Refresher - Random Variables and Probability Distributions

Department of Statistics, University of Warwick

August 2026

1 Random Variables

As with the other refresher courses, you will possibly have seen the material below. If you have not, don't worry, every student's journey is different. This material will all be taught again once you arrive

1.1 What is a Random Variable?

Let's look at some examples.

Watch the video: Random Variables

Example 1.1 *In the background material, we described a football match between Manchester United and Manchester City as an experiment, with the result of the match expressed as outcomes. Instead of representing the outcomes as Win, Lose or Draw, we could represent as the number of points United received for each outcome, namely 3 for Win, 1 for Draw and 0 for Lose. We could also represent the result as the final score, and use a positive or negative number for the difference in scores; e.g. -1 would mean City scored one more goal than United, 0 would be a draw, etc. The difference in score tells us who won, and thus how many points they got, while also providing information on performance.*

English Premier League

2025-2026		GP	W	D	L	F	A	GD	P
1	Arsenal	2	2	0	0	6	0	+6	6
2	Tottenham Hotspur	2	2	0	0	5	0	+5	6
3	Liverpool	2	2	0	0	7	4	+3	6
4	Chelsea	2	1	1	0	5	1	+4	4
5	Nottingham Forest	2	1	1	0	4	2	+2	4
6	Manchester City	2	1	0	1	4	2	+2	3

The English Premier League table two games into the 2025/2026 season. The "GD" column is Goal Difference. Source: ESPN

Outcomes are more useful to us mathematically. If we represent each outcome as the points received, then summing all outcomes over the season gives the final points earned. If we sum the goal difference in each game, we get the goal difference over the entire season.

Example 1.2 *A student answers 5 questions in an exam. We could represent the exam as a sequence of correct (C) or wrong (W) answers, e.g. $CCWWC$. Instead, let $C = 1$ and $W = 0$ and write the sequence as 11001. Now, summing the sequence tells how many correct answers, 3, were given, summarising the final exam result in just one integer. We could call this random variable X.*

However, summarising in this way loses information on the questions themselves. The sequences 11001 and 00111 both lead to 3 correct answers but in a totally different way. Instead, we could consider the sequence as a binary representation of an integer, and making that conversion allows us to ***uniquely*** *represent each possible sequence as a single integer. We could call this random variable Y.*

For example, the sequence 11001 *would be represented by:*

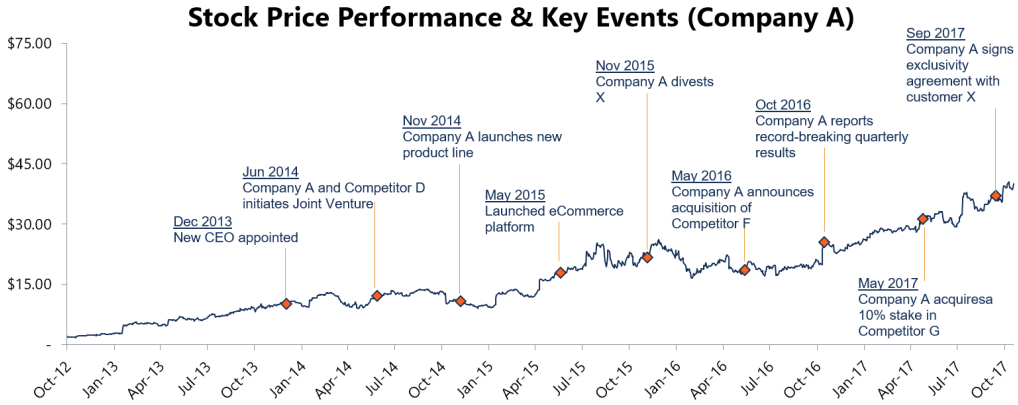
$$1 \times 2^4 + 1 \times 2^3 + 0 \times 2^2 + 0 \times 2^1 + 1 \times 2^0 = 16 + 8 + 0 + 0 + 1 = 25$$

while the sequence 00111 *would be:*

$$0 \times 2^4 + 0 \times 2^3 + 1 \times 2^2 + 1 \times 2^1 + 1 \times 2^0 = 0 + 0 + 4 + 2 + 1 = 7$$

Thus, we can represent the entire exam as one unique integer for each possible sequence of answers. That is, if I told you $Y = 28$, you would know exactly how the student answered the exam.

Example 1.3 *The overall performance and outlook of a company depends on several factors, often interacting with each other in complex and random ways. However, its share (or stock) price says how much outside investors are willing to pay for the company, giving a summary of both past performance and future outlook. One price represents all of the complex underlying factors.*



An example historical performance indicated by share price. Source: Investopedia

Not every experiment results in a number, with the above 3 examples demonstrating how useful it is to translate outcomes into numbers. A random variable makes this transformation.

Definition 1.1 *A **random variable** X is a function that applies a numerical value to the outcomes of a random experiment. We call an observed value of X a **realisation**, denoted $\boldsymbol{x}$.*

The random variable now represents the experiment mathematically, with any outcome being a realisation. We define the probability of an outcome happening as $P(X = x)$.

This definition of a random variable as a function is rather weird and difficult to conceptualise but is one you will see in University. Often it will be as simple as just labelling each outcome with a corresponding number. In Example 1.2, the function we used was to convert the sequence of answers into a binary sequence and then into a unique integer. We now have a way of representing a set of outcomes as one number.

Hence, I often just think of a random variable as the outcomes of an experiment represented as numbers, ignoring the "function" aspect.

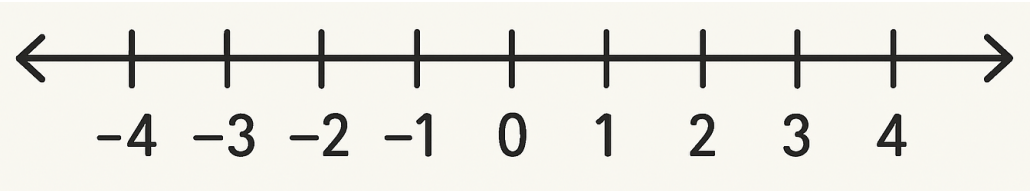
1.2 Types of Random Variables

Random variables are functions and so we use appropriate terminology. The sample space of the experiment Ω is the **domain** of the random variable X. The **range** of X, $\text{Range}(X)$, is the set of numbers that represent the outcomes. For example, if $\Omega = \{W, D, L\}$ for the football match, we might let $\text{Range}(X) = \{3, 1, 0\}$, the corresponding points received. Here the range has a finite number of outcomes. If X corresponds to the number of correct answers out of 5, then $\text{Range}(X) = \{0, 1, \ldots, 5\}$, again finite. If Y is the binary representation of the sequence, the range is $\{0, \ldots, 31\}$, which makes sense given there are $2^5 = 32$ possible ways of answering the exam.

If for the football match X is the goals difference instead of points, the range is then both positive and negative integers. In theory, there is no upper (or lower) bound to the goal difference, and so we could let $\text{Range}(X) = \mathbb{Z}$. The set of integers is infinite but it is **discrete**. Sets like $\mathbb{N}$ and $\mathbb{Q}$ are also discrete; you will learn why but the following definition is enough for now.

Definition 1.2 *A random variable X is called **discrete** if its range is finite or a subset of $\mathbb{Q}$.*

Recall that $\mathbb{Z}$ and $\mathbb{N}$ are subsets of $\mathbb{Q}$. One intuition for what makes a set discrete is if there are "gaps"; we can mark each number down distinctly. Another intuition is that we can "list" all the numbers.

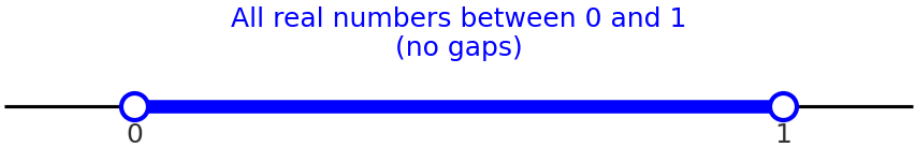


Notice the gaps; there are plenty of numbers between 0 and 1 but none of them are integers.

However, for the share price example, the range is all positive real numbers $\mathbb{R}^+$, not just $\mathbb{N}$. Imagine in the football game we measured the distance all players ran; this would also not be discrete. We call such random variables **continuous**.

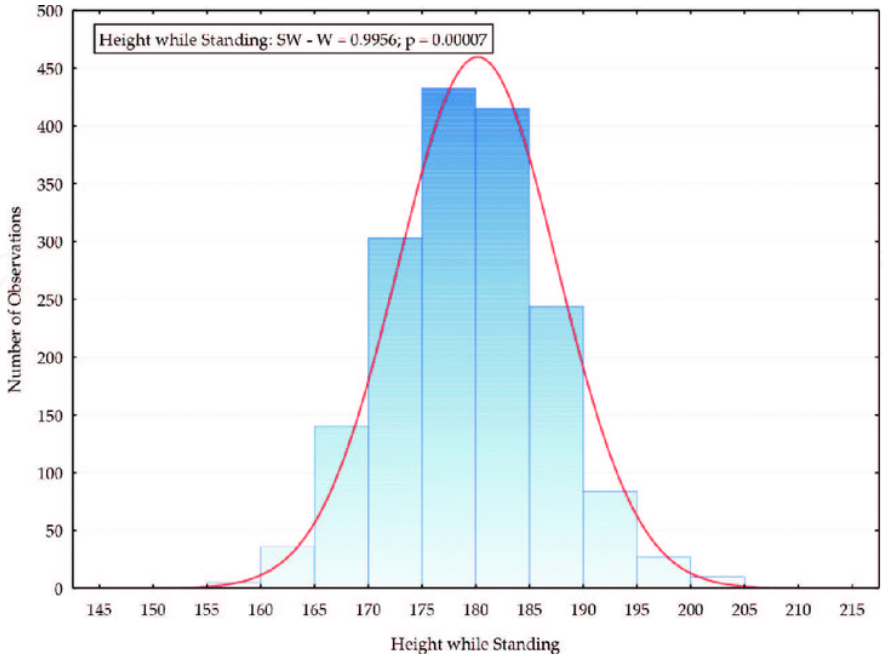
A continuous range does not need to go as far as ∞, a common example is the interval (a, b); *all* numbers between a and b. This is not discrete as there are no gaps; if we choose any two numbers, we can always find a number between them that is also in the interval. Despite not being infinitely long, there are infinitely many numbers within.

Definition 1.3 *A random variable X is called **continuous** if it has a continuous range.*



You will see a more detailed explanation later this year.

Measuring the exact time an exam takes or the heights of students in a class are further examples of continuous random variables. However, we often measure these things in rounded units; using seconds or centimetres but can always get more and more accurate.



A histogram of male heights. The histogram represents the discrete data collected, while the curve represents the true continuous distribution. Source: here

Key Takeaways:

- A random variable is a mathematical formulation of an experiment
- The elements of the range of a random variable are the numbers representing the sample space of the experiment
- A discrete random variable has a range that is finite or infinite with "gaps"
- A continuous random variable has a continuous range with no "gaps".

