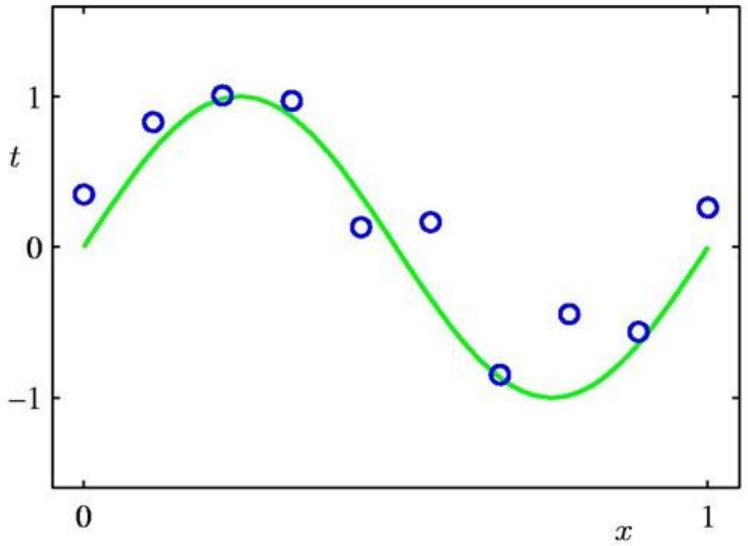
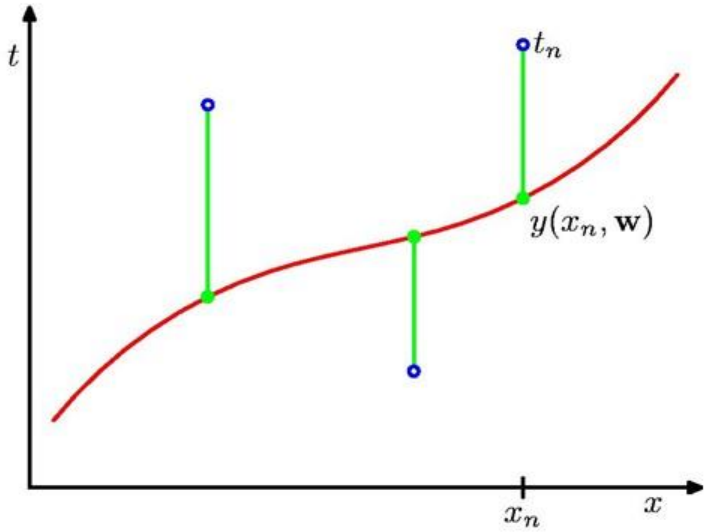
Introduction to Probability distributions

Polynomial Curve Fitting



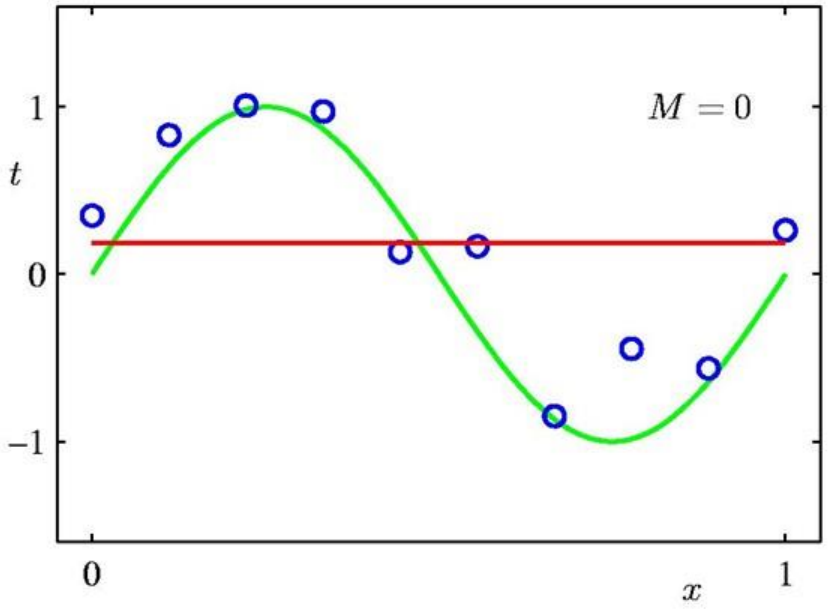
$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

Sum-of-Squares Error Function

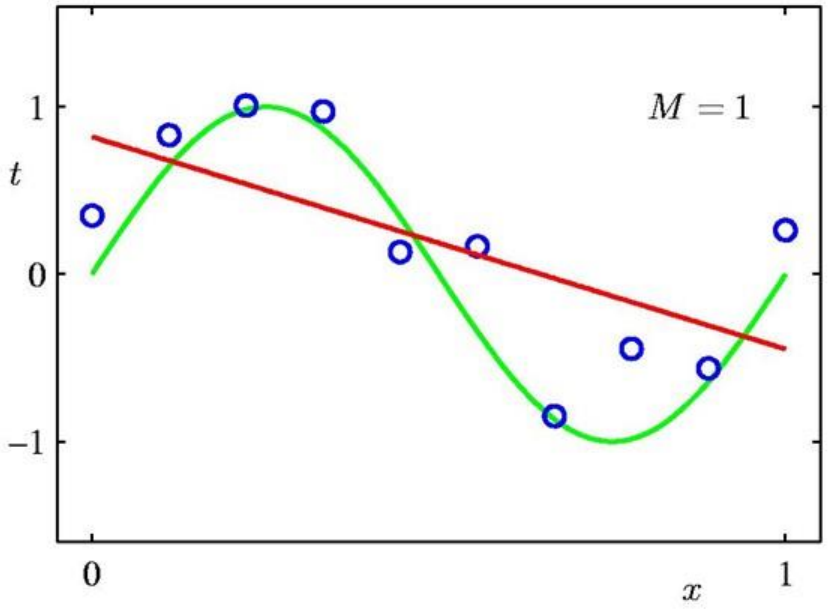


$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2$$

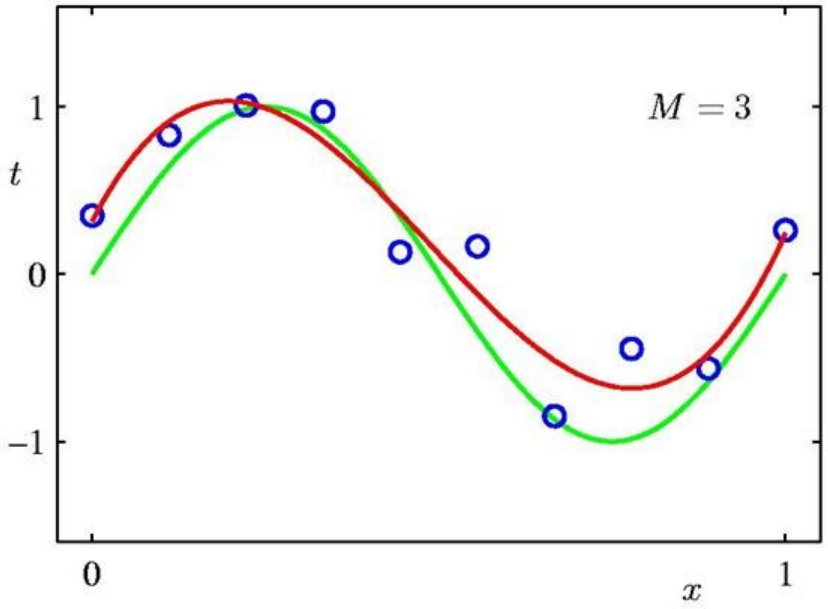
0^{th} Order Polynomial



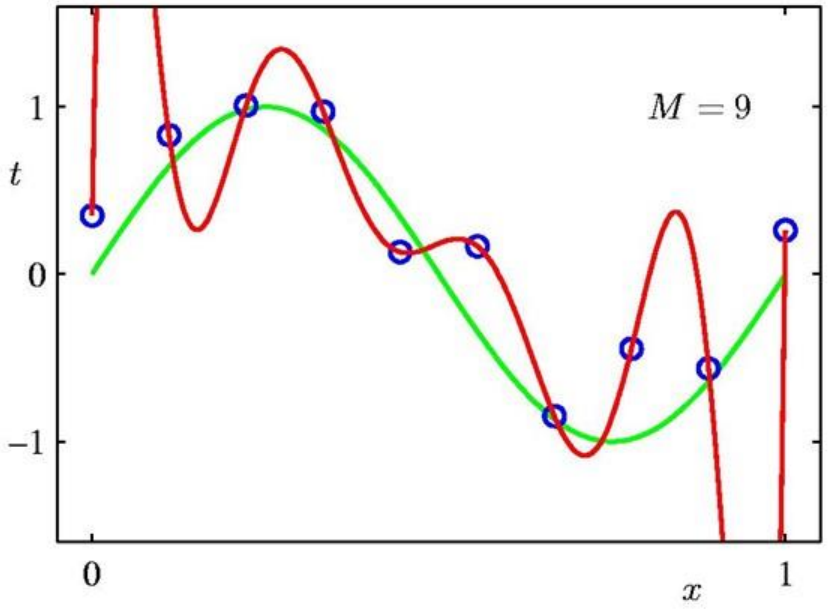
1st Order Polynomial



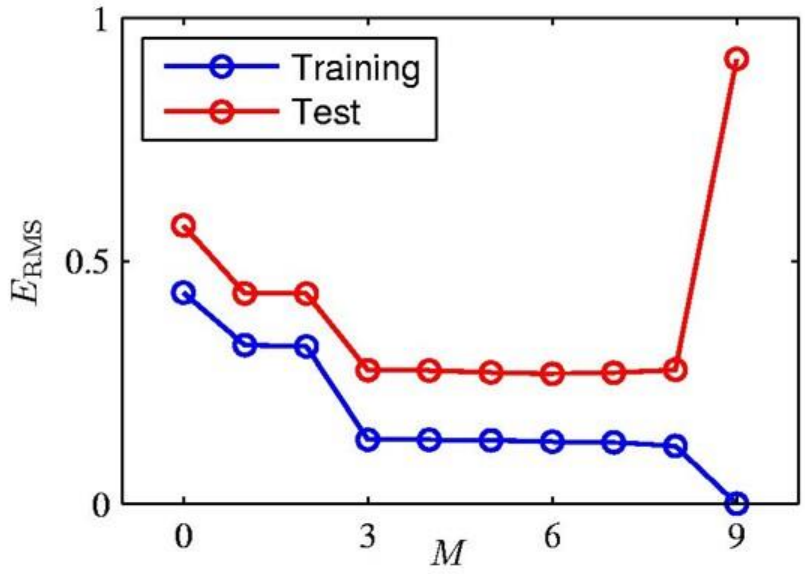
3rd Order Polynomial



9th Order Polynomial



Over-fitting



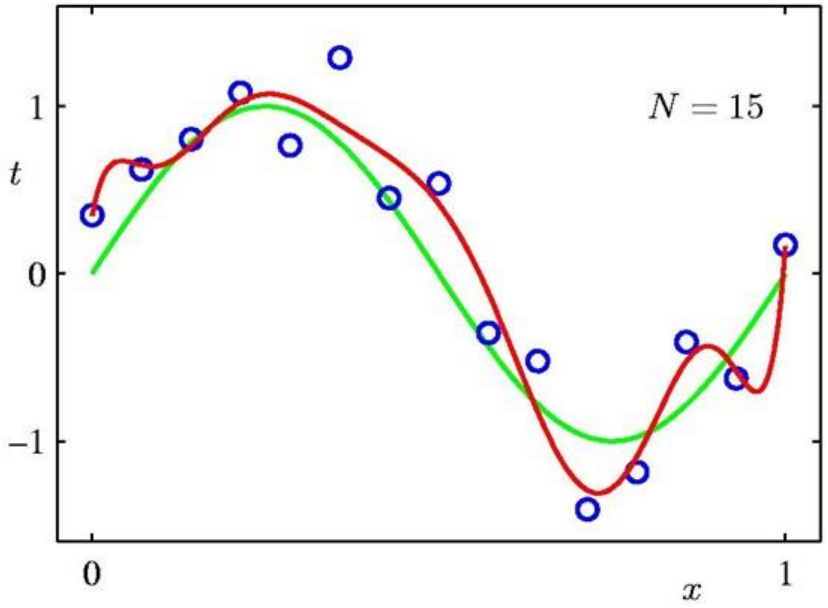
Root-Mean-Square (RMS) Error: $E_{\mathrm{RMS}} = \sqrt{2E(\mathbf{w}^\star)/N}$

Polynomial Coefficients

	$M=0$	$M=1$	$M=3$	$M=9$
$w_0^\star$	0.19	0.82	0.31	0.35
$w_1^\star$		-1.27	7.99	232.37
$w_2^\star$			-25.43	-5321.83
$w_3^\star$			17.37	48568.31
$w_4^\star$				-231639.30
$w_5^\star$				640042.26
$w_6^\star$				-1061800.52
$w_7^\star$				1042400.18
$w_8^\star$				-557682.99
$w_9^\star$				125201.43

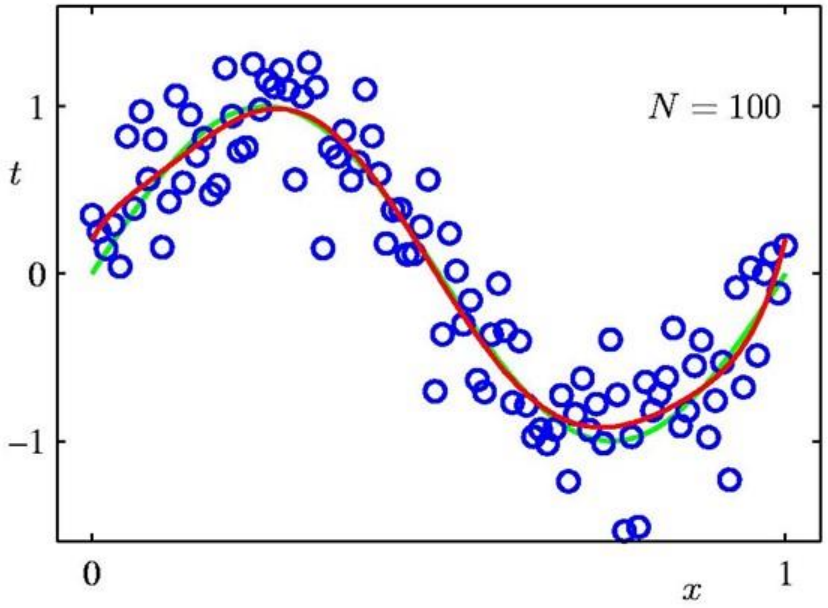
Data Set Size: $N = 15$

9th Order Polynomial



Data Set Size: $N = 100$

9th Order Polynomial

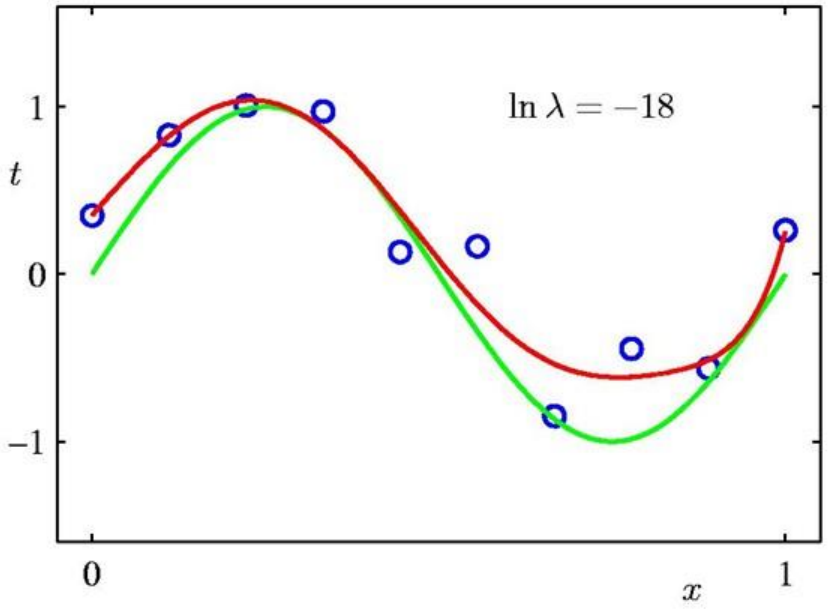


Regularization

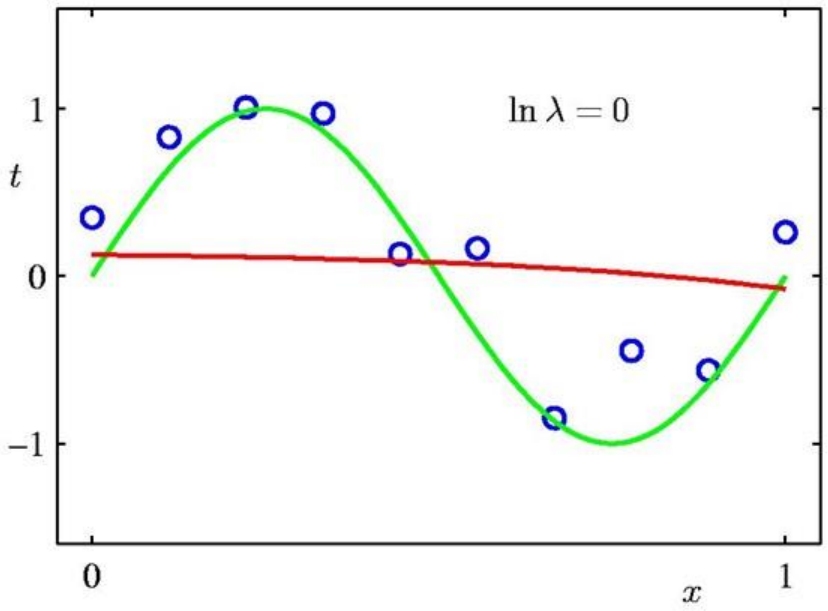
Penalize large coefficient values

$$\widetilde{E}(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2$$

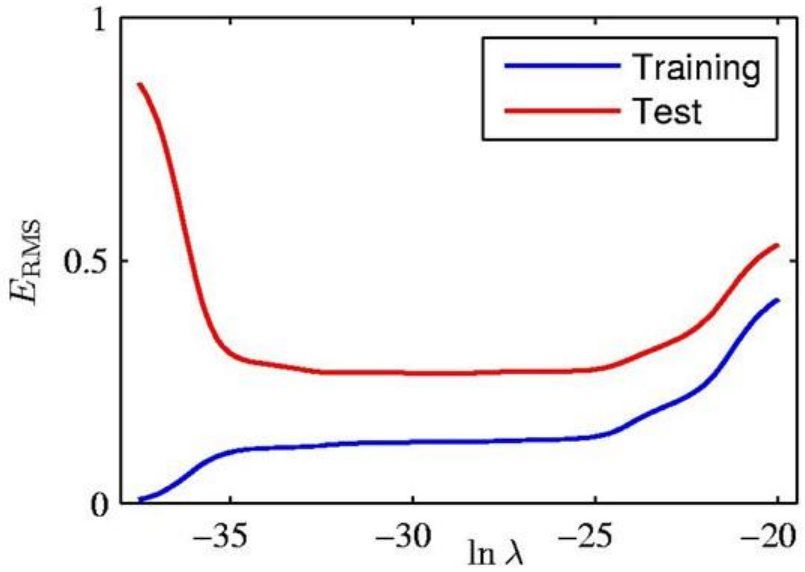
Regularization: $\ln \lambda = -18$



Regularization: $\ln \lambda = 0$



Regularization: E_{RMS} vs. $\ln \lambda$

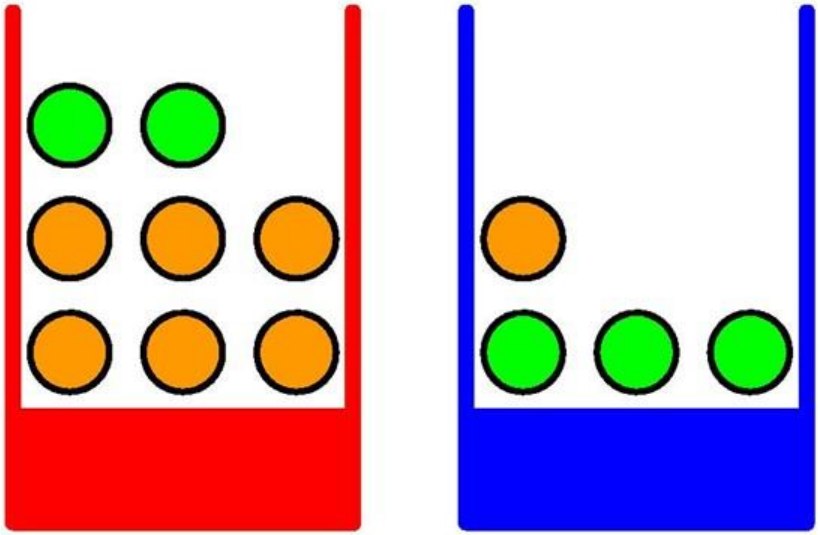


Polynomial Coefficients

	$\ln \lambda = -\infty$	$\ln \lambda = -18$	$\ln \lambda = 0$
$w_0^\star$	0.35	0.35	0.13
$w_1^\star$	232.37	4.74	-0.05
$w_2^\star$	-5321.83	-0.77	-0.06
$w_3^\star$	48568.31	-31.97	-0.05
$w_4^\star$	-231639.30	-3.89	-0.03
$w_5^\star$	640042.26	55.28	-0.02
$w_6^\star$	-1061800.52	41.32	-0.01
$w_7^\star$	1042400.18	-45.95	-0.00
$w_8^\star$	-557682.99	-91.53	0.00
$w_9^\star$	125201.43	72.68	0.01

Probability Theory

Apples and Oranges



- Consider two random variables X and Y
  (which could for instance be the Box and Fruit variables considered above).

- We shall suppose that
  X can take any of the values x_i where $i = 1, \ldots, M$,
  and
  Y can take the values y_j where $j = 1, \ldots, L$.

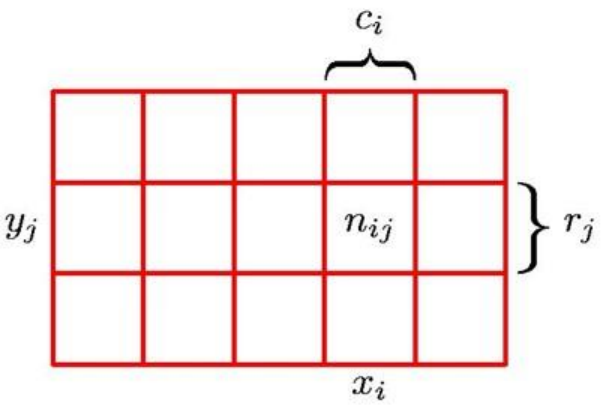
- Consider a total of N trials in which we sample both of the variables X and Y , and

- Let the number of such trials in which $X = x_i$ and $Y = y_j$ be n_{ij} .

- Also, let the number of trials in which X takes the value x_i (irrespective of the value that Y takes) be denoted by c_i ,

- and similarly let the number of trials in which Y takes the value y_j be denoted by r_j.

Probability Theory



Marginal Probability

$$p(X = x_i) = \frac{c_i}{N}.$$

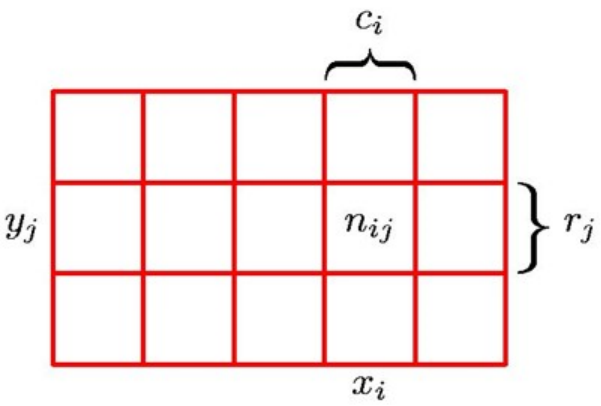
Joint Probability

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$$

Conditional Probability

$$p(Y = y_j | X = x_i) = \frac{n_{ij}}{c_i}$$

Probability Theory



Sum Rule

$$p(X = x_i) = \frac{c_i}{N} = \frac{1}{N}\sum_{j=1}^{L} n_{ij}$$

$$= \sum_{j=1}^{L} p(X = x_i, Y = y_j)$$

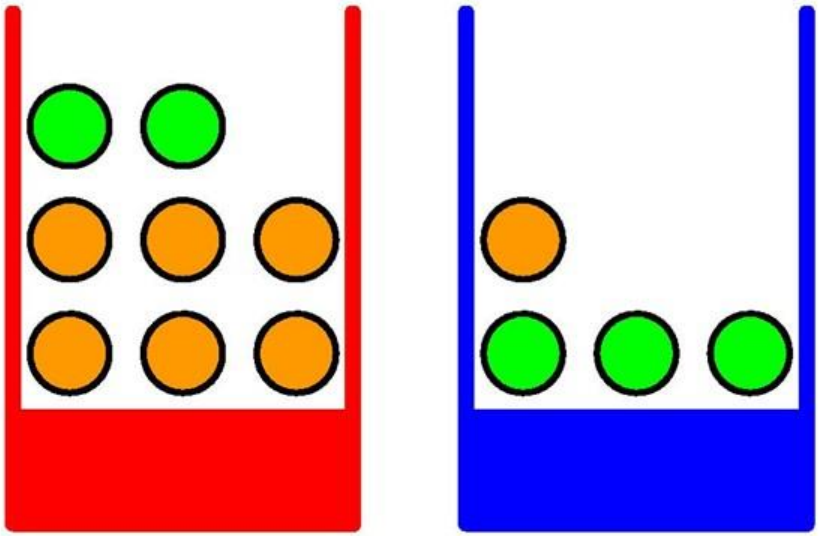
Product Rule

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N} = \frac{n_{ij}}{c_i} \cdot \frac{c_i}{N}$$

$$= p(Y = y_j | X = x_i)p(X = x_i)$$

Probability Theory

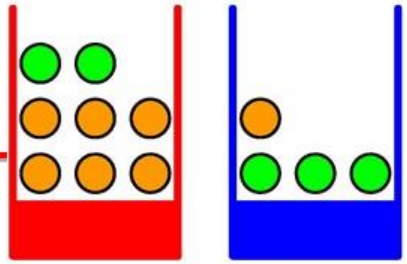
Apples and Oranges



- Let us suppose that we pick (with replacement)
- the **red box 40% of the time** and we pick the **blue box 60% of the time**, and
- that when we remove an item of fruit from a box we are **equally likely to select any of the pieces** of fruit in the box.

- Let random variable **B** denotes the identity of Box chosen.
- This random variable **B** can take two values:
  r (corresponding to the red box) or
  b (corresponding to the blue box).
- Similarly, let random variable **F** denoted the identity of the fruits

  random variable **F** can take values:
  a (for apple) or
  o (for orange).



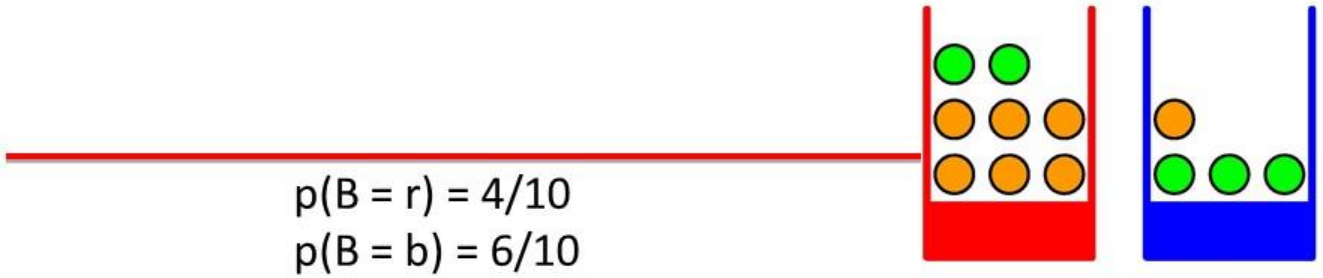
- Probability of choosing a box:

$p(B = r) = 4/10$ and $p(B = b) = 6/10$

- Now Let us find answers of followings:

(i) "What is the overall probability of picking an apple?",

(ii) "Given that we have chosen an orange,
what is the probability that it is from blue box?"



p(B = r) = 4/10
p(B = b) = 6/10

- Note that these satisfy p(B = r) + p(B = b) = 1.

- Now suppose that we pick a box at random, and it turns out to be the blue box.

- Then the probability of selecting an apple is just the fraction of apples in the blue box which is 3/4, and so

$$p(F = a \mid B = b) = 3/4.$$

- In fact, we can write out all four conditional probabilities for each type of fruit, given the selected box:

$$p(F= a \mid B = r) = 1/4$$
$$p(F= o \mid B = r) = 3/4$$
$$p(F= a \mid B = b) = 3/4$$
$$p(F= o \mid B = b) = 1/4$$

- Again, note that these probabilities are normalized so that

$$p(F = a \mid B = r) + p(F = o \mid B = r) = 1$$
$$p(F = a \mid B = b) + p(F = o \mid B = b) = 1.$$

and similarly

- We can now use the sum and product rules of probability to evaluate the overall probability of choosing an apple.

- The overall probability of choosing an apple

$$p(F = a) = p(F = a|B = r)p(B = r) + p(F = a|B = b)p(B = b)$$

$$p(F=a)=11/20$$

- from which it follows, using the sum rule, than

$$p(F = o) = 1 - 11/20 = 9/20.$$

- Suppose instead we are told that:

  a piece of fruit has been selected and it is an orange,

  and we would like to know which box it came from.

- This **requires** that we evaluate

  the probability distribution of boxes given the identity of the fruit,

whereas the probabilities **calculated earlier** give

the **probability distribution of fruit** **given** **the identity of the box**.

We can solve the problem of **reversing the conditional probability** by using Bayes' theorem to give:

$$p(B = r \mid F = o) = p(F = o \mid B = r)p(B = r)/p(F = o)$$

$$p(B = r \mid F = o) = 2/3$$

From the sum rule, it then follows that $p(B = b \mid F = o) = 1 - 2/3 = 1/3$.

- So far we have been quite careful to make a distinction between:

  a random variable (such as the box **B** in the fruit example)

  and

  the values that the random variable can take
  (for example r if the box were the red one).

- Thus the probability that **B** takes the value **r** is denoted **p(B = r).**

- Although this helps to avoid ambiguity, it leads to a rather cumbersome notation, and
- in many cases there will be no need for such pedantry.
- Instead, we may simply write **p(B) to denote a distribution over the random variable B**,

  or

- **p(r)** to denote the **distribution** evaluated **for the particular value r**,

  provided that the interpretation is clear from the context.

Interpretation of Bayes' theorem:

- If it is asked "which box had been chosen"
  before being told the identity of the selected item of fruit,

  then the most complete information we have is provided by

  the probability p(B).

- We call this the ***prior probability*** because

  it is the probability available before we observe the identity of the fruit.

Interpretation of Bayes' theorem:

- Once we are told that the **fruit is an orange**,
- we can then use Bayes' theorem to compute the probability $p(B|F)$,
- which we shall call the ***posterior probability*** because it is the probability obtained after we have observed F.

Interpretation of Bayes' theorem...

- Note that in this example, the prior probability of selecting the red box was 4/10,

  so that we were more likely to select the blue box than the red one.

- However, once we have observed that the piece of selected fruit is an orange,

  we find that the posterior probability of the red box is now 2/3,

- so that it is now more likely that the box we selected was in fact the red one.

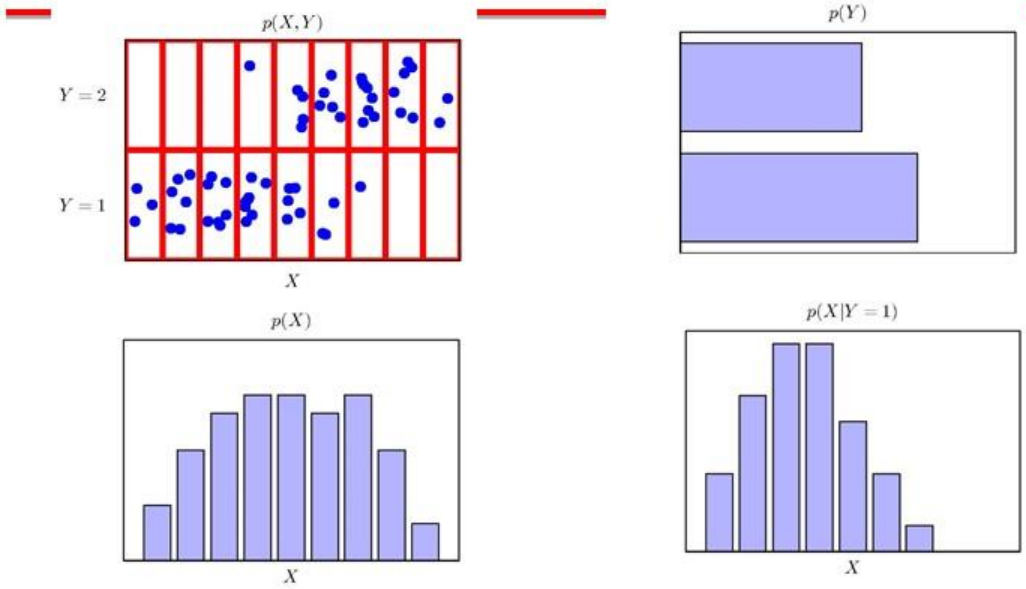
Interpretation of Bayes' theorem...

- This result accords with our intuition,
- as the **proportion of oranges is much higher in the red box** than it is in the blue box,

  and so the observation that the fruit was an orange provides significant evidence favouring the red box.

- In fact, the **evidence is sufficiently strong that it outweighs the prior** and

  makes it more likely that the red box was chosen rather than the blue one.



An illustration of a distribution over two variables, X, which takes 9 possible values, and Y , which takes two possible values. The top left figure shows a sample of 60 points drawn from a joint probability distribution over these variables. The remaining figures show histogram estimates of the marginal distributions p(X) and p(Y), as well as the conditional distribution p(X|Y = 1) corresponding to the bottom row in the top left figure.

