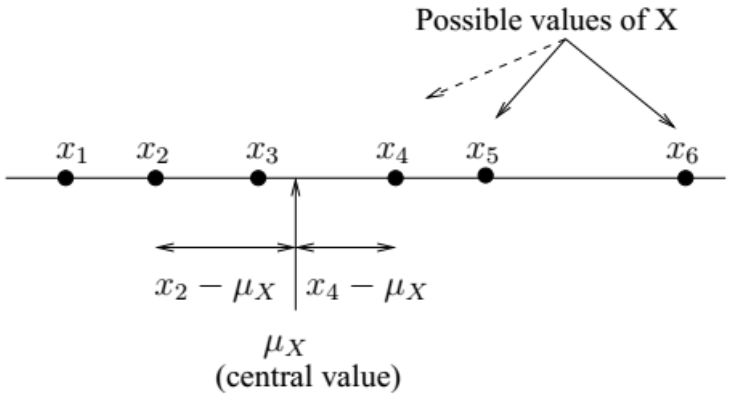
$$\mathrm{Var}(g(X)) = \mathbb{E}\left[\Big(g(X) - \mathbb{E}(g(X))\Big)^2\right].$$

Note: The variance *is the square of the standard deviation of X, so*

$$sd(X) = \sqrt{\mathrm{Var}(X)} = \sqrt{\sigma_X^2} = \sigma_X.$$

Variance as the average squared distance from the mean

The variance is a measure of how *spread out* are the values that X can take. It is the *average squared distance between a value of X and the central (mean) value, μ_X.*



$$\text{Var}(X) = \underbrace{\mathbb{E}}_{(2)}\ \underbrace{[(X-\mu_X)^2]}_{(1)}$$

(1) *Take distance from observed values of X to the central point, μ_X. Square it to balance positive and negative distances.*

(2) *Then take the average over all values X can take: ie. if we observed X many times, find what would be the average squared distance between X and μ_X.*

Note: The mean, μ_X, and the variance, σ_X^2, of X are just *numbers: there is nothing random or variable about them.*

Variance as the average squared distance from the mean

Example: Let $X = \begin{cases} 3 & \text{with probability } 3/4, \\ 8 & \text{with probability } 1/4. \end{cases}$

Then

$$\mathbb{E}(X) = \mu_X = 3 \times \frac{3}{4} + 8 \times \frac{1}{4} = 4.25$$

$$\textit{Var}(X) = \sigma_X^2 = \frac{3}{4} \times (3 - 4.25)^2 + \frac{1}{4} \times (8 - 4.25)^2$$

$$= 4.6875.$$

When we observe X, we get either *3 or 8: this is random.*
But μ_X *is fixed at 4.25, and* σ_X^2 *is fixed at 4.6875, regardless of the outcome of X.*

Variance as the average squared distance from the mean

For a discrete random variable,

$$\textit{Var}(X) = \mathbb{E}\left[(X - \mu_X)^2\right] = \sum_x (x - \mu_X)^2 f_X(x) = \sum_x (x - \mu_X)^2 \mathbb{P}(X = x).$$

This uses the definition of the expected value of a function of X:

$$\textit{Var}(X) = \mathbb{E}(g(X)) \quad \textit{where} \quad g(X) = (X - \mu_X)^2.$$

Variance as the average squared distance from the mean

$$\boxed{\textit{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}X)^2 = \mathbb{E}(X^2) - \mu_X^2}$$

Proof:

$$\begin{aligned}
\mathrm{Var}(X) &= \mathbb{E}\left[(X-\mu_X)^2\right] \quad \text{by definition} \\
&= \mathbb{E}[\underbrace{X^2}_{\text{r.v.}} - 2\underbrace{X}_{\text{r.v.}}\underbrace{\mu_X}_{\text{constant}} + \underbrace{\mu_X^2}_{\text{constant}}] \\
&= \mathbb{E}(X^2) - 2\mu_X\mathbb{E}(X) + \mu_X^2 \quad \text{by Thm 2.7} \\
&= \mathbb{E}(X^2) - 2\mu_X^2 + \mu_X^2 \\
&= \mathbb{E}(X^2) - \mu_X^2 \,. \qquad \square
\end{aligned}$$

Note: $\mathbb{E}(X^2) = \sum_x x^2 f_X(x) = \sum_x x^2\,\mathbb{P}(X=x)$. This is not the same as $(\mathbb{E}X)^2$:

e.g.
$$X = \begin{cases} 3 & \text{with probability } 0.75, \\ 8 & \text{with probability } 0.25. \end{cases}$$

Then $\mu_X = \mathbb{E}X = 4.25$, so $\mu_X^2 = (\mathbb{E}X)^2 = (4.25)^2 = 18.0625$.
But
$$\mathbb{E}(X^2) = \left(3^2 \times \frac{3}{4} + 8^2 \times \frac{1}{4}\right) = 22.75.$$

Thus $\boxed{\mathbb{E}(X^2) \neq (\mathbb{E}X)^2 \textit{ in general.}}$

Properties of Variance

If a and b are constants and $g(x)$ is a function, then

i) $\mathit{Var}(aX+b) = a^2\,\mathit{Var}(X)$.

ii) $\mathit{Var}(a\,g(X)+b) = a^2\,\mathit{Var}\{g(X)\}$.

