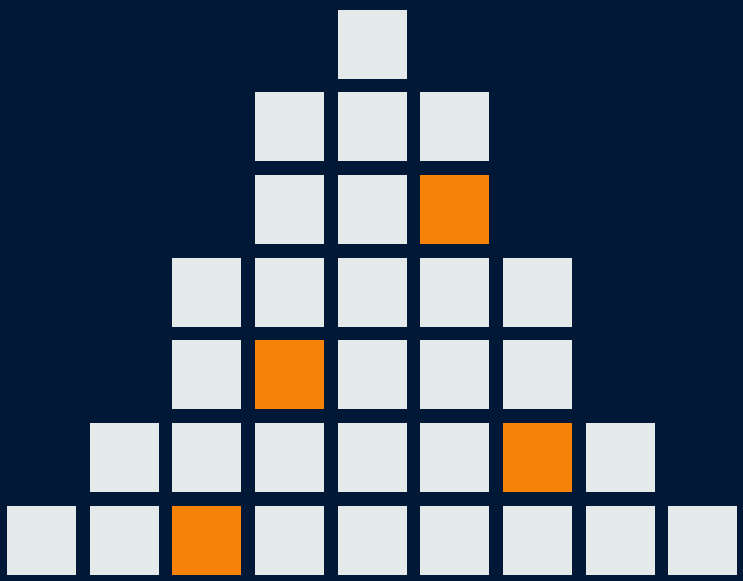
DATA SCIENCE INFINITY

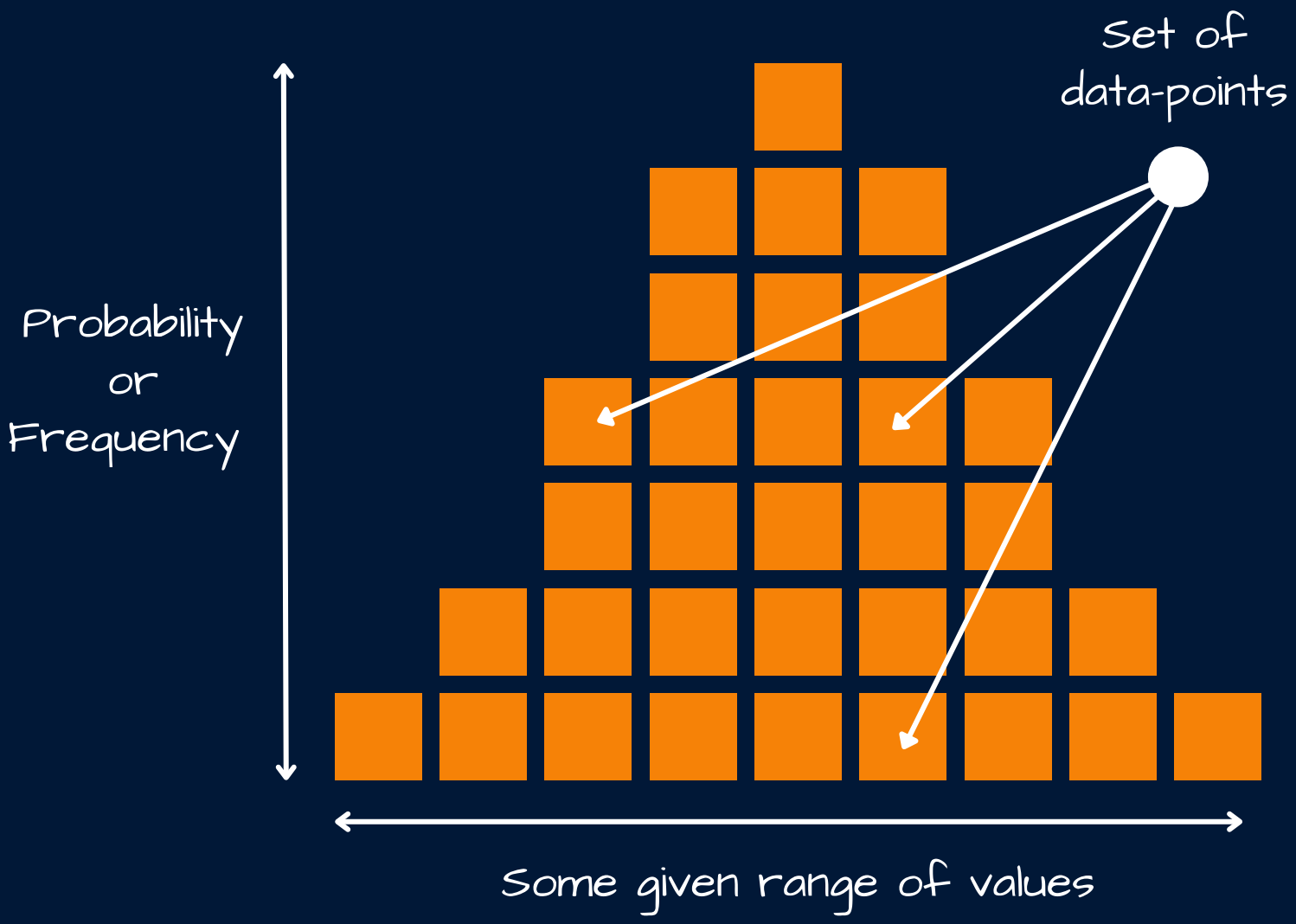
COMMON STATISTICAL DISTRIBUTIONS



AN OVERVIEW

What?

In Statistics, a **distribution** is simply a way to understand how a set of data points are spread over some given range of values.

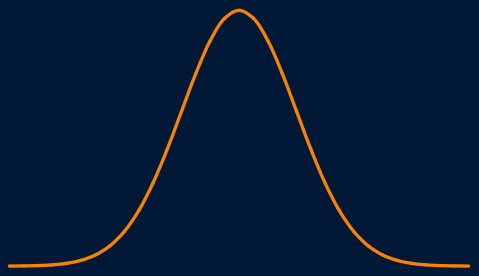


DATA SCIENCE INFINITY

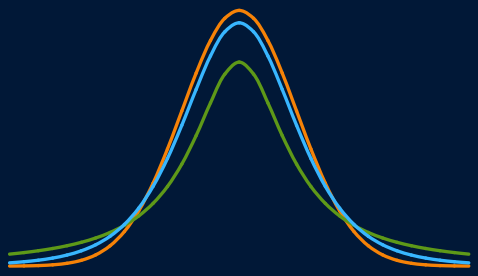
Types of distribution...

There are many, many different types of statistical distribution - each of which represent different types of data, and/or serve different purposes...

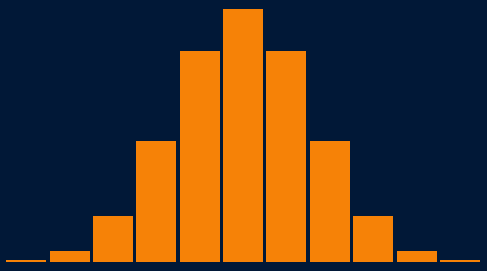
Here we will cover several commonly used distributions...



