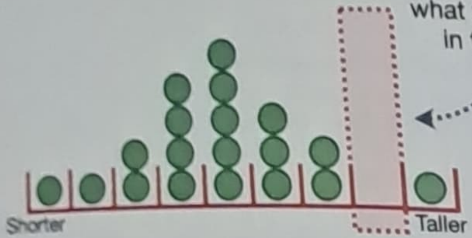
Continuous Probability Distributions: Main Ideas

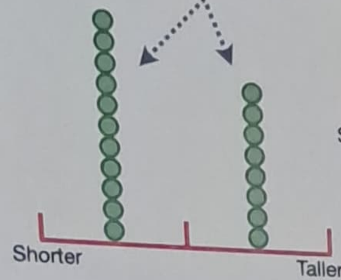
1) **The Problem:** Although they can be super useful, beyond needing a lot of data, histograms have two problems when it comes to continuous data:

1) it's not always clear what to do about gaps in the data and...

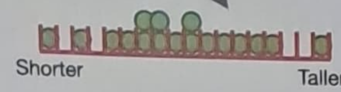


...**2)** histograms can be very sensitive to the size of the bins.

If the bins are too wide, then we lose all of the precision...

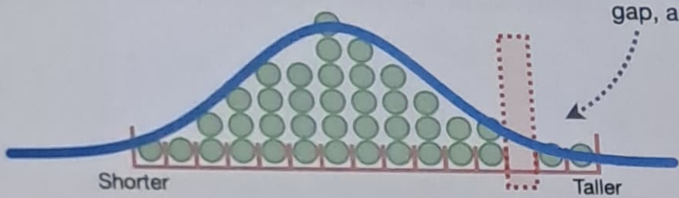


...and if the bins are too narrow, it's impossible to see trends.



2) **A Solution:** When we have continuous data, a **Continuous Distribution** allows us to avoid all of these problems by using mathematical formulas just like we did with **Discrete Distributions**.

In this example, we can use a **Normal Distribution**, which creates a bell-shaped curve, instead of a histogram. It doesn't have a gap, and there's no need to fiddle with bin size.



There are lots of commonly used **Continuous Distributions**. Now we'll talk about the most useful of all, the **Normal Distribution**.

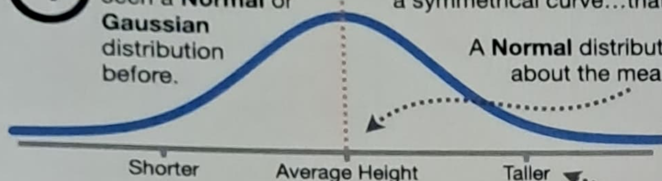
The Normal (Gaussian) Distribution: Main Ideas Part 1

1 Chances are you've seen a **Normal** or **Gaussian** distribution before.

It's also called a **Bell-Shaped Curve** because it's a symmetrical curve...that looks like a bell.

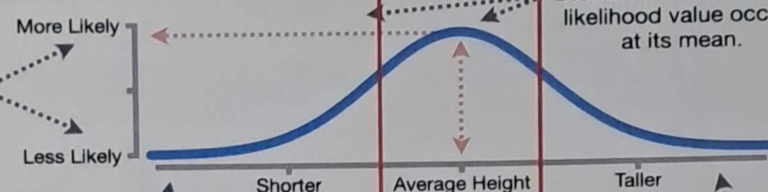
A **Normal** distribution is symmetrical about the mean, or average, value.

In this example, the curve represents human Height measurements.



2 The y-axis represents the **Likelihood** of observing any specific Height.

The **Normal Distribution**'s maximum likelihood value occurs at its mean.

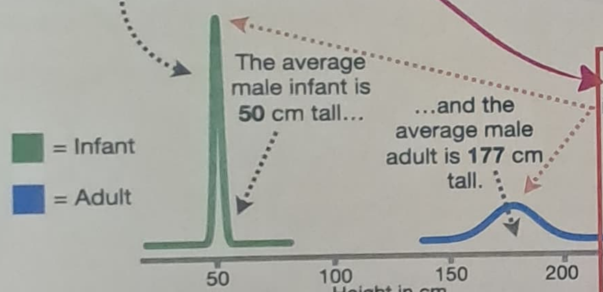


For example, it's relatively rare to see someone who is super short...

...relatively common to see someone who is close to the average height...

...and relatively rare to see someone who is super tall.

