



Machine Learning

Maximum Likelihood Estimation(MLE)

Dr. Mehran Safayani

safayani@iut.ac.ir

safayani.iut.ac.ir



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



https://github.com/safayani/machine_learning_course



Data distribution

- **Data distribution** refers to the way data points are spread or distributed across possible values in a dataset. It describes the frequency or probability of occurrence of each value or range of values in the data. Understanding data distribution is essential in statistics, data analysis, and machine learning because it provides insights into the underlying structure, patterns, and behavior of the data.

Why data distribution is important in machine learning?

- Data generation: GAN, Chat-gpt
- Anomaly detection
- Model assumptions:
 - Many machine learning algorithms make assumptions about the distribution of the data.
 - For example:
 - Linear regression assumes that the residuals are normally distributed.
 - Gaussian Naive Bayes assumes that features follow a Gaussian distribution.
 - Violating these assumptions can lead to poor model performance.

Discrete Data vs Continuous Data

- **Discrete Data** — Data that is defined by specific finite values that have no continuous space between one another. Discrete data is *counted*.
 - Example: marital status {single, married, divorced}
- **Continuous Data** — Data that exists in a continuous space and can take on any value within that space. Continuous data is *measured*.
 - Example: size of a house, {210.25 square meters}

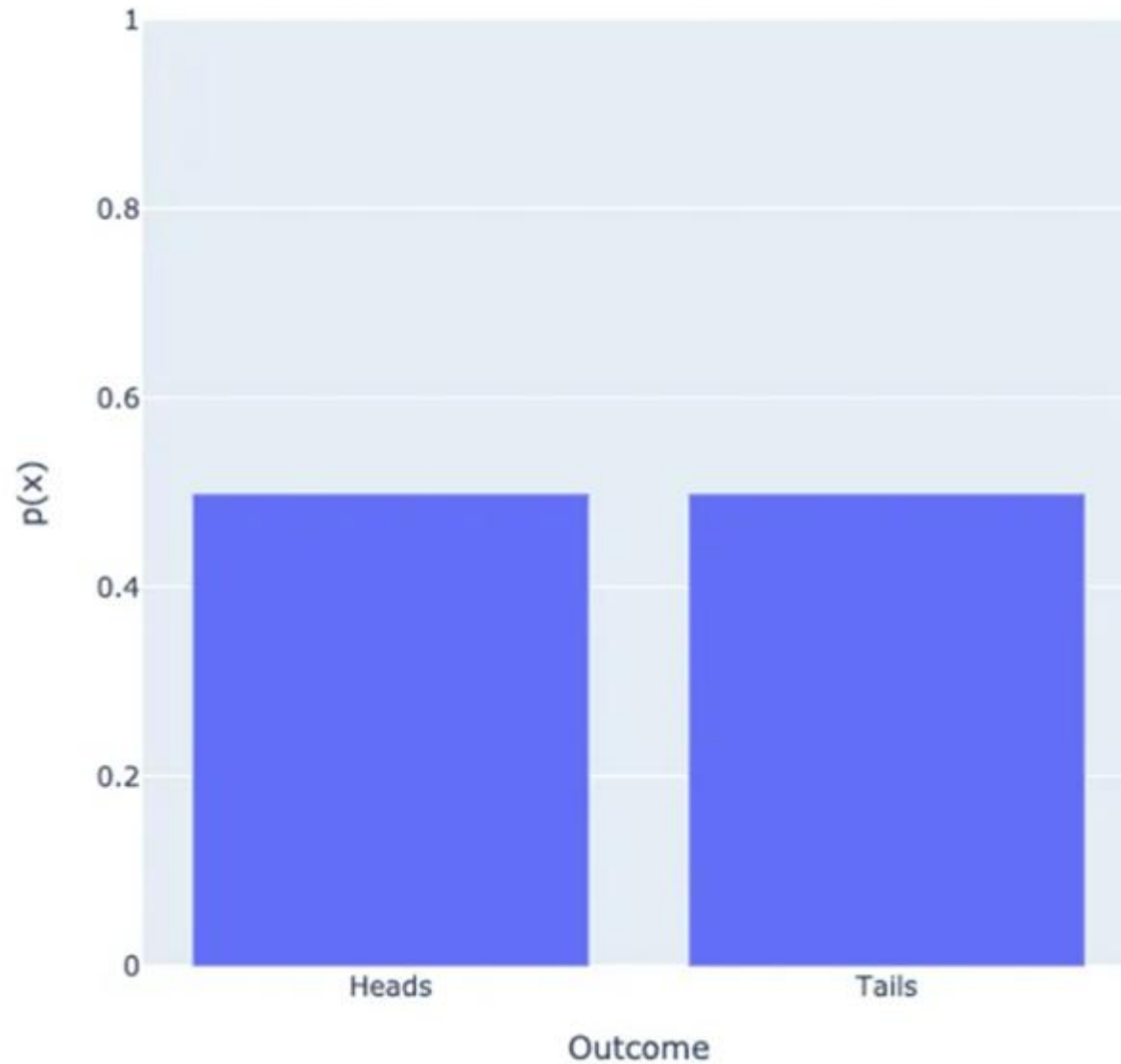
PMF vs PDF vs CDF

- ***PMF (Probability Mass Function)*** — a mathematical formula to measure the *probability of drawing a specific value from a discrete data distribution*.
- ***PDF (Probability Density Function)*** — a mathematical formula to measure the *probability density of different values across a continuous space*.

Discrete Data Distributions

- **Bernoulli Distribution**
- The Bernoulli Distribution captures the probability of receiving one of two outcomes (often called success or failure) given a single trial. It is actually just a special case of the Binomial distribution where $n=1$.
- Example:
- Flipping a coin: Head or Tail

Bernoulli Distribution



Bernoulli PMF Formula

$$P_x = \begin{cases} 1 - p & \text{for } x = 0 \\ p & \text{for } x = 1 \end{cases}$$

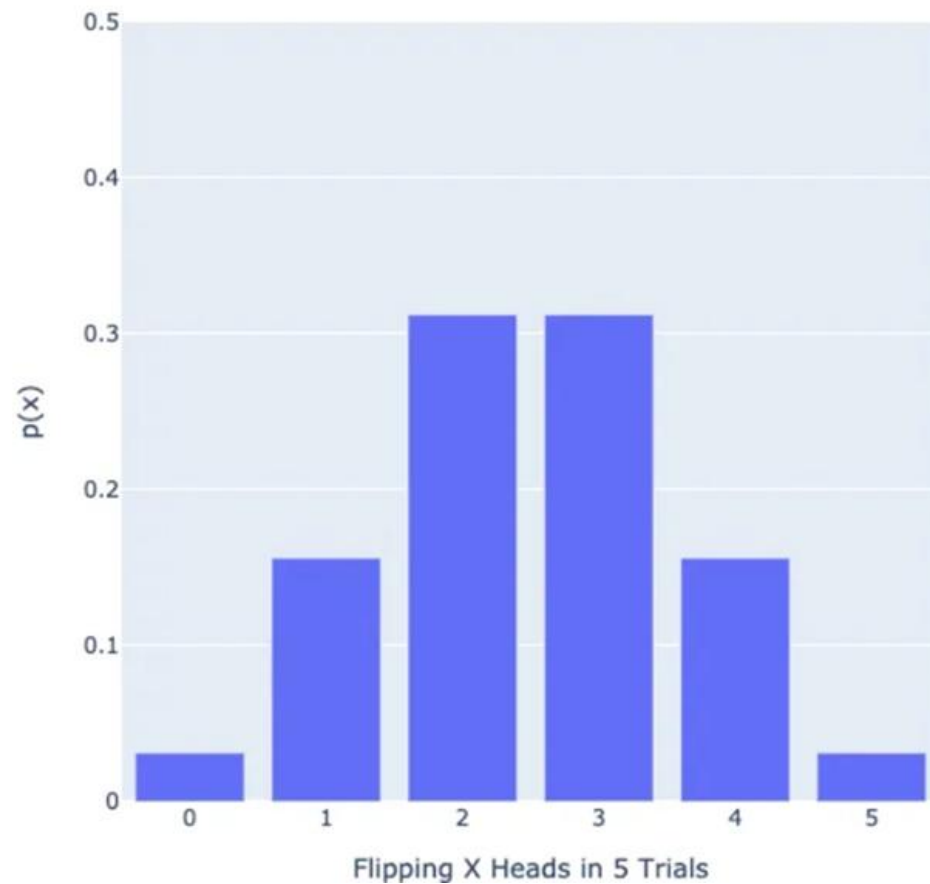
Our Example:

$$P_x = 0.5 = 50\%$$

Binomial Distribution

- The binomial distribution is just taking Bernoulli one step further. We still have trials that result in one of two outcomes (success or failure), but now we are looking at the **probability that a specific number of outcomes (x) occur in n trials** instead of a single trial.
- **Example: flipping a coin n times**

Binomial Distribution



Binomial PMF Formula

$$P_x = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$$

Our Example:

$$P_2 = \frac{5!}{2!(5-2)!} 0.5^2 (1-0.5)^{5-2}$$

$$P_2 = 10 * 0.25 * 0.125$$

$$P_2 = 0.3125 = 31.25\%$$

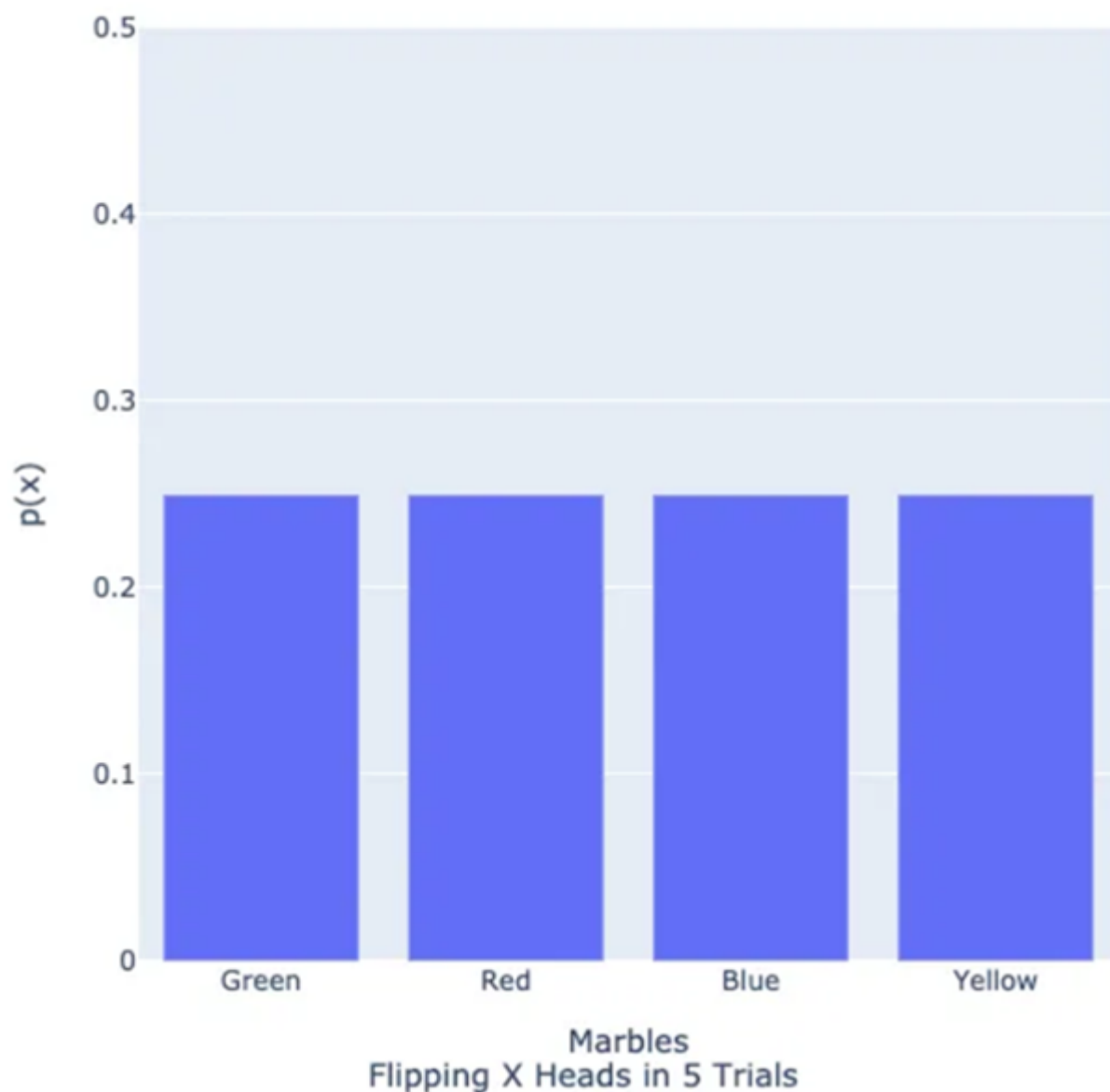
Discrete Uniform Distribution

- A discrete uniform distribution is a simple distribution where we have a set of potential outcomes (n), each of which has an equal likelihood of occurring.

Example:

- You blindly reach into a bag of marbles that contains a green marble, a red marble, a blue marble, and a yellow marble. What are the chances of picking the yellow marble?