Normal Distribution



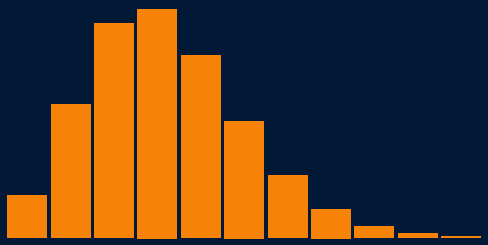
t-Distribution



Binomial Distribution

Bernoulli Distribution

Uniform Distribution



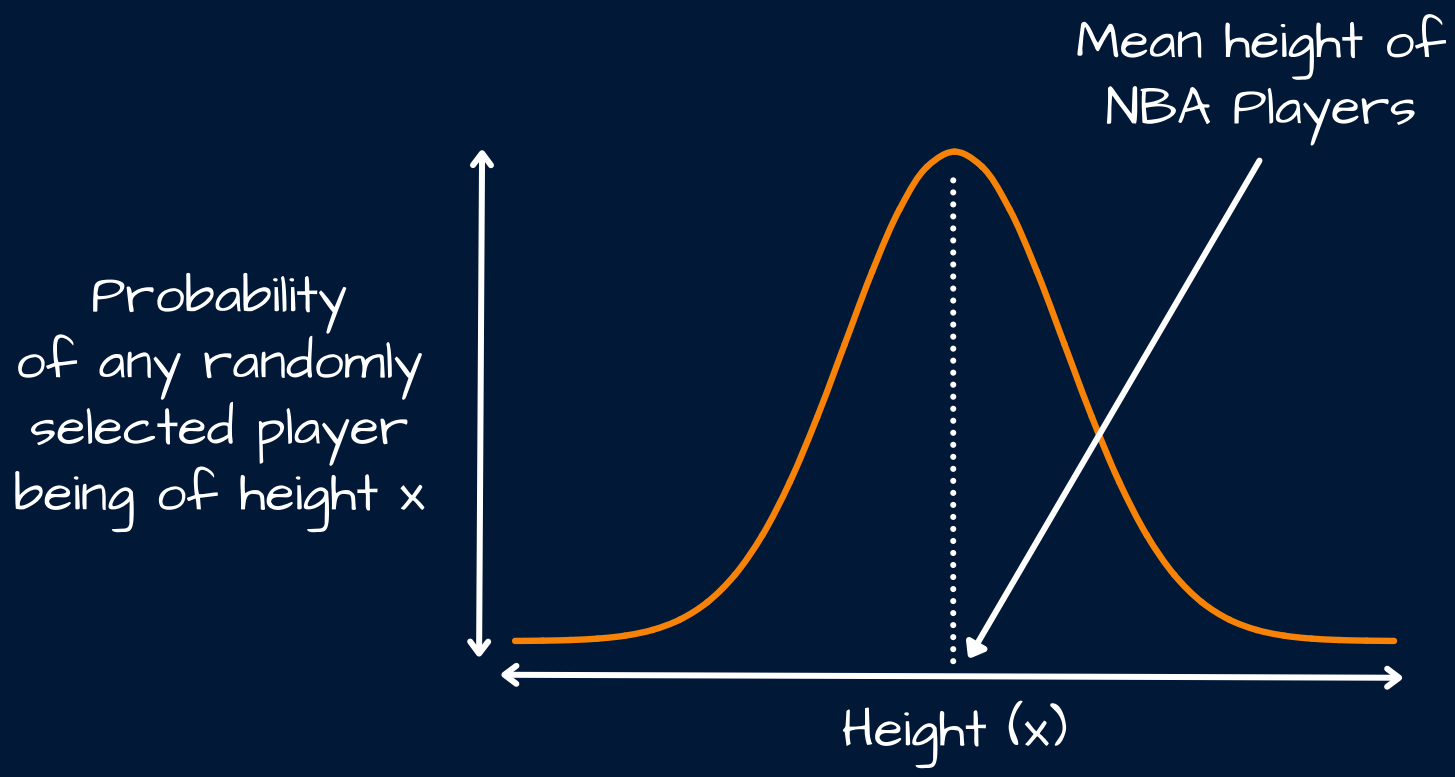
Poisson Distribution

Normal Distribution 1

A normal distribution shows the **probability density** for a population of **continuous** data (for example *height in cm* for all NBA players)

In other words, it shows how likely is it that any player from the NBA is of a certain height. Most players are around the mean/average height, fewer are much taller, or much shorter.