3 Here are two **Normal Distributions** of the heights of male infants and adults.

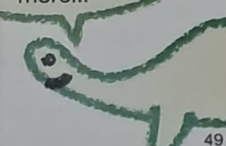


The average male infant is **50** cm tall...

...and the average male adult is **177** cm tall.

Because the normal distribution for infants has a higher peak than the one for adults, we can see that there's a higher likelihood that an infant will be close to its mean than an adult will be close to its mean. The difference in peak height tells us there's less variation in how tall an infant is compared to how tall an adult is.

Lots of things can be approximated with **Normal Distributions**: Height, birth weight, blood pressure, job satisfaction, and many more!!!



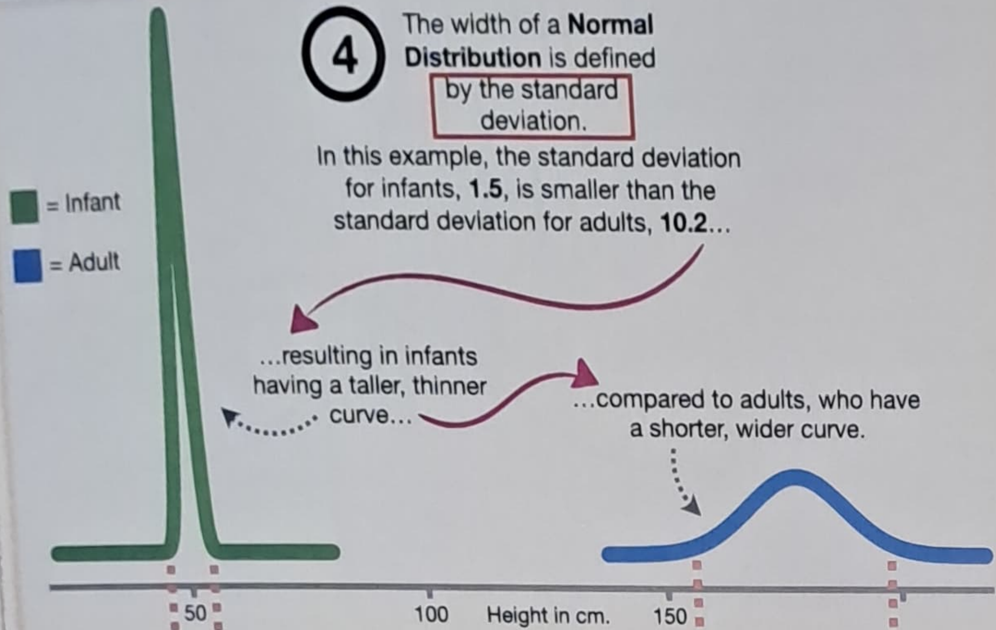
The Normal (Gaussian) Distribution: Main Ideas Part 2

4 The width of a **Normal Distribution** is defined by the standard deviation.

In this example, the standard deviation for infants, **1.5**, is smaller than the standard deviation for adults, **10.2**...

...resulting in infants having a taller, thinner curve...

...compared to adults, who have a shorter, wider curve.



5 Knowing the standard deviation is helpful because normal curves are drawn such that about **95%** of the measurements fall between **+/- 2 Standard Deviations** around the **Mean**.

Because the mean measurement for infants is **50** cm, and **2 x the standard deviation = 2 x 1.5 = 3**, about **95%** of the infant measurements fall between **47** and **53** cm.

Because the mean adult measurement is **177** cm, and **2 x the standard deviation = 2 x 10.2 = 20.4**, about **95%** of the adult measurements fall between **156.6** and **197.4** cm.

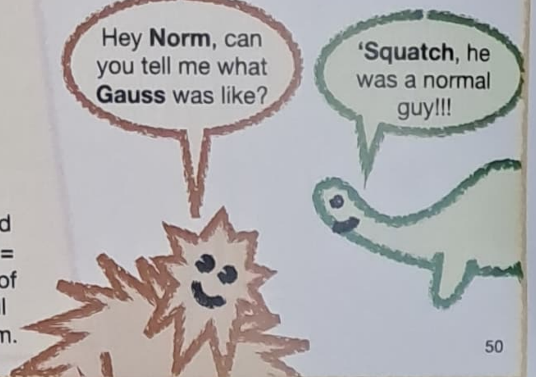
To draw a **Normal Distribution**, you need to know:

1) The **Mean** or average measurement. This tells you where the center of the curve goes.

2) The **Standard Deviation** of the measurements. This tells you how tall and skinny, or short and fat, the curve should be.