Discrete Uniform Distribution



Discrete Uniform PMF Formula

$$P_x = \frac{1}{n}$$


Our Example:

$$P_x = \frac{1}{4} = 25\%$$

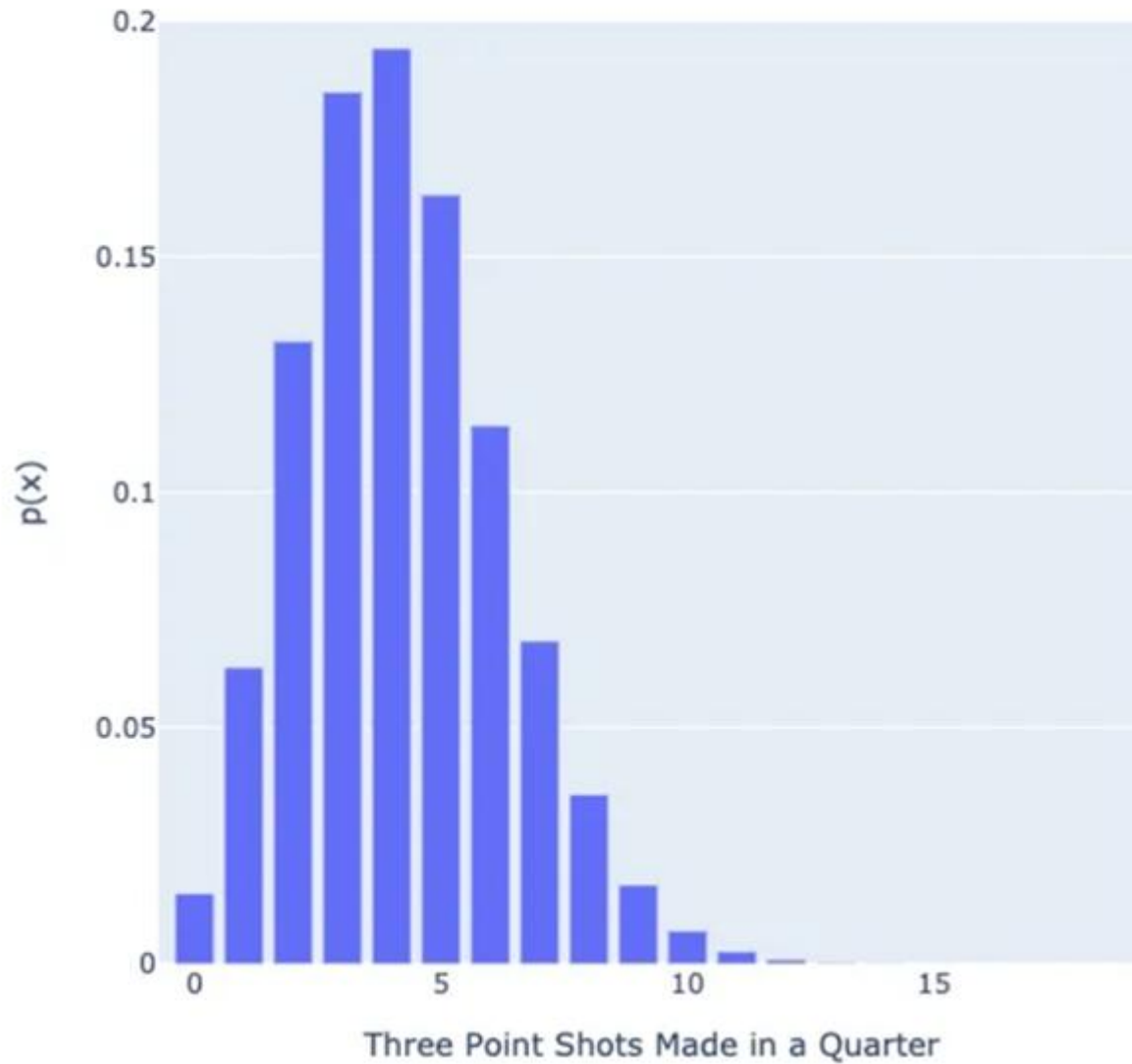
Poisson Distribution

- The Poisson distribution is used to answer the question **how many times is an event likely to occur over a given period of time?** The Poisson distribution is defined by a *rate parameter* (λ), which is the mean number of occurrences of that event in a single unit of the observed time.

Example:

- Let's say that a basketball team scores an average of 4.2 three point shots per quarter  If that is true, then what is the likelihood that this team will score exactly 7 three point shots in a quarter?

Poisson Distribution



Poisson PMF Formula

$$P_x = \frac{\lambda^x}{x!} e^{-\lambda}$$

Our Example:

$$P_7 = \frac{4.2^7}{7!} e^{-4.2}$$

$$P_7 = 0.0686 = 6.86\%$$

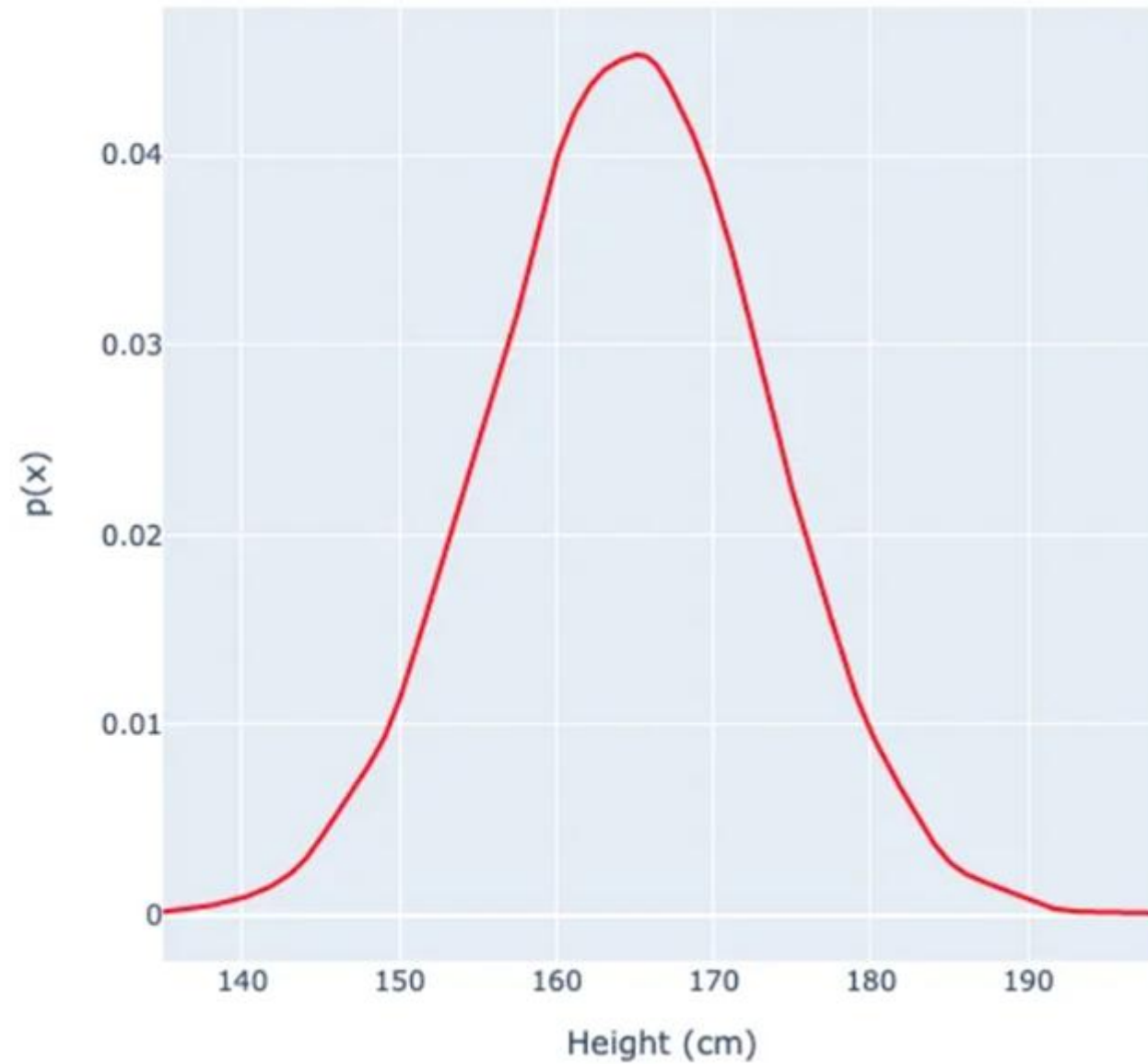
Continuous Data Distributions

- **Gaussian (Normal) Distribution**
- The Gaussian distribution is probably the **most widely known and observed distribution in nature**. It is defined by a *mean value* (μ) and a *standard deviation* (σ).
- Also now that we are talking about continuous values, we can no longer say “what is the likelihood of this *exact* value occurring” because technically there is no *exact* values in a continuous space. Instead we ask the question “what is the likelihood of a sample *falling within a given range of values*?”

Example:

- Let's say I am friends with every single person in the world (🧐) and everyone volunteers their height information to me. The mean height of the population turns out to be 164.58cm with a standard deviation of 8.83cm.
- Given the information above, what is the probability of someone being taller than 175cm?

Gaussian Distribution



Gaussian PDF Formula

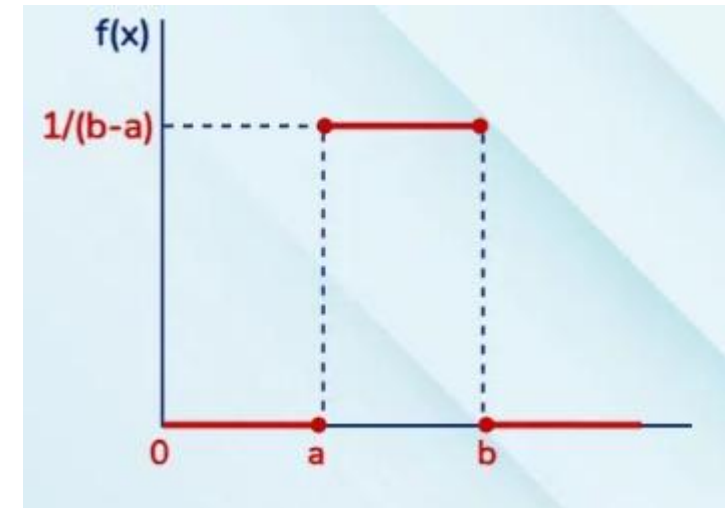
$$p_x = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Continuous uniform distribution

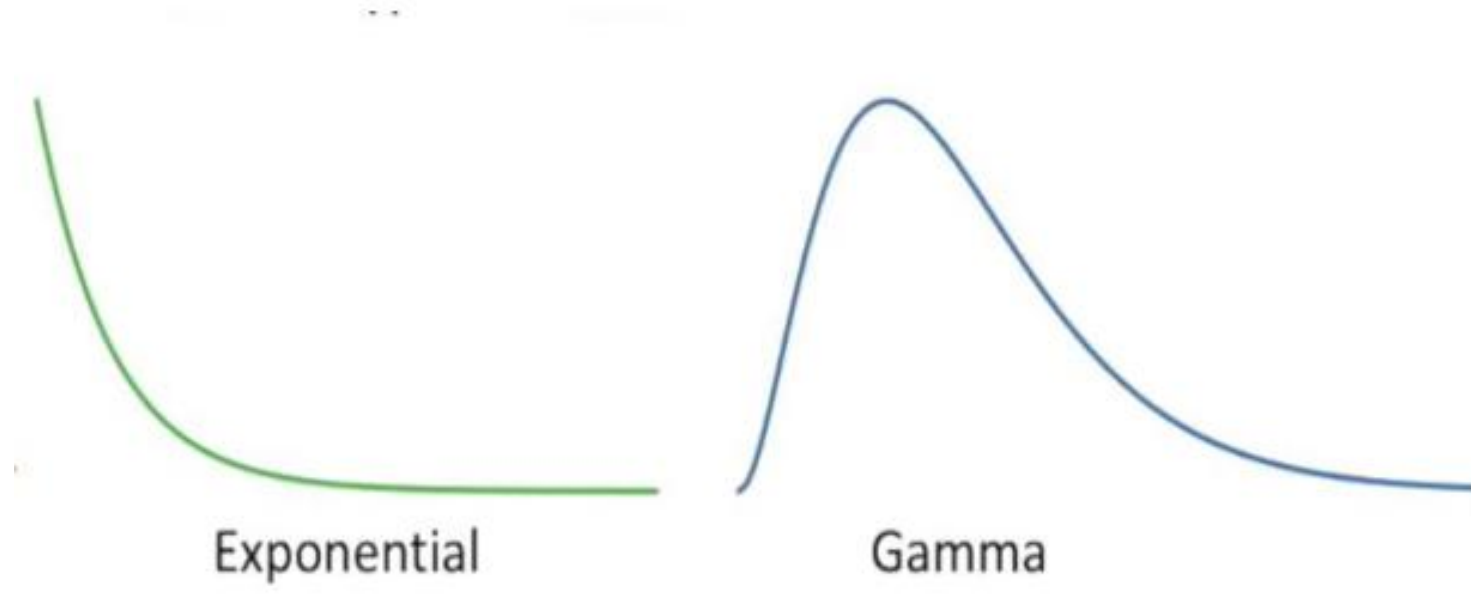
A continuous uniform distribution can be seen in situations where every possible outcome within a specific range is equally likely to occur,

Let's consider a continuous uniform distribution where the random variable X can take any value between $a=2$ and $b=5$.

$$f(x) = \begin{cases} \frac{1}{5-2} = \frac{1}{3} & \text{if } 2 \leq x \leq 5, \\ 0 & \text{otherwise.} \end{cases}$$