A normal distribution is symmetrical both sides of the mean. You might also see this referred to as a Gaussian Distribution!

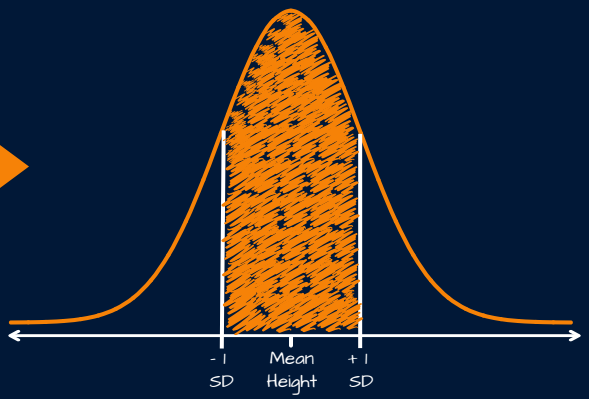


Normal Distribution 2

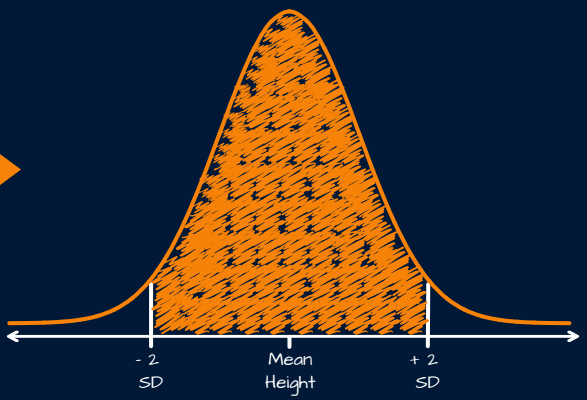
The spread of the values in our population is measured using a metric called **standard deviation**.

The **Empirical Rule** tells us that...

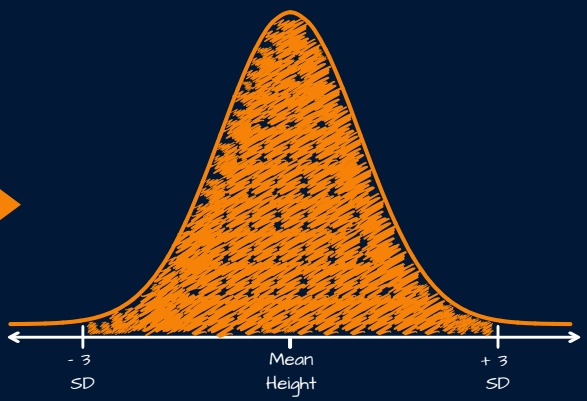
68.3% of the values will fall between **1 standard deviation** above and below the mean



95.5% of the values will fall between **2 standard deviations** above and below the mean



99.7% of the values will fall between **3 standard deviations** above and below the mean



