Machine Learning

Dr. Mehran Safayani

safayani@iut.ac.ir

safayani.iut.ac.ir

https://www.aparat.com/mehran.safayani

https://github.com/safayani/machine_learning_course

Department of Electrical and computer engineering, Isfahan university of technology, Isfahan, Iran

MLE (Maximum Likelihood Estimation)

یادآوری

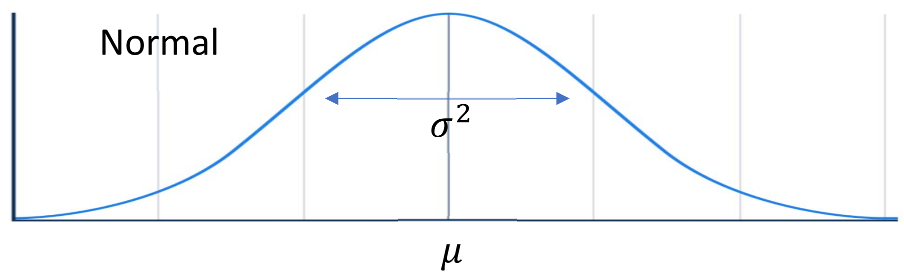
توزیع گاوسی:

$$P(y \mid \mu , \sigma^2) = N(y \mid \mu , \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left(\frac{-(y-\mu)^2}{2\sigma^2}\right)$$

اگر y به صورت بردار باشد دارای میانگین $\vec{\mu}$ و کوواریانس Σ است:

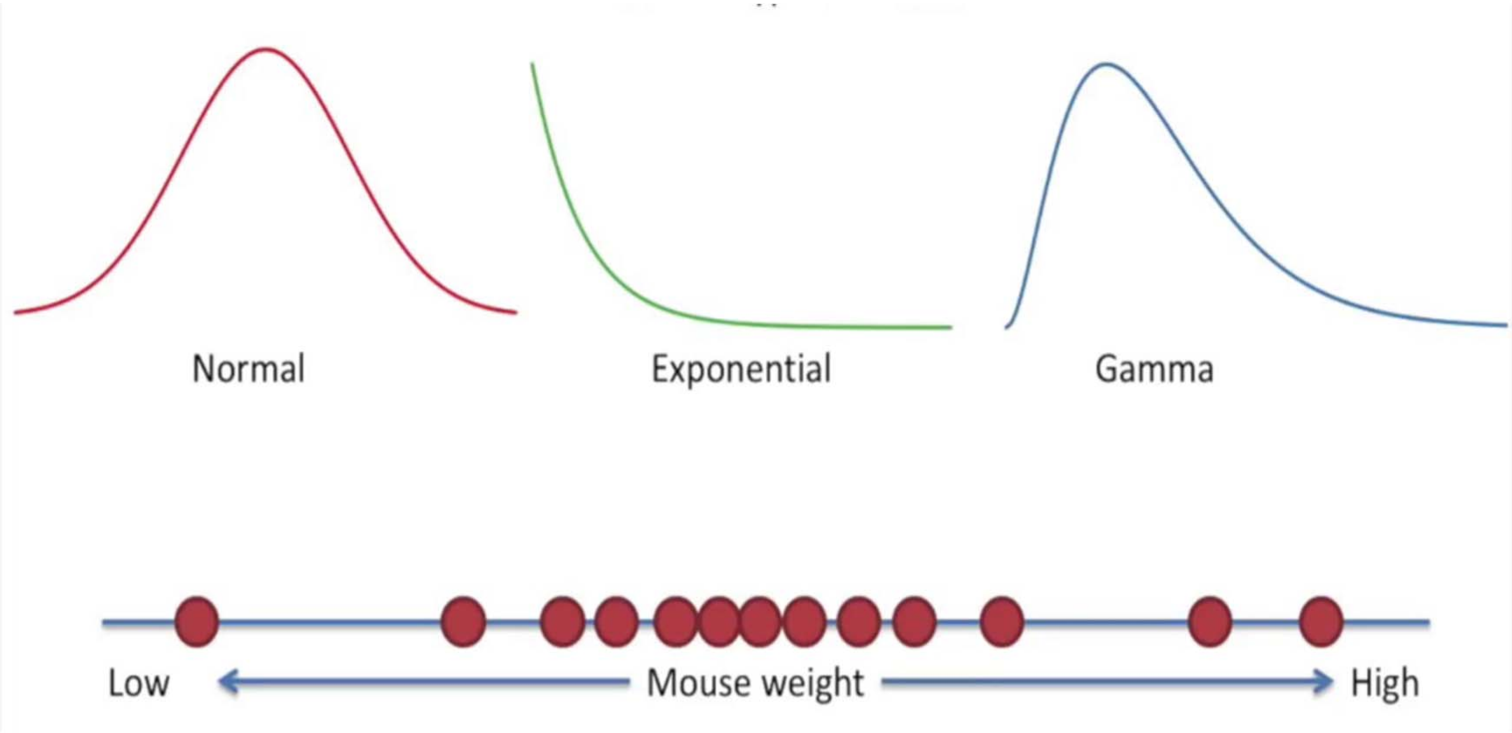
$$N(y \mid \mu , \Sigma) = \frac{1}{\sqrt{(4\pi)^D \det(\Sigma)}} \exp\left(\frac{-1}{2} (y-\mu)^T \, \Sigma^{-1} \, (y-\mu)\right)$$

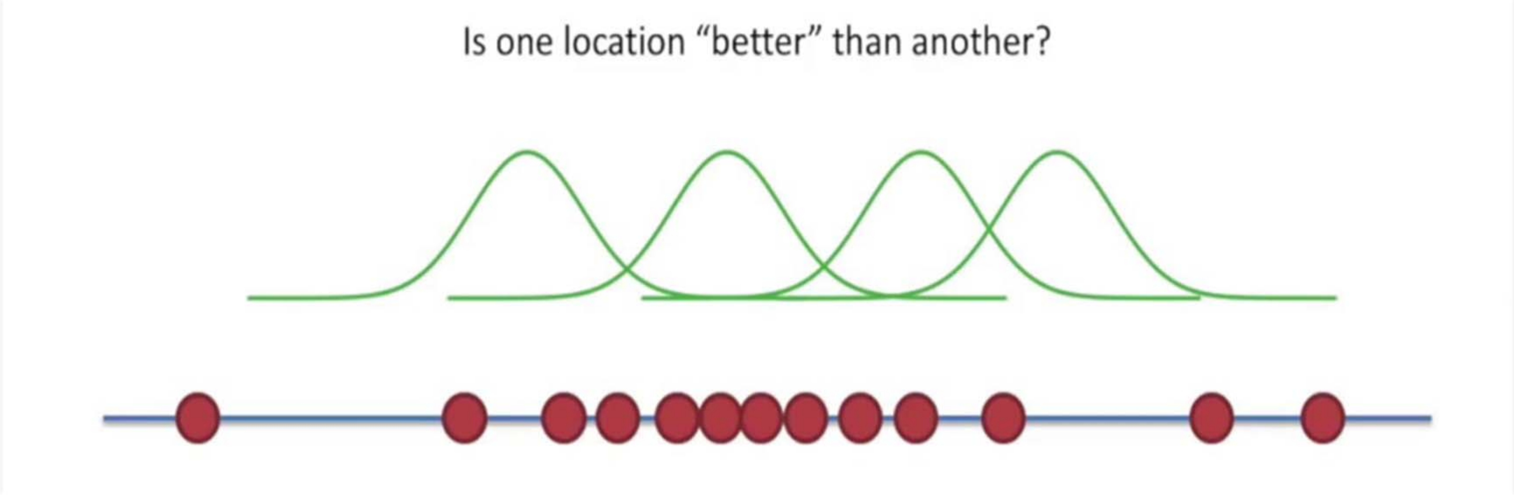
x , y دو متغیر تصادفی مستقل هستند اگر:



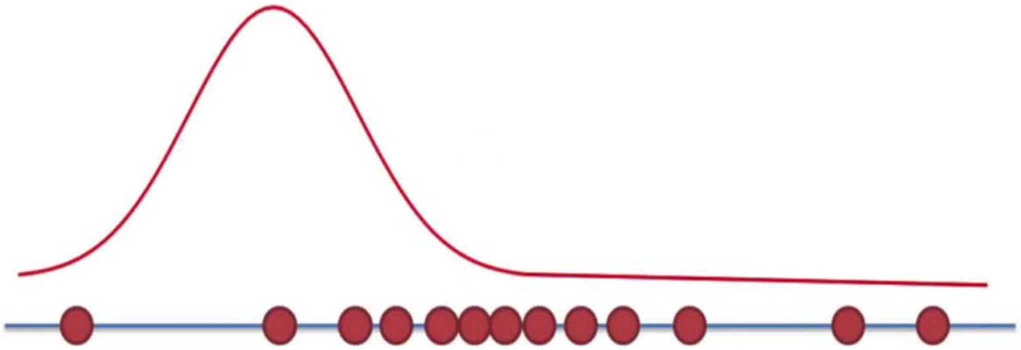
$$P(x , y) = P(x) \, P(y)$$

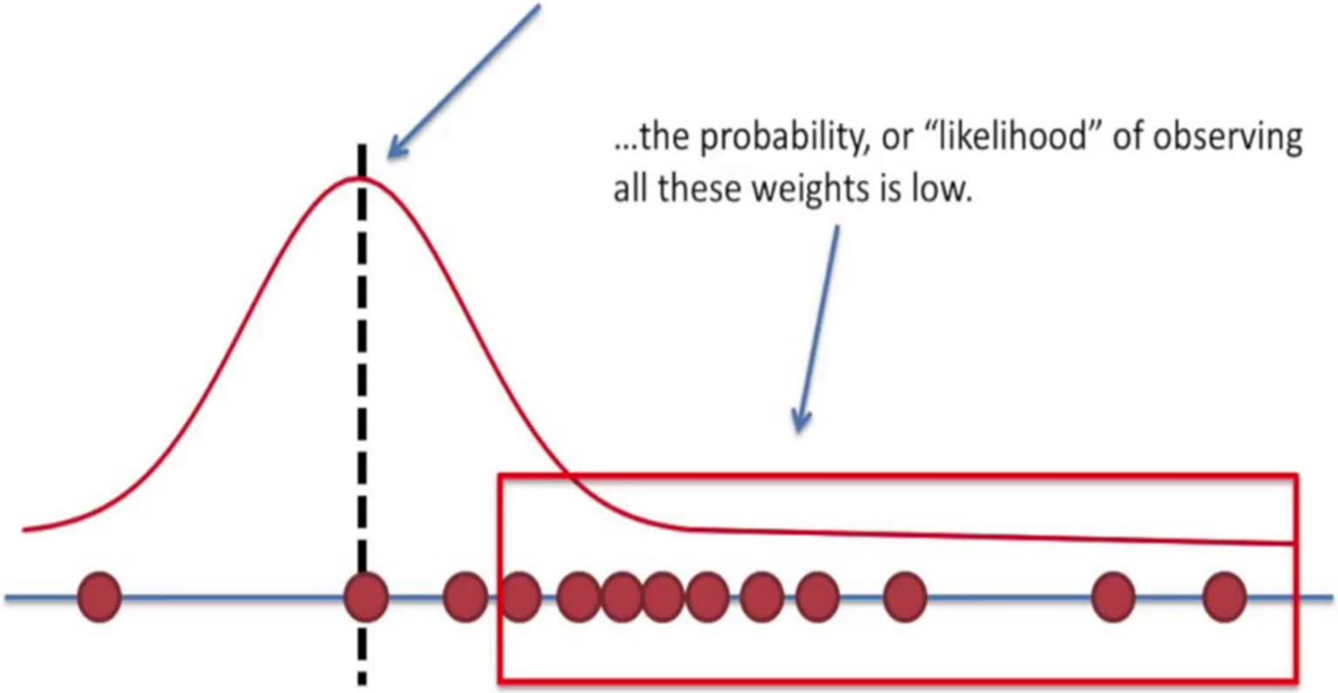
MLE example





Is one location “better” than another?