The Rules of Probability

Sum Rule $$p(X) = \sum_{Y} p(X, Y)$$

Product Rule $$p(X, Y) = p(Y|X)p(X)$$

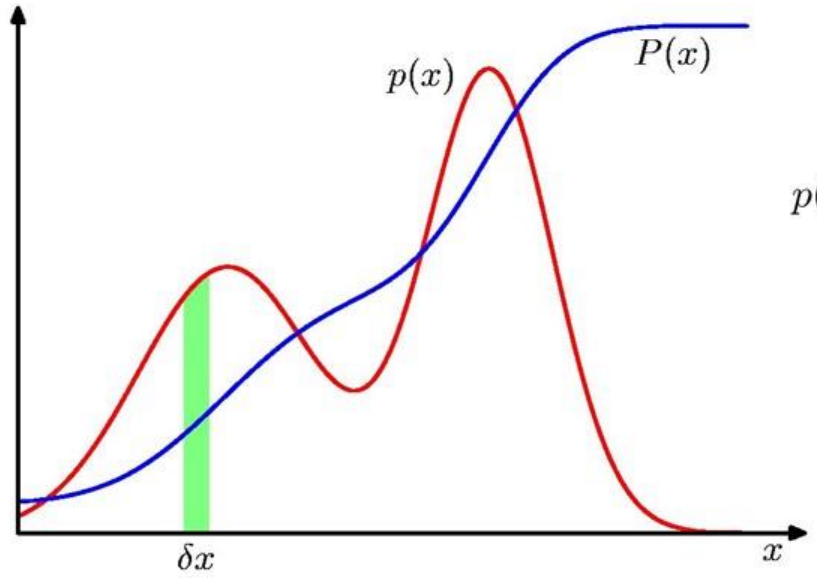
Bayes' Theorem

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}$$

$$p(X) = \sum_{Y} p(X|Y)p(Y)$$

posterior $\propto$ likelihood × prior

Probability Densities

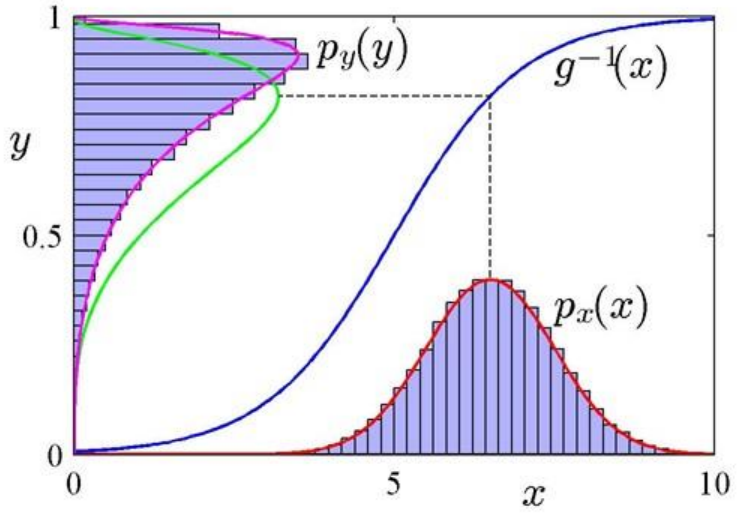


$$p(x \in (a,b)) = \int_a^b p(x)\,\mathrm{d}x$$

$$P(z) = \int_{-\infty}^{z} p(x)\,\mathrm{d}x$$

$$p(x) \geqslant 0 \qquad \int_{-\infty}^{\infty} p(x)\,\mathrm{d}x = 1$$

Transformed Densities



$$p_y(y) = p_x(x)\left|\frac{\mathrm{d}x}{\mathrm{d}y}\right| = p_x(g(y))\,|g'(y)|$$

Expectations

$$\mathbb{E}[f] = \sum_x p(x) f(x)$$

$$\mathbb{E}[f] = \int p(x) f(x)\, \mathrm{d}x$$

$$\mathbb{E}_x[f|y] = \sum_x p(x|y) f(x)$$

Conditional Expectation (discrete)

$$\mathbb{E}[f] \simeq \frac{1}{N} \sum_{n=1}^{N} f(x_n)$$

Approximate Expectation (discrete and continuous)

Variances and Covariances

$$\mathrm{var}[f] = \mathbb{E}\left[(f(x) - \mathbb{E}[f(x)])^2\right] = \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2$$

$$\begin{aligned}
\mathrm{cov}[x, y] &= \mathbb{E}_{x,y}\left[\{x - \mathbb{E}[x]\}\{y - \mathbb{E}[y]\}\right] \\
&= \mathbb{E}_{x,y}[xy] - \mathbb{E}[x]\mathbb{E}[y]
\end{aligned}$$

$$\begin{aligned}
\mathrm{cov}[\mathbf{x}, \mathbf{y}] &= \mathbb{E}_{\mathbf{x},\mathbf{y}}\left[\{\mathbf{x} - \mathbb{E}[\mathbf{x}]\}\{\mathbf{y}^{\mathrm{T}} - \mathbb{E}[\mathbf{y}^{\mathrm{T}}]\}\right] \\
&= \mathbb{E}_{\mathbf{x},\mathbf{y}}[\mathbf{x}\mathbf{y}^{\mathrm{T}}] - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{y}^{\mathrm{T}}]
\end{aligned}$$

Curve Fitting: Bayesian Approach

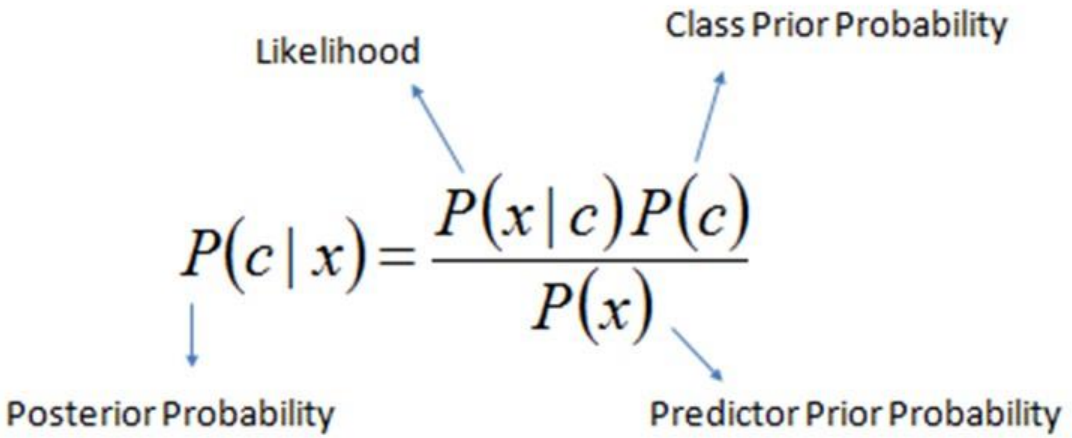
Consider the example of polynomial curve fitting discussed earlier.

Lets model uncertainty in selecting appropriate parameters $\boldsymbol{w}$ (Bayesian perspective)

Recall that in the boxes of fruit example, the **observation of the identity of the fruit** provided relevant information that **altered** the probability that the **chosen box was the red one**.

In that example, Bayes' theorem was used to **convert** a **prior** probability into a **posterior** probability **by incorporating the evidence** provided by the observed data.

As we shall see in detail later, we can adopt a similar approach when making inferences about quantities such as the parameters **w** in the polynomial curve fitting example.



$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

2.1 Decision Using Posterior Probability

- Posterior Probabilities

Define $P(\omega_j|x) = \left(\text{the probability of } \omega \text{ being } \omega_j \text{ given that } x \text{ has been measured}\right)$

Bayes rule derives

$$P(\omega_j|x) = \frac{p(x|\omega_j)P(\omega_j)}{p(x)} \quad (1)$$

- Decision Rule (1) Minimizing error probability

$$\textbf{Decide} \quad \begin{array}{ll} \omega_1 & \text{if } P(\omega_1|x) > P(\omega_2|x) \\ \omega_2 & \text{if } P(\omega_1|x) < P(\omega_2|x) \end{array} \quad (2)$$

- Decision Rule (2) ***Likelihood ratio***

$$\textbf{Decide} \quad \begin{array}{ll} \omega_1 & \\ \omega_2 & \end{array} \text{ if } \frac{p(x|\omega_1)}{p(x|\omega_2)} \begin{array}{l} > \\ < \end{array} \frac{P(\omega_1)}{P(\omega_2)} \quad (3)$$

independent of observation x

10

Continuous Random Variable

- Continuous random variables can have an infinite number (uncountable) of possible values within their defined interval or range.
- Unlike discrete random variables, which can only assume distinct, countable values.
- Continuous random variables are characterized by probability density functions (PDFs).

Continuous Random Variable

- PDFs describe the likelihood or density of the random variable assuming different values within the range.
- The PDF specifies the relative likelihood of observing the variable at any specific value or within any specific interval.

Probability of a Specific Value vs Range of Values (interval)

- In continuous probability distributions, the probability of a specific, exact value is technically zero.
- This is because continuous random variables can take on an infinite number of values within a given range, and
- Thus, the probability associated with any single, isolated point is infinitesimal.

Probability of a Specific Value vs Range of Values (interval)

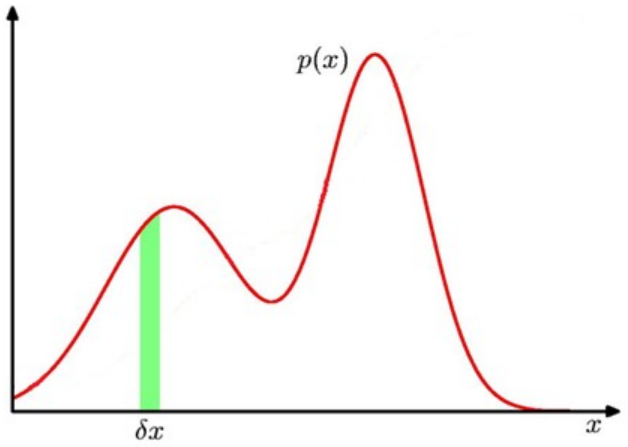
- For example: You measured the height of students (H) in a University campus with very high precision of measurement.
- The height can be 170 cm, 170.01 cm, and 170.00006 cm etc.
- Indeed between any two value say 170 cm and 170.01, there are infinite possible values of height.

Probability of a Specific Value vs Range of Values (interval)

- In case of continuous random variable, it is meaningless to talk about probability of a specific value say $p(H = 170\,\text{cm})$.
- As $p(H = 170\,\text{cm}) = 0$, $H = 170\,\text{cm}$ is one value against infinite possible values.
- Thus, we calculate the probability of the variable falling within a specified interval or range of values in case of continuous random variables.

Probability Density Function (PDF)

- It describes the probability distribution of a continuous random variable and
- It provides information about how the values of that variable are likely to be distributed over a specified range.



- PDF is always non negative.

$$p(x) \geqslant 0$$

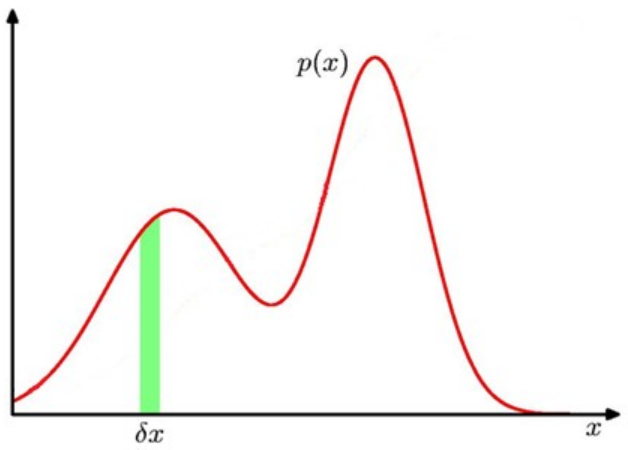
- The total area under the PDF curve over the entire range is equal to 1.

$$\int_{-\infty}^{\infty} p(x)\,\mathrm{d}x = 1$$

Probability in Intervals

- To find the probability that the continuous random variable falls within a specific interval, integrate the PDF over that interval.

- Probability of a continuous random variable falling within a specified interval or range of values, say (*a*, *b*) is



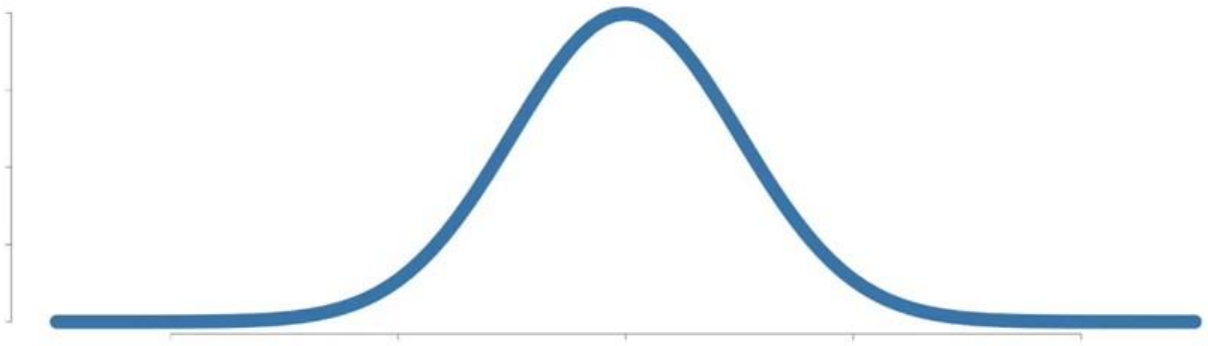
$$p(x) \geqslant 0$$

$$p(x \in (a,b)) = \int_a^b p(x)\,\mathrm{d}x$$

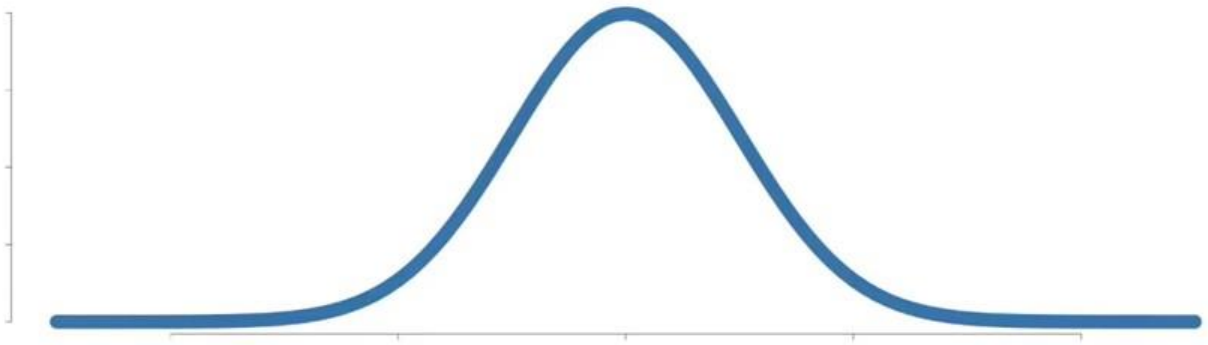
$$\int_{-\infty}^{\infty} p(x)\,\mathrm{d}x = 1$$

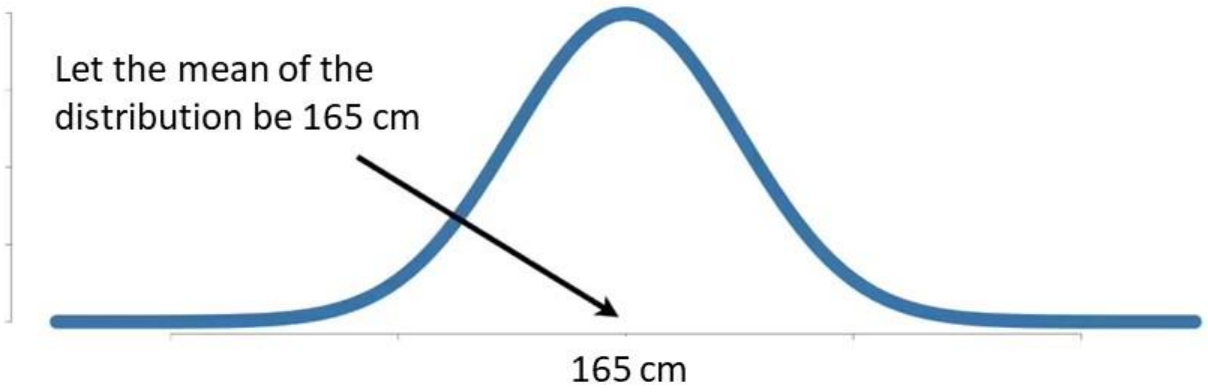
Likelihood

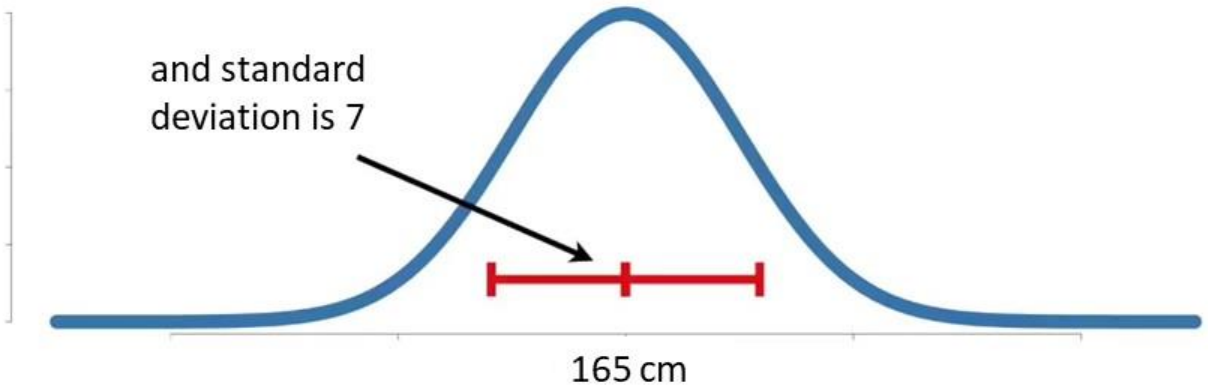
So let's start by looking at probability with respect to a normal distribution (keeping in mind that this concept applies to all continuous distributions).

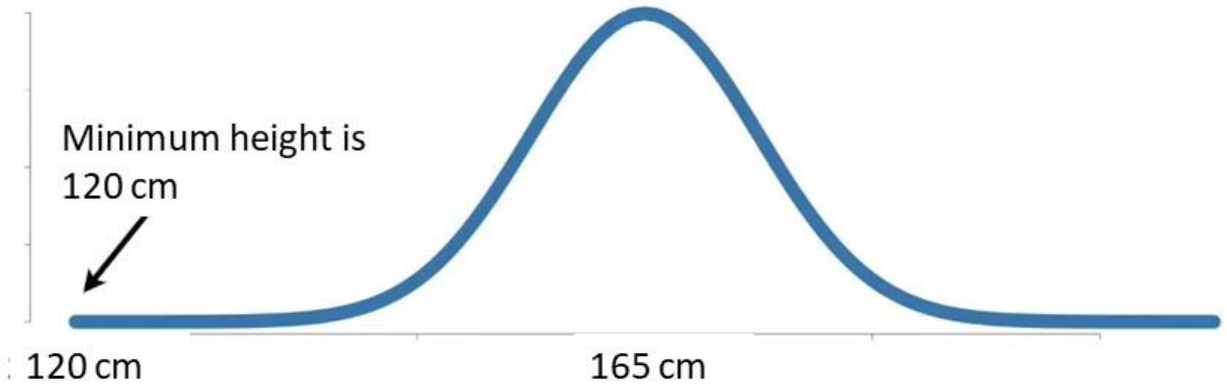


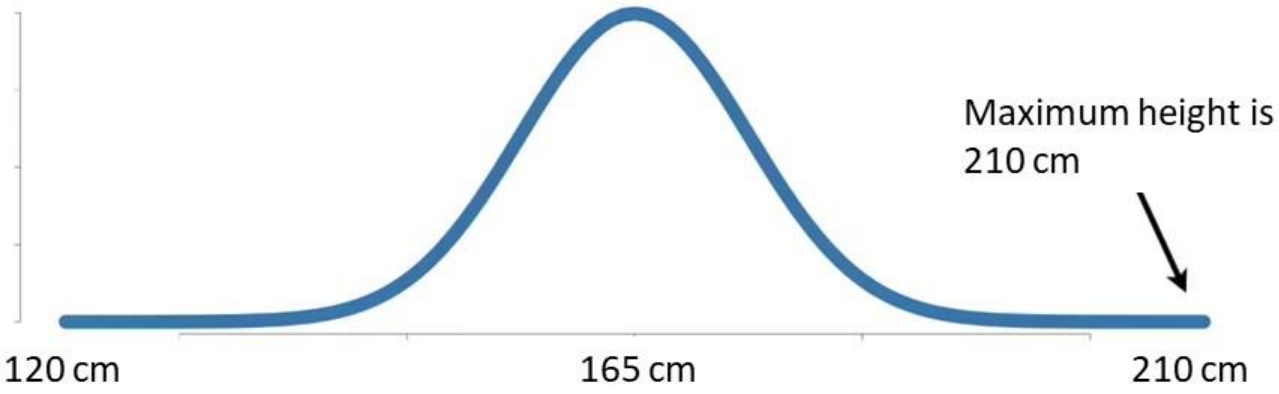
Let it be a distribution of students' height (measured with very high precision) of a University Campus



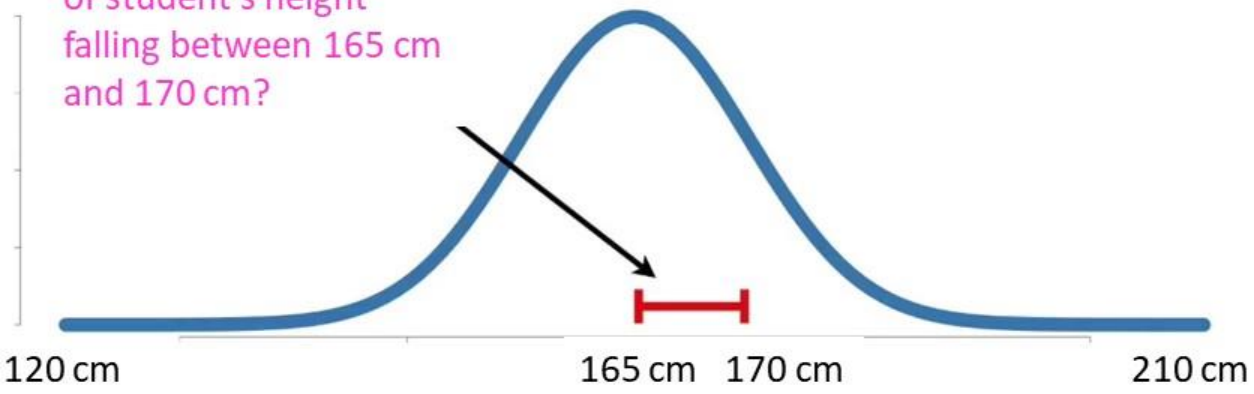

Let the mean of the
distribution be 165 cm
165 cm

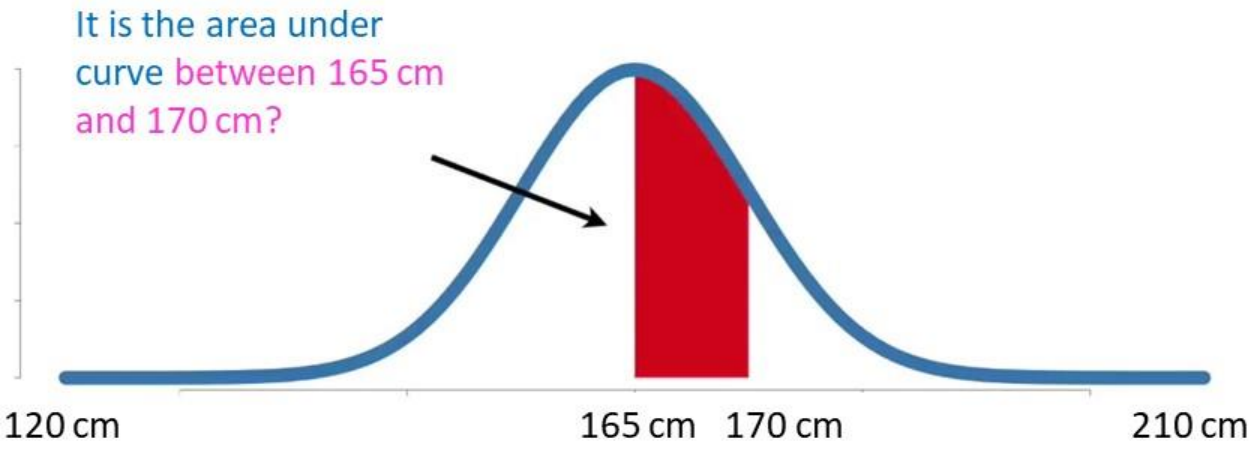

and standard
deviation is 7
165 cm

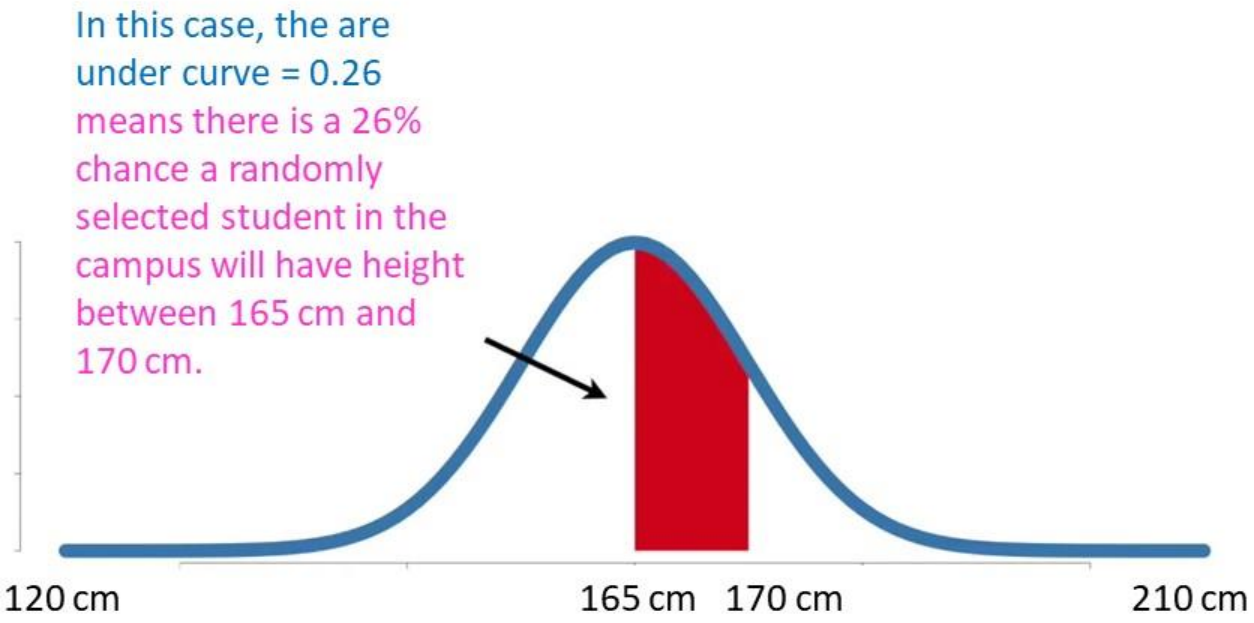

Minimum height is
120 cm
120 cm
165 cm


Maximum height is
210 cm
120 cm
165 cm
210 cm

If we select a random student then what is the probability of student's height falling between 165 cm and 170 cm?

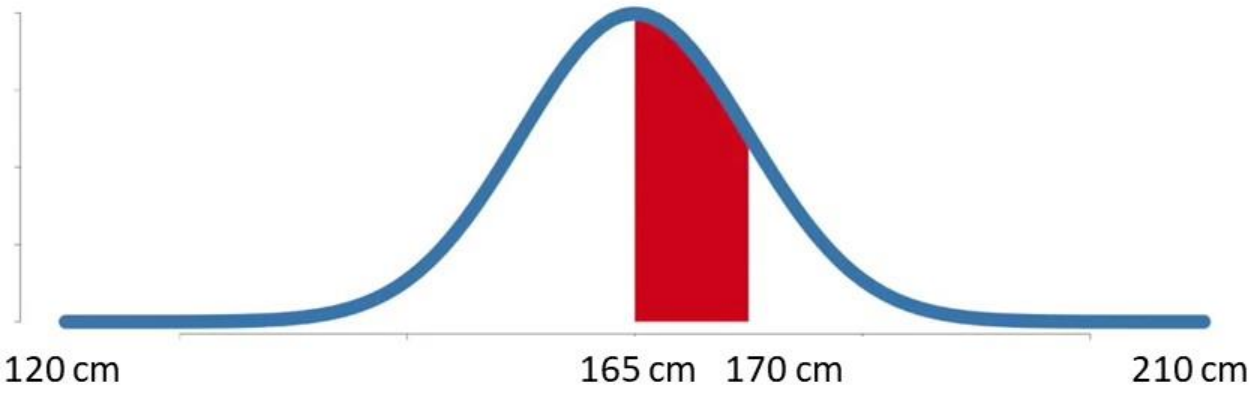



It is the area under
curve between 165 cm
and 170 cm?
120 cm
165 cm
170 cm
210 cm

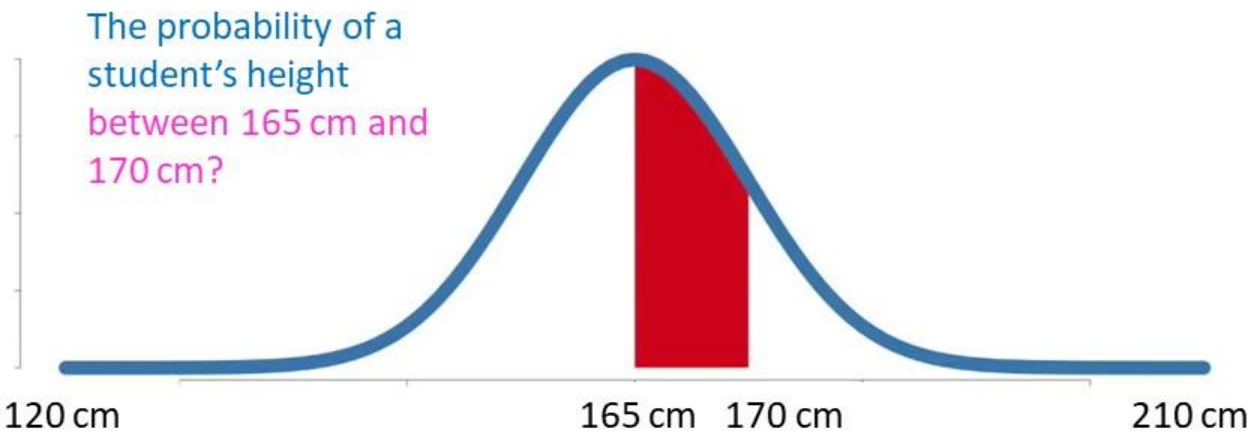

In this case, the are under curve = 0.26 means there is a 26% chance a randomly selected student in the campus will have height between 165 cm and 170 cm.
120 cm
165 cm
170 cm
210 cm

It is denoted as:

p(height between 165 and 170 cm | mean = 165 and SD = 7)

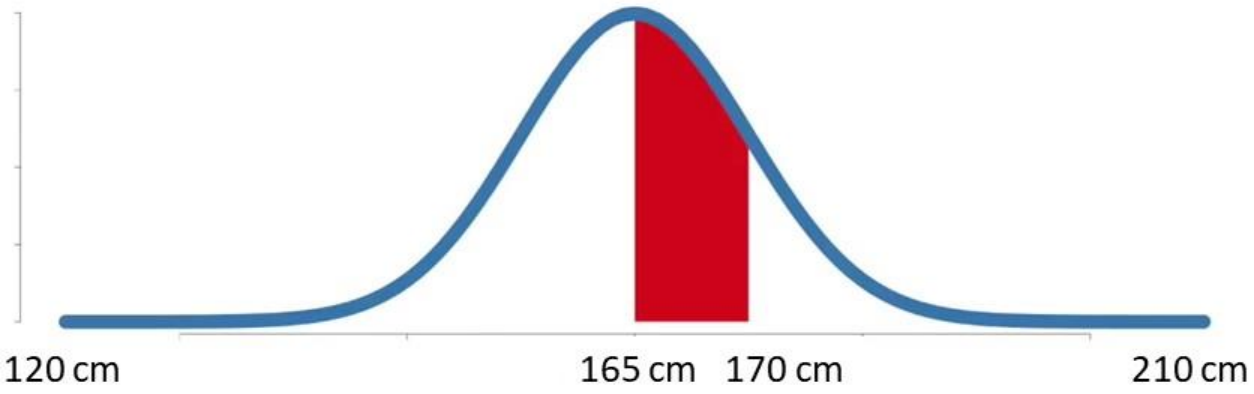


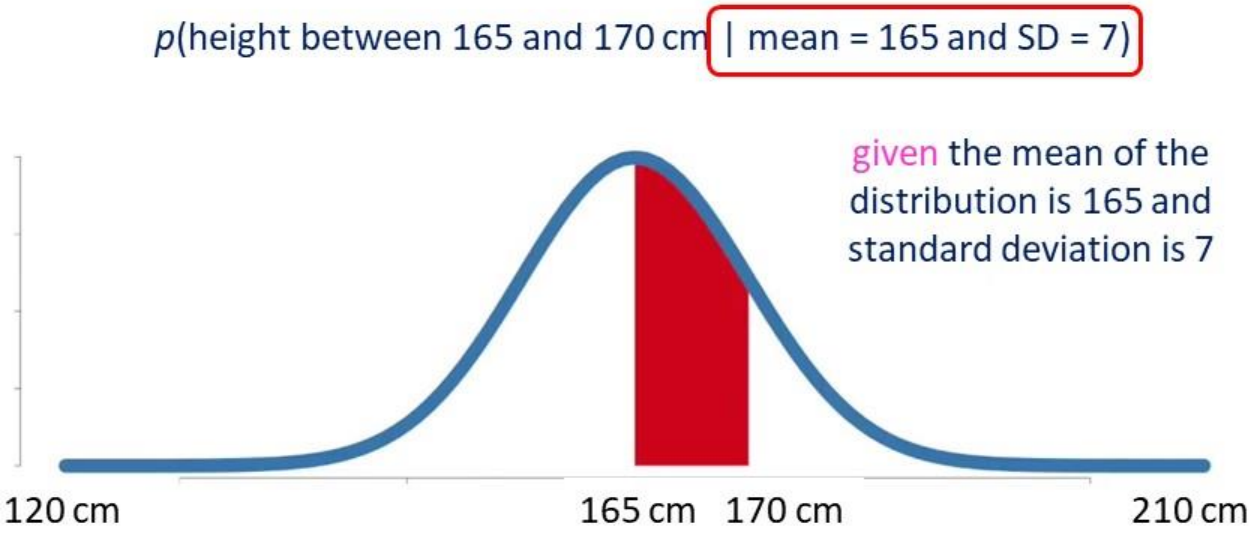
p(height between 165 and 170 cm | mean = 165 and SD = 7)

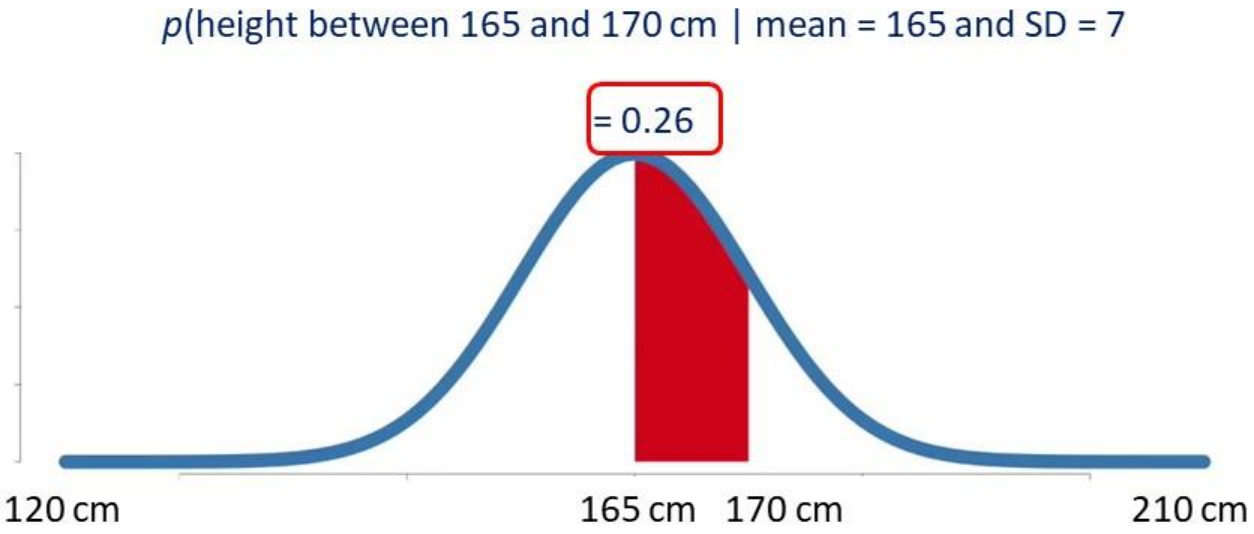


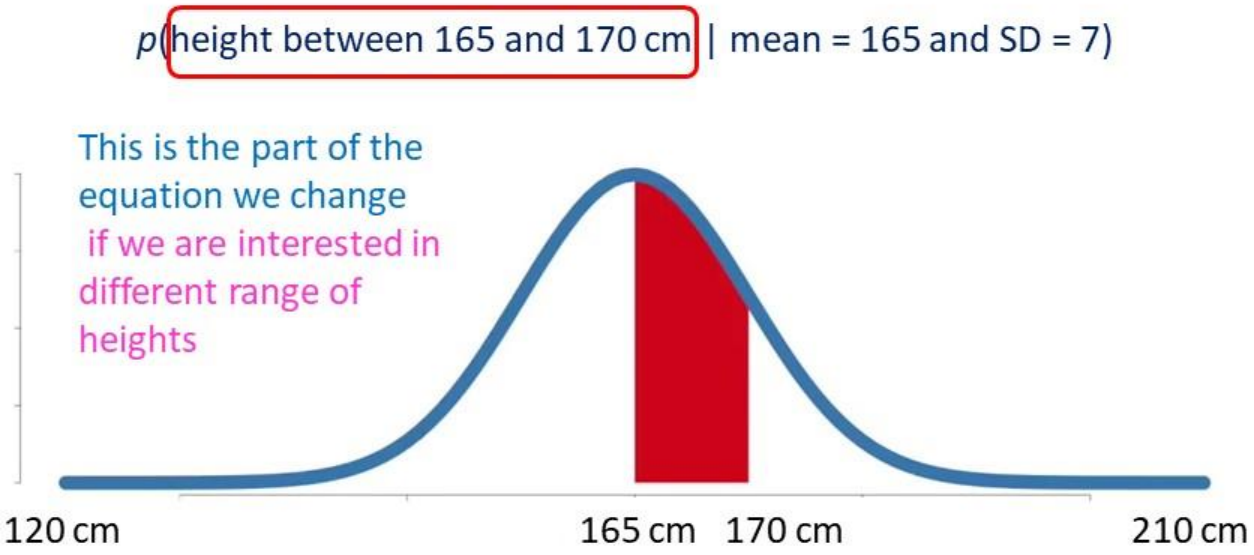
given

p(height between 165 and 170 cm | mean = 165 and SD = 7)