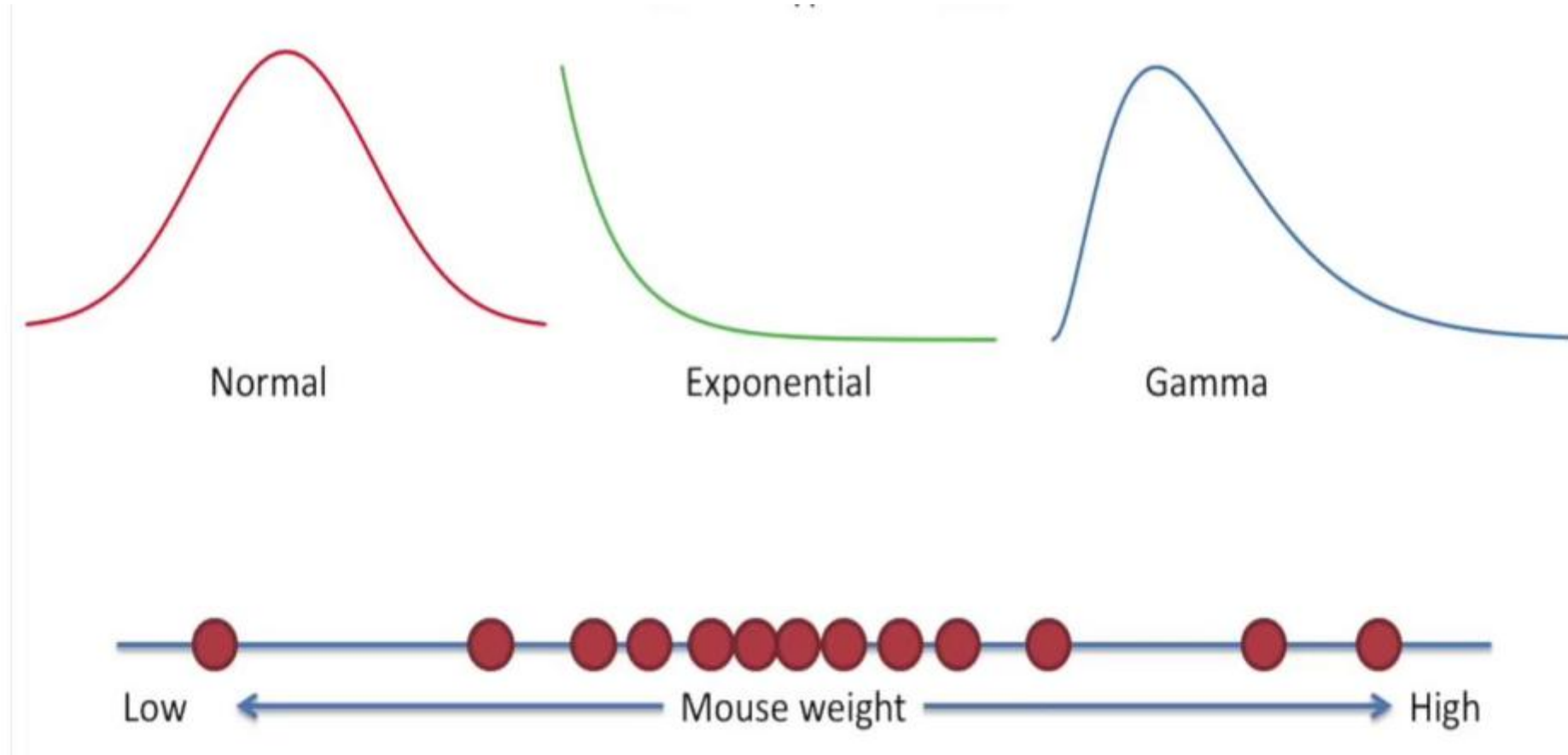
Other data distributions



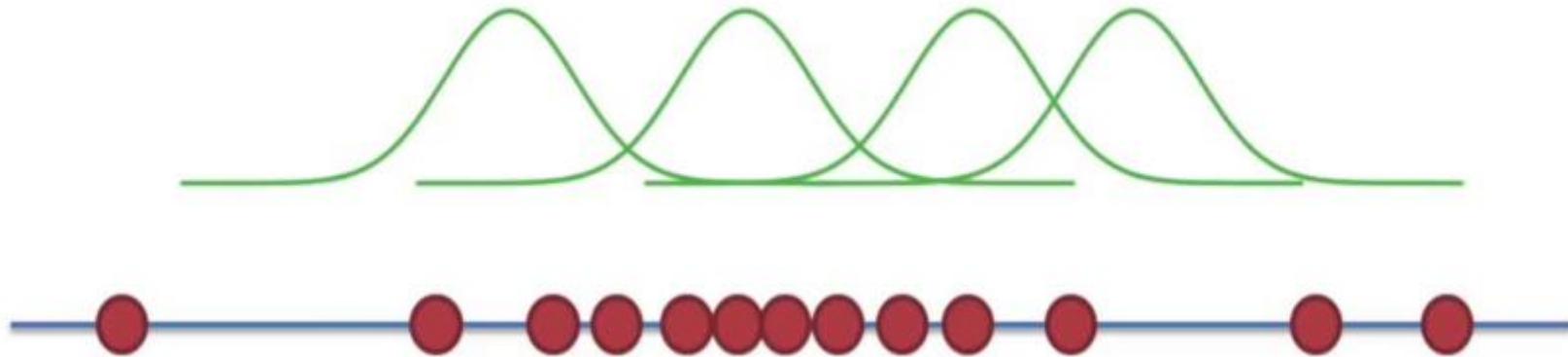
Maximum likelihood Estimation

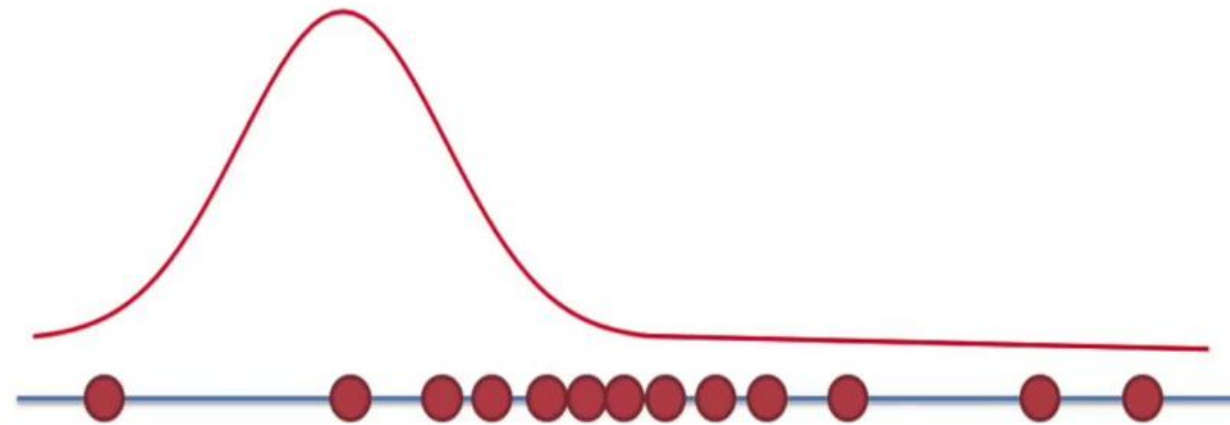
- **Maximum Likelihood Estimation (MLE)** is a statistical method used to estimate the parameters of a probability distribution or a statistical model by maximizing the **likelihood function**. The likelihood function measures how well the model explains the observed data for a given set of parameters. MLE finds the parameter values that make the observed data most probable under the assumed model.

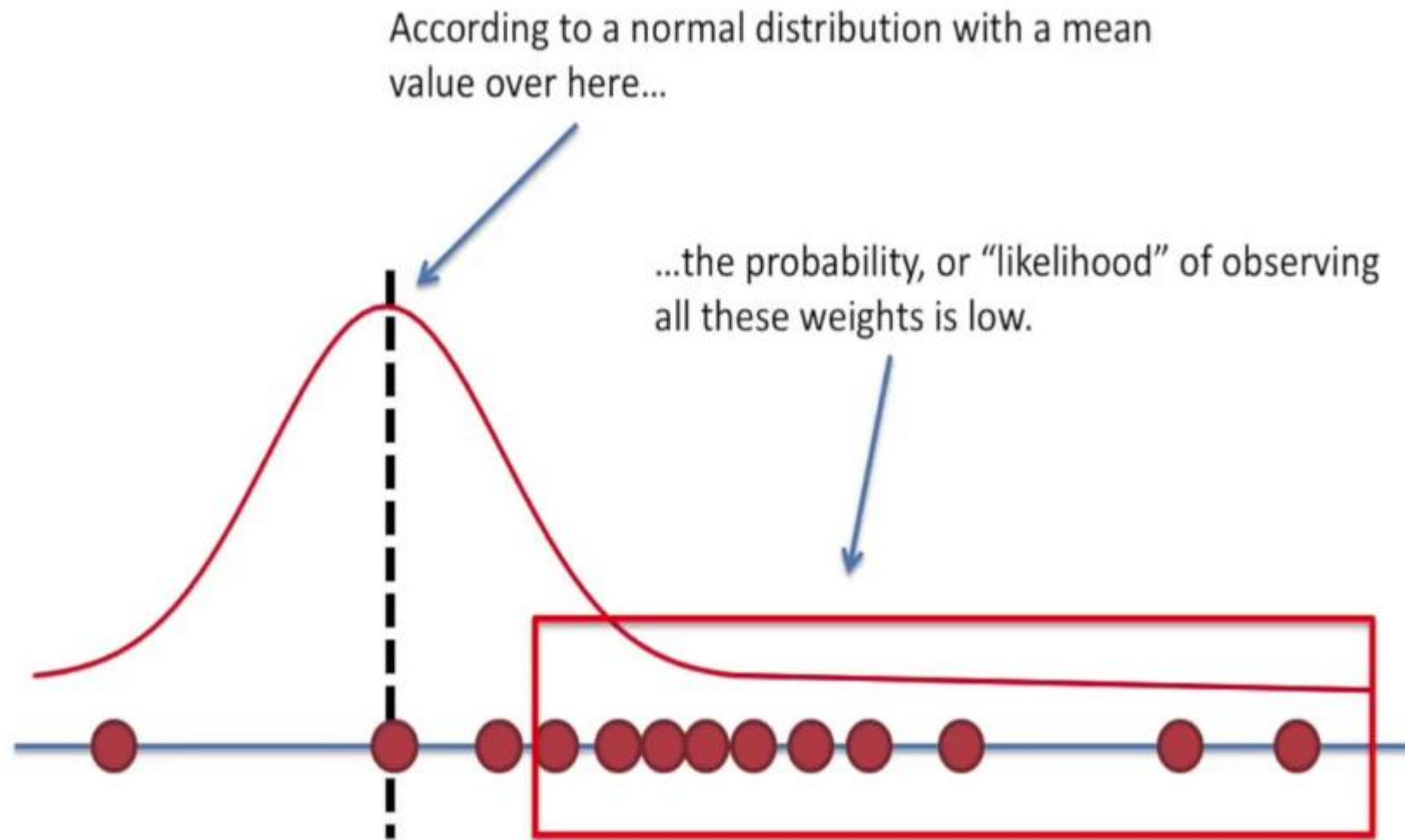
MLE example



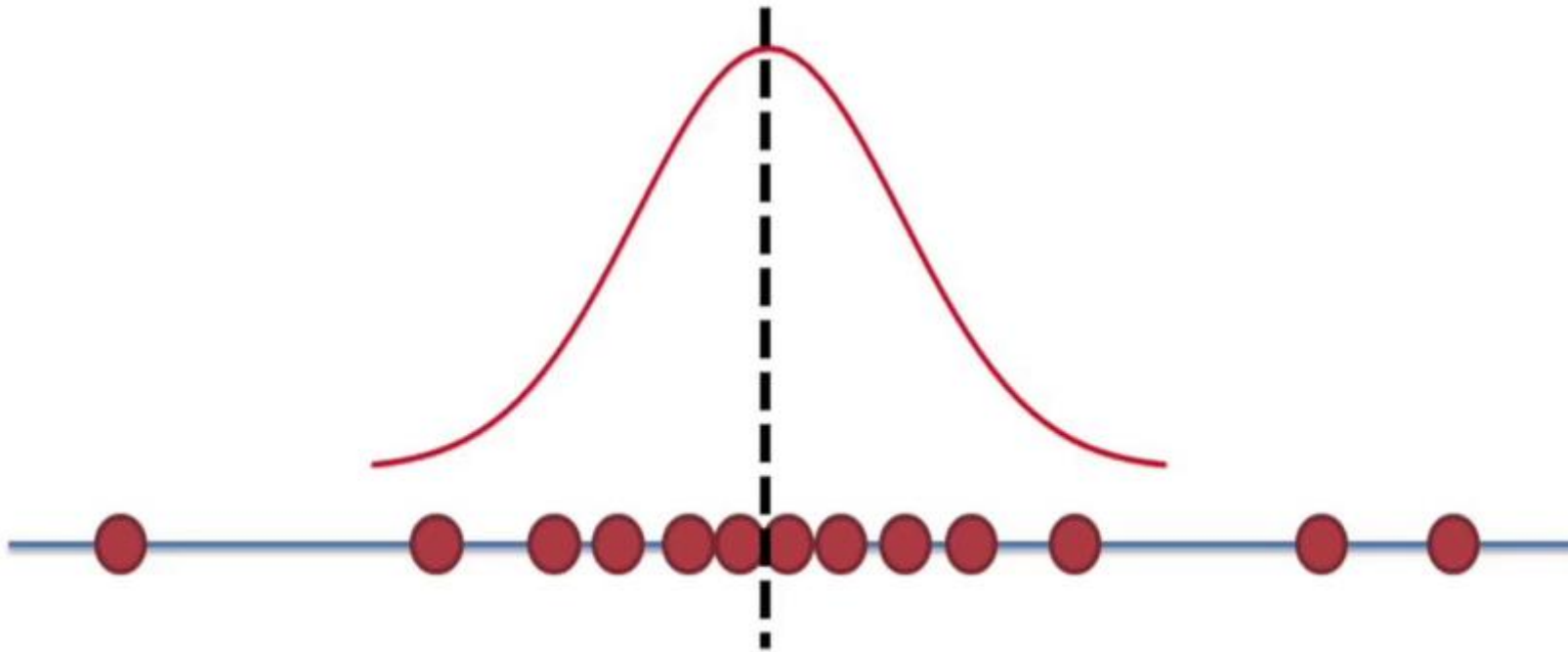
Is one location “better” than another?



