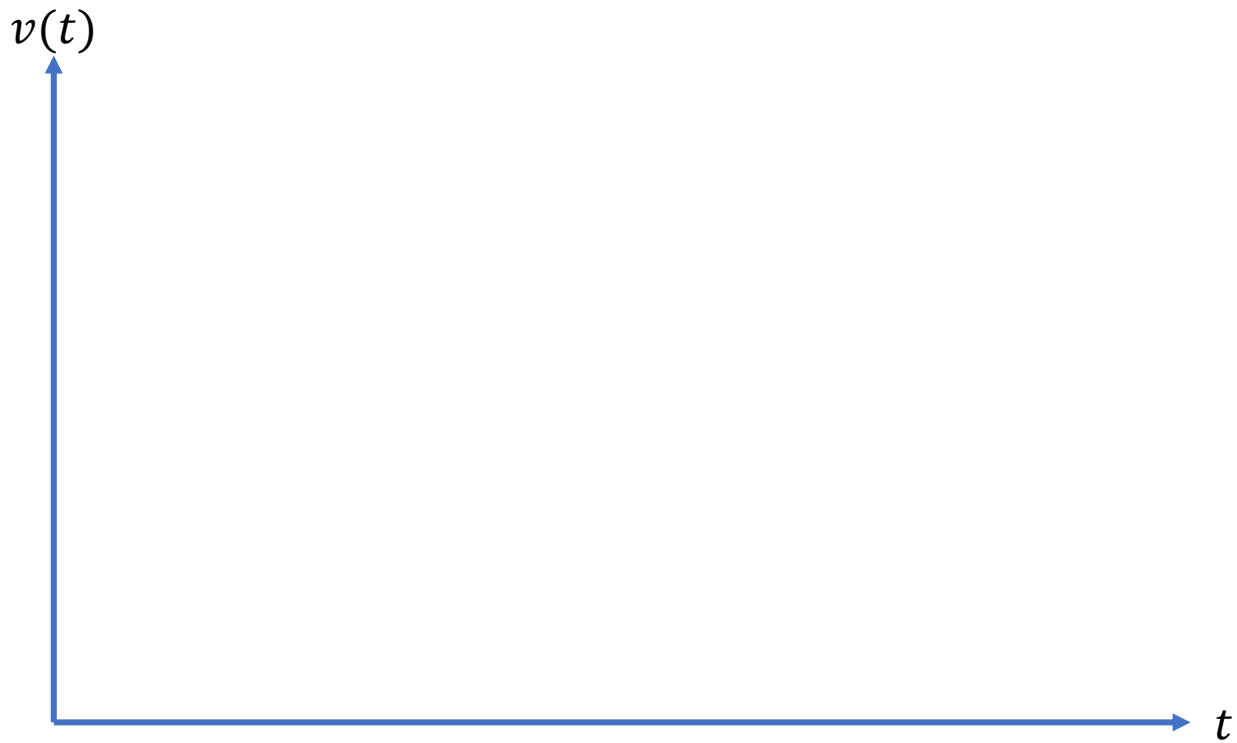
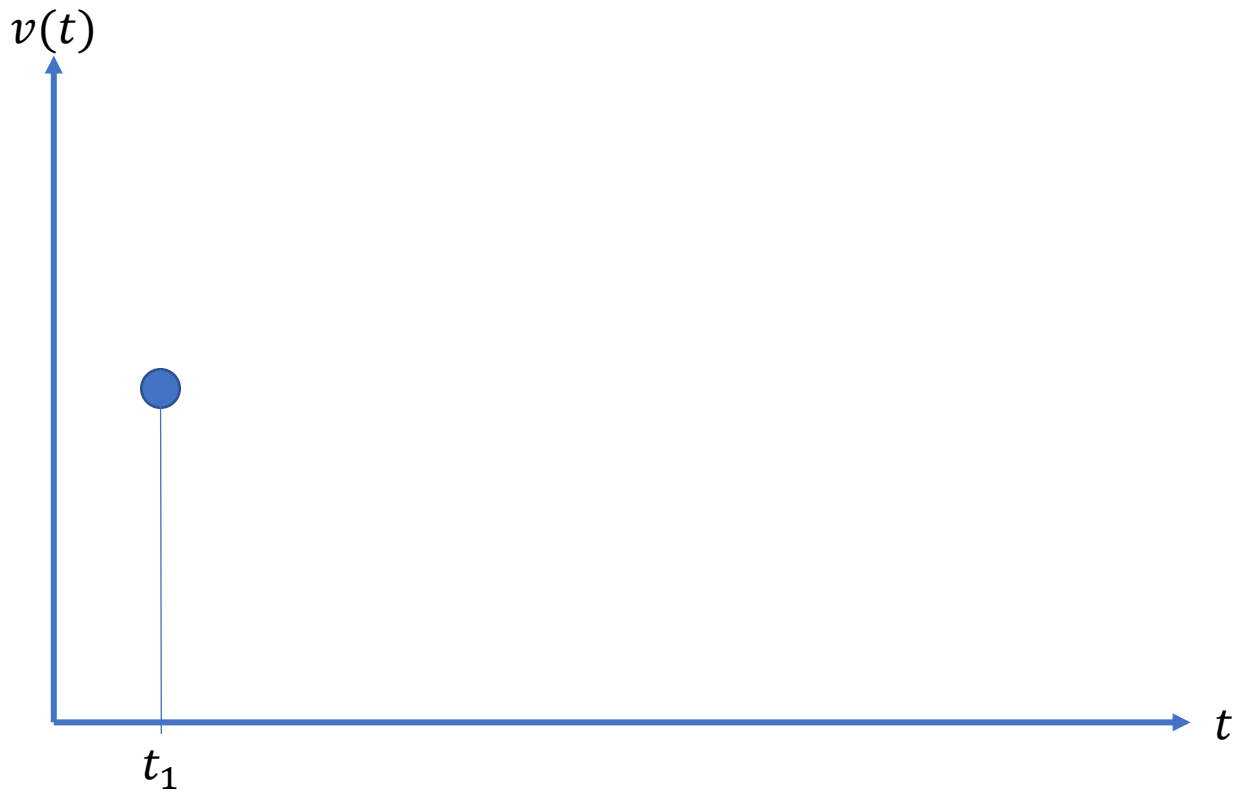
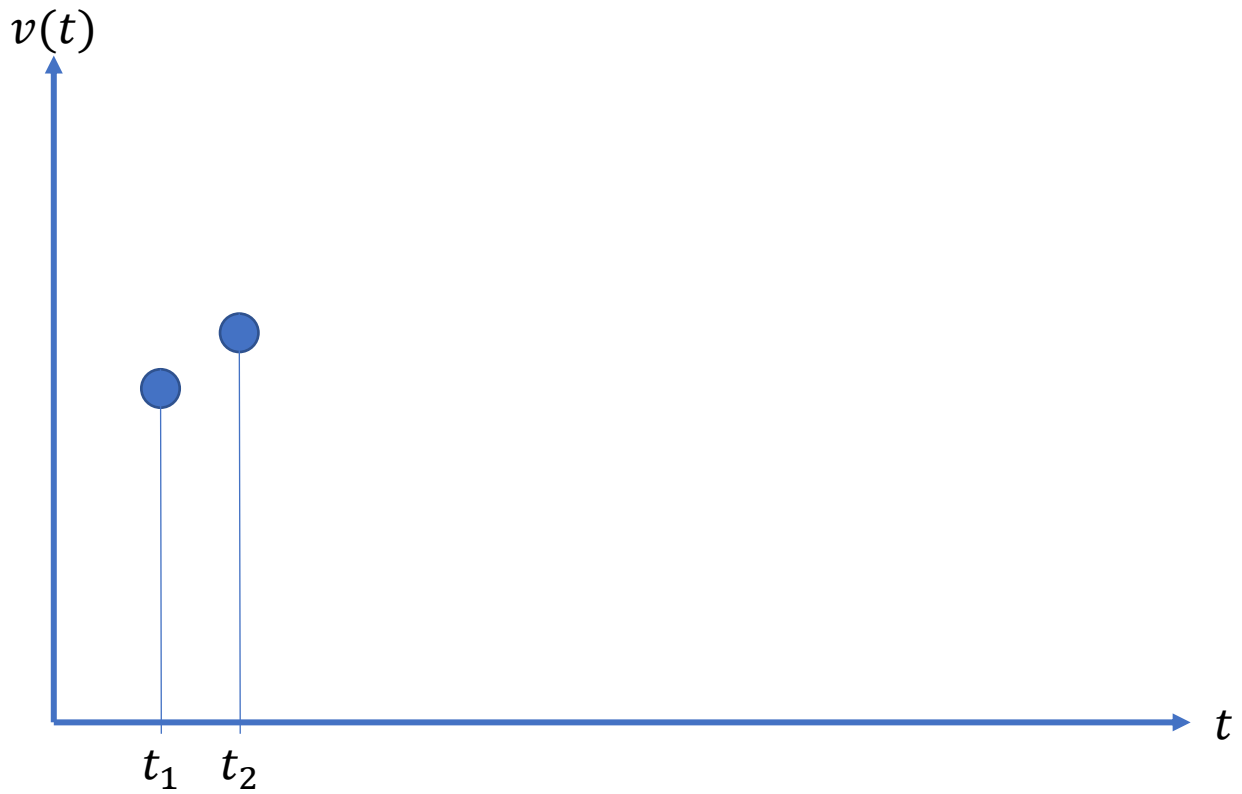
Parameter Estimation

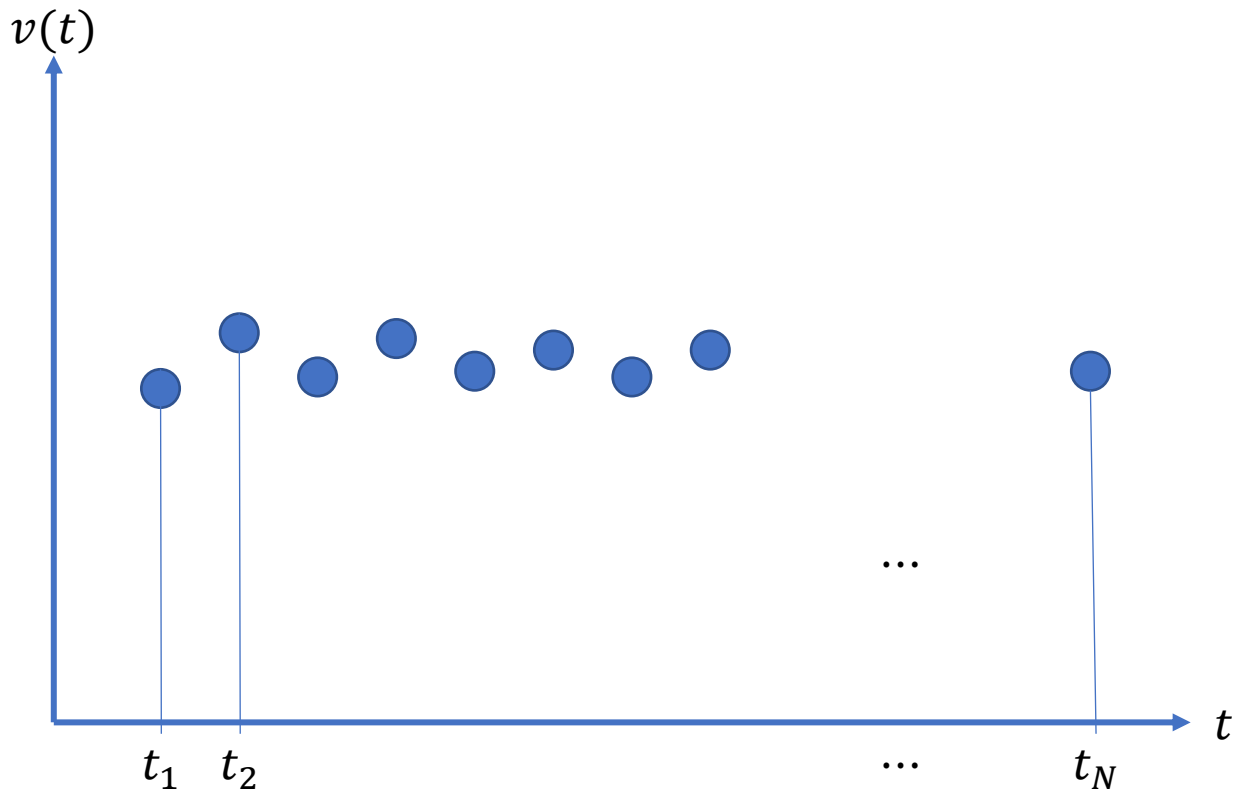
Stochastic Phenomena

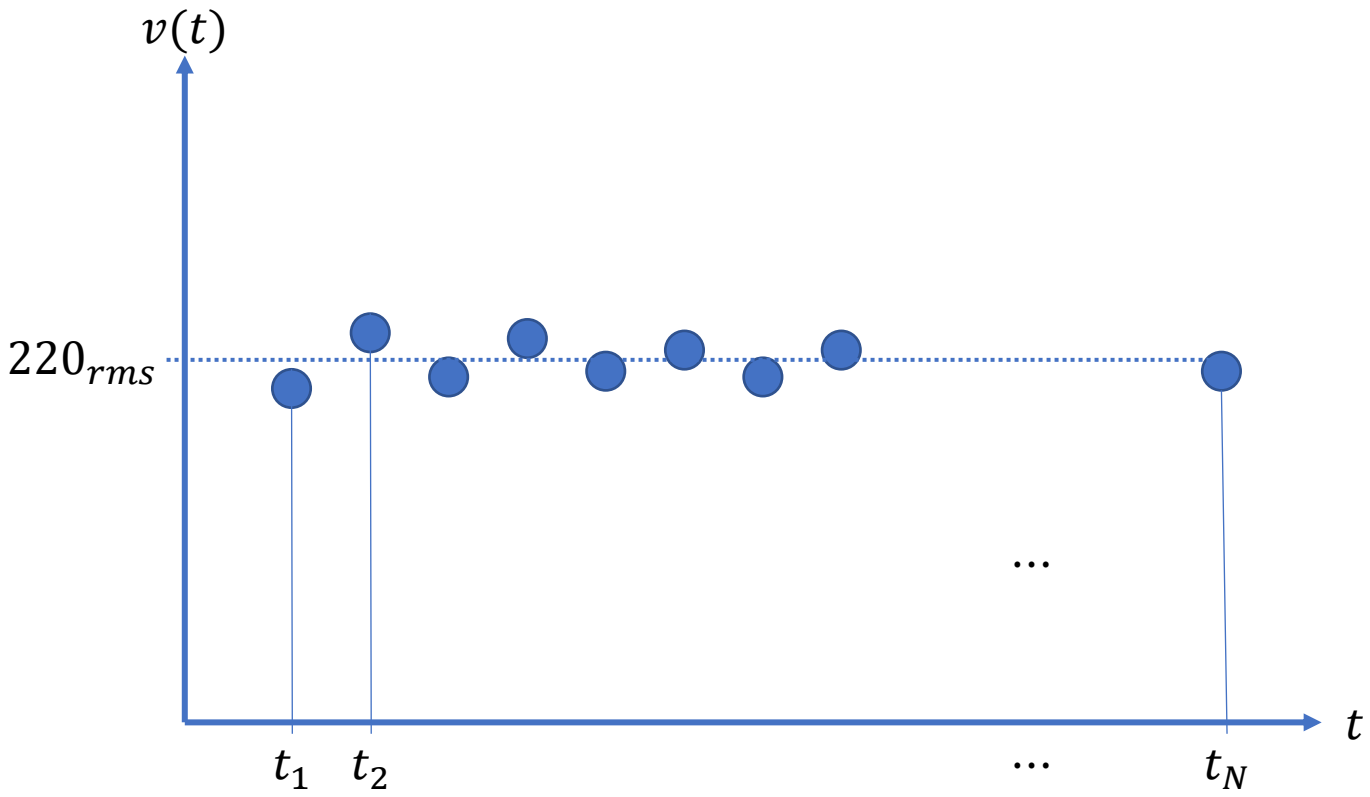
- Daily temperature in a city.
- Depth of a river.
- Voltage in an electric socket.

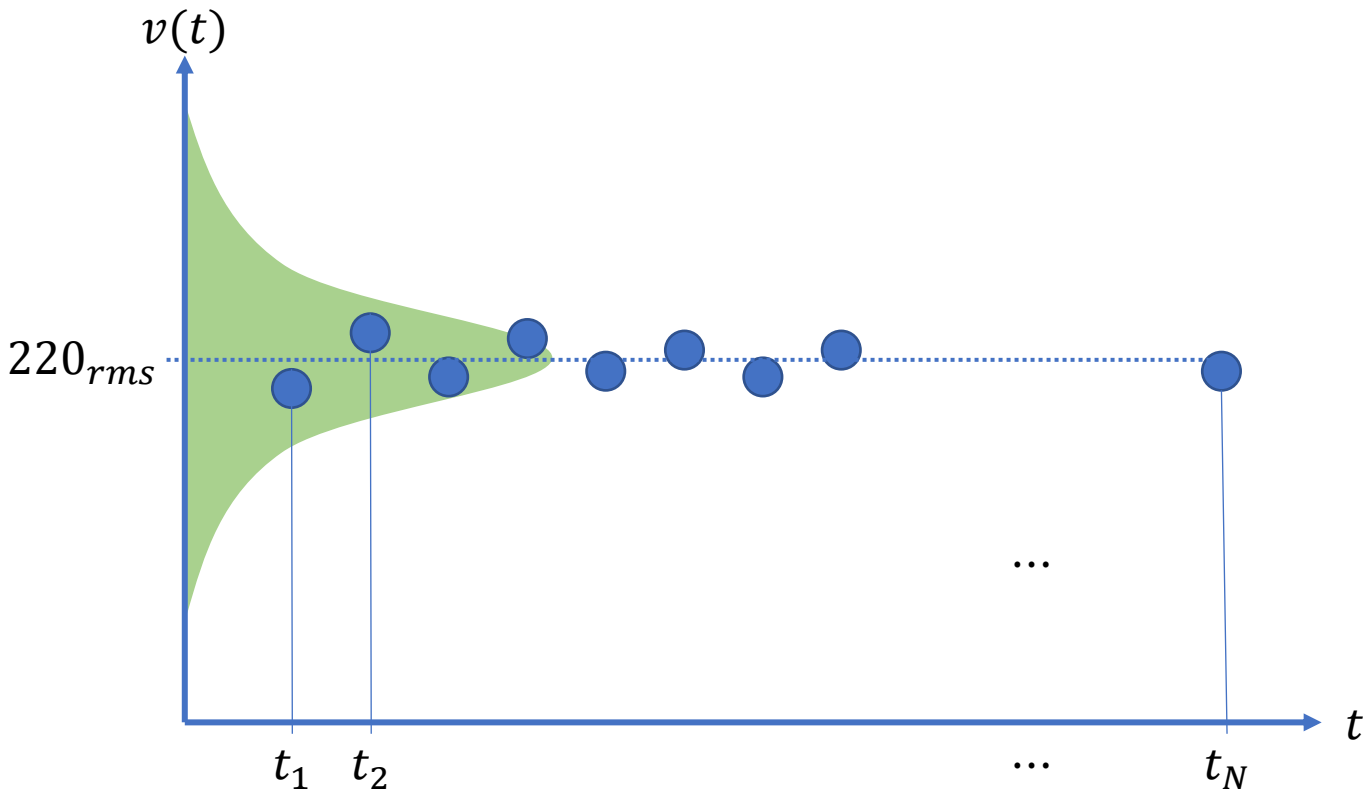

$v(t)$
t


$v(t)$
t
t_1

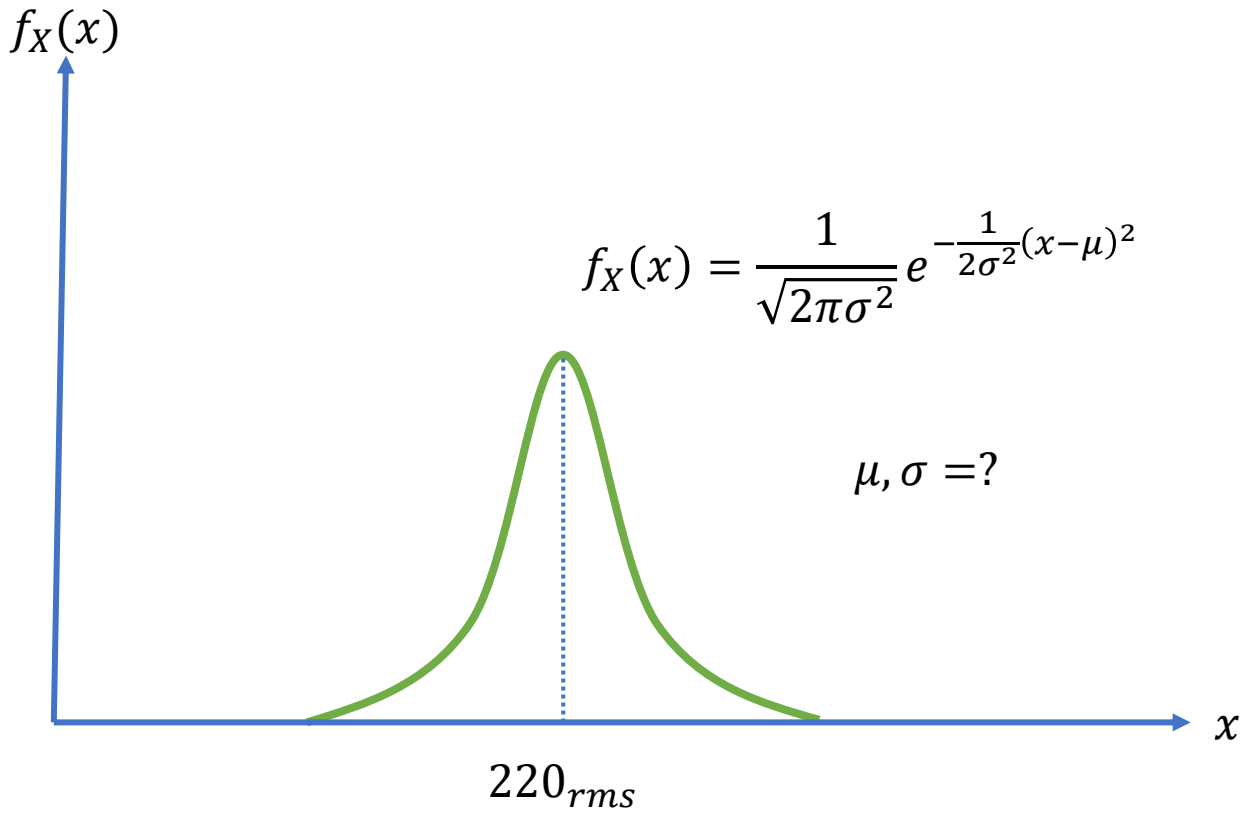

$v(t)$
t
t_1
t_2


$v(t)$
t
t_1
t_2
$\cdots$
$\cdots$
t_N


$v(t)$
220_{rms}
t
t_1
t_2
...
...
t_N


$v(t)$
220_{rms}
t
t_1
t_2
...
...
t_N

- The actual function $f(t)$ originating such observations is practically impossible to determine in a general.
- In some (simple) cases, we can determine $f(t)$.
- In general, though: We make some assumptions, s.a.:
  - Assume a “shape” (e.g. Gaussian).
  - Estimate the “parameters” of said shape (e.g., mean and variance of a normal distribution).
- The electric socket is too complex. Therefore, we represent it by means of a probability density function $f_X(x)$.


$f_X(x)$
$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$
μ, σ =?
x
220_{rms}

Maximum Likelihood Estimation

Maximum Likelihood Estimation (MLE)

- How to estimate the parameters of the distribution, given that estimating the distribution itself is so hard?
  - Assume a “shape”.
  - Estimate the “parameters”, e.g., mean and variance.
    - Make observations.
    - Collect samples $x_i, i = 1,2, \dots, N$.
    - Use these samples to estimate the parameters $\Theta = [\theta_1 \quad \theta_2]^T$ of the distribution.

Maximum Likelihood Estimation (MLE)

Problem statement:

- Given $X = \{x_1, x_2, \dots, x_N\}$.
- Obtain the “best”estimation $\widehat{\Theta}$ of real parameters Θ.

Maximum Likelihood Estimation (MLE)

Ideal solution:

- $\widehat{\Theta} = \Theta$.
- Most often unfeasible.
- Instead, consider an error $e = \widehat{\Theta}(X) - \Theta$.
- We want to make $E[e] \mapsto 0$ and its variance as small as possible by coming up with the best estimator $\widehat{\Theta}$ (according to the MLE criterion).
- How do we define a satisfactory estimator?

Maximum Likelihood Estimation (MLE)

Possible strategy:

- Minimization of the Mean Squared Error (MSE).
- However, we cannot directly attribute a “minimum value” to a random variable.

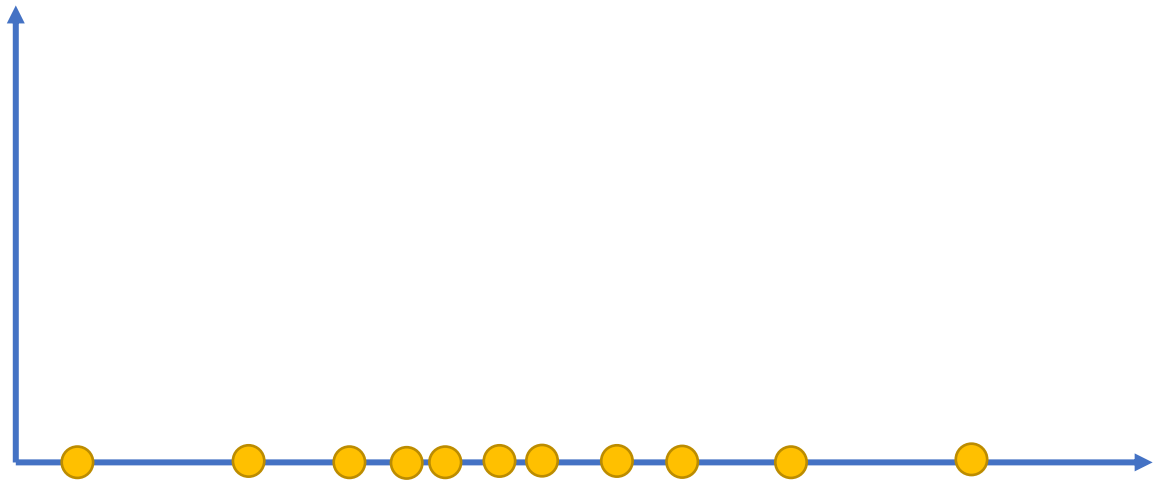
Maximum Likelihood Estimation (MLE)

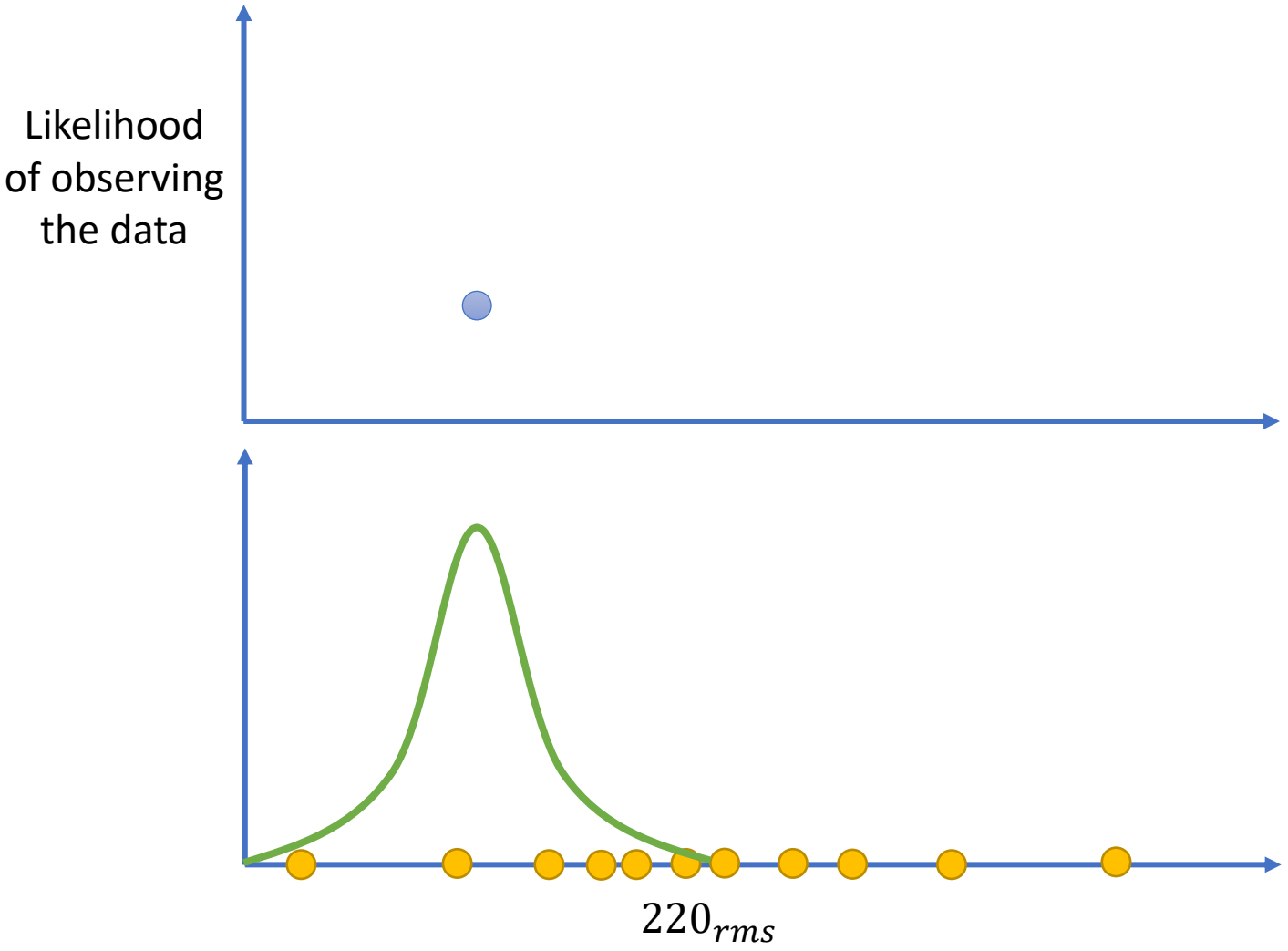
A more fundamental approach:

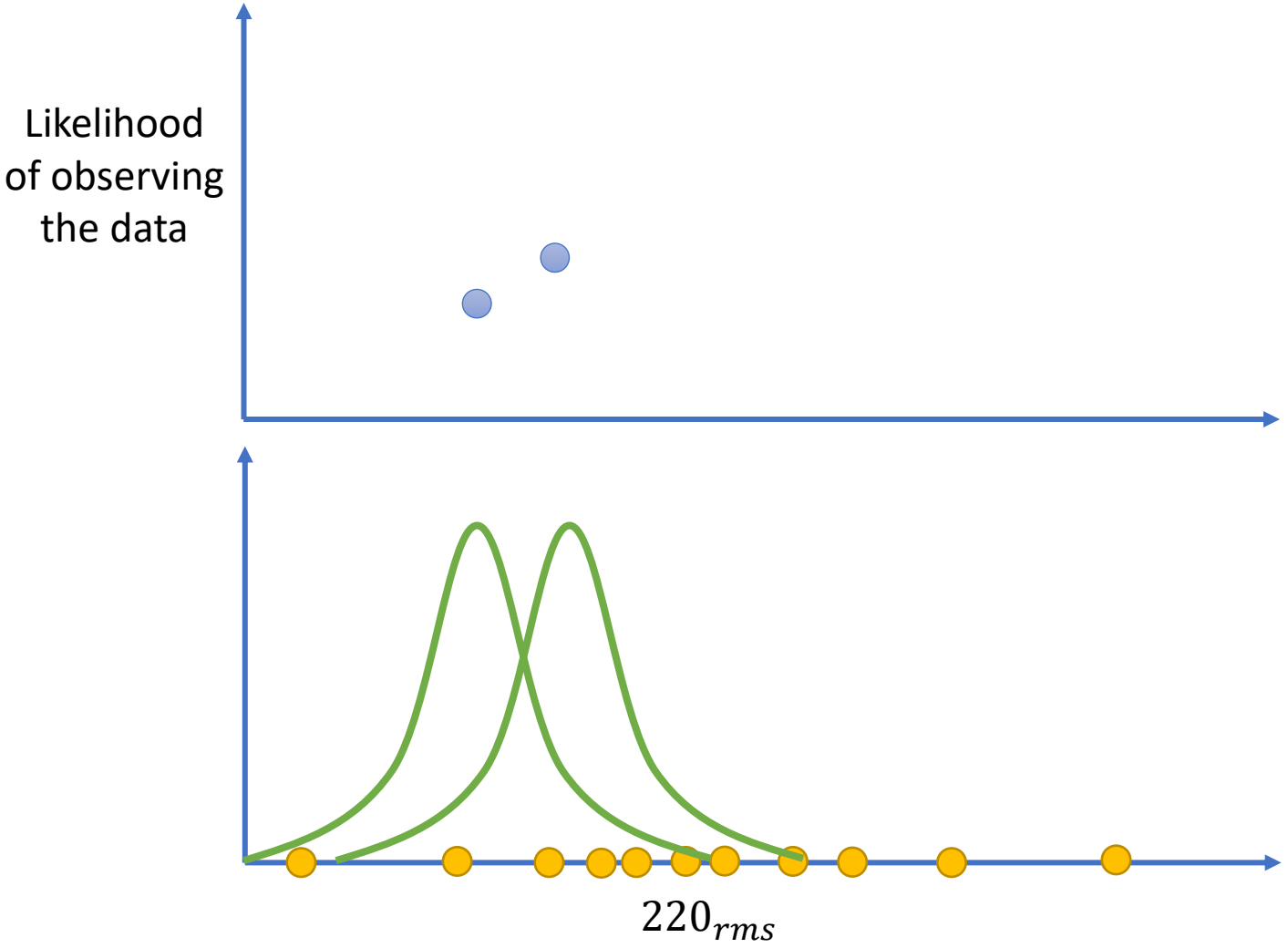
- Maximum Likelihood Estimation.

