# Bio:

George Igwegbe
Machine Learning Engineer

George Igwegbe is the Machine Learning Engineer at **Hankali Labs**. He has experience in **designing IoT hardware systems in metering (electricity)**. He's also worked in Hardware and Machine Learning.

He received his Bachelor's degree in Mechanical Engineering from **UNILAG**. He is a certified **Tensorflow Developer** and co-organizer of **TinyML Nigeria**. Curator of "**tinyml-papers-and-projects**".

George is interested in **Machine Learning on Embedded Systems (Safety Critical Systems)** and **Video Analytics**.

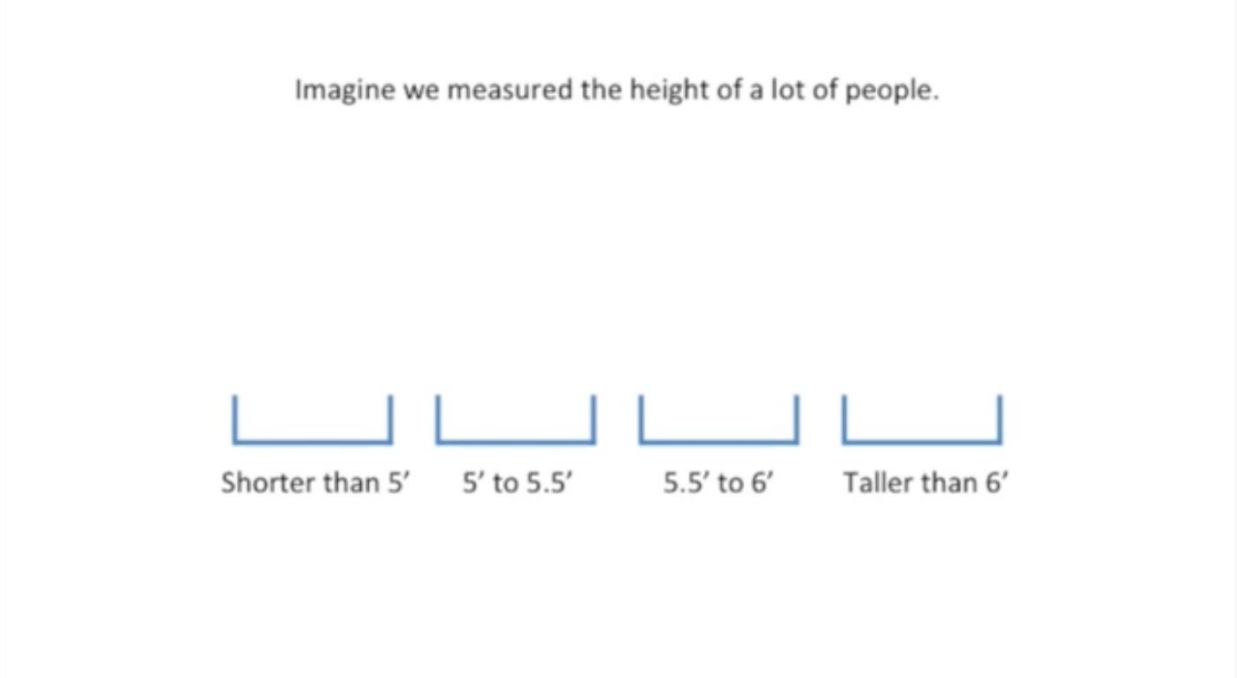
HANKALI LABS

Probability Distribution - Maximum Log Likelihood Estimation(MLE)

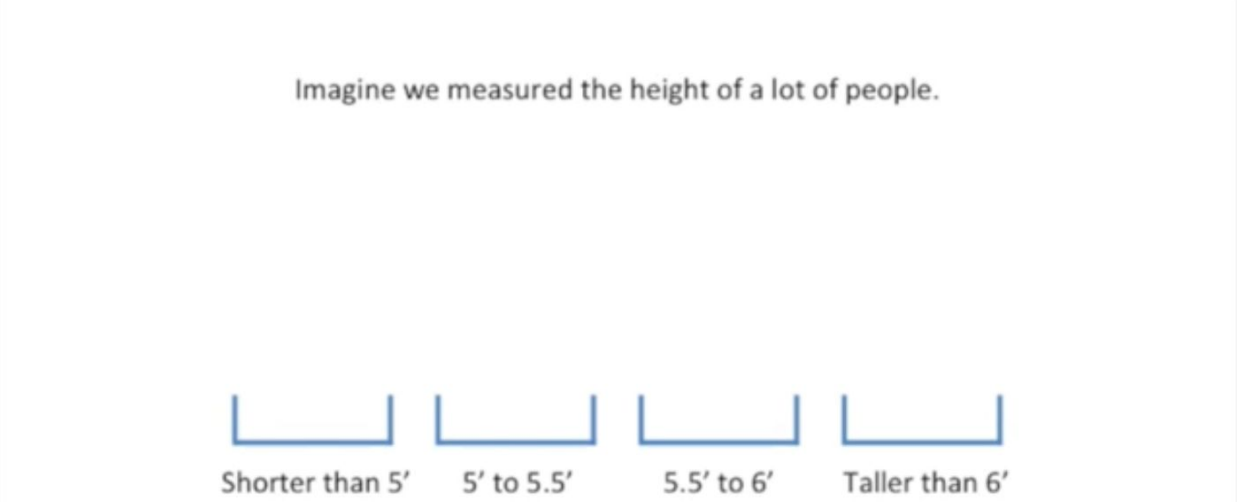
# Agenda

- What is a distribution in statistics
- MLE for Normal Distribution
- MLE for Normal Distribution in python

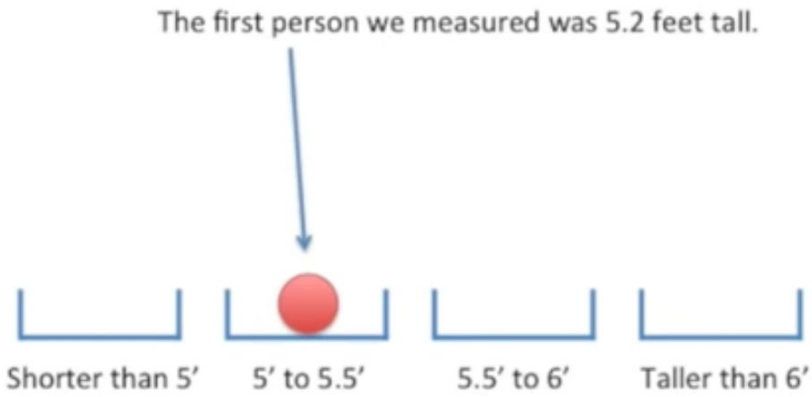
- What is a distribution in statistics?

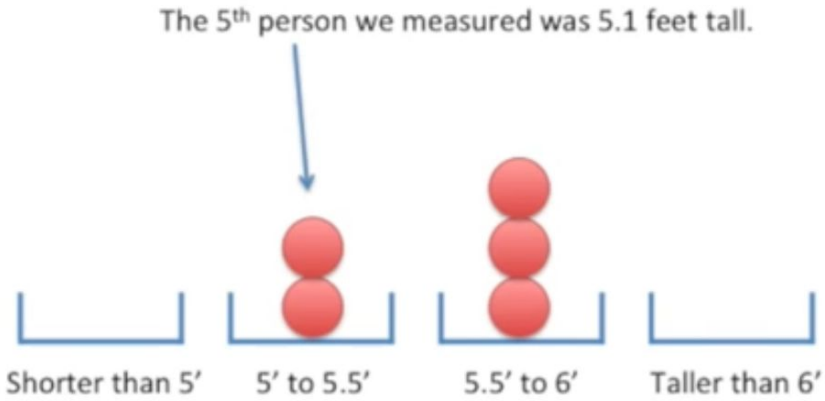


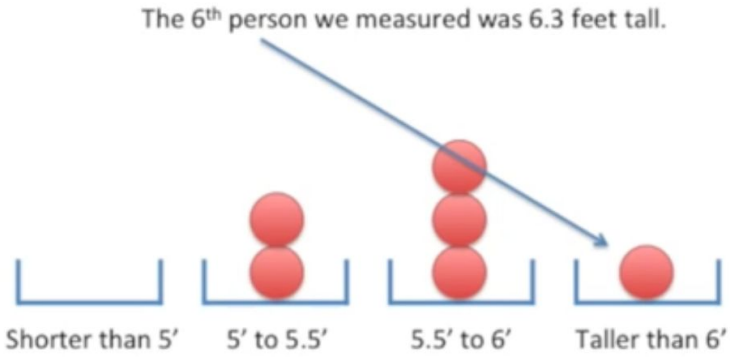
- What is a distribution in statistics?

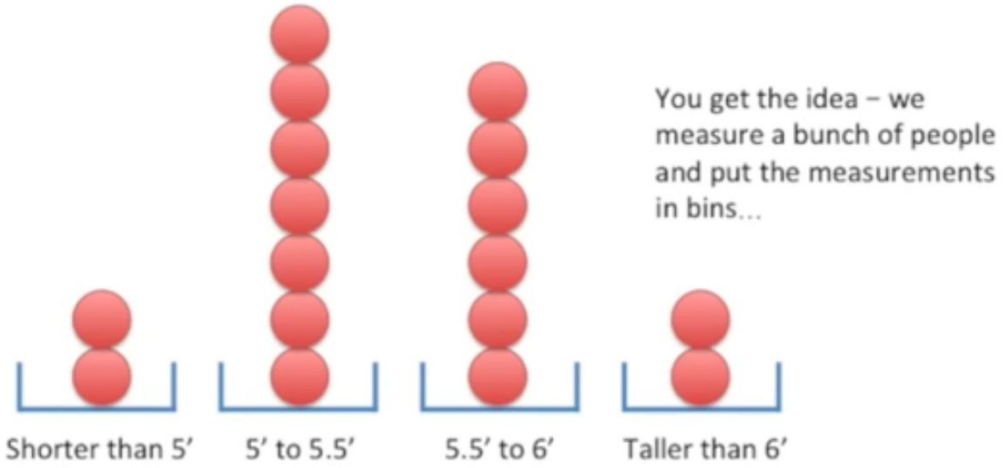



"A distribution is a function that shows the possible values for a variable and how often they occur"


The first person we measured was 5.2 feet tall.
Shorter than 5’
5’ to 5.5’
5.5’ to 6’
Taller than 6’

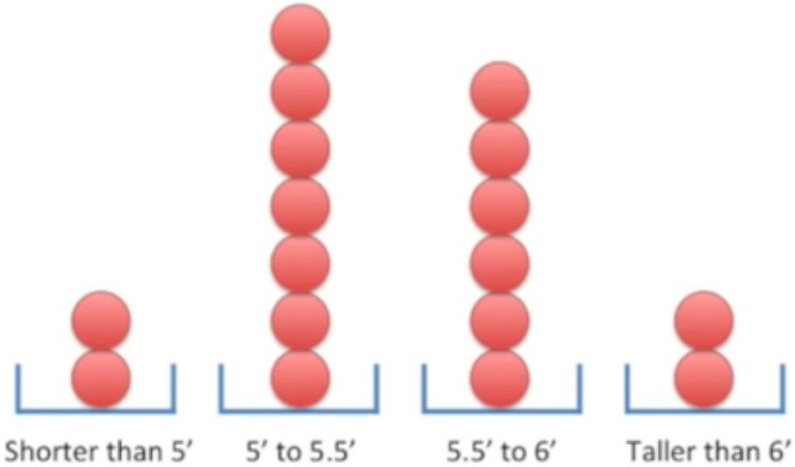

The 5th person we measured was 5.1 feet tall.
Shorter than 5’
5’ to 5.5’
5.5’ to 6’
Taller than 6’

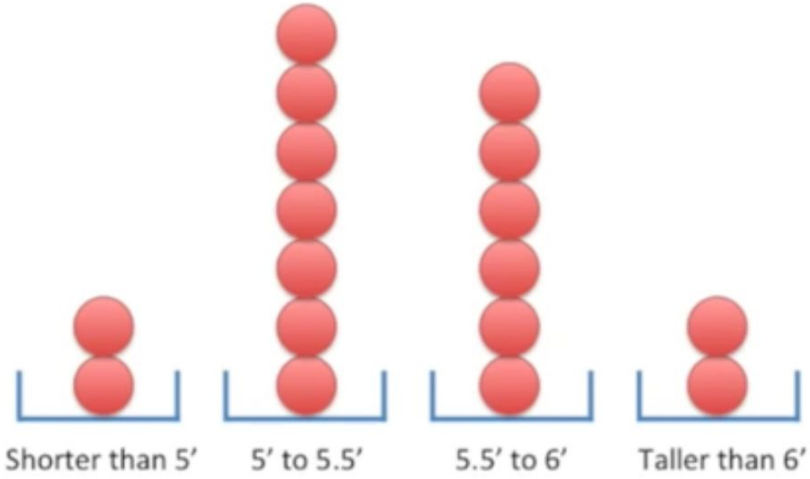

The 6th person we measured was 6.3 feet tall.
Shorter than 5’
5’ to 5.5’
5.5’ to 6’
Taller than 6’




You get the idea – we measure a bunch of people and put the measurements in bins…

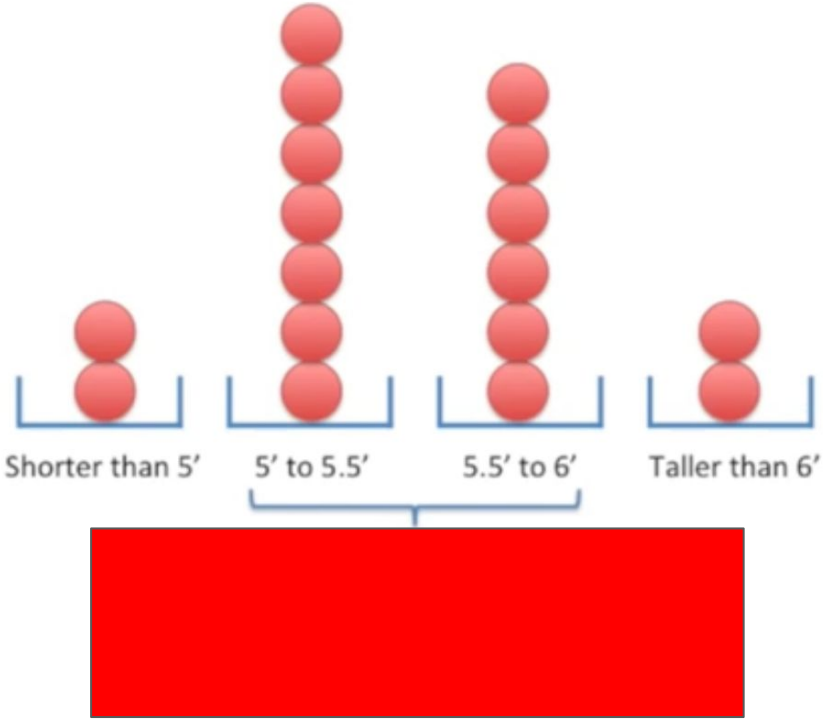
- What kind of chart is below?

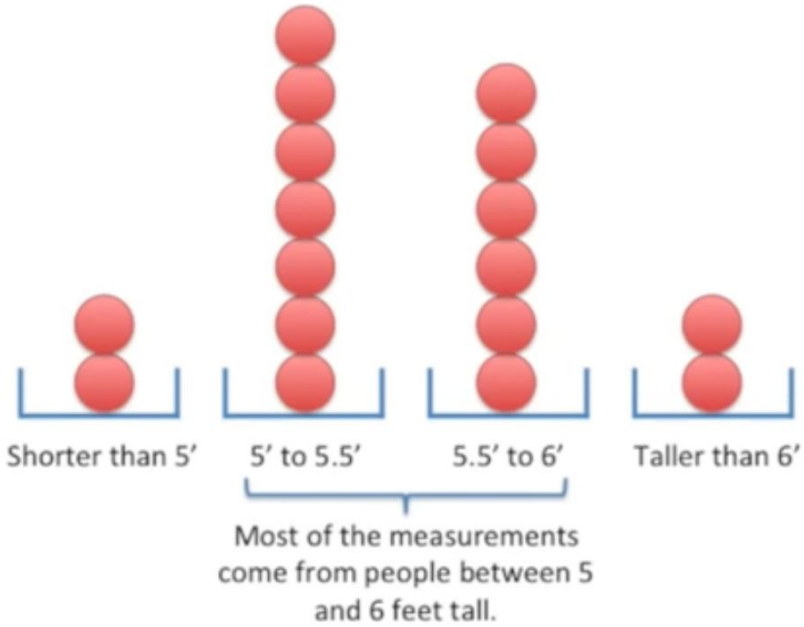


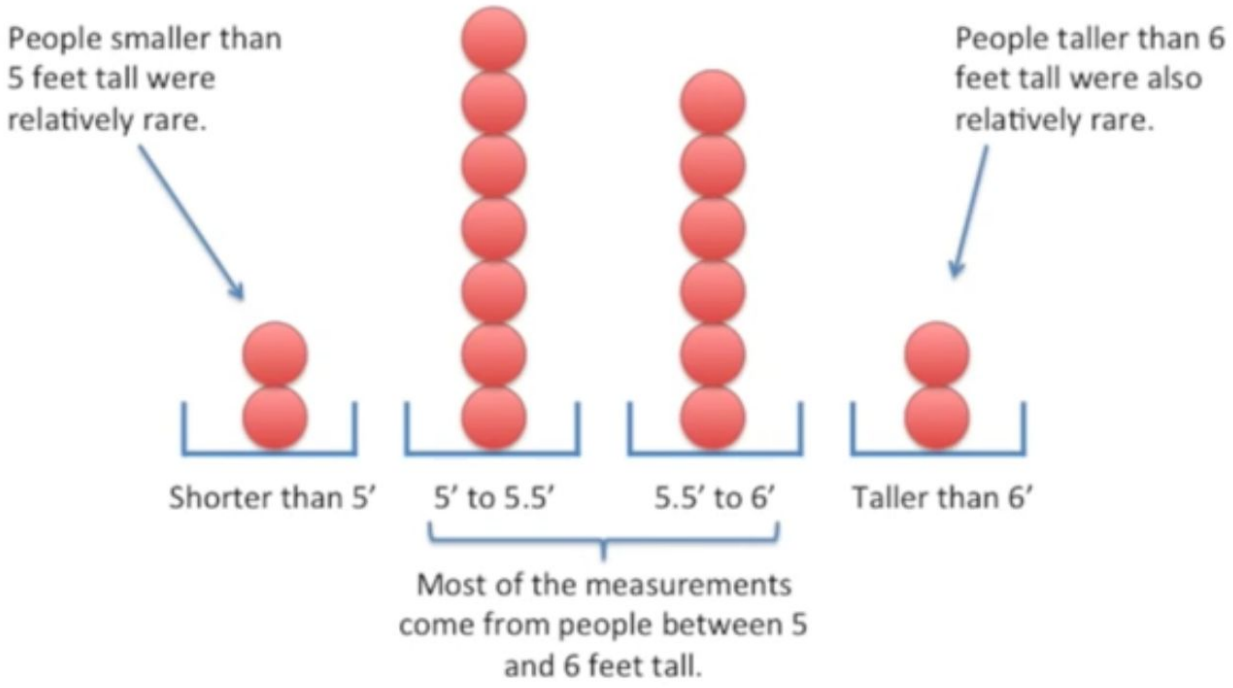


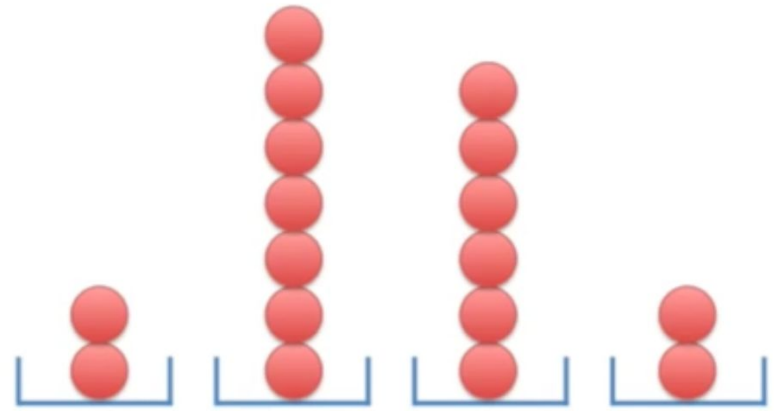

When you stack a bunch of measurements into bins like this, you get a **histogram**.

- What can we say about the data below?

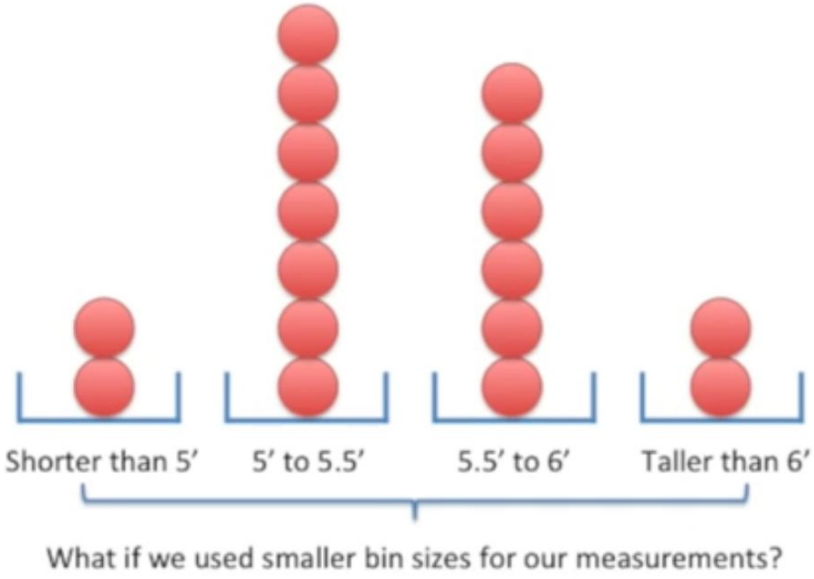


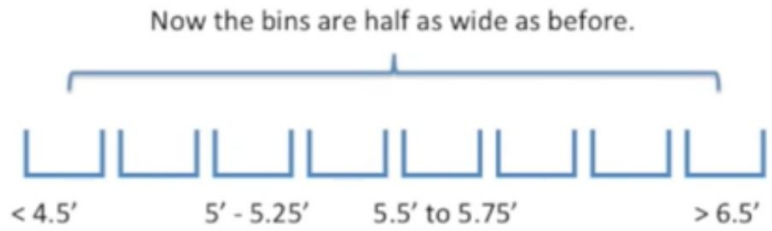

Shorter than 5’
5’ to 5.5’
5.5’ to 6’
Taller than 6’
Most of the measurements come from people between 5 and 6 feet tall.