Proof:

(part (i))

$$\begin{aligned}
\mathit{Var}(aX+b) &= \mathbb{E}\Big[\{(aX+b)-\mathbb{E}(aX+b)\}^2\Big] \\
&= \mathbb{E}\Big[\{aX+b-a\mathbb{E}(X)-b\}^2\Big] \quad \textit{by Thm 2.7} \\
&= \mathbb{E}\Big[\{aX-a\mathbb{E}(X)\}^2\Big] \\
&= \mathbb{E}\Big[a^2\{X-\mathbb{E}(X)\}^2\Big] \\
&= a^2\mathbb{E}\Big[\{X-\mathbb{E}(X)\}^2\Big] \quad \textit{by Thm 2.7} \\
&= a^2\,\mathit{Var}(X)\,.
\end{aligned}$$

Part (ii) follows similarly.

Note: These are very different from the corresponding expressions for expectations (Theorem 2.7). Variances are more difficult to manipulate than expectations.

Example: finding expectation and variance from the probability function

Example: Recall Mr Chance and his balloon-hire business from page 74. Let X be the height of balloon selected by a randomly chosen customer. The probability function of X is:

height, x (m)	2	3	4
$\mathbb{P}(X = x)$	0.5	0.3	0.2

(a) What is the average amount of gas required per customer?

Gas required was $X^3/2$ from page 74.
Average gas per customer is $\mathbb{E}(X^3/2)$.

$$\mathbb{E}\left(\frac{X^3}{2}\right) = \sum_x \frac{x^3}{2} \times \mathbb{P}(X = x)$$

$$= \frac{2^3}{2} \times 0.5 + \frac{3^3}{2} \times 0.3 + \frac{4^3}{2} \times 0.2$$

$$= 12.45 \ m^3 \text{ gas.}$$

Find $\text{Var}(Y)$.

Example: finding expectation and variance from the probability function

Recall Mr Chance's balloons from page 74. The random variable Y is the amount of gas required by a randomly chosen customer. The probability function of Y is:

gas, y (m^3)	4	13.5	32
$\mathbb{P}(Y = y)$	0.5	0.3	0.2

Find $\text{Var}(Y)$.

First method: *use* $\text{Var}(Y) = \mathbb{E}[(Y - \mu_Y)^2]$*:*

$$\begin{aligned} \text{Var}(Y) &= (4 - 12.45)^2 \times 0.5 + (13.5 - 12.45)^2 \times 0.3 + (32 - 12.45)^2 \times 0.2 \\ &= 112.47. \end{aligned}$$

Second method: *use* $\mathbb{E}(Y^2) - \mu_Y^2$*: (usually easier)*

$$\begin{aligned} \mathbb{E}(Y^2) &= 4^2 \times 0.5 + 13.5^2 \times 0.3 + 32^2 \times 0.2 \\ &= 267.475. \end{aligned}$$

So $\text{Var}(Y) = 267.475 - (12.45)^2 = 112.47$ *as before.*

Variance of a sum of random variables: Var(X + Y)

There are two cases when finding the variance of a sum:

1. **General case:**

> *For general X and Y,*
> *$\mathrm{Var}(X+Y)$ is NOT equal to $\mathrm{Var}(X)+\mathrm{Var}(Y)$.*

We have to find $\mathrm{Var}(X+Y)$ using their covariance (see later courses).

2. **Special case:** when X and Y are *INDEPENDENT:*

> *When X and Y are INDEPENDENT,*
> *$\mathrm{Var}(X+Y)=\mathrm{Var}(X)+\mathrm{Var}(Y)$.*

Class task for Expectation and Variance

A common problem for fire stations is the number of hoax calls received. Each hoax call must be answered as if it were a real fire, so it uses up fire station resources and personnel. A particular problem is that hoaxes are often more common at times when real fires are common, because the real fires put the idea into the hoaxers' minds. This means that fire stations can be most plagued by hoax calls at the times that they are busiest with real fires.

This question suggests how a fire station might consider modelling the occurrences of real fires and hoax calls, so that it can plan how to allocate its resources.

Let Y be the number of **real fires** that will occur over the next month.

Let X be the number of **hoax calls** that will occur over the next month.

Suppose that

$$\begin{aligned} Y &\sim \text{Poisson}(\lambda)\,, \\ X \,|\, Y &\sim \text{Poisson}(\alpha + \beta Y)\,. \end{aligned}$$

(a) State $E(Y)$ and $\text{var}(Y)$. Hence state $E(Y^2)$. [Hint: $\text{var}(Y) = E(Y^2) - (EY)^2$.]

(b) State $E(X \,|\, Y)$ and $\text{var}(X \,|\, Y)$. (Note that you should leave your result in terms of Y.)

(c) Using the formula for conditional expectation, show that $E(X) = \alpha + \beta\lambda$.

(d) The fire station needs to budget for the **total number of calls**, $X + Y$, because each hoax call must be answered as if it were real. Show that

$$E(X + Y) = \alpha + \beta\lambda + \lambda.$$

Class task for Expectation and Variance

(e) **Note:** the procedure used in this question is required

To find $\text{var}(X+Y)$, we will need to find $\text{cov}(X,Y) = E(XY) - E(X)E(Y)$. Using the formula for conditional expectation, we have

$$E(XY) = E_Y\left\{ E(XY \,|\, Y) \right\}.$$

Conditional on Y, we can take the Y outside the inner expectation as a constant:

$$E(XY) = E_Y\left\{ Y \times E(X \,|\, Y) \right\}.$$

Complete the working to show that

$$E(XY) = \alpha\lambda + \beta(\lambda + \lambda^2).$$

[Hint: you will need to use the result for $E(Y^2)$ from part (a).]