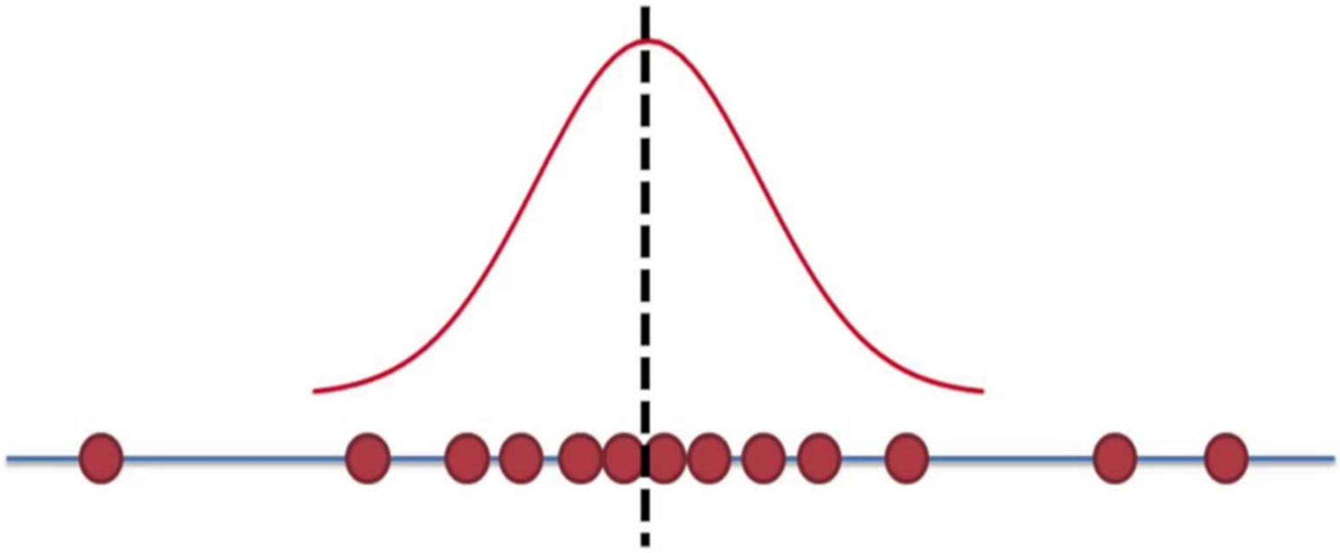


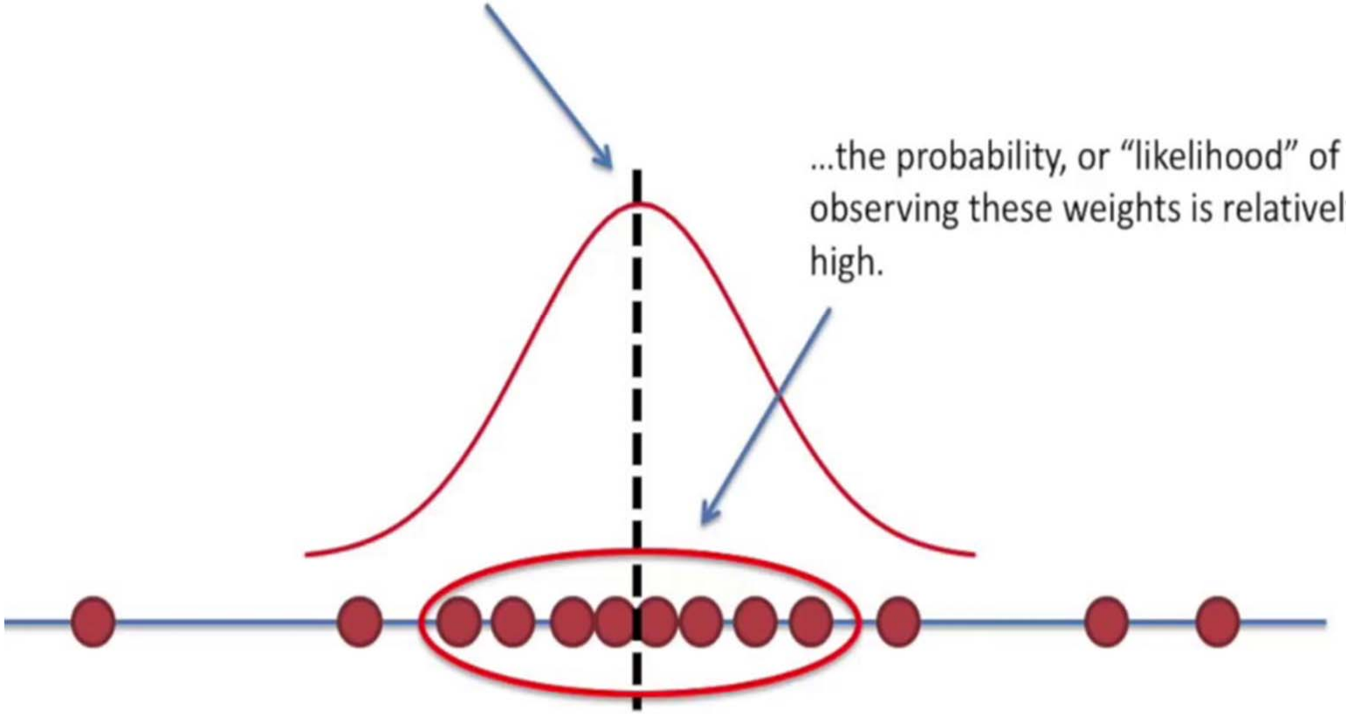


According to a normal distribution with a mean value over here…

…the probability, or "likelihood" of observing all these weights is low.

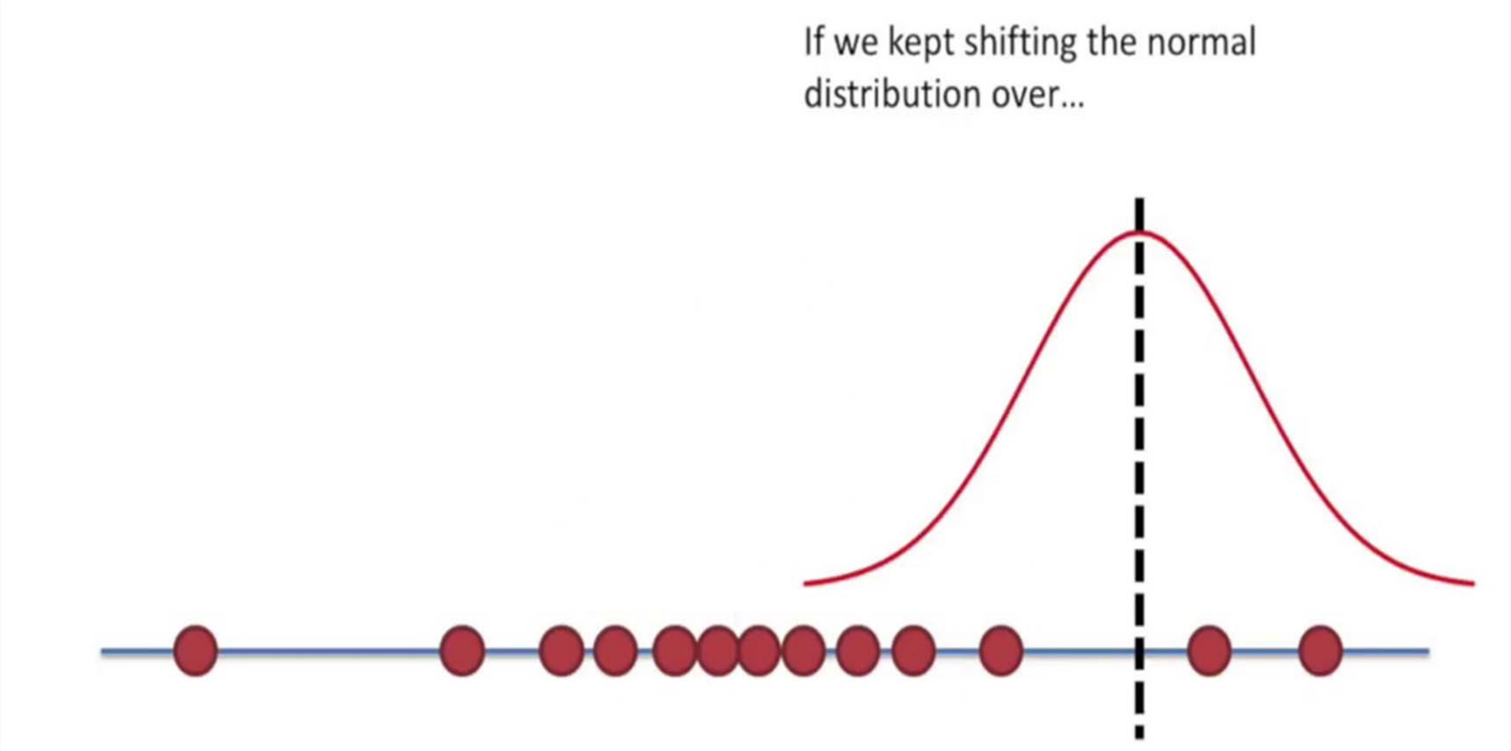
What if we shifted the normal distribution over, so that its mean was the same as the average weight?