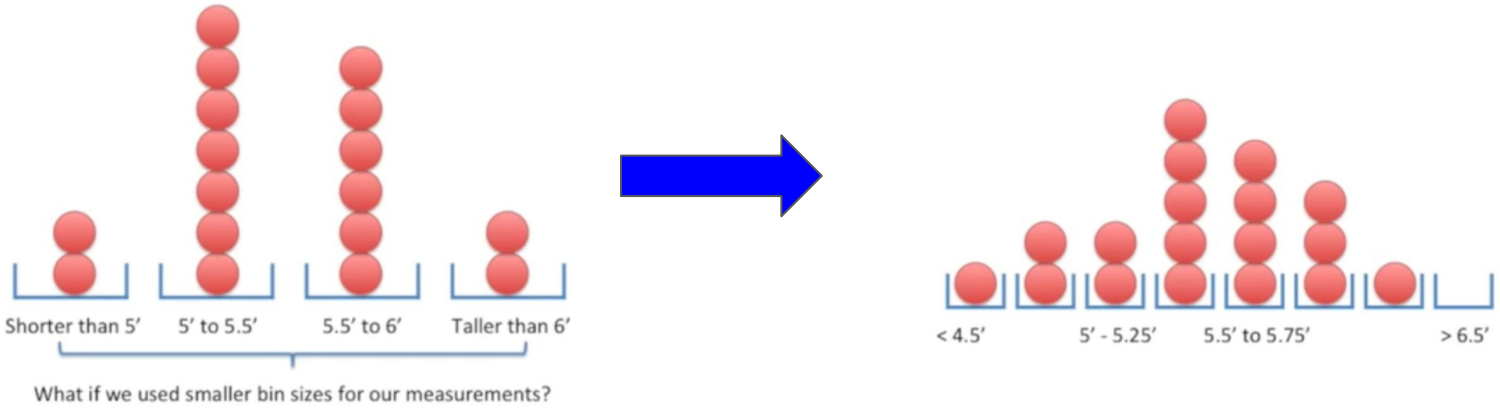

People smaller than 5 feet tall were relatively rare.
People taller than 6 feet tall were also relatively rare.
Shorter than 5'
5' to 5.5'
5.5' to 6'
Taller than 6'
Most of the measurements come from people between 5 and 6 feet tall.



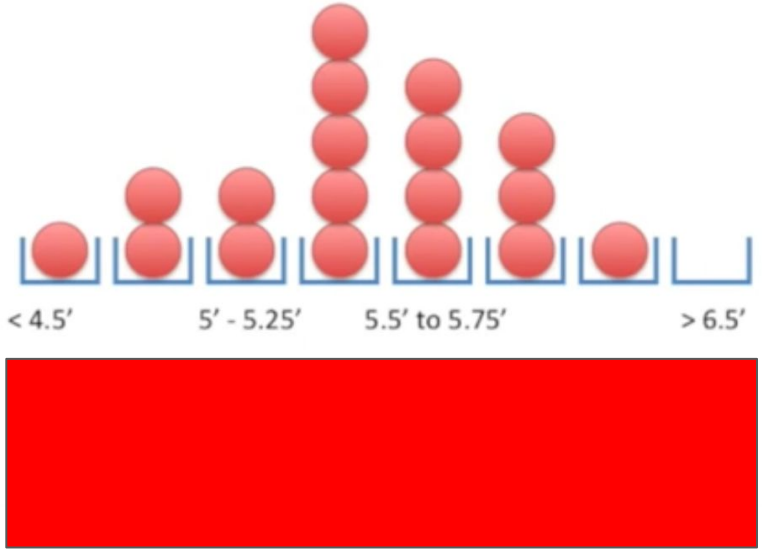

In other words, if you picked one measurement at random, there is a good chance it would be between 5 and 6 feet tall.

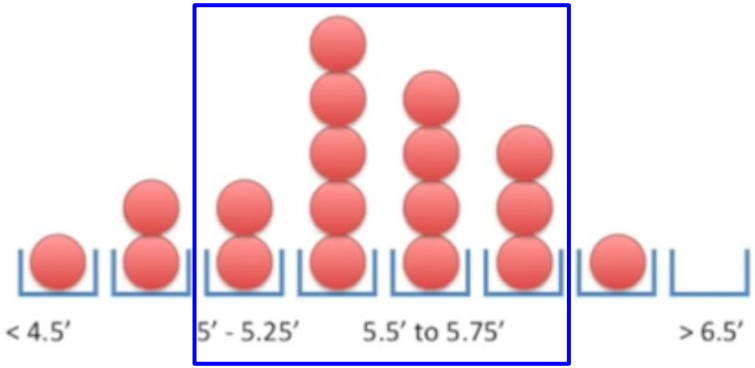

Shorter than 5'
5' to 5.5'
5.5' to 6'
Taller than 6'
What if we used smaller bin sizes for our measurements?


Now the bins are half as wide as before.
< 4.5'
5' - 5.25'
5.5' to 5.75'
> 6.5'

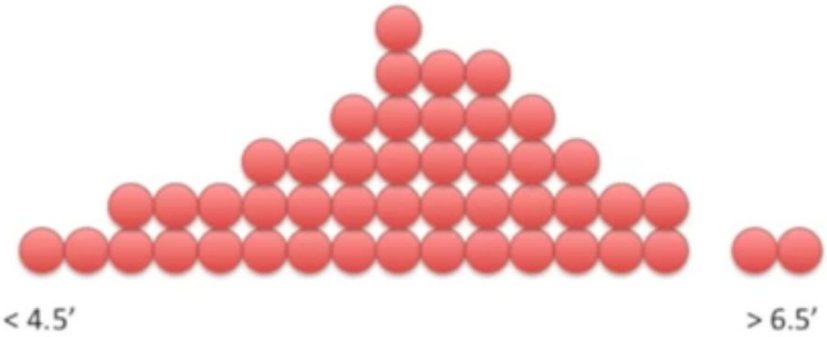



"Increase the resolution of the distribution by increase the bin size"


< 4.5’
5’ - 5.25’
5.5’ to 5.75’
> 6.5’

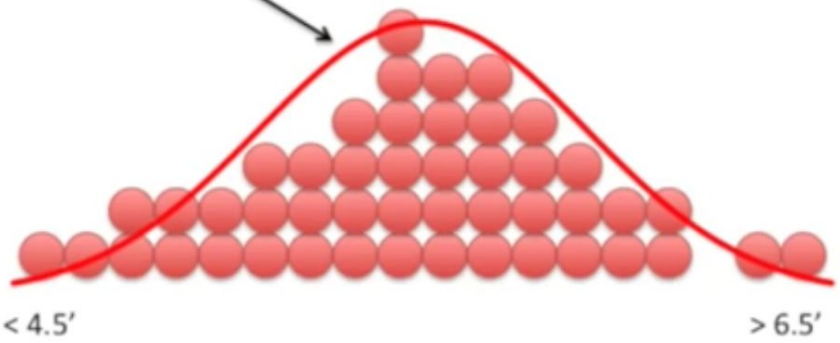


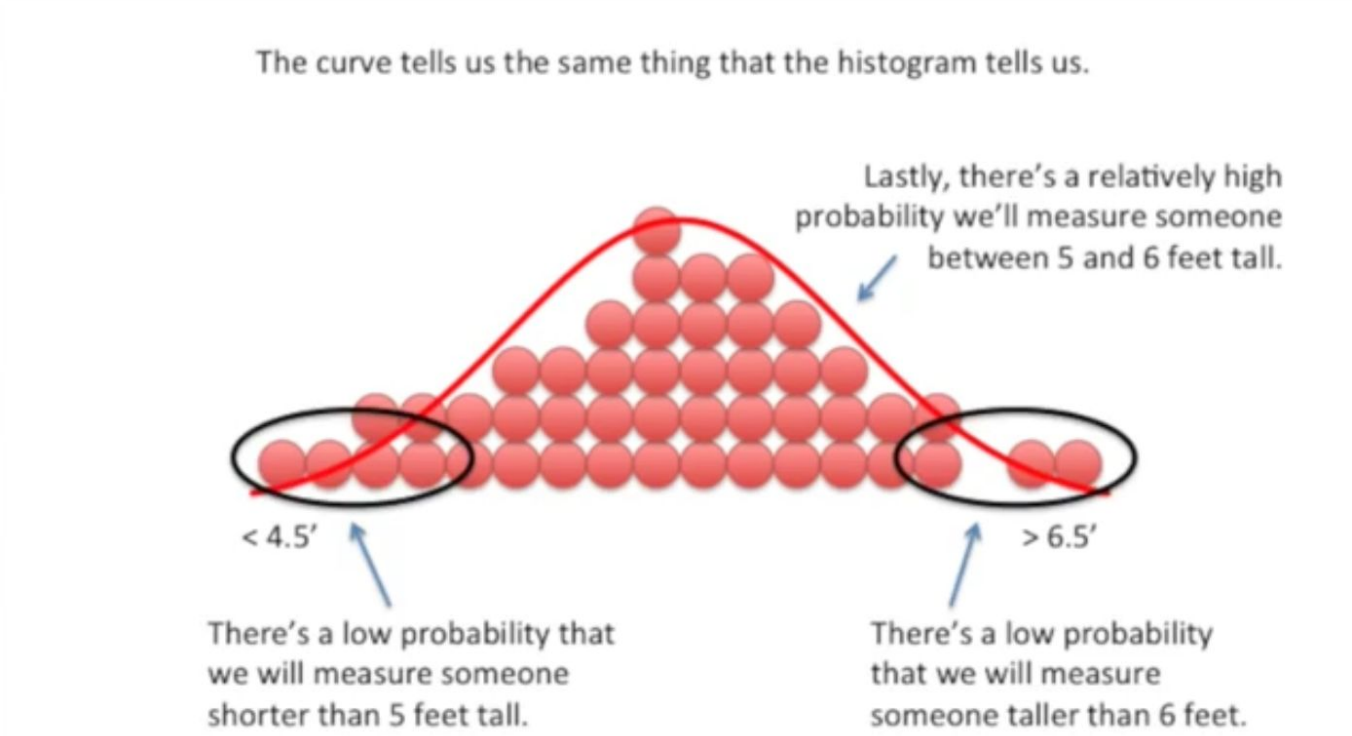

Again, most measurements are between 5 and 6 feet tall, but we can be more precise and say half of the people are between 5.25' and 5.75'




By measuring more people and using smaller bins, we get a more accurate and more precise estimate of how heights are distributed.

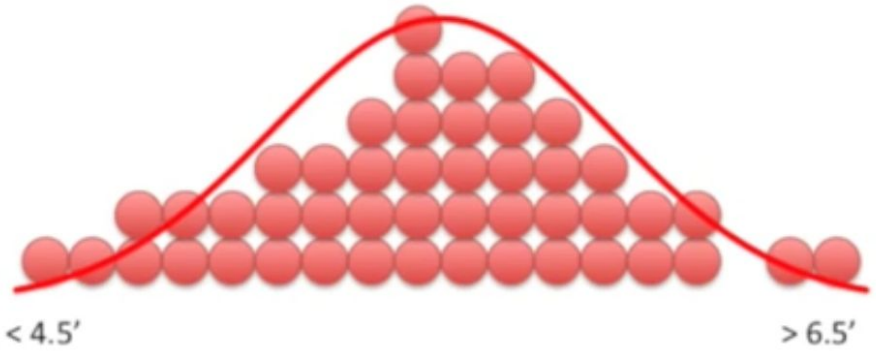
We can use a curve to approximate the histogram.

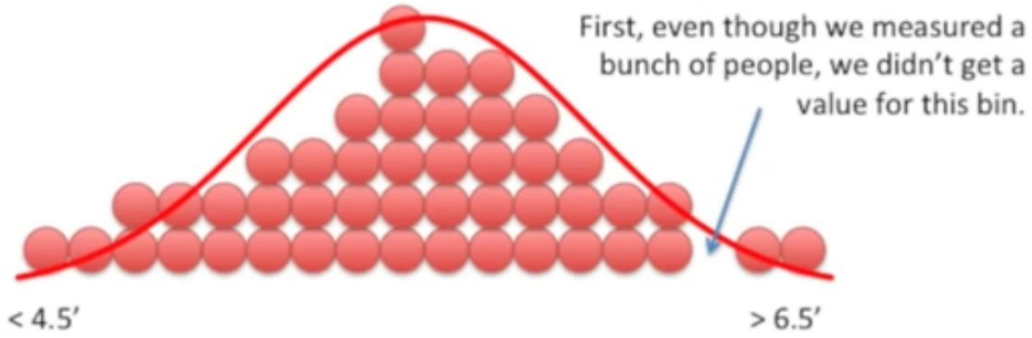


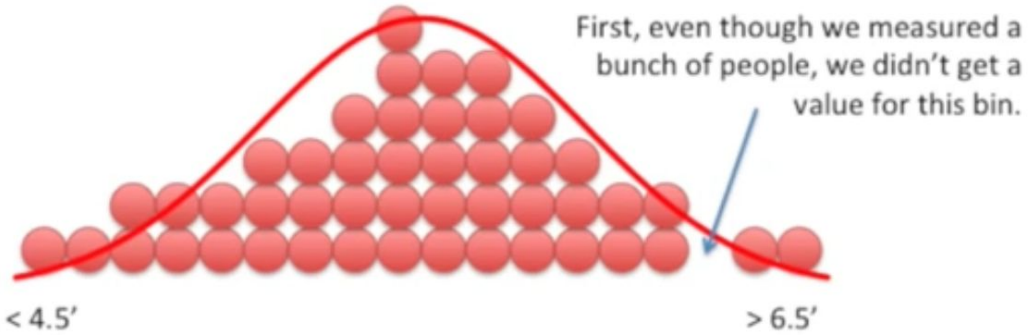



"A distribution is a function that shows the possible values for a variable and how often they occur"

However, the curve has a few advantages over the histogram.

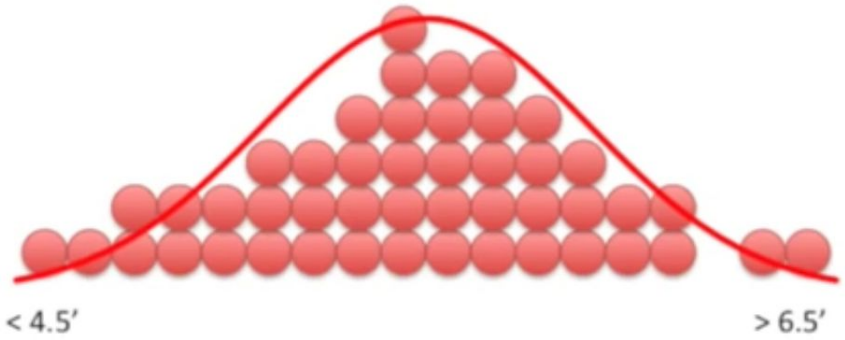



First, even though we measured a bunch of people, we didn't get a value for this bin.
< 4.5'
> 6.5'

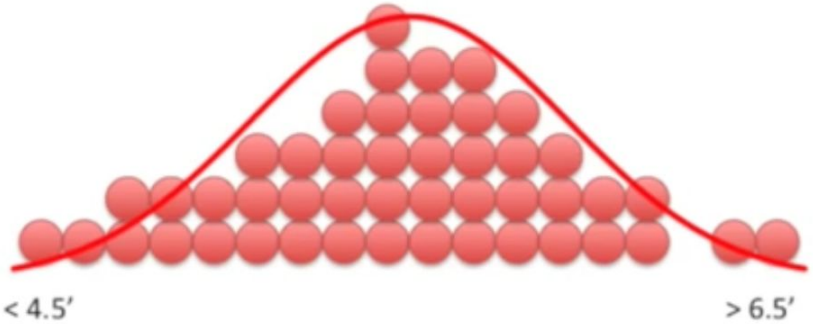



Since we can't calculate that probability with the histogram, does that mean that we will never get a measurement that fits into that bin? **No**

Another advantage is that the curve is not limited by the width of the bins.

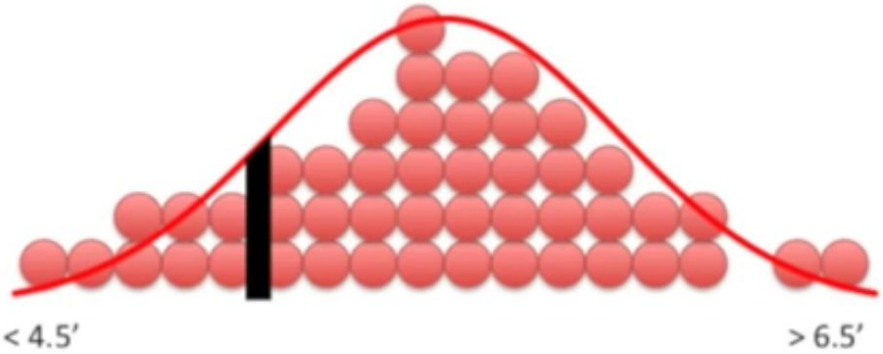


Another advantage is that the curve is not limited by the width of the bins.




If we wanted to know the probability of measuring someone between 5.021 and 5.317, we could use calculus (or a computer) to calculate this, without having to round to the nearest bin size.

Another advantage is that the curve is not limited by the width of the bins.




If we wanted to know the probability of measuring someone between 5.021 and 5.317, we could use calculus (or a computer) to calculate this, without having to round to the nearest bin size.

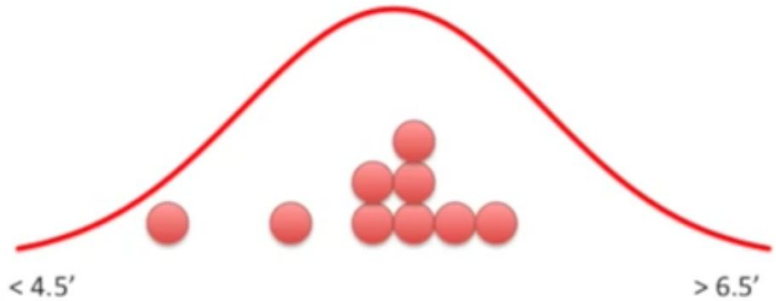
Lastly, if we don't have enough time or money to get a ton of measurements…

< 4.5'

> 6.5'

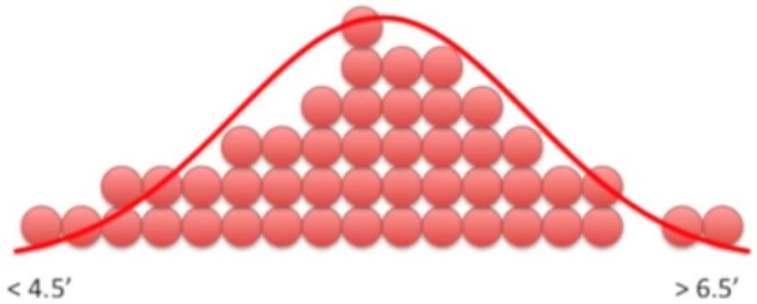
"Low/Insufficient sampling"

Lastly, if we don't have enough time or money to get a ton of measurements… the approximate curve (based on the mean and standard deviation of the data we were able to collect), is usually good enough.




Thus, using the curve can save us a lot of time and money.

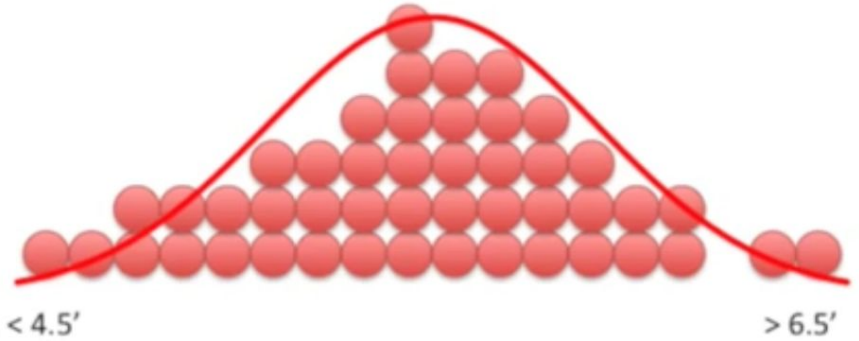
Both the histogram and the curve are "**distributions**"




"A distribution is a function that shows the possible values for a variable and how often they occur"

Both the histogram and the curve are "**distributions**"

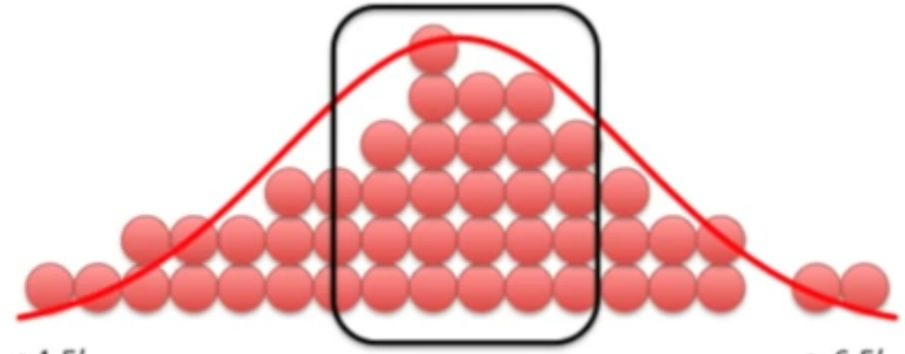
**They show us how the probabilities of measurements are distributed.**




"A distribution is a function that shows the possible values for a variable and how often they occur"

Both the histogram and the curve are "**distributions**"

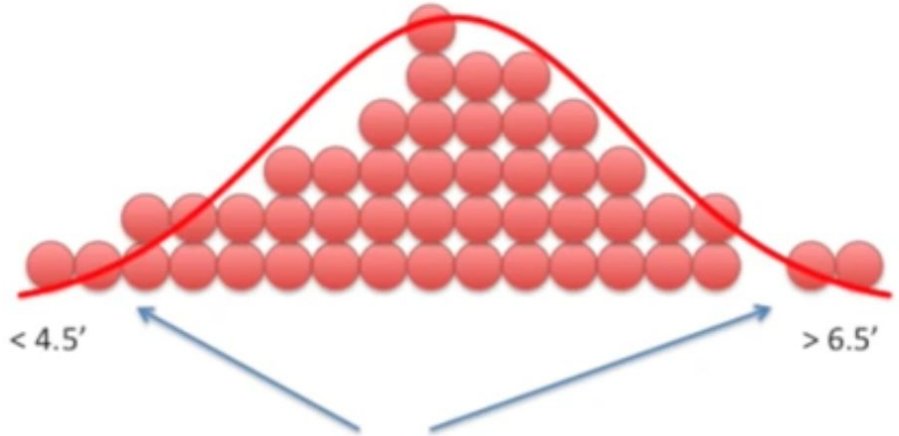
**They show us how the probabilities of measurements are distributed.**




The tallest part of the histogram, or curve, shows the region where measurements are most likely.