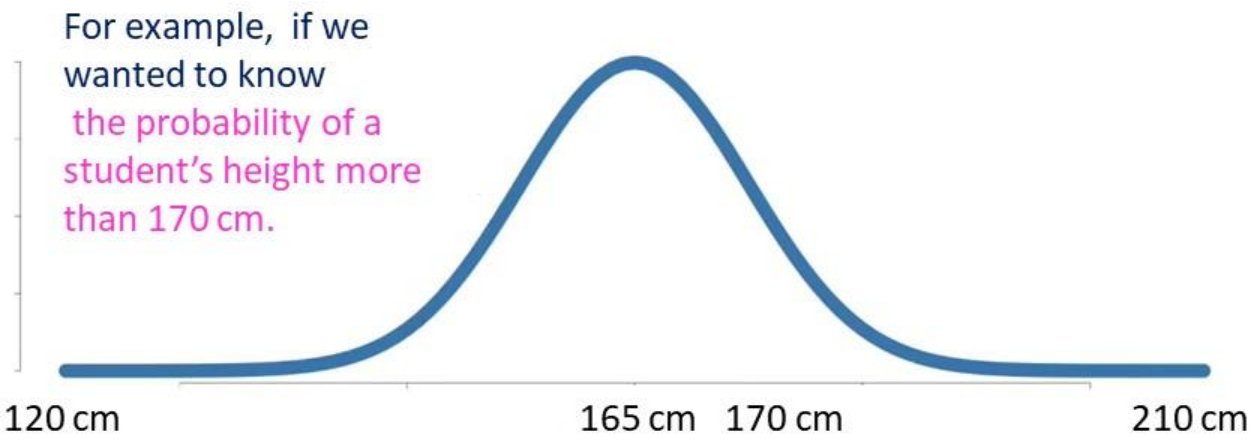



p(height between 165 and 170 cm | mean = 165 and SD = 7)
given the mean of the distribution is 165 and standard deviation is 7
120 cm
165 cm
170 cm
210 cm

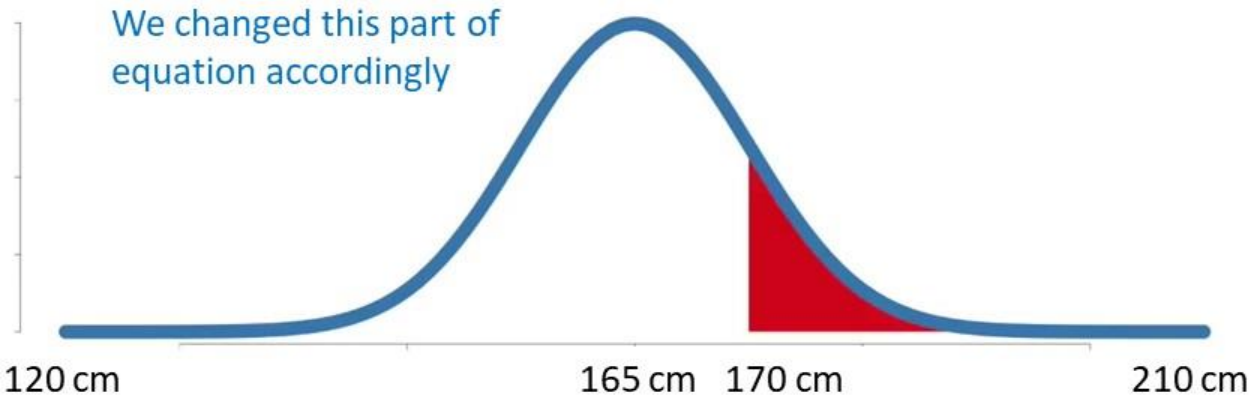

p(height between 165 and 170 cm | mean = 165 and SD = 7
= 0.26
120 cm
165 cm
170 cm
210 cm


p(height between 165 and 170 cm | mean = 165 and SD = 7)
This is the part of the equation we change if we are interested in different range of heights
120 cm
165 cm
170 cm
210 cm

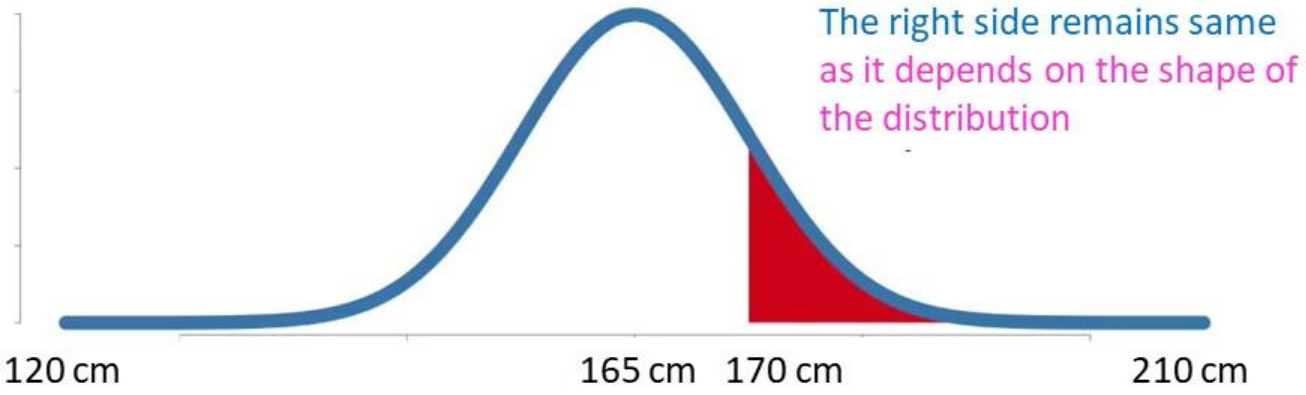
p(height >170 cm | mean = 165 and SD = 7)



p(height >170 cm | mean = 165 and SD = 7)

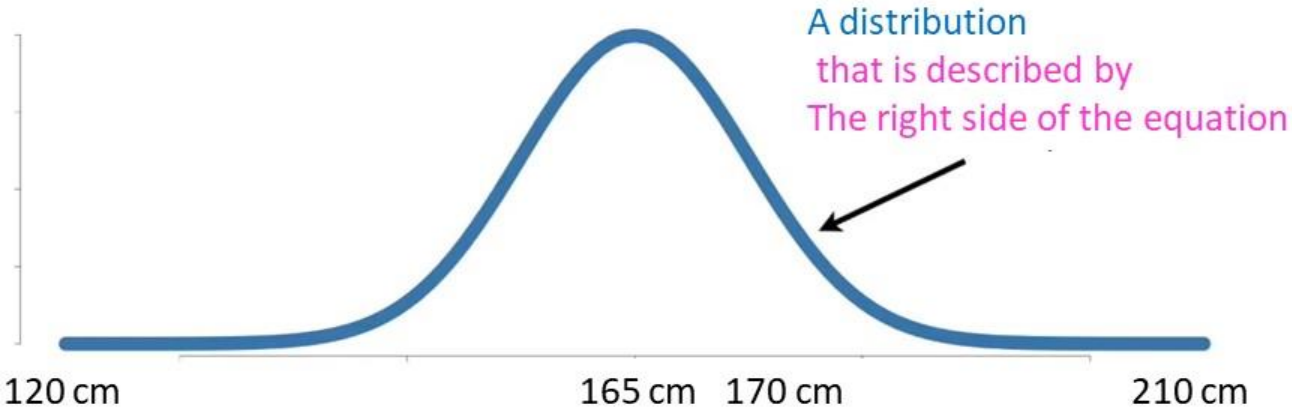


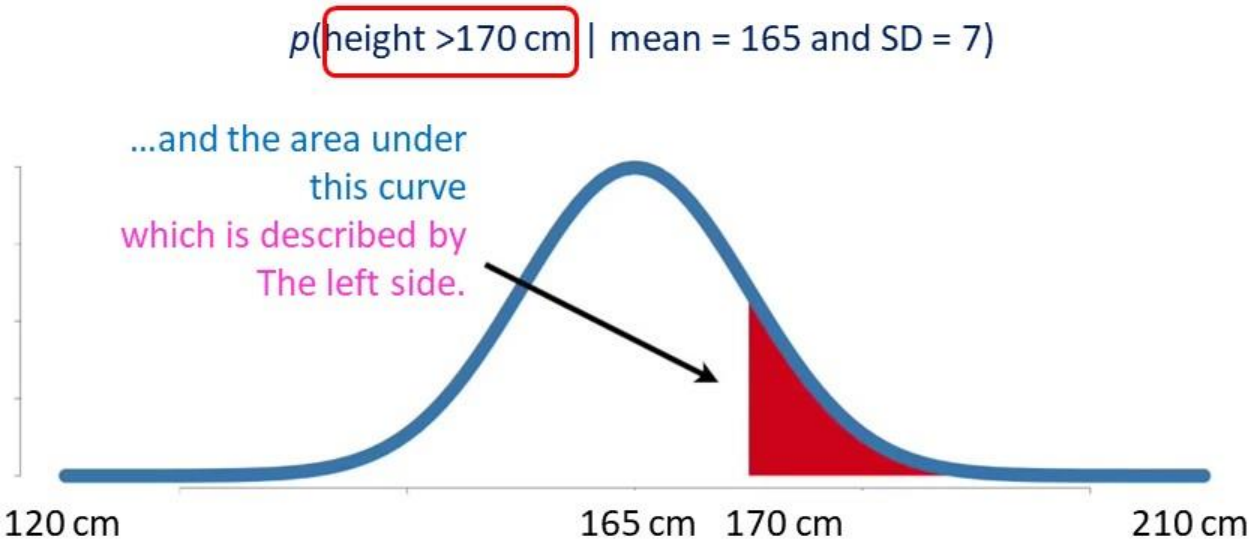
p(height >170 cm | mean = 165 and SD = 7)

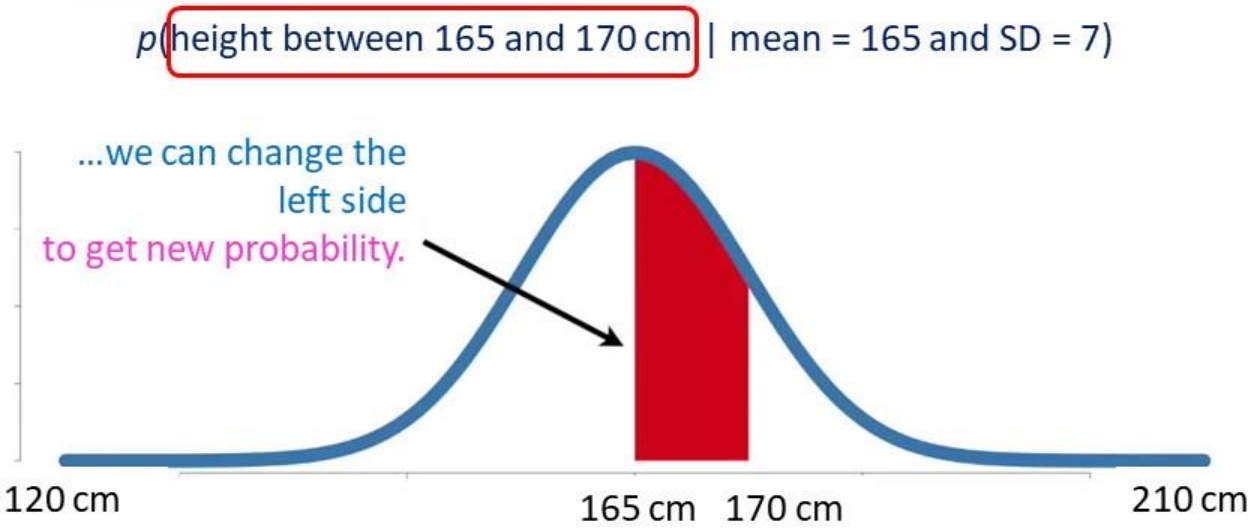


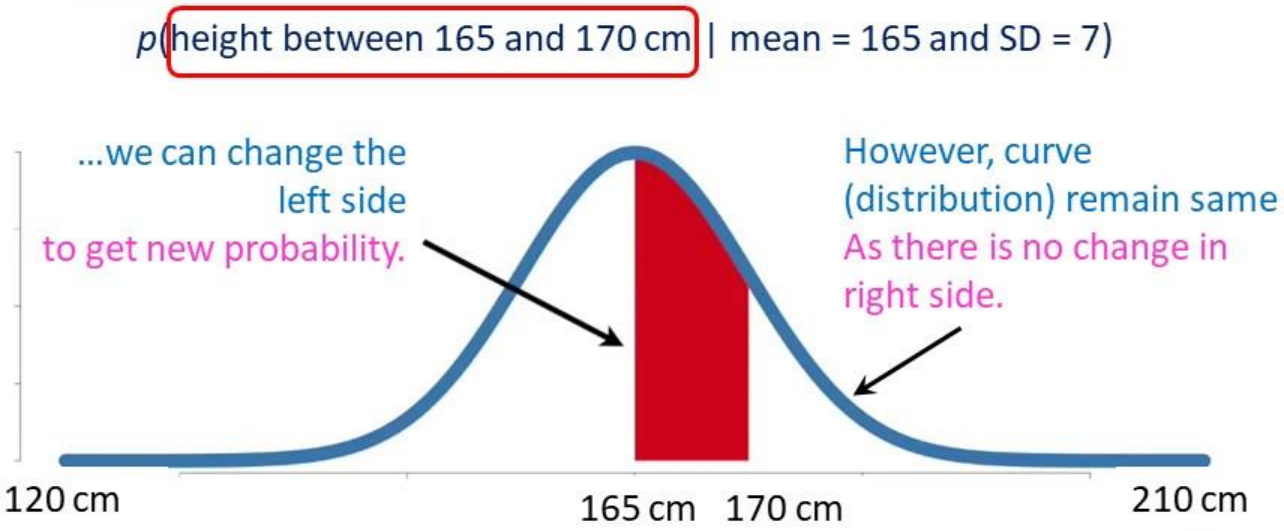
p(height >170 cm | mean = 165 and SD = 7)

p(height >170 cm | mean = 165 and SD = 7)

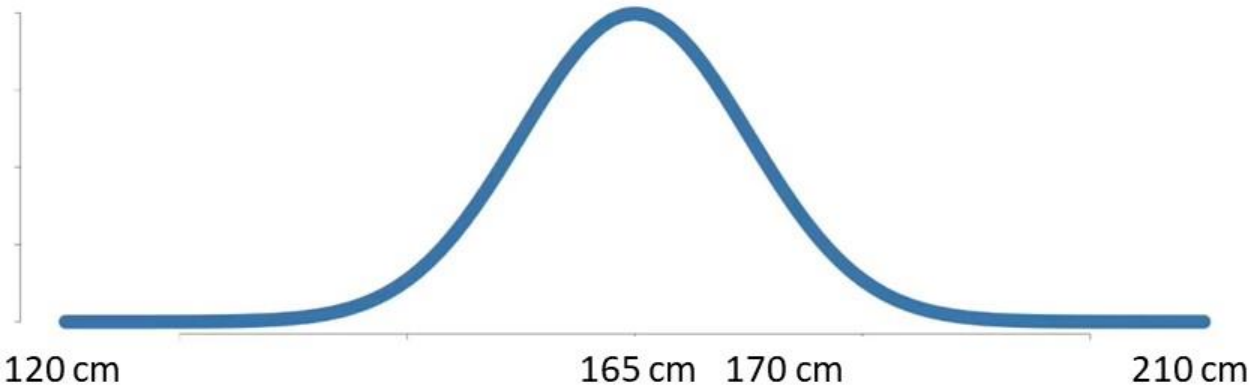


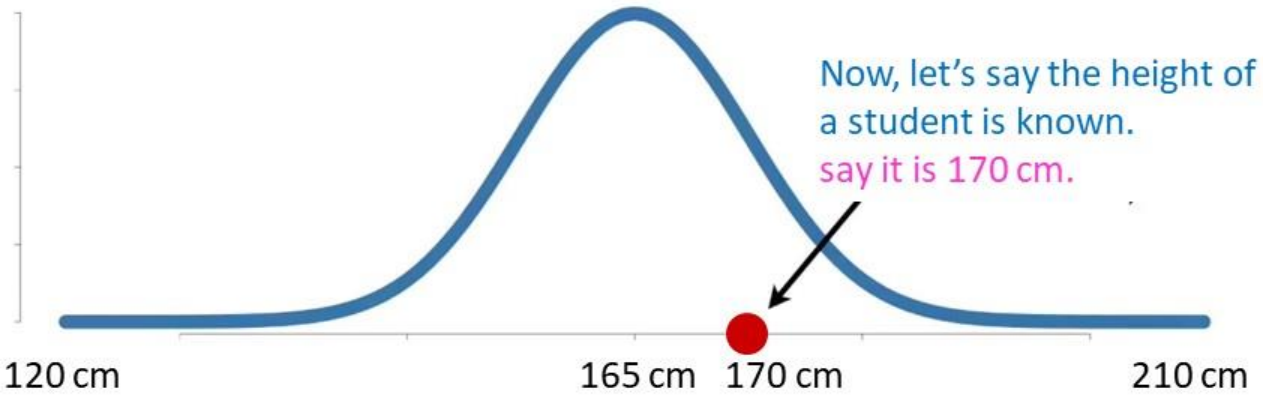

p(height >170 cm | mean = 165 and SD = 7)
...and the area under
this curve
which is described by
The left side.
120 cm
165 cm
170 cm
210 cm


p(height between 165 and 170 cm | mean = 165 and SD = 7)
...we can change the
left side
to get new probability.
120 cm
165 cm
170 cm
210 cm

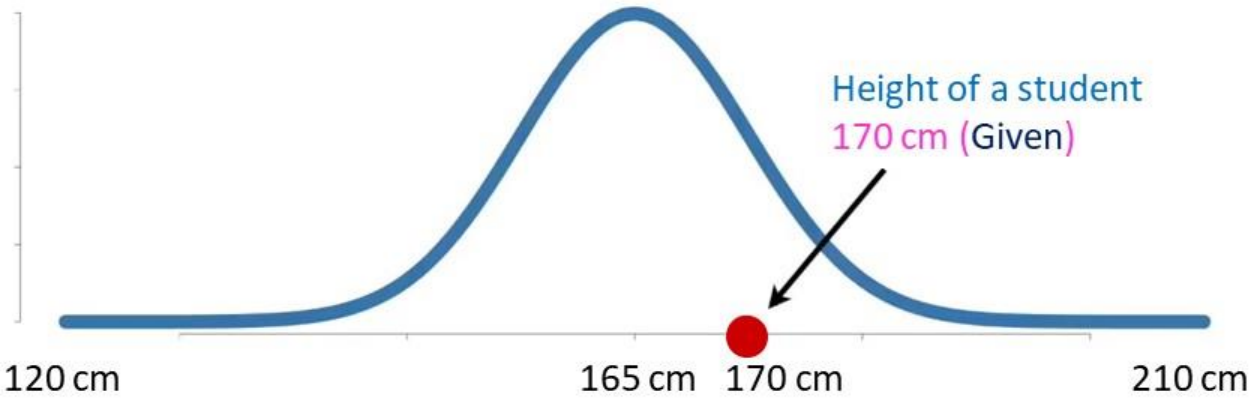

p(height between 165 and 170 cm | mean = 165 and SD = 7)
...we can change the left side to get new probability.
However, curve (distribution) remain same As there is no change in right side.
120 cm
165 cm
170 cm
210 cm

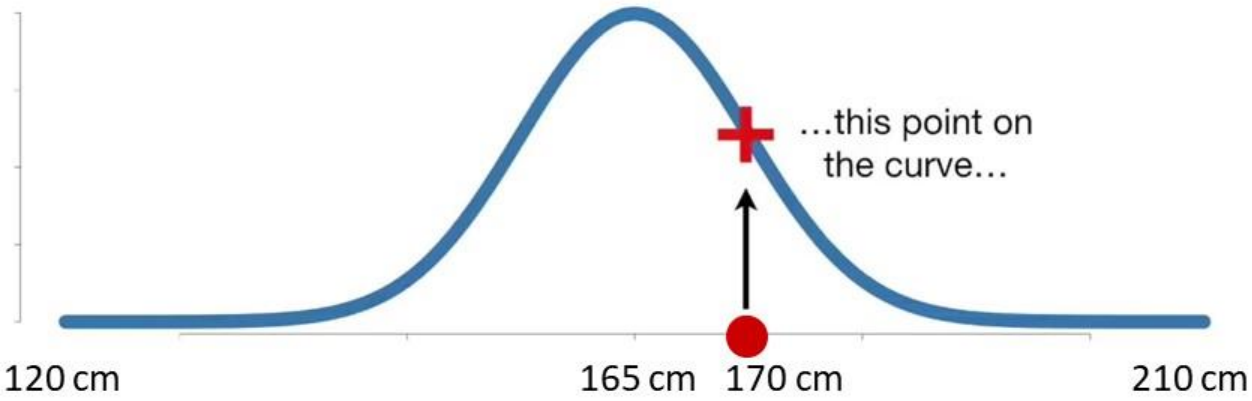
We have discussed about probability,
Let's discuss about
likelihood...



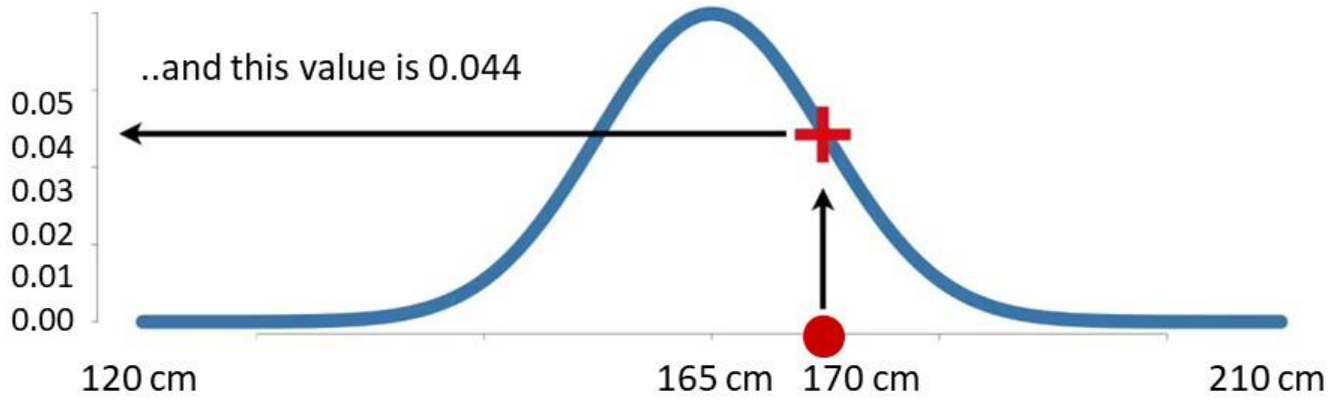

Now, let's say the height of
a student is known.
say it is 170 cm.
120 cm
165 cm
170 cm
210 cm

Now, we want to know
the likelihood of this instance (170 cm)
belonging to a distribution



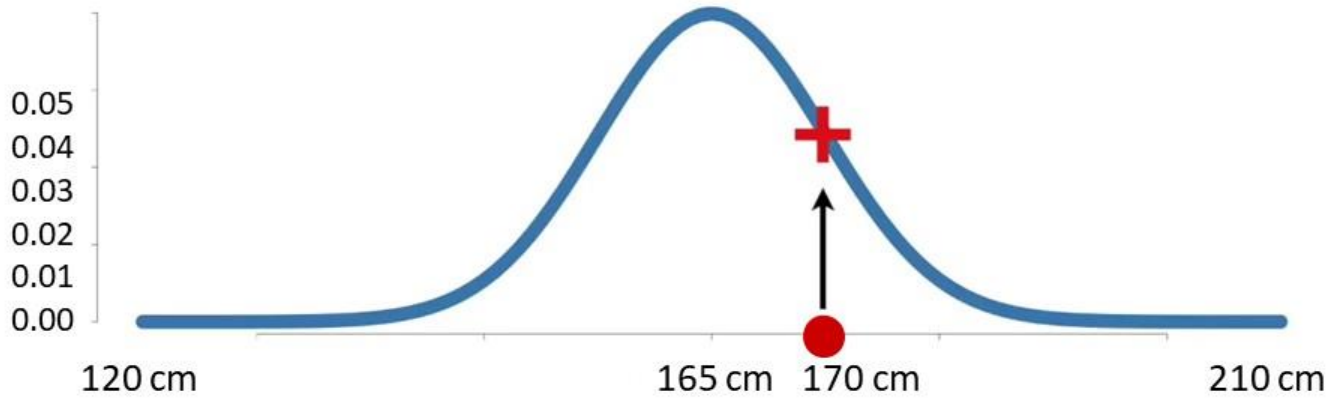

…this point on
the curve…
120 cm
165 cm
170 cm
210 cm

The likelihood of this instance (170 cm)
belonging to a distribution



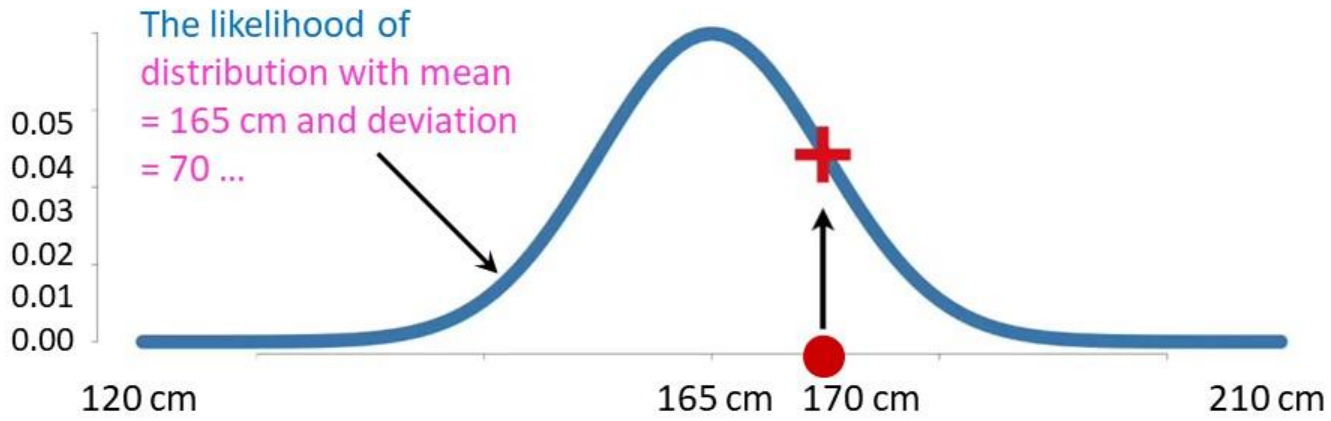
Mathematically expressed as:

L(mean = 165 and standard deviation = 7 | height of student = 170 cm)



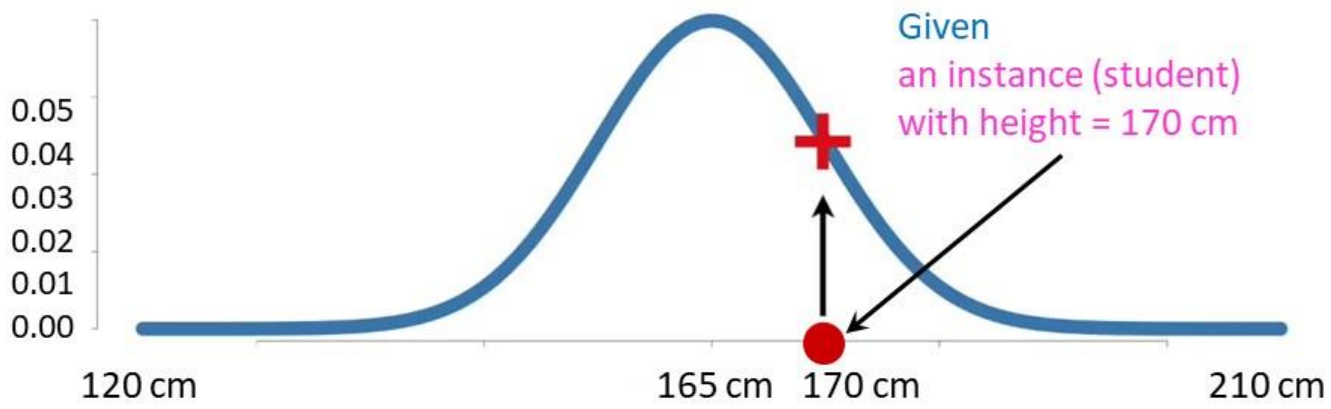
Mathematically expressed as:

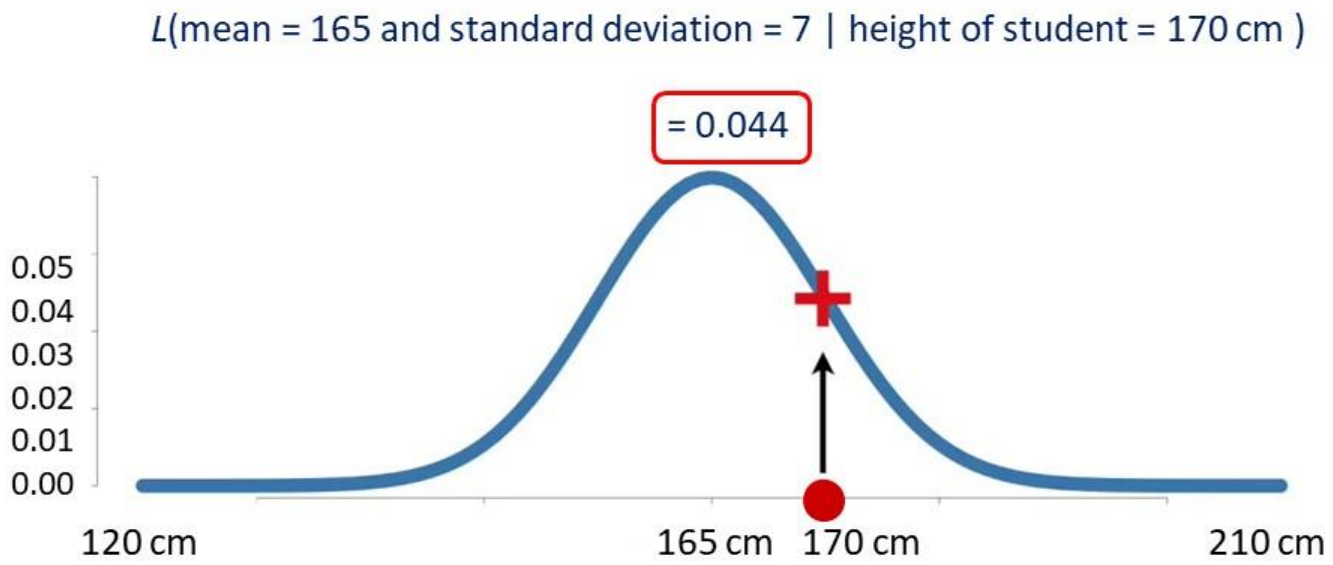
L(mean = 165 and standard deviation = 7 | height of student = 170 cm)

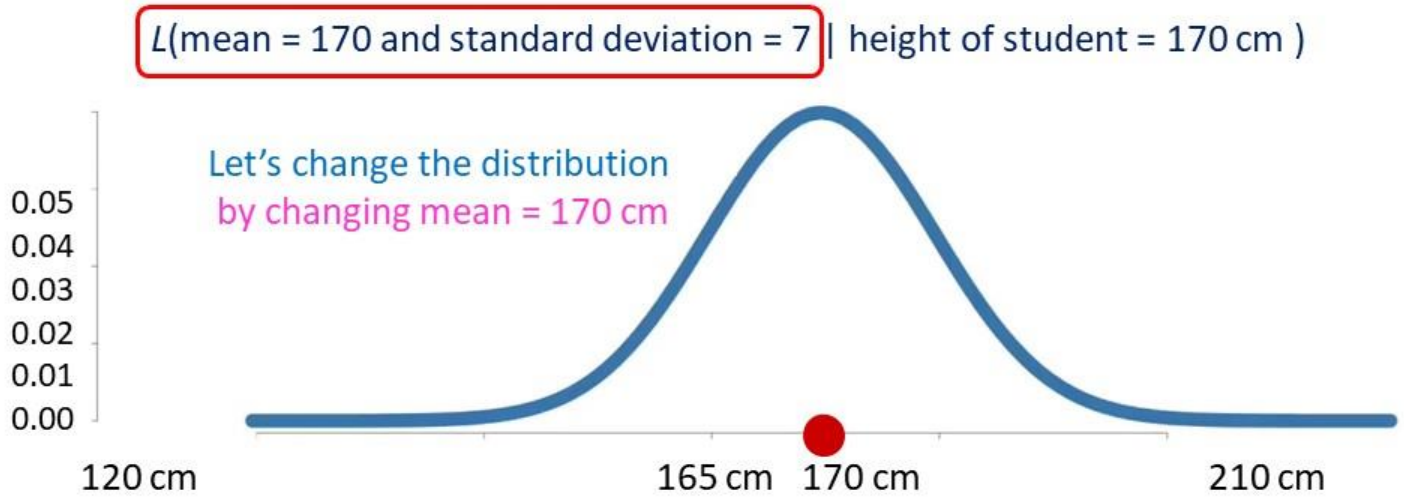


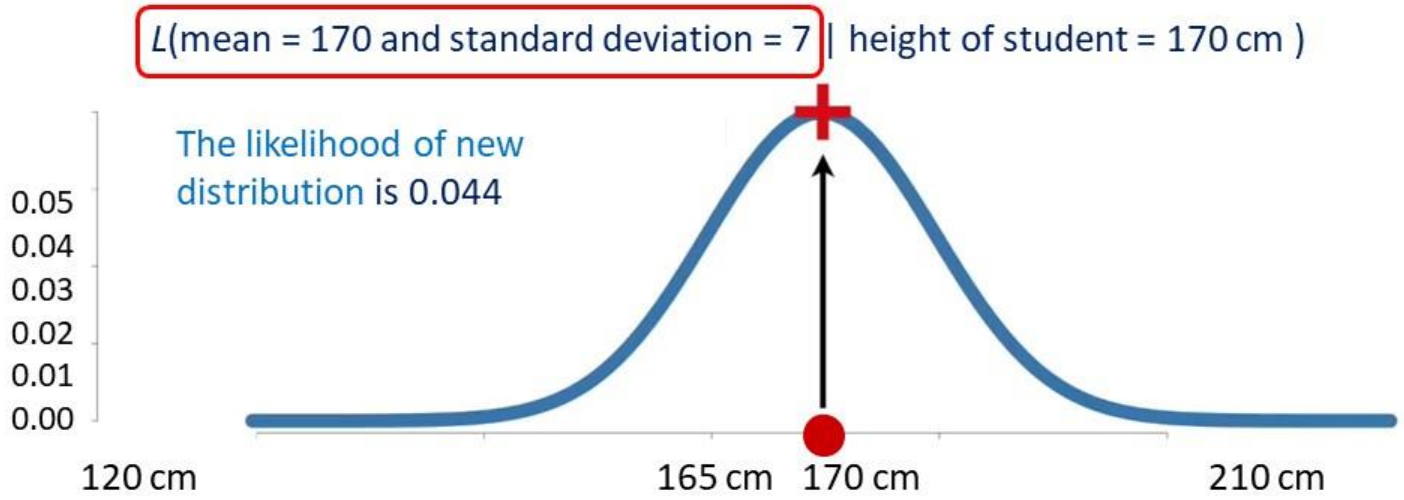
Mathematically expressed as:

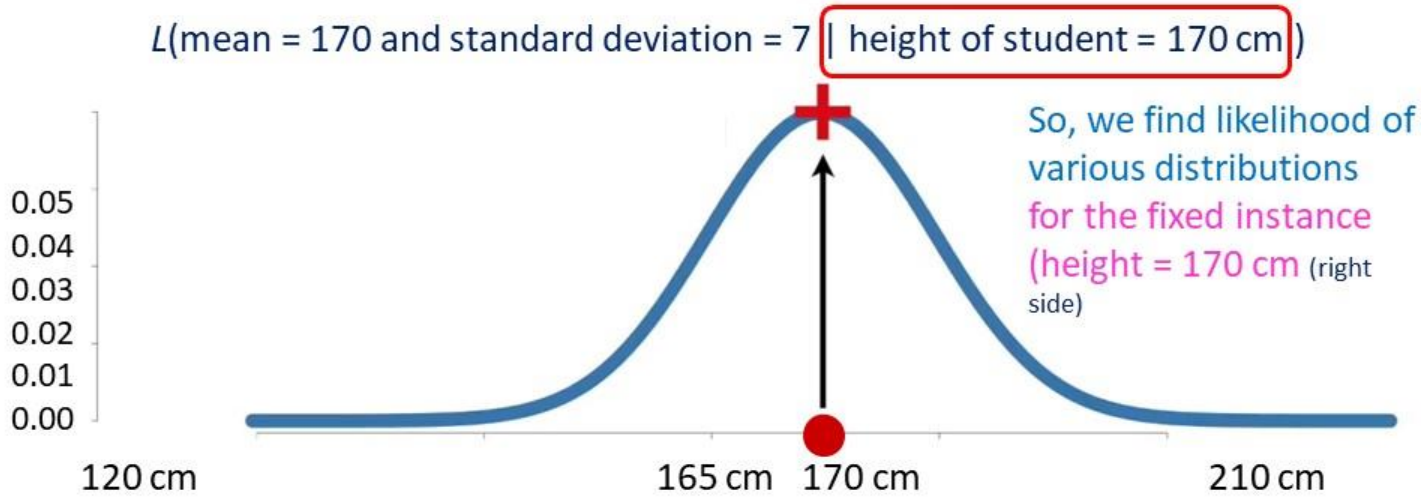
L(mean = 165 and standard deviation = 7 | height of student = 170 cm)