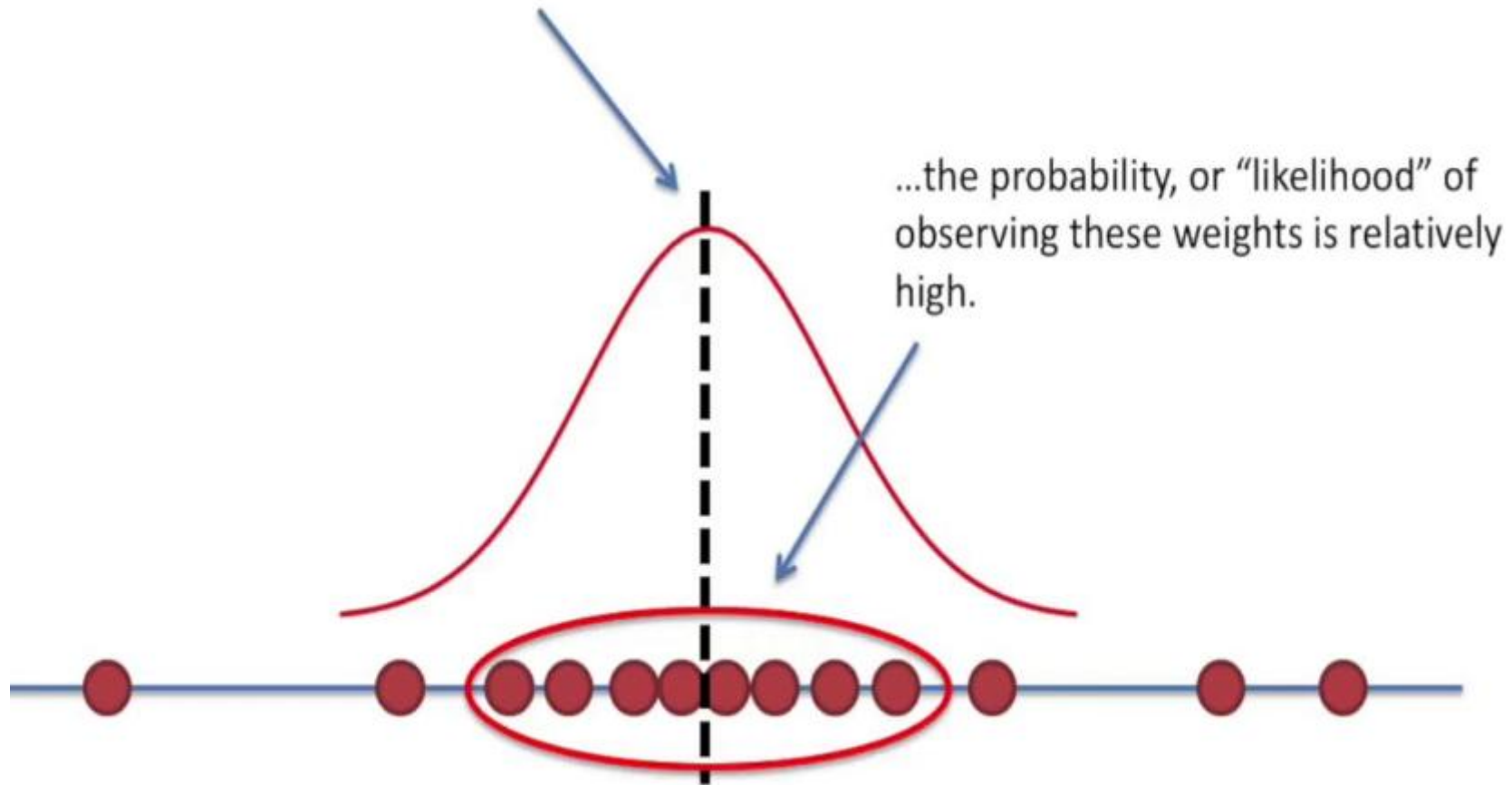




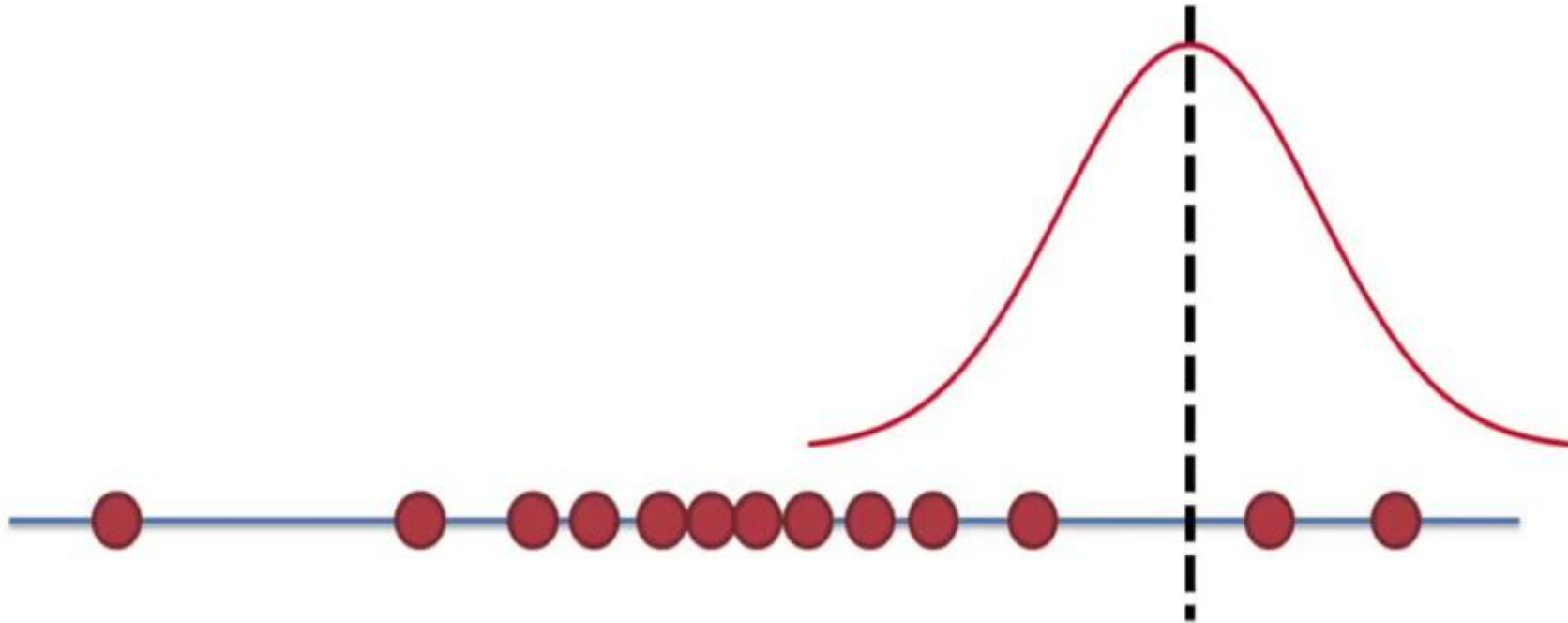
What if we shifted the normal distribution over, so that its mean was the same as the average weight?



According to a normal distribution
with a mean value here...

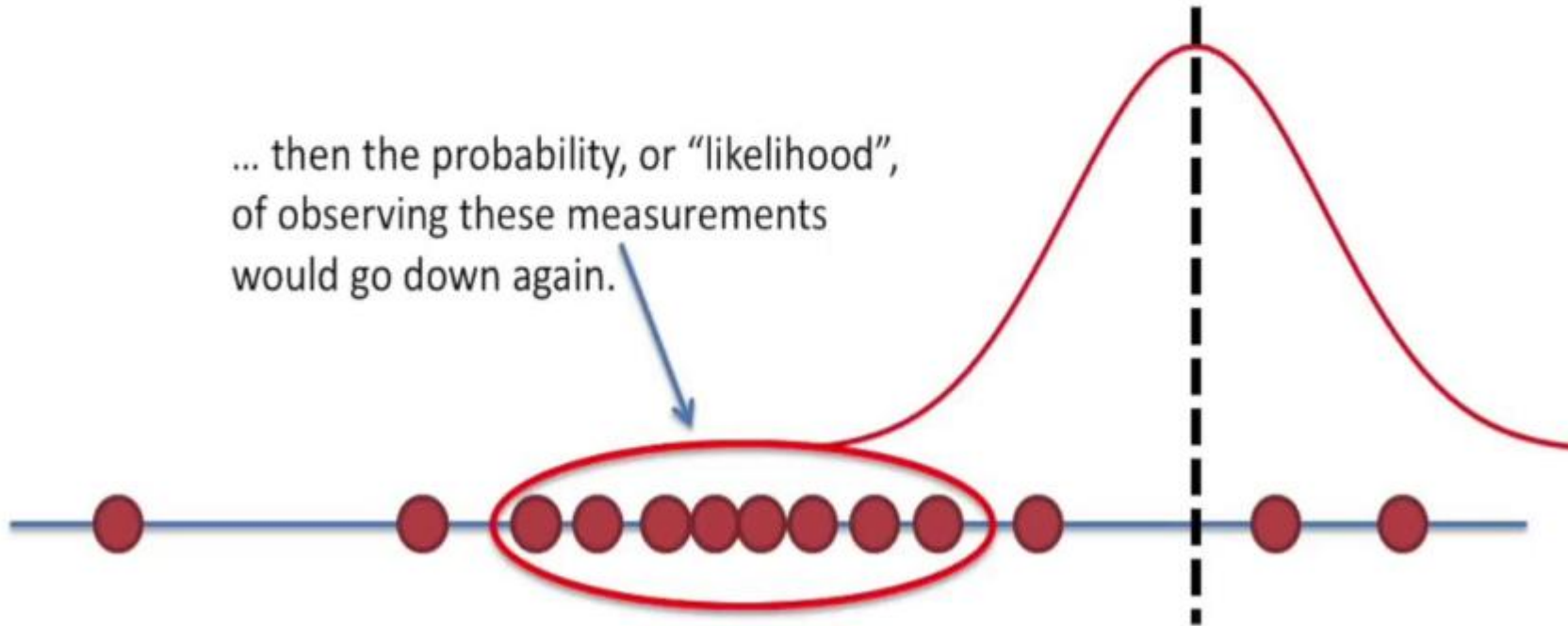


If we kept shifting the normal distribution over...



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.

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

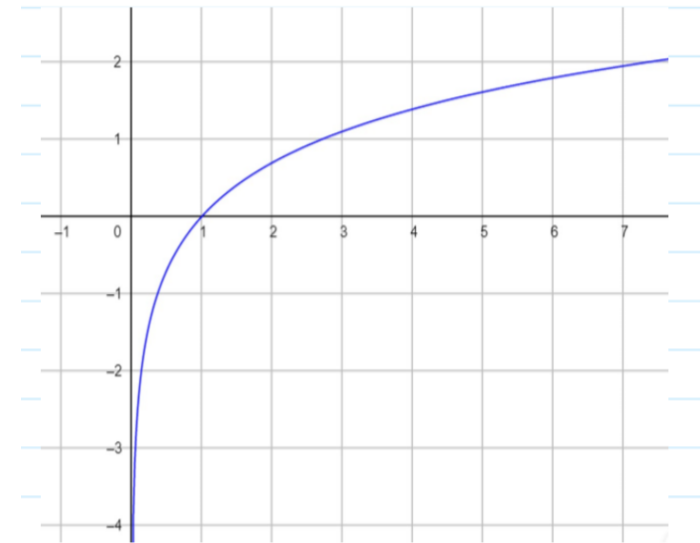
فرض می کنیم هر داده از توزیع $P(x | \theta)$ به صورت مستقل به دست آمده اند.