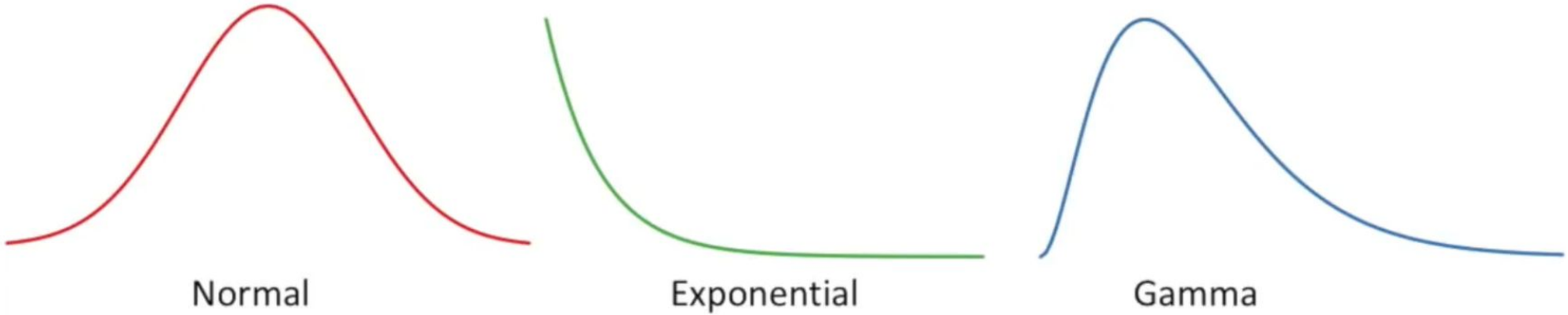
Both the histogram and the curve are "**distributions**"

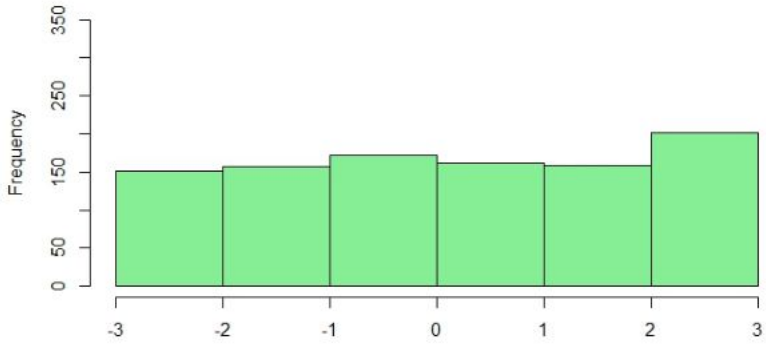
**They show us how the probabilities of measurements are distributed.**




The low parts of the histogram, or curve, show where measurements are less likely.

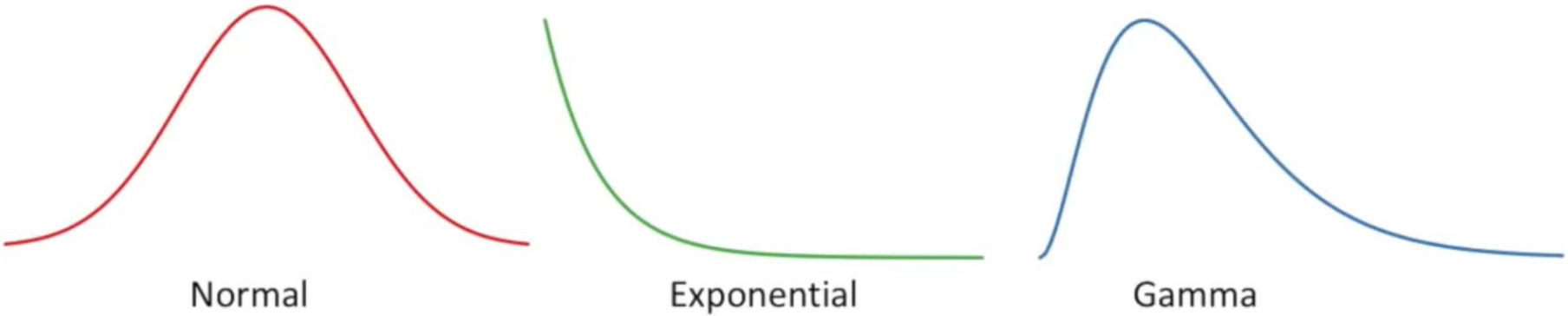
We've been talking about how height measurements are distributed, but there are all kinds of distributions with all kinds of interesting shapes.

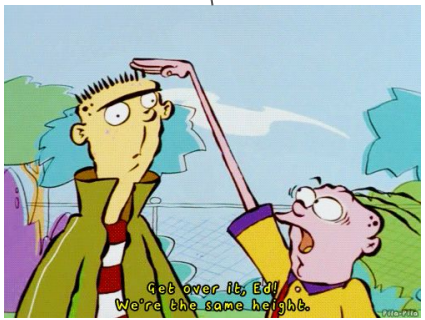



"What kind of distribution is the above?"

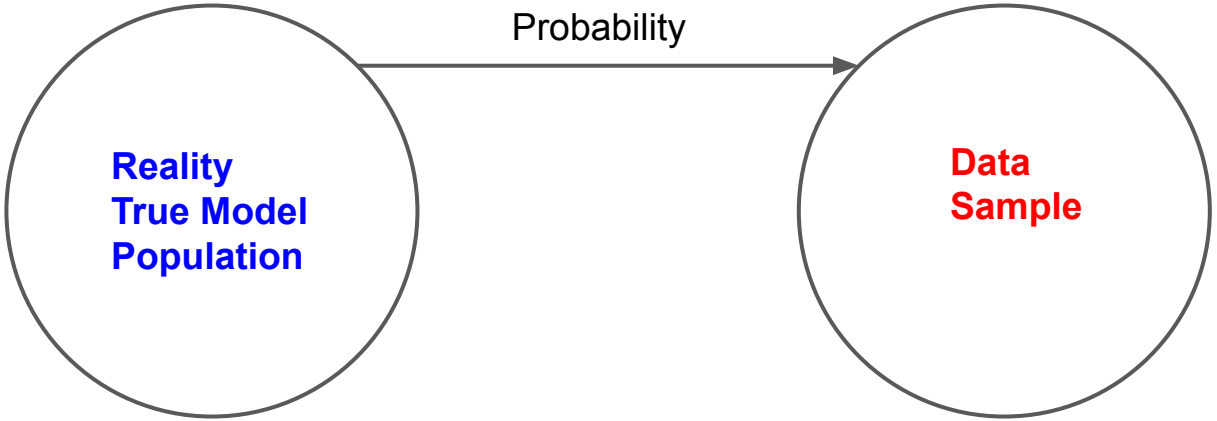
# Real-life applications of distribution



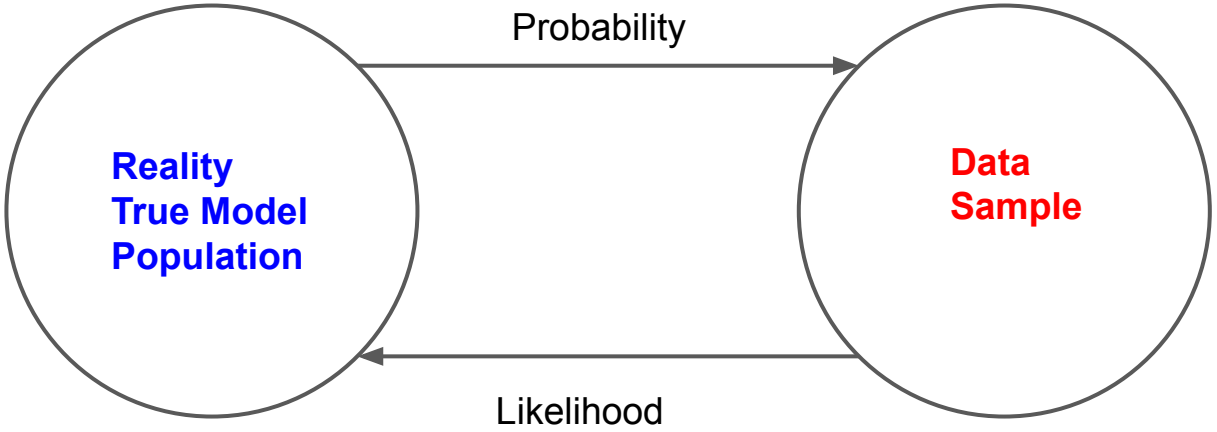

# Break and QA

# What is Likelihood?




- Probability: What is the chance of observing data or sample given a specific model or population?
- If the height is normal(mua, sigma) what is the chance of observing x?

# What is Likelihood?




- Likelihood: given observed data, what is the chance that a given reality or model  is true?
- If you observe x, what is the best normal distribution (mua, sigma)?

# Maximum Likelihood Estimation

A procedure to:

1. Determine best model parameters(reality) that fit given data
2. Compare multiple models to determine the best fit to data

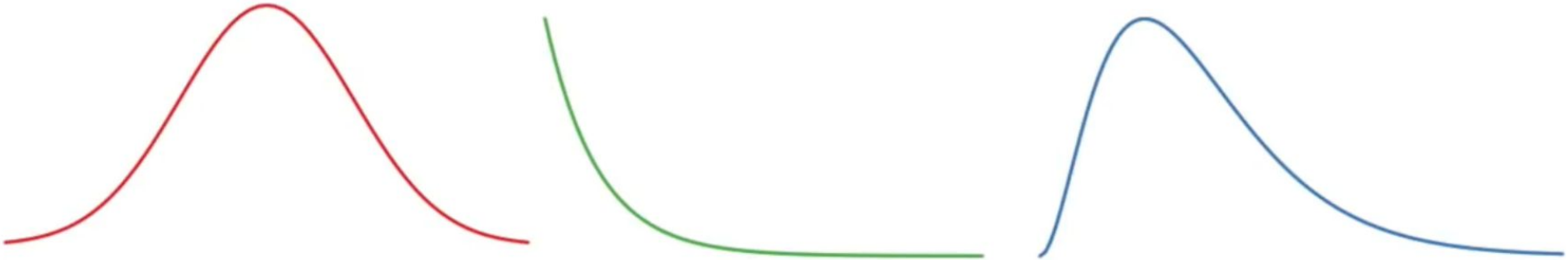
What it does:

1. Maximizes log-likelihood function to estimate parameters

Let’s say we weighed a bunch of mice…

The goal of maximum likelihood is to find the optimal way to fit a distribution to the data.

There are lots of different types of distributions
for different types of data…



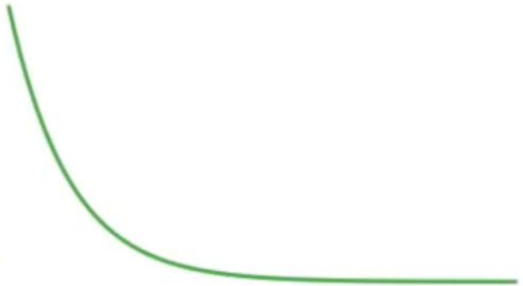
Normal

Exponential

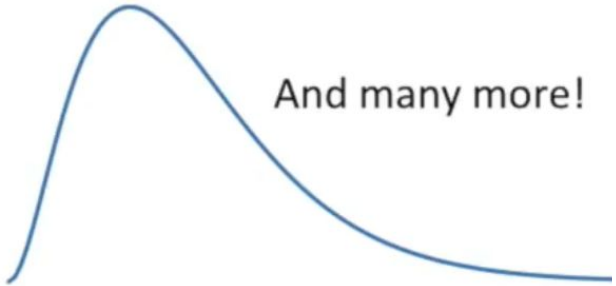
Gamma

There are lots of different types of distributions
for different types of data...

Normal



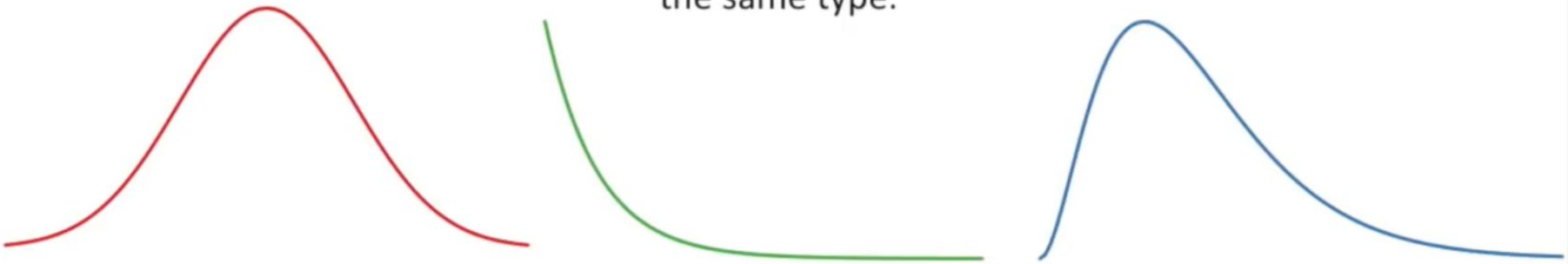
Exponential



Gamma

And many more!

The reason you want to fit a distribution to your data is it can be easier to work with and it is also more general - it applies to every experiment of the same type.



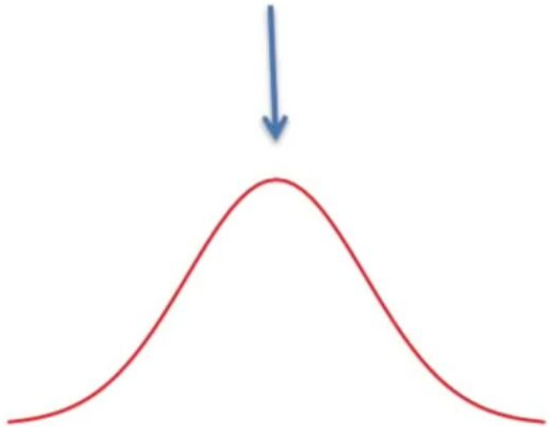
Normal

Exponential

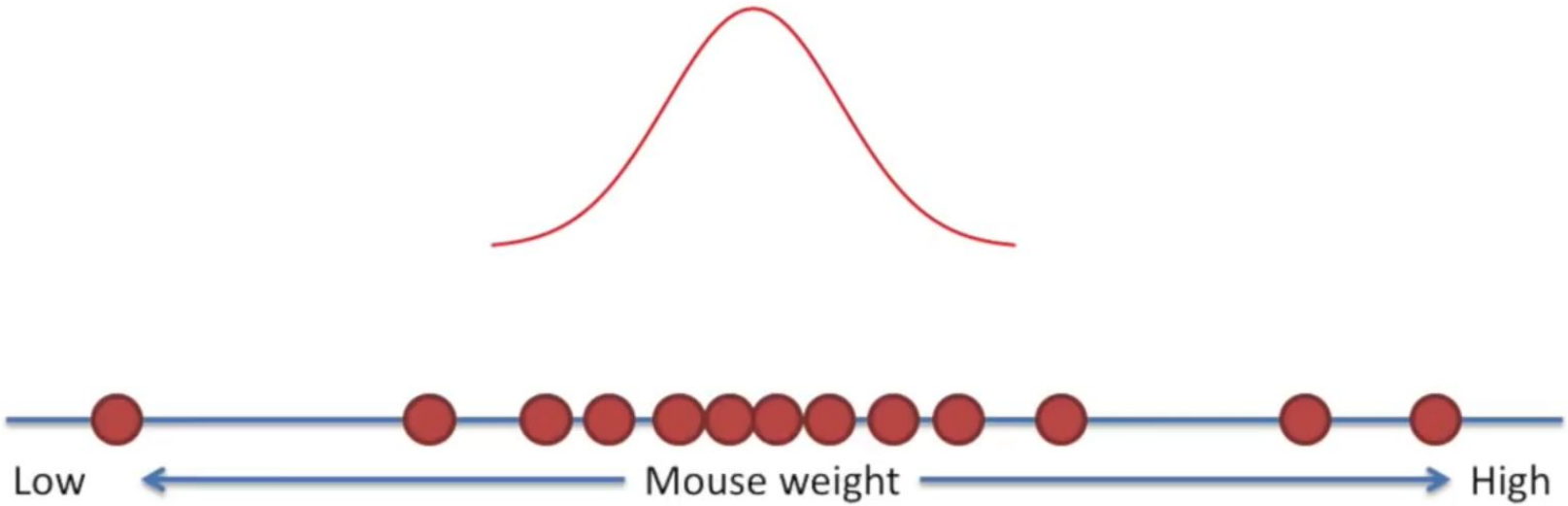
Gamma

In this case, we think that the weights might be normally distributed…

That means we think it came from this type of distribution:

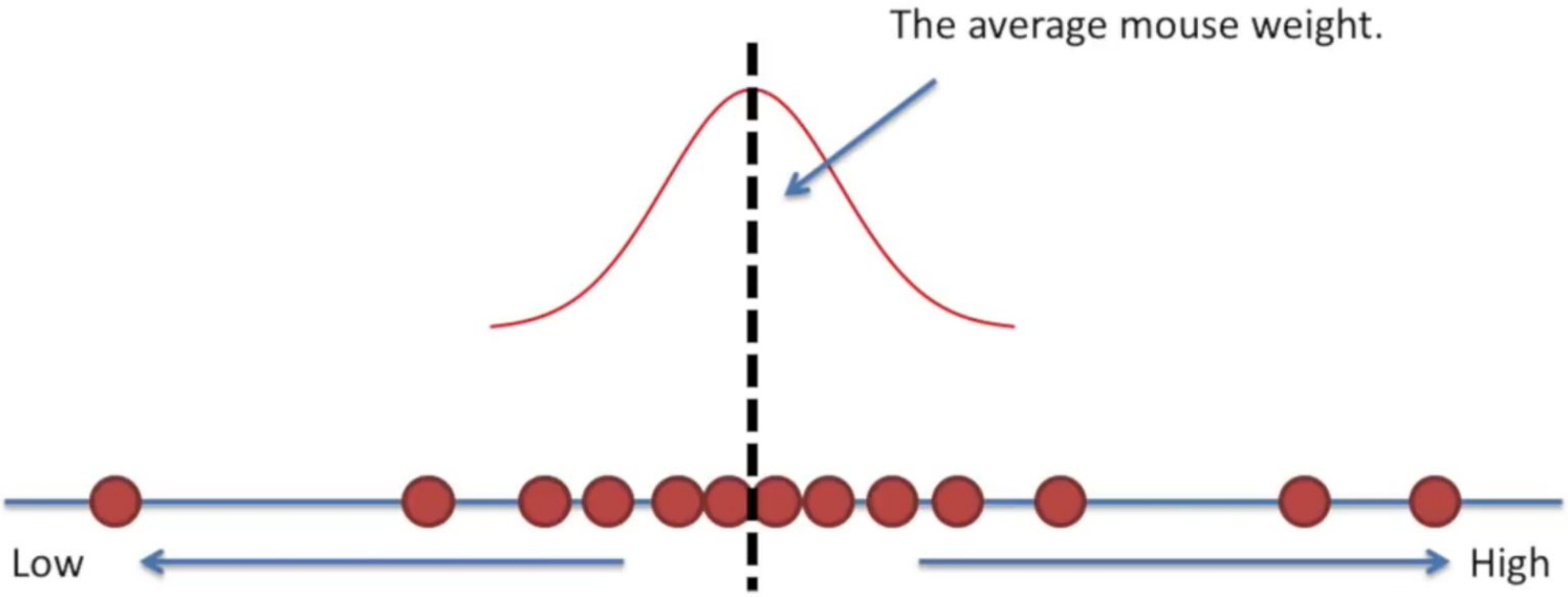


“Normally distributed” means a number of things:



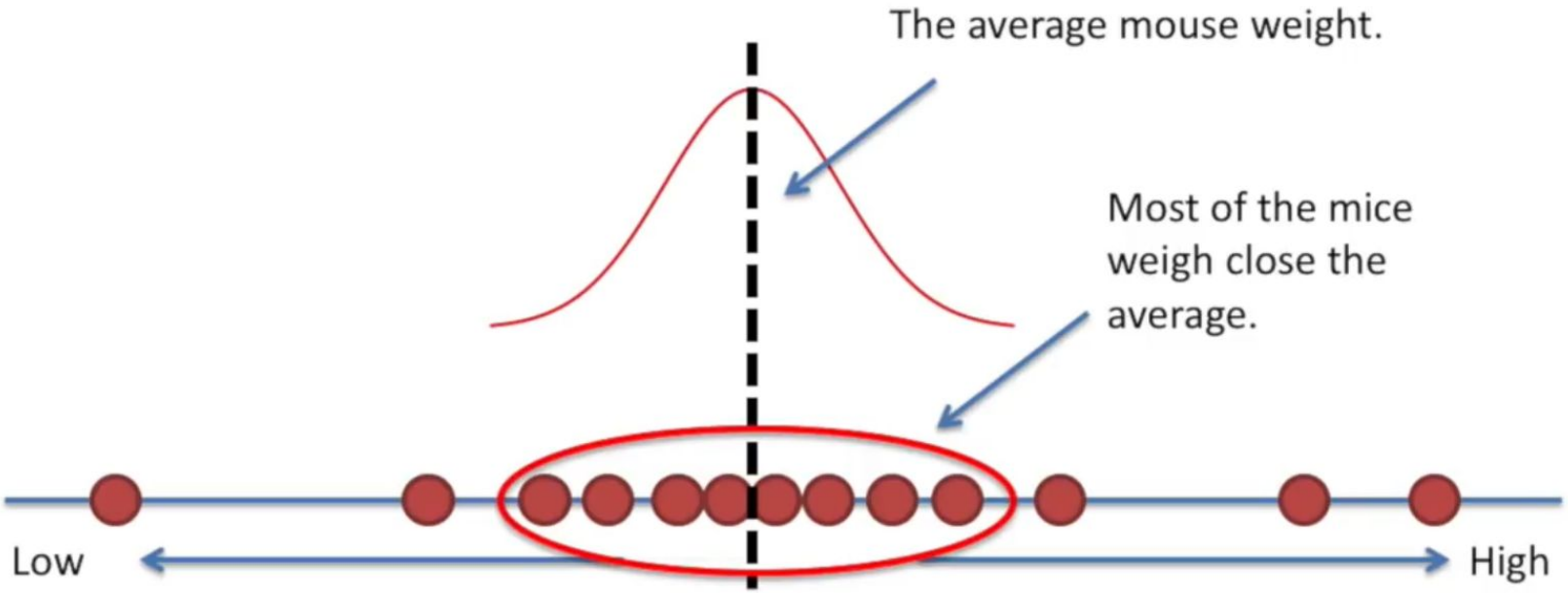
"Normally distributed" means a number of things:

1) We expect most of the measurements (mouse weights) to be close to the mean (average).