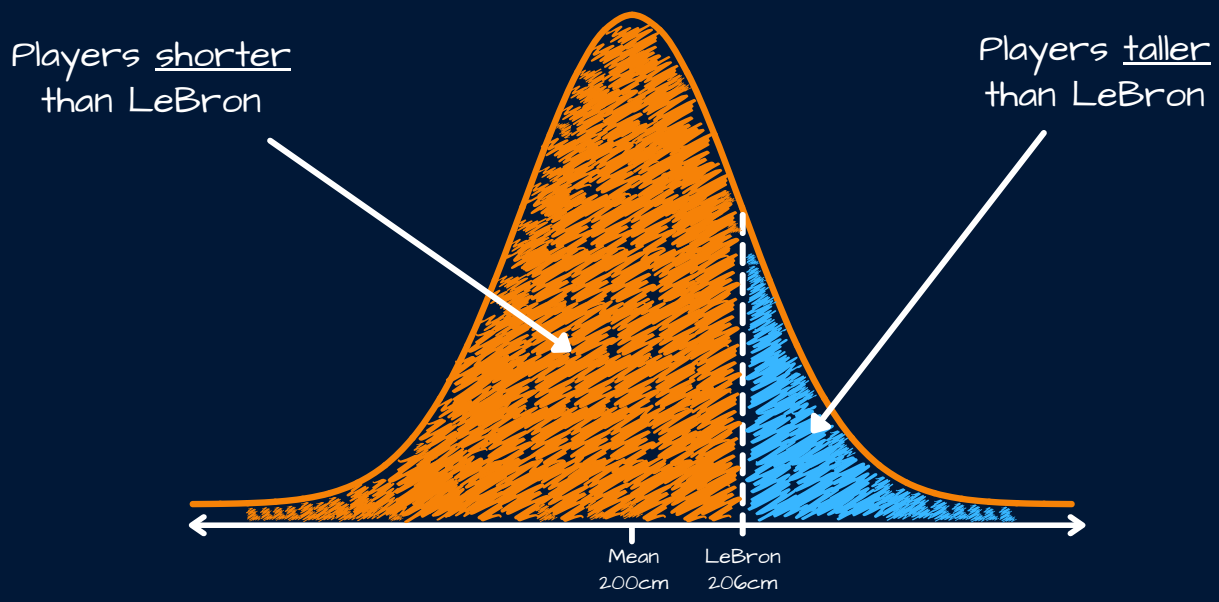
Normal Distribution 3

As an example...

Say we know that for all players in the NBA, the **mean height is 200cm** and the **standard deviation is 7cm**.

If LeBron James is 206cm tall - what proportion of NBA players is he taller than? We can figure this out!

LeBron is 6cm taller than the mean (206cm - 200cm). Since the standard deviation is 7cm, he is 0.86 standard deviations (6cm / 7cm) *above the mean.*