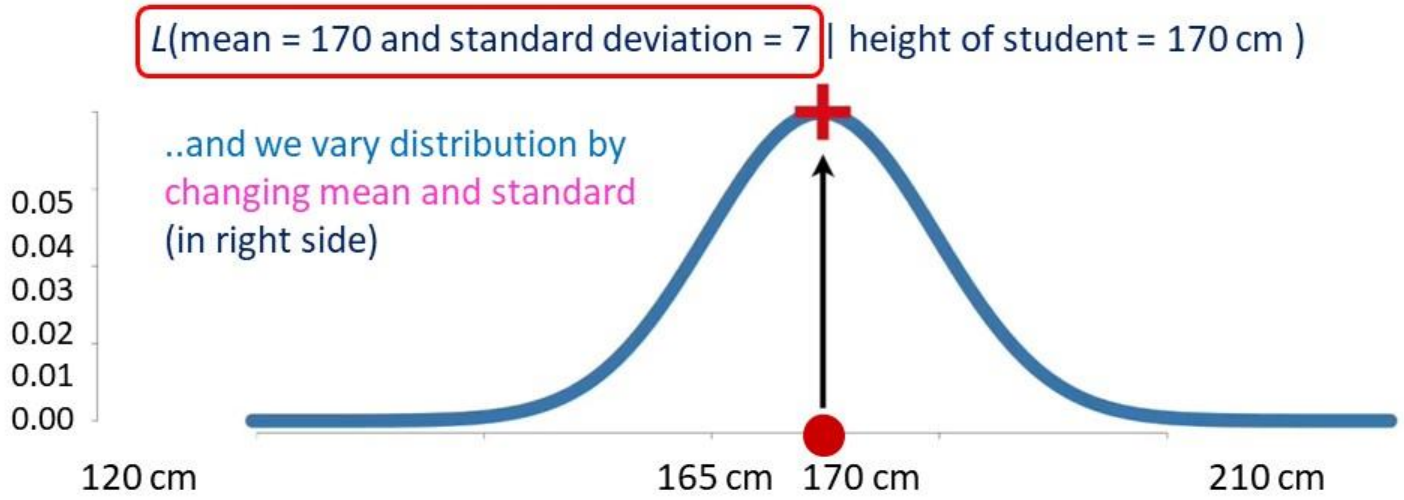


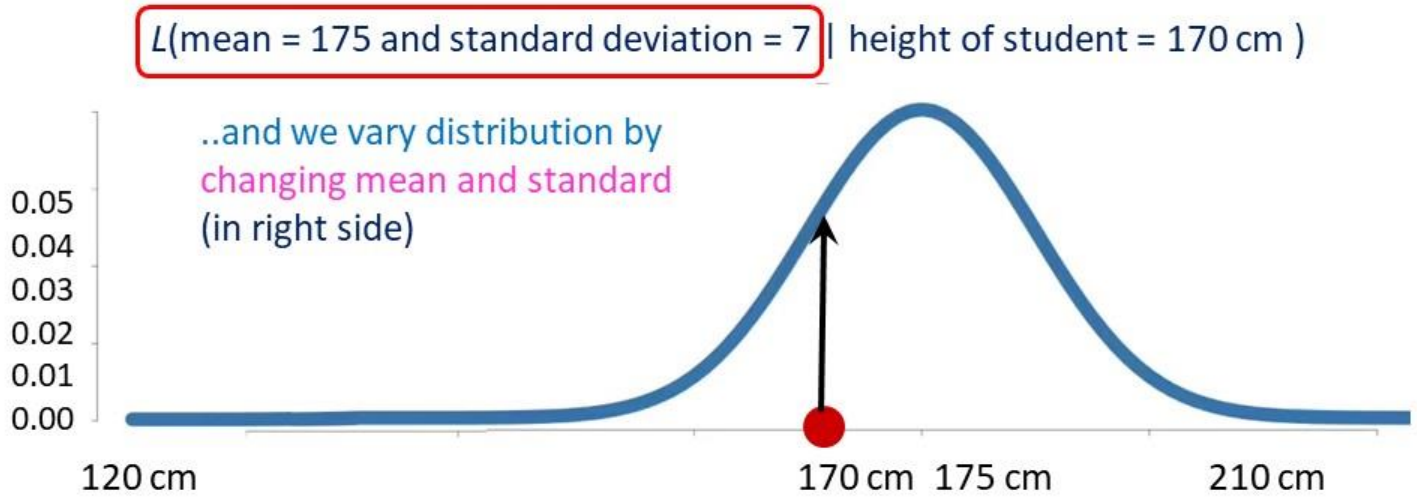

L(mean = 165 and standard deviation = 7 | height of student = 170 cm)
= 0.044
0.05
0.04
0.03
0.02
0.01
0.00
120 cm
165 cm
170 cm
210 cm


L(mean = 170 and standard deviation = 7 | height of student = 170 cm)
Let's change the distribution
by changing mean = 170 cm
0.05
0.04
0.03
0.02
0.01
0.00
120 cm
165 cm
170 cm
210 cm

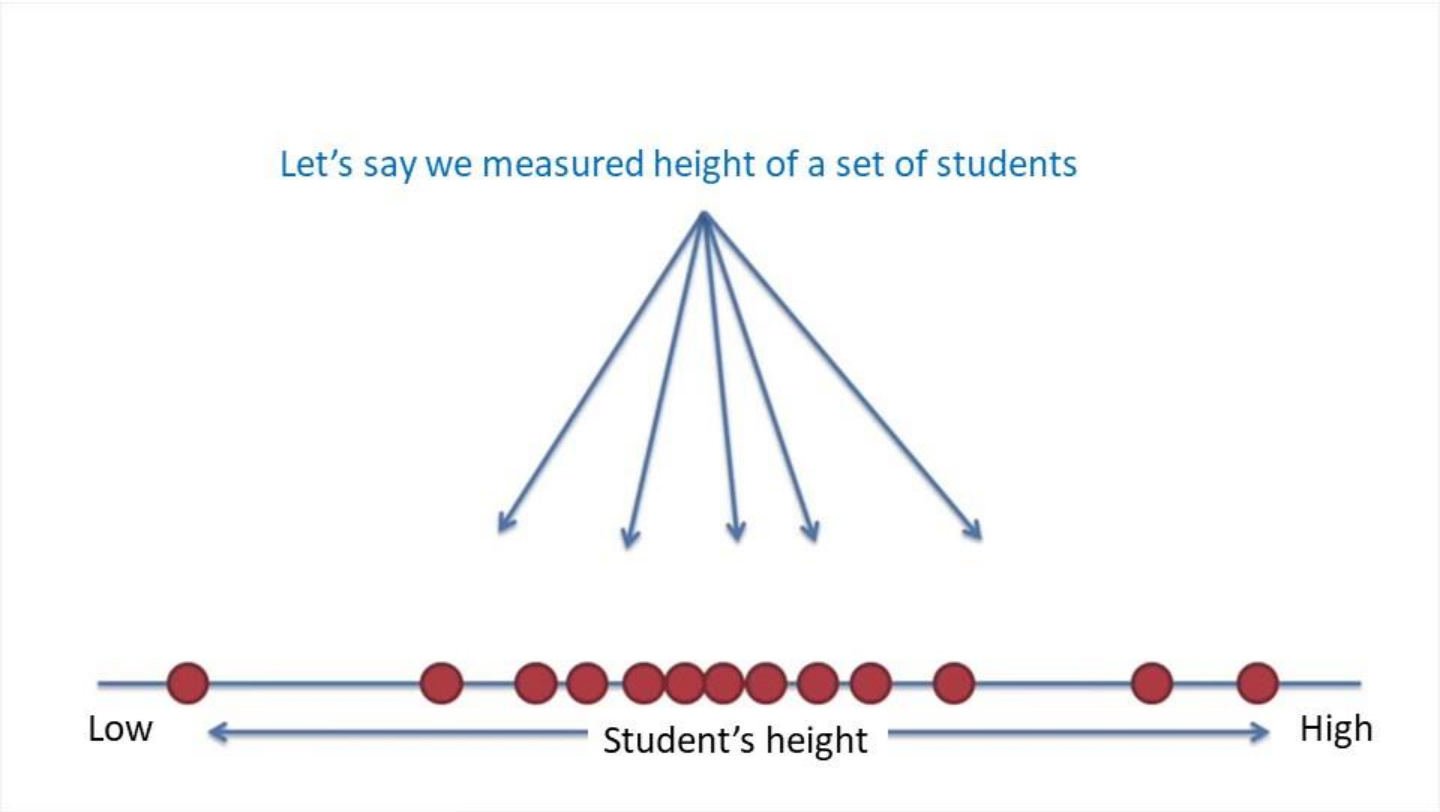

L(mean = 170 and standard deviation = 7 | height of student = 170 cm)
The likelihood of new distribution is 0.044
0.05
0.04
0.03
0.02
0.01
0.00
120 cm
165 cm
170 cm
210 cm


L(mean = 170 and standard deviation = 7 | height of student = 170 cm)
So, we find likelihood of various distributions for the fixed instance (height = 170 cm (right side)
0.05
0.04
0.03
0.02
0.01
0.00
120 cm
165 cm
170 cm
210 cm


L(mean = 170 and standard deviation = 7 | height of student = 170 cm)
..and we vary distribution by
changing mean and standard
(in right side)
0.05
0.04
0.03
0.02
0.01
0.00
120 cm
165 cm
170 cm
210 cm


L(mean = 175 and standard deviation = 7 | height of student = 170 cm)
..and we vary distribution by
changing mean and standard
(in right side)
0.05
0.04
0.03
0.02
0.01
0.00
120 cm
170 cm
175 cm
210 cm

Maximum Likelihood Estimation

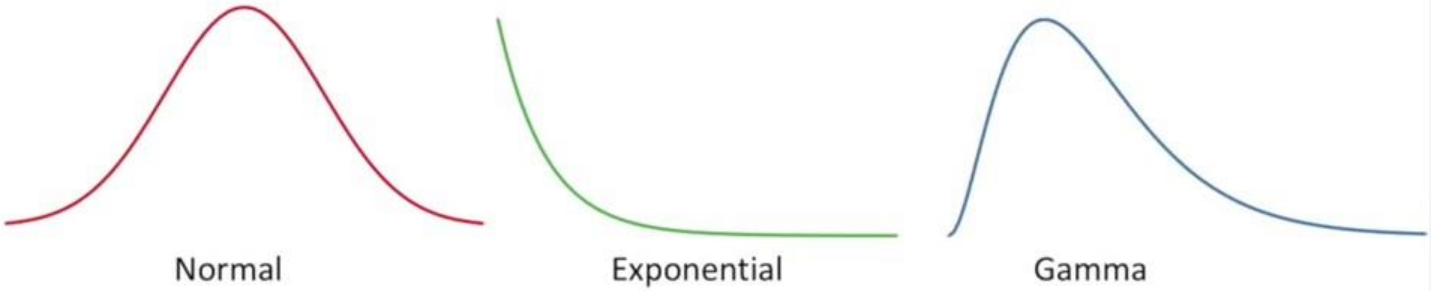

Let’s say we measured height of a set of students
Low
Student’s height
High

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

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

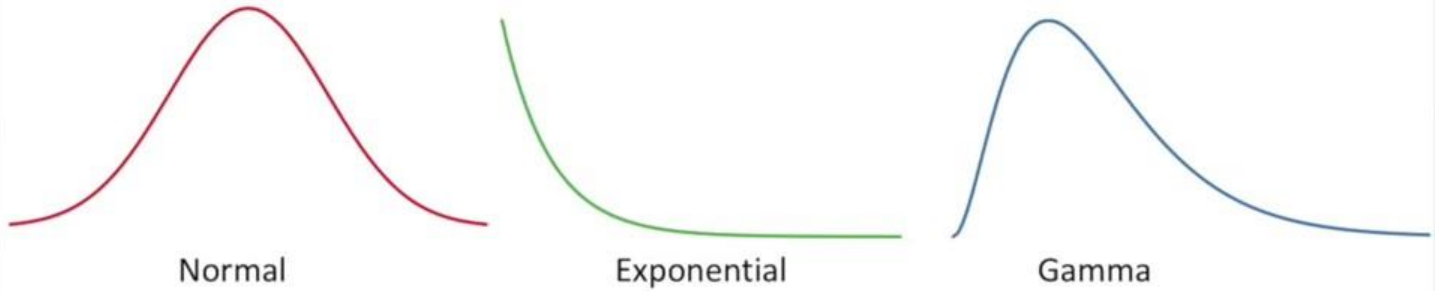


The reason we want to fit a distribution to the data is

Distribution is much more than data

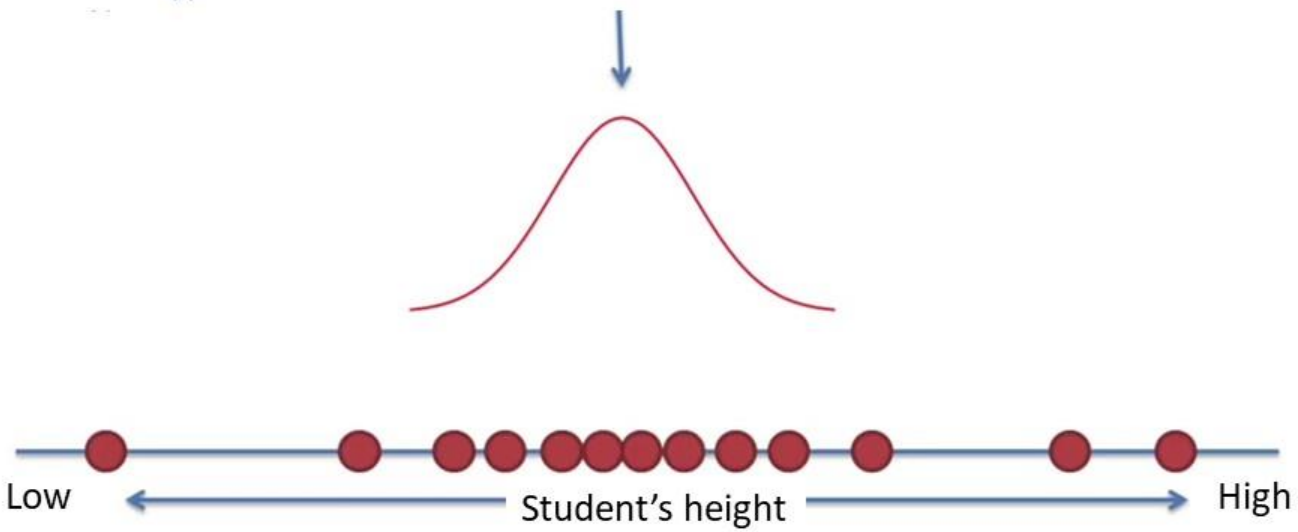
As distribution represent whole pollution,

while data is a sample of the population



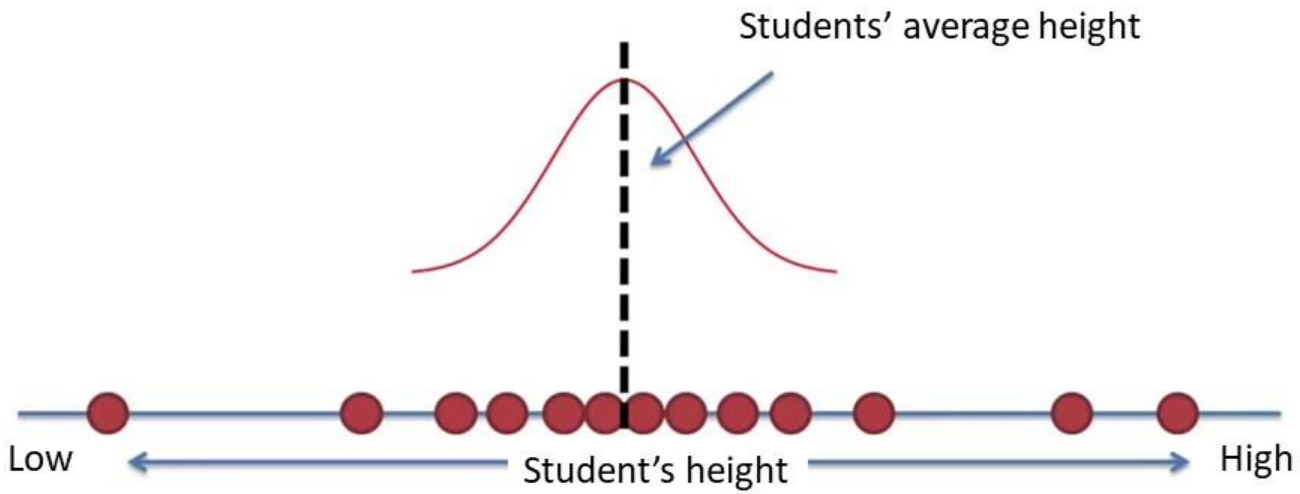
In this case, we may assume that heights are
normally distributed

That means, it has came from this type of distribution



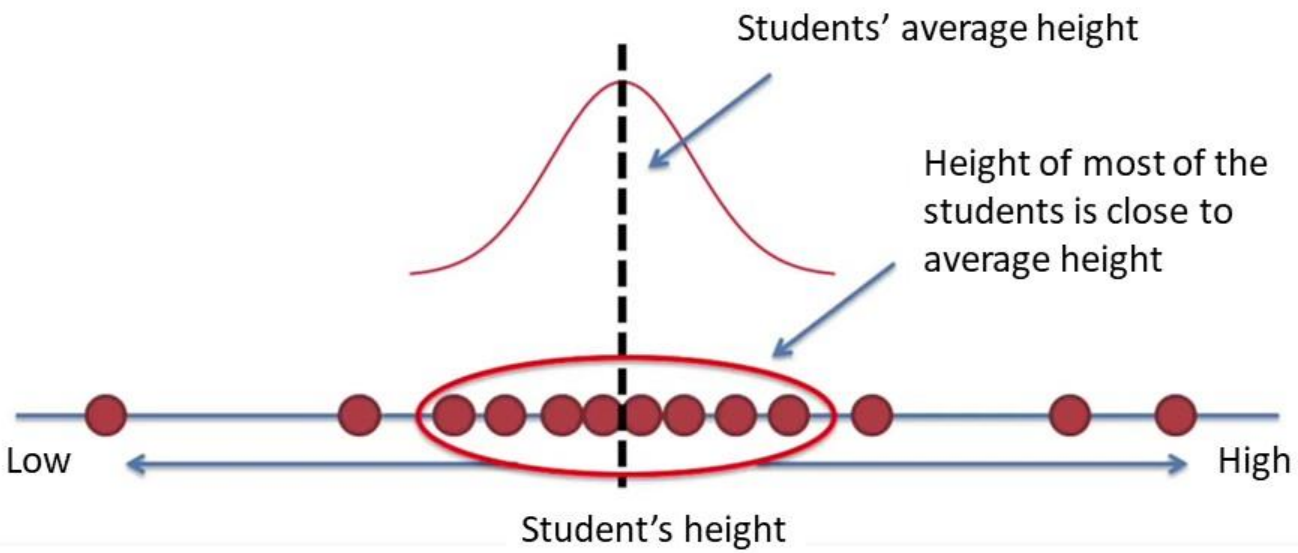
"Normally distributed" means:

1) We expect most of the students have height close to mean (average)



"Normally distributed" means:

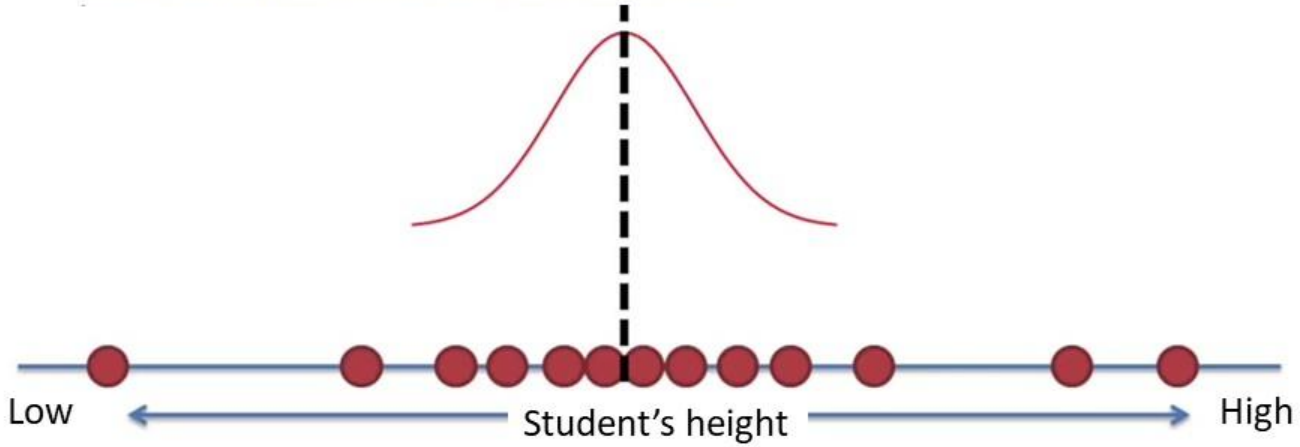
1) We expect most of the students have height close to mean (average)



"Normally distributed" means:

1) We expect most of the students have height (measurements) close to mean (average)

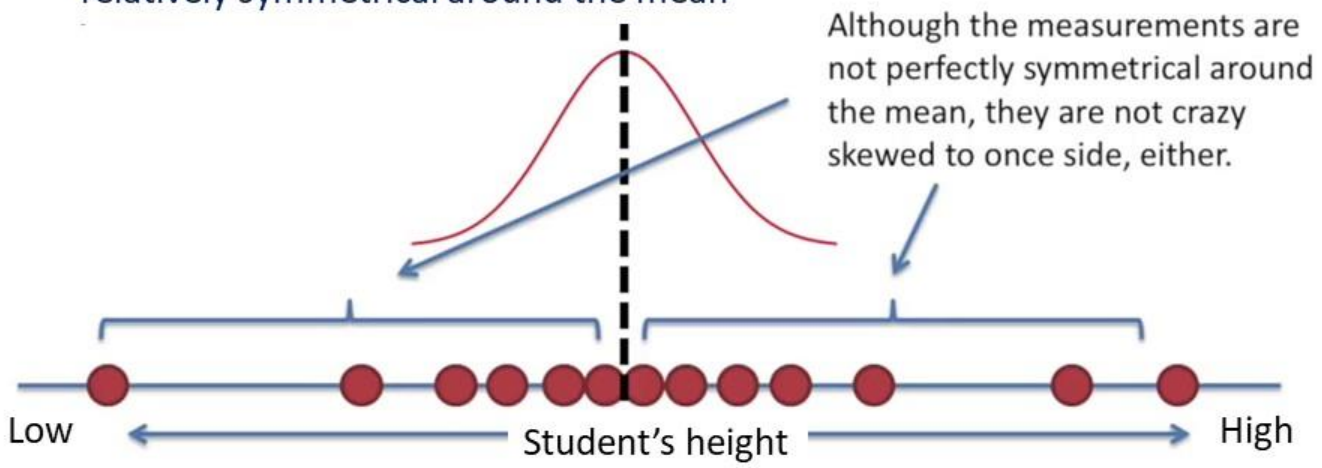
2) We expect the measurements to be relatively symmetrical around the mean



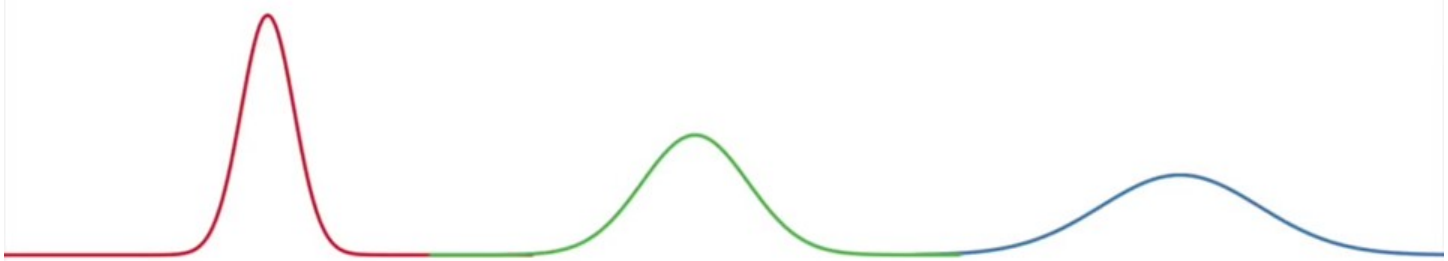
"Normally distributed" means:

1) We expect most of the students have height (measurements) close to mean (average)

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

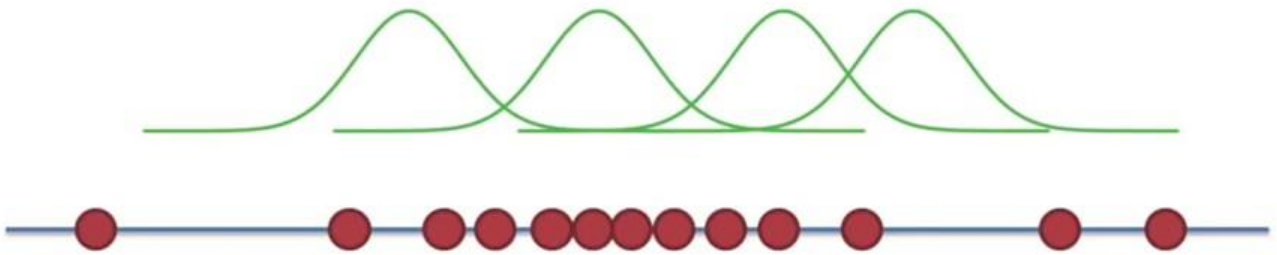


Normal distribution may have different kinds of shapes and sizes...

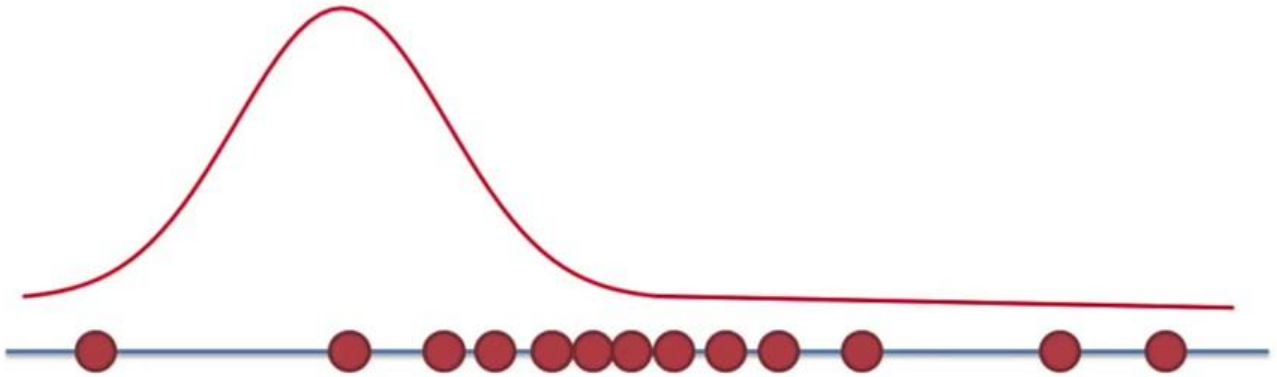


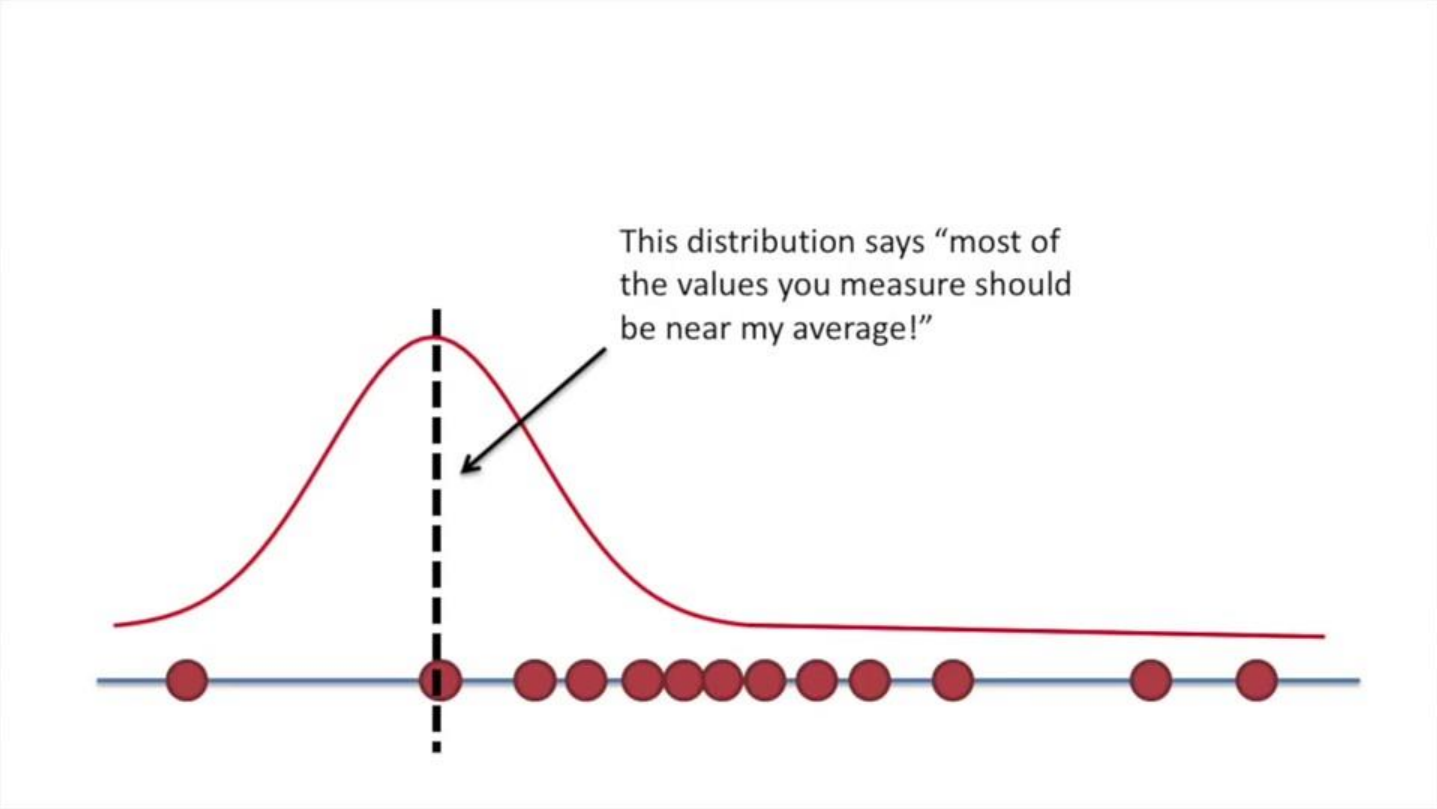
Once we decide the shape of the distribution, we need to figure out its center (position)In this case, we may assume that heights are

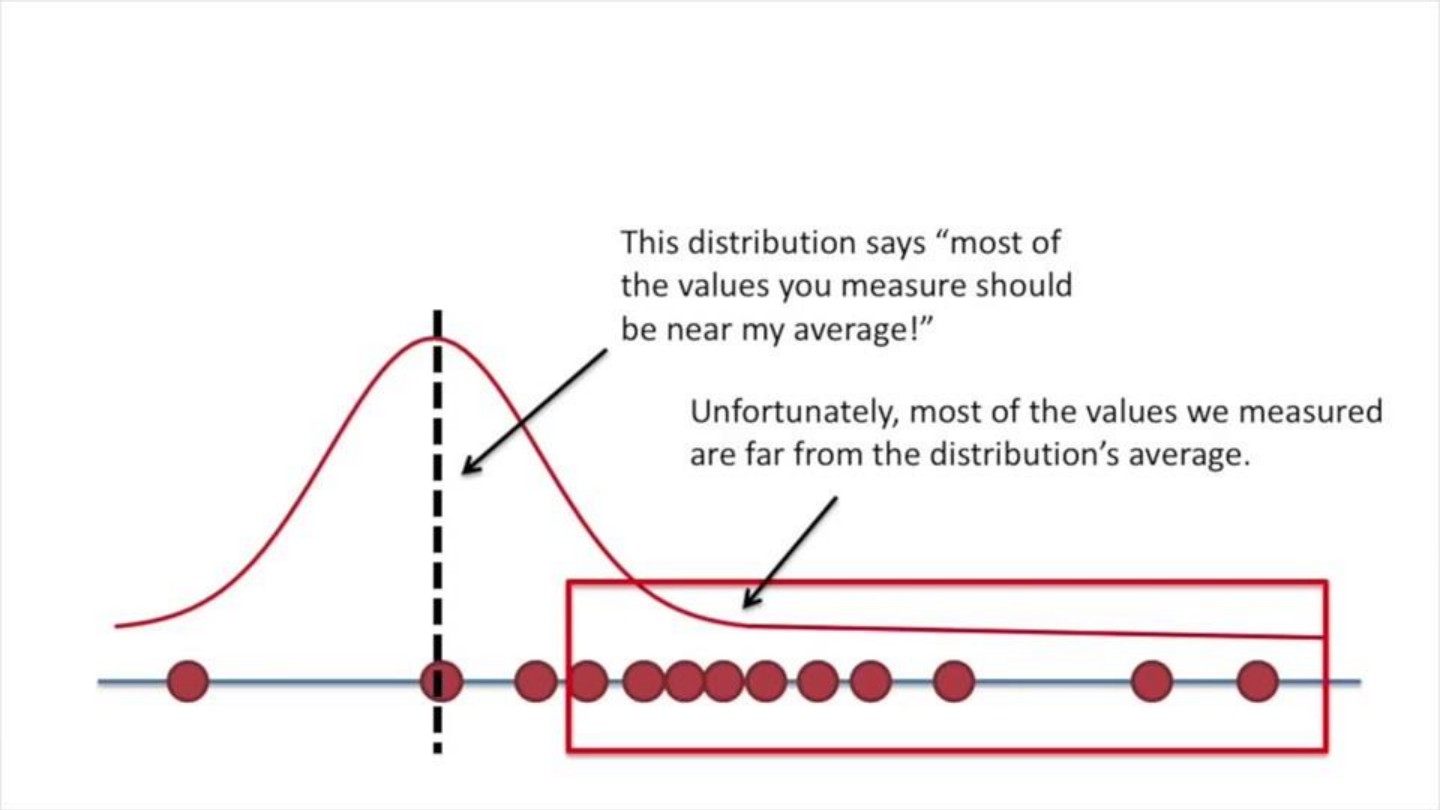
Is one location is “better” than another?

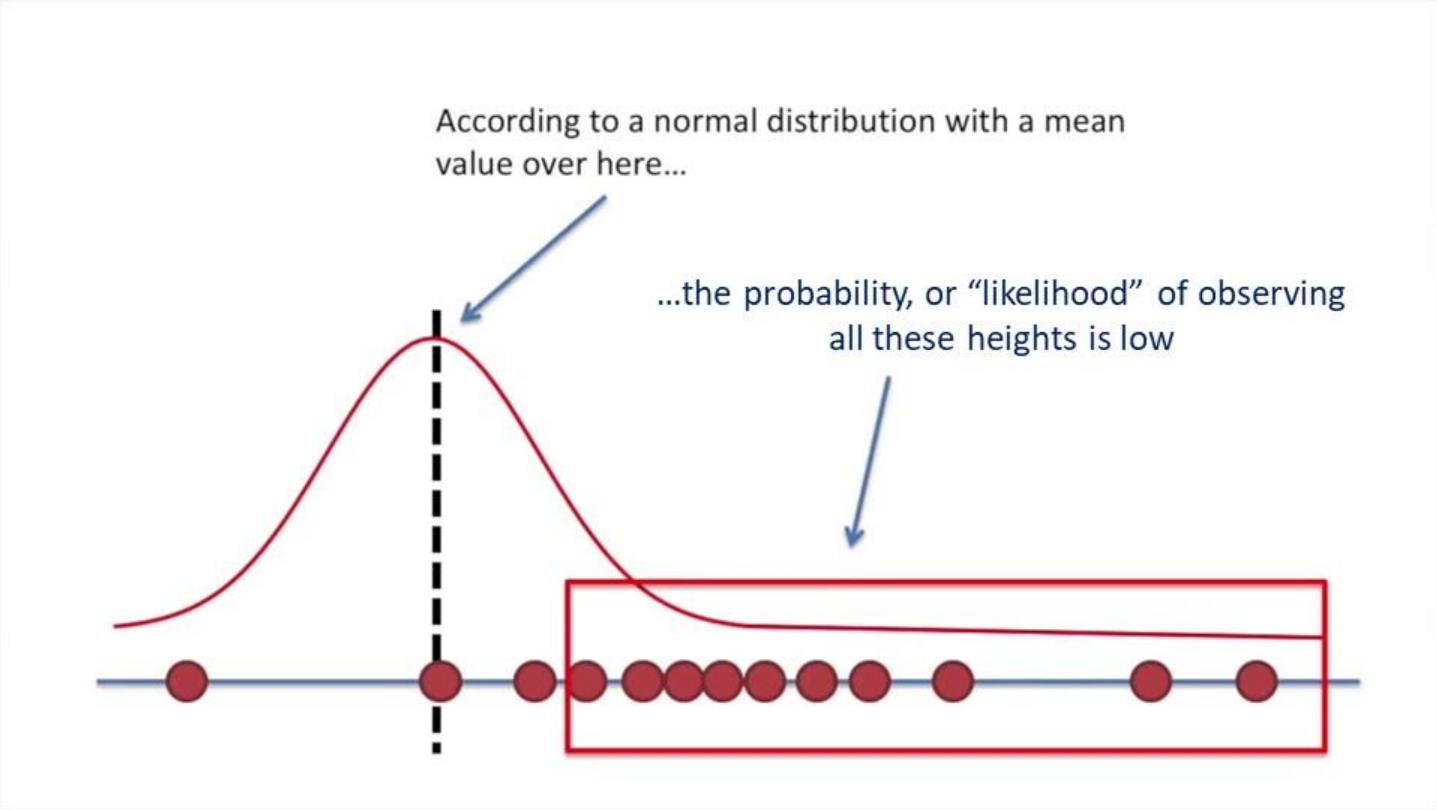


Before we get too technical, lets just pick any old normal distribution and see how well it fits the data.