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

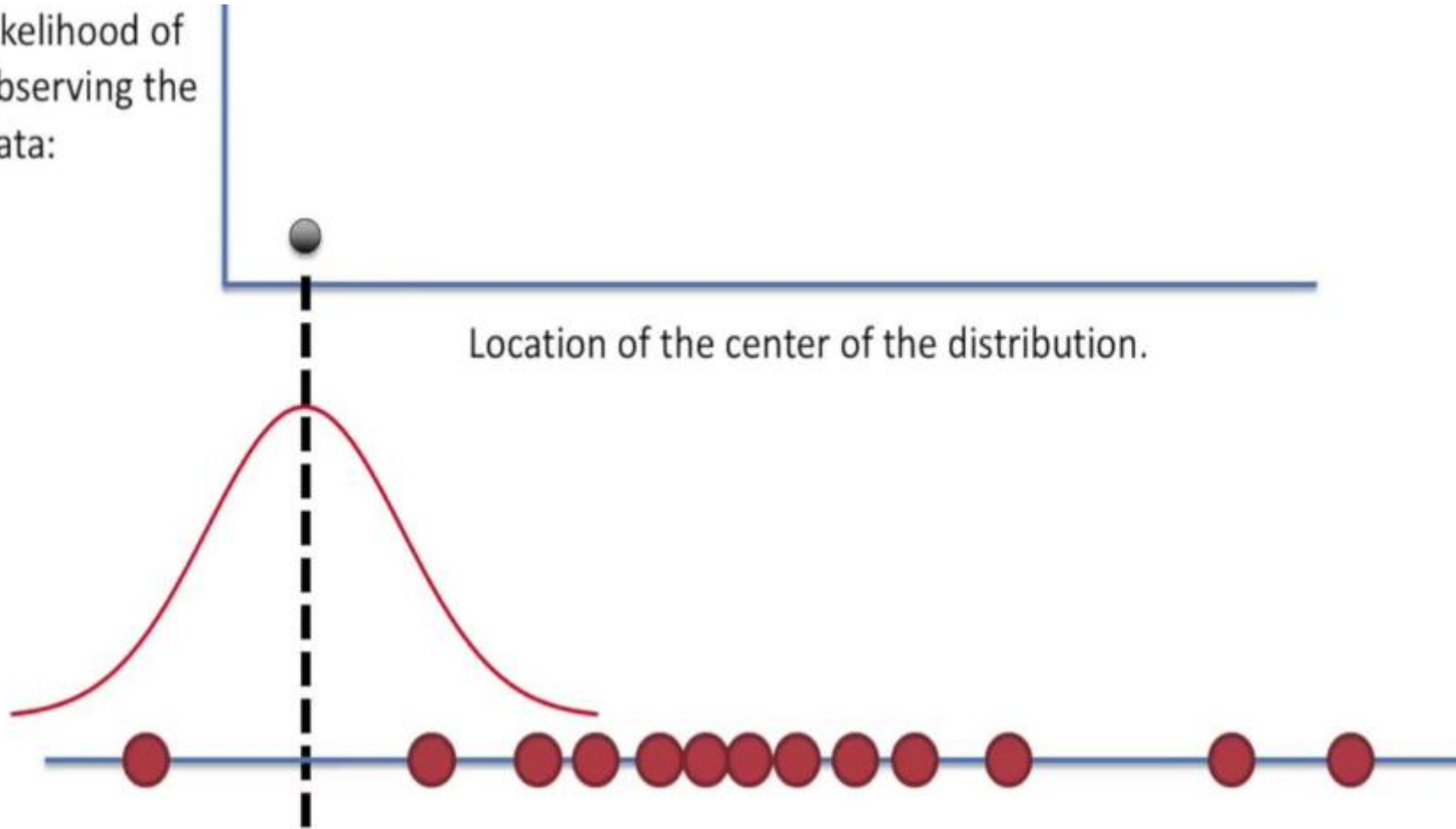
$$\hat{\theta} = \operatorname{argmax}_{\theta} L(\theta)$$

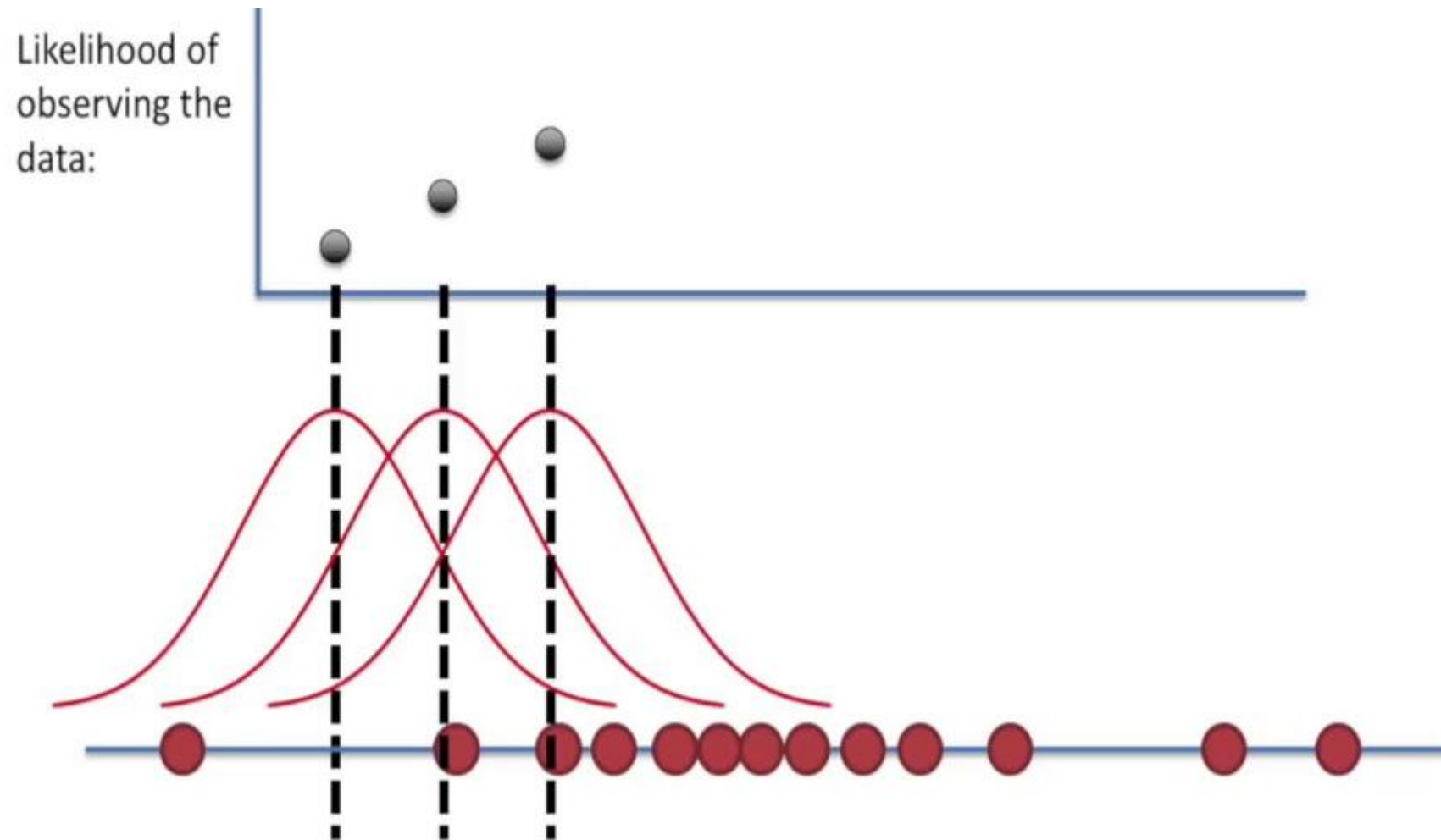
$$\log P(X | \theta) = \sum_{n=1}^N \log P(x^n | \theta) = \log L(\theta)$$

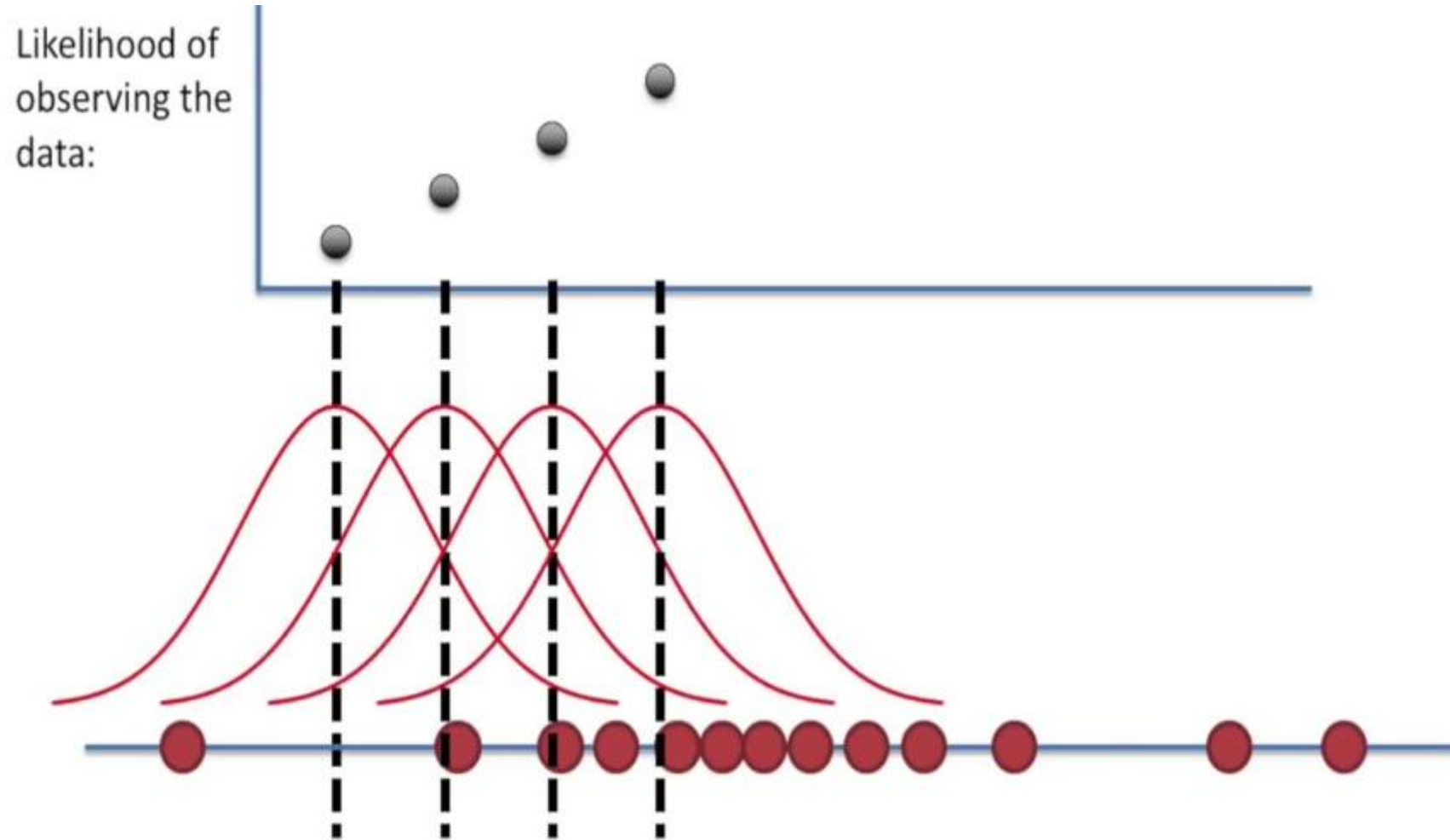
$$\hat{\theta} = \operatorname{argmax}_{\theta} L(\theta) = \operatorname{argmax}_{\theta} \log L(\theta)$$