If you don't already know about the **Mean** and **Standard Deviation**, check out **Appendix B**.

BAM!!!



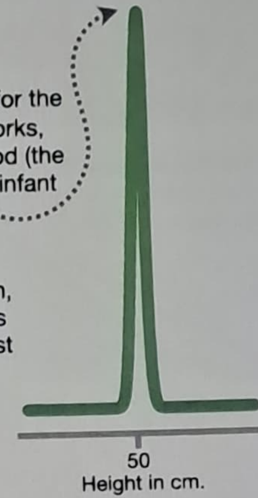
The Normal (Gaussian) Distribution: Details

1 The equation for the **Normal Distribution** looks scary, but, just like every other equation, it's just a matter of plugging in numbers and doing the math.

$$f(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$$

2 To see how the equation for the **Normal Distribution** works, let's calculate the likelihood (the y-axis coordinate) for an infant that is **50** cm tall.

Since the mean of the distribution is also **50** cm, we'll calculate the y-axis coordinate for the highest part of the curve.



3 ***x*** is the x-axis coordinate. So, in this example, the x-axis represents **Height** and ***x*** **= 50**.

The Greek character ***μ, mu***, represents the mean of the distribution. In this case, ***μ*** **= 50**.

Lastly, the Greek character ***σ***, ***sigma***, represents the standard deviation of the distribution. In this case, ***σ*** **= 1.5**.

$$f(x = 50 \mid \mu = 50, \sigma = 1.5) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$$

Now, we just do the math....

$$= \frac{1}{\sqrt{2\pi 1.5^2}} e^{-(50-50)^2/(2\times 1.5^2)}$$

$$= \frac{1}{\sqrt{14.1}} e^{-0^2/4.5}$$

$$= \frac{1}{\sqrt{14.1}} e^{0}$$

$$= \frac{1}{\sqrt{14.1}}$$

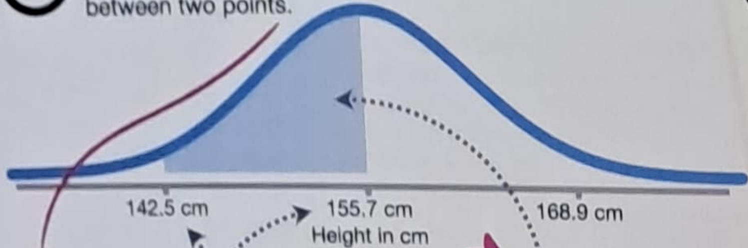
$$\boxed{= 0.27}$$

...and we see that the likelihood, the y-axis coordinate, for the tallest point on the curve, is **0.27**.

Remember, the output from the equation, the y-axis coordinate, is a **likelihood**, ***not*** a **probability**. In **Chapter 7**, we'll see how likelihoods are used in **Naive Bayes**. To learn how to calculate probabilities with **Continuous Distributions**, read on...

Calculating Probabilities with Continuous Probability Distributions: Details

1 For **Continuous Probability Distributions**, probabilities are the **area under the curve** between two points.

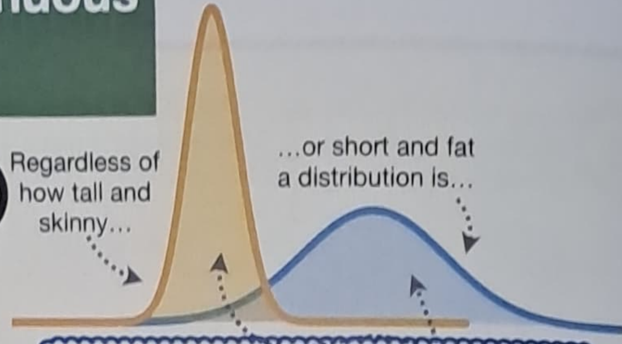


For example, given this **Normal Distribution** with **mean = 155.7** and **standard deviation = 6.6**, the probability of getting a measurement between **142.5** and **155.7** cm...

...is equal to this area under the curve, which in this example is **0.48**. So, the probability is **0.48** that we will measure someone in this range.

2 Regardless of how tall and skinny...

...or short and fat a distribution is...



...the total area under its curve is **1**. Meaning, the probability of measuring anything in the range of possible values is **1**.