"Normally distributed" means a number of things:

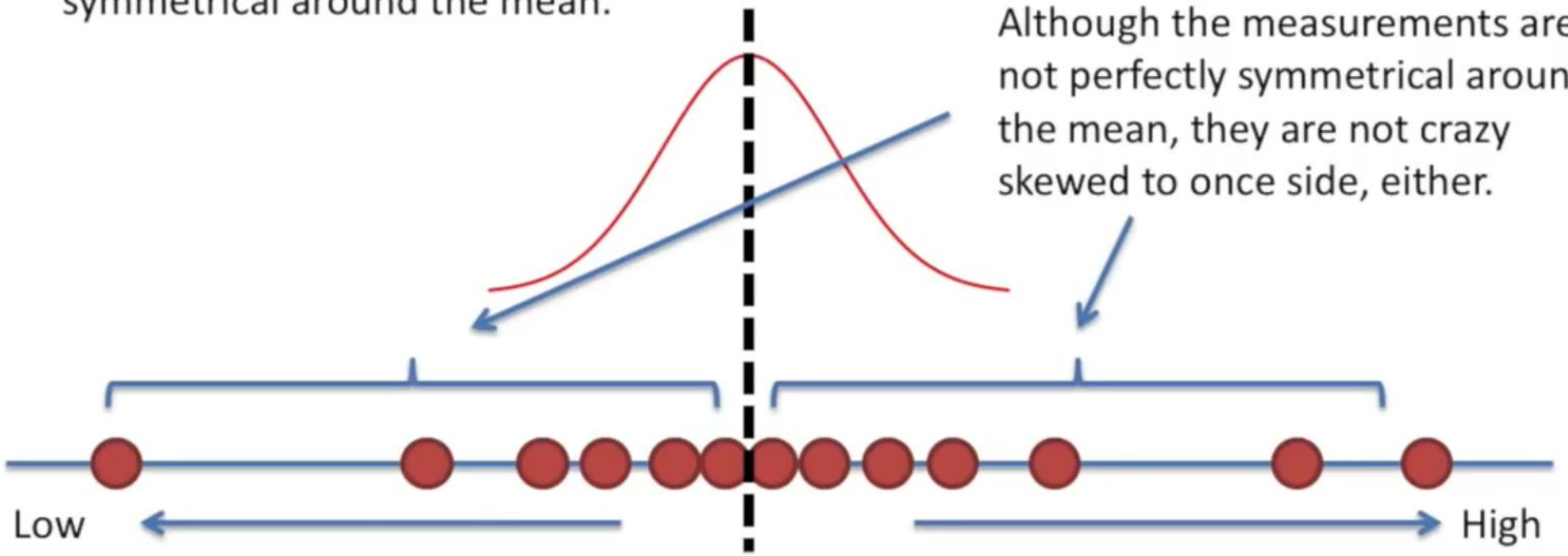
1) We expect most of the measurements (mouse weights) to be close to the mean (average).



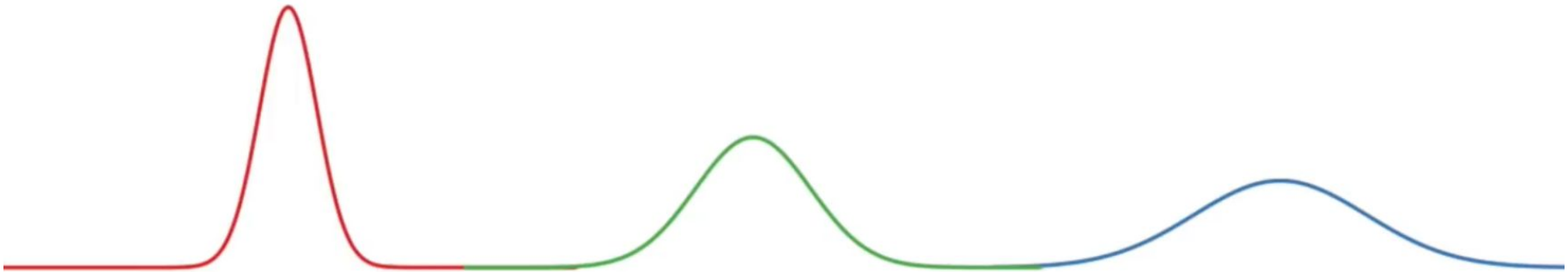
"Normally distributed" means a number of things:

1) We expect most of the measurements (mouse weights) to be close to the mean (average).

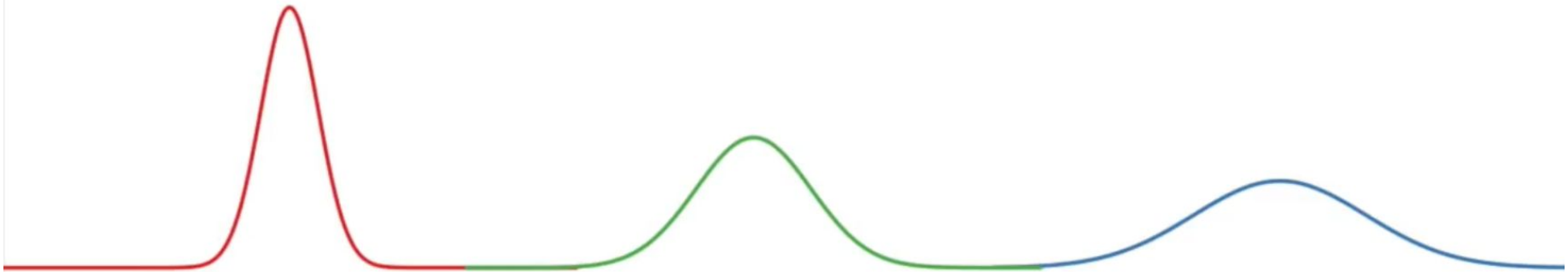
2) We expect the measurements to be relatively symmetrical around the mean.



Normal distributions come in all kinds of shapes and sizes…



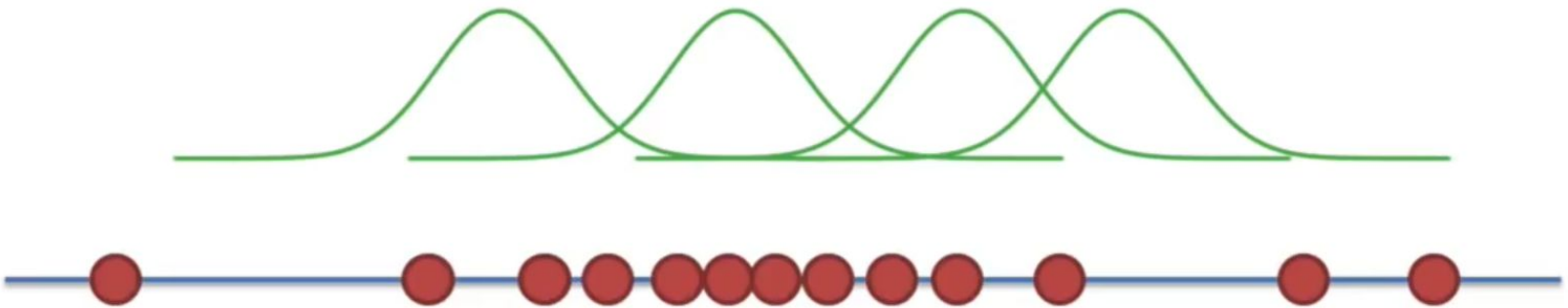
Normal distributions come in all kinds of shapes and sizes…



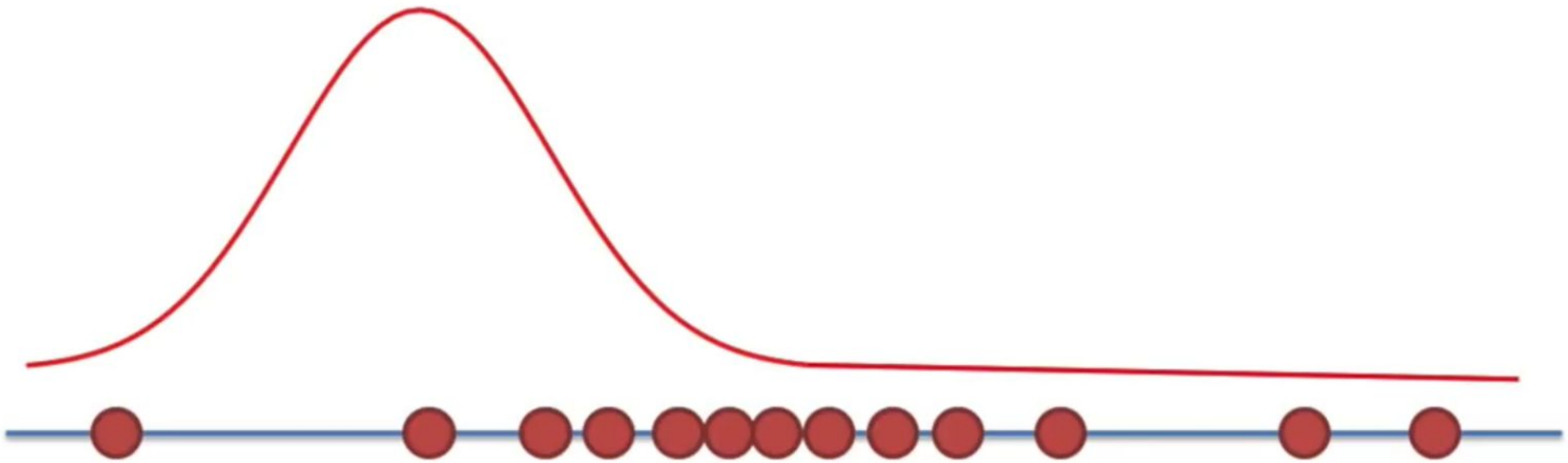
“Skinny” “Medium” “Large Boned”

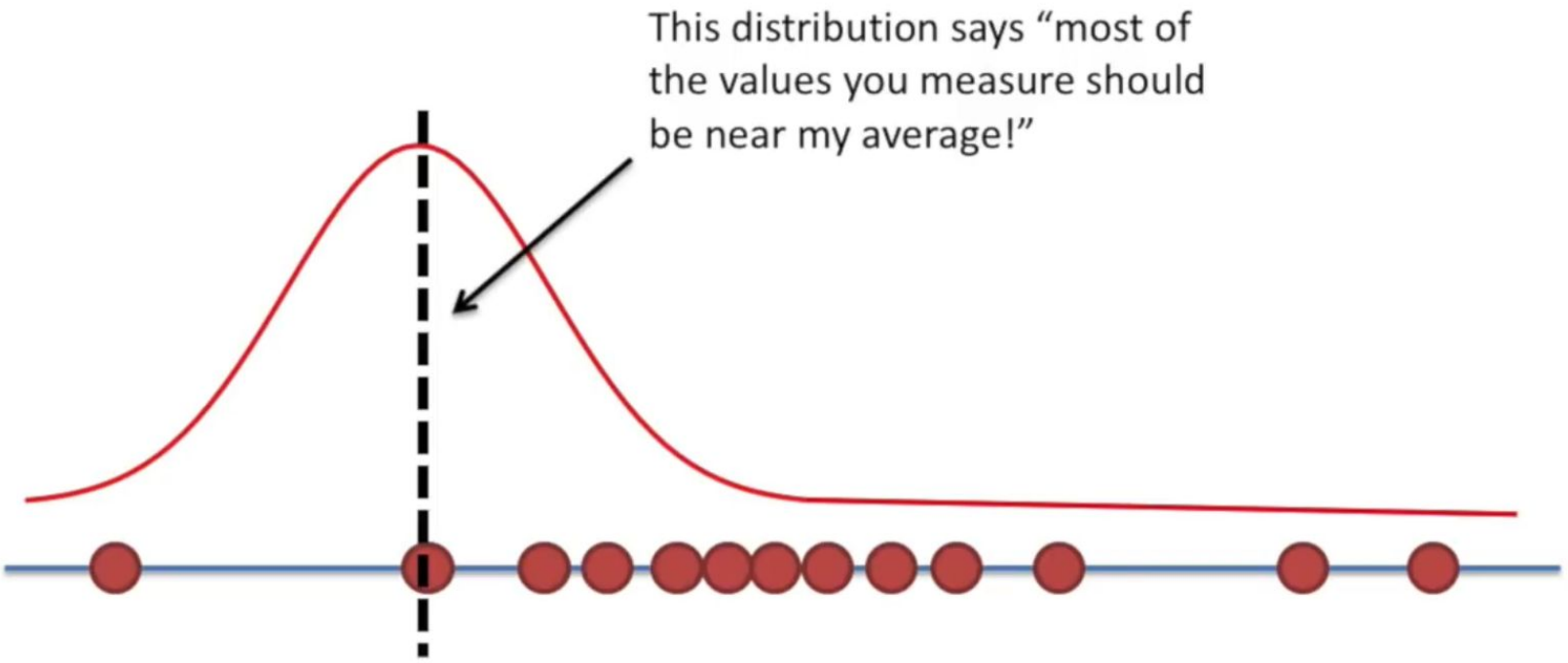
Once we settle on the shape, we have to figure out where to center the thing…

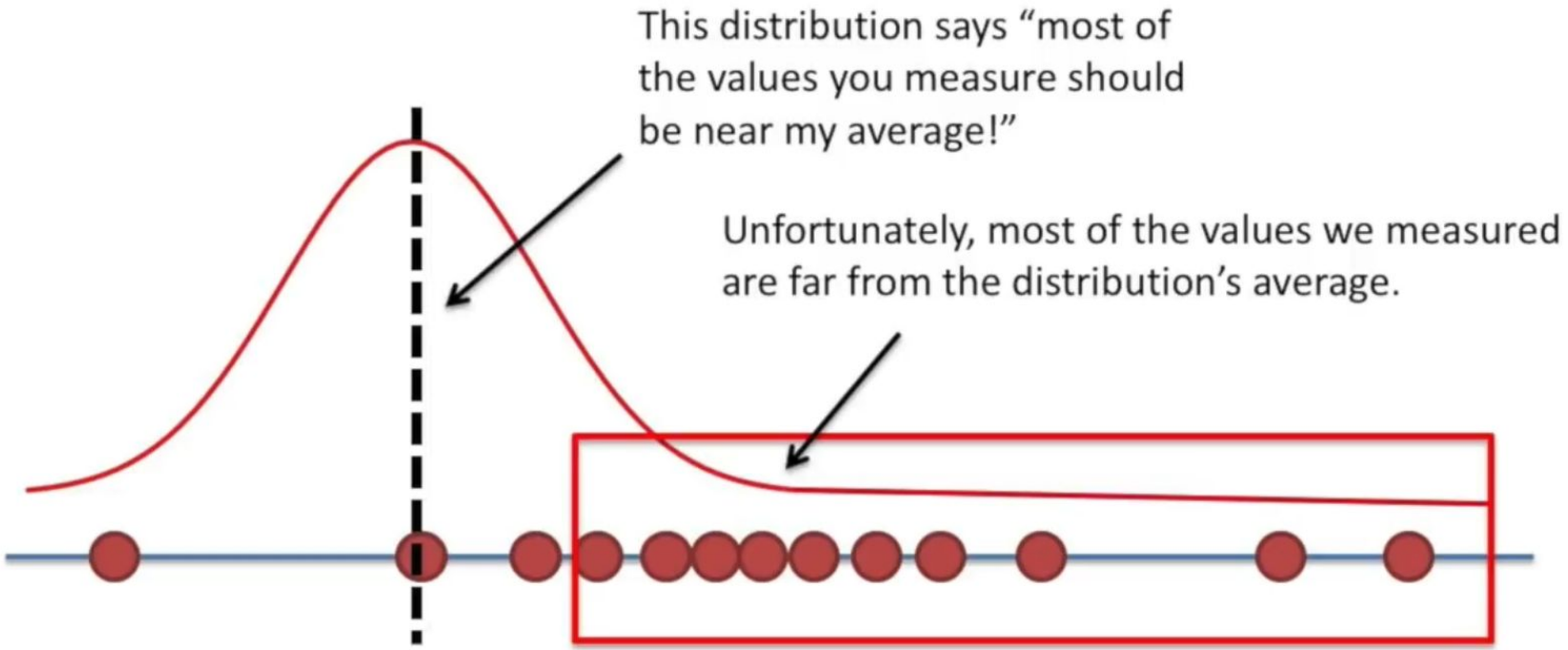
Is one location “better” than another?

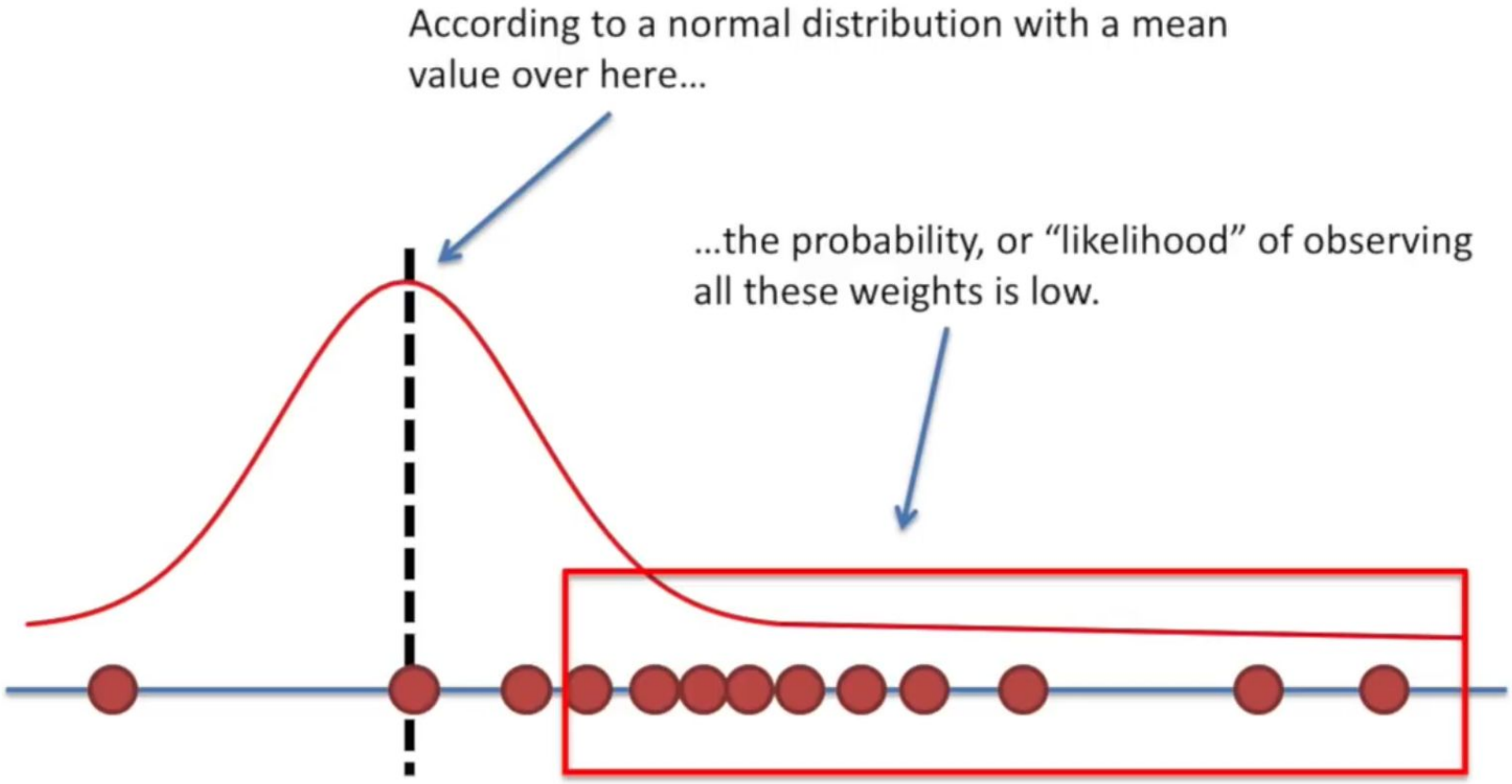


Before we get too technical, lets just pick any old normal distribution and see how well it fits the data.