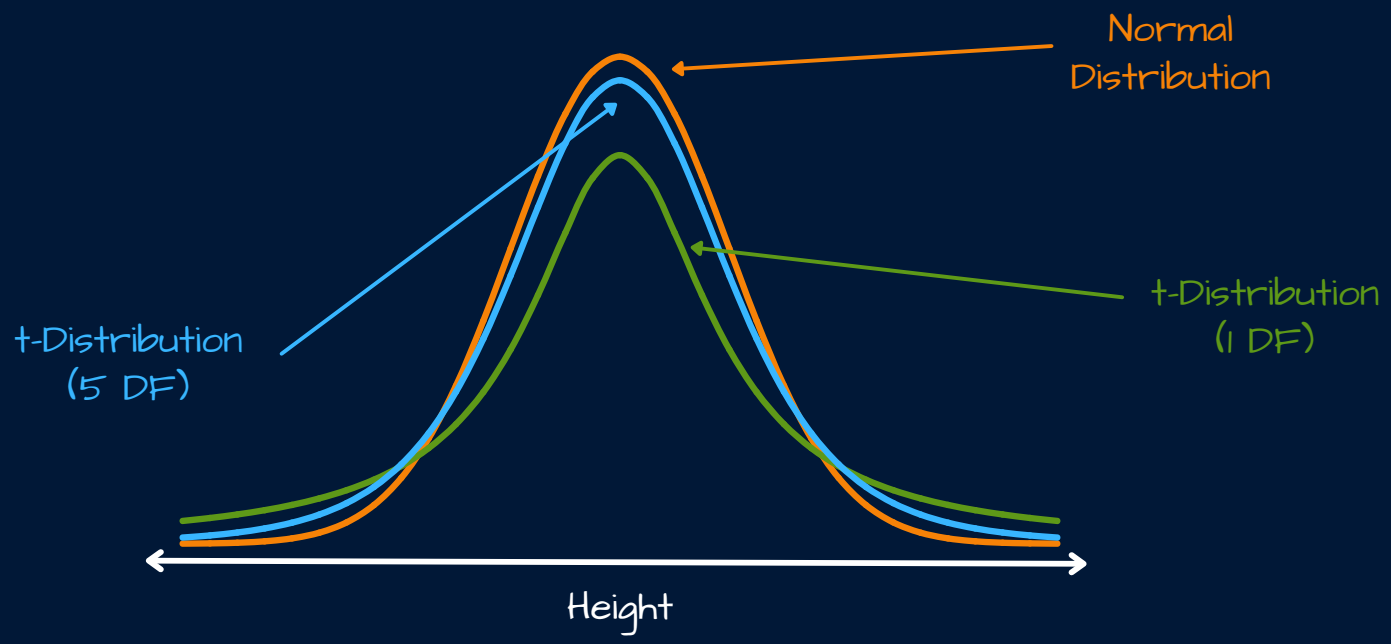


Our value of 0.86 standard deviations is called the **z-score**. This can be converted to a percentile using the probability density function (or a lookup table) giving us our answer. **LeBron James is taller than 80.5% of players in the NBA!**

t-Distribution

Just like a normal distribution, a **t-distribution** is symmetrical around the mean, and the breadth is based around the deviation within the data.

While a normal distribution works with a population - a t-distribution is designed for situations where **sample size is small**. The shape of the t-distribution becomes broader as the sample size decreases, to take into account the extra uncertainty we are faced with.



The shape of a t-distribution relates to the number of **degrees of freedom** which is calculated as the **sample size minus one**.

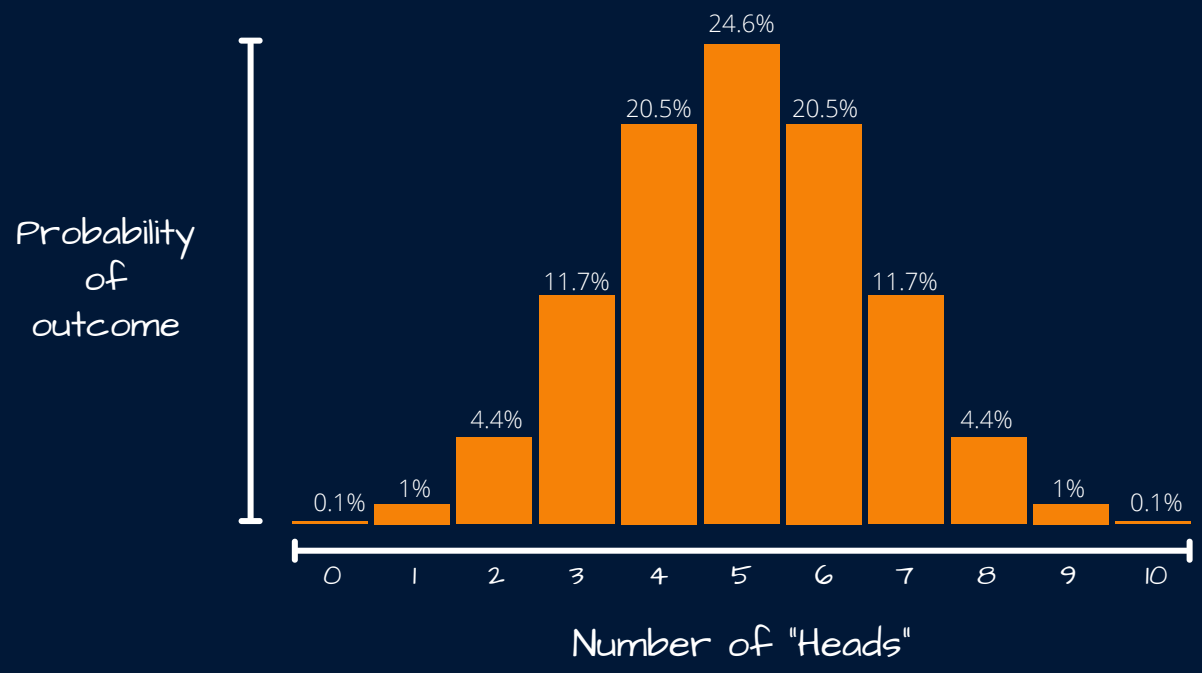
As the sample size, and thus the degrees of freedom gets larger, the t-distribution tends towards a normal distribution - as with a larger sample we're more certain around estimating the true population statistics.