2 Probability Distributions

It's helpful to translate experiments into numbers, but we also want to have some idea how often these occur; their probability.

In football, we want to know how often goals are scored or the probability a team wins a certain number of games. When it comes to share prices, we want to know the probability of the price increasing by a certain amount, or at what price we should sell. When measuring the heights of students in a class, we want to know what proportion of students should be below a certain height. For a random variable, we describe the probabilities of specific realisations through a **probability distribution**.

2.1 Probability Mass Function

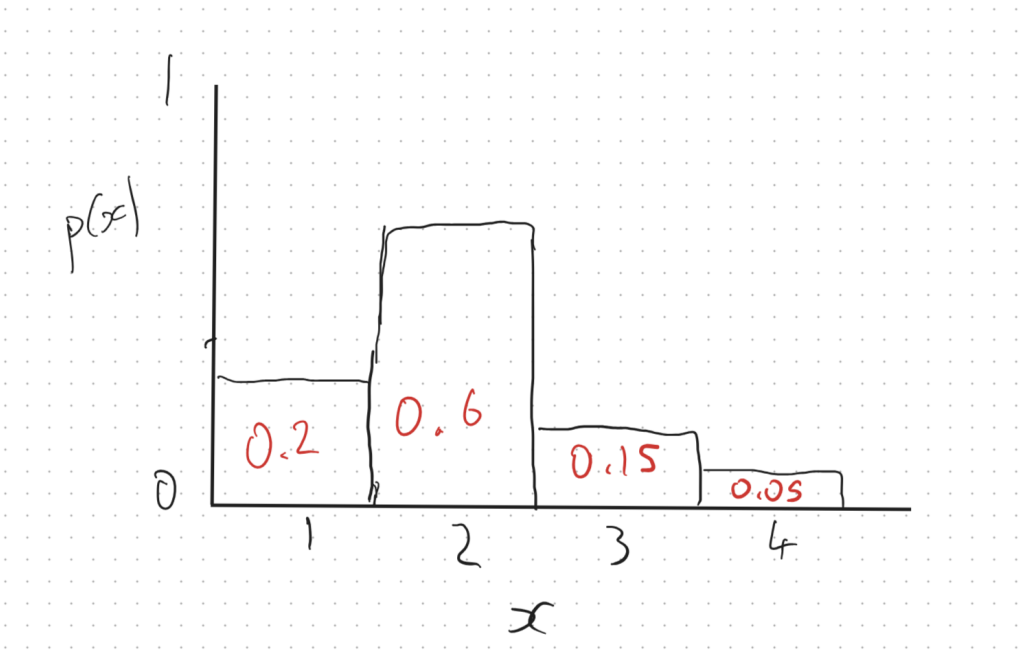
Watch the video: pmf and cdf

We have already seen an example of a probability distribution in the refresher. When we roll a fair die, there are six equally likely outcomes. We give each of these outcomes a probability of $\frac{1}{6}$ and put them in a table as below (now with random variable notation).

x	1	2	3	4	5	6
$P(X=x)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

The roll of a die has a finite range and so we write each $P(X=x)$ explicitly in a table. When the range is something like $\mathbb{Z}$, we need to be a bit more general (which we will see later).

Definition 2.1 *For a discrete random variable X, we define the **probability mass function (pmf)** $p : \mathbb{R} \to [0,1]$ such that $p(x) = P(X=x)$, a function assigning a probability to each element of the range.*



We can write the pmf as a histogram also.

Note: The domain of $p(x)$ is $\mathbb{R}$ rather than just Range(X). For any real numbers not in Range(X) we define their probability to be 0 and don't need to list them.

In order to be a pmf, $p(x)$ must satisfy:

- $0 \leq p(x) \leq 1, \forall x$: each probability must be between 0 and 1
- $\sum_x p(x) = 1$: the sum of all probabilities must be 1.

We can use the pmf to calculate probabilities on subsets of the range by summing as before.

2.2 Cumulative Distribution Function

For a continuous random variable, it is impossible to list each probability explicitly. In fact, each outcome must have a probability of 0; $P(X = x) = 0\ \forall x$. Why? There are too many options to think any are actually possible to happen. This is one of the trickier concepts when thinking of infinity and so don't worry yet if you don't understand.

The probability a share price rises to *exactly* £280 is 0; there are too many possibilities. However, we can find the probability the price is *between* £279.99 and £280.01. So, rather than $P(X = x)$, think of $P(X \in (a, b))$ where $(a, b) \subseteq \text{Range}(X)$.

Definition 2.2 *For a random variable X, we define the* ***cumulative distribution function (cdf)*** $F : \mathbb{R} \to [0, 1]$ *as* $F(x) = P(X \leq x) = P(X \in (-\infty, x])$.

The cdf is the probability of X taking any value up to and including x; the cumulative probability up to x. Writing outcomes as numbers allows us to use set notation here.

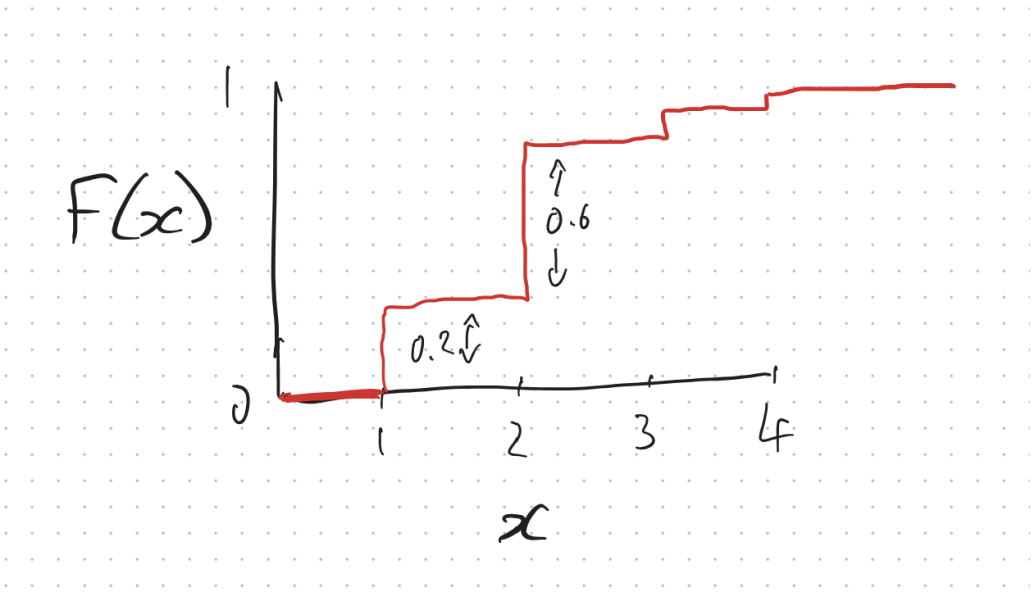
We use the cdf to find the probability of X being in any interval as:

$$\begin{aligned} P(a \leq X \leq b) &= P(X \leq b) - P(X \leq a) \\ &= F(b) - F(a) \end{aligned}$$

Note: The set $\{X \leq x\}$ can be written as $\{X < x\} \cup \{x\}$. For a continuous random variable, knowing $P(X = x) = 0$ implies $P(X \leq x) = P(X < x)$.

The same definition of the cdf works for discrete random variables, and can actually be calculated explicitly as:

$$F(x) = P(X \leq x) = \sum_{y : y \leq x} p(y)$$



The cdf for a discrete random variable.

We see the stepwise nature of the cdf for a discrete random variable; at each possible realisation there is an increase after adding on the corresponding probability. We sum the bars of the histogram for the pmf as we go along.

Just like with the pmf, we write the cdf as a table or create a general formula. See below for a table corresponding to the die roll.

x	1	2	3	4	5	6
$P(X \leq x)$	$\frac{1}{6}$	$\frac{2}{6}$	$\frac{3}{6}$	$\frac{4}{6}$	$\frac{5}{6}$	1

A key difference from the continuous case is that because individual outcomes may have nonzero probabilities in the discrete case, $P(X \leq x) \neq P(X < x)$ for discrete X.

If we are asked to find $P(X > x)$, we use the fact that probabilities sum to 1 to get:

$$\begin{aligned} P(X > x) &= 1 - P(X \leq x) \\ &= 1 - F(x) \end{aligned}$$

However, if asked for $P(X \geq x)$, the discrete case has to add on $P(X = x)$.

Example 2.1 *A random variable X can take values 0, 1, 2 and 3. Please see below for a table detailing the pmf $p(x)$. Use this table to create a new table for the cdf $F(x)$.*

x	*0*	*1*	*2*	*3*
$p(x)$	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{5}$	$\frac{3}{10}$

First, we double check that this satisfies the characteristics of a pmf. Each probability is nonnegative and less than 1. Further, the total sum $\frac{1}{6} + \frac{1}{3} + \frac{1}{5} + \frac{3}{10} = 1$.

Then, we calculate each element of the cdf table explicitly. Note there is a recursive nature to the calculation; rather than calculating the whole sum for each x, we just take the cumulative probability up to the previous realisation and add $p(x)$. So:

$$\begin{aligned}
F(0) &= p(0) = \tfrac{1}{6} \\
F(1) &= p(0) + p(1) = F(0) + p(1) = \tfrac{1}{2} \\
F(2) &= p(0) + p(1) + p(2) = F(1) + p(2) = \tfrac{7}{10} \\
F(3) &= p(0) + p(1) + p(2) + p(3) = F(2) + p(3) = 1
\end{aligned}$$

This gives the following table:

x	*0*	*1*	*2*	*3*
$F(x)$	$\frac{1}{6}$	$\frac{1}{2}$	$\frac{7}{10}$	1

The cdf of the largest outcome should always be 1, as everything must be less than or equal to it. Also, note how the cdf is a non-decreasing function; it either stays the same or increases at every jump.

A cdf table gives you the pmf by inverting the process.

Example 2.2 *A discrete random variable X can take values of 0, 2 and 4. The cdf is detailed below.*

x	*0*	*2*	*4*
$F(x)$	*0.3*	*0.7*	1

What is $p(2) = P(X = 2)$?

We have that $F(2) = P(X \leq 2) = 0.7$ and $F(0) = 0.3$. We also know that:

$$\begin{aligned}
F(0) + p(2) &= F(2) \\
\Rightarrow p(2) &= F(2) - F(0) \\
&= 0.7 - 0.3 \\
&= 0.4
\end{aligned}$$

So $P(X = 2) = p(2) = 0.4$.

2.3 Probability Density Function

So for discrete random variables, we can find the probability, or "mass", of both points and intervals. For continuous random variables, only the mass of intervals is possible; individual points have mass 0. So, in the continuous case, we instead think of "density" of points. The mass of the interval is then a function of the density of the points within and the width or "volume" of the interval.