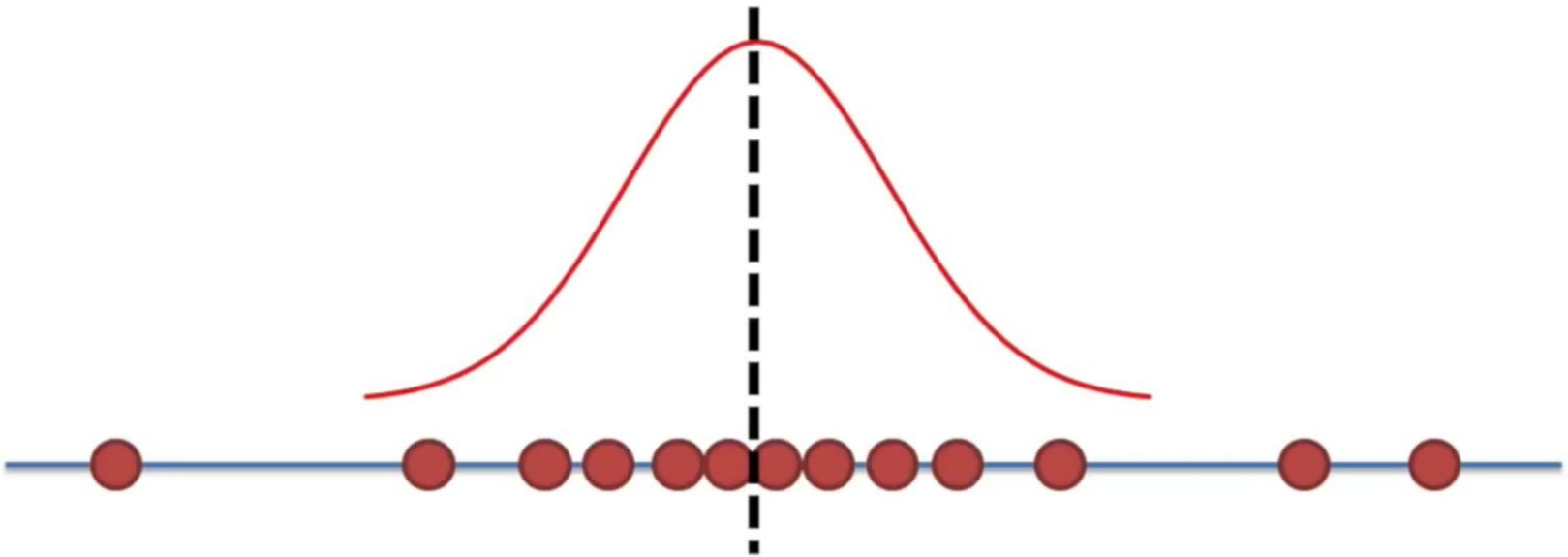


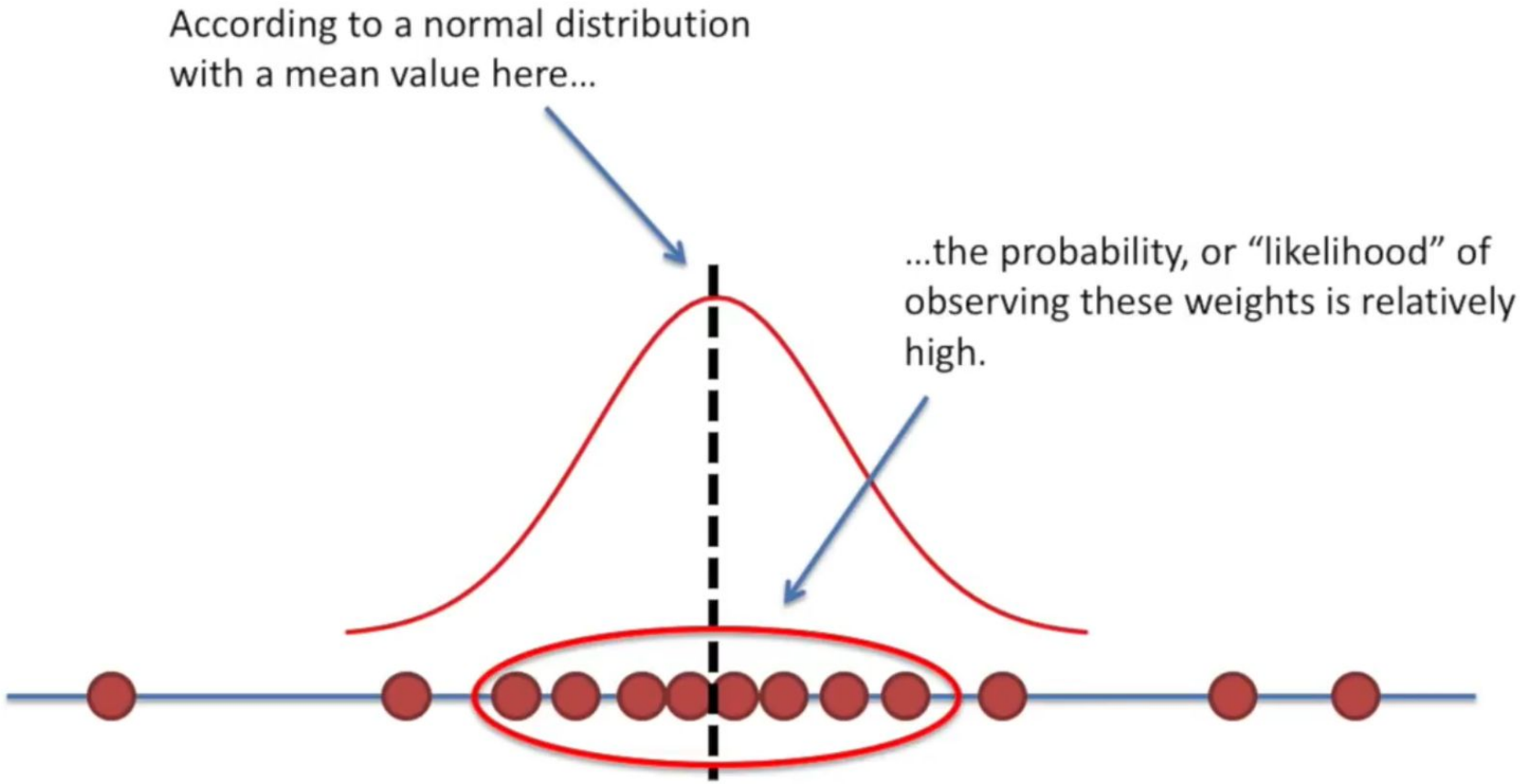

This distribution says “most of the values you measure should be near my average!”


This distribution says “most of the values you measure should be near my average!”
Unfortunately, most of the values we measured are far from the distribution’s average.


According to a normal distribution with a mean value over here…
…the probability, or “likelihood” of observing all these weights is low.

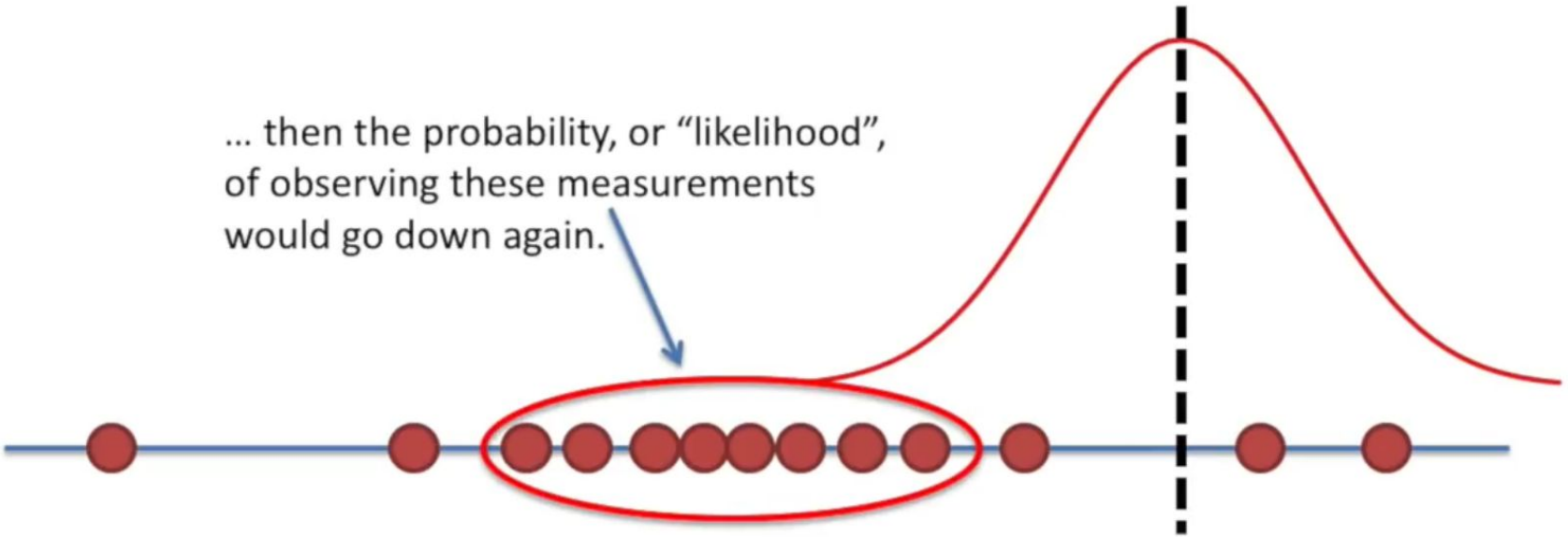
What if we shifted the normal distribution over, so that its mean was the same as the average weight?



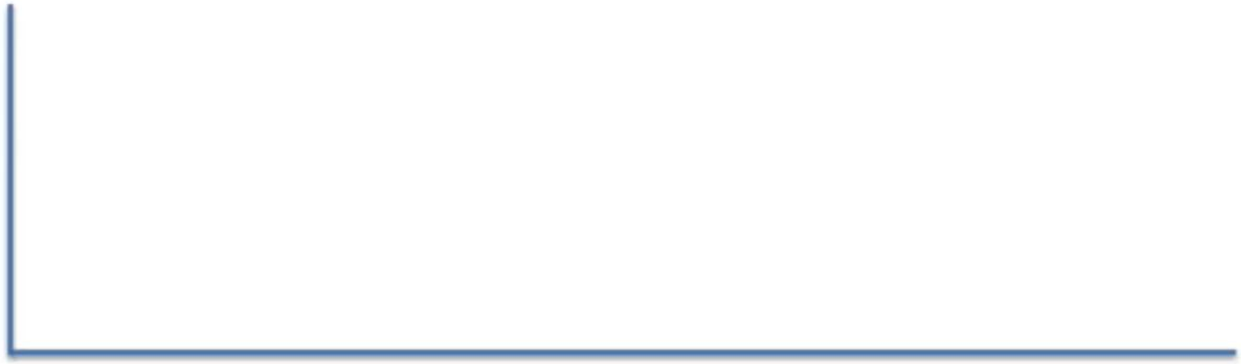

According to a normal distribution with a mean value here…
…the probability, or “likelihood” of observing these weights is relatively high.

If we kept shifting the normal distribution over…

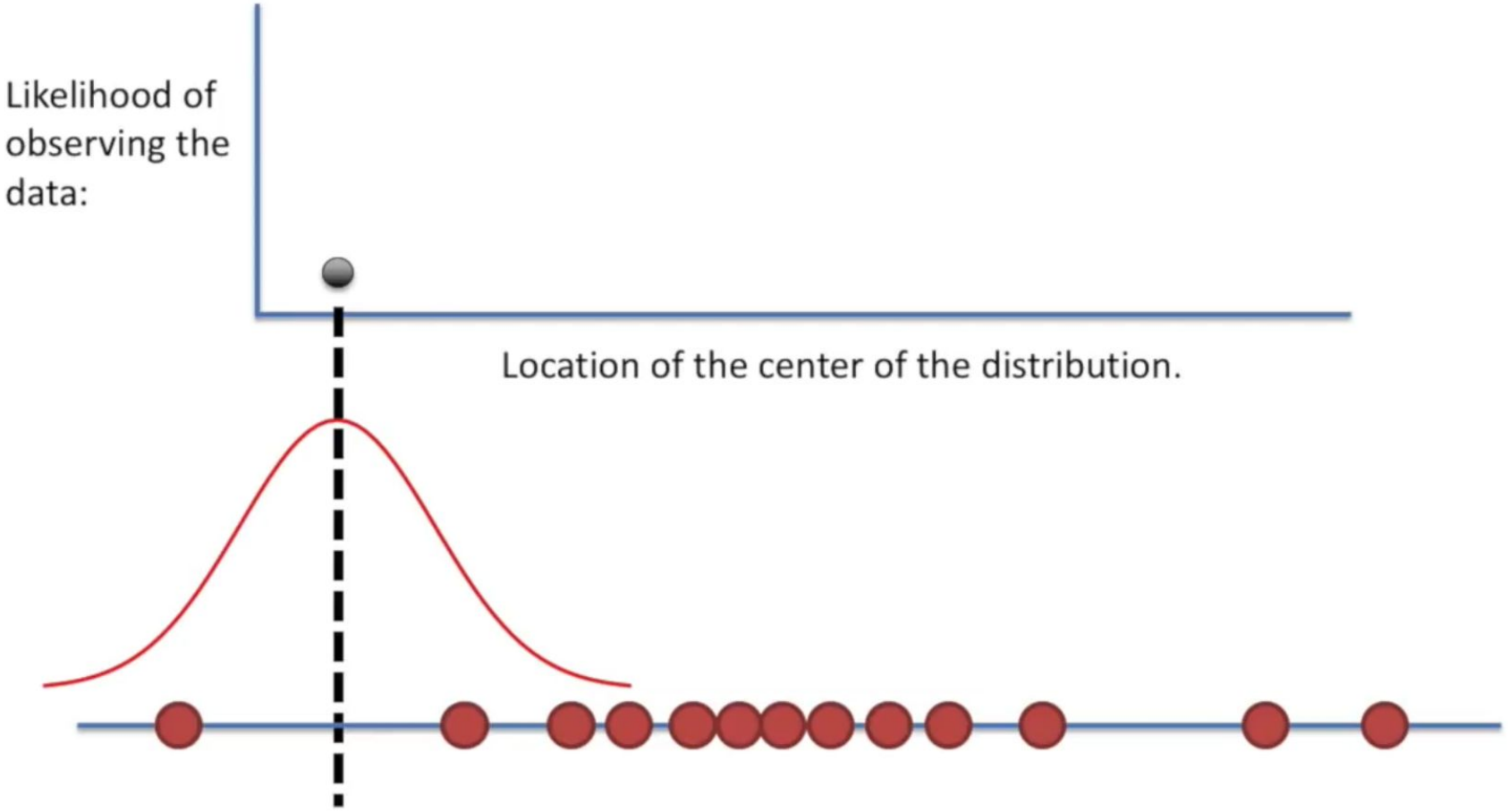
… then the probability, or “likelihood”, of observing these measurements would go down again.

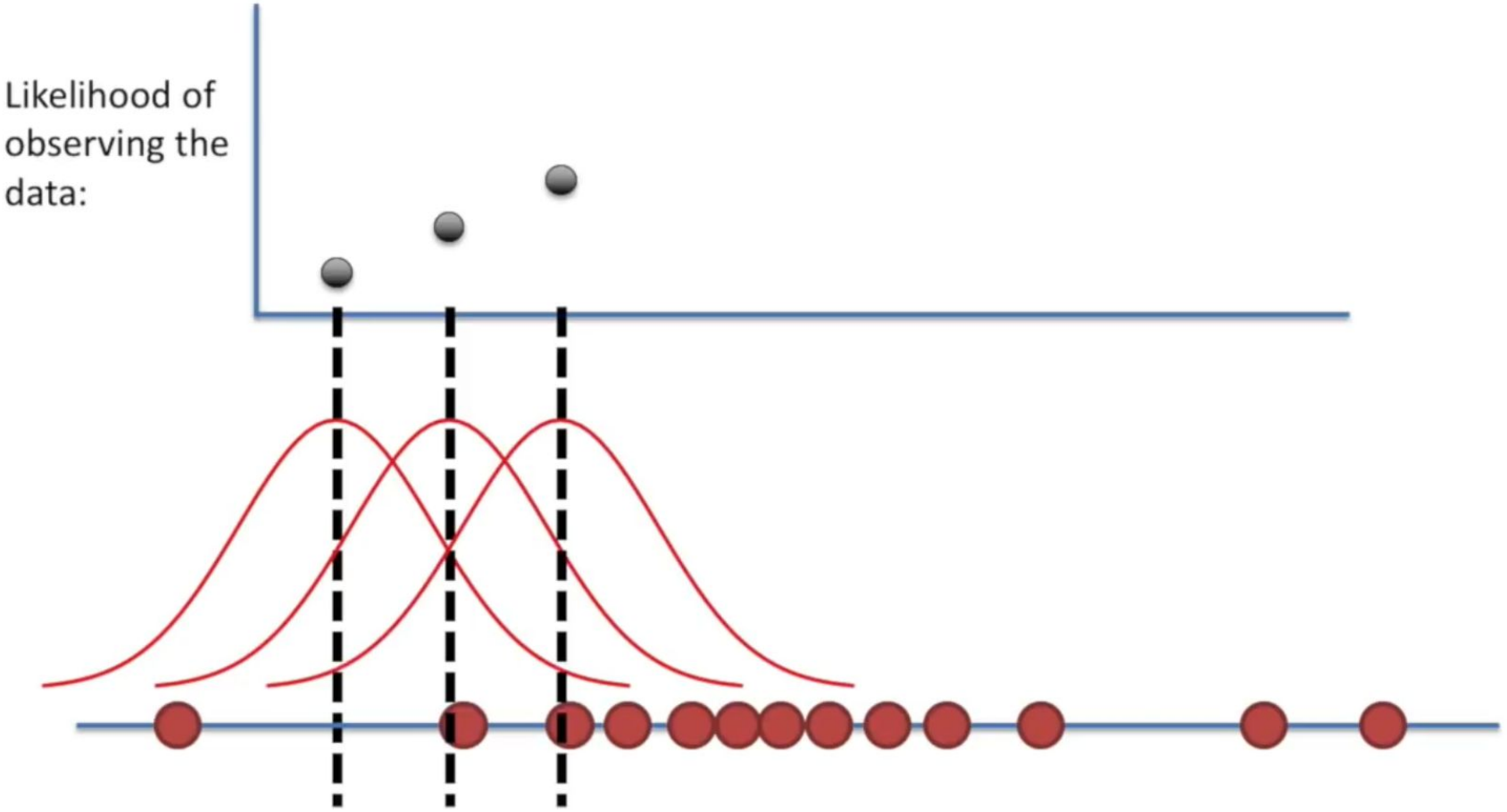


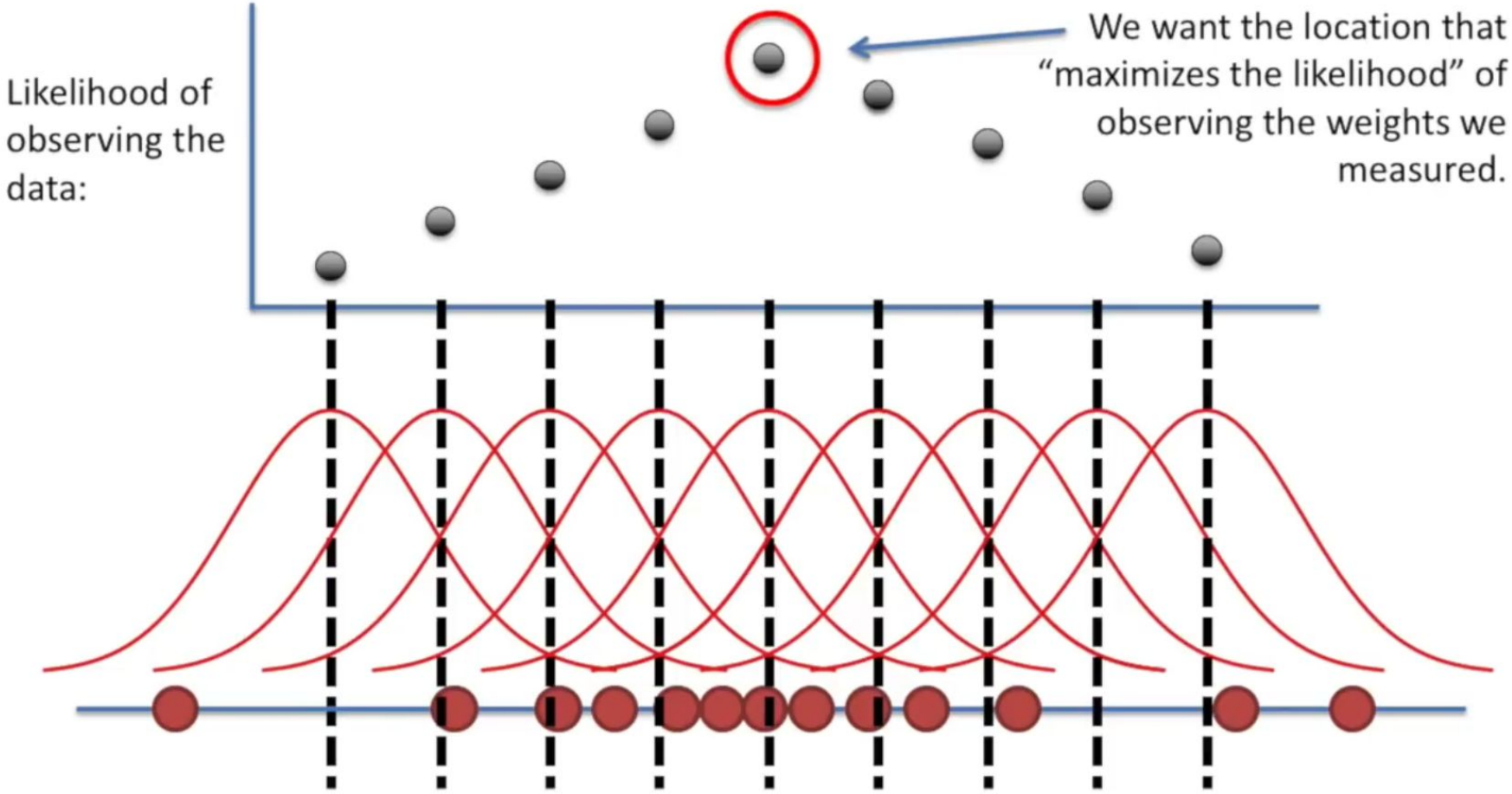
Likelihood of observing the data:

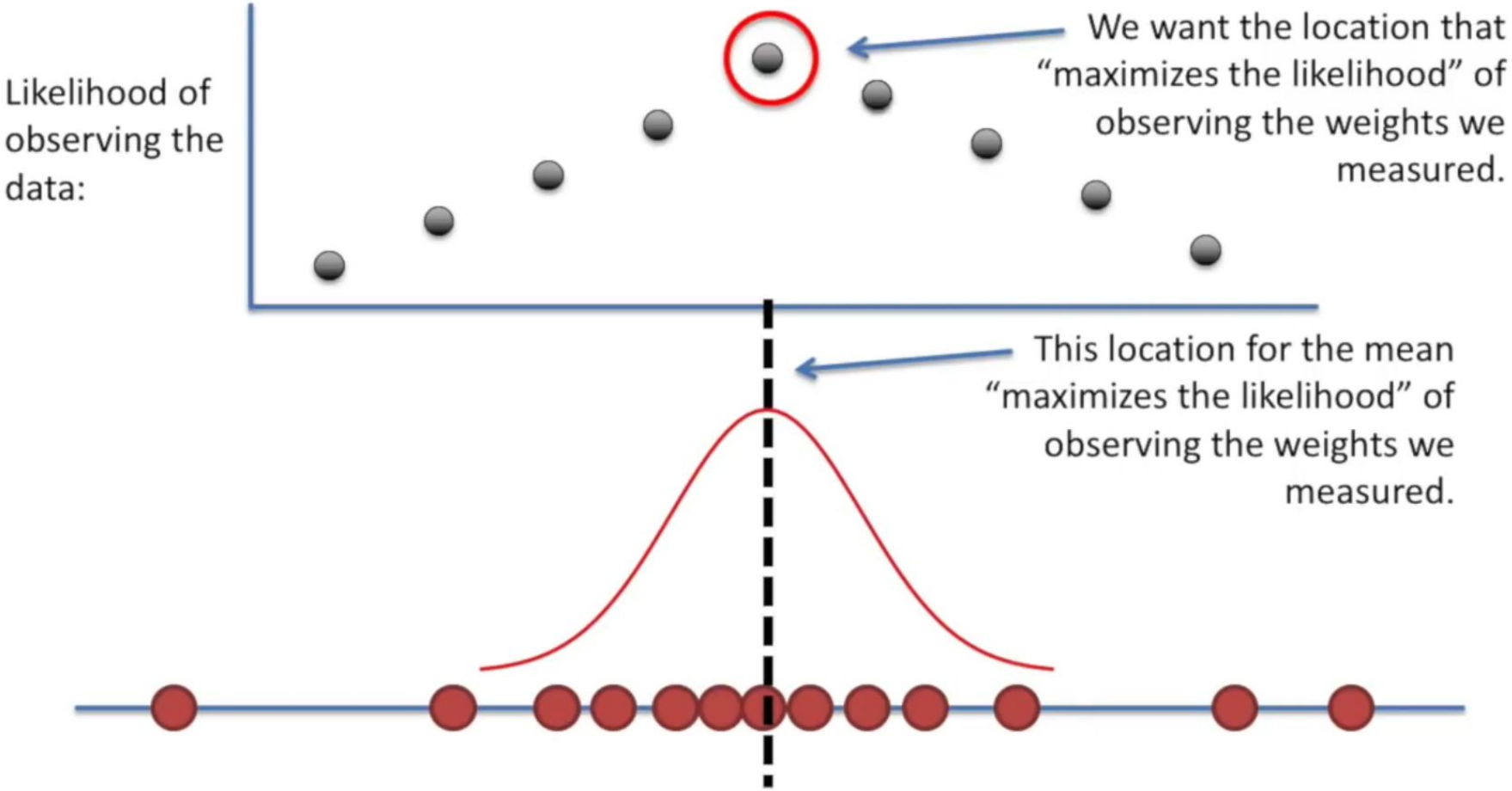


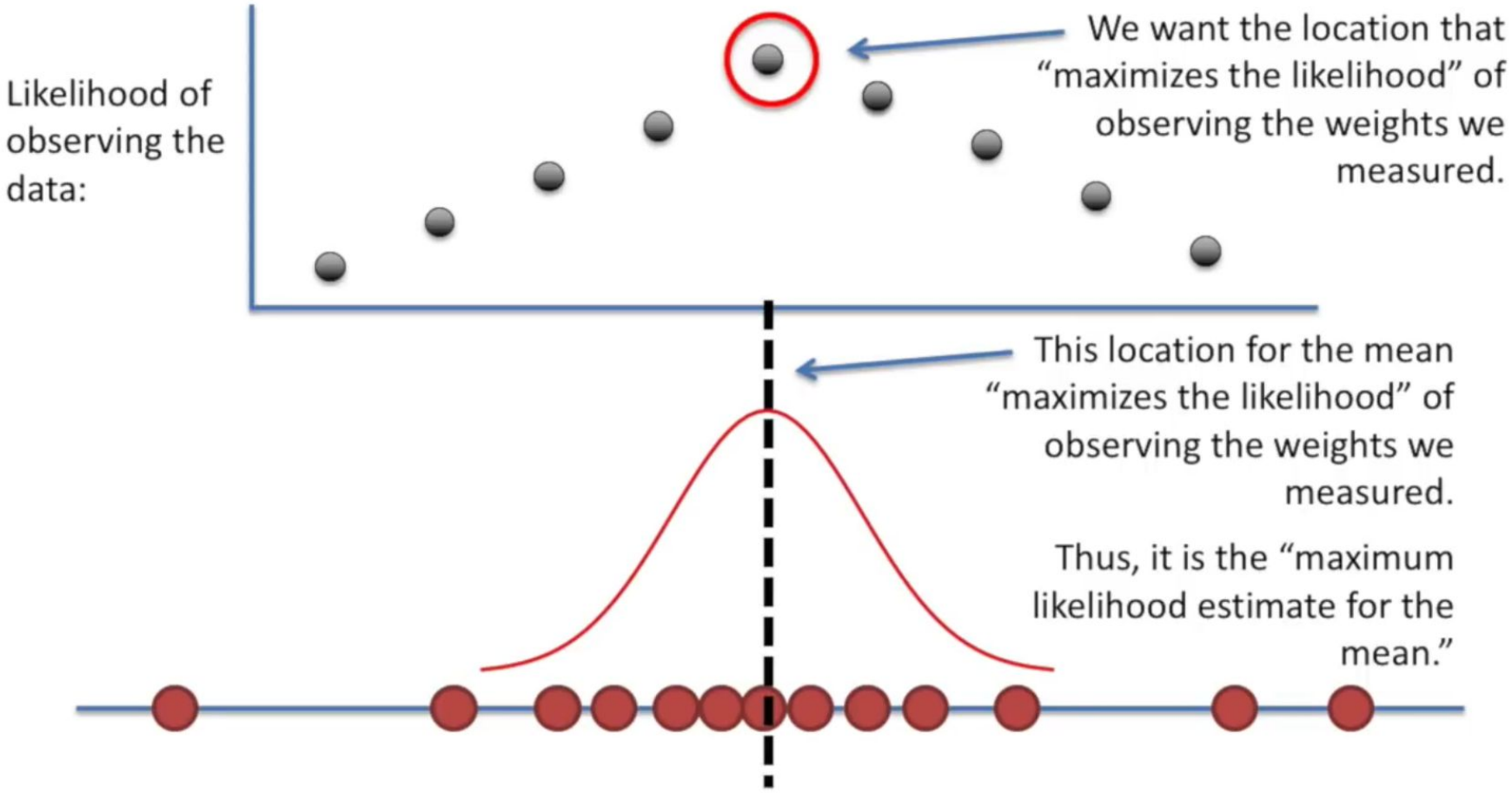
Location of the center of the distribution.


Likelihood of observing the data:
Location of the center of the distribution.

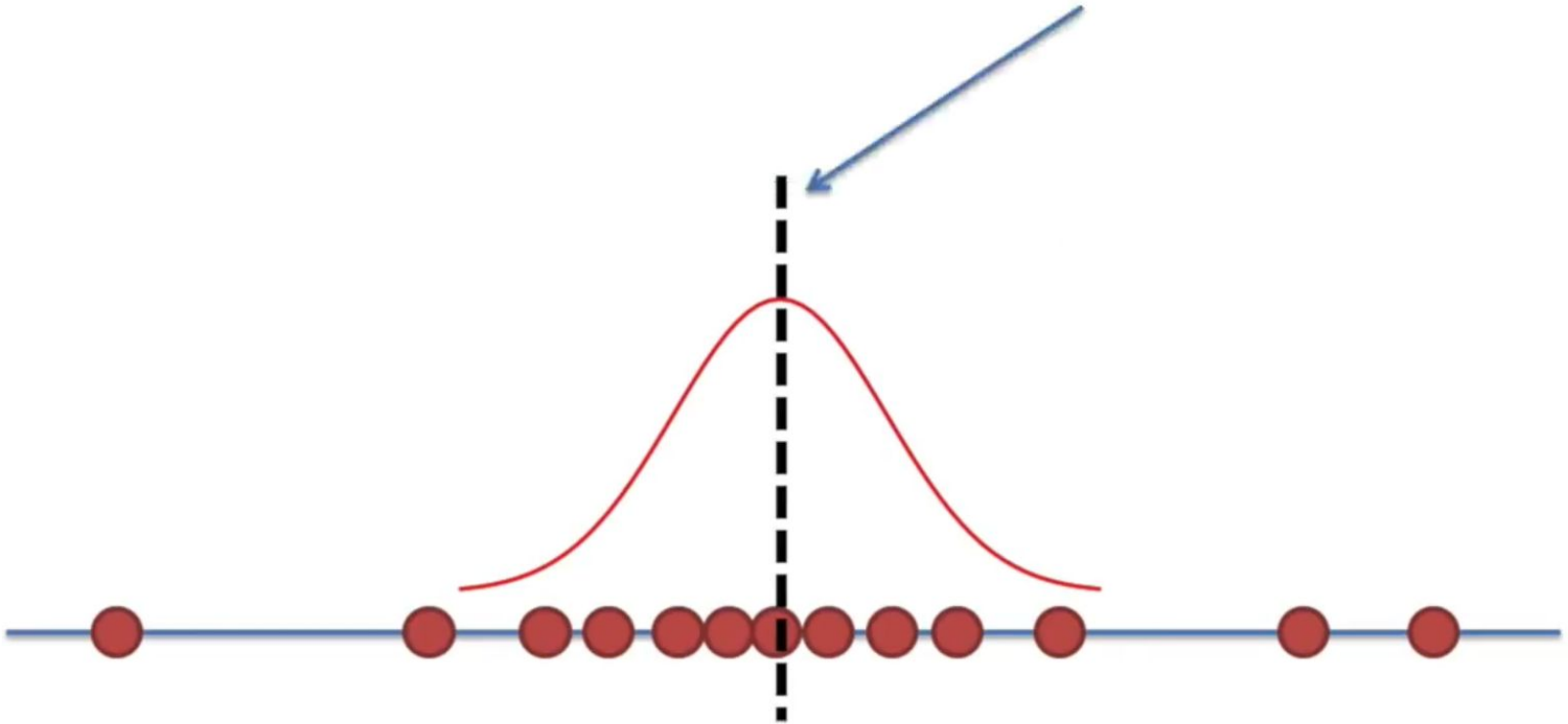

Likelihood of observing the data:


Likelihood of observing the data:
We want the location that "maximizes the likelihood" of observing the weights we measured.

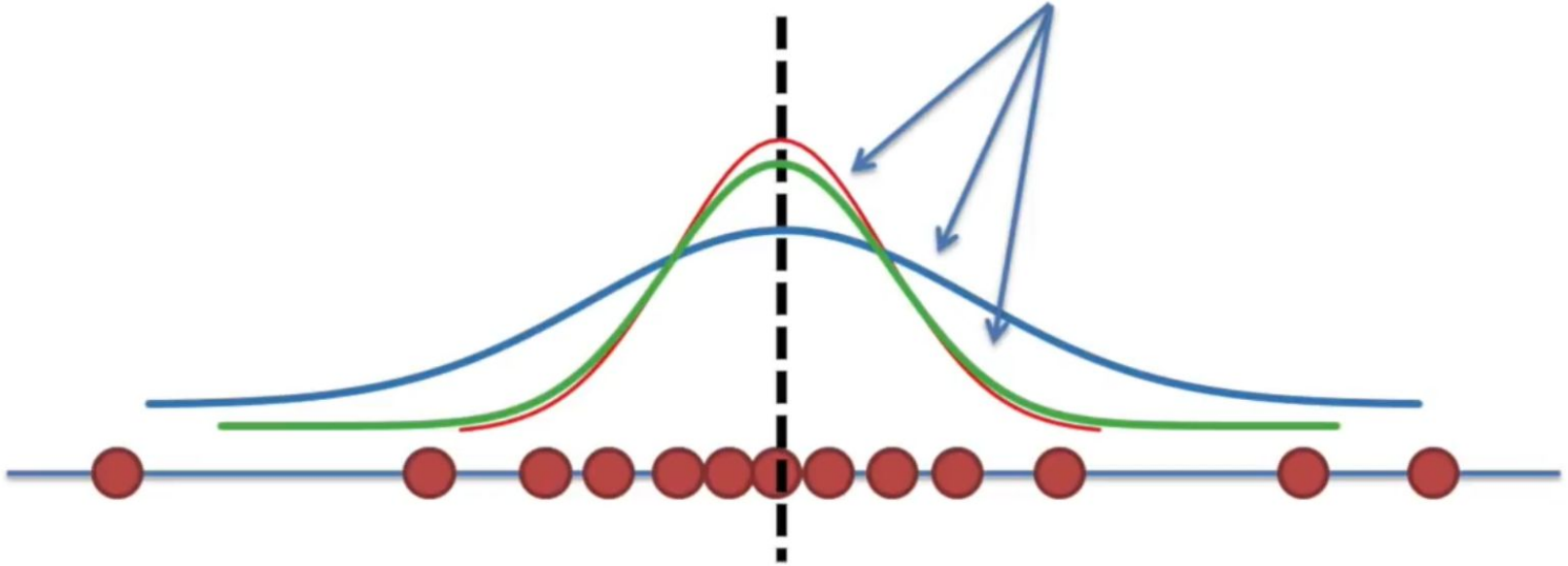

Likelihood of observing the data:
We want the location that "maximizes the likelihood" of observing the weights we measured.
This location for the mean "maximizes the likelihood" of observing the weights we measured.

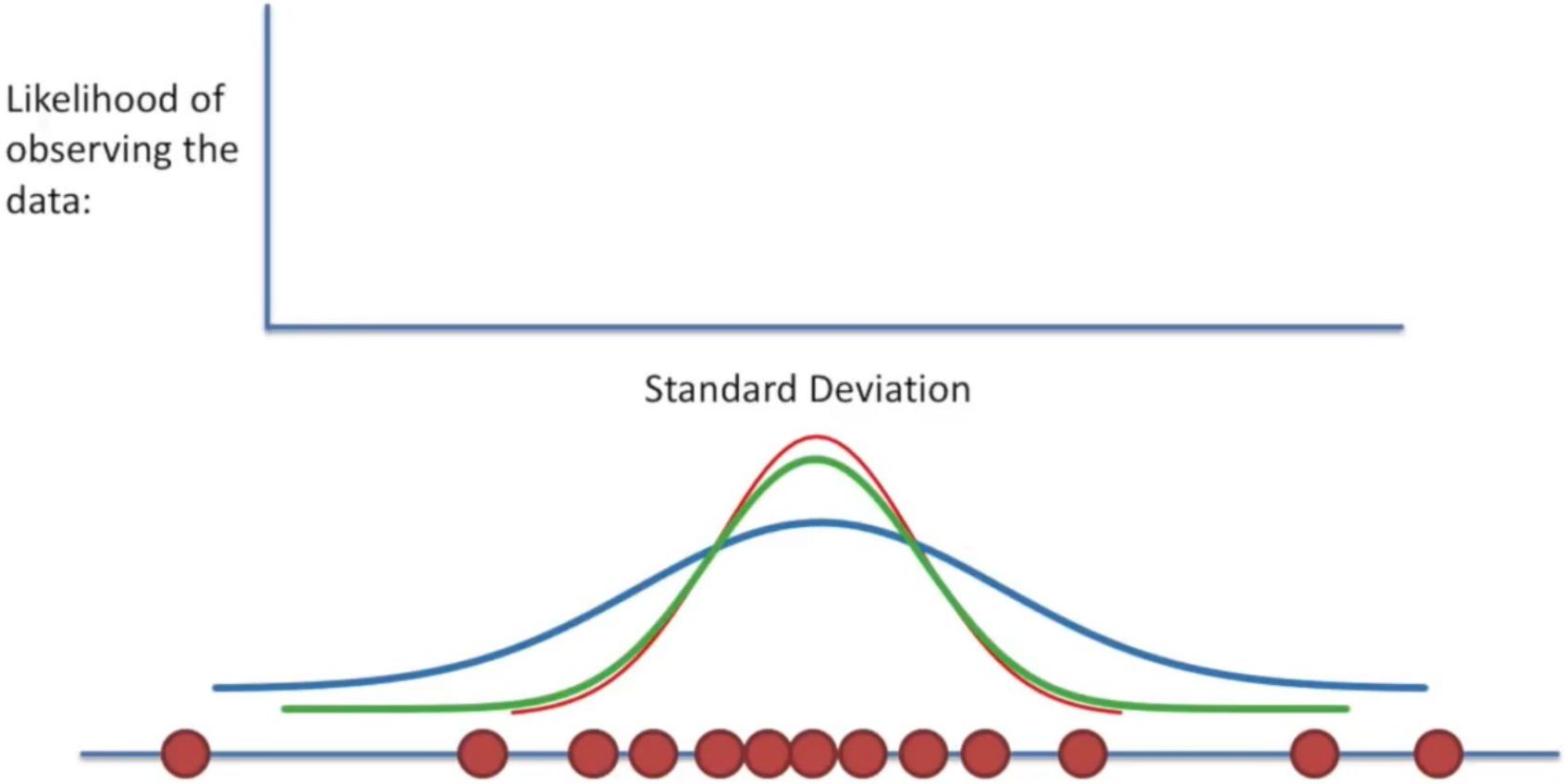

Likelihood of observing the data:
We want the location that "maximizes the likelihood" of observing the weights we measured.
This location for the mean "maximizes the likelihood" of observing the weights we measured.
Thus, it is the "maximum likelihood estimate for the mean."

Great! Now we have figured out the
maximum likelihood estimate for the mean!

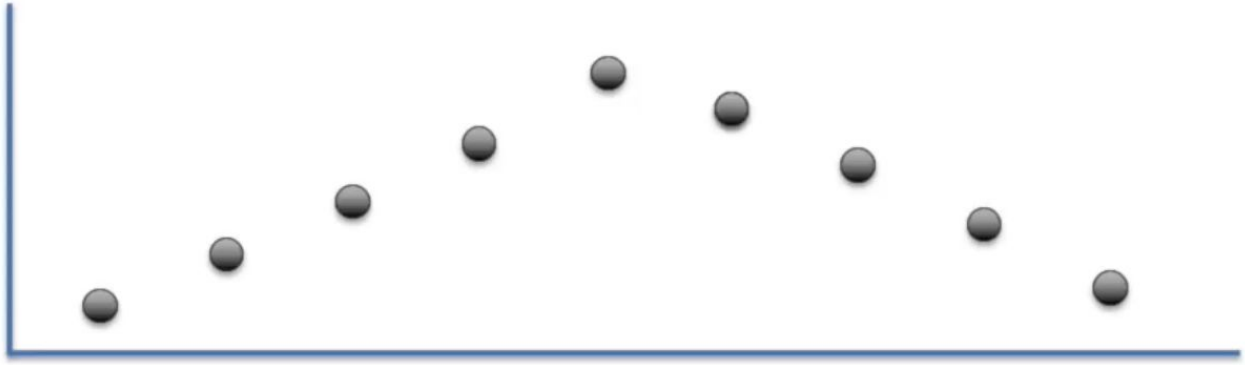


Now we have to figure out the "maximum likelihood estimate for the standard deviation…."