According to a normal distribution with a mean value here…

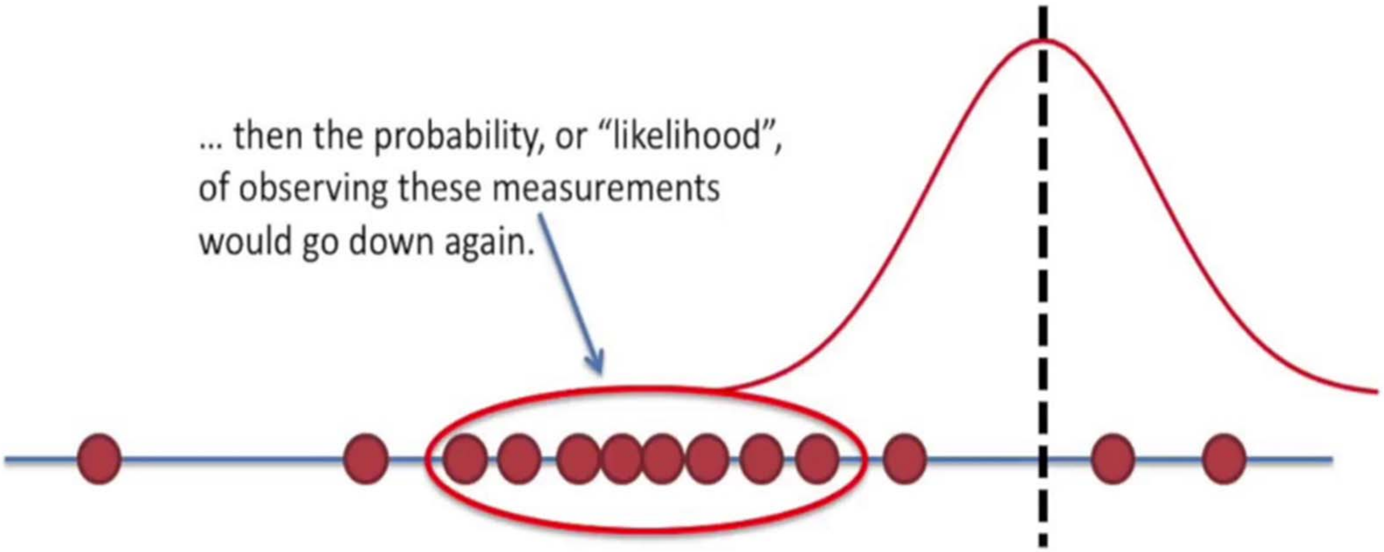
…the probability, or "likelihood" of observing these weights is relatively high.



If we kept shifting the normal distribution over...

If we kept shifting the normal distribution over…

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



می خواهیم تابع چگالی P(x) را به دست آوریم.

یک تابع چگالی احتمالی پارامتری برای P(x) تعریف میکنیم.

با فرض وجود مشاهده های $X = (x^1 , \dots , x^n)$ سعی می کنیم پارامترهای مدل را بهینه کنیم تا احتمال P(x) بیشینه گردد.

$$X = (x^1 , \dots , x^n)$$

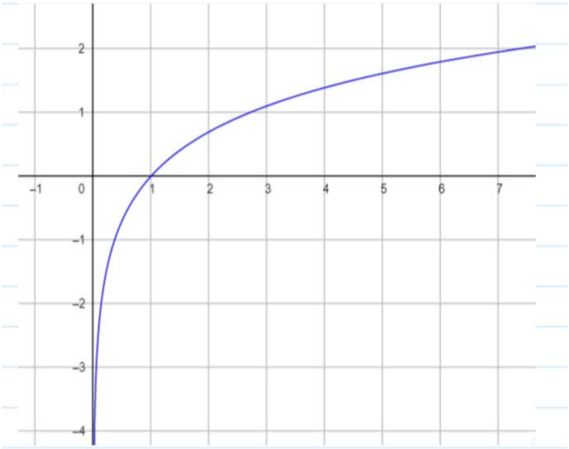
فرض می کنیم هر داده از توزیع $P(x \mid \theta)$ به صورت مستقل به دست آمده اند.

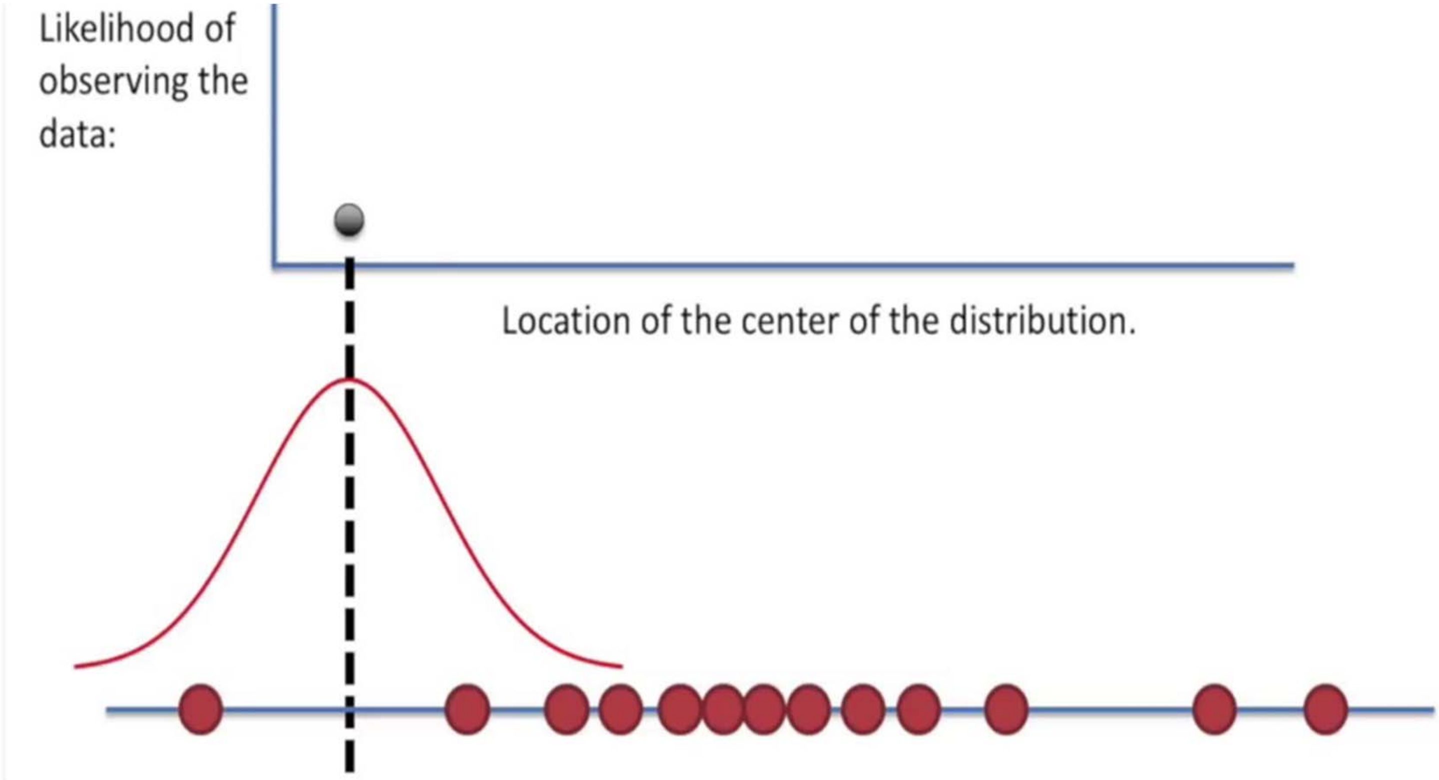
$$P(X \mid \theta) = \prod_{n=1}^{N} P(x^n|\theta) = L(\theta)$$

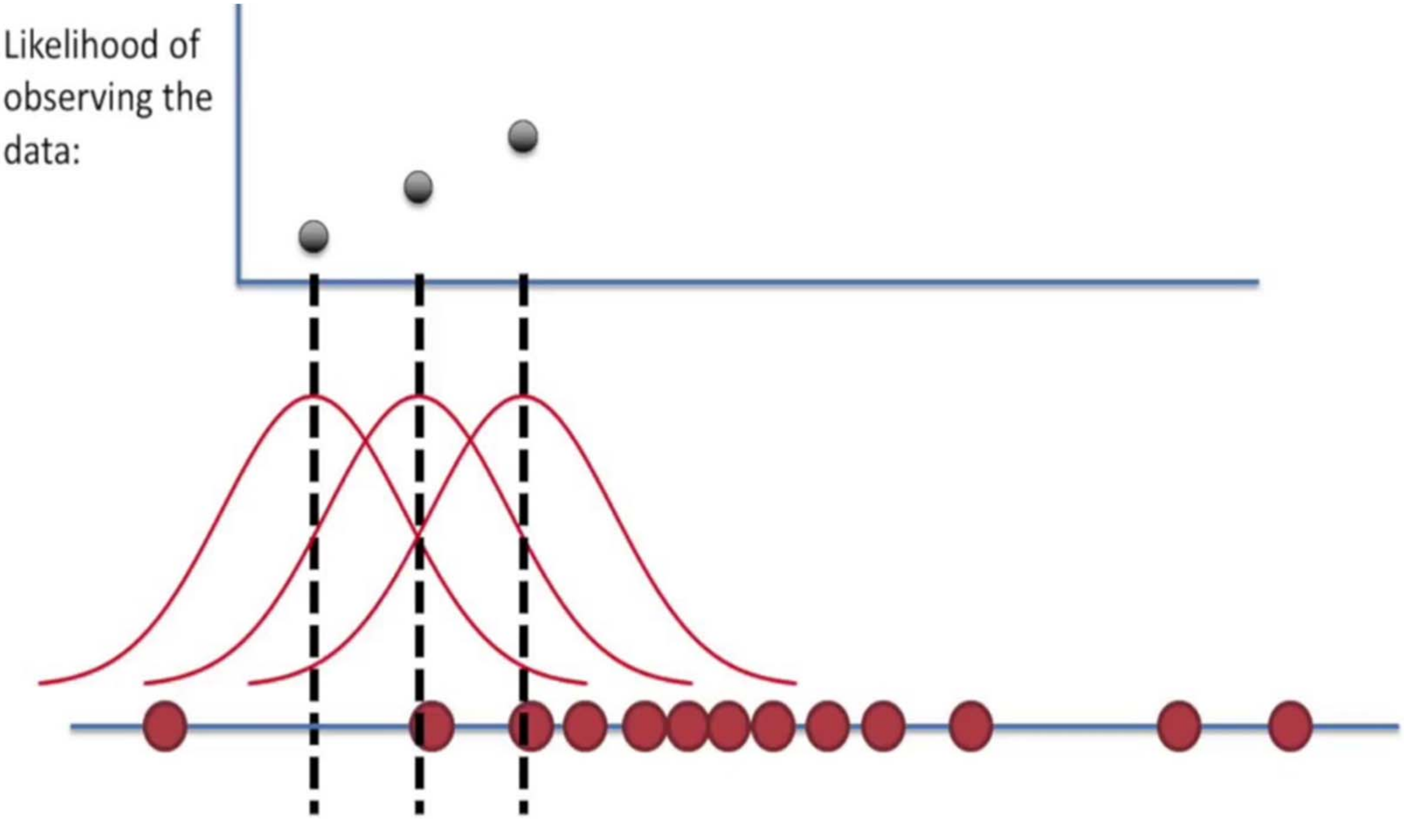
$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\ L(\theta)$$