Watch the video: pdfs

To get from the pmf to cdf we summed. Integration is basically summing over an interval.

Definition 2.3 *For a continuous random variable X, we define its* ***probability density function (pdf)*** *$f(x) : \mathbb{R} \to [0, 1]$ such that:*

$$F(x) = \int_{-\infty}^{x} f(a)da$$

Note: We use a in the integral as a "dummy" variable because it would not make sense to have x both as the variable being integrated and in the limits of the integral.

Conversely, we have:

$$f(x) = \frac{dF(x)}{dx}$$

So, we can get the pdf from the cdf or vice versa.
The pdf will usually be presented in the following way:

$$f(x) = \begin{cases} \text{some function of } x, \text{if } x \in \text{Range}(X) \\ 0, \text{otherwise} \end{cases}$$

See that f is a function of the whole real line; any value not within the range is given density zero.
As with the pmf, we need the "sum", namely integral, to be 1:

- $\int_{-\infty}^{\infty} f(x)dx = 1$

However, as we are dealing with density not mass, the pdf is not restricted to being between 0 and 1 like the pmf is.

Note that our definition for f allowing for any real number also allows us to define integration from $-\infty$ to ∞. If we were to plot the pdf $f(x)$, these requirements corresponds to a function that will always be nonnegative and with an area under the curve of 1.

The following is Question 1 in the November 2020 Further Maths statistics exam.

Question 2.1 *The continuous random variable X has pdf:*

$$f(x) = \begin{cases} \frac{1}{5}, 1 \leq x \leq 6, \\ 0, \textit{otherwise} \end{cases}$$

What is $P(X \geq 3)$?

Solution 2.1 *First, we find $F(3) = P(X \leq 3)$.*

$$\begin{aligned} F(3) &= \int_{-\infty}^{3} f(x)dx \\ &= \int_{1}^{3} \tfrac{1}{5}dx \\ &= \left[\tfrac{x}{5}\right]_1^3 \\ &= \tfrac{3}{5} - \tfrac{1}{5} = \tfrac{2}{5} \end{aligned}$$

Therefore, $P(X \geq 3) = 1 - F(3) = \frac{3}{5}$.

The following was Question 8 in the 2023 Further Maths Statistics paper and has a video solution accompanying it.

Question 2.2 *For a continuous random variable X, you are given the following pdf $f(x)$:*

$$f(x) = \begin{cases} k\sin(2x), 0 \leq x \leq \frac{\pi}{6} \\ 0, \textit{otherwise} \end{cases}$$

a) What is k?

b) Find $F(x)$

c) Find the median of X to 3 significant figures

d) Find $E[X]$ in the form $\frac{b\sqrt{3}-\pi}{a}$, where $a, b \in \mathbb{Z}$

Solution 2.2 *Watch the video solution: 2023 Further Maths Question 8*

We refer to the pmf or pdf of a random variable as its **probability distribution**. We often use f rather than p to refer to the function in general.

Key Takeaways:

- A discrete random variable has a pmf which gives the probability of any realisation
- A continuous random variable has a pdf which gives the density of any realisation, and probabilities exist only over intervals.
- The cdf sums the pmf or integrates the pdf to get probabilities of sets
- A pmf/pdf must always sum/integrate to 1
- A pmf must always have values between 0 and 1

3 Expectation and Variance

A probability distribution tells us how likely it is that United score 3 goals against City. It tells us what proportion of students should be taller than 180cm. It tells us the probability a share price falls below £200.

However, it's often more interesting to know what happens *on average*. What is the average number of goals United score? What is the average height in the class? What is the average result I get when I roll a die? What is the average share price?

When we know these average values, it's also interesting to know how variable the data is around this average. If United score 2 goals a game on average, are they scoring 2 goals consistently or do they sometimes score 0 and sometimes 4? Does my class have some very tall and very short students, or is everyone roughly the same? Is the share price very volatile, or is it a safe investment?

We explore ways you may have seen before of summarising the properties of probability distributions and random variables. If you have not seen these concepts in such detail, don't worry, it will be taught once you arrive.

3.1 Expectation

For a random variable X, we are interested in the average result, or what we *expect* to happen. Think about calculating the mean from a frequency distribution table, except with probability rather than frequency.

Definition 3.1 *The **expectation** or **expected value** of a random variable X, $E[X]$, is the **mean** of the outcomes X can take weighted by their probabilities.*

For a discrete random variable:

$$E[X] = \sum_x x \cdot p(x)$$

$E[X]$ is a generalisation of the weighted average: we take each outcome x and weight it by its corresponding probability. We then sum all the weighted outcomes.

So, when calculating the expect number of goals United score, we weight the amount of goals by the probability they score that many goals and sum everything.

E[X] extends to continuous random variables as we might expect:

$$E[X] = \int_{-\infty}^{\infty} x \cdot f(x)dx$$

For outcomes that lie outside the range, their probability and thus weight is 0. From this point onwards, we will use the pdf/integral definition for convenience unless otherwise stated; for discrete use a sum instead of integral.

Example 3.1 *Let X be the result of rolling a fair 6-sided die. What is $E[X]$?*
In this case, $p(x) = \frac{1}{6}$, $\forall x$. Thus:

$$\begin{aligned} E[X] &= \sum_{k=1}^{6} k \cdot P(X = k) = \sum_{k=1}^{6} k \cdot \tfrac{1}{6} \\ E[X] &= 1 \cdot \tfrac{1}{6} + 2 \cdot \tfrac{1}{6} + 3 \cdot \tfrac{1}{6} + 4 \cdot \tfrac{1}{6} + 5 \cdot \tfrac{1}{6} + 6 \cdot \tfrac{1}{6} \\ &= 3.5 \end{aligned}$$

As the die is fair, the weighted and simple averages are the same.

For some function $g(x)$, we can also define the expectation as:

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)dx$$

This is is often called the **law of the unconcious statistician**, or **LOTUS** (although you probably have not seen it called this). For example, if $g(x) = x^2$, then:

$$E[X^2] = \int_{-\infty}^{\infty} x^2 f(x)dx$$

Note: We only replace the first x with $g(x)$ in the formula, we do not change the $f(x)$. We change the value but not the weight.

Using the formula for $E[g(X)]$ (feel free to try prove the following yourself), for some $a, b \in \mathbb{R}$ we get:

$$E[aX + b] = aE[X] + b \tag{1}$$

We call this **linearity of expectation**. If $b = 0$, this shows $E[cX] = cE[X]$; we can take the constant factor outside of the expectation.

Think about if we subtracted 5cm from the height of every student in the class. The new expected value would just be the previous minus 5cm. If we doubled their heights, we would double the average.

Suppose we have two random variables X and Y. Then:

$$E[X + Y] = E[X] + E[Y]$$

Combining all our rules together we have:

$$E[aX + bY] = aE[X] + bE[Y]$$

3.2 Intuition - Games

In the background material, we mentioned how games (and gambling) were often a driving force for the development of probability. Games also provide some intuition for expectations.

In a game, think of $E[X]$ as the average score. So, for a 6-sided die, I expect to get 3.5 each time I roll, despite the fact I can never actually roll 3.5. This is important when the score relates to some monetary payout; if I get £x when I roll an x, then I make £3.50 on average each time I play the game.

$E[X]$ gives a fair price for the game. If I have to pay £3 to play this game then I should always play; I make £0.50 on average every time. If it's £4 pound, I should find something else to do as I lose £0.50 each time (all casino games that are pure chance like roulette are built this way). £3.50 would be a fair price; playing and not playing end up the same on average and it's up to you how much risk you want to take. This line of thinking will be covered in greater detail once you start studying game theory and financial mathematics.

Instead, suppose I roll the die and the amount I receive is the result squared. Now, my average payout will be $E[X^2]$; I calculate how much I would get each time (x^2) and weight it by the probability.

Now suppose the game consists of rolling a 6-sided die and a 12-sided die. I add both numbers together and receive the total. If I let X be the 6-sided and Y be the 12-sided, my return is $X + Y$. My average return will then be $E[X + Y]$ and so I can calculate both expectations separately and add them together.

3.3 Median and Mode

When looking at the heights of students in a class, the expected value is just one measure of central tendency. Rather than thinking of the average height, instead we can think of the most common height. Alternatively, we might be interested in the height which is right in the middle; 50% of students are taller, 50% shorter. You may have seen these concepts before when analysing data sets, but they extend to probability distributions.

We begin with the most common outcome, the mode.

Definition 3.2 *For a random variable X, we define its* ***mode*** *$Mode(X)$ as the value x that maximises the probability distribution, or:*

$$Mode(X) = \arg\max_x f(x)$$

The arg max stands for "argument max", the value of x that maximises $f(x)$; you have probably not seen this notation before. If we just had max, it would be the actual maximum probability rather than the x value.

If we plot a probability distribution, the mode will be the x value corresponding to the highest point in the graph. To calculate the mode, we use techniques from calculus to find maximum points. In a game, the mode is the most likely outcome.

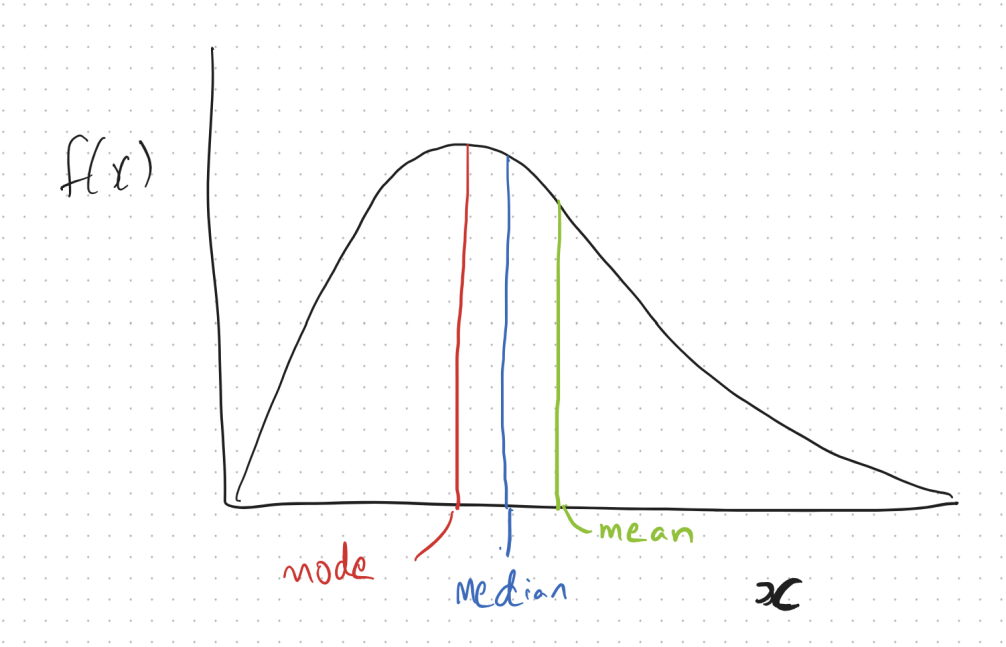
The median of a dataset is the value right in the middle, with half the possible outcomes on either side.