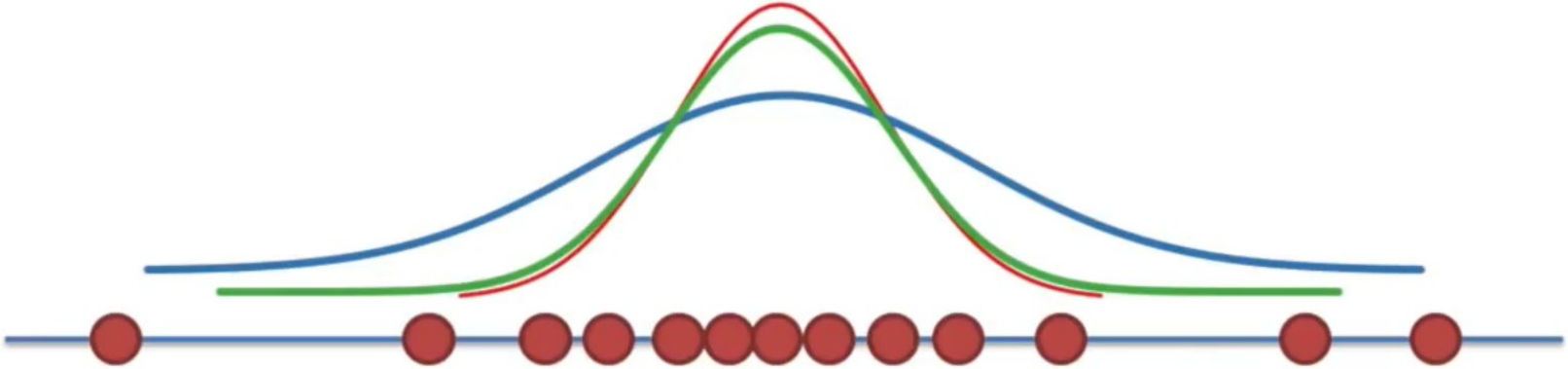



Likelihood of observing the data:
Standard Deviation

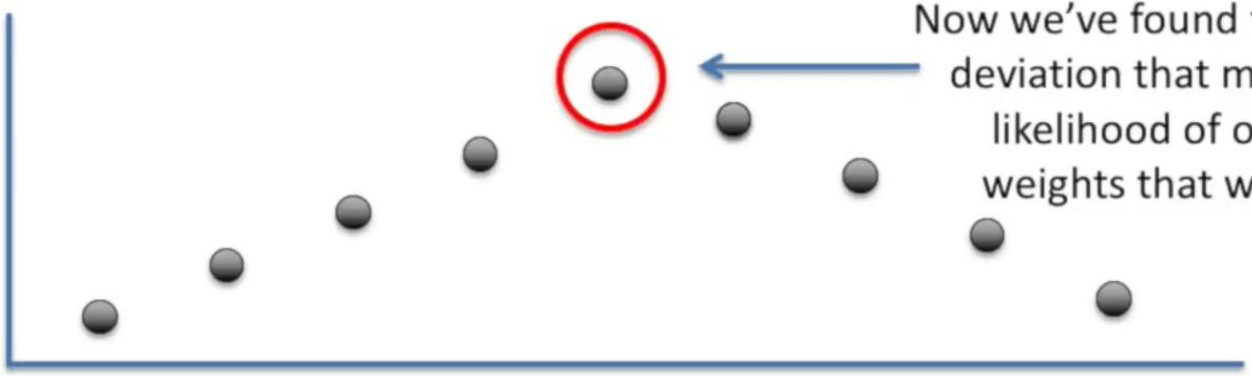
Likelihood of observing the data:



Standard Deviation

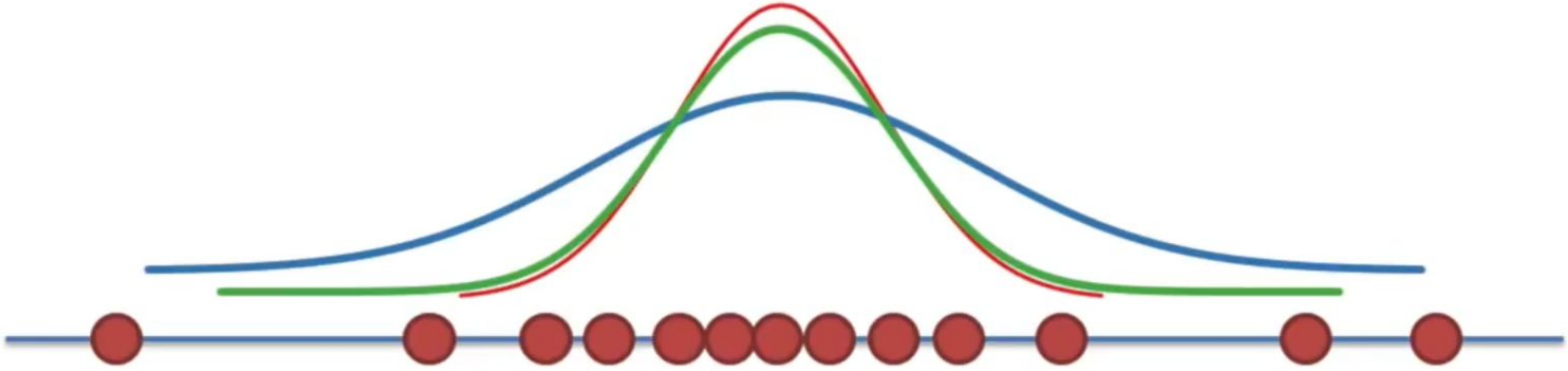


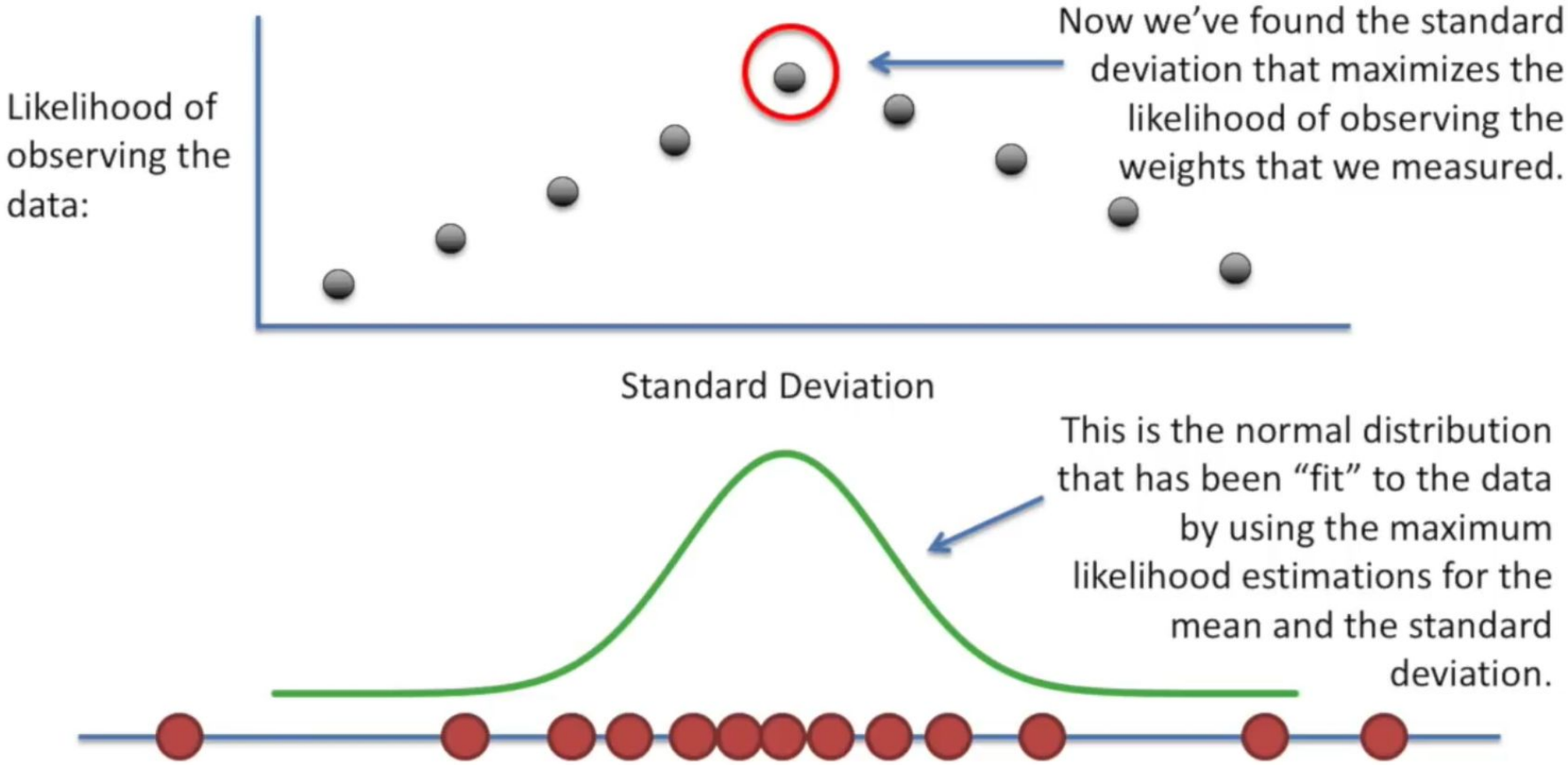
Likelihood of observing the data:



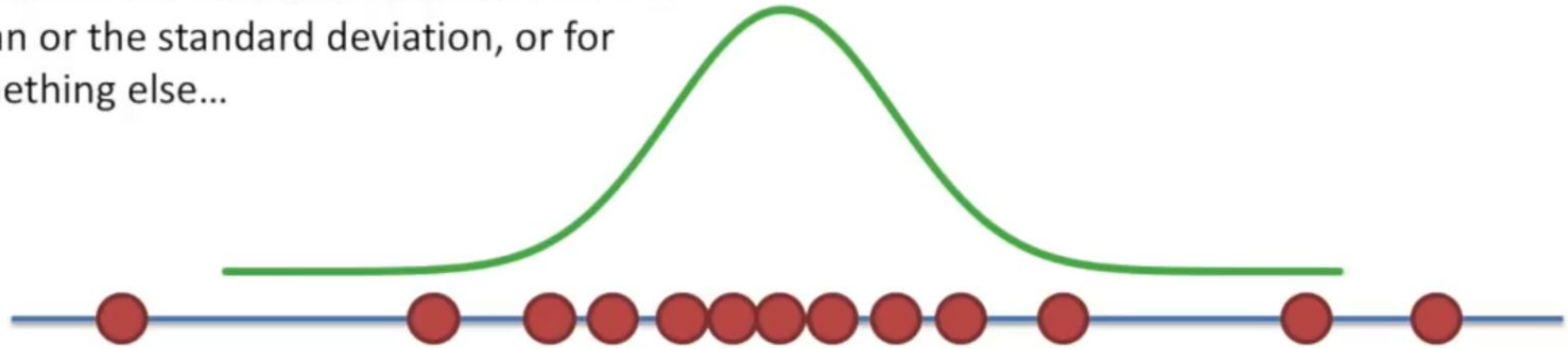
Now we've found the standard deviation that maximizes the likelihood of observing the weights that we measured.

Standard Deviation

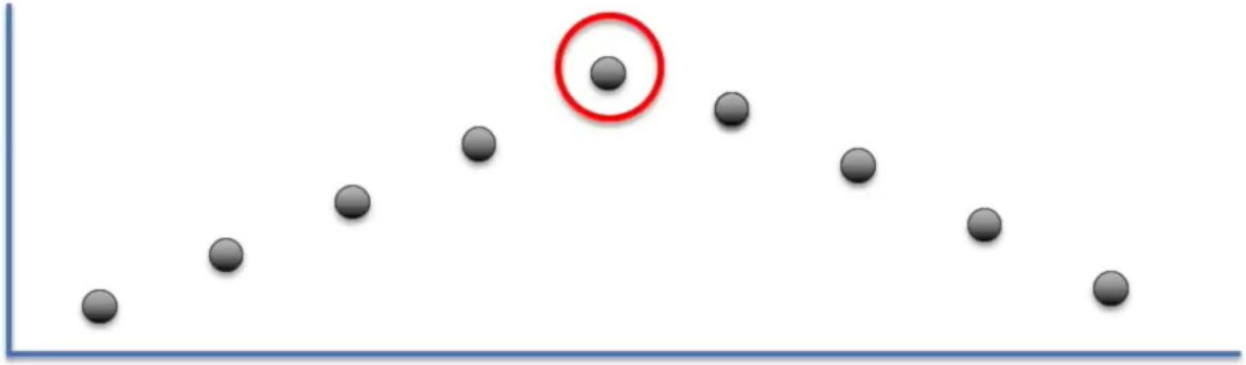



Likelihood of observing the data:
Now we've found the standard deviation that maximizes the likelihood of observing the weights that we measured.
Standard Deviation
This is the normal distribution that has been "fit" to the data by using the maximum likelihood estimations for the mean and the standard deviation.

Now when someone says that they have the maximum likelihood estimates for the mean or the standard deviation, or for something else…

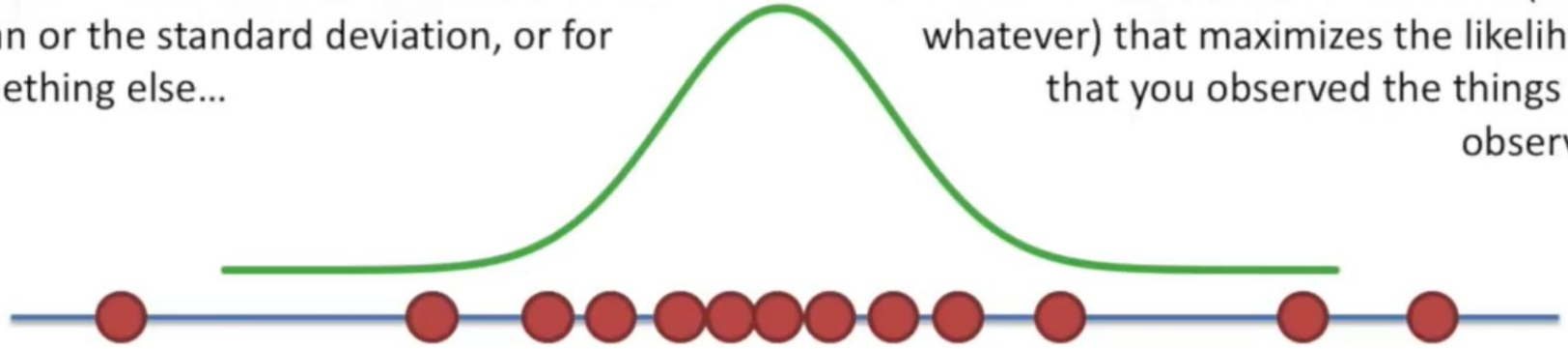


Likelihood of observing the data:



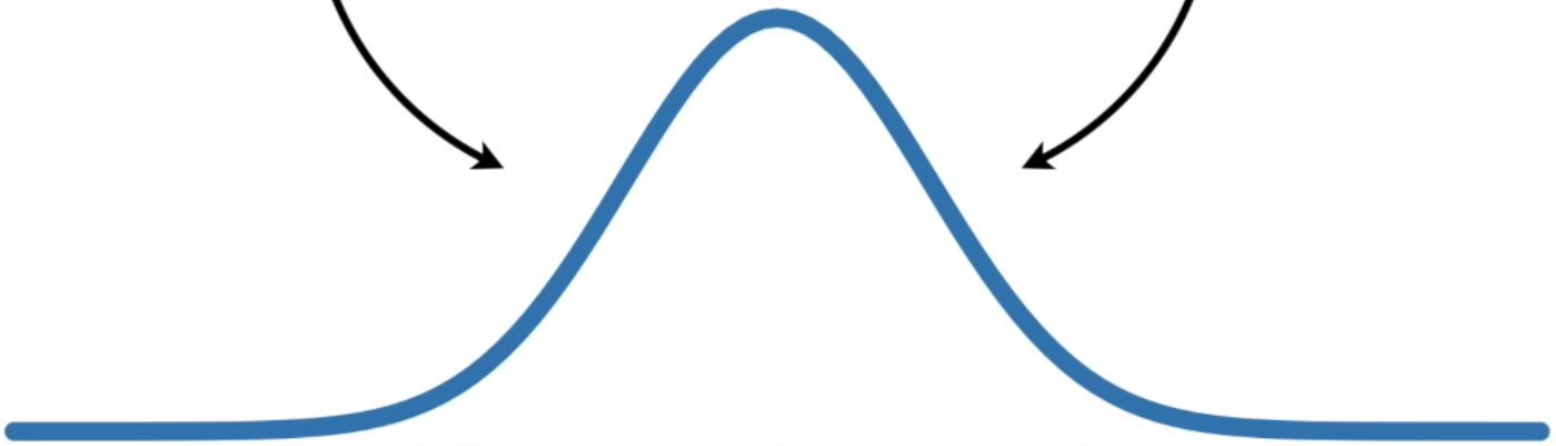
Now when someone says that they have the maximum likelihood estimates for the mean or the standard deviation, or for something else…

… you know that they found the value for the mean or the standard deviation (or for whatever) that maximizes the likelihood that you observed the things you observed.



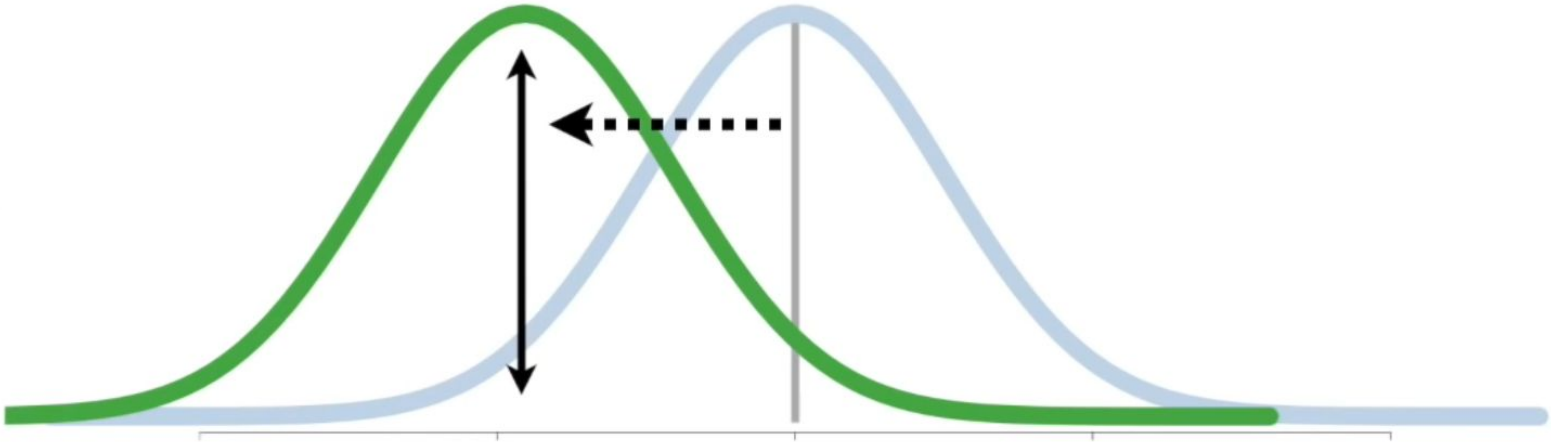
…it's the equation for the **normal distribution**, or **normal curve**.

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



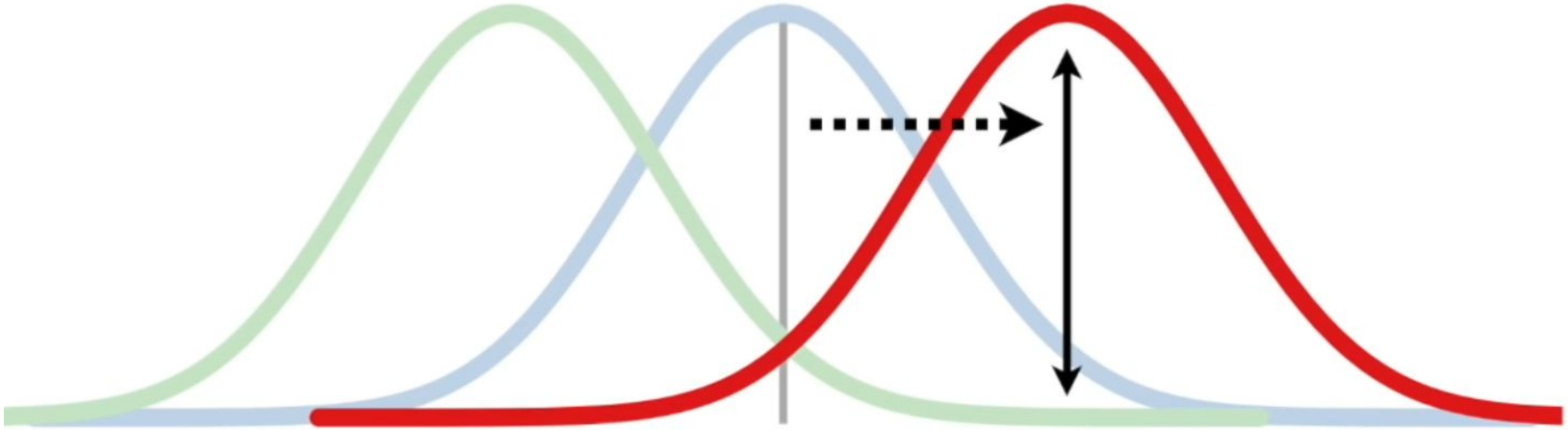
A smaller value for $\mu$ moves the **mean** of the distribution to the left…

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



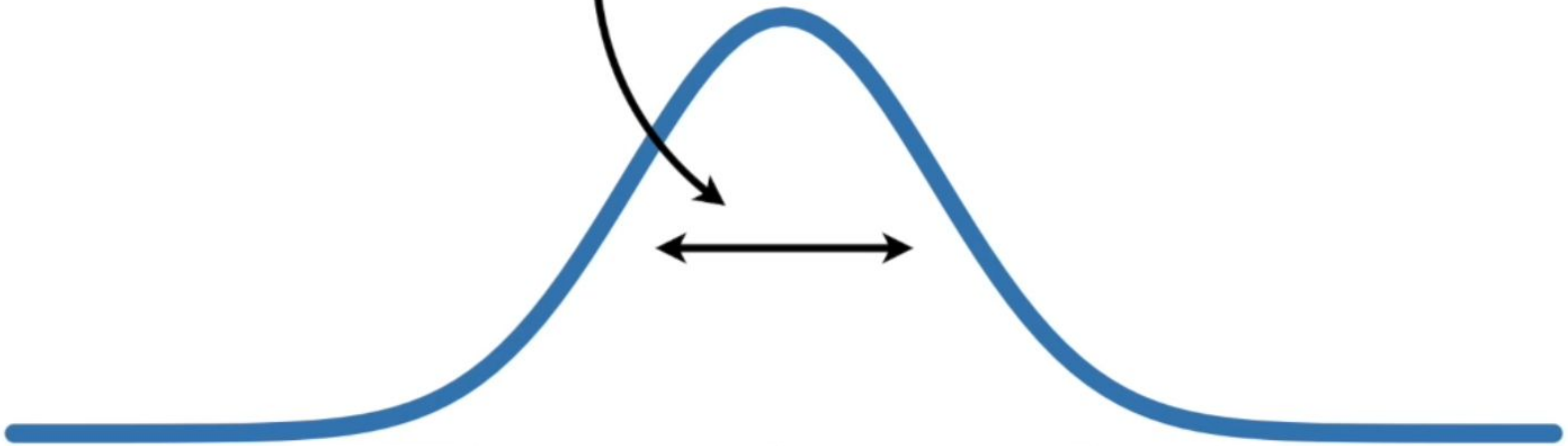
…and a larger value for $\mu$ moves the **mean** of the distribution to the right.