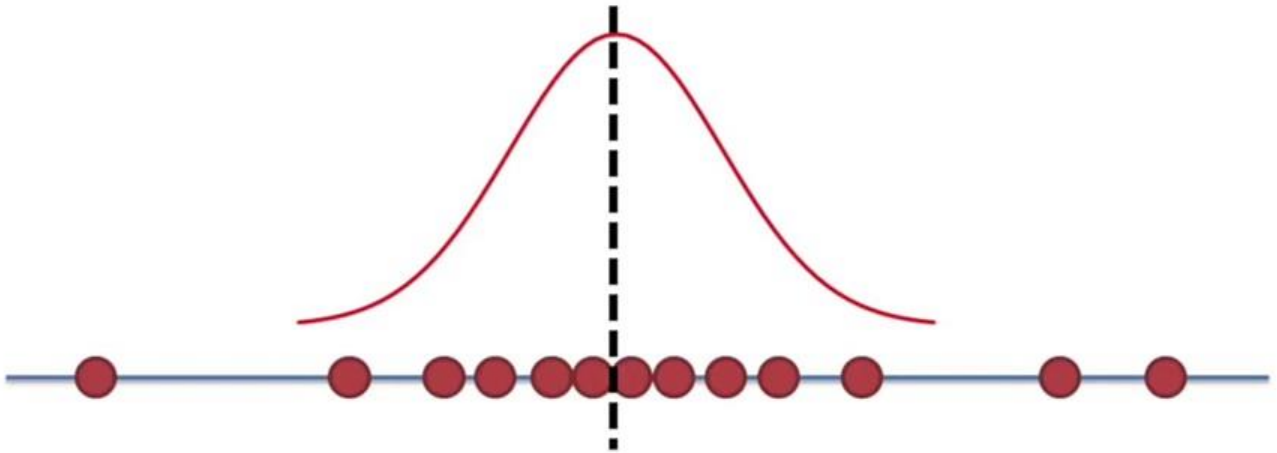


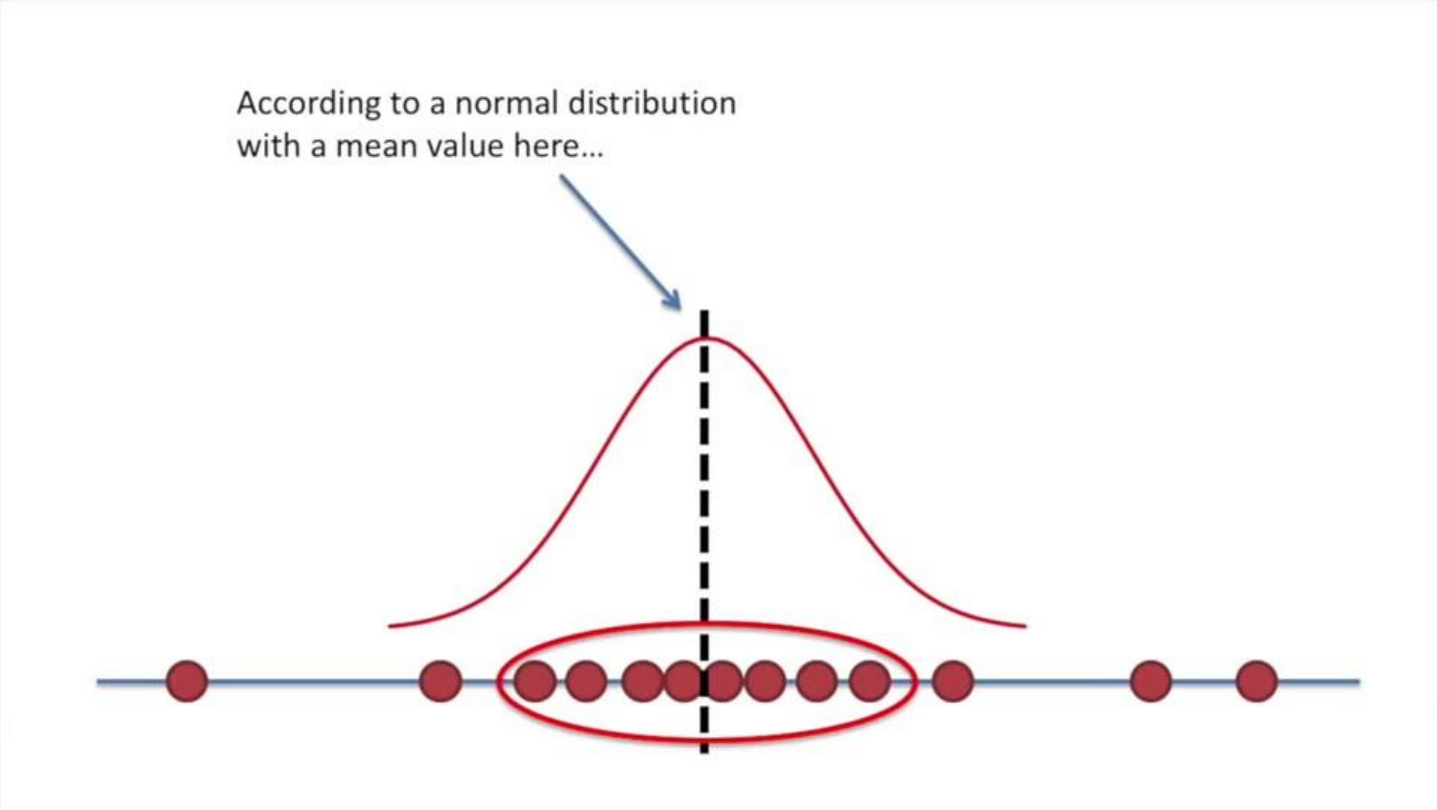

This distribution says “most of the values you measure should be near my average!”

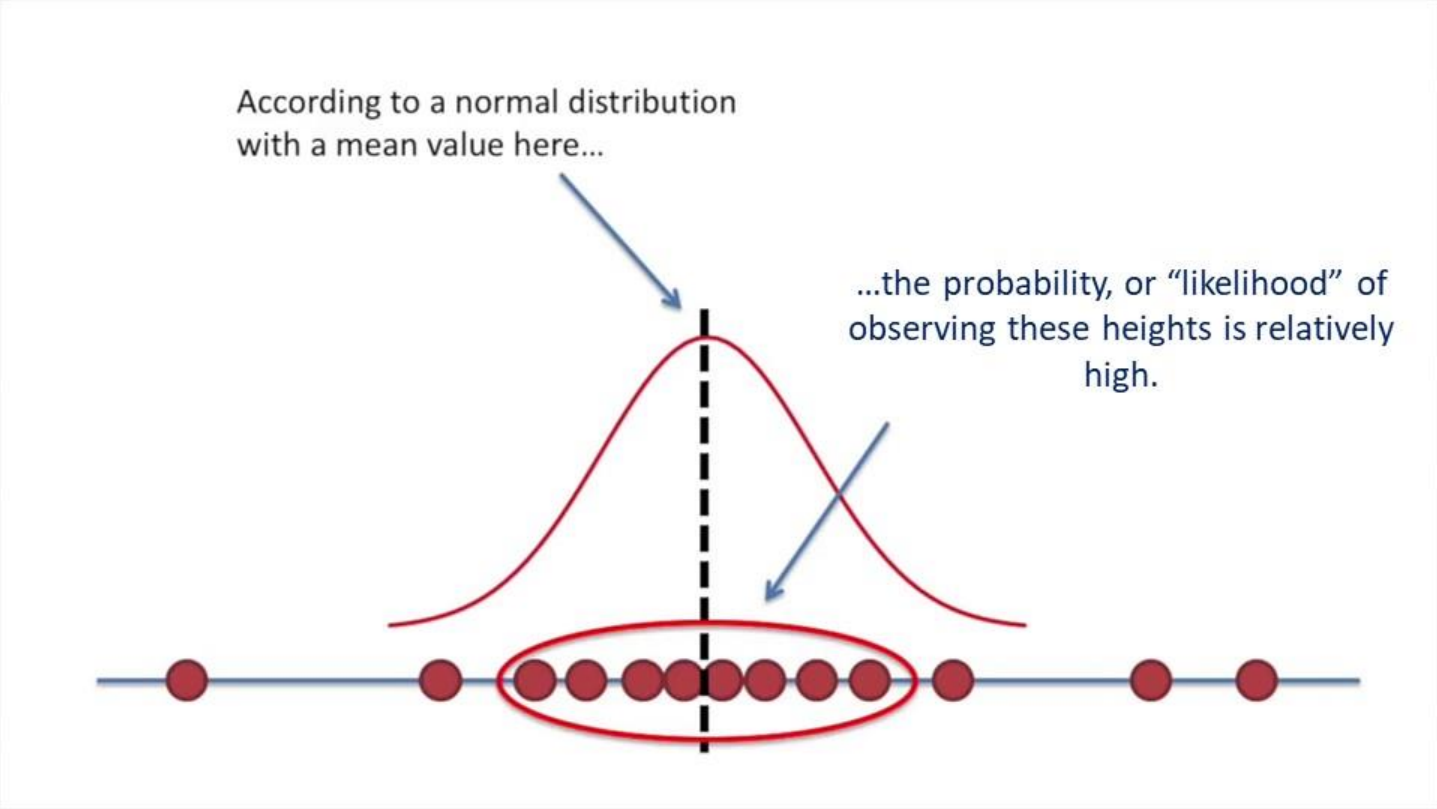

This distribution says “most of the values you measure should be near my average!”
Unfortunately, most of the values we measured are far from the distribution’s average.

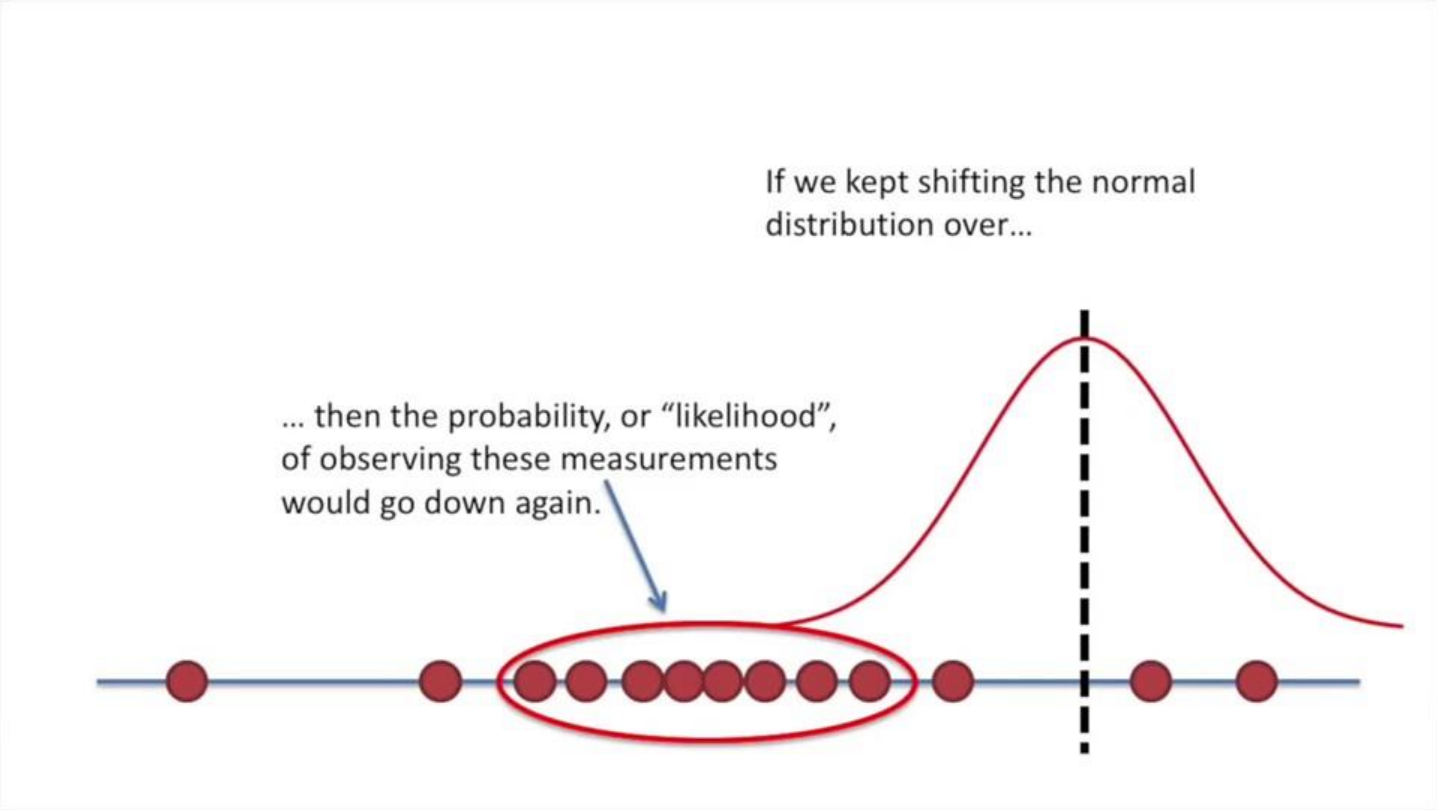

According to a normal distribution with a mean value over here...
...the probability, or "likelihood" of observing all these heights is low

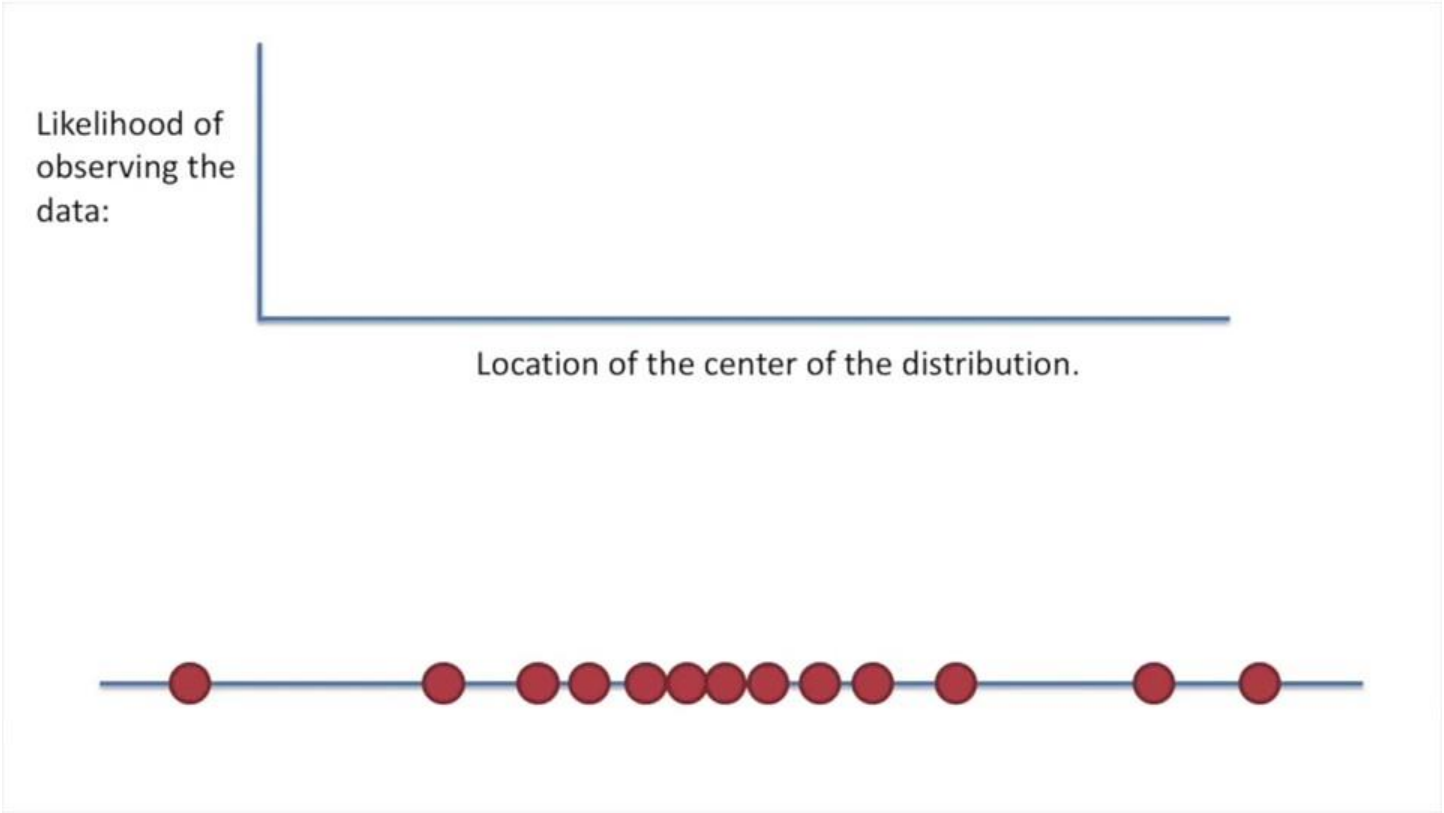
What if we shift the normal distribution over, so that its mean was the same as the average weight?

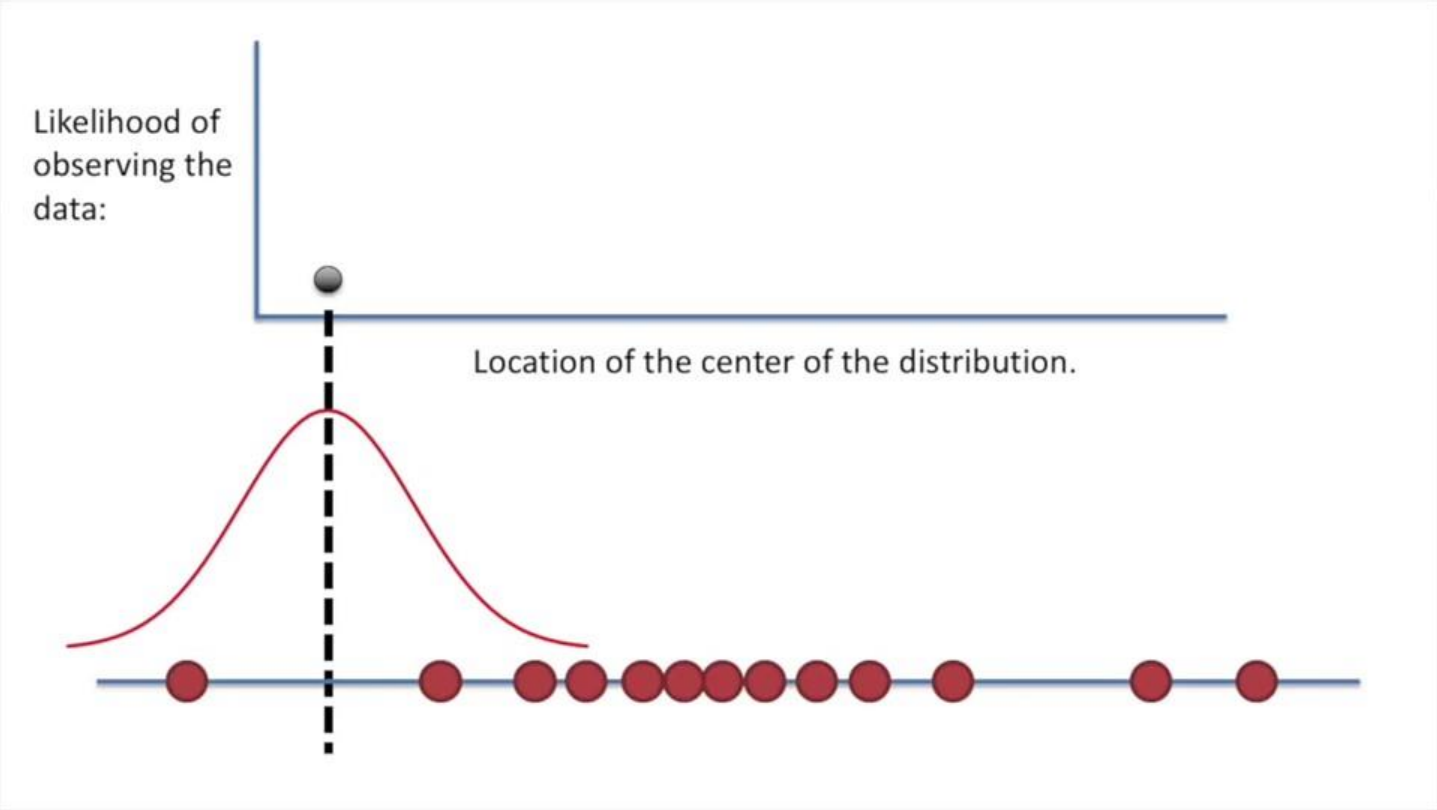


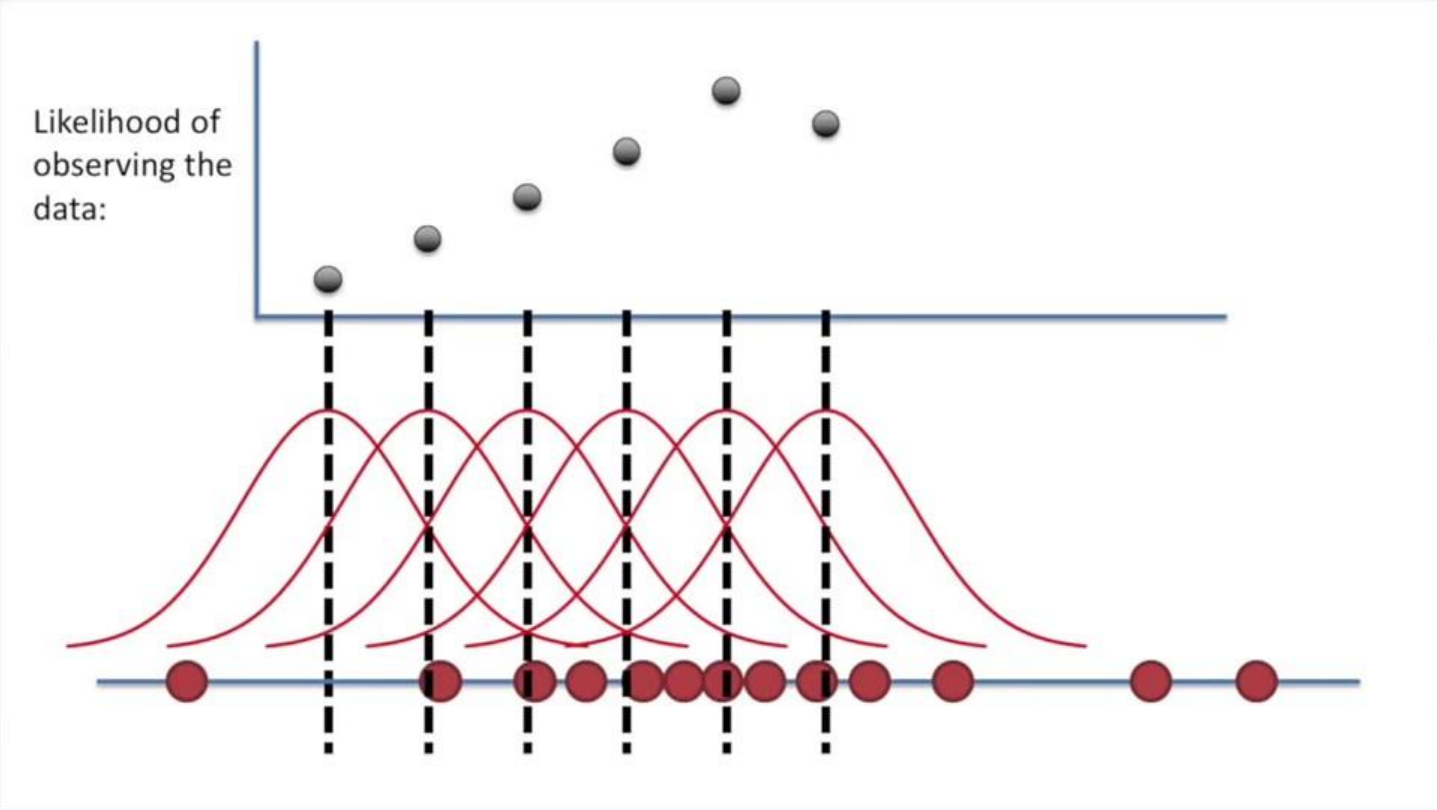

According to a normal distribution
with a mean value here…

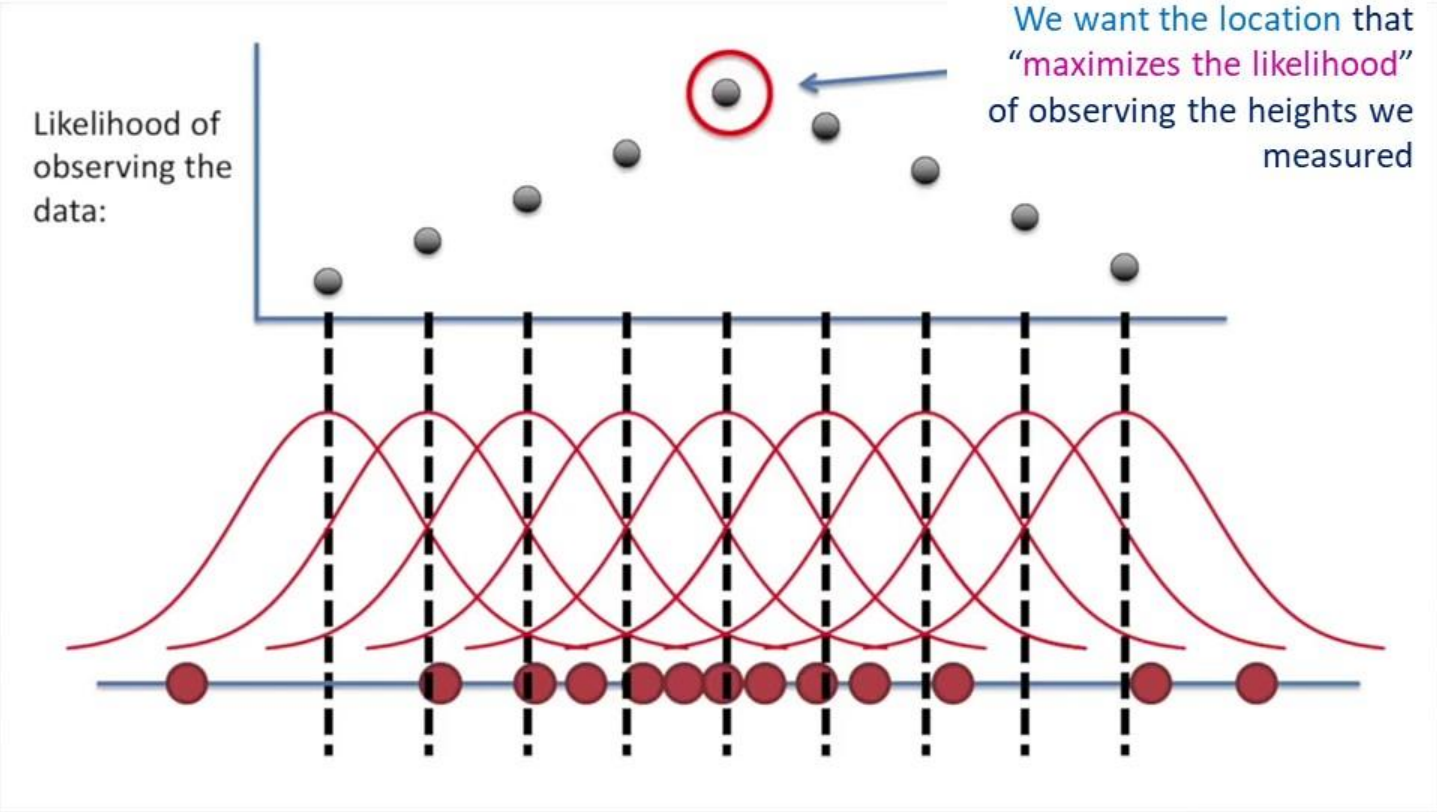

According to a normal distribution
with a mean value here…
…the probability, or "likelihood" of
observing these heights is relatively
high.

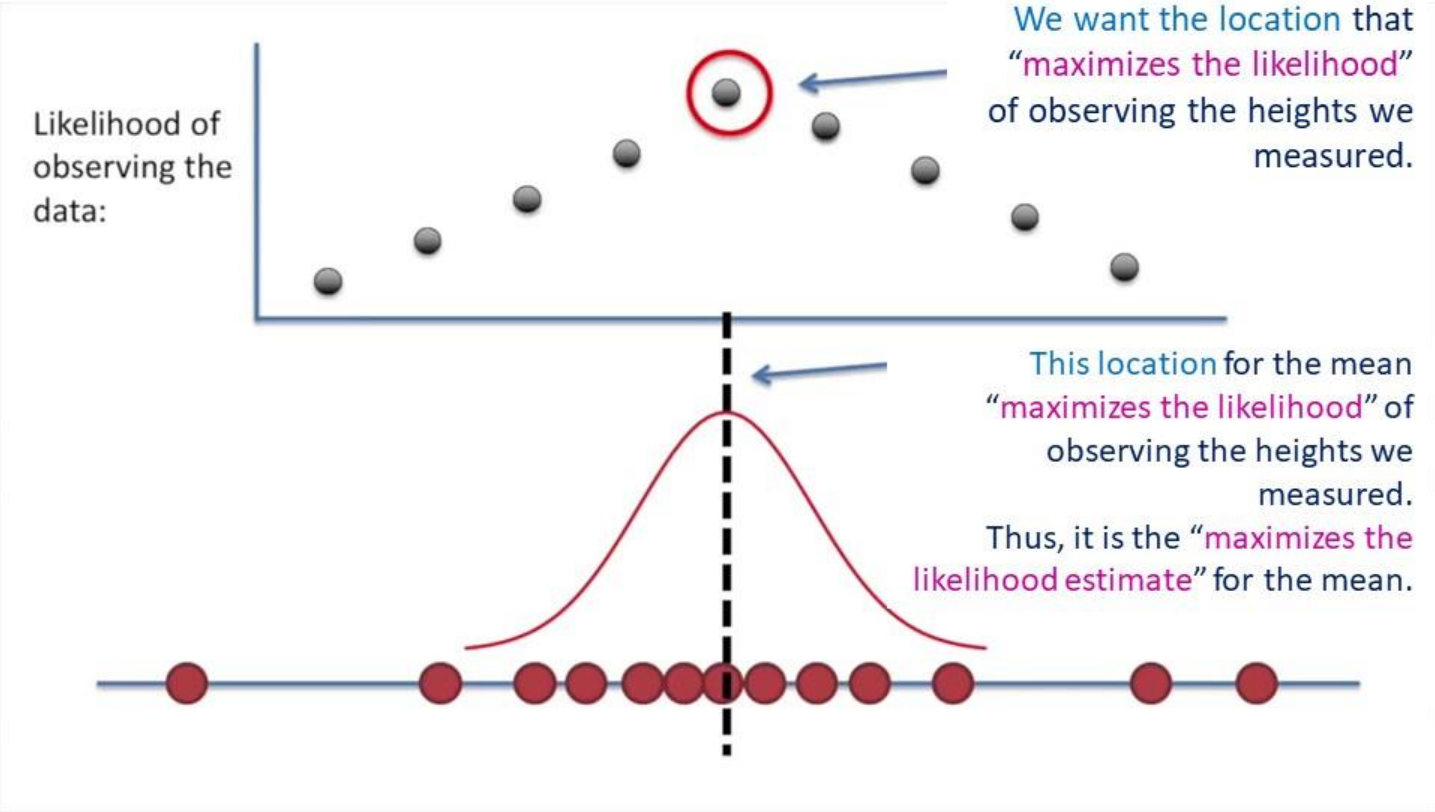

If we kept shifting the normal distribution over…
… then the probability, or "likelihood", of observing these measurements would go down again.

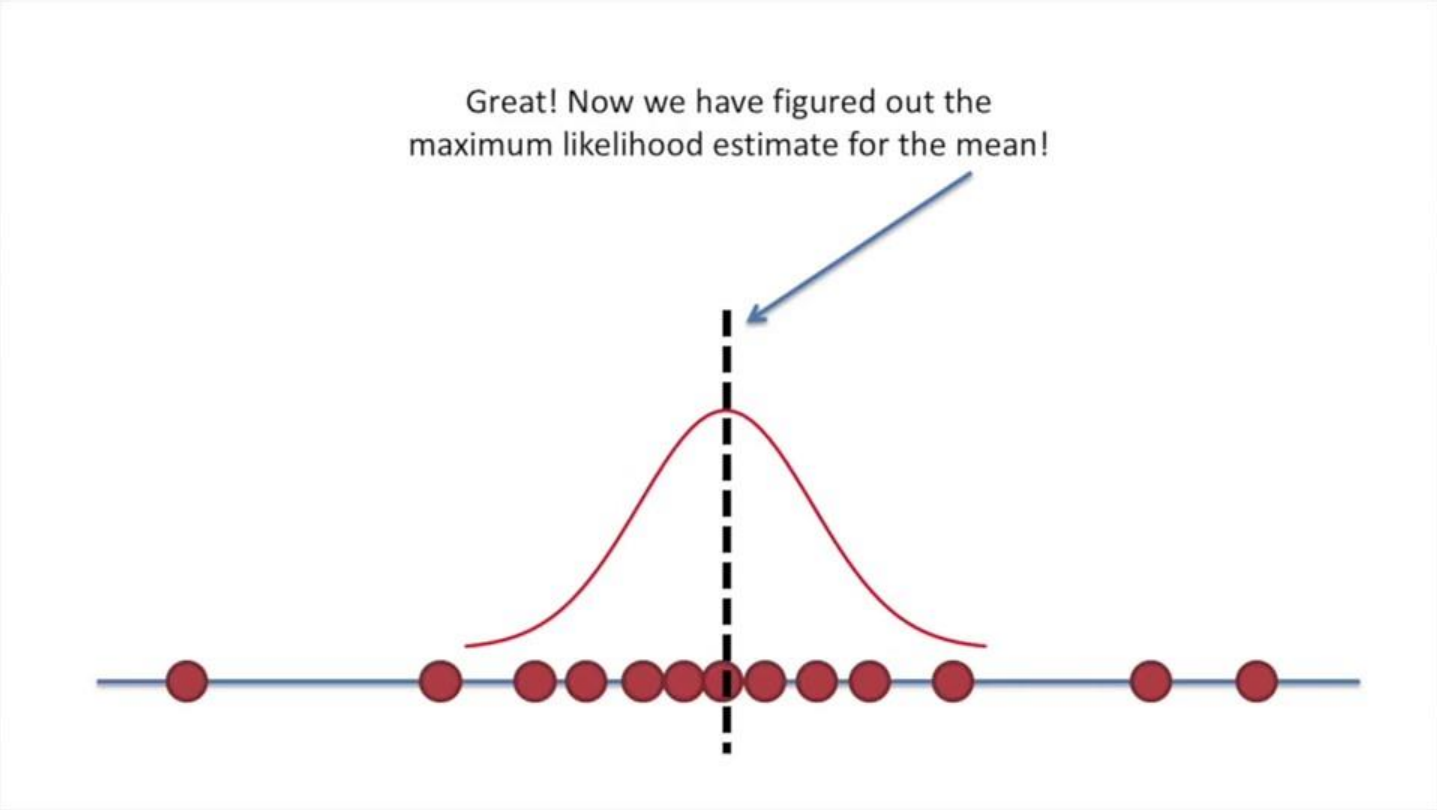

Likelihood of observing the data:
Location of the center of the distribution.