Definition 3.3 *For a random variable X, we define its* ***median*** *$Median(X)$ as the value x that has a cumulative probability of* 0.5*, or:*

$$P(X \leq Median(X)) = 0.5$$

So, the median is the value with a probability of 0.5 of being lower (and thus higher) than it. In a game, it would be the score that half the time you beat, half the time you don't.

In a plot of the probability distribution, the median will be the x which has half the area under the graph on either side of it. To find the median, we solve the algebraic equation $P(X \leq Median(X)) = 0.5$.

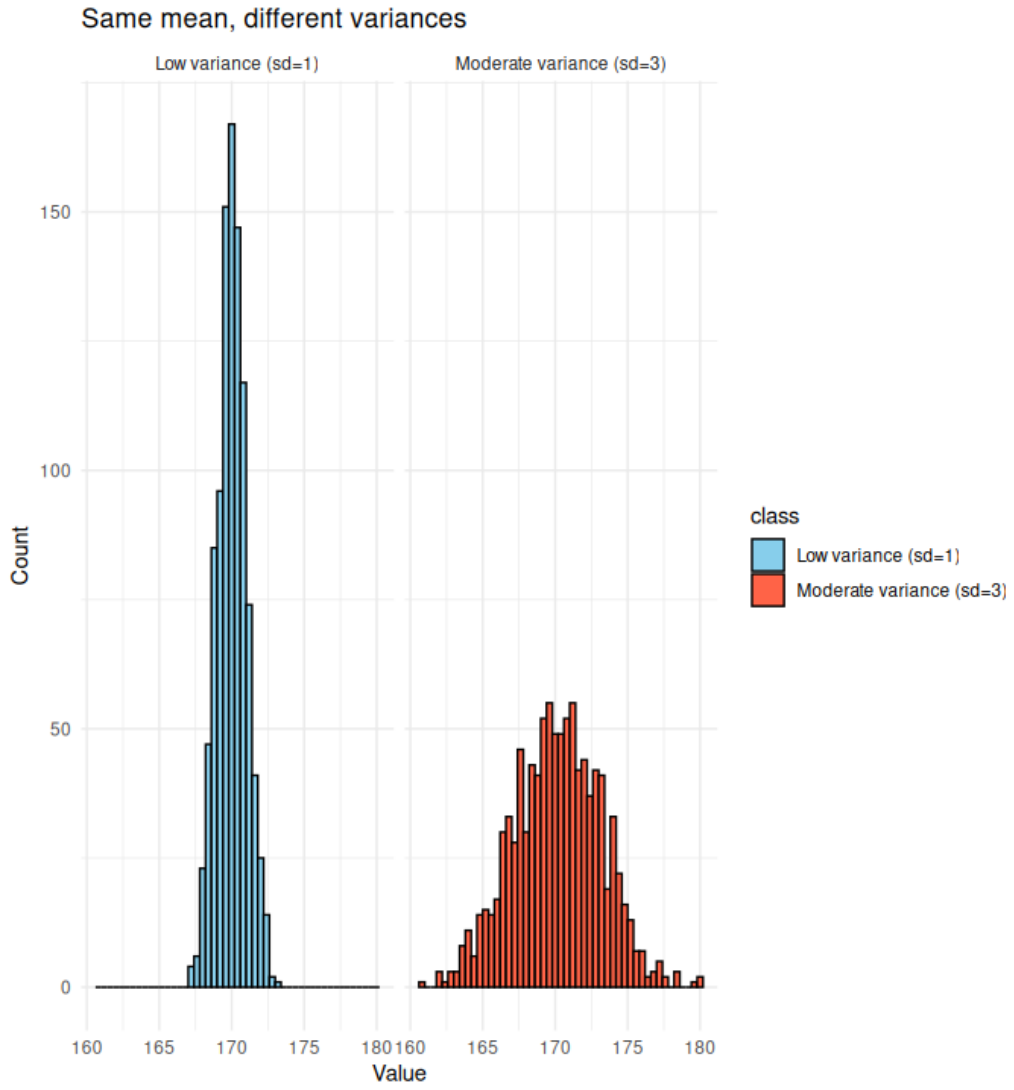


Here we see a skewed distribution with the mean, median and mode labelled. The mode is the highest point in the graph. The median is moved slightly to the right, allowing for half the area under the curve to be either side of it. The mean is even further to the right than the median, as it depends not just on the probabilities but also the values; bigger values will increase the mean but not the median.

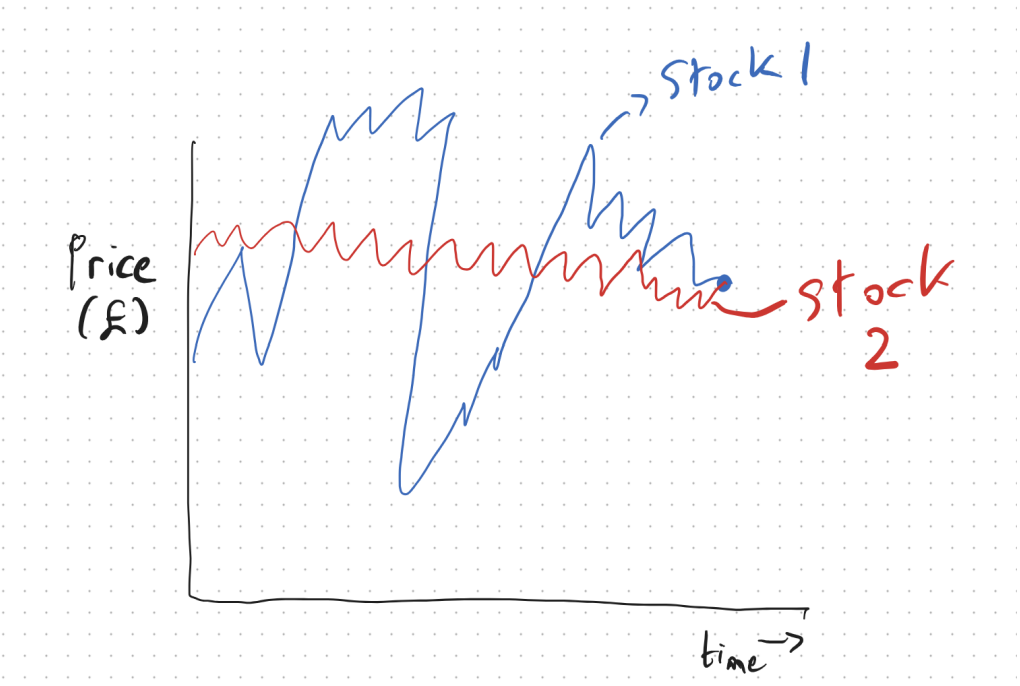
3.4 Variance

Knowing what happens on average can only tell us so much. Knowing how far the outcomes can be from the average is also valuable.

Suppose I record the heights of two (very large) classes of students and they both have the same mean. In Class 1, most students are pretty close to the expectation. In Class 2, there is a large spread to the heights; most students are in the middle but some are short and some are tall. Notice below how the increased spread decreases the height (of the curve, not the students).



Suppose instead I am deciding between two stocks to invest in, both with the same price. Stock 1 has major fluctuations in value while stock 2 is pretty consistent. Some way of measuring the volatility and comparing the two is very useful.



We want some way to describe the spread of the data. Well, think about the average distance from the mean; a large average distance would mean a large spread. So, we calculate $X - E[X]$, the deviation from the mean. However, we generally square distances when calculating, and so we take the squared deviation from the mean, $(X - E[X])^2$. Then, we want the weighted average so take expectation. This gives the variance.

Definition 3.4 *For a random variable X, its **variance** $Var(X)$ is the expectation of the squared deviation from the mean, or:*

$$Var(X) = E[(X - E[X])^2]$$

Note: $E[X]$ is a constant in the above formula, so $g(x) = (x - E[X])^2$ for LOTUS.

The larger the variance, the more likely outcomes far from $E[X]$ are to occur. A share price with high variance is risky, with a chance of making a big return but also a big loss. A share price with low variance is more predictable and probably give a steady but unspectacular return.

Similar qualities to linearity of expectation exist for variance. For $a, b \in \mathbb{R}$:

$$Var(aX + b) = a^2 Var(X)$$

As the variance of a constant is zero, adding it to X has no effect on variance. Think about subtracting 5cm from the height of every student to account for the machine used; it won't change the actual variability, just the values.

However, multiplying X by a will increase the variance by a factor of a^2; this makes sense given the power of 2 in $Var(X)$.

We often use a different formula to calculate the variance:

$$Var(X) = E[X^2] - (E[X])^2 \tag{2}$$

Generally, we know $E[X]$ and it is easier to calculate $E[X^2]$ than $E[(X - E[X])^2]$.

Example 3.2 *We want to calculate the variance when rolling a 6-sided die. We know $E[X] = 3.5$ so we find $E[X^2]$:*

$$\begin{aligned} E[X^2] &= \sum_{k=1}^{6} k^2 \cdot \tfrac{1}{6} \\ E[X^2] &= 1^2 \cdot \tfrac{1}{6} + 2^2 \cdot \tfrac{1}{6} + 3^2 \cdot \tfrac{1}{6} + 4^2 \cdot \tfrac{1}{6} + 5^2 \cdot \tfrac{1}{6} + 6^2 \cdot \tfrac{1}{6} \\ &= \tfrac{91}{6} \end{aligned}$$

Then, we have:

$$\begin{aligned} Var(X) &= E[X^2] - E[X]^2 \\ &= \tfrac{91}{6} - (\tfrac{7}{2})^2 \\ &= \tfrac{91}{6} - \tfrac{49}{4} \\ &= \tfrac{35}{12} \approx 2.92 \end{aligned}$$

Suppose we have two random variables X and Y. Then, if they are **independent**, we have:

$$Var(X + Y) = Var(X) + Var(Y)$$

If they are dependent in some way, there is an extra covariance term which we will not go into.

This gives us a general formula for X and Y independent and $a, b \in \mathbb{R}$:

$$Var(aX + bY) = a^2 Var(X) + b^2 Var(Y)$$

Note: Even if we are looking at $X - Y$ (so $a = 1,\ b = -1$), the variance will increase as b^2 is positive. If we include a new random variable, it can only get more variable.

Example 3.3 *Imagine X and Y are throws of two fair 6-sided dice. Then $E[X - Y] = E[X] - E[Y] = 0$. However, $Var(X - Y) = Var(X) + Var(Y) = \frac{35}{12} + \frac{35}{12} = \frac{35}{6}$.*

Think about what the range of $X - Y$ is. It can go as high as 5, when $X = 6$ and $Y = 1$, but also go as low -5. That is, Range$(X - Y) = \{-5, -4, \ldots, 0, \ldots, 4, 5\}$. Each of these outcomes will have a probability based on the pairs of numbers that could generate them. The range of outcomes is now larger than before and so is the maximum distance from the expectation; it is now 5, it used to be 2.5. Hence, the variance has increased.