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



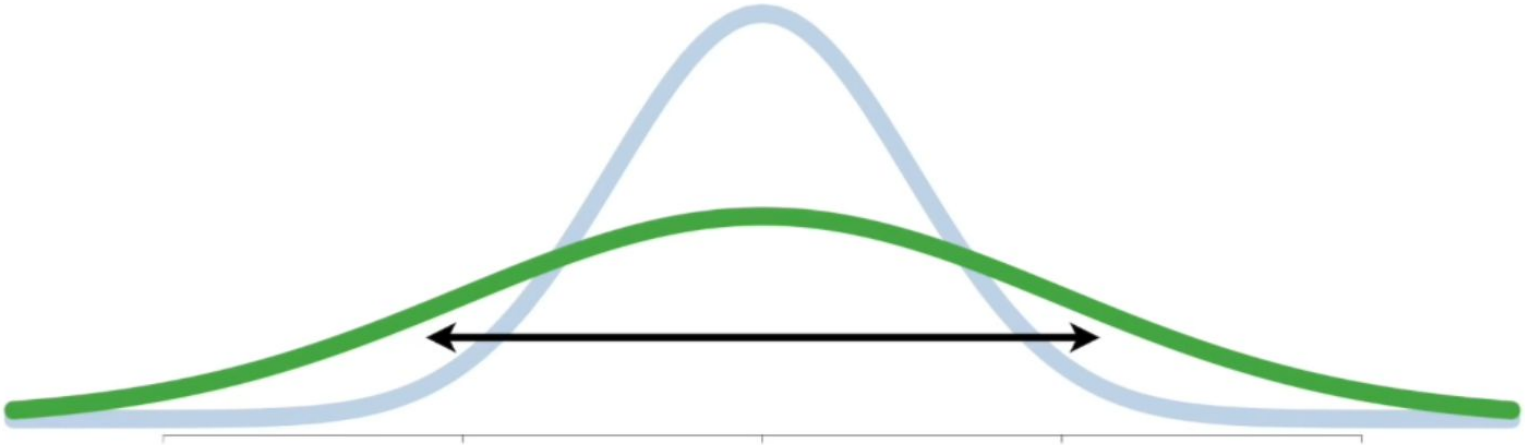
The second parameter, the Greek character $\sigma$, is the **standard deviation** and determines the normal distribution's width.

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



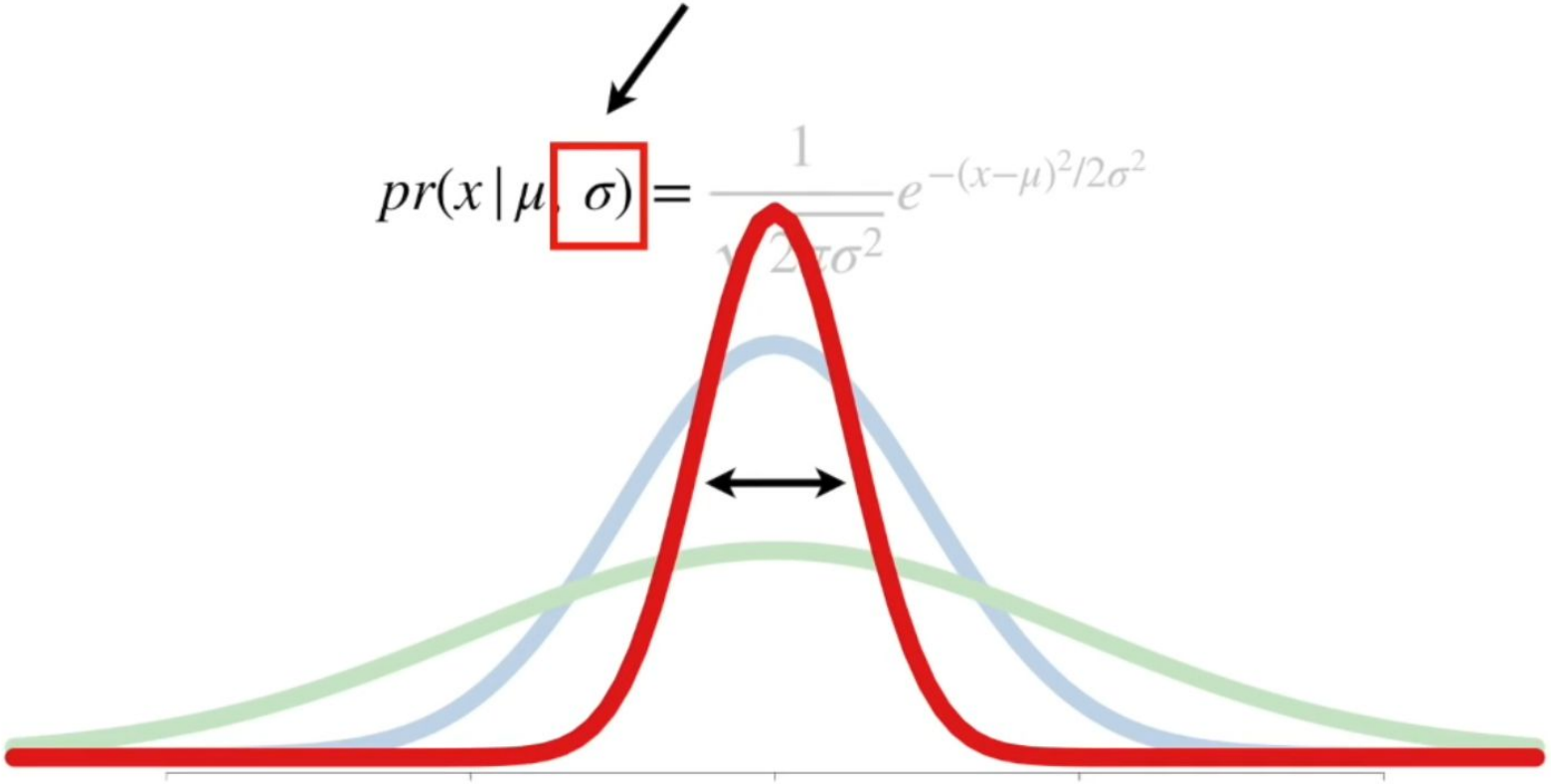
A larger value for $\boldsymbol{\sigma}$ makes the normal curve shorter and wider…

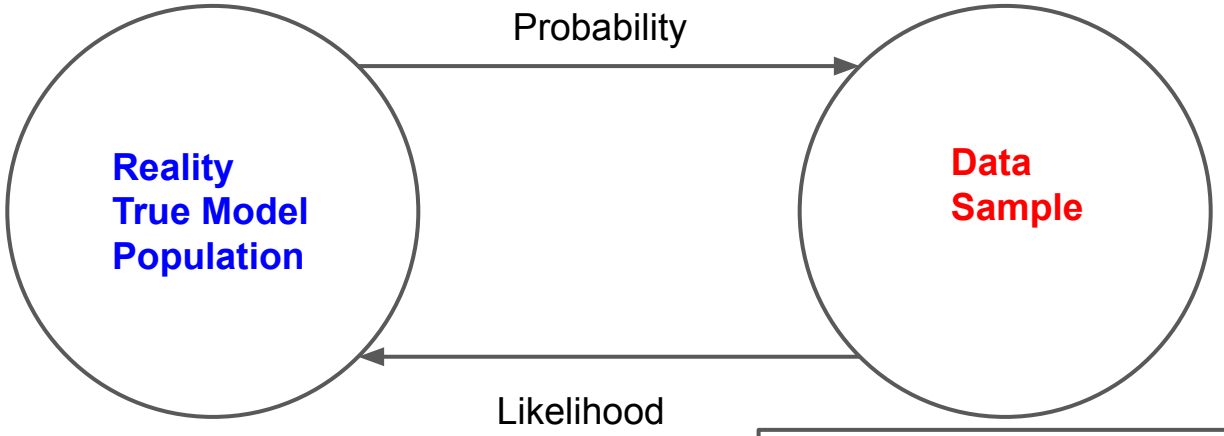
$$pr(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$$



…and a smaller value for $\sigma$ makes the normal curve taller and narrower.

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






…to find the optimal values for $\mu$ (the mean) and $\sigma$ (the standard deviation) given some data, $x$

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

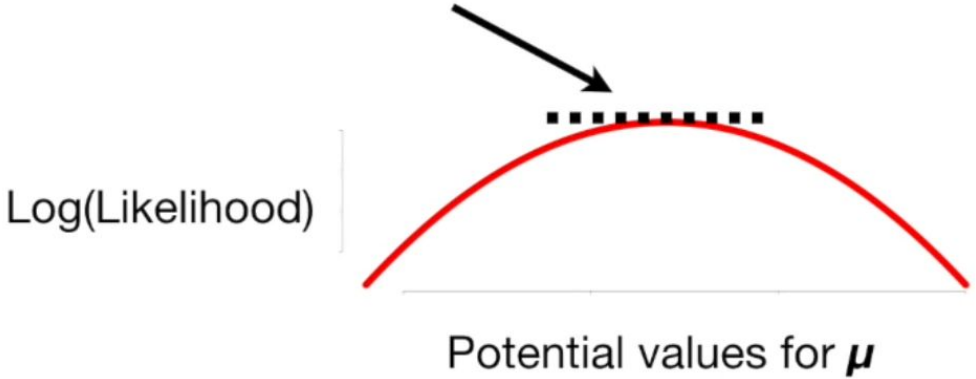
$$L(\mu, \sigma \mid x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$$

$$\ln\left[L(\mu, \sigma \mid x_1, \ \ldots, x_n)\right]$$

$$= -\frac{n}{2}\ln(2\pi) - n\ln(\sigma) - \frac{(x_1-\mu)^2}{2\sigma^2} - \ldots - \frac{(x_n-\mu)^2}{2\sigma^2}$$

$$\frac{\partial}{\partial\mu}\ln\left[L(\mu, \sigma \mid x_1, \ \ldots, x_n)\right]$$

…and we'll use it to find the peak, aka where the slope = **0**.



$$\ln\left[L(\mu, \sigma \mid x_1, \ \ldots, x_n)\right]$$

$$= -\frac{n}{2}\ln(2\pi) - n\ln(\sigma) - \frac{(x_1-\mu)^2}{2\sigma^2} - \ldots - \frac{(x_n-\mu)^2}{2\sigma^2}$$

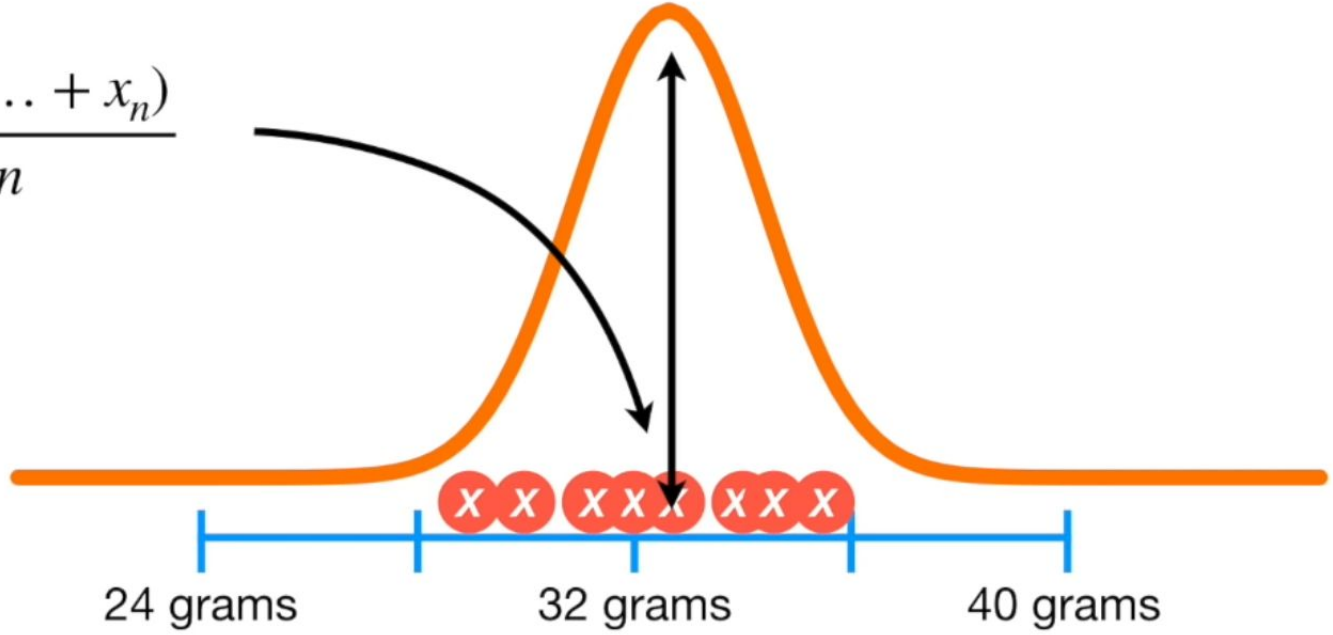
$$\frac{\partial}{\partial\sigma}\ln\left[L(\mu, \sigma \mid x_1, \ \ldots, x_n)\right]$$

This derivative is the slope function for the Log(likelihood) curve…

After a long derivation……..

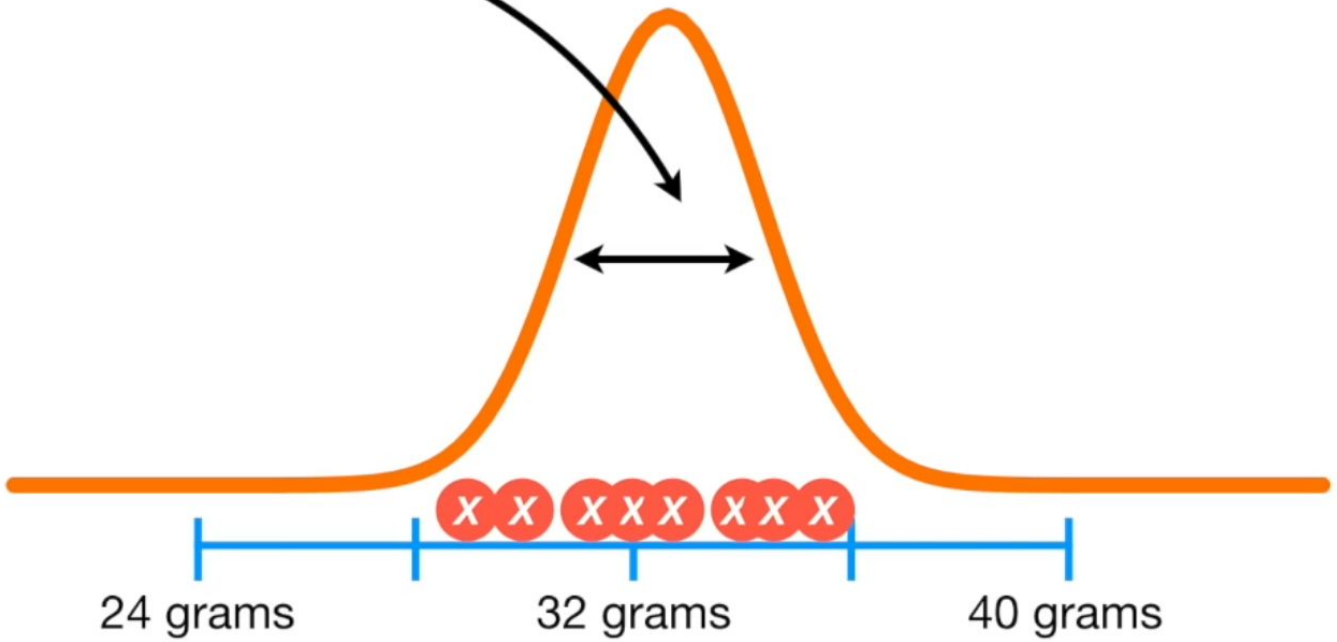
The mean of the data is the maximum likelihood estimate for where the center of the normal distribution should go…

$$\mu = \frac{(x_1 + \ldots + x_n)}{n}$$



$$\sigma = \sqrt{\frac{(x_1 - \mu)^2 + \ldots + (x_n - \mu)^2}{n}}$$

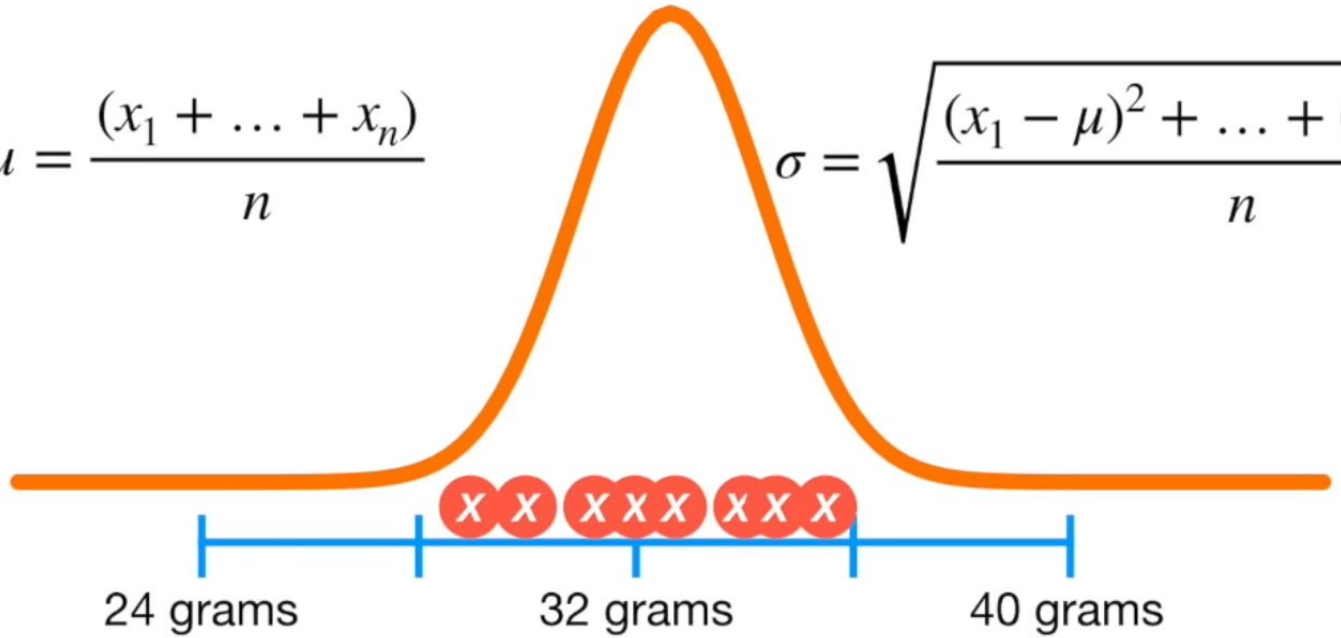
Thus, we use the formula for the standard deviation to determine the width of the normal curve that, given the data, has the maximum likelihood.



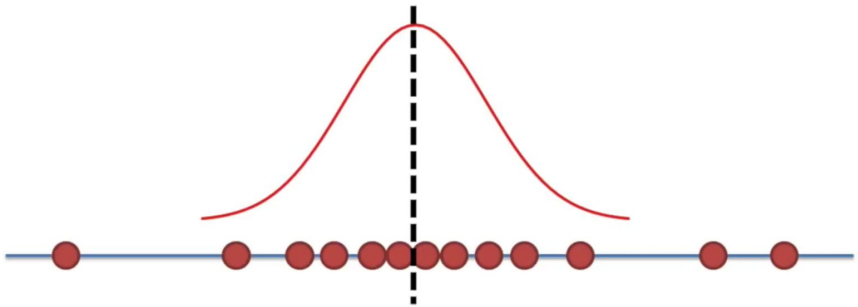
**NOTE:** These solutions may be obvious, but now we have the math that proves that our intuition is correct.

$$\mu = \frac{(x_1 + \ldots + x_n)}{n}$$

$$\sigma = \sqrt{\frac{(x_1 - \mu)^2 + \ldots + (x_n - \mu)^2}{n}}$$



In everyday conversation, “probability” and “likelihood” mean the same thing. However, in Statistics, “likelihood” specifically refers to this situation, we have covered here; where you are trying to find the **optimal value** for the **mean or standard deviation for a distribution** given a bunch of observed measurements.



# Maximum Likelihood Estimation

A procedure to:

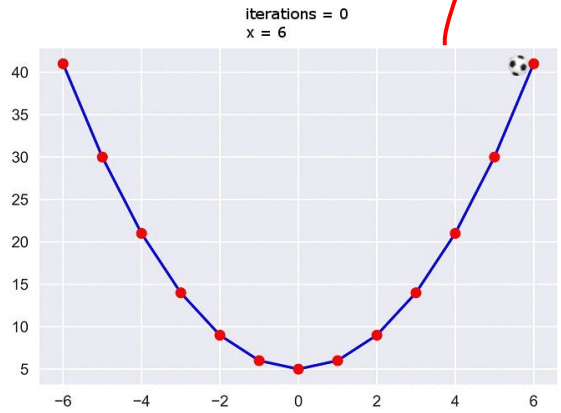
1. Determine best model parameters(reality) that fit given data
2. Compare multiple models to determine the best fit to data

What it does:

1. Maximizes log-likelihood function to estimate parameters

- Example in Python

  - Generate random data for human heights in cm (range from 120cm to 180cm)
  - Using the Log Likelihood equation obtain the model parameters(mean and standard dev) using an optimizer
  - Compare the obtained model parameters with the means and standard deviation of the measured data.
  - Discuss the result.




Notebook: Link

Reference:

- The Main Ideas behind Probability Distributions
- The Normal Distribution, Clearly Explained!!!
- Probability is not Likelihood. Find out why!!!
- Maximum Likelihood For the Normal Distribution, step-by-step!!!
- Machine learning - Maximum likelihood and linear regression