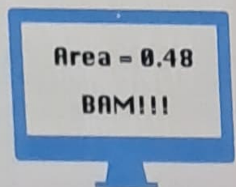
3 There are two ways to calculate the area under the curve between two points:

1) The hard way, by using calculus and integrating the equation between the two points ***a*** and ***b***.

$$\int_a^b f(x)\,dx$$

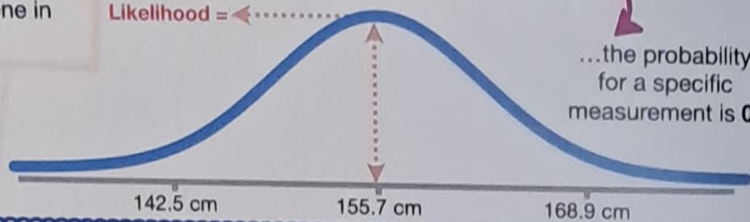
UGH!!! NO ONE ACTUALLY DOES THIS!!!

2) The easy way, by using a computer. See **Appendix C** for a list of commands.



4 One confusing thing about **Continuous Distributions** is that the while the likelihood for a specific measurement, like **155.7**, is the y-axis coordinate and > 0...

...the probability for a specific measurement is **0**.



One way to understand why the probability is **0** is to remember that probabilities are areas, and the area of something with no width is **0**.

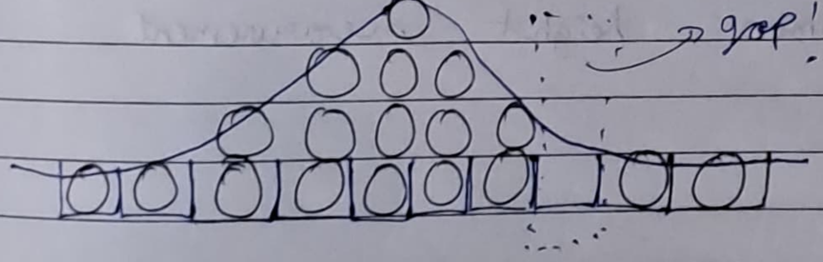
Another way is to realize that a continuous distribution has infinite precision, thus, we're really asking the probability of measuring someone who is *exactly* **155.7000000000000000000000 0000000000000000000000000000000000 00000000000000000000000000000...** tall.

Note: The Binomial Distribution is useful for anything that has binary outcomes (wins and losses, yeses & noes etc). but there are lots of other discrete probability distributions!

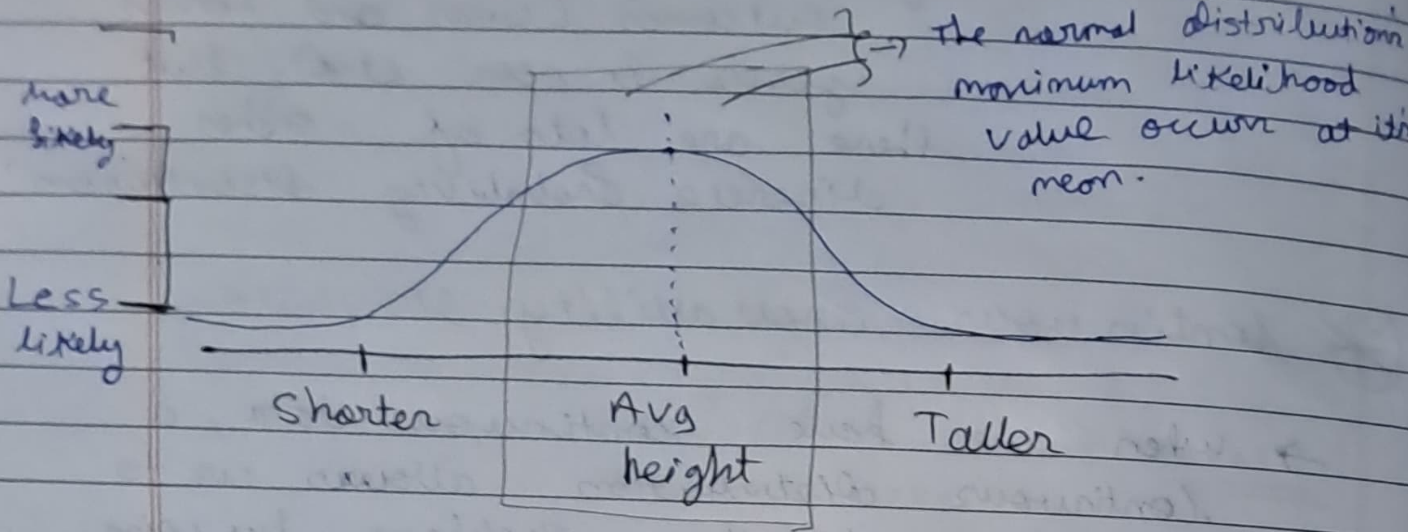
Continuous Probability Distribution

→ when we have continuous data, a continuous distribution allows us to avoid all of these problems by using mathematical formulas just like we did with discrete Distributions.