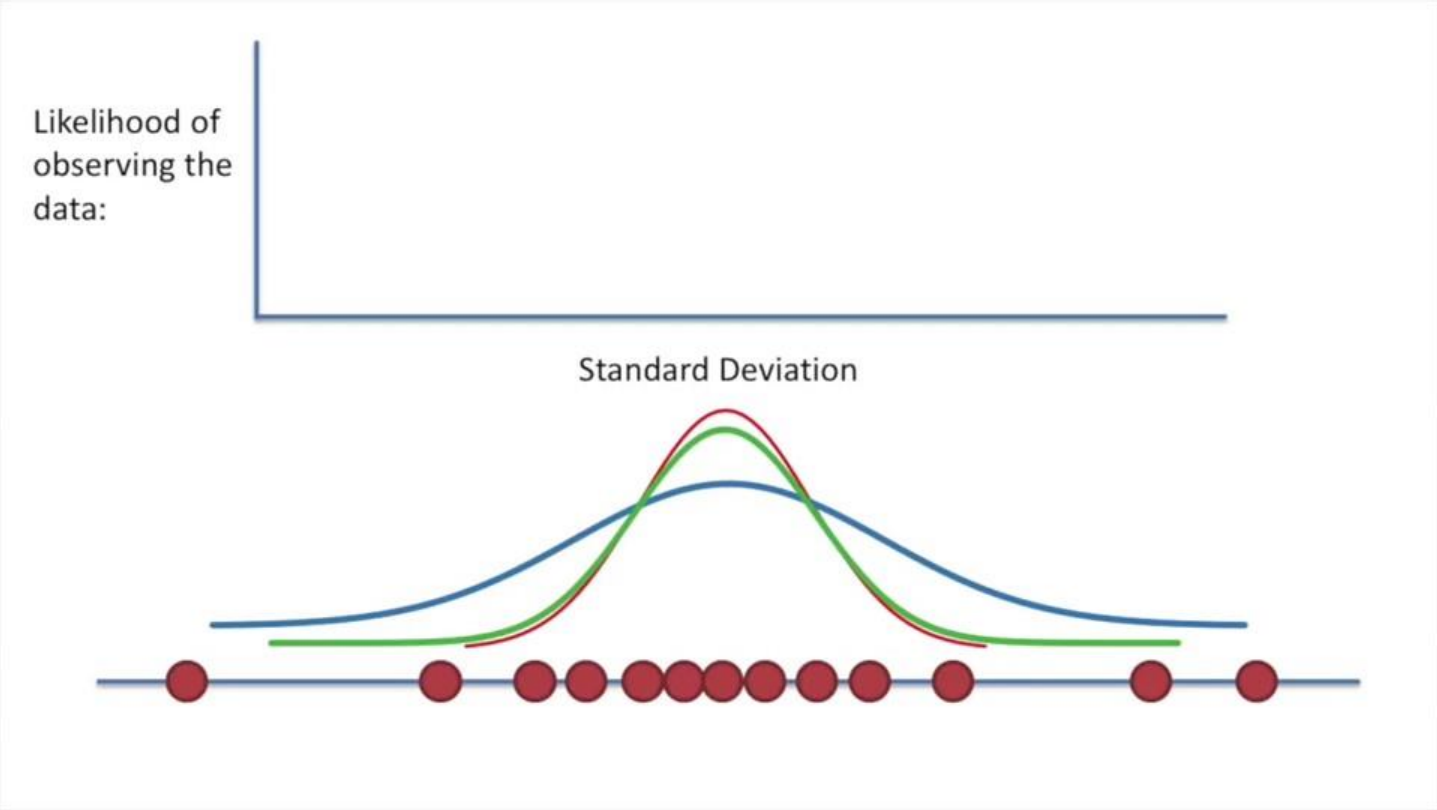

Likelihood of observing the data:
Location of the center of the distribution.

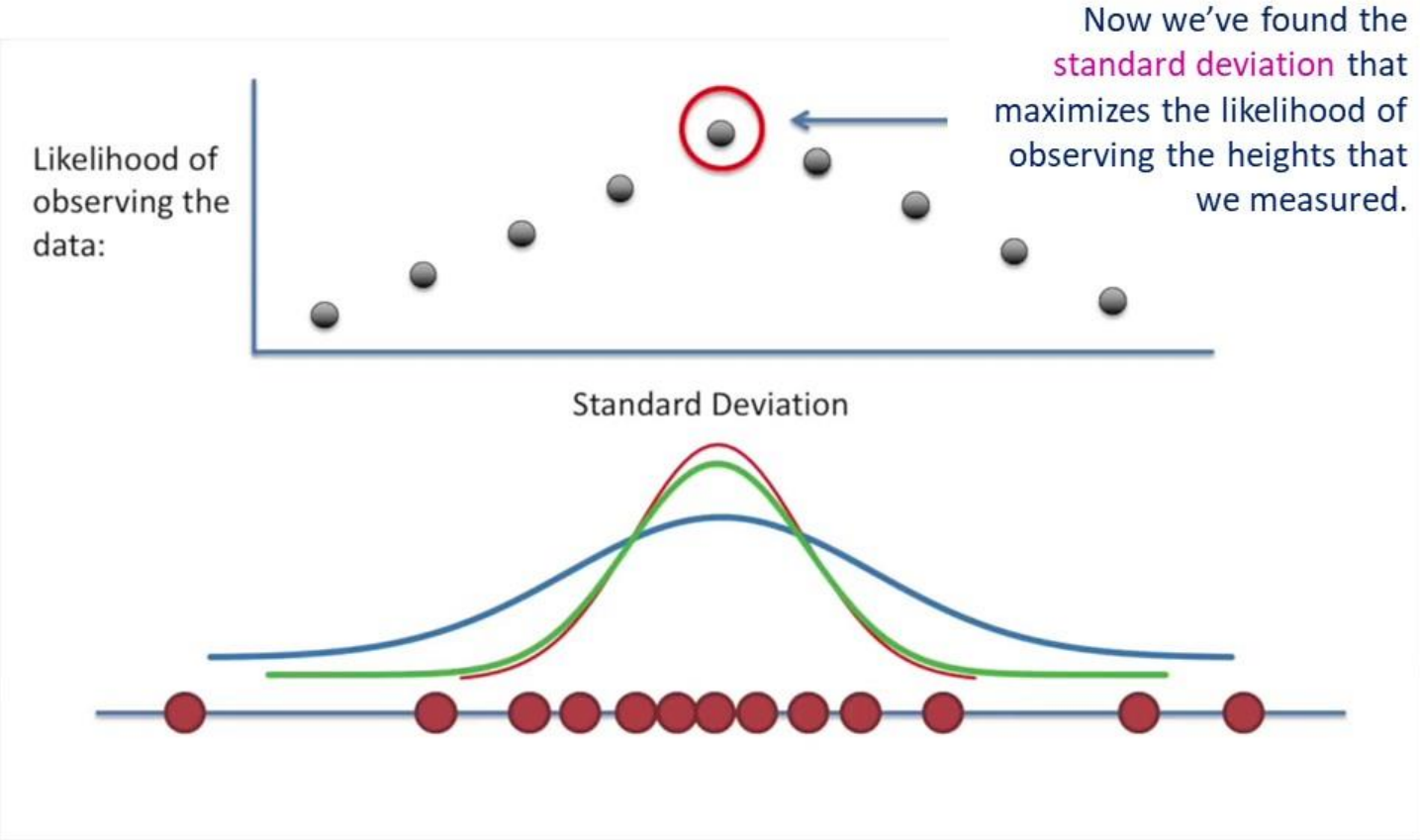

Likelihood of
observing the
data:

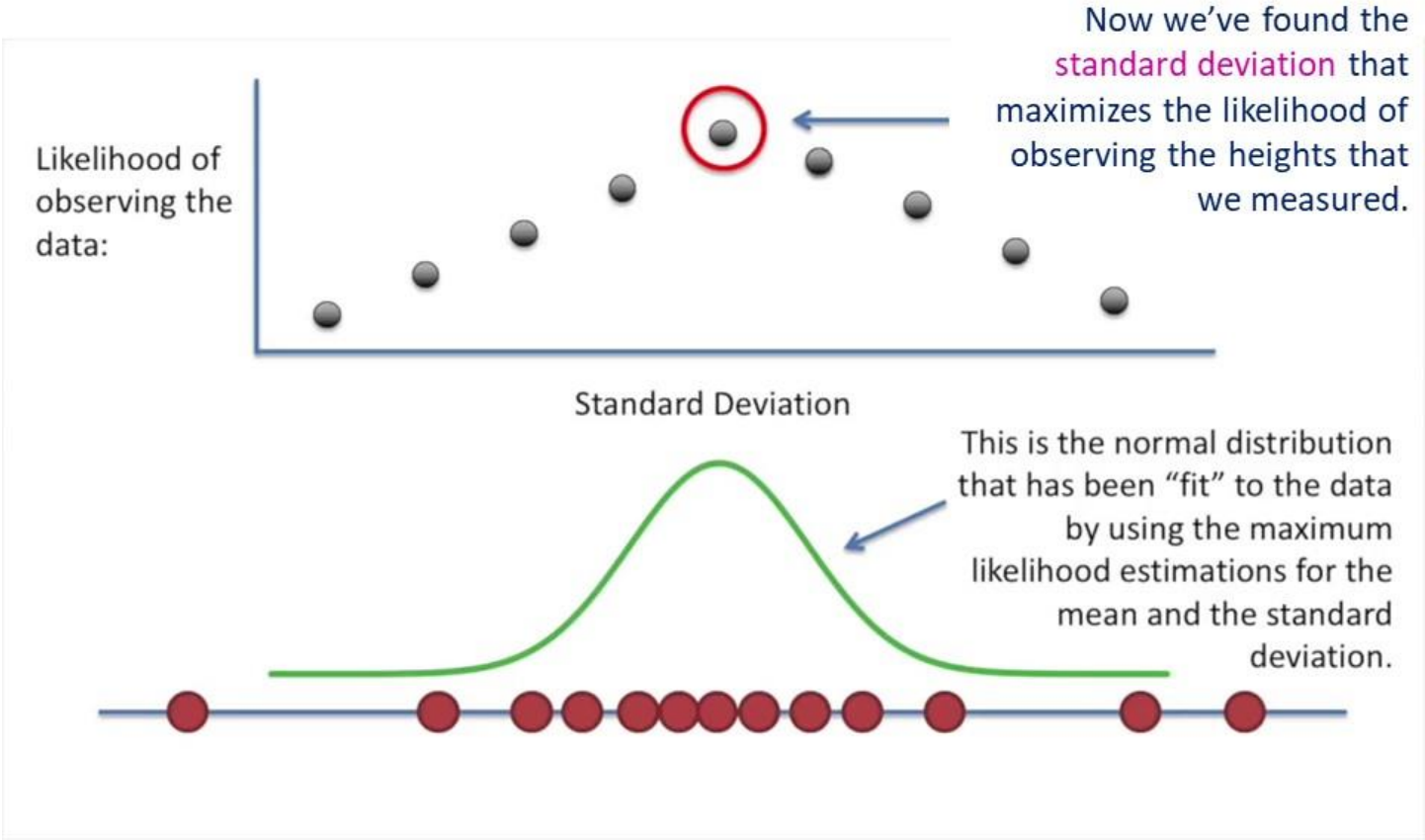

We want the location that
"maximizes the likelihood"
of observing the heights we
measured
Likelihood of
observing the
data:


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

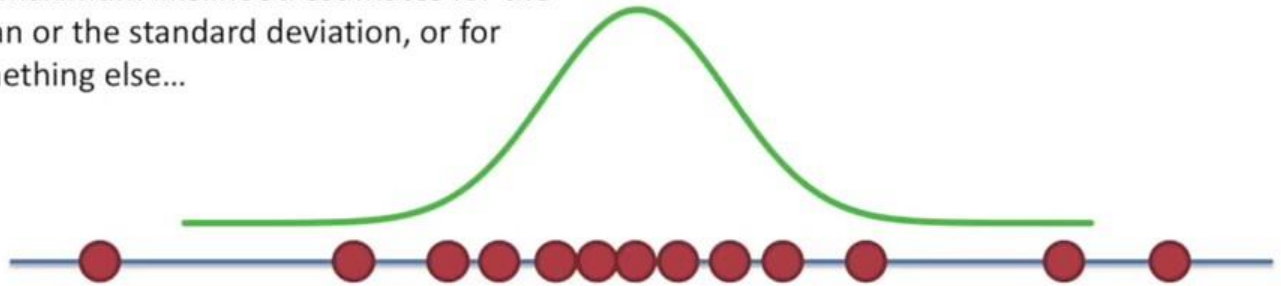

Great! Now we have figured out the
maximum likelihood estimate for the mean!

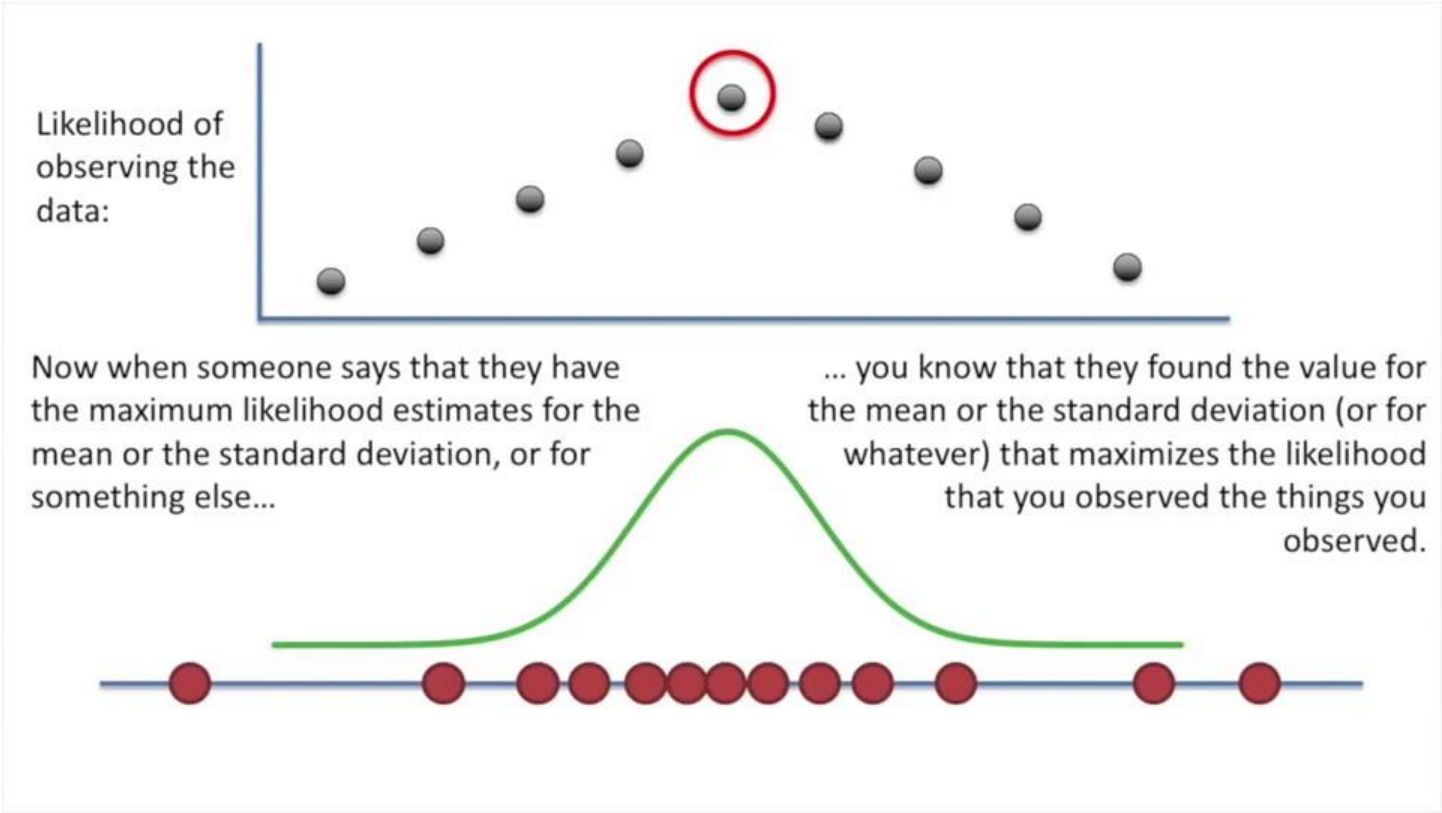

Likelihood of observing the data:
Standard Deviation


Now we've found the standard deviation that maximizes the likelihood of observing the heights that we measured.
Likelihood of observing the data:
Standard Deviation


Now we've found the standard deviation that maximizes the likelihood of observing the heights that we measured.
Likelihood of observing the data:
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.

Curve Fitting: Bayesian Approach

We capture our assumptions about $\mathbf{w}$, before observing the data, in the form of a prior probability distribution $p(\mathbf{w})$.

The effect of the observed data $D = \{t_1, \ldots, t_N\}$ is expressed through the conditional probability $p(D|\mathbf{w})$.

Bayes' theorem, which takes the form

$$p(\mathbf{w}|D) = p(D|\mathbf{w})\, p(\mathbf{w}) / p(D)$$

Thus, allows us to evaluate the uncertainty in w after we have observed D in the form of the posterior probability $p(\mathbf{w}|D)$.

Curve Fitting: Bayesian Approach

The quantity $p(D|\boldsymbol{w})$ on the right-hand side of Bayes' theorem is evaluated for the observed data set D

It can be viewed as a **function of** the parameter vector $\boldsymbol{w}$, in which case it is called the **likelihood function**.

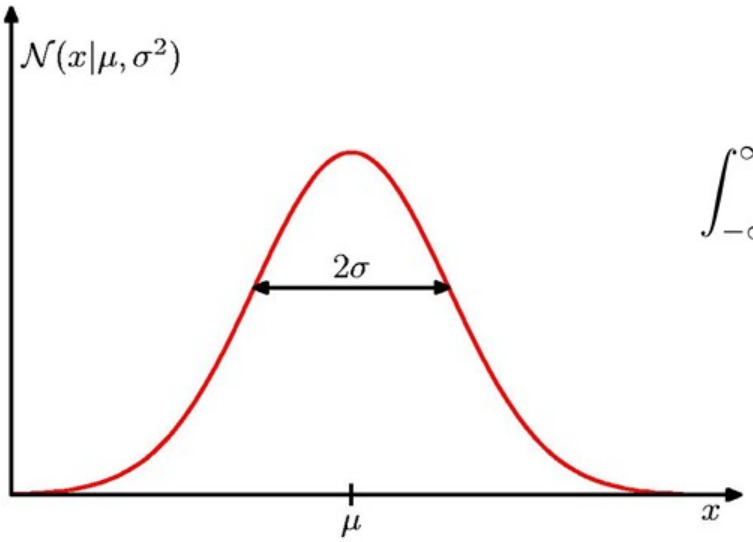
It expresses **how probable** the observed data set is **for different settings** of the parameter vector $\boldsymbol{w}$.

Note that the likelihood is not a probability distribution over $\boldsymbol{w}$, and its integral with respect to $\boldsymbol{w}$ does not (necessarily) equal one.

$$posterior \propto likelihood \times prior$$

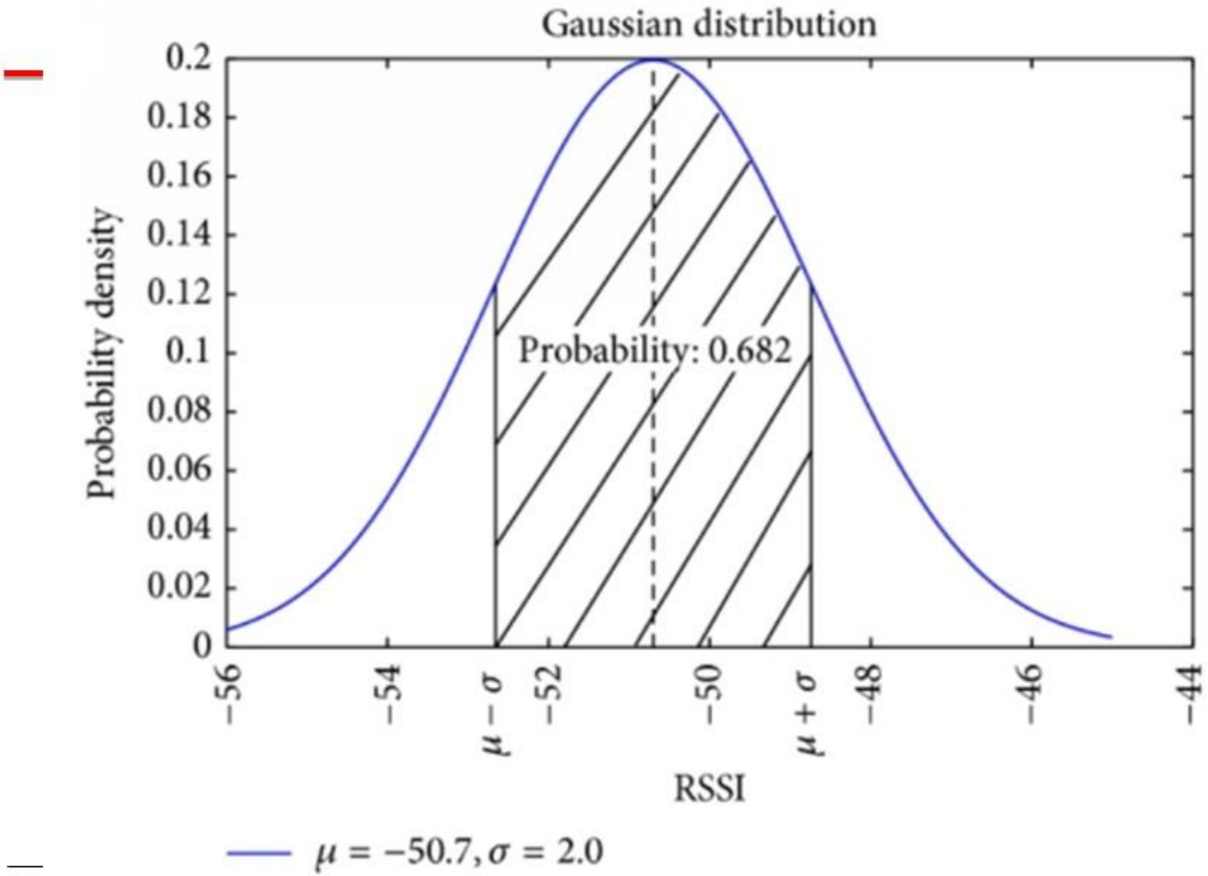
The Gaussian Distribution

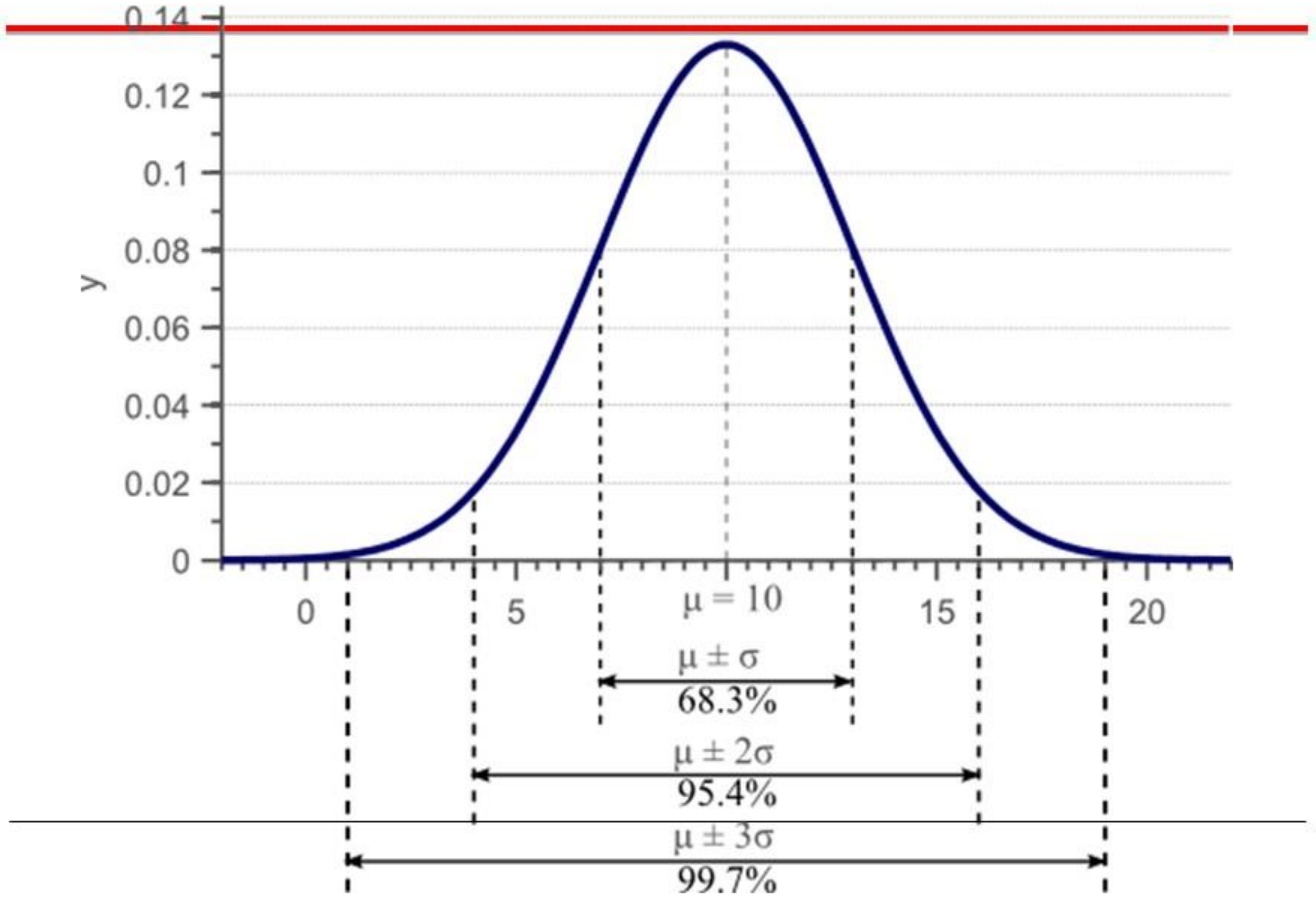
$$\mathcal{N}\left(x|\mu,\sigma^2\right) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

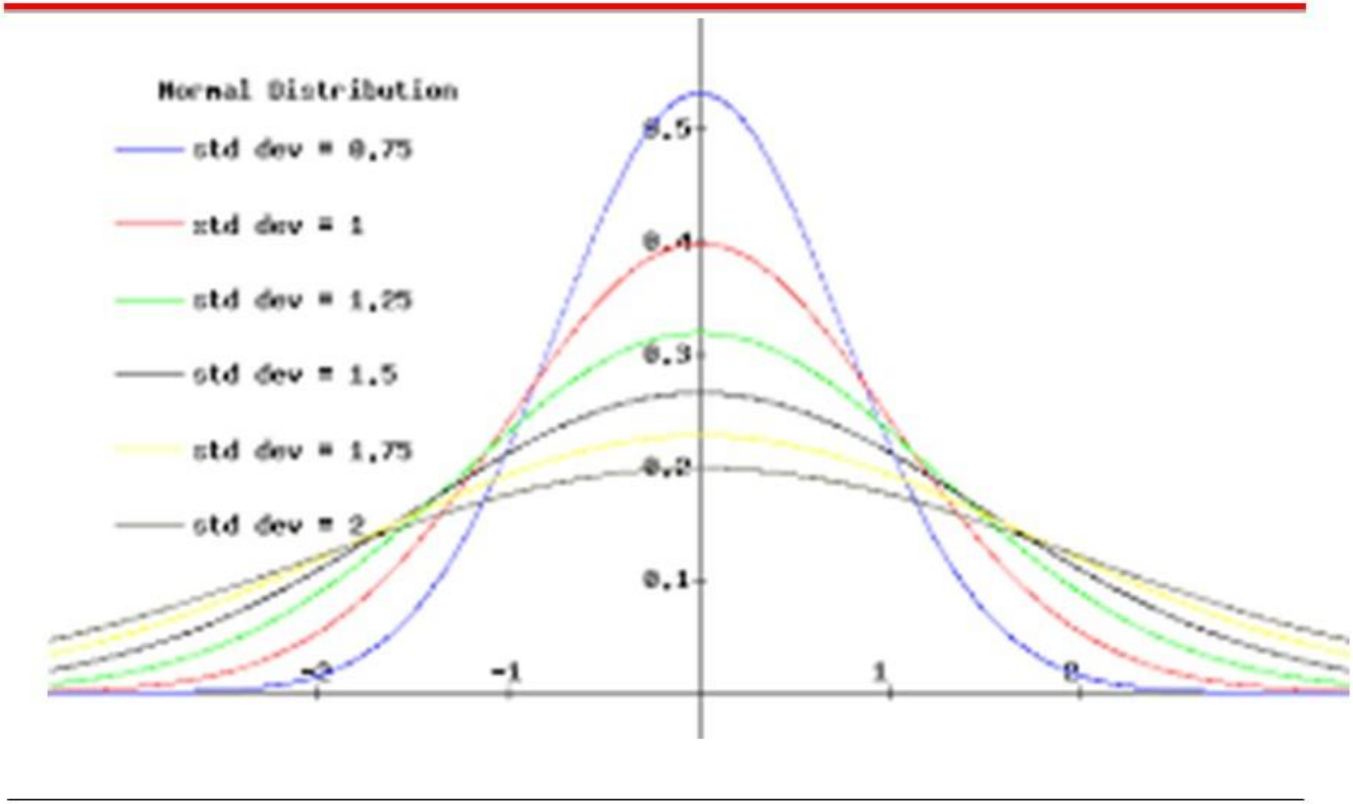


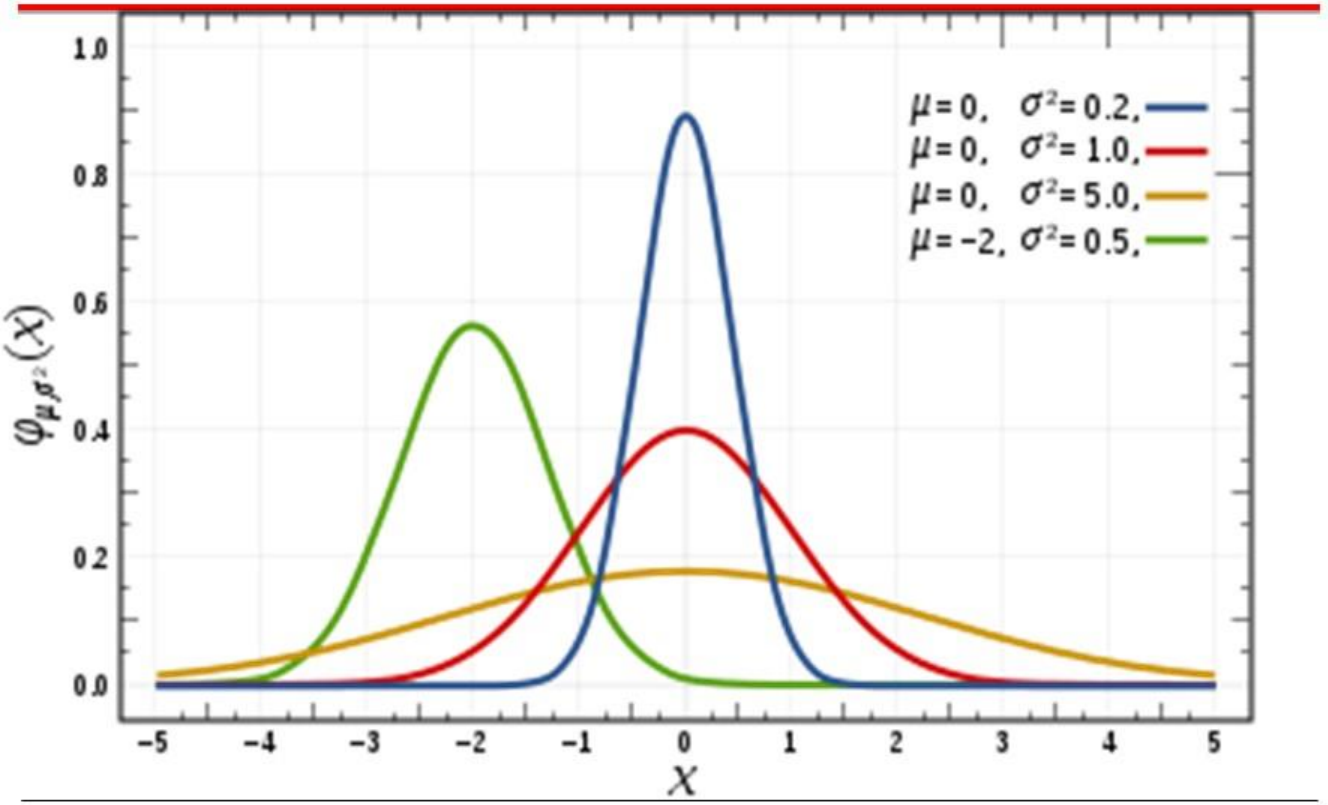
$$\mathcal{N}(x|\mu,\sigma^2) > 0$$

$$\int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu,\sigma^2\right)\,\mathrm{d}x = 1$$


Gaussian distribution
Probability density
0.2
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
Probability: 0.682
−56
−54
μ − σ
−52
−50
μ + σ
−48
−46
−44
RSSI
μ = −50.7, σ = 2.0


0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
y
0
5
μ = 10
15
20
μ ± σ
68.3%
μ ± 2σ
95.4%
μ ± 3σ
99.7%


Normal Distribution
std dev = 0.75
std dev = 1
std dev = 1.25
std dev = 1.5
std dev = 1.75
std dev = 2
0.5
0.4
0.3
0.2
0.1
-2
-1
1
2


1.0
0.8
0.6
0.4
0.2
0.0
$\varphi_{\mu,\sigma^2}(X)$
−5
−4
−3
−2
−1
0
1
2
3
4
5
X
$\mu = 0, \quad \sigma^2 = 0.2,$
$\mu = 0, \quad \sigma^2 = 1.0,$
$\mu = 0, \quad \sigma^2 = 5.0,$
$\mu = -2, \quad \sigma^2 = 0.5,$

Gaussian Mean and Variance

$$\mathbb{E}[x] = \int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu,\sigma^2\right) x \, \mathrm{d}x = \mu$$

$$\mathbb{E}[x^2] = \int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu,\sigma^2\right) x^2 \, \mathrm{d}x = \mu^2 + \sigma^2$$

$$\mathrm{var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2 = \sigma^2$$

The Multivariate Gaussian

$$N(\mathbf{X}|\mu,\Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{X}-\mu)^T\Sigma^{-1}(\mathbf{X}-\mu)\right\}$$