Now try the following quick question; the first statistics question from 2019 Further Mathematics A Level.

Question 3.1 *We know that $Var(X) = 5$. What is $Var(4X - 3)$?*

Solution 3.1

$$\begin{aligned} Var(4X - 3) &= 4^2 Var(X) \\ &= 16 \times 5 = 80 \end{aligned}$$

The following is a modified version of Q4 in the 2020 Further Maths paper.

Question 3.2 *The discrete random variable X corresponds to rolling a fair n-sided die. Define the random variable $Y = 2X$. Using the standard results for $\sum n$, $\sum n^2$ and $Var(aX + b)$, show:*

$$Var(Y) = \frac{n^2 - 1}{3}$$

Solution 3.2 *The standard results referred to are:*

$$\sum_{k=1}^{n} k = \frac{n(n+1)}{2}$$
$$\sum_{k=1}^{n} k^2 = \frac{n(n+1)(2n+1)}{6}$$

We first find $Var(X)$, which requires $E[X]$ and $E[X^2]$. Each outcome $k \in \{1, \ldots, n\}$ has probability $\frac{1}{n}$ and so $E[X]$ is:

$$\begin{aligned} E[X] &= \sum_{k=1}^{n} k \tfrac{1}{n} \\ &= \tfrac{1}{n} \sum_{k=1}^{n} k \\ &= \frac{n(n+1)}{2n} = \frac{n+1}{2} \end{aligned}$$

Similarly, $E[X^2]$ is:

$$\begin{aligned} E[X^2] &= \sum_{k=1}^{n} k^2 \tfrac{1}{n} \\ &= \frac{n(n+1)(2n+1)}{6n} = \frac{(n+1)(2n+1)}{6} \end{aligned}$$

Then:

$$\begin{aligned} Var(X) &= E[X^2] - E[X]^2 \\ &= \frac{(n+1)(2n+1)}{6} - \frac{(n+1)^2}{4} \\ &= \frac{2n^2 + 3n + 1}{6} - \frac{n^2 + 2n + 1}{4} \\ &= \frac{2n^2 - 2}{24} \\ &= \frac{n^2 - 1}{12} \end{aligned}$$

Now, $Var(Y) = Var(2X) = 4Var(X)$, and so $Var(Y) = \frac{n^2-1}{3}$, as required.

Key Takeaways:

- The expected value $E[X]$ of a random variable is a weighted average
- The mode is the value with highest probability

- The median is the value that has a probability of 0.5 either side of it
- The variance measures the spread of the distribution
- $E[aX + b] = aE[X] + b$, $Var(aX + b) = a^2Var(X)$
- $E[X + Y] = E[X] + E[Y]$
- $Var(X + Y) = Var(X) + Var(Y)$ when X and Y are **independent**

4 Specific Distributions

So far, we have seen examples of experiments and random variables, and talked about distributions. However, we have not discussed specific distributions for those examples.

In this section I introduce several such distributions. If you studied Further Maths for A-levels you may have seen them all. If you studied Maths, you may have seen just the first two. Nonetheless, no matter what your mathematical journey until now, it is ok to not have seen (or remember) some of these. I will recap them now as they are necessary to solve some of the sample exam questions but they will be covered again in detail once you get to Warwick.

4.1 Binomial Distribution

Many experiments have two possible outcomes; success and failure. This could be answering a question on an exam, playing a tennis match, buying a car at an auction. Any experiment with two possible outcomes is called a **Bernoulli Trial**, where p is the probability of success.

Swiss Mathematician Daniel Bernoulli (1700-1782). Source: Wikipedia

Flipping a coin is the classic example of a Bernoulli trial. Let X be the random variable where $H = 1$, a success, and $T = 0$, a failure. For a fair coin, $p = \frac{1}{2}$, and X is a Bernoulli random variable.

However, often we are more interested in a series of Bernoulli trials. Rather than flipping a coin once, we flip it ten times and count the heads. Rather than answering one question we answer a whole exam and check the final score. Rather than playing one tennis match we play a whole tournament. If each of these is a Bernoulli random variable and 1 is a success, we add all values of the random variable together to count the number of successes. Then, we can find the probability of a particular number of success (5 heads out of 10, 40% final score, 10 wins out of 12 matches).

The sum of Bernoulli random variables is called a Binomial random variable, under two conditions. First, the number of trials, n, must be fixed. We need to know how many coins will be flipped or how many questions are on the exam. The tennis fans among you will know that usually tournaments are knockout; you play until you lose. In this case, n is not fixed and a Binomial is not a suitable distribution.

The full bracket for Wimbledon 2025. Source: Bleacher Report

Second, the probability p is fixed between trials. In other words, each trial is independent of the previous. If we were to draw a tree diagram, the edge probabilities would always be the same. In the coin flip example, flipping a heads should have no effect on the rest of the flips. In the exam, however, maybe the questions get more difficult as you go through, and so p actually decreases with each trial; the Binomial would not be appropriate.

Definition 4.1 *Suppose X is the number of successes recorded in n Bernoulli trials with success probability p. Then, X follows a **Binomial** distribution if:*

- *The number of trials n is fixed*
- *The trials are independent of each other (p is fixed).*

We write $X \sim B(n,p)$ or $Bin(n,p)$.

The Binomial is a discrete random variable, as $\text{Range}(X) = \{0, 1, \ldots, n\}$. The probability of recording k successes in n trials is given by the pmf:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \ k \in \{0, 1, \ldots, n\}$$

where $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ is the **Binomial coefficient**.

Note: We don't use the $p(x)$ notation for the pmf here so as not to get confused with the probability of success p.

For a $B(n,p)$ distribution, we have:

- $E[X] = np$
- $Var(X) = np(1-p)$

The expectation is simply the number of trials multiplied by probability of success. The variance is maximised (check yourself) when $p = \frac{1}{2}$; success and failure being equally likely gives most uncertainty.

As mentioned in the Tree Diagrams refresher, any discrete random variable can be transformed into a Bernoulli by picking a set of outcomes as success and the rest as failure. When rolling a die, let even numbers be success and odd be failure. When spinning a roulette wheel, let our chosen numbers be success and the rest be failure. In a football match, let winning be a success and losing or drawing be a failure. This make the Binomial distribution extremely flexible once the conditions on n and p are satisfied.

The following is a modified version of Q15 on the 2024 A level mathematics paper 3.

Question 4.1 $X \sim B(48, 0.175)$

a) Find $E[X]$

b) Show $Var(X) = 6.93$

c) Find $P(X \geq 6)$

d) Find $P(9 \leq X \leq 15)$

Solution 4.1 *(a)* $E[X] = np = 48 \times 0.175 = 8.4$
(b) $Var(X) = np(1-p) = 48 \times (0.175) \times (1 - 0.175) = 6.93$
(c) To get $P(X \geq 6)$ *we could find the probability for each number 6 and above and add them together. It is quicker to use the relation* $P(X \geq 6) = 1 - P(X < 6)$*. To get* $P(X < 6)$*, we get the probabilities for* $\{0, 1, 2, 3, 4, 5\}$.

$$P(X = 0) = 0.0001$$
$$P(X = 1) = 0.001$$
$$P(X = 2) = 0.005$$
$$P(X = 3) = 0.0161$$
$$P(X = 4) = 0.0385$$
$$P(X = 5) = 0.0718$$

Summing these together gives $P(X < 6) = 0.1325$ *and so* $P(X \geq 6) = 0.8675$.
(d) To get $P(9 \leq X \leq 15)$*, we find the probability for each number between 9 and 15 and sum them. This gives us* 0.3198.

4.2 Normal Distribution

The Binomial is the most common discrete random variable you may have encountered up until now. When dealing with continuous data, the Normal distribution is most common.

German mathematician and scientist Carl Friedrich Gauss (1777-1855).

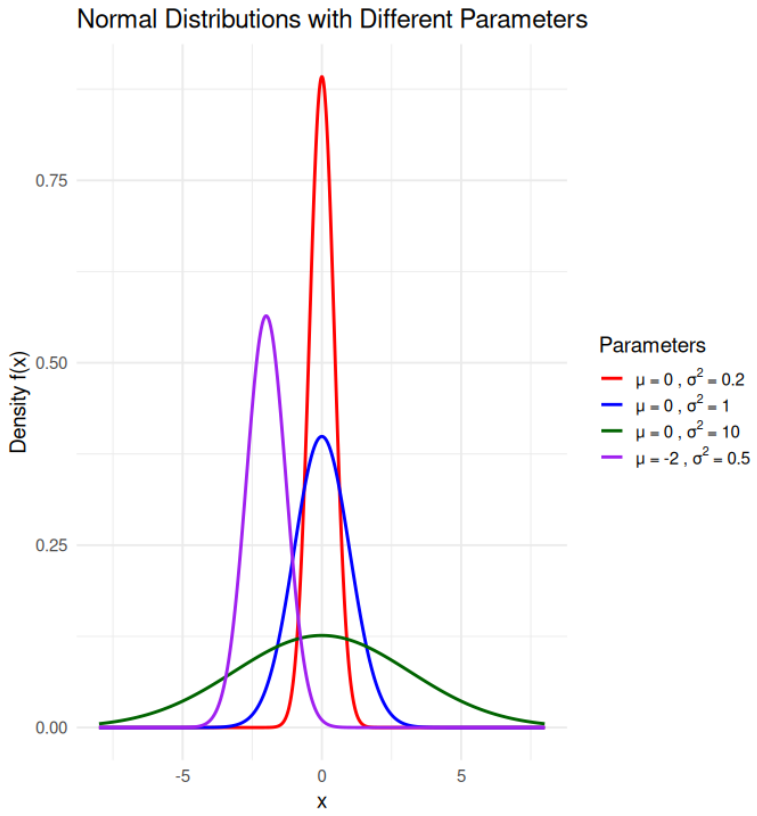
When measuring heights of students in a class, a reasonable first assumption to make on the distribution is **symmetry**; a student is as likely to be tall as they are short. Similarly, assume a share price is just as likely to increase as decrease. For a symmetric distribution, the mean and median are equal.

A second assumption we might make is **unimodality**; there is only one mode, a single peak in the curve. This assumes we have only one height in the class that is most common, or one share price that is most likely. For a symmetric unimodal distribution, the mean and mode (and thus median) are equal.

Definition 4.2 *A random variable X follows a* ***Normal*** *or* ***Gaussian*** *distribution with mean μ and variance σ^2, written $N(\mu, \sigma^2)$, if it has the following pdf:*

$$f(x) = \frac{1}{\sqrt{(2\pi\sigma^2)}} \exp(-\frac{(x-\mu)^2}{2\sigma^2})$$

The cdf of the Normal Distribution is not available in a nice, easy form, and so is instead written as an integral of the pdf. The graph of the pdf for multiple values of μ and σ^2 is below; μ dictates location and σ dictates the width/spread of the graph.