In the example below, we can use a Normal Distribution, which creates a bell-shaped curve, instead of a histogram. It doesn't have a gap, & there's no need to fiddle with bin size.



(1) The <u>Normal (Gaussian) Distribution</u>:-



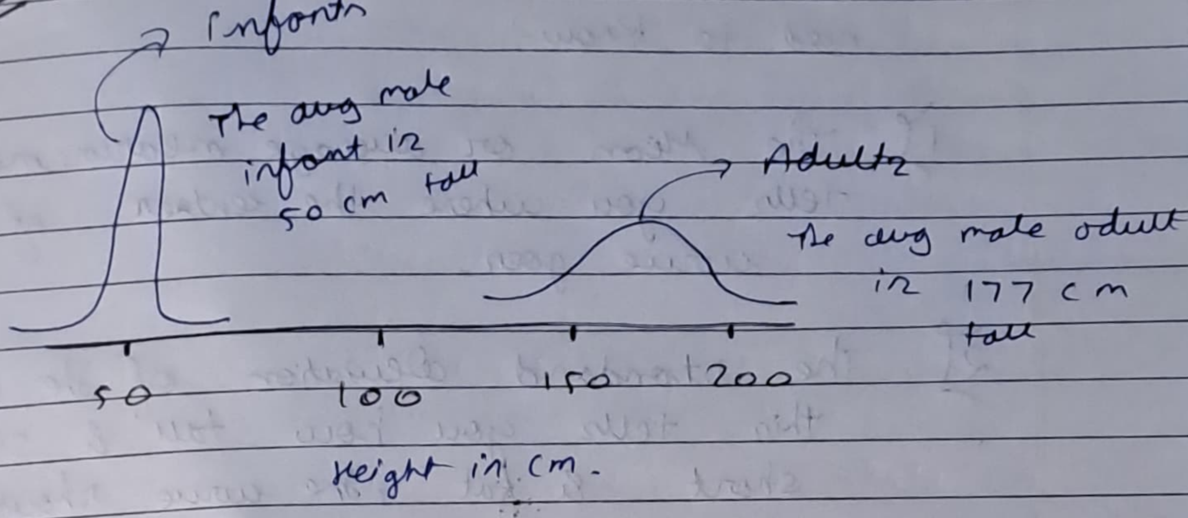
∴ It's also called a Bell-Shaped curve because it's a symmetric curve ... that looks like a bell.

∴ A Normal Distribution is symmetrical about the mean or average value.

∴ In this example, the curve represents human height measurement classmate

Date
Page

i.e.



∴ Because the normal distribution for infants has a higher peak than the one for adults, we can see that there's a higher likelihood that an infant will be close to its mean than an adult will be close to its mean. The difference in peak height tells us there's less variation in how tall an infant is compared to how tall an adult is.

∴ To draw a Normal Distribution, you need to know :-

1] The Mean or average measurement. This tells you where the center of the curve goes.

2] The standard deviation of the measurement this tells you how tall & skinny or short & fat, the curve should be.

IMP

(i) The width of a Normal Distribution is defined by the standard deviation.

(ii) Knowing the standard deviation is helpful because normal curves are drawn such that about 95% of the measurement fall between +/- 2 standard deviations around the mean.

classmate

Date

Page

$$\boxed{f(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}}$$

$x \rightarrow$ x-axis co-ordinate, so in this example x-axis represents height & $x = 50$.

$\mu \rightarrow$ mu $\rightarrow$ the mean of the distribution. in this case $\mu = 50$.

$\sigma \rightarrow$ sigma $\rightarrow$ standard deviation of the distribution. In this case. $\sigma = \underline{\underline{1.5}}$

$f(x = 50 \mid \mu = 50, \sigma = 1.5)$

$$= \frac{1}{\sqrt{2\pi(1.5)^2}} e^{-((50-50)^2/2\times(1.5)^2)}$$

$$= \frac{1}{\sqrt{14.1}} \cdot e^{-0^2/4.5}$$

$$= \underline{\underline{0.27}}$$

$\therefore$ we see that the likelihood, the y-axis co-ordinate, for the tallest point on the curve is $\underline{\underline{0.27}}$