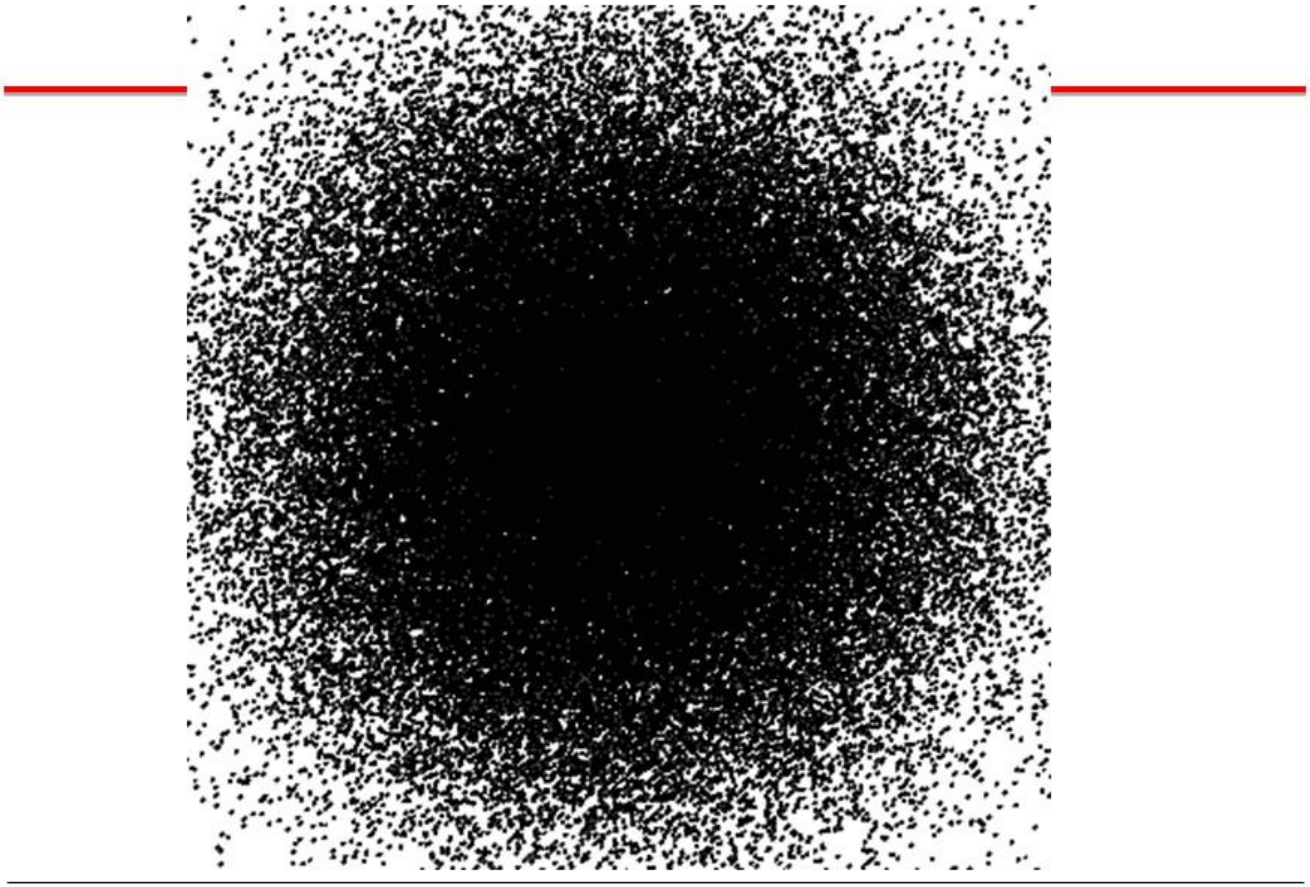
$\mathbf{X}$ is n dimensional vector

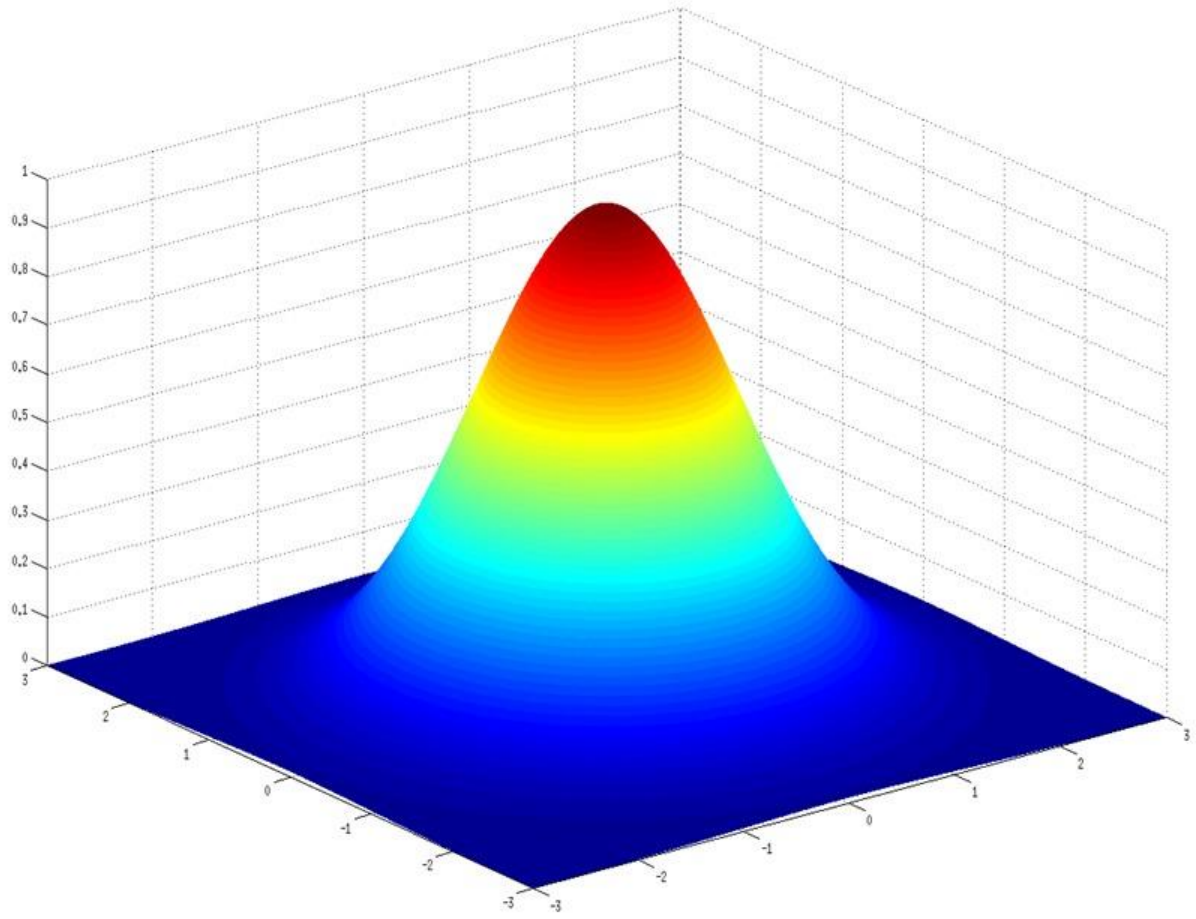
μ is mean vector of dimension n

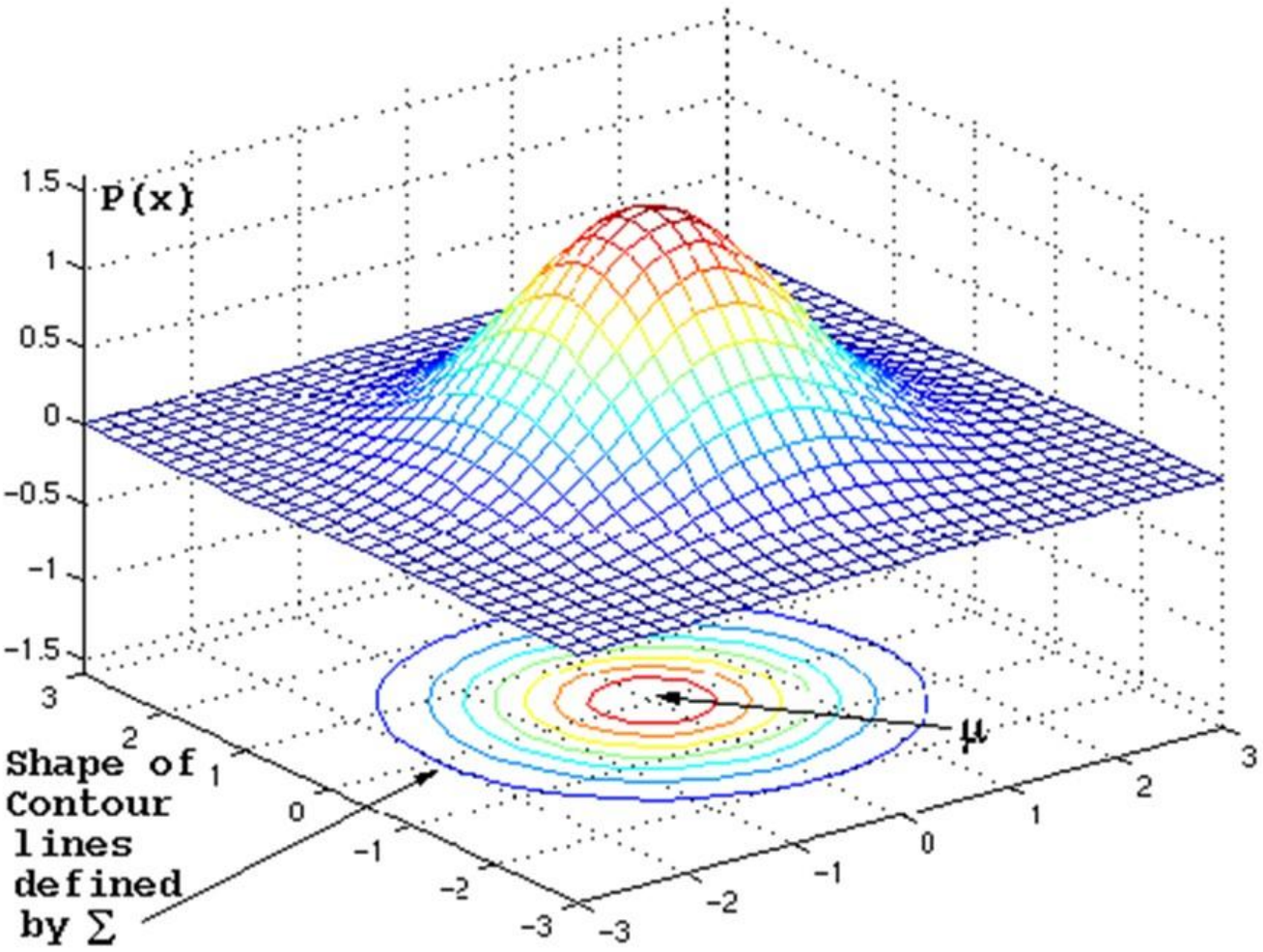
Σ is n x n covariance matrix

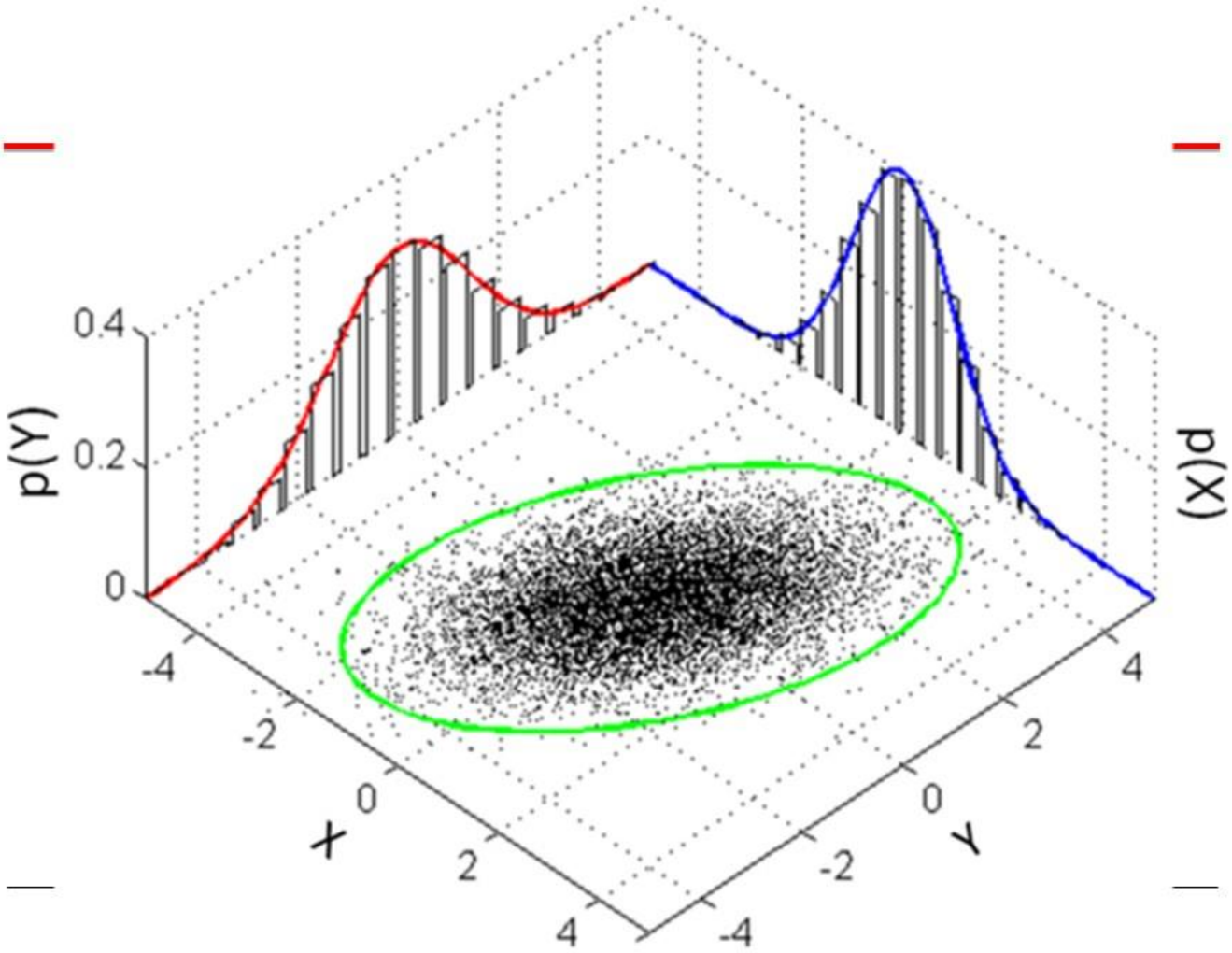
Note: 1. Σ is a symmetric matrix

2. $(\mathbf{X}-\mu)^T\Sigma^{-1}(\mathbf{X}-\mu)$ is a scalar quantity (1x n x n x n x n x 1 = 1x1 =1). It is termed as Mahalanobis distance (squared).



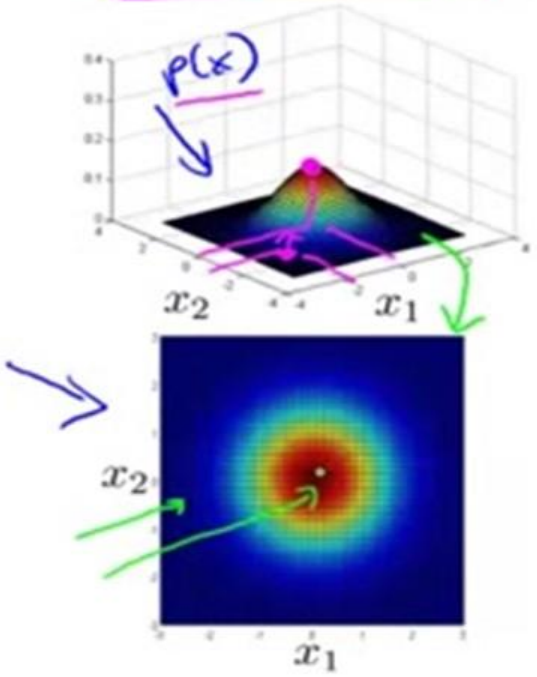

1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
3
2
1
0
-1
-2
-3
-3
-2
-1
0
1
2
3


P(x)
1.5
1
0.5
0
-0.5
-1
-1.5
3
2
1
0
-1
-2
-3
-3
-2
-1
0
1
2
3
μ
Shape of Contour lines defined by Σ


p(Y)
0.4
0.2
0
-4
-2
0
2
4
X
4
2
0
-2
-4
Y
p(X)

Multivariate Gaussian (Normal) examples

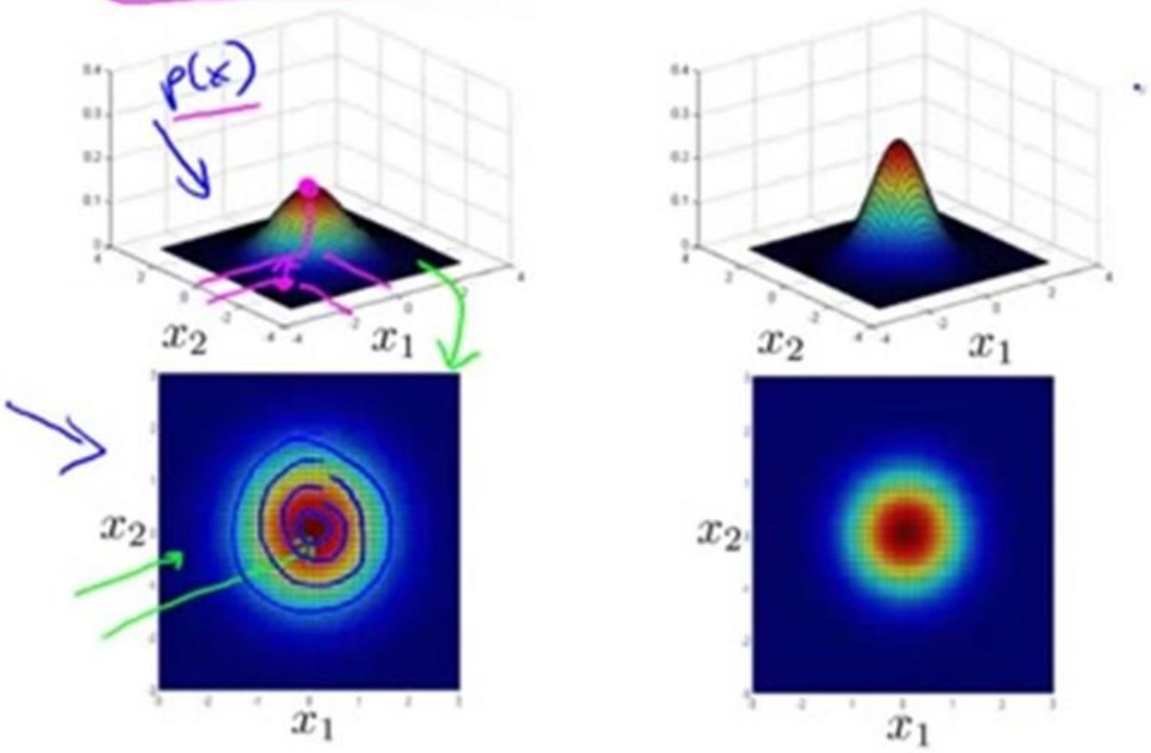
$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$



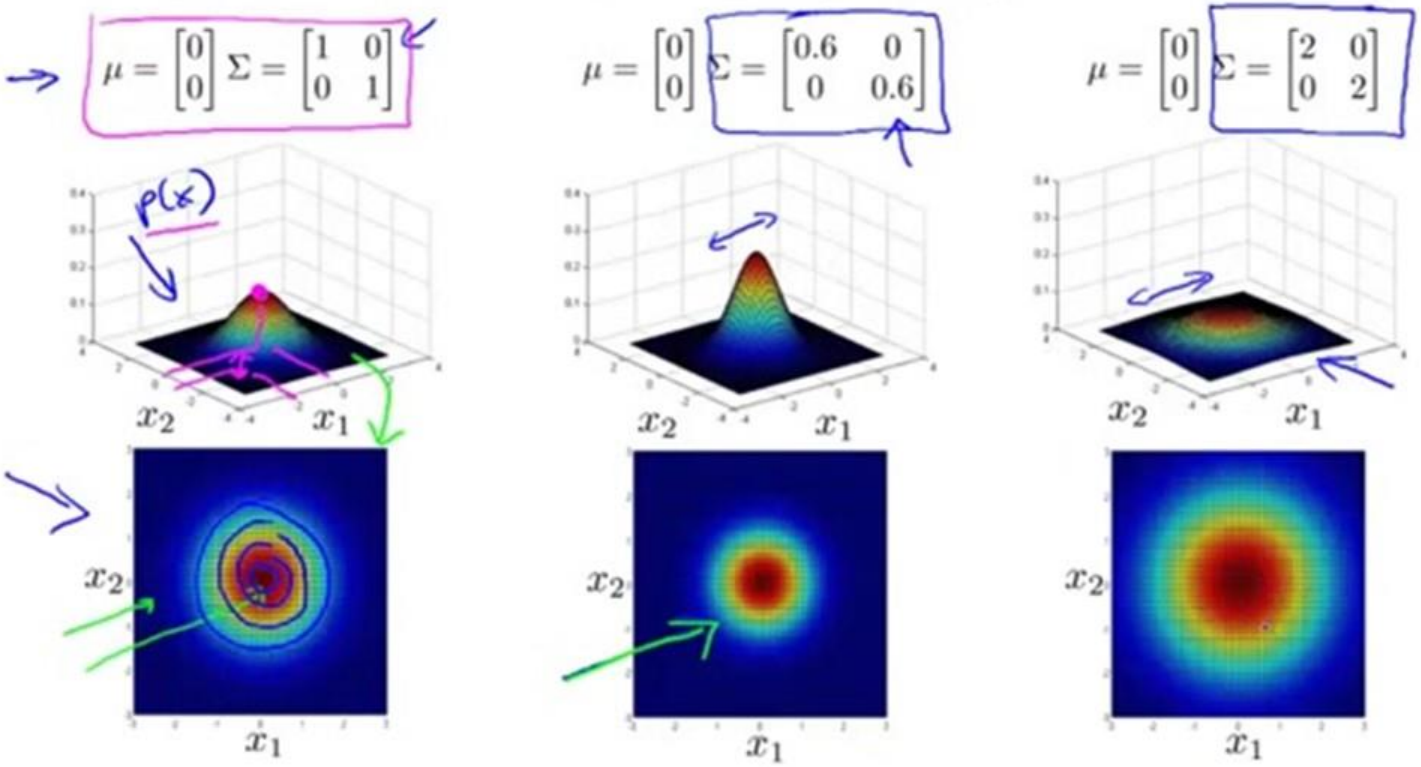
Multivariate Gaussian (Normal) examples

$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 0.6 & 0 \\ 0 & 0.6 \end{bmatrix}$$

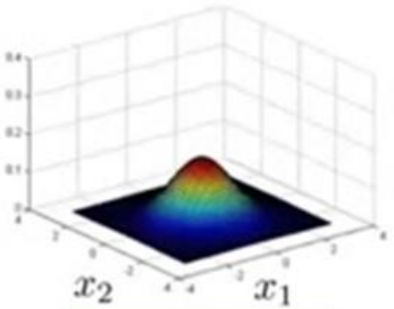


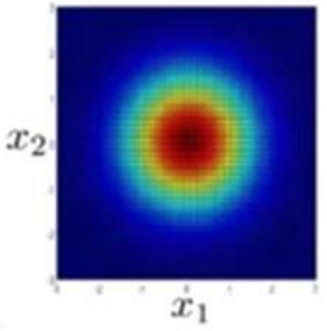
Multivariate Gaussian (Normal) examples



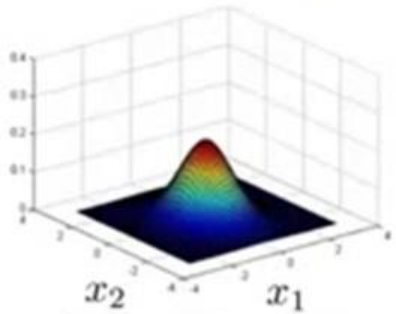
Multivariate Gaussian (Normal) examples

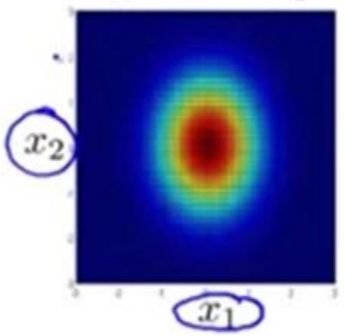
$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$





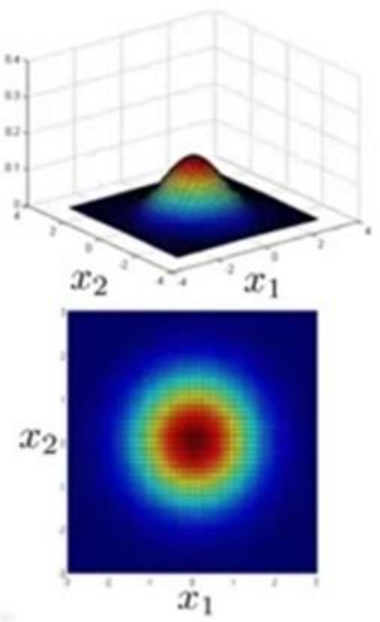
$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 0.6 & 0 \\ 0 & 1 \end{bmatrix}$$



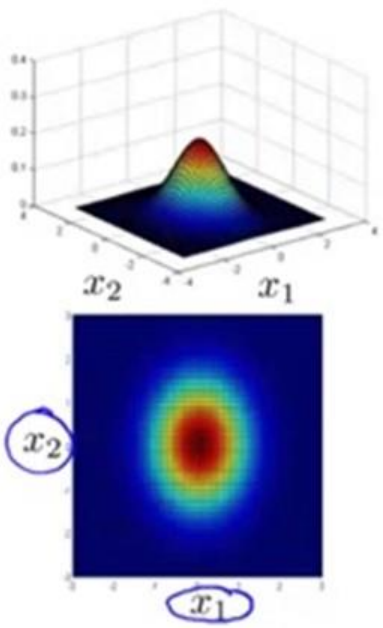


Multivariate Gaussian (Normal) examples

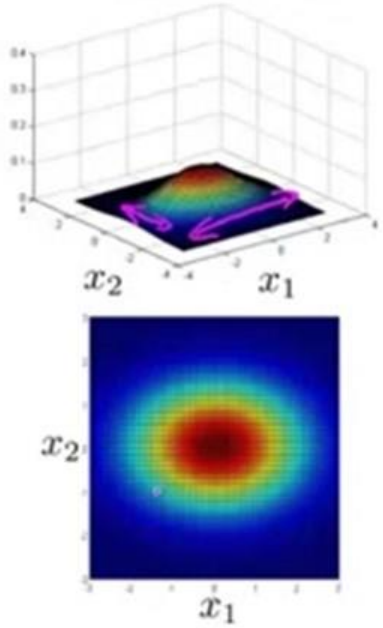
$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$



$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 0.6 & 0 \\ 0 & 1 \end{bmatrix}$

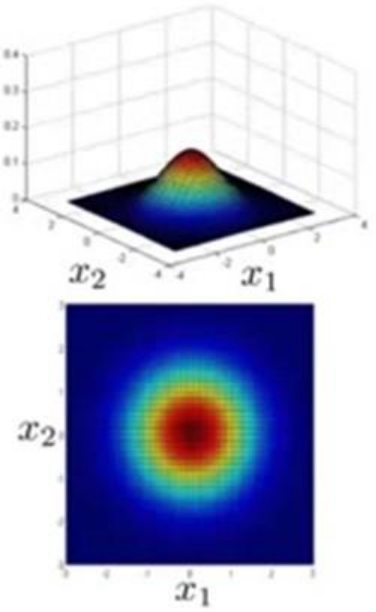


$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$

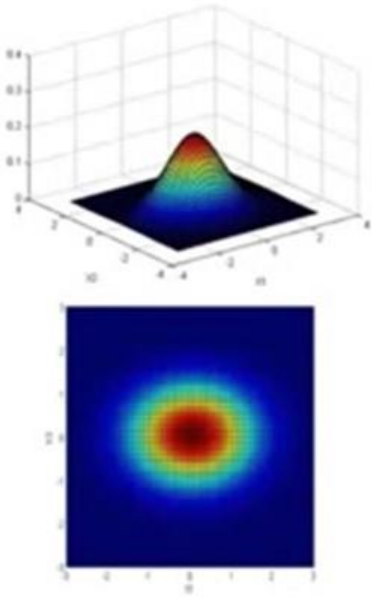


Multivariate Gaussian (Normal) examples

$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

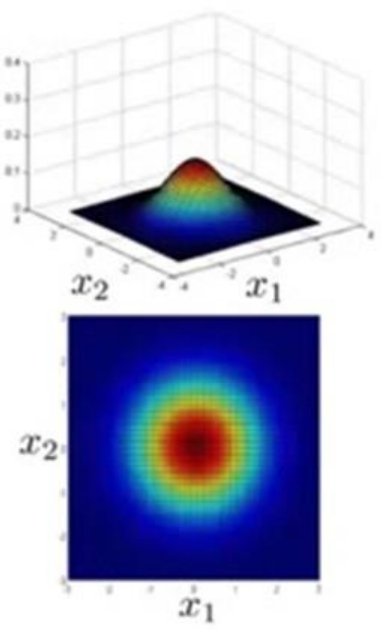


$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 0.6 \end{bmatrix}$

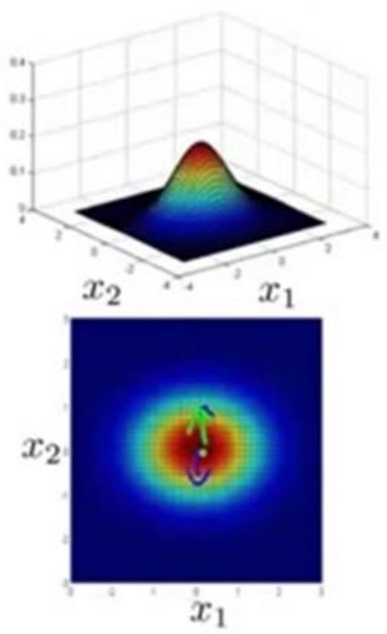


Multivariate Gaussian (Normal) examples

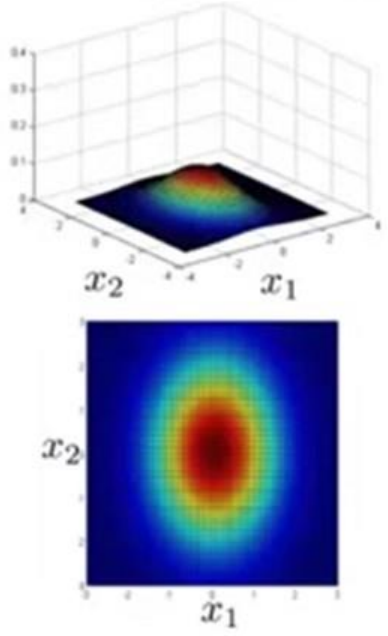
$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$



$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 0.6 \end{bmatrix}$

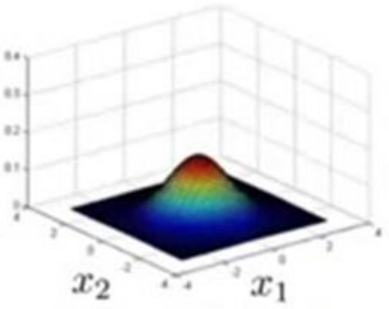


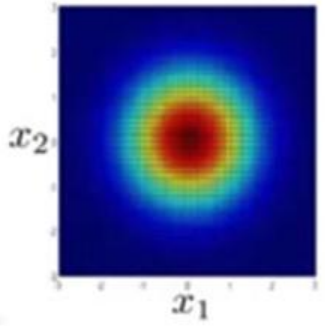
$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$



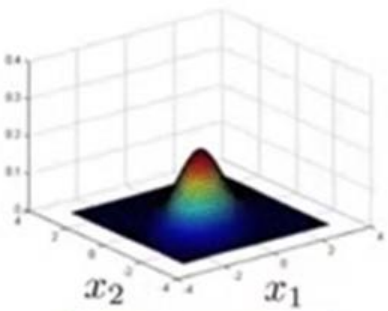
Multivariate Gaussian (Normal) examples

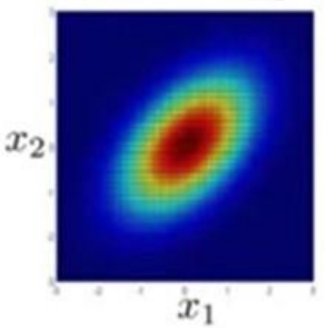
$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$





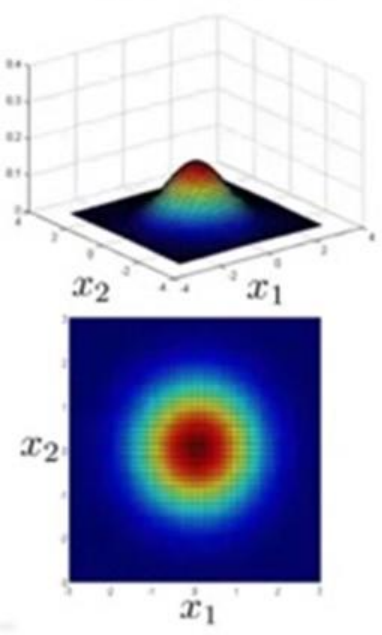
$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$



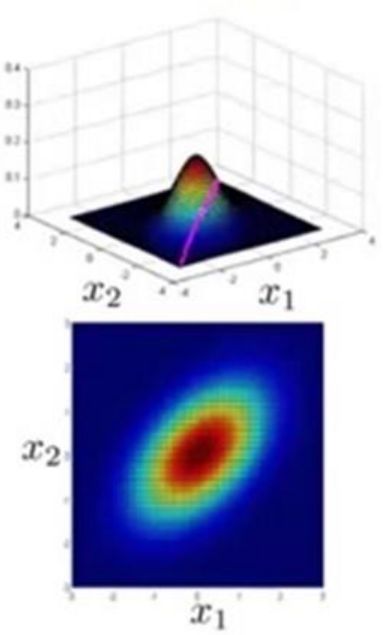


Multivariate Gaussian (Normal) examples

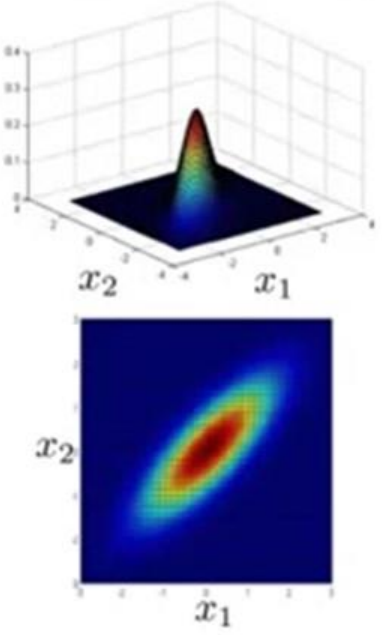
$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$



$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$

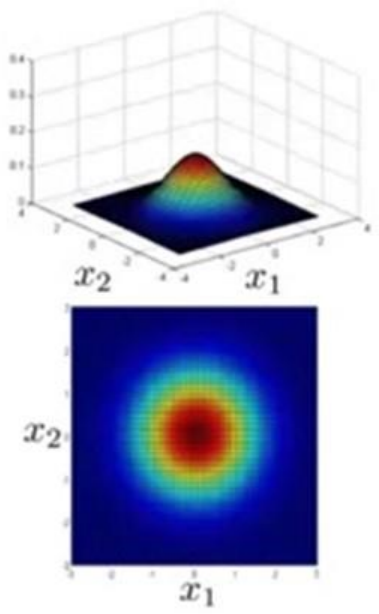


$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}$

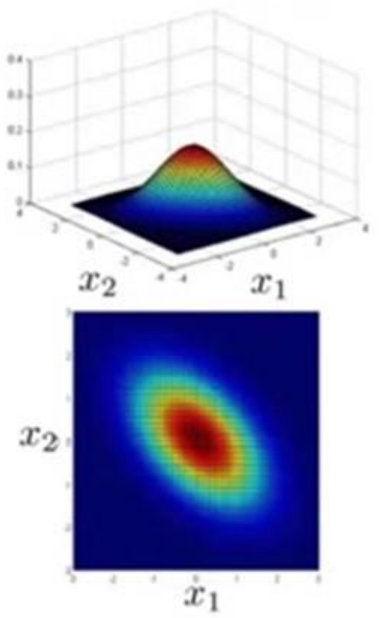


Multivariate Gaussian (Normal) examples

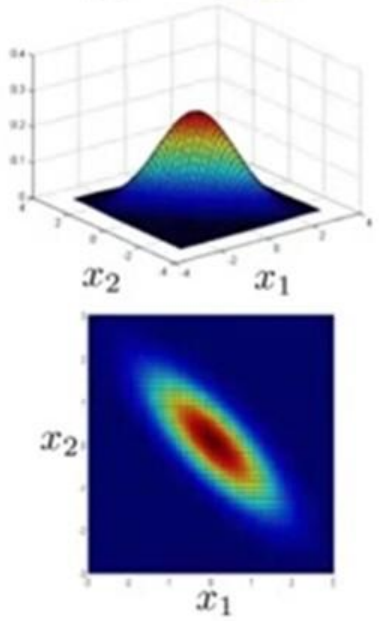
$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$



$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}$

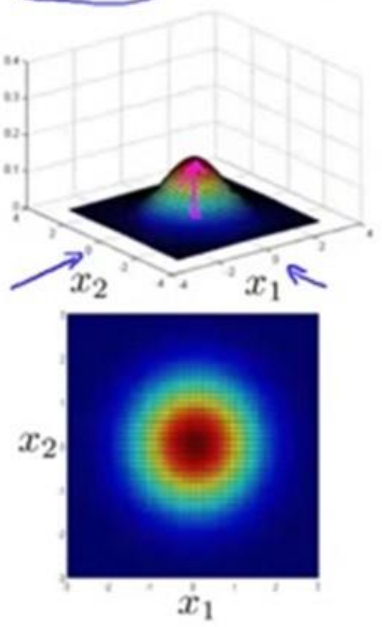


$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & -0.8 \\ -0.8 & 1 \end{bmatrix}$

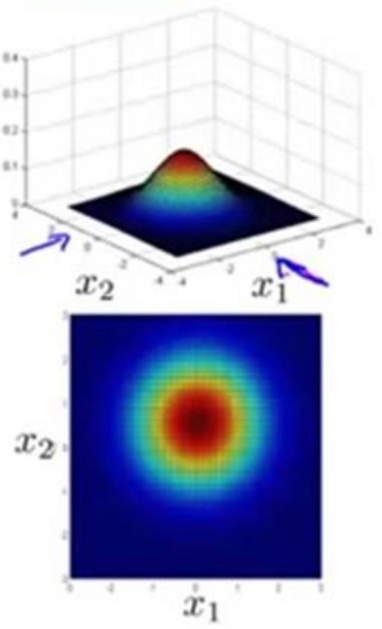


Multivariate Gaussian (Normal) examples

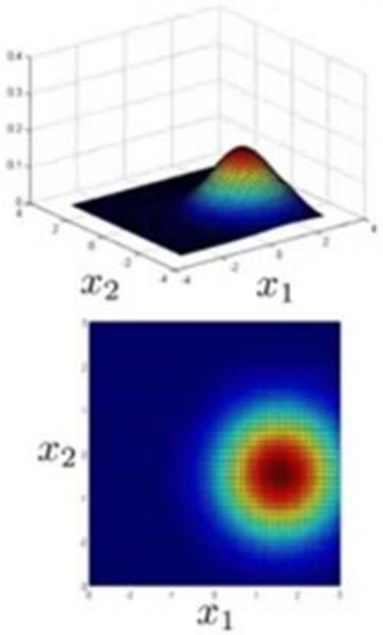
$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$