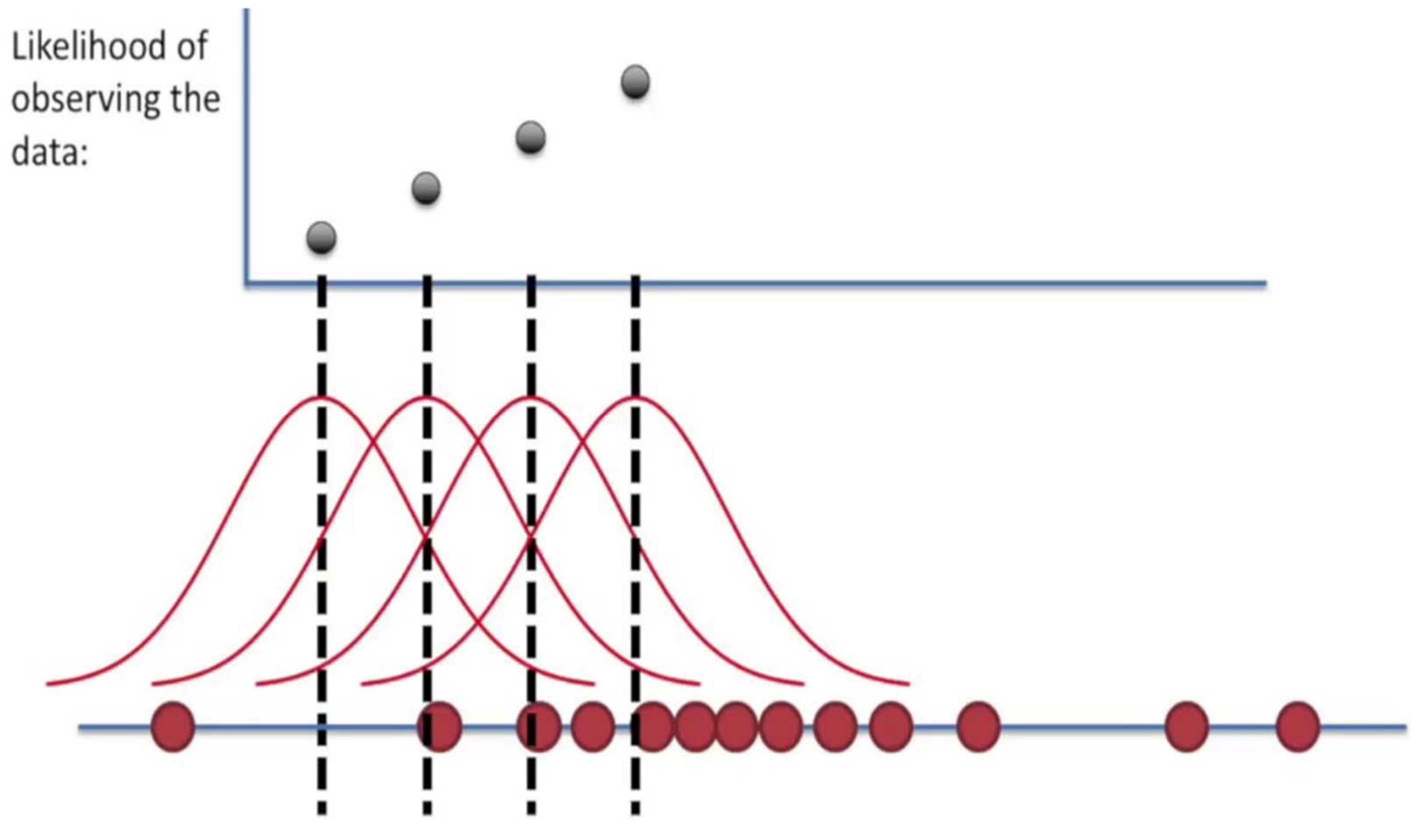
$$\text{Log } P(X \mid \theta) = \sum_{n=1}^{N} \log P(x^n \mid \theta) = \log L(\theta)$$

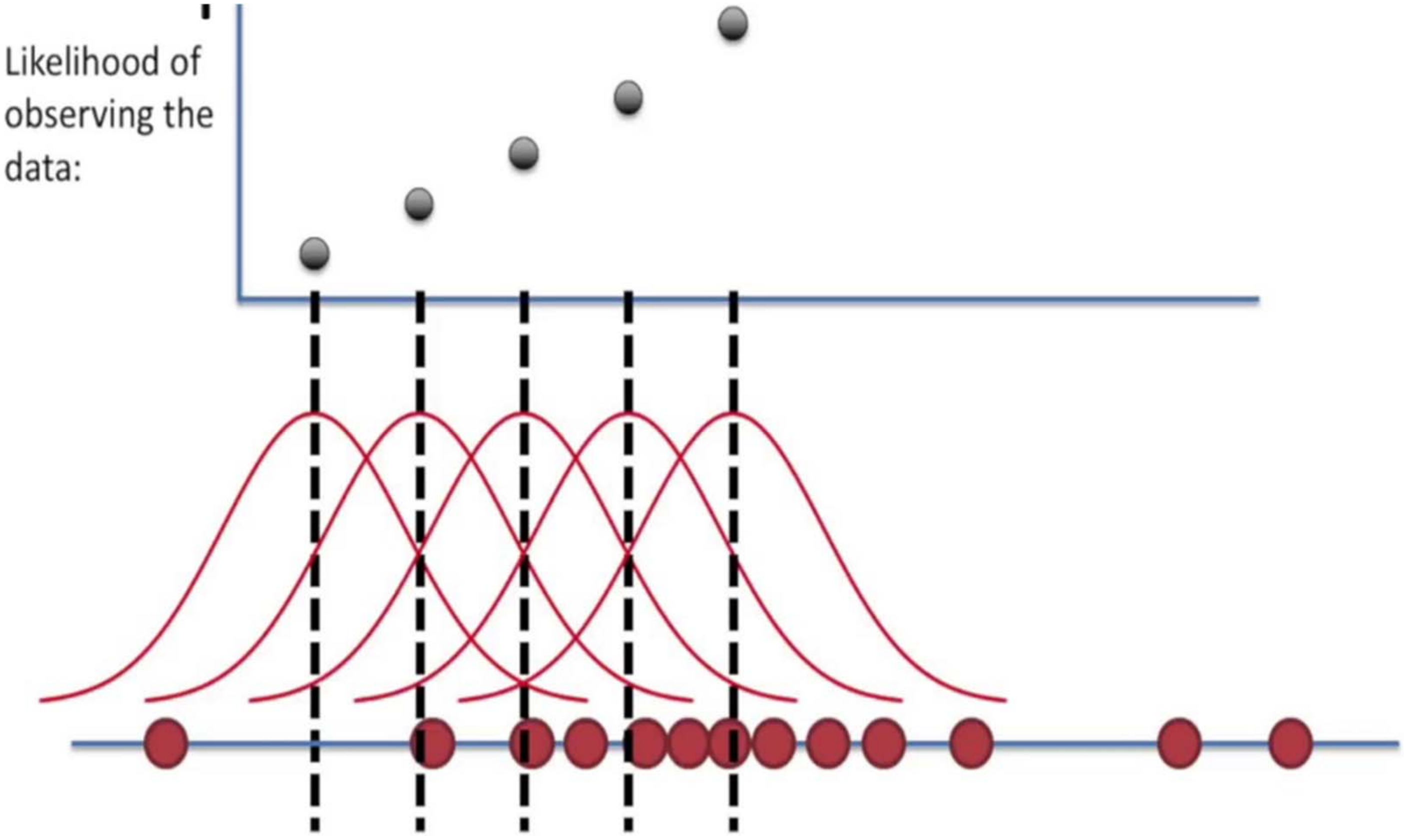
$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\ L(\theta) = \underset{\theta}{\operatorname{argmax}}\ \log L(\theta)$$