Binomial Distribution

A Binomial Distribution can end up looking a lot like the shape of a normal distribution. The main difference is that instead of plotting continuous data, it instead plots a distribution of **two possible discrete outcomes** for example, the results from flipping a coin.

Imagine flipping a coin 10 times, and from those 10 flips, noting down how many were "Heads". It could be any number between 1 and 10.

Now imagine repeating that task 1,000 times...



If the coin we are using is indeed fair (not biased to heads or tails) then the distribution of outcomes should start to look the plot above. In the vast majority of cases we get 4, 5, or 6 "heads" from each set of 10 flips, and the likelihood of getting more extreme results is much more rare!

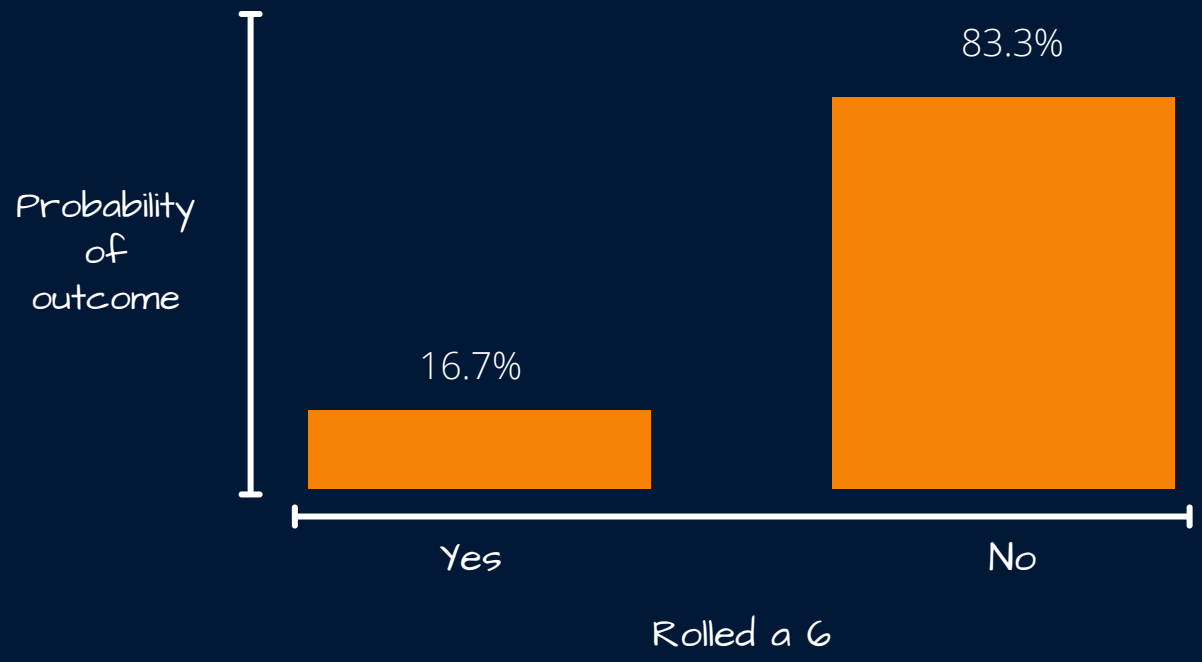
Bernoulli Distribution

The **Bernoulli Distribution** is a special case of the Binomial Distribution. It considers only two possible outcomes, success or failure, true or false.

It's a really simple distribution, but worth knowing!