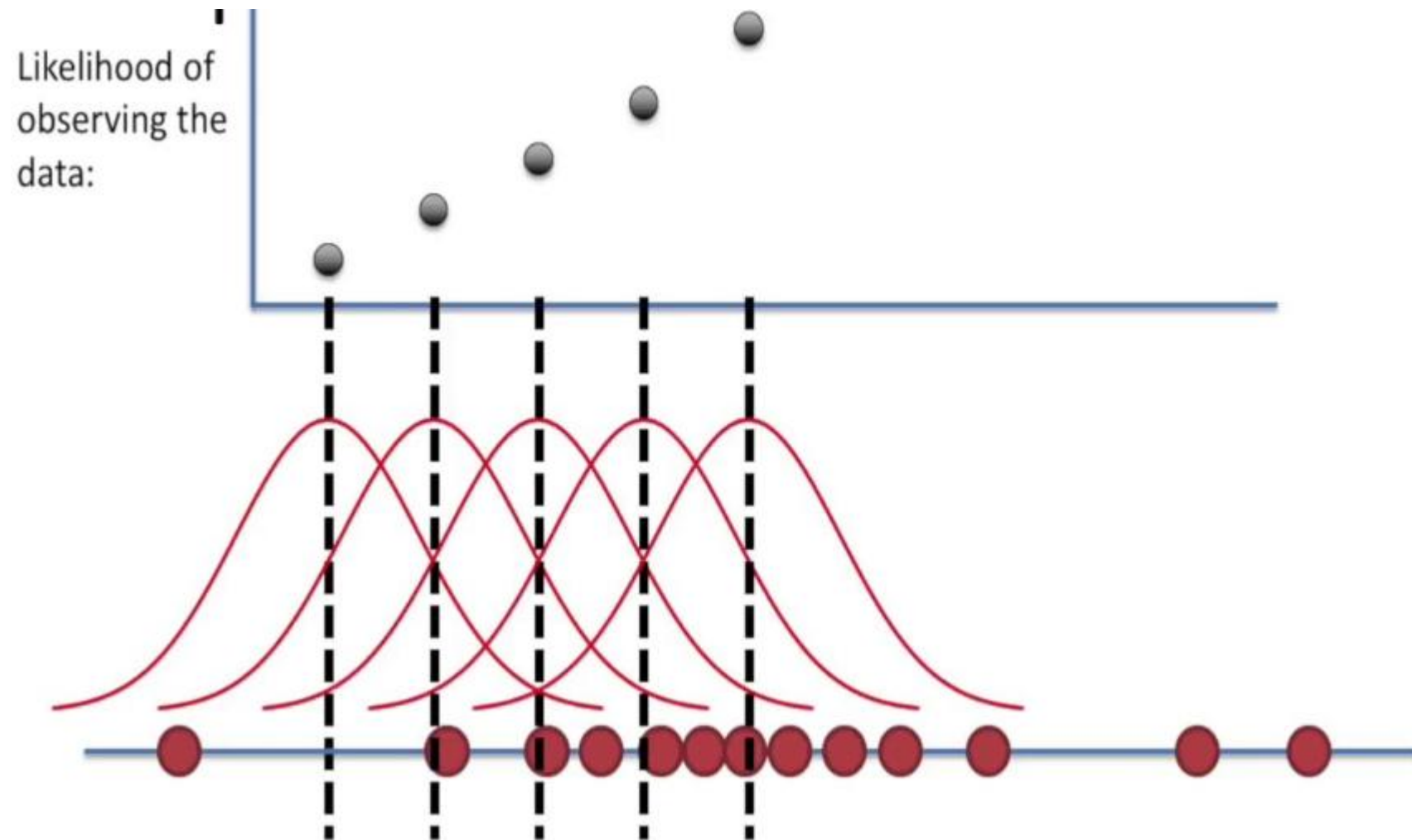


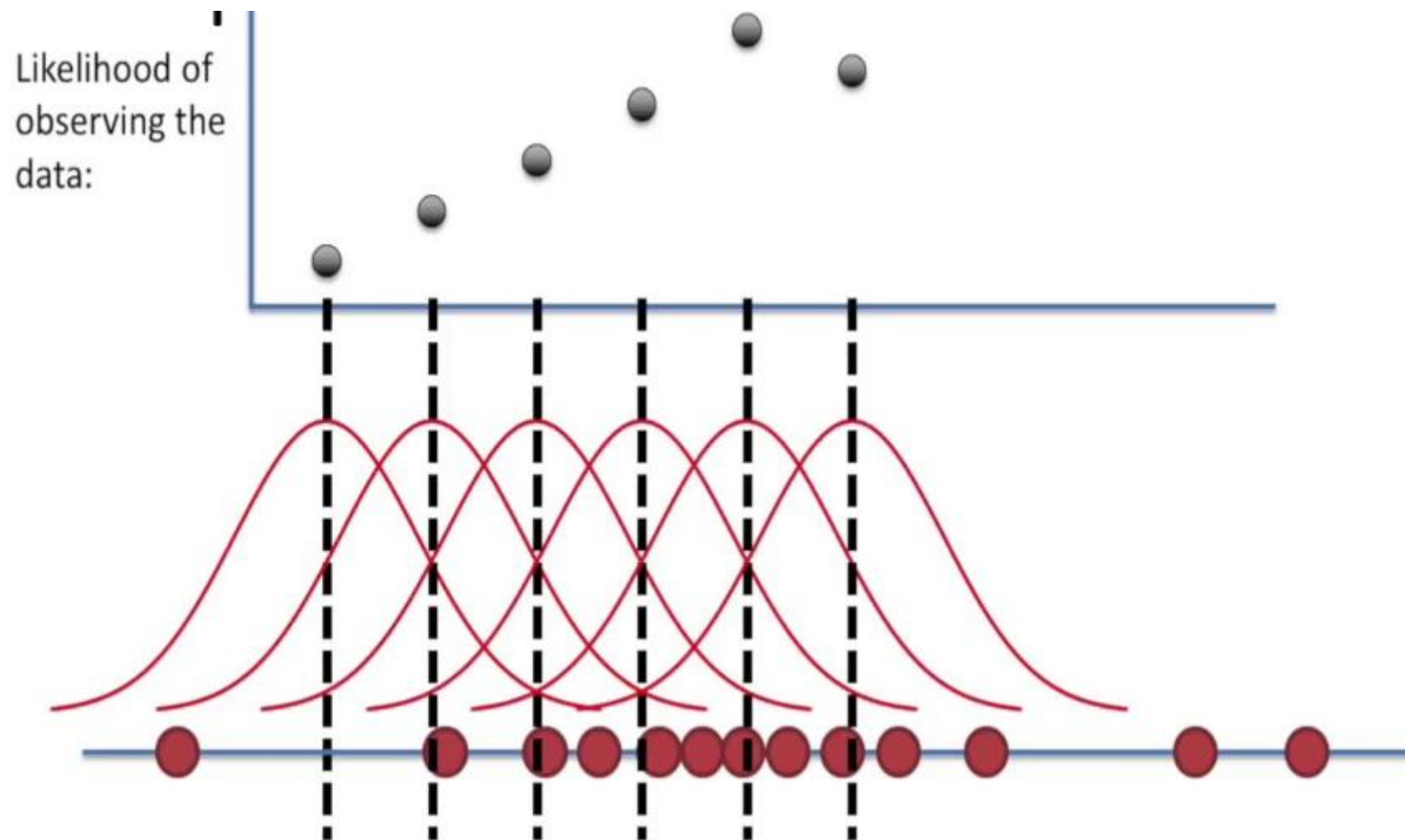
Likelihood of
observing the
data:

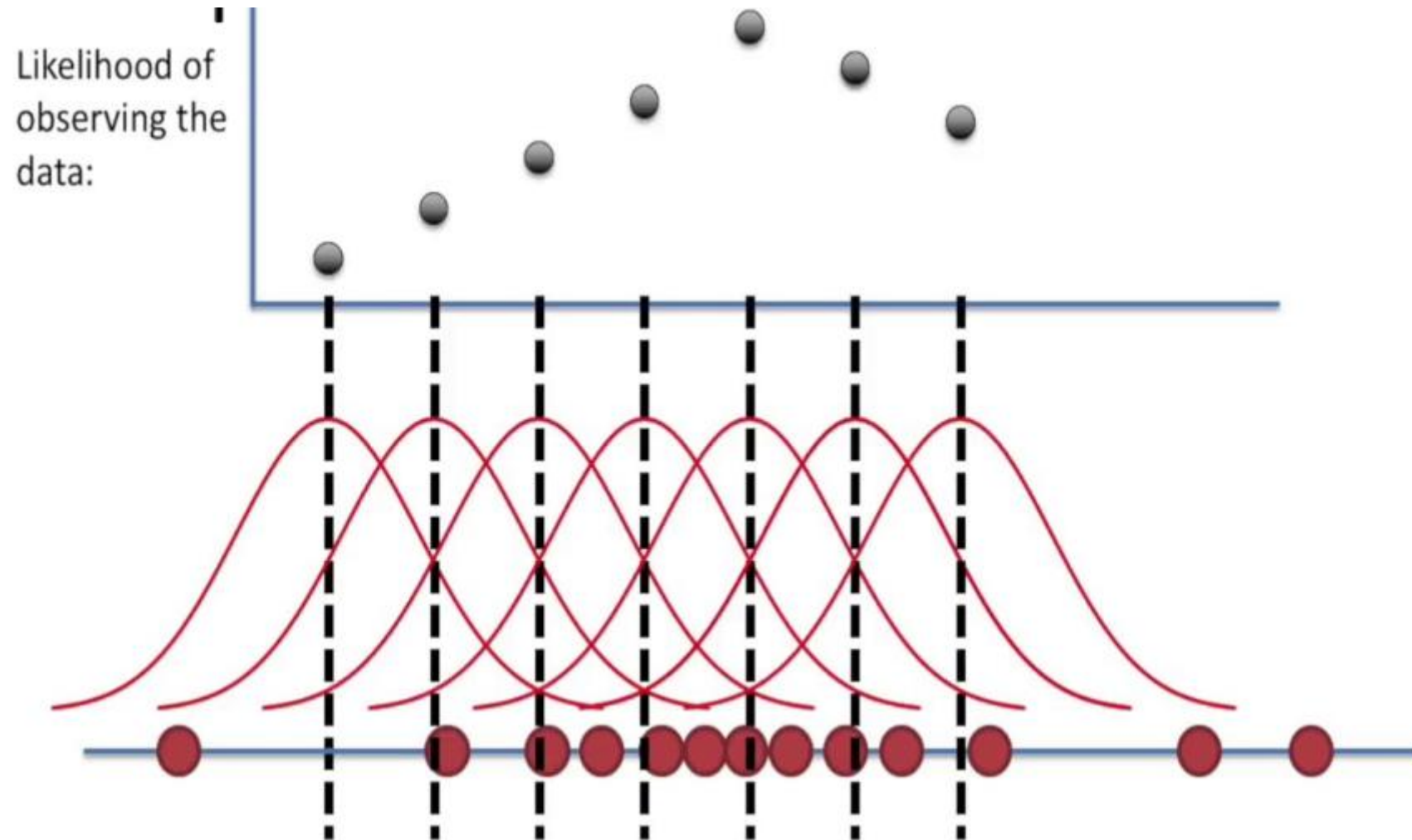


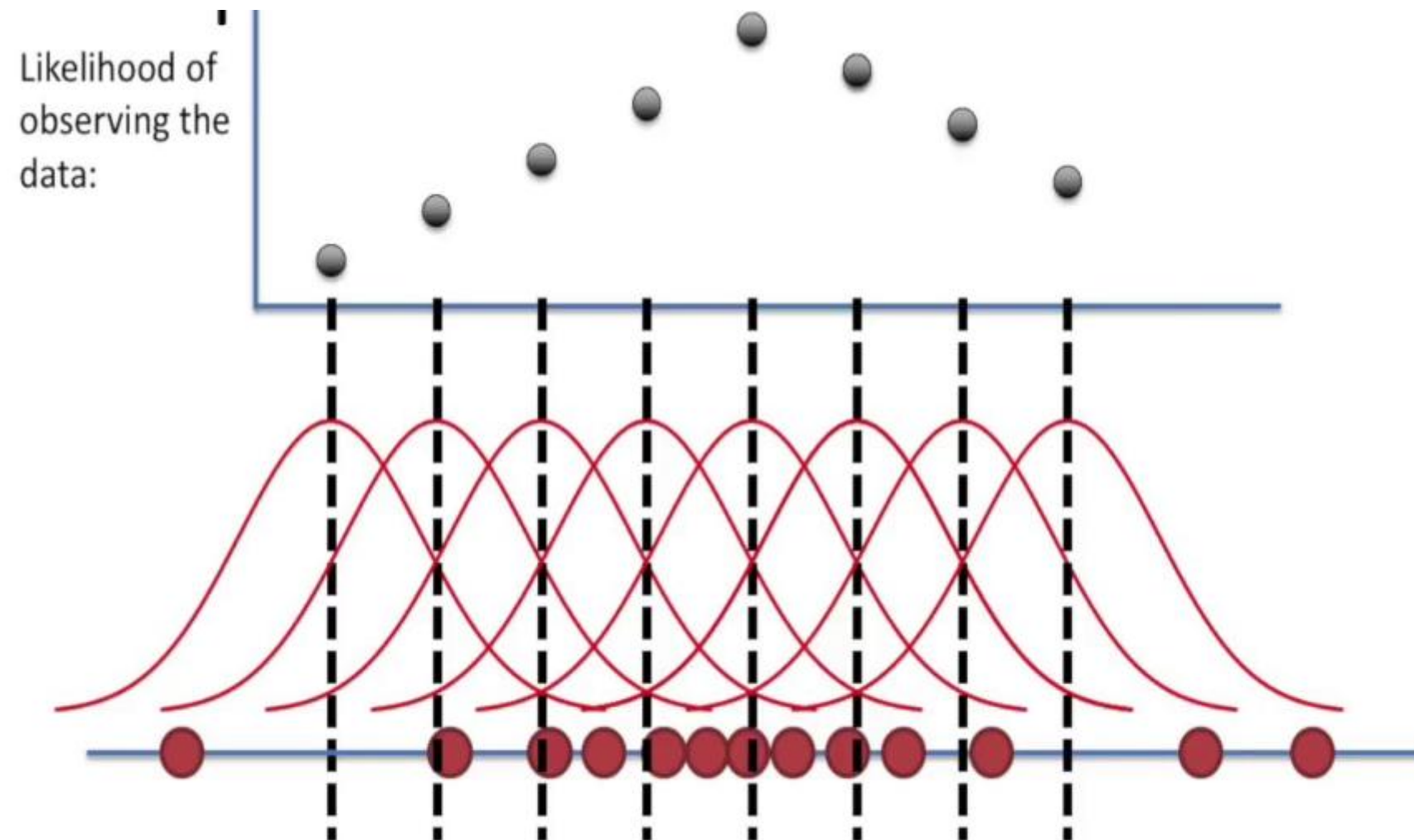






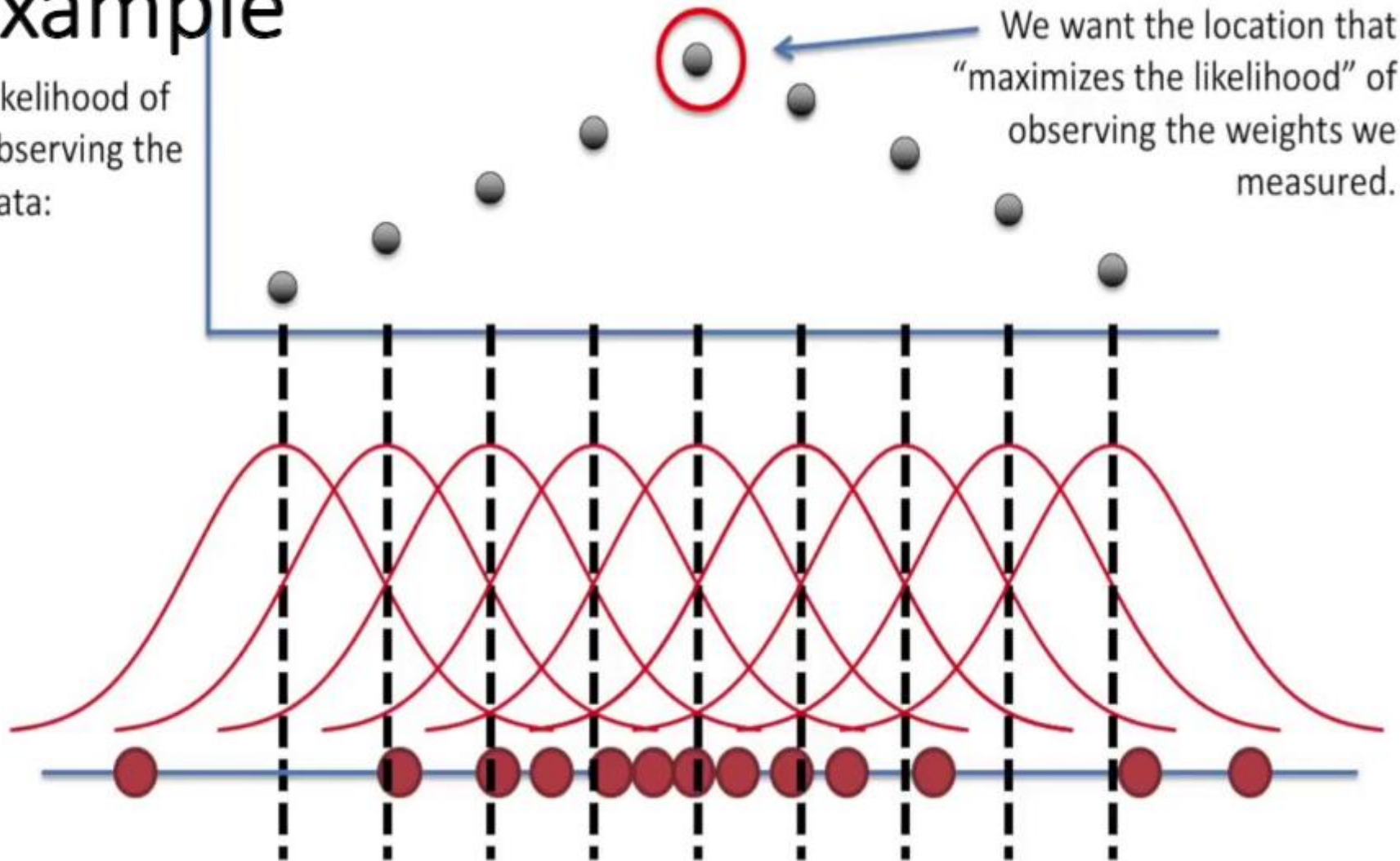






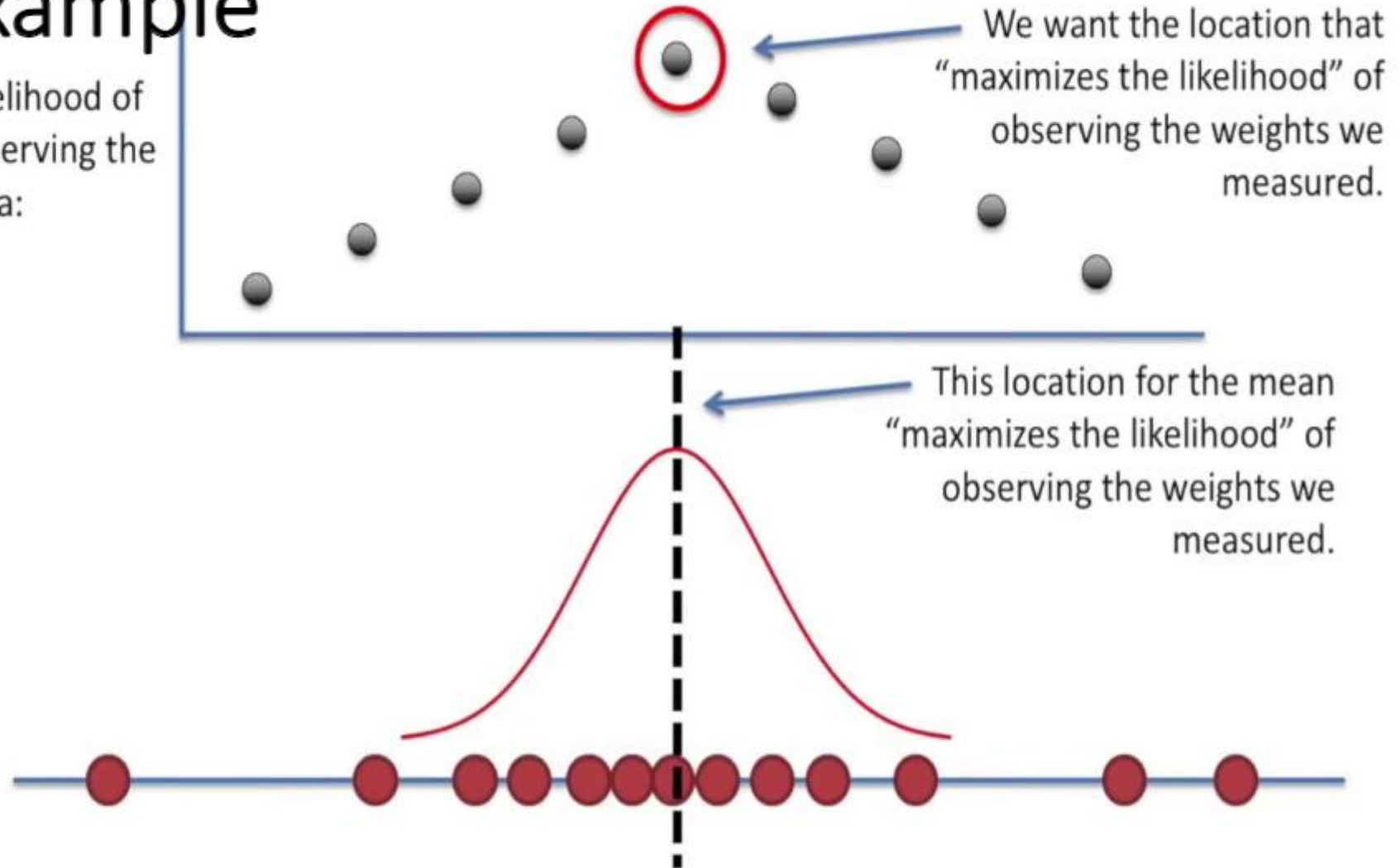
MLE Example

Likelihood of
observing the
data:

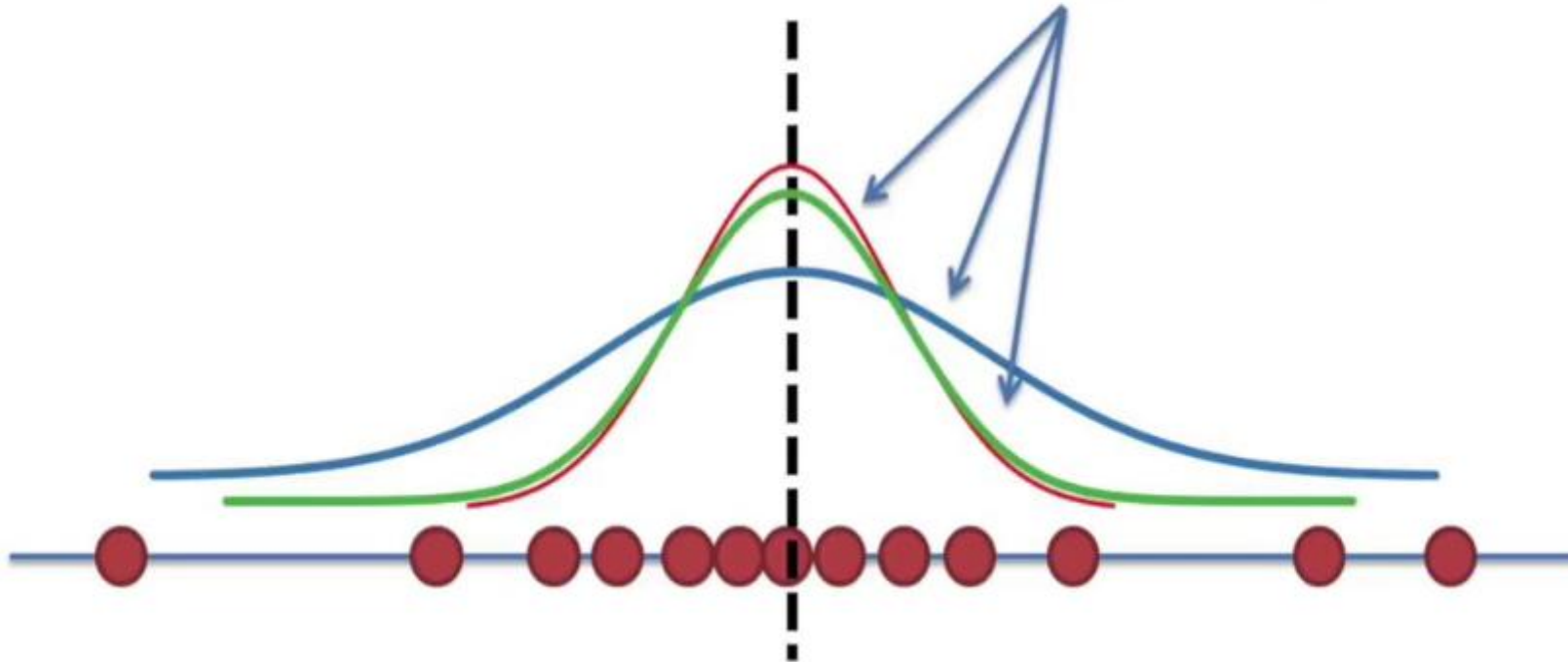


MLE Example

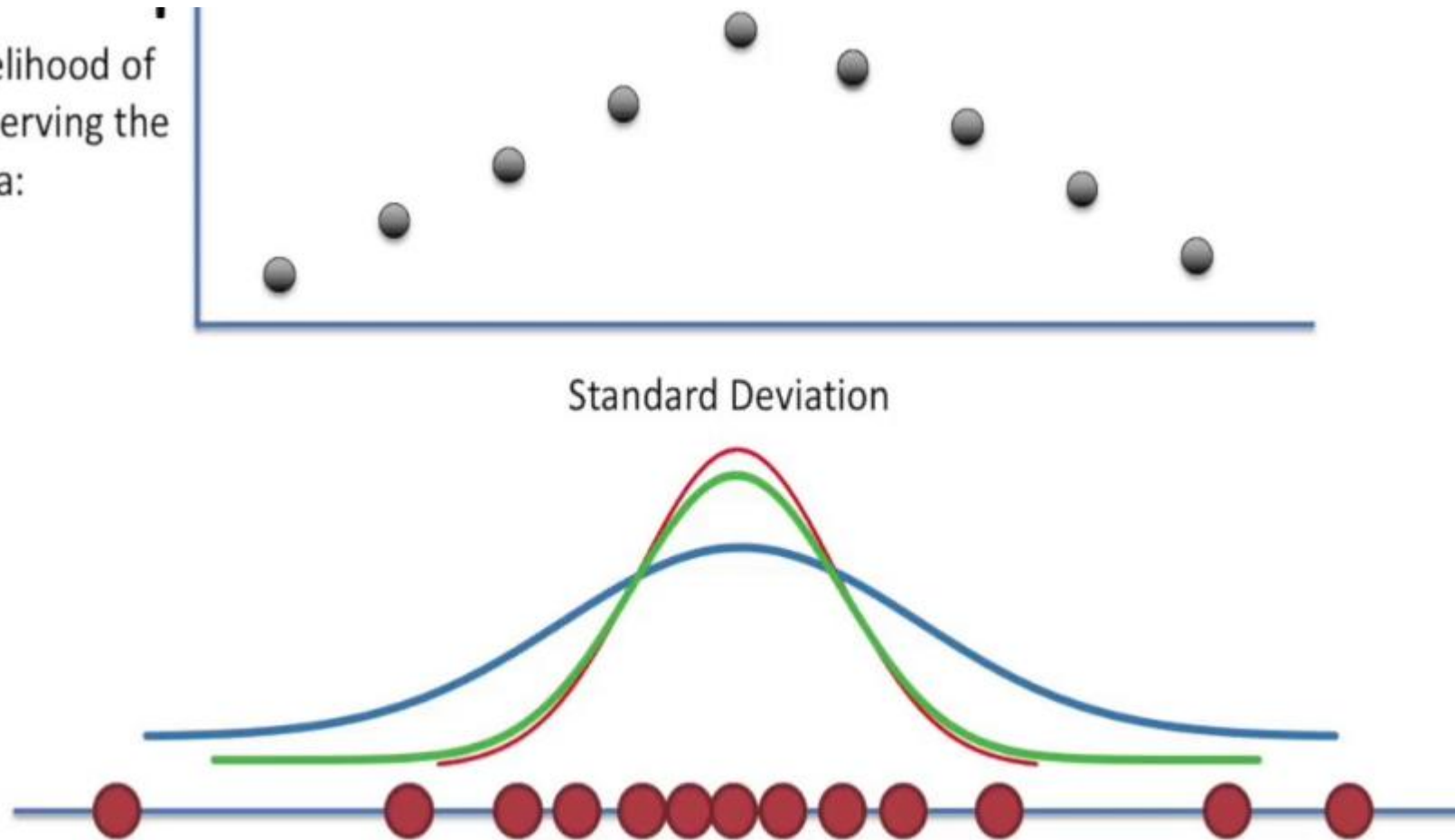
Likelihood of
observing the
data:



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

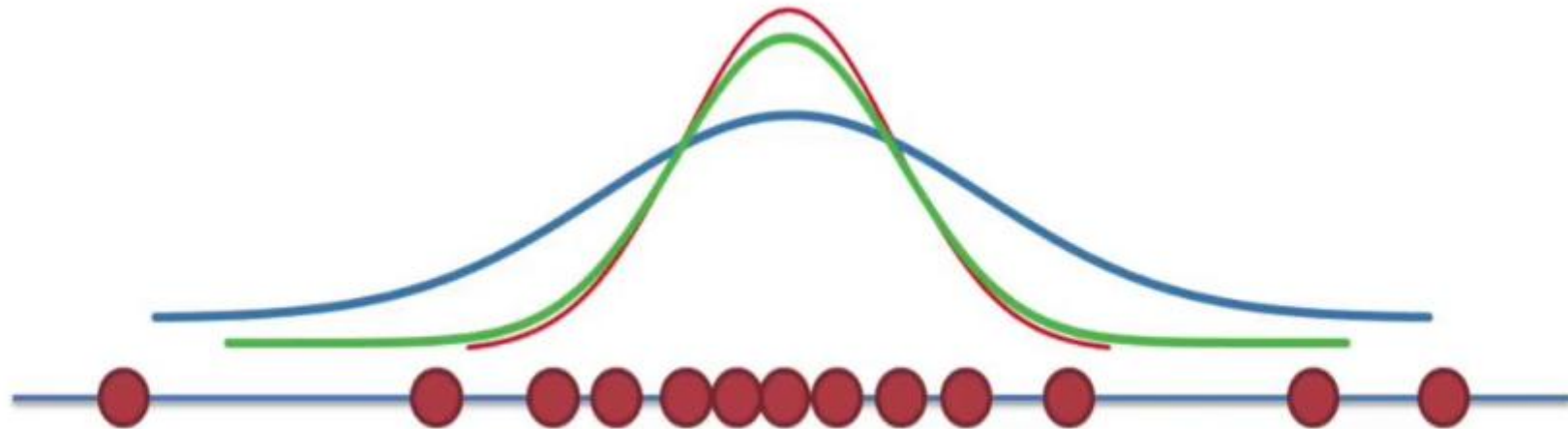
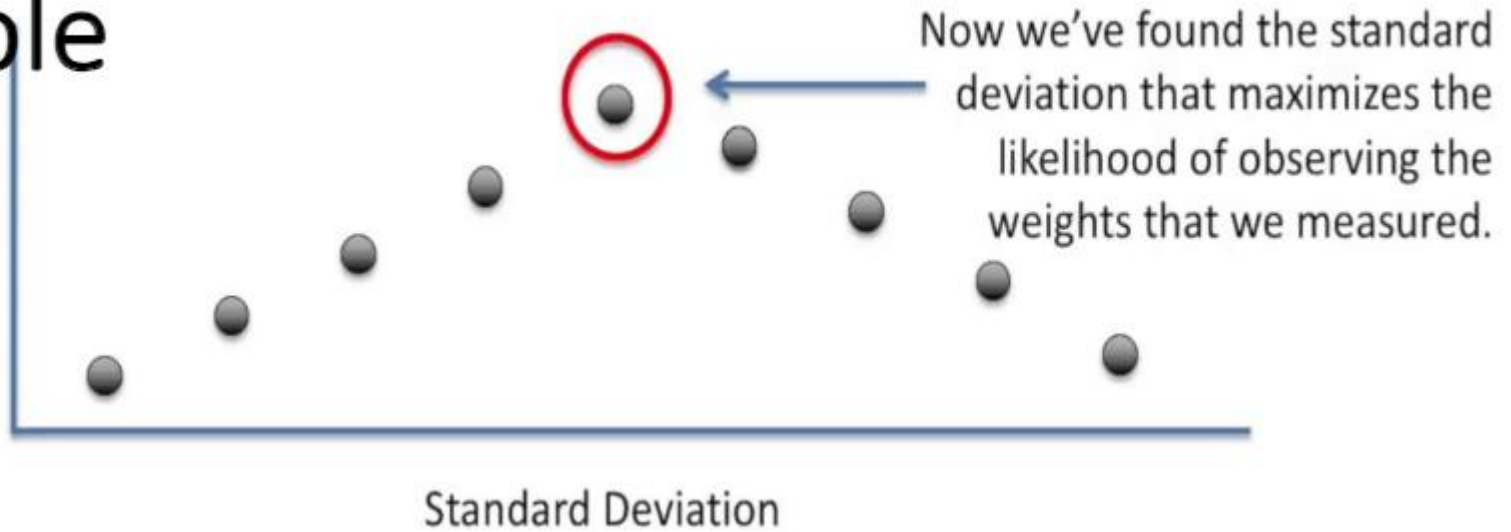


Likelihood of
observing the
data:



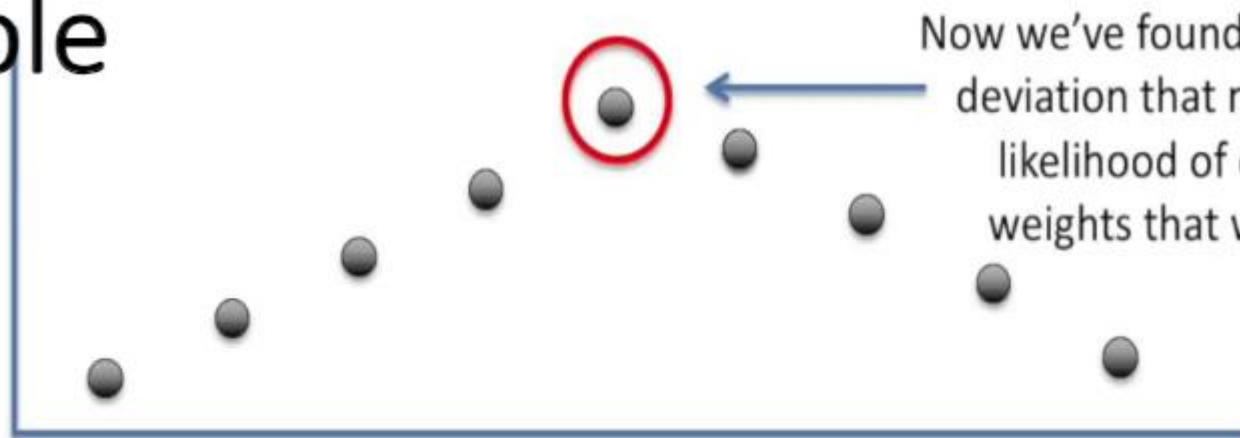
MLE Example

Likelihood of
observing the
data:



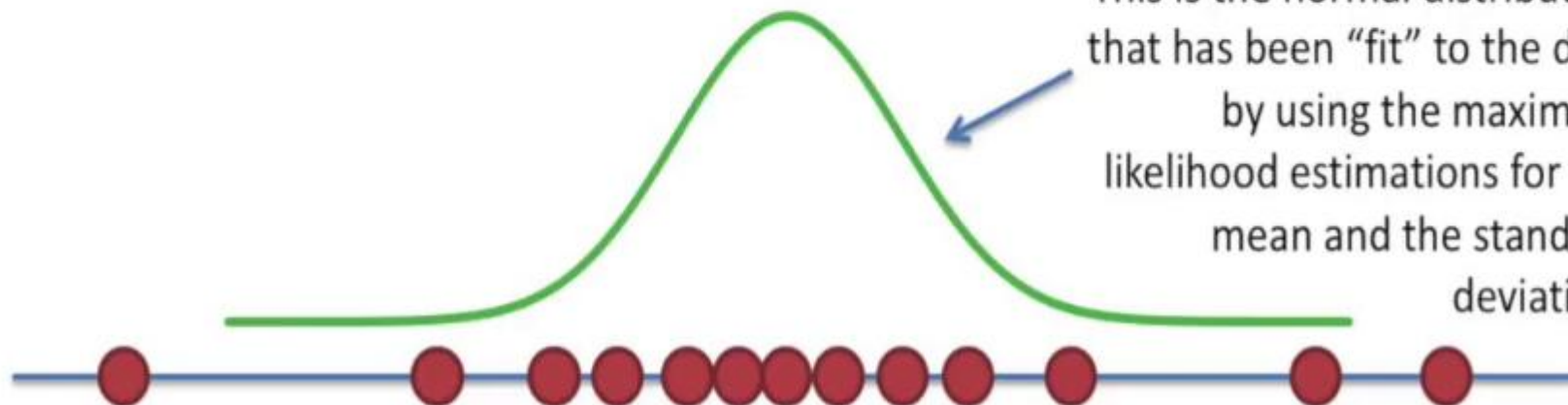
MLE Example

Likelihood of
observing the
data:



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

Standard Deviation



This is the normal distribution
that has been "fit" to the data
by using the maximum
likelihood estimations for the
mean and the standard
deviation.

Calculating the MLE

- Probability of observing a single data point x

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

- 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.

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

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

Derivation with respect to μ

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

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

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

- The same can be done for the standard deviation.

Let x_1, x_2, \dots, 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 $\theta_1 = \mu$ and $\theta_2 = \sigma^2$

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

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:

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

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 θ_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 $2\theta_2^2$:

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

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:

$$y_n = x_n^T w + \varepsilon_n$$

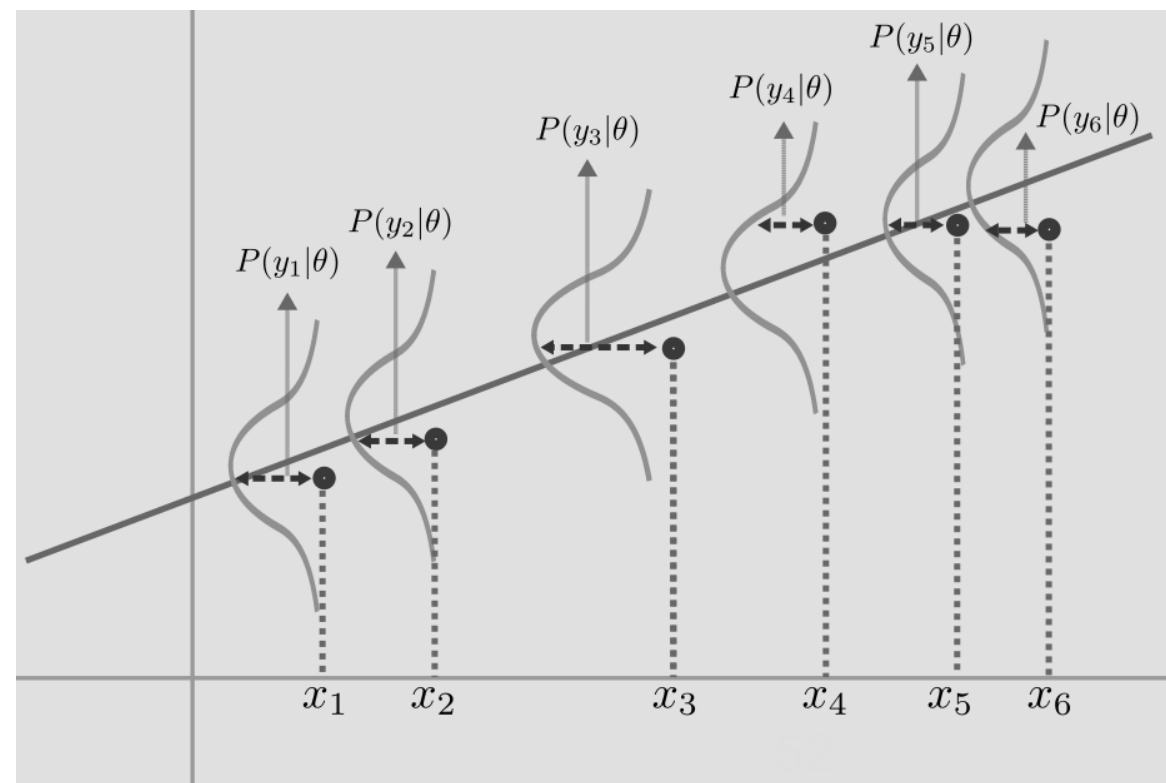
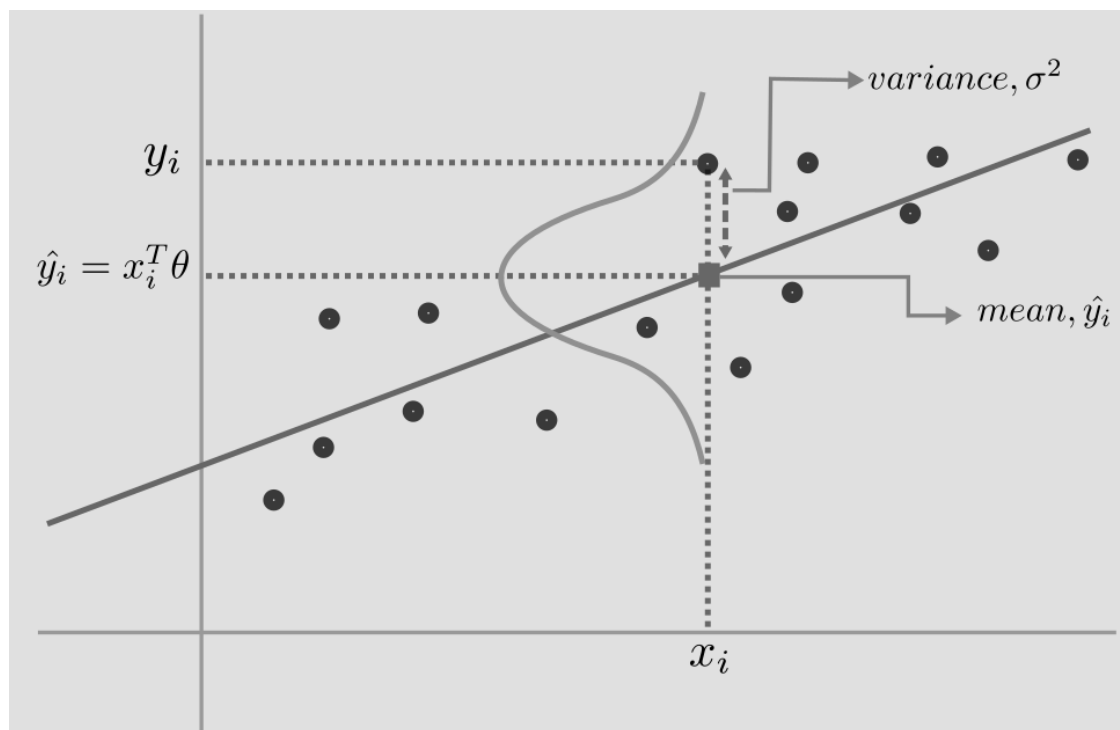
$$\varepsilon_n \sim N(\mu, \sigma^2):$$

ε_n یک نویز گاوسی با میانگین صفر و واریانس σ^2 است.
نویز با نمونه های تبدیل یافته جمع می شود و مستقل از نمونه هاست.

w : پارامترهای مدل است

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

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



ادامه یک مدل احتمالاتی برای LS

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

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

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

رابطه LS , log-likelihood

Log Likelihood:

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

LS:

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

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