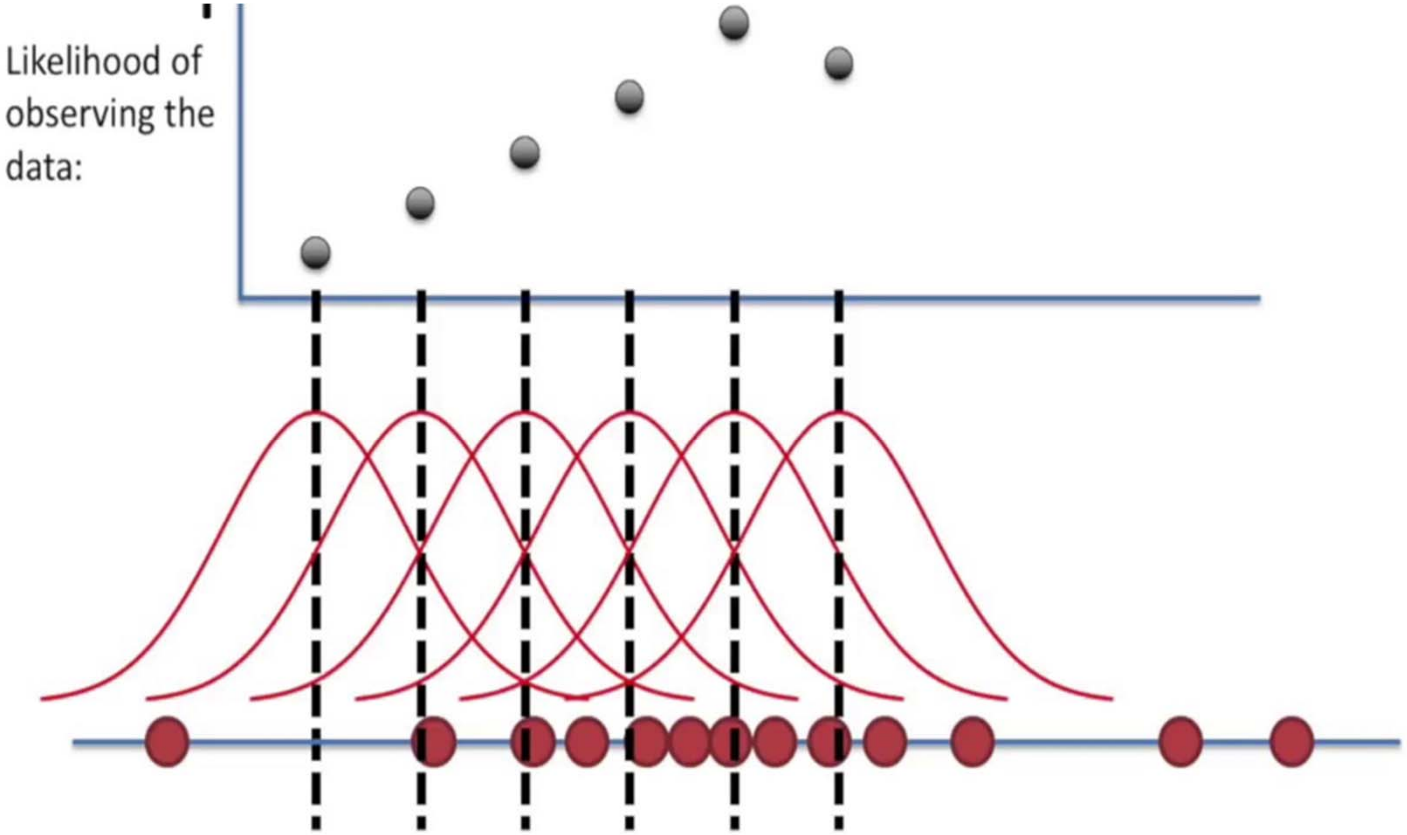


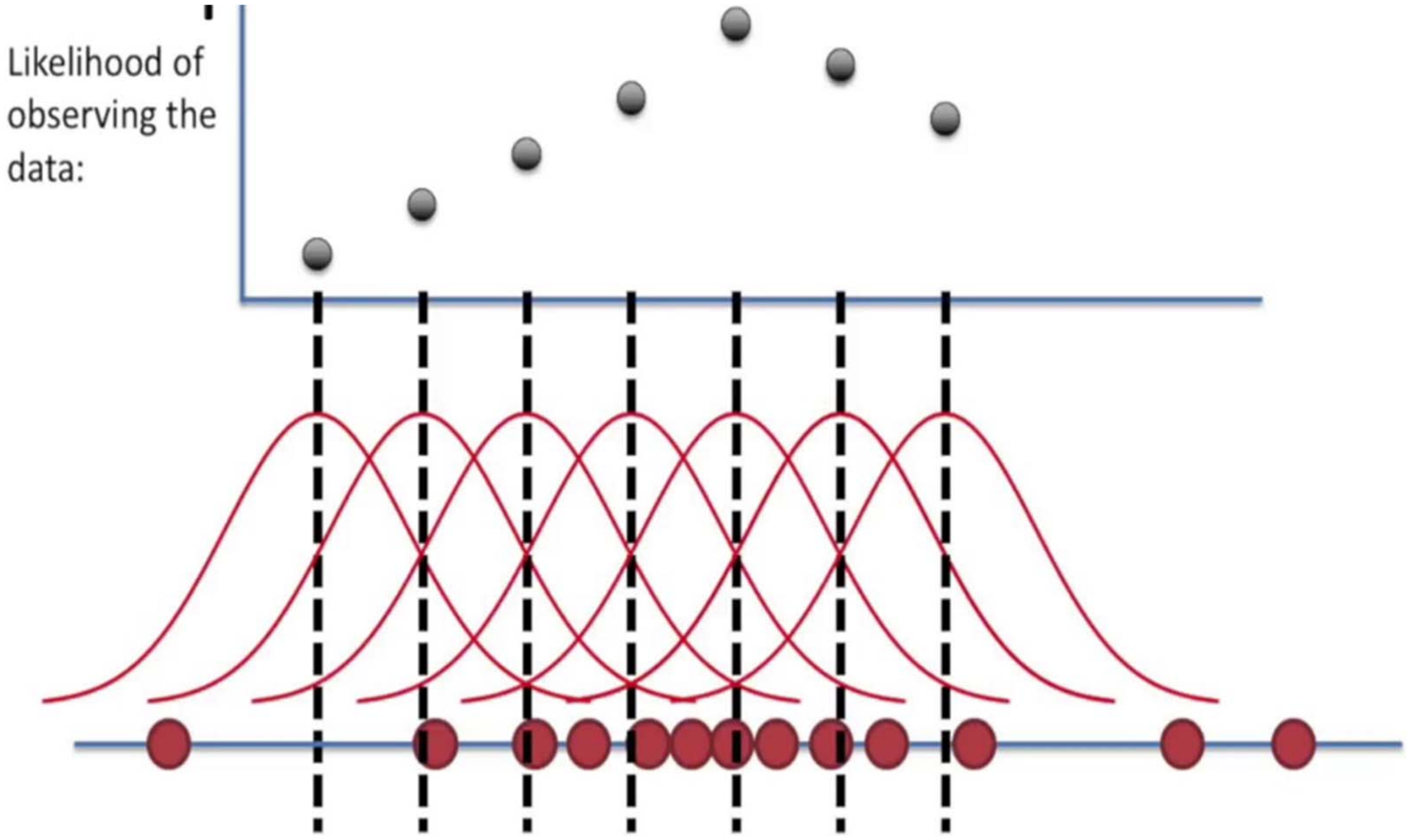

Likelihood of observing the data:
Location of the center of the distribution.

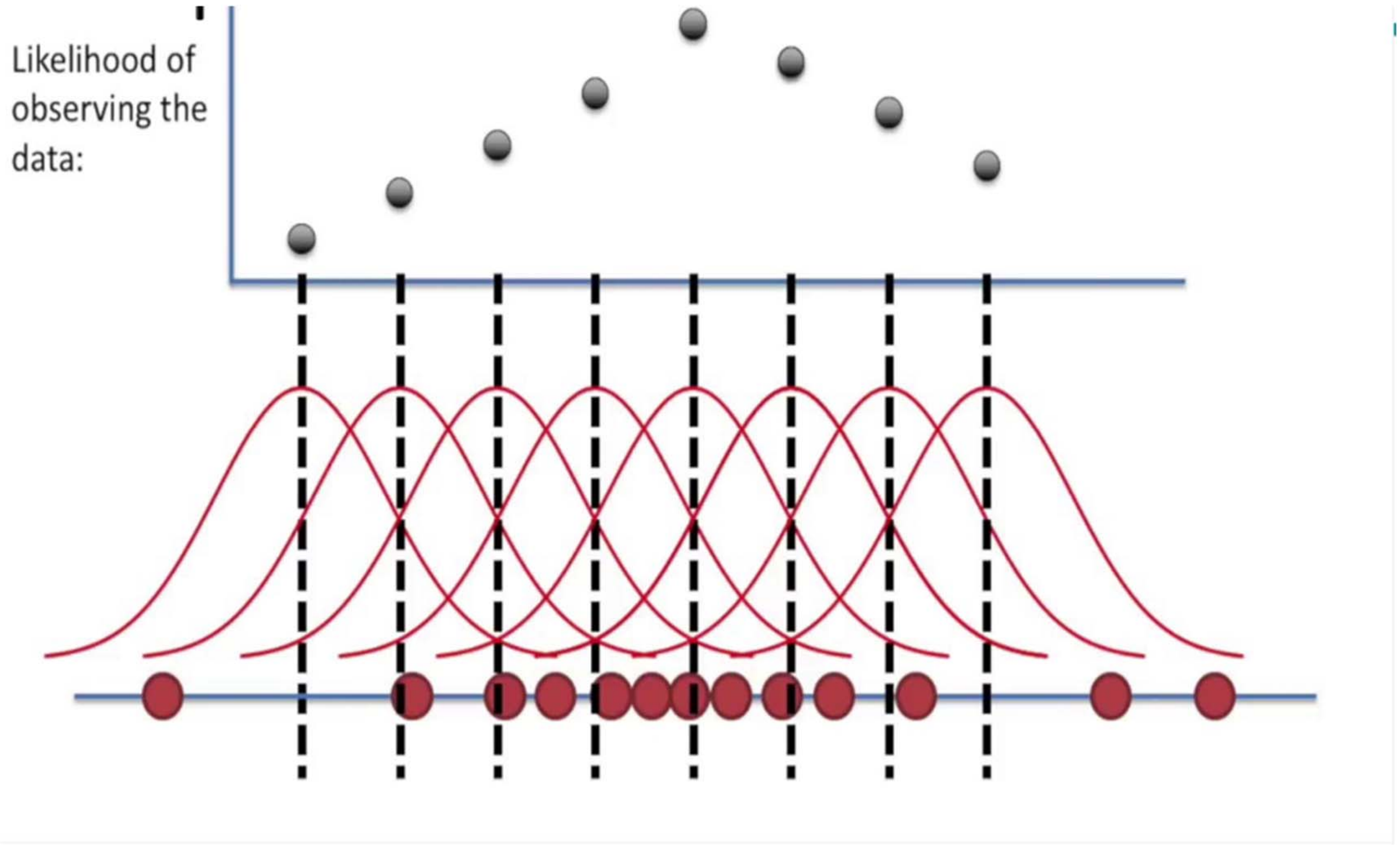

Likelihood of observing the data:


Likelihood of observing the data:

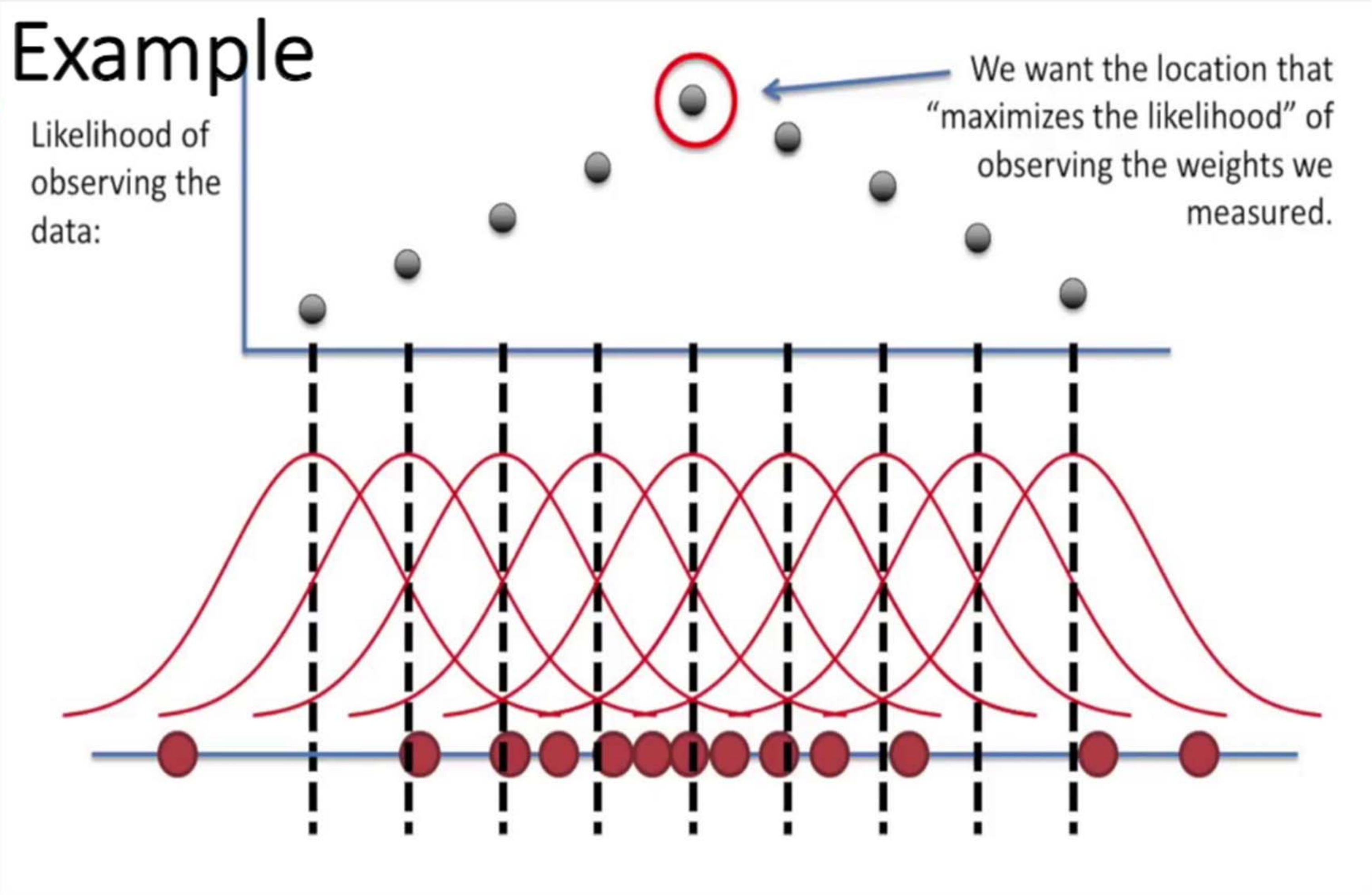

Likelihood of observing the data:


Likelihood of observing the data:


Likelihood of observing the data:


Likelihood of observing the data:

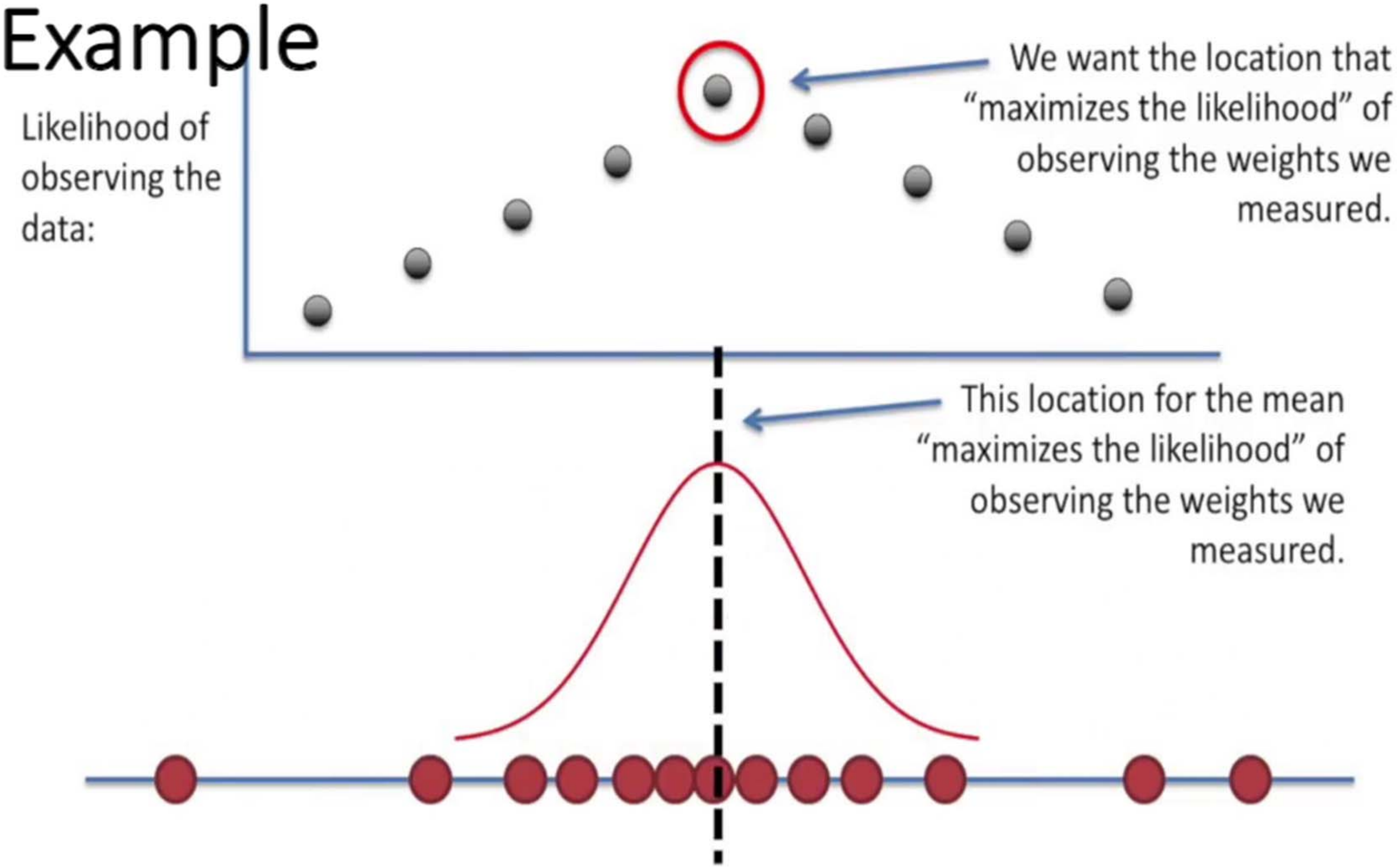
MLE Example



Likelihood of observing the data:

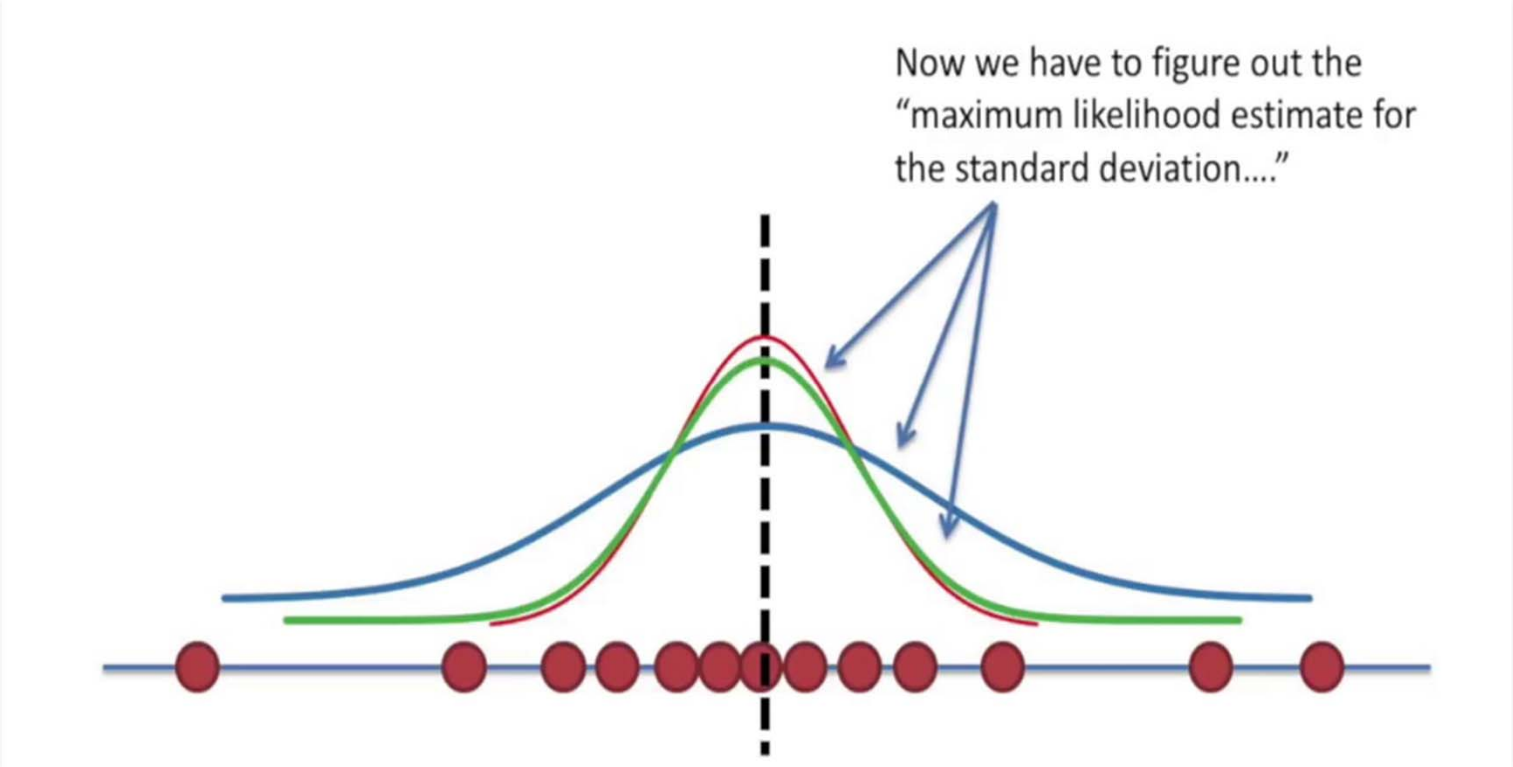
We want the location that "maximizes the likelihood" of observing the weights we measured.