In the example below we're looking at the probability of **rolling a 6** with a standard die.

If we roll a die many, many times, we should end up with a probability of rolling a 6, 1 out of every 6 times (or 16.7%) and thus a probability of not rolling a 6, in other words rolling a 1,2,3,4 or 5, 5 times out of 6 (or 83.3%) of the time!



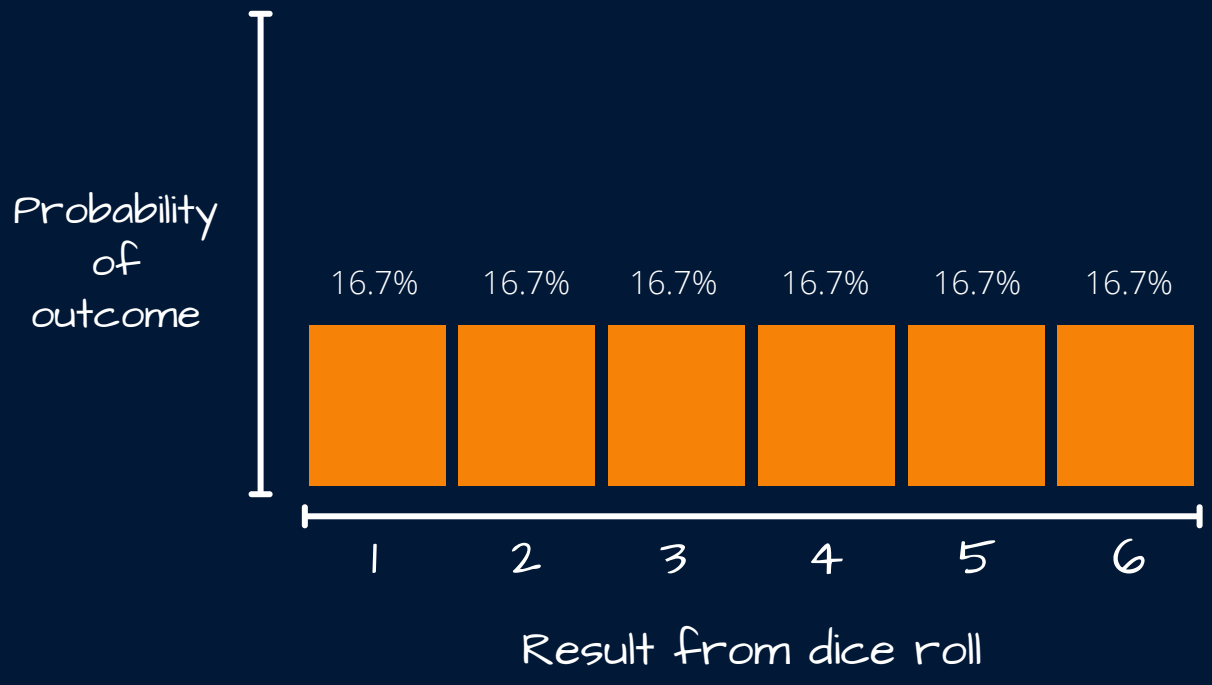
Uniform Distribution

A **Uniform Distribution** is a distribution in which all events are equally likely to occur.

Below, we're looking at the results from rolling a die many, many times.

We're looking at which number we got on each roll and tallying these up.

If we roll the die enough times (and the die is fair) we should end up with a completely uniform probability where the chance of getting any outcome is exactly the same



Poisson Distribution

A **Poisson Distribution** is a **discrete** distribution similar to the Binomial Distribution (in that we're plotting the probability of whole numbered outcomes)

Unlike the other distributions we have seen however, this one is **not symmetrical** - it is instead bounded between 0 and infinity

The Poisson distribution describes the number of events or outcomes that occur during some fixed interval. Most commonly this is a time interval like in our example below where we are plotting the distribution of **sales per hour** in a shop.