Four example Normal distributions. Look at the values of μ and σ^2 and see how the curves change.

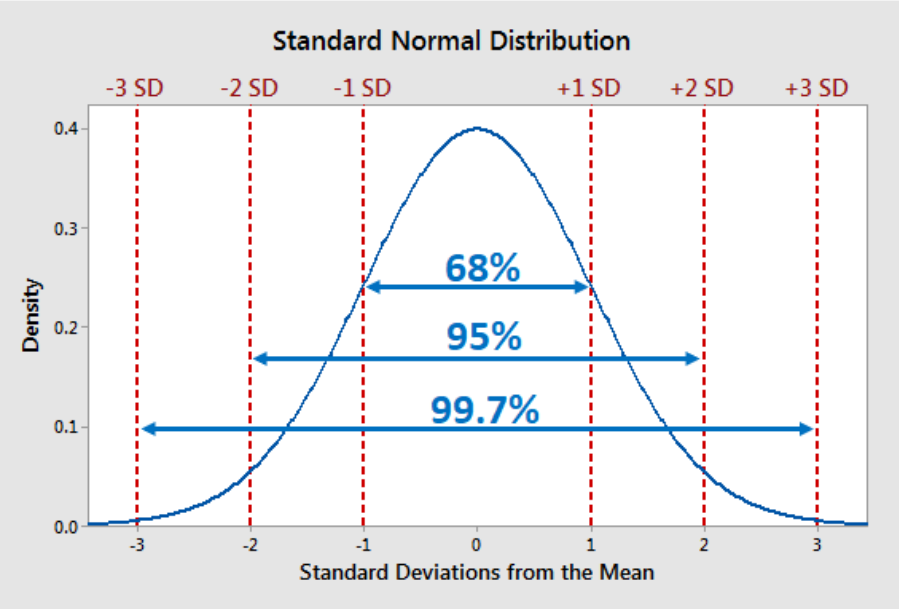
The shape of the distribution is a **bell curve**. Often, the phrases "Normal" and "bell curve" are used interchangeably. However, other distributions have bell curve shaped graphs too, so use Normal or Gaussian to be specific.

The mean μ tells us where the center of the bell curve is, while the **standard deviation** σ describes its width. A large σ makes it more likely to observe extreme values and thus less likely to observe values close to μ. We see how a bigger σ (the green curve) leads to a more "squashed" graph. As the area under the curve is fixed to 1, increasing the width decreases the height.

The Normal is continuous with Range$(X) = \mathbb{R}$; technically, any real number is possible. However, as we get further from the expectation, the probability decreases to the point where it is *essentially* impossible. The larger the variance, the larger this range of *reasonable* values.

The standard deviation leads to the following **empirical rule**:

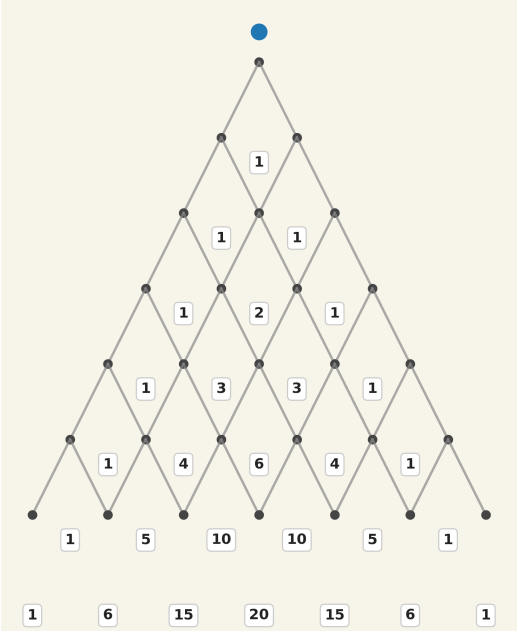
- c.68% of the distribution is within one standard deviation of the mean
- c.95% of the distribution is within two standard deviations of the mean
- c.99.7% of the distribution is within three standard deviations of the mean



A graphical representation of the empirical rule. Source: Statistics by Jim

This is why hypothesis tests often use a 95% significance level; it is two standard deviations either side.

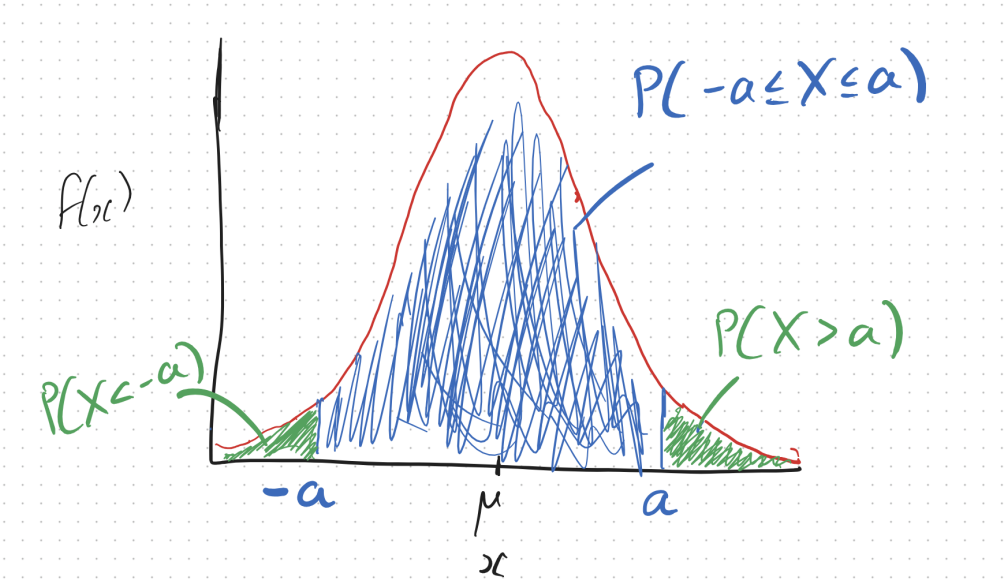
We can assume X is a Normal random variable, but until we record real data it is impossible to know if this is appropriate. A histogram of the data checks whether the requirements are met. Nonetheless, the Normal is a good default choice for a distribution due to its nice properties (which we will see) and ubiquity in nature. I often think of dropping marbles to form Pascal's triangle; most will settle into the middle but some will spread out evenly.



The symmetry of the Normal distribution gives us the following property:

$$P(X \leq -a) = P(X \geq a) = 1 - P(X \leq a)$$

Thus, complex probability statements can be generated from simply knowing $P(X \leq a)$.



4.2.1 The Standard Normal Distribution

The ability to break down complex statements is important because we cannot directly integrate the pdf. Instead, we use probabilities already given to us (such as in log tables). However, it is impossible to have these probabilities available for each possible μ and σ, so we use a reference.

Rather than thinking of the heights of students, think of their difference from the average height. Then negative values will indicate shorter students and so on. For a share price, think of the change from some starting price instead.

Definition 4.3 *We say Z follows the **Standard Normal** distribution if it is a Normal distribution with $\mu = 0$ and $\sigma^2 = 1$.*

Note: We usually represent the Standard Normal with a Z but Z can also be used for any Normal. Sometimes, ϕ is used for the Standard Normal and Φ for its cdf.

To get from any Normal to the Standard Normal, we **standardise**.

Definition 4.4 *Suppose $X \sim N(\mu, \sigma^2)$. Let $Z = \frac{x-\mu}{\sigma}$. Then, $Z \sim N(0, 1)$.*

So, to standardise any Normal random variable X we:

- **Shift** by subtracting the expected value
- **Scale** by dividing by the standard deviation

Using the previous properties to help intuition:

$$E[\tfrac{X-\mu}{\sigma}] = \tfrac{1}{\sigma}(E[X] - \mu) = 0$$
$$Var(\tfrac{X-\mu}{\sigma}) = \tfrac{1}{\sigma^2}Var(X - \mu) = \tfrac{1}{\sigma^2}Var(X) = 1$$

Standardising the Normal is very useful because we only need the specific probabilities for this one distribution; everything else can be derived from them.

4.2.2 The Normal Approximation

The Normal distribution is widely studied not just because of its simplicity and ubiquity, but also because it has a very useful property for large sample sizes.

Suppose X counts how many twos you get in 1000 rolls of a fair 6-sided die, such that $X \sim Bin(1000, \frac{1}{6})$. We can easily find $P(X = 50)$. However, getting $P(20 \leq X \leq 100)$ would require individually calculating the probability of all 81 numbers in the set. There is a quicker way.

Definition 4.5 *We say a discrete random variable X has a **Normal approximation** if $X \approx N(E[X], Var(X))$ under certain conditions.*

By approximating the number of twos with a Normal distribution, we find $P(20 \leq X \leq 100)$ through two calculations, $P(X \leq 100)$ and $P(X \leq 20)$, rather than 81.

Note: It is often recommended to perform a **continuity correction** when approximating a discrete random variable with the Normal. This entails adding a small number, usually 0.5, to the value of interest to account for the starting point of the interval. I will not do this in my solutions for simplicity.

The Binomial is the most common distribution to approximate. It becomes more symmetric as n grows and it is tedious to sum probabilities. Thus, when n is large, the Normal approximation of the Binomial is appropriate.

Note: It is common to make statements about "large" n, but often what constitutes large is not given or justified. What it actually means is that it is true for the limit as $n \to \infty$, and so we choose an n along the way to get close. The larger n is, the better the approximation.

The reasons for the Normal approximation are rooted in the **Central Limit Theorem** and **Law of Large Numbers**; you will explore these in great detail but these are core properties as to why the Normal distribution is so popular.

The following is a modified version of Q17 from the 2024 A level Maths Paper 3.

Question 4.2 *In 2019, the lengths of babies at a newborn clinic can be modelled by a Normal distribution with mean 50 and standard deviation 4.*

a) State the probability that the length of a new-born baby is less than 50 cm.

b) Find the probability that the length of a new-born baby is more than 56 cm

c) Find the probability that the length of a new-born baby is more than 40 cm but less than 60 cm.

d) Determine the length exceeded by 95% of all new-born babies at the clinic.

Solution 4.2 *We let X be the length of new born babies such that $X \sim N(50, 4^2)$. Note that we write 4^2 to be clear that this is the variance and 4 is the standard deviation.*
(a) We want $P(X < 50)$. Notice that 50 is the mean and by symmetry also the median. Hence, the probability is 0.5.
(b) We want $P(X > 56) = 1 - P(X \leq 56)$. To find this, we standardise.

$$\begin{aligned} z = \frac{x - \mu}{\sigma} &= \frac{56 - 50}{4} \\ &= 1.5 \end{aligned}$$

So, we need $P(Z \leq 1.5)$, which is 0.9332. Thus, $P(Z \geq 1.5) = 1 - 0.9332 = 0.0668$
(c) To find $P(40 < X < 60)$ we standardise both end points. This gives the interval $P(-2.5 < Z < 2.5)$. Using some algebra:

$$\begin{aligned} P(-2.5 < Z < 2.5) &= P(Z < 2.5) - P(Z < -2.5) \\ &= P(Z < 2.5) - (1 - P(Z < 2.5)) \\ &= 2 \times P(Z < 2.5) - 1 \\ &= 2 \times 0.9938 - 1 \\ &= 0.9876 \end{aligned}$$

(d) This is the length x such that $P(X > x) = 0.95$. This is equivalent to $P(X \leq x) = 0.05$ and thus z such that $P(Z \leq z) = 0.05$.