MLE Example

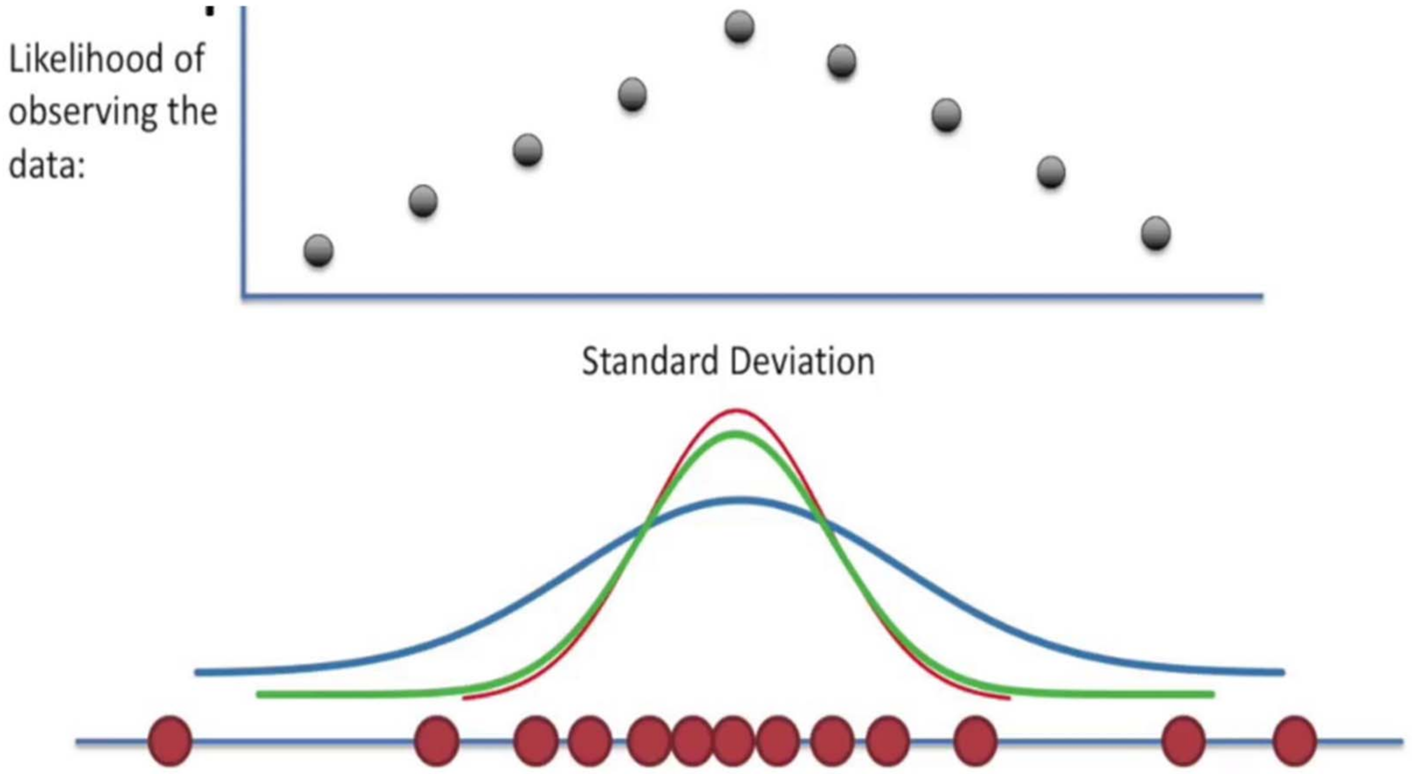


We want the location that "maximizes the likelihood" of observing the weights we measured.

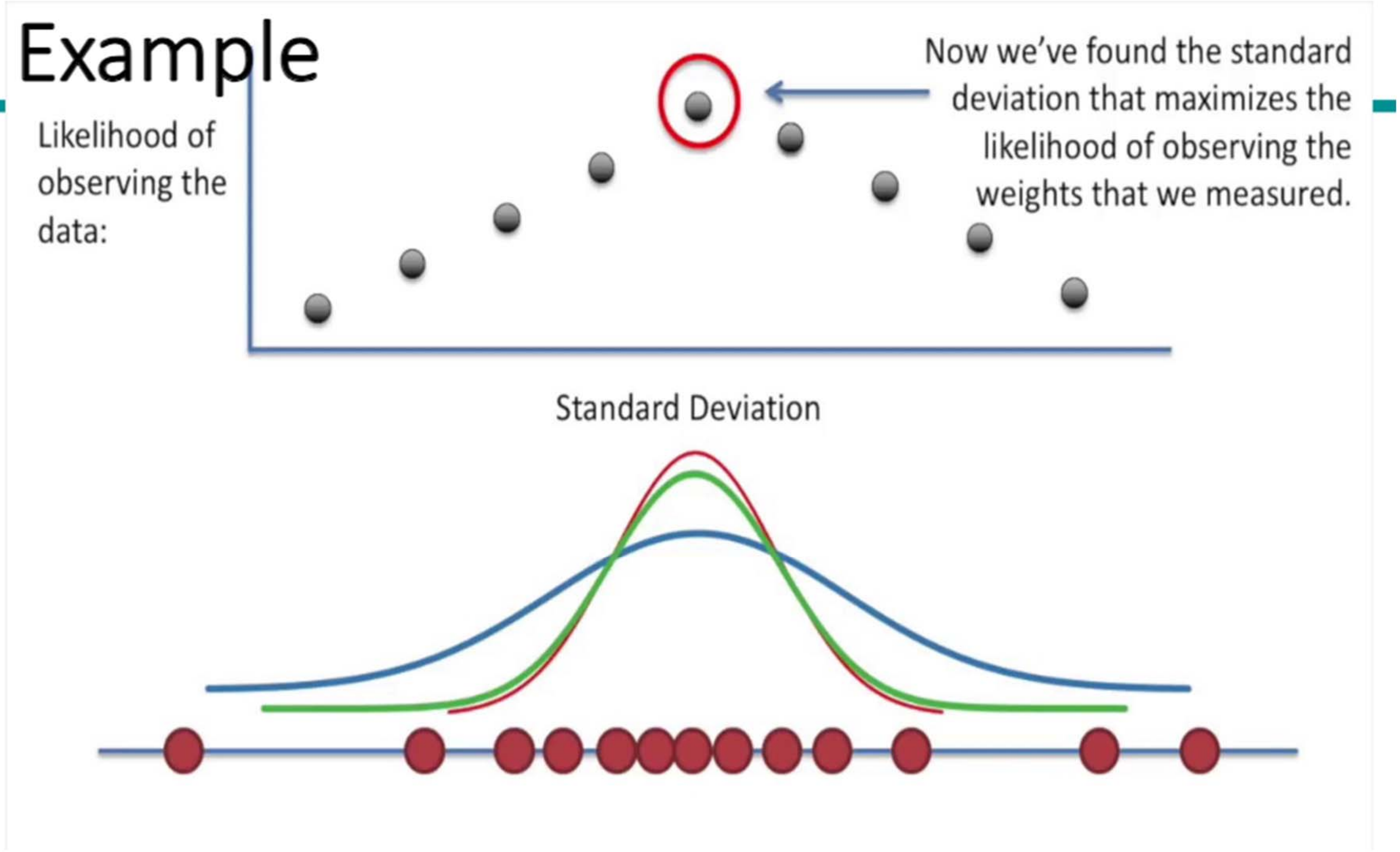
This location for the mean "maximizes the likelihood" of observing the weights we measured.



Now we have to figure out the “maximum likelihood estimate for the standard deviation....”

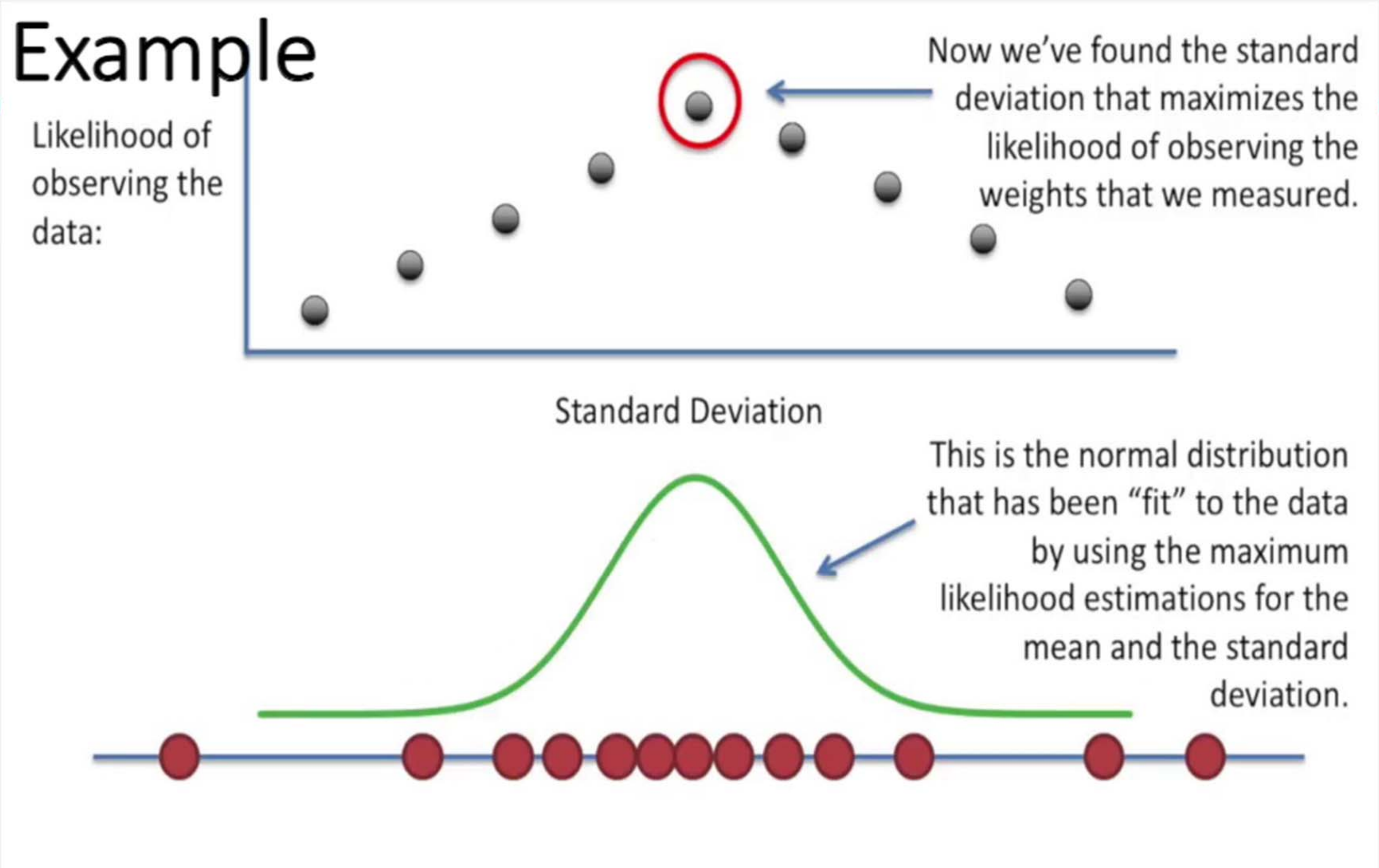

Likelihood of observing the data:
Standard Deviation

MLE Example



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

MLE Example



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

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.