$\mu = \begin{bmatrix} 0 \\ 0.5 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$



$\mu = \begin{bmatrix} 1.5 \\ -0.5 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$



Geometry of the Gaussian

$$\Delta^2 = (x-\mu)\Sigma^{-1}(x-\mu)^T$$

Oval shows constant Δ^2 value…

Write Σ in terms of eigenvectors…

$$\Sigma = U\Lambda U^T$$

$$\Sigma = \begin{bmatrix} u_1 & u_2 \\ \downarrow & \downarrow \end{bmatrix} \begin{bmatrix} \lambda_1 & \\ & \lambda_2 \end{bmatrix} \begin{bmatrix} u_1 & \rightarrow \\ u_2 & \rightarrow \end{bmatrix}$$

Then…

$$y_i = u_i^T(x-\mu)$$

$$\Delta^2 = \frac{y_1^2}{\lambda_1} + \frac{y_2^2}{\lambda_2}$$

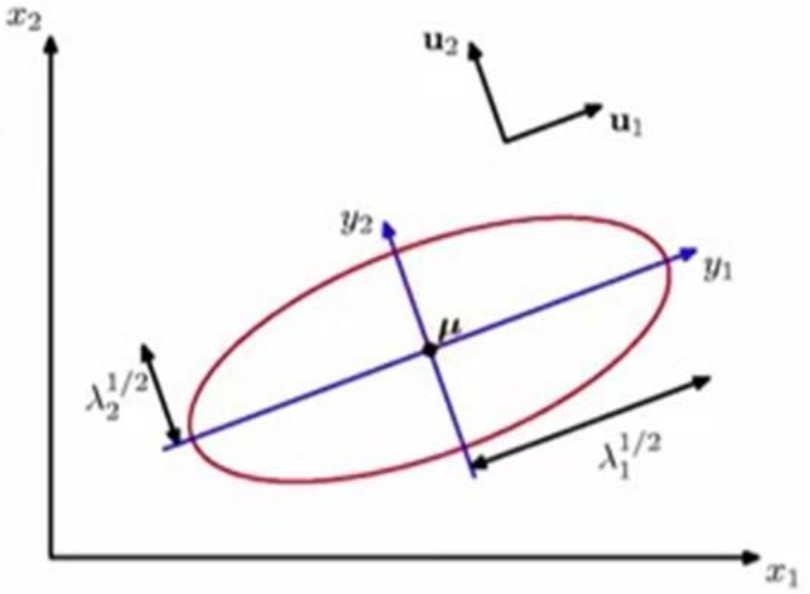
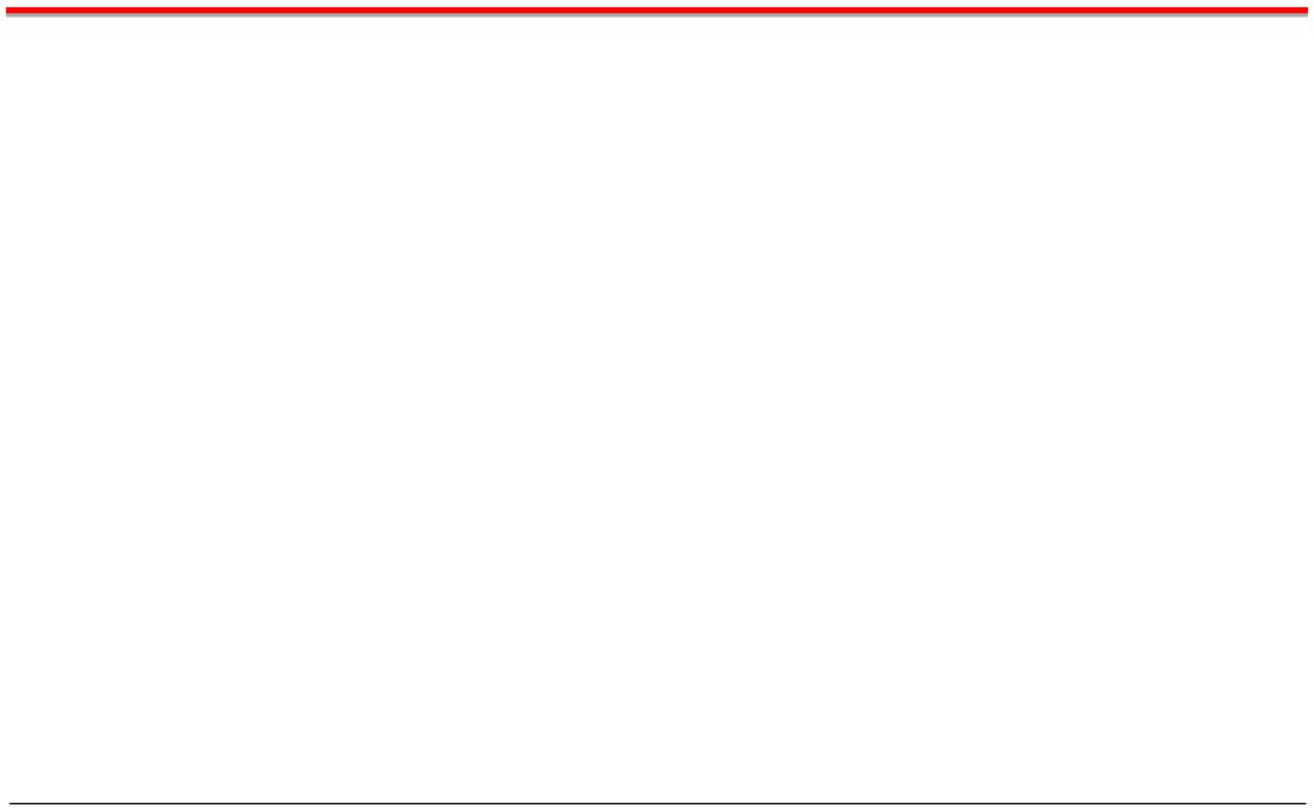
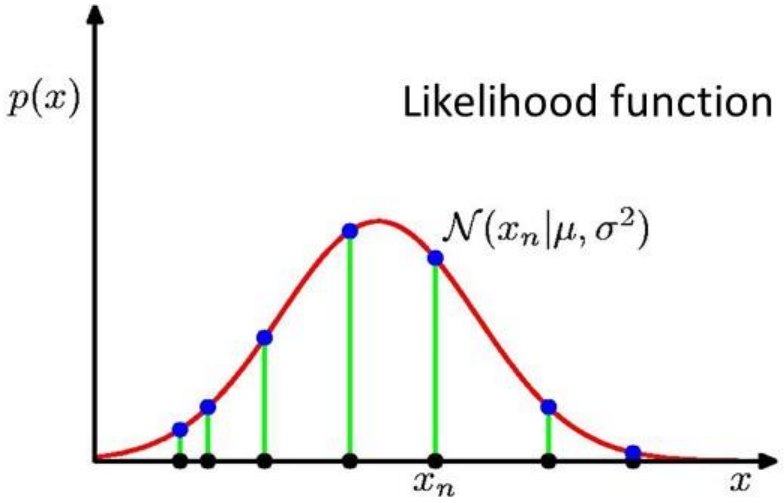


Figure from Chris Bishop's slides



Gaussian Parameter Estimation



$$p(\mathbf{x}|\mu,\sigma^2) = \prod_{n=1}^{N} \mathcal{N}\left(x_n|\mu,\sigma^2\right)$$

Maximum (Log) Likelihood

$$\ln p\left(\mathbf{x}|\mu,\sigma^2\right) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln(2\pi)$$

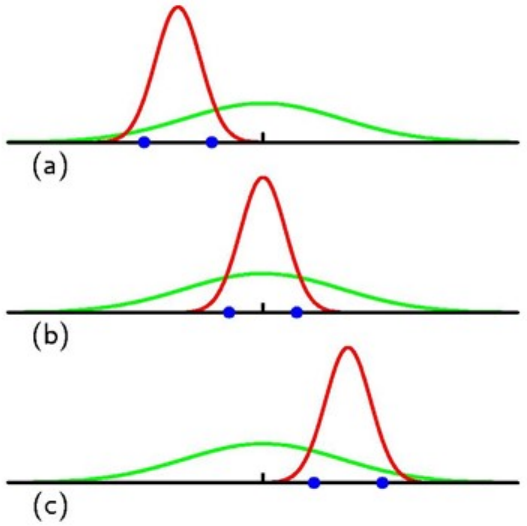
$$\mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}x_n \qquad \sigma^2_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}(x_n-\mu_{\mathrm{ML}})^2$$

Properties of μ_{ML} and σ^2_{ML}

$$\mathbb{E}[\mu_{\mathrm{ML}}] = \mu$$

$$\mathbb{E}[\sigma^2_{\mathrm{ML}}] = \left(\frac{N-1}{N}\right)\sigma^2$$

$$\begin{aligned}\widetilde{\sigma}^2 &= \frac{N}{N-1}\sigma^2_{\mathrm{ML}} \\ &= \frac{1}{N-1}\sum_{n=1}^{N}(x_n - \mu_{\mathrm{ML}})^2\end{aligned}$$



Curve Fitting Re-visited

The goal in the curve fitting problem is to be able to
make predictions for the target variable t
given some new value of the input variable x
on the basis of a set of training data comprising N input values
$\mathbf{X} = \{x_1, \ldots, x_N\}$ and their corresponding target values $t = \{t_1, \ldots, t_N\}$.

We can **express our uncertainty** over the value of the **target variable** using a **probability distribution**.

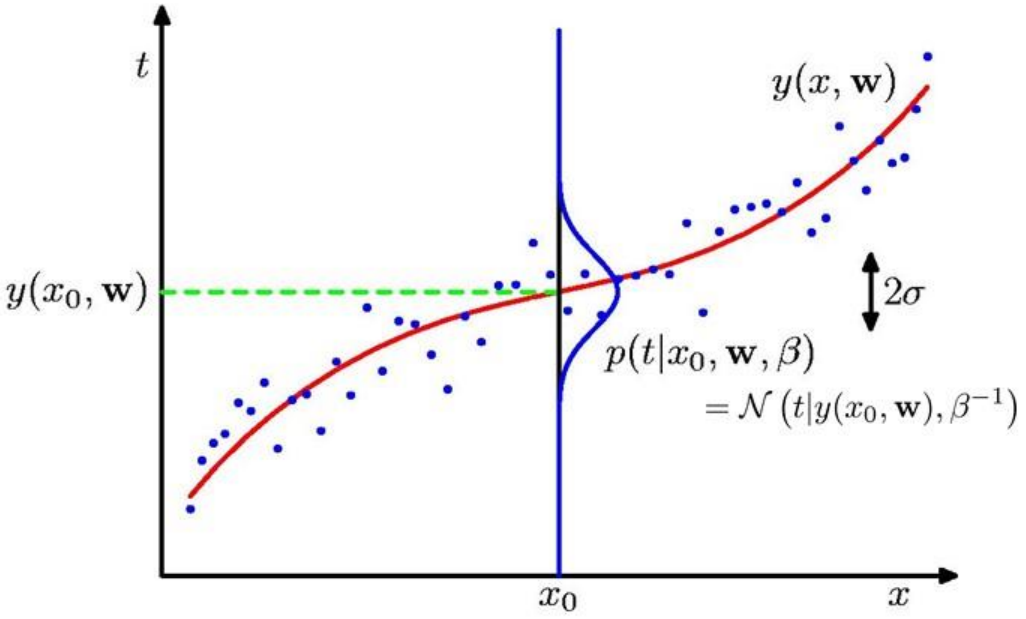
For this purpose, we **shall assume** that, given the value of x, the corresponding **value of** t has a **Gaussian distribution** with a **mean** equal to the value $y(x, \boldsymbol{w})$ of the polynomial curve given by (1.1).

Thus we have: $$p(t|x, \mathbf{w}, \beta) = N\left(t \middle| y(x, \mathbf{w}), \beta^{-1}\right)$$

Where $\beta = 1/\sigma^2$ is defined as precision parameter

Assume that data (target values) are drawn from above distribution independently.

Curve Fitting Re-visited…



Maximum Likelihood

We can use the taring date {**X**, **t**} to determine the values of the unknown parameters **w** and β by maximum likelihood. The likelihood function is given by

$$p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}\left(t_n|y(x_n, \mathbf{w}), \beta^{-1}\right)$$

$$\ln p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = \underbrace{-\frac{\beta}{2} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2}_{\beta E(\mathbf{w})} + \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi)$$

For obtaining $\mathbf{w}_{\mathrm{ML}}$, maximize above equation with respect to **w**.

we can omit the last two terms as they do not depend on **w**.

Also, we note that scaling the log likelihood by a positive constant coefficient does not alter the location of the maximum with respect to **w**, and so we can replace the coefficient $\beta/2$ with $1/2$.

Finally, instead of maximizing the log likelihood, we can equivalently minimize the negative log likelihood.

Maximum Likelihood

We therefore see that maximizing likelihood is equivalent, so far as determining **w** is concerned, to minimizing the *sum-of-squares error function* defined by (1.2).

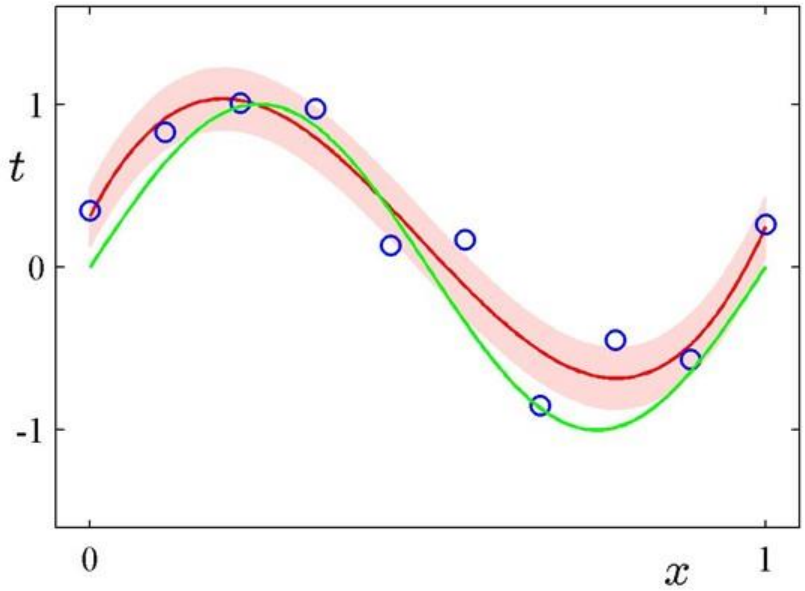
Thus the sum-of-squares error function has arisen as a consequence of maximizing likelihood under the assumption of a Gaussian noise distribution.

We can also use maximum likelihood to determine the precision parameter β of the Gaussian conditional distribution. Maximizing likelihood with respect to β gives

$$\frac{1}{\beta_{\mathrm{ML}}} = \frac{1}{N}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}_{\mathrm{ML}}) - t_n\}^2$$

Predictive Distribution

$$p(t|x, \mathbf{w}_{\mathrm{ML}}, \beta_{\mathrm{ML}}) = \mathcal{N}\left(t|y(x, \mathbf{w}_{\mathrm{ML}}), \beta_{\mathrm{ML}}^{-1}\right)$$



MAP: A Step towards Bayes

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}) = \left(\frac{\alpha}{2\pi}\right)^{(M+1)/2} \exp\left\{-\frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}\right\}$$

$$p(\mathbf{w}|\mathbf{x}, \mathbf{t}, \alpha, \beta) \propto p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta)p(\mathbf{w}|\alpha)$$

$$\beta\widetilde{E}(\mathbf{w}) = \frac{\beta}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}$$

Determine $\mathbf{w}_{\mathrm{MAP}}$ by minimizing regularized sum-of-squares error, $\widetilde{E}(\mathbf{w})$.

Bayesian Curve Fitting

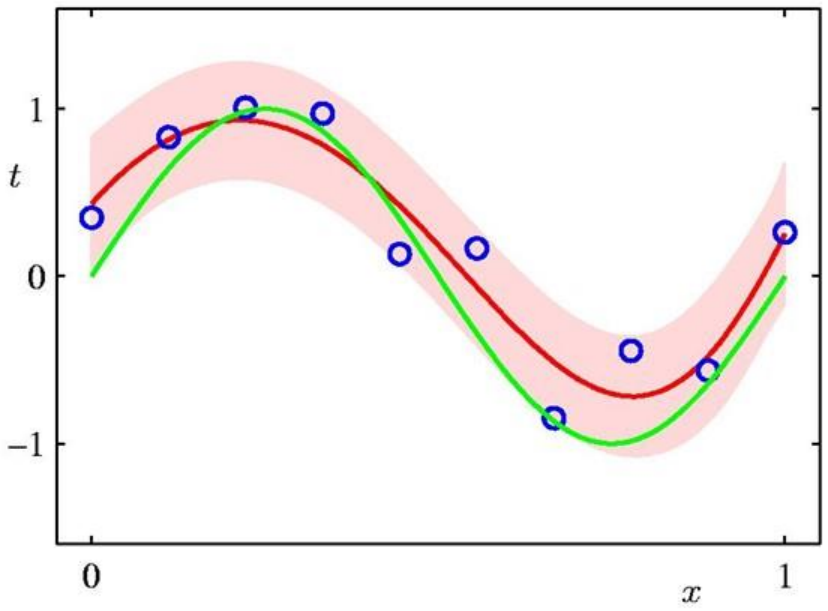
$$p(t|x,\mathbf{x},\mathbf{t}) = \int p(t|x,\mathbf{w})p(\mathbf{w}|\mathbf{x},\mathbf{t})\,\mathrm{d}\mathbf{w} = \mathcal{N}\left(t|m(x),s^2(x)\right)$$

$$m(x) = \beta\boldsymbol{\phi}(x)^{\mathrm{T}}\mathbf{S}\sum_{n=1}^{N}\boldsymbol{\phi}(x_n)t_n \qquad s^2(x) = \beta^{-1} + \boldsymbol{\phi}(x)^{\mathrm{T}}\mathbf{S}\boldsymbol{\phi}(x)$$

$$\mathbf{S}^{-1} = \alpha\mathbf{I} + \beta\sum_{n=1}^{N}\boldsymbol{\phi}(x_n)\boldsymbol{\phi}(x_n)^{\mathrm{T}} \qquad \boldsymbol{\phi}(x_n) = \left(x_n^0,\ldots,x_n^M\right)^{\mathrm{T}}$$

Bayesian Predictive Distribution

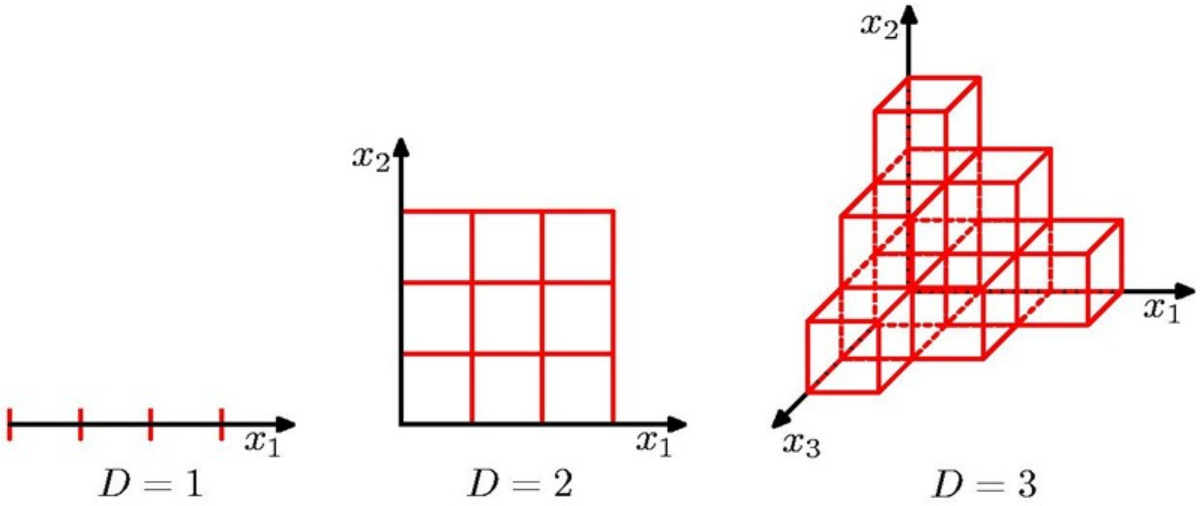
$$p(t|x, \mathbf{x}, \mathbf{t}) = \mathcal{N}\left(t|m(x), s^2(x)\right)$$



Model Selection

Cross-Validation

Curse of Dimensionality

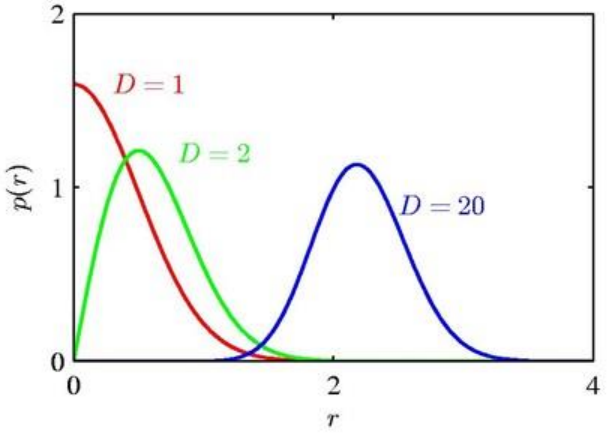


Curse of Dimensionality

Polynomial curve fitting, M = 3

$$y(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{i=1}^{D} w_i x_i + \sum_{i=1}^{D}\sum_{j=1}^{D} w_{ij} x_i x_j + \sum_{i=1}^{D}\sum_{j=1}^{D}\sum_{k=1}^{D} w_{ijk} x_i x_j x_k$$

Gaussian Densities in higher dimensions



Decision Theory

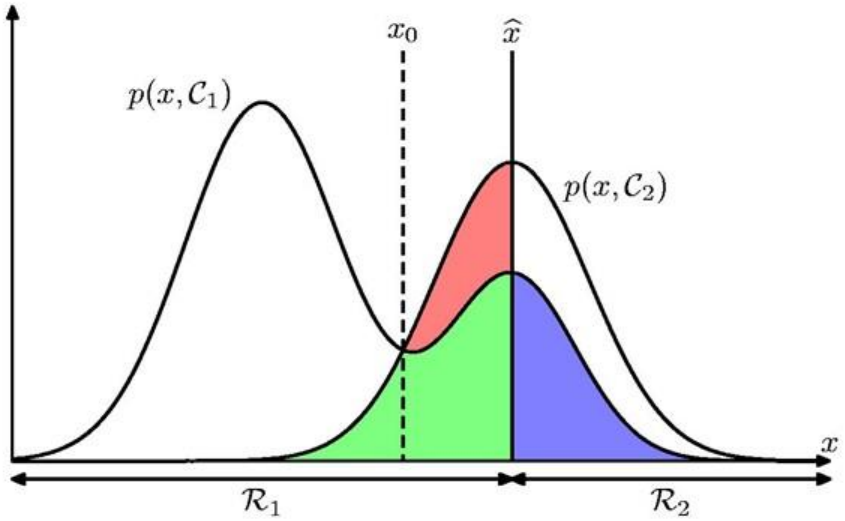
Inference step

Determine either $p(t|\mathbf{x})$ or $p(\mathbf{x}, t)$.

Decision step

For given x, determine optimal t.

Minimum Misclassification Rate



$$
\begin{aligned}
p(\text{mistake}) &= p(\mathbf{x} \in \mathcal{R}_1, \mathcal{C}_2) + p(\mathbf{x} \in \mathcal{R}_2, \mathcal{C}_1) \\
&= \int_{\mathcal{R}_1} p(\mathbf{x}, \mathcal{C}_2)\, \mathrm{d}\mathbf{x} + \int_{\mathcal{R}_2} p(\mathbf{x}, \mathcal{C}_1)\, \mathrm{d}\mathbf{x}.
\end{aligned}
$$

Minimum Expected Loss

Example: classify medical images as 'cancer' or 'normal'

		Decision: cancer	Decision: normal
Truth	cancer	0	1000
Truth	normal	1	0

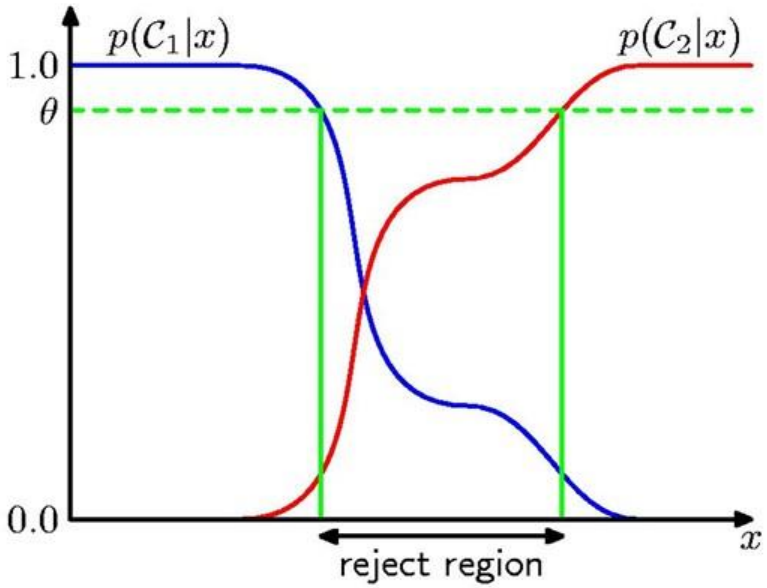
Minimum Expected Loss

$$\mathbb{E}[L] = \sum_k \sum_j \int_{\mathcal{R}_j} L_{kj} p(\mathbf{x}, \mathcal{C}_k) \, \mathrm{d}\mathbf{x}$$

Regions $\mathcal{R}_j$ are chosen to minimize

$$\mathbb{E}[L] = \sum_k L_{kj} p(\mathcal{C}_k|\mathbf{x})$$

Reject Option



Why Separate Inference and Decision?

- Minimizing risk (loss matrix may change over time)
- Reject option
- Unbalanced class priors
- Combining models

Decision Theory for Regression

Inference step

Determine $p(\mathbf{x}, t)$.

Decision step

For given x, make optimal prediction, y(x), for t.

Loss function: $\mathbb{E}[L] = \iint L(t, y(\mathbf{x}))p(\mathbf{x}, t)\,\mathrm{d}\mathbf{x}\,\mathrm{d}t$

The Squared Loss Function

$$\mathbb{E}[L] = \iint \{y(\mathbf{x}) - t\}^2 p(\mathbf{x}, t)\, \mathrm{d}\mathbf{x}\, \mathrm{d}t$$

$$\begin{aligned}\{y(\mathbf{x}) - t\}^2 &= \{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}] + \mathbb{E}[t|\mathbf{x}] - t\}^2 \\ &= \{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\}^2 + 2\{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\}\{\mathbb{E}[t|\mathbf{x}] - t\} + \{\mathbb{E}[t|\mathbf{x}] - t\}^2\end{aligned}$$

$$\mathbb{E}[L] = \int \{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\}^2 p(\mathbf{x})\, \mathrm{d}\mathbf{x} + \int \mathrm{var}\,[t|\mathbf{x}]\, p(\mathbf{x})\, \mathrm{d}\mathbf{x}$$

$$y(\mathbf{x}) = \mathbb{E}[t|\mathbf{x}]$$

Generative vs Discriminative

Generative approach:

Model $p(t, \mathbf{x}) = p(\mathbf{x}|t)p(t)$

Use Bayes' theorem $p(t|\mathbf{x}) = \dfrac{p(\mathbf{x}|t)p(t)}{p(\mathbf{x})}$

Discriminative approach:

Model $p(t|\mathbf{x})$ directly

Entropy

$$\mathrm{H}[x] = -\sum_{x} p(x) \log_2 p(x)$$

Important quantity in

- coding theory
- statistical physics
- machine learning

Entropy

Coding theory: x discrete with 8 possible states; how many bits to transmit the state of x?

All states equally likely

$$\mathrm{H}[x] = -8 \times \frac{1}{8} \log_2 \frac{1}{8} = 3 \text{ bits.}$$

Entropy

x	a	b	c	d	e	f	g	h
$p(x)$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{64}$	$\frac{1}{64}$	$\frac{1}{64}$	$\frac{1}{64}$
code	0	10	110	1110	111100	111101	111110	111111

$$\begin{aligned} \mathrm{H}[x] &= -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{4}\log_2\frac{1}{4} - \frac{1}{8}\log_2\frac{1}{8} - \frac{1}{16}\log_2\frac{1}{16} - \frac{4}{64}\log_2\frac{1}{64} \\ &= 2 \text{ bits} \end{aligned}$$

$$\begin{aligned} \text{average code length} &= \frac{1}{2}\times 1 + \frac{1}{4}\times 2 + \frac{1}{8}\times 3 + \frac{1}{16}\times 4 + 4\times\frac{1}{64}\times 6 \\ &= 2 \text{ bits} \end{aligned}$$

Entropy

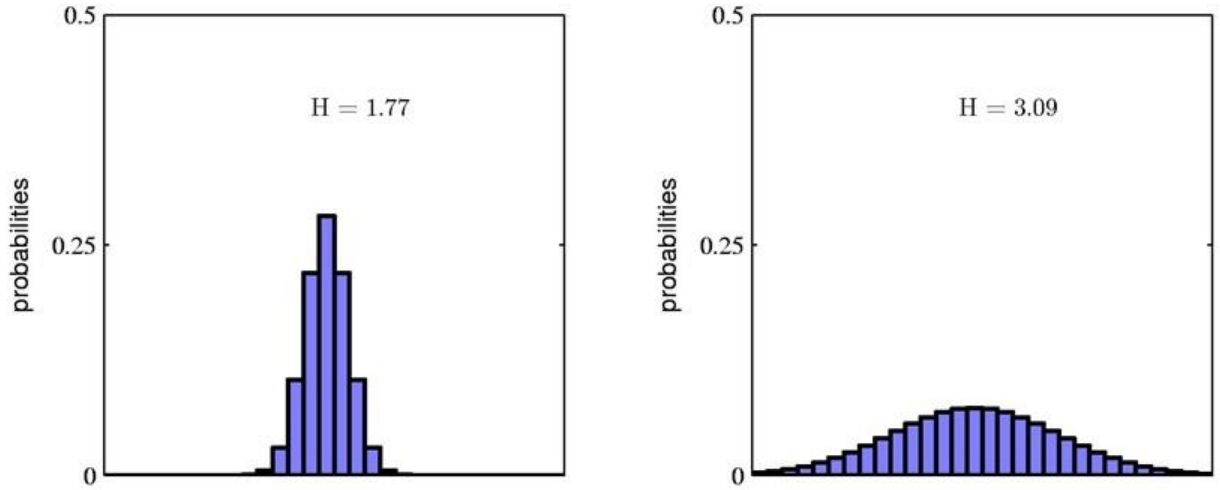
In how many ways can N identical objects be allocated M bins?

$$W = \frac{N!}{\prod_i n_i!}$$

$$\mathrm{H} = \frac{1}{N} \ln W \simeq - \lim_{N \to \infty} \sum_i \left(\frac{n_i}{N} \right) \ln \left(\frac{n_i}{N} \right) = - \sum_i p_i \ln p_i$$

Entropy maximized when $\forall i : p_i = \frac{1}{M}$

Entropy



Differential Entropy

Put bins of width ¢ along the real line

$$\lim_{\Delta \to 0} \left\{ -\sum_i p(x_i) \Delta \ln p(x_i) \right\} = -\int p(x) \ln p(x) \, \mathrm{d}x$$

Differential entropy maximized (for fixed σ^2) when

$$p(x) = \mathcal{N}(x|\mu, \sigma^2)$$

in which case

$$\mathrm{H}[x] = \frac{1}{2} \left\{ 1 + \ln(2\pi\sigma^2) \right\}.$$

Conditional Entropy

$$\mathrm{H}[\mathbf{y}|\mathbf{x}] = -\iint p(\mathbf{y},\mathbf{x}) \ln p(\mathbf{y}|\mathbf{x}) \, \mathrm{d}\mathbf{y} \, \mathrm{d}\mathbf{x}$$

$$\mathrm{H}[\mathbf{x},\mathbf{y}] = \mathrm{H}[\mathbf{y}|\mathbf{x}] + \mathrm{H}[\mathbf{x}]$$

The Kullback-Leibler Divergence

$$\begin{aligned}\mathrm{KL}(p\|q) &= -\int p(\mathbf{x}) \ln q(\mathbf{x})\,\mathrm{d}\mathbf{x} - \left(-\int p(\mathbf{x}) \ln p(\mathbf{x})\,\mathrm{d}\mathbf{x}\right) \\ &= -\int p(\mathbf{x}) \ln \left\{\frac{q(\mathbf{x})}{p(\mathbf{x})}\right\} \mathrm{d}\mathbf{x}\end{aligned}$$

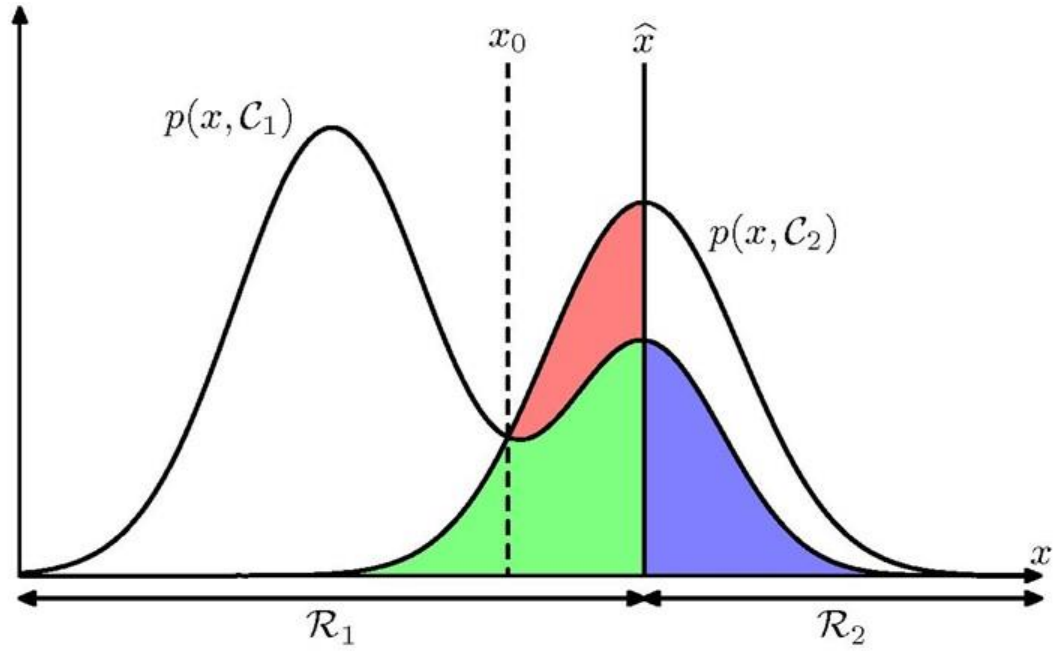
$$\mathrm{KL}(p\|q) \simeq \frac{1}{N}\sum_{n=1}^{N} \{-\ln q(\mathbf{x}_n|\boldsymbol{\theta}) + \ln p(\mathbf{x}_n)\}$$

$$\mathrm{KL}(p\|q) \geqslant 0 \qquad\qquad \mathrm{KL}(p\|q) \not\equiv \mathrm{KL}(q\|p)$$

Mutual Information

$$\begin{aligned} \mathrm{I}[\mathbf{x},\mathbf{y}] &\equiv \mathrm{KL}(p(\mathbf{x},\mathbf{y})\|p(\mathbf{x})p(\mathbf{y})) \\ &= -\iint p(\mathbf{x},\mathbf{y}) \ln\left(\frac{p(\mathbf{x})p(\mathbf{y})}{p(\mathbf{x},\mathbf{y})}\right) \mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{y} \end{aligned}$$

$$\mathrm{I}[\mathbf{x},\mathbf{y}] = \mathrm{H}[\mathbf{x}] - \mathrm{H}[\mathbf{x}|\mathbf{y}] = \mathrm{H}[\mathbf{y}] - \mathrm{H}[\mathbf{y}|\mathbf{x}]$$


x_0
$\widehat{x}$
$p(x, \mathcal{C}_1)$
$p(x, \mathcal{C}_2)$
x
$\mathcal{R}_1$
$\mathcal{R}_2$