Certain Normal tables might give you this number, but generally we must take advantage of symmetry and find $P(Z \leq -z) = 0.95$. We find $-z = 1.645$ and thus $z = -1.645$.

Finding z is the first step, we must now transform it back to x:

$$\begin{aligned} -1.645 &= \frac{x - 50}{4} \\ -6.58 &= x - 50 \\ x &= 43.42 \end{aligned}$$

So 95% of babies are longer than 43.42 when born.

Note: Part (d) references 95% but this is not 2 standard deviations away like the empirical rule. There is a difference between the number with 95% above it, as in (d), and the 2 numbers with 95% between them. Those two numbers have 2.5% either side.

4.3 A Quick Note

The following three distributions were not covered in A level Maths but may have been in Further Maths. They will be used to solve questions from Further Maths or STEP later but don't panic if this is your first time seeing them.

4.4 Uniform Distribution

The simplest form of any probability distribution is to assume each outcome is equally likely. When we roll a fair die or flip a fair coin, we have no reason to believe any outcome is more likely than another; we assume the probabilities are uniform.

Definition 4.6 *A random variable X follows a **uniform** distribution if every outcome in the range has the same probability of occurring.*

For a discrete random variable with a finite range, we get the pmf by assigning each outcome the same probability of 1 divided by the number of outcomes. An example would be flipping a fair coin or rolling a fair die; the word fair suggests uniformity. However, we cannot define the pmf this way for a range such as $\mathbb{Z}$; think about how you could prove this.

Nonetheless, we still define the uniform distribution for a continuous random variable. Suppose Range(X) $= [a, b]$. Then, we assume each number in the interval is equally likely and we get the pdf to be:

$$f(x) = \begin{cases} \frac{1}{b-a}, \ x \in [a, b], \\ 0, \text{otherwise} \end{cases}$$

The cdf is then defined as:

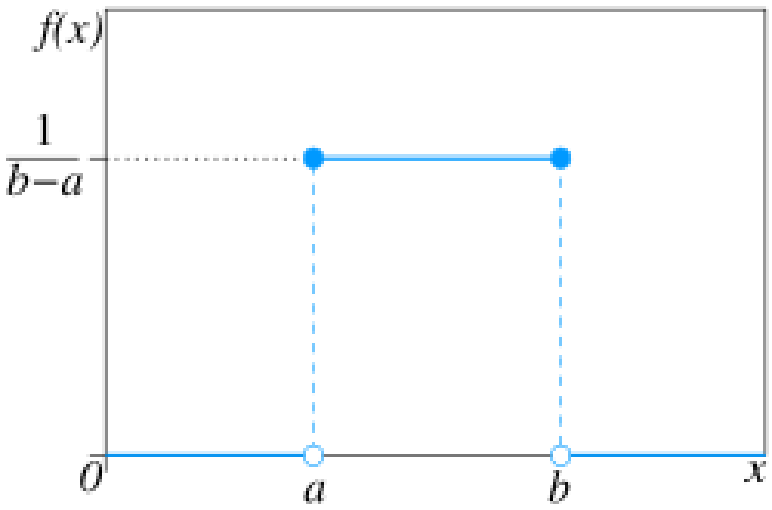
$$F(x) = \begin{cases} 0, \ x < a, \\ \frac{x-a}{b-a}, \ x \in [a, b], \\ 1, x > b \end{cases}$$

We write $X \sim U[a, b]$, with:

- $E[X] = \frac{a+b}{2}$,
- $Var(X) = \frac{(b-a)^2}{12}$

As we might expect, the expectation is the average of the two end points, so right in the middle.

The Uniform distribution is often called the **rectangular distribution** because of how the graph of the pdf looks



The Uniform distribution. Source: Wikipedia

We see how it is a horizontal line between a and b, and then zero elsewhere. The area under the curve being 1 is also clear: the base is $b-a$ and the height is $\frac{1}{b-a}$

A random number generator is a common example of a continuous uniform distribution. The following was Q6 in the 2023 Further Maths paper.

Question 4.3 *A game consists of two rounds. The first round of the game uses a random number generator to output the score X, a real number between 0 and 10.*

a) Find $P(X > 4)$

The second round of the game uses an unbiased 6-sided die to give the score Y. The variables X and Y are independent.

b) Find the mean total score of the game

c) Find the variance of the total score of the game

Solution 4.3 *(a) X has a uniform distribution on $[0, 10]$, and so $F(x) = \frac{x-0}{10-0} = \frac{x}{10}$. Hence:*

$$\begin{aligned} P(X > 4) &= 1 - P(X \leq 4) \\ &= 1 - F(4) \\ &= 1 - \tfrac{4}{10} \\ &= 0.6 \end{aligned}$$

There is a 60% chance of getting a number bigger than 4.
(b) The total score is $X + Y$. We know $E[X + Y] = E[X] + E[Y]$. Using the formula for a uniform, $E[X] = \frac{0+10}{2} = 5$. We previously found $E[Y] = 3.5$, and so $E[X + Y] = 8.5$.
(c) We also have that $Var(X + Y) = Var(X) + Var(Y)$. We previously found $Var(Y) = 2.92$. Then, for a continuous uniform distribution:

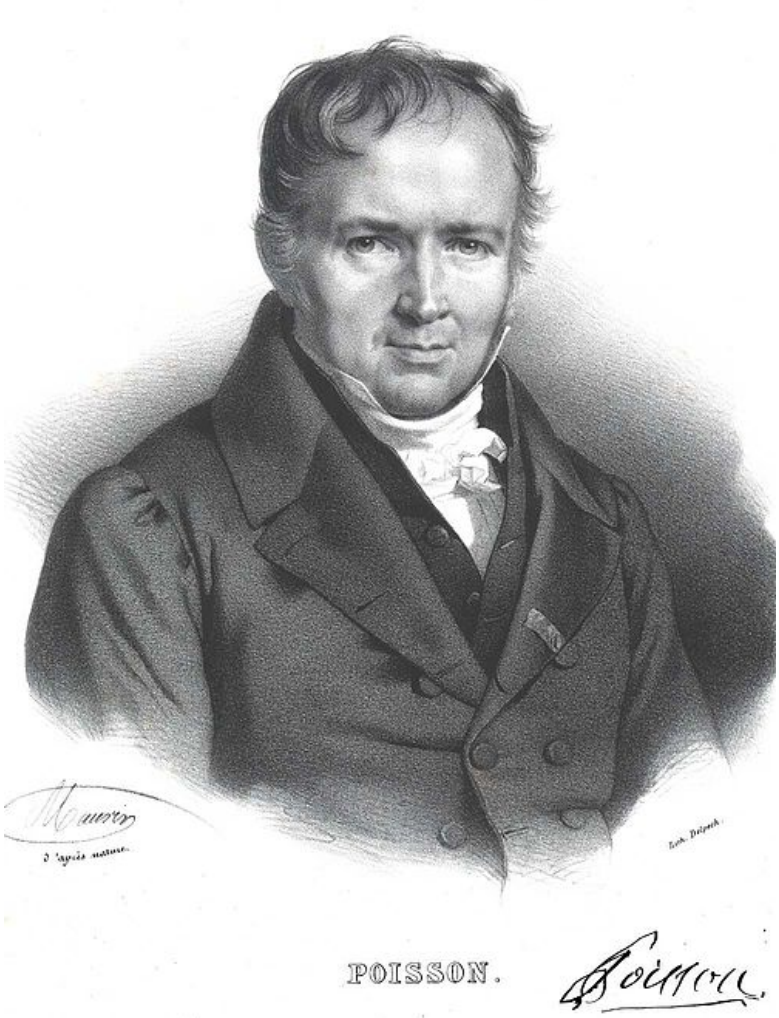
$$\begin{aligned} Var(X) &= \tfrac{(b-a)^2}{12} \\ &= \tfrac{10^2}{12} \\ &= 8.33 \end{aligned}$$

Hence, $Var(X + Y) = 8.33 + 2.92 = 11.25$

Question 3.2 was originally presented in the form of a discrete uniform distribution rather than as an n-sided die; they are equivalent.

4.5 Poisson Distribution

Often, we don't have a fixed number of trials but instead a fixed interval of time, and we wish to count the number of events that happen in that interval. The Poisson distribution is applicable in these cases.



French mathematician Siméon Denis Poisson (1781 - 1840).

Rather than considering whether United win or lose a match, suppose we want to predict how many goals they score in a game. There is no fixed number of trials; the Binomial can model how many times in 10 games they score more than 2 goals, for example, but not how likely a number of goals is in one game.

Instead, we consider the average number of goals they score in a game and use that to find the probability. This is the Poisson and requires the following assumptions.

As there is no fixed number of trials, there is no upper limit to the amount of events that can happen. *In theory*, United could score any amount of goals, although physical (and tactical) limitations prevent that in reality.

The average rate of goals should be assumed consistent within the interval. If United score 1.5 goals on average. this rate should be the same no matter what point in the game we are at. This is often a simplification; it is more common to score goals near the end of a game than the start.

Finally, the time taken for a goal should be independent of when the last goal was scored. It doesn't matter if the last goal was 5 minutes or 50 minutes ago, the probability I score now is the same.

Essentially, the Poisson is applicable when we have an average rate of occurrence, each occurrence has no effect on the others, and there is no maximum that can happen.

Definition 4.7 *Suppose a random variable X counts how often an event occurs over a fixed interval. X follows a* ***Poisson distribution*** *if:*

- $Range(X) = \mathbb{Z}^+ = \{0, 1, 2, \dots\}$
- *X has a constant mean rate λ over the fixed interval*
- *X is independent of the time since the last event*

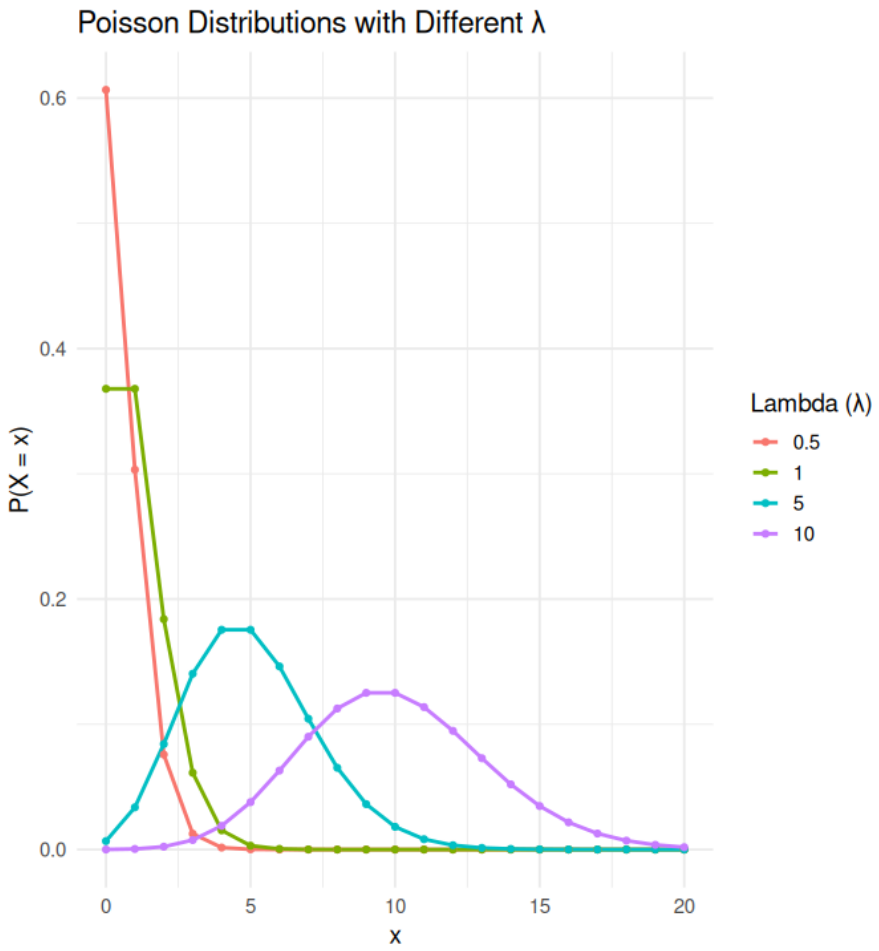
We write this as $X \sim Poi(\lambda)$. To calculate the probability we observe the event k times over the interval, we use the following pmf:

$$P(X = k) = \frac{\lambda^k}{k!} \exp(-\lambda)$$

For a $Poi(\lambda)$ distribution, we have:

- $E[X] = \lambda$
- $Var(X) = \lambda$

So for a Poisson distribution the mean and variance are equal.



The Poisson distribution for four different λ. Note for $\lambda = 10$ the distribution is starting to look Normal.

Note: X is not just a count, but a count *over a fixed interval*. It is the average number of goals scored over a 90 minute game. If the rate is given for one interval but we are interested in another, we can scale the rate λ. For example, we can divide λ by 90 to get the goals per minute and go from there. It is important to make sure our rate matches the interval.

Note: The interval of interest for the Poisson is generally time, but it could be something like length (such as accidents over a stretch of road) or area (such as trees in a forest). In linear modeling courses you will encounter such cases.

There is a link between the Poisson and Binomial.

Example 4.1 *Suppose we are interested in how often customers visit a shop. One approach would be to follow* 100 *people who walk by the shop, and count how many come in and how many walk by. This would be a* $Bin(100, p)$ *random variable, where* p *is the probability someone who walks by comes in and would be our final estimate.*

Another approach would be to wait for one hour in the shop and count how many people enter. This would be a $Poi(\lambda)$ *random variable, where* λ *is the average amount of people who come in every hour and would be our final estimate.*

In fact, when n is large and p is small (so $1 - p$ is close to 1 and $np(1 - p) \approx np$), a $Poi(np)$ distribution is a good approximation of the Binomial distribution. I like to think of the Poisson as the Binomial with an infinite number of trials; in each tiny interval of time you are checking whether the event happens or not. So when n is large they are similar.

So when n is large, the Poisson and Normal are both good approximations of the Binomial. As you might expect, this means the Normal is a good approximation of the Poisson as well when $\lambda = np$ is large.

4.6 Exponential Distribution

We saw the Poisson distribution as a way of counting how many events occur in a fixed interval. An alternative but equivalent way of thinking about this problem is as the time between events.

The number of customers entering a queue is often modeled with a Poisson. Suppose we know in one hour period that 2 people entered. Instead, I could say that 15 minutes in, one person came. Then, 35 minutes after that, another person came. 20 minutes after that a third person arrived. This third person came after the hour and so we know only two people came in the first hour.

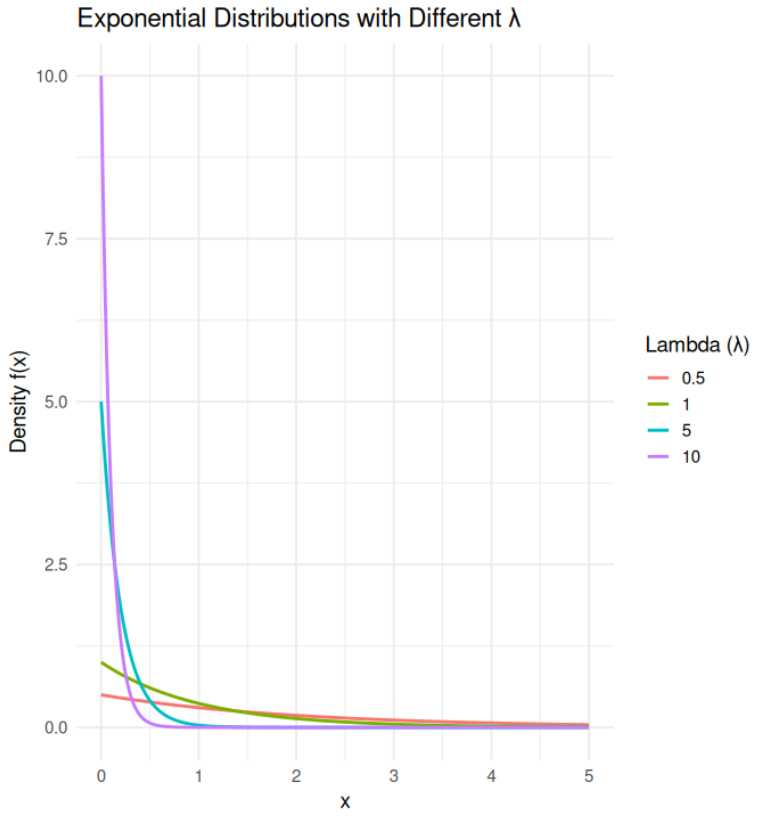
Definition 4.8 *Suppose $X \sim Poi(\lambda)$ and let T be the time between events. Then, T is a continuous random variable that follows the **Exponential** distribution with rate parameter λ, range $[0, \infty)$, and pdf*

$$f(t) = \begin{cases} \lambda e^{-\lambda t}, & t \geq 0 \\ 0, & t \leq 0 \end{cases}$$

The cdf is then:

$$F(t) = \begin{cases} 1 - e^{-\lambda t}, & t \geq 0 \\ 0, & t \leq 0 \end{cases}$$

We write $T \sim \exp(\lambda)$.



The Exponential distribution for the same four λ as the Poisson. For the two large λ, the waiting times tend to be shorter, as expected.

Note: Just as with the Poisson, the rate parameter is the rate *per unit time*. It is common to transition between the Exponential and Poisson distributions and you should be careful with what λ represents.

The Exponential distribution has a property similar to the Poisson.

Definition 4.9 *We say a random variable T has the **memoryless property** if:*

$$P(T > t + s | X > s) = P(T > t), \forall s, t \geq 0$$

That is, it doesn't matter how long has passed to the current point, we only need to look from now onwards. We have for $T \sim \exp(\lambda)$:

- $E[T] = \frac{1}{\lambda}$

- $Var(T) = \frac{1}{\lambda^2}$

The result for expectation should make sense when we think of the link to the Poisson; if we expect λ events to occur in one unit of time and the rate is constant, then we should expect the average time between each event to be $\frac{1}{\lambda}$.

The Exponential distribution is very important when measuring times between events, such as in risk analysis for insurance companies.

Key Takeaways:

- The Binomial distribution is for counting the number of successes in n independent Bernoulli trials
- The Normal distribution is common, simple and a good approximation of some discrete distributions
- The Uniform distribution assumes each outcome is equally likely
- The Poisson distribution counts the number of events in a fixed interval of time
- The Exponential distribution measures the time between events, rather than counting them over time like the Poisson

5 Worked Questions

The following was Q8 in 2019 Further Maths Paper 3.

Question 5.1 *The phone calls an office receives is modelled by a Poisson distribution with a mean rate of 3 calls per 10 minutes.*

a) Find the probability exactly 2 calls are received in 10 mins.

b) Find the probability the office receives more than 30 calls in an hour.

The office manager splits an hour into 6 10-minute periods and records the number of calls taken in each period.

c) Find the probability that the office receives exactly 2 calls in a 10-minute period exactly twice within an hour.

The office has just received a call.

d) Find the the probability the next call is received more than 10 minutes later

Mahah arrives at the office 5 minutes after the last call was received.

e) State the probability that the next call received by the office is received more than 10 minutes later. Explain your answer.

Solution 5.1 *Watch the video solution: 2019 Further Maths Paper 3 Question 8*

The following is Q11 in STEP II 2023.

Question 5.2

11 **(i)** X_1 and X_2 are both random variables which take values $x_1, x_2, \ldots, x_n$, with probabilities $a_1, a_2, \ldots, a_n$ and $b_1, b_2, \ldots, b_n$ respectively.

The value of random variable Y is defined to be that of X_1 with probability p and that of X_2 with probability $q = 1 - p$.

If X_1 has mean μ_1 and variance σ_1^2, and X_2 has mean μ_2 and variance σ_2^2, find the mean of Y and show that the variance of Y is $p\sigma_1^2 + q\sigma_2^2 + pq(\mu_1 - \mu_2)^2$.

(ii) To find the value of random variable B, a fair coin is tossed and a fair six-sided die is rolled. If the coin shows heads, then $B = 1$ if the die shows a six and $B = 0$ otherwise; if the coin shows tails, then $B = 1$ if the die does **not** show a six and $B = 0$ if it does. The value of Z_1 is the sum of n independent values of B, where n is large.

Show that Z_1 is a Binomial random variable with probability of success $\frac{1}{2}$.

Using a Normal approximation, show that the probability that Z_1 is within 10% of its mean tends to 1 as $n \longrightarrow \infty$.

(iii) To find the value of random variable Z_2, a fair coin is tossed and n fair six-sided dice are rolled, where n is large. If the coin shows heads, then the value of Z_2 is the number of dice showing a six; if the coin shows tails, then the value of Z_2 is the number of dice **not** showing a six.

Use part **(i)** to write down the mean and variance of Z_2.

Explain why a Normal distribution with this mean and variance will not be a good approximation to the distribution of Z_2.

Show that the probability that Z_2 is within 10% of its mean tends to 0 as $n \longrightarrow \infty$.

Solution 5.2 *Watch the video solution: 2023 Step II Question 11*