Calculating the MLE

- Probability of observing a single data point *x*

$$P(x;\underset{\uparrow}{\mu},\underset{\uparrow}{\sigma}) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Parameters

- Example: $$P(9, 9.5, 11; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(9-\mu)^2}{2\sigma^2}\right) \times \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(9.5-\mu)^2}{2\sigma^2}\right) \times \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(11-\mu)^2}{2\sigma^2}\right)$$

The Log likelihood

- Maximum is found by differentiation, i.e., find the derivative of the function w.r.t. a variable, set it to zero and find the required value.
- Since the previous expression is not easy to differentiate, we simplify the calculus considering the natural logarithm of the expression.

$$\ln(P(x;\mu,\sigma)) = \ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \frac{(9-\mu)^2}{2\sigma^2} + \ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \frac{(9.5-\mu)^2}{2\sigma^2} + \ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \frac{(11-\mu)^2}{2\sigma^2}$$

$$\ln(P(x;\mu,\sigma)) = -3\ln(\sigma) - \frac{3}{2}\ln(2\pi) - \frac{1}{2\sigma^2}\left[(9-\mu)^2 + (9.5-\mu)^2 + (11-\mu)^2\right]$$

Derivation with respect to mu

- This expression can be easily differentiated to find the maximum.

$$\frac{\partial \ln(P(x;\mu,\sigma))}{\partial \mu} = \frac{1}{\sigma^2}\left[9 + 9.5 + 11 - 3\mu\right].$$

$$\mu = \frac{9 + 9.5 + 11}{3} = 9.833$$

- The same can be done for the standard deviation.

Let x_1, x_2 , … , x_{n} be a random sample from a normal distribution with unknown mean μ and variance σ^2 .

Find Maximum Likelihood estimators of mean μ and variance σ^2 .

Answer

In finding the estimators , the first thing we will do is write the probability density function as a function of θ_1 = μ and θ_2 = σ^2

$$f(x_i\ ;\theta_1\ ,\theta_2) = \frac{1}{\sqrt{\theta_2}\ \sqrt{2\pi}} \exp\ [\frac{-(x_i-\theta_1)^2}{2\theta_2}]$$

For $-\infty < \theta_1 < \infty$ and $0 < \theta_2 < \infty$. We do this so as not to cause confusion when taking the derivative of the likelihood with respect to σ^2. Now , that makes the likelihood function:

$$L(\theta_1\ ,\theta_2) = \prod_{i=1}^{n} f(x_i\ ;\theta_1\ ,\theta_2) = \theta_2^{-n/2}\ (2\pi)^{-n/2} \exp[\frac{-1}{2\theta_2} \sum_{i=1}^{n} (x_i - \theta_1)^2]$$

And therefore the log of the likelihood function:

$$\text{Log L}(\theta_1\ ,\ \theta_2) = \frac{-n}{2}\log\theta_2 - \frac{n}{2}\log(2\pi) - \frac{\sum(x_i-\theta_1)^2}{2\theta_2}$$

Now, upon taking the partial derivative of the log likelihood with respect to θ_1 , and setting to 0 , we see that a few things cancel each other out , leaving us with:

$$\frac{\partial\ \text{Log L}(\theta_1\ ,\ \theta_2)}{\partial\theta_1} = \frac{-2\ \sum(x_i-\theta_1)\ (-1)}{2\theta_2} \equiv 0$$

Now, multiplying through by θ_2 and distributing the summation , we get:

$$\sum(x_i - n\theta_1) = 0$$

Now , solving for $\theta_1$and putting on its hat we have shown that the maximum likelihood estimate of θ_1 is :

$$\hat{\theta}_1 = \hat{\mu} = \frac{\sum x_i}{n} = \bar{x}$$

Now for θ_2 taking the partial derivative of the log likelihood with respect to θ_2 , and setting to 0 , we get:

$$\frac{\partial \text{ Log L}(\theta_1 , \theta_2)}{\partial \theta_2} = \frac{-n}{2\theta_2} + \frac{\sum(x_i - \theta_1)^2}{2\theta_2^2} = 0$$

Multiplying through by $2\theta_2^2$:

$$\frac{\partial \text{ Log L}(\theta_1 , \theta_2)}{\partial \theta_2} = [\frac{-n}{2\theta_2} + \frac{\sum(x_i - \theta_1)^2}{2\theta_2^2} = 0] * 2\theta_2^2$$

We get:

$$-n\theta_2 + \sum(x_i - \theta_1)^2 = 0$$

And , solving for θ_2 , and putting on its hat , we have shown that the maximum likelihood estimate of θ_2 is:

$$\hat{\theta}_2 = \hat{\sigma}^2 = \frac{\sum(x_i - \bar{x})^2}{n}$$

ارتباط روش LS , MLE

Least square (LS)

یک تابع هزینه تعریف کردیم و با توجه به داده ها مدلی را پیدا کردیم که تابع هزینه را کمینه می کرد.

در این قسمت یک نگاه جدید داریم و می خواهیم از منظر مدل های احتمالاتی به این مسئله نگاه کنیم و به عبارتی یک تعبیر احتمالاتی از مسئله LS داشته باشیم.

یک مدل احتمالاتی برای LS

فرض کنید داده های ما توسط مدل زیر تولید می شوند:

$$y_n = x_n^T \text{ w} + \varepsilon_n$$

$\varepsilon_n \sim \text{N}(\mu , \sigma^2)$:

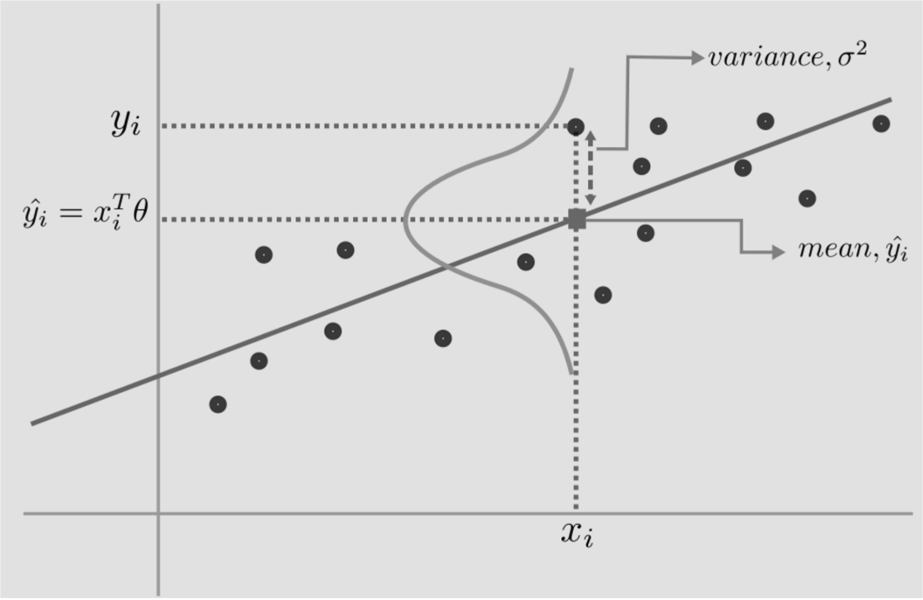
ε_n [illegible] σ^2 [illegible]

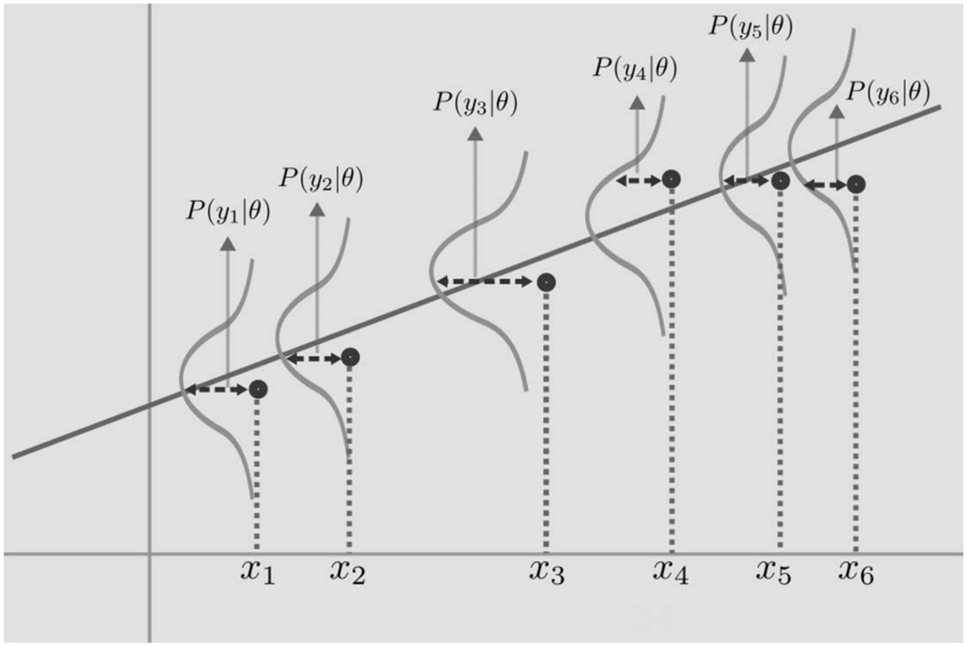
نویز با نمونه های تبدیل یافته جمع می شود و مستقل از نمونه هاست.

w: پارامترهای مدل است

$$\text{P}(y_n \mid x_n, \text{w}) = \text{N}(x_n^T \text{ w} , \sigma^2)$$

$$P(y_n \mid x_n, w) = N(x_n^T w \ , \sigma^2)$$





ادامه یک مدل احتمالاتی برای LS

به شرط N نمونه درست نمایی (Likelihood) برای داده $Y = (y_1 , y_2 ,..., y_n)$ با داشتن ورودی های X (هر سطر یک داده) و پارامترهای مدل w به صورت زیر است:

$$P(Y \mid X , w) = \prod_{n=1}^{N} P(y_n \mid x_n, w) = \prod_{n=1}^{N} N(y_n \mid x_n^T w , \sigma^2)$$

ما بایستی این Likelihood را نسبت به پارامترهای مدل w بیشینه کنیم. یعنی بهترین مدل مدلی است که این درست نمایی را بیشینه کند.

LS , log-likelihood رابطه

Log Likelihood:

L_{LL}(w) = log P(y| X , w) = $\frac{-1}{2\sigma^2} \sum_{n=1}^{N} (y_n - x_n^T \text{ w})^2$ + con

LS:

L_{MSE}(w) = $\frac{1}{2N} \sum_{n=1}^{N} (y_n - x_n^T \text{ w})^2$

$$\underset{\text{w}}{\text{argmin}}\, L_{MSE}(\text{w}) = \underset{\text{w}}{\text{argmax}}\, L_{LL}(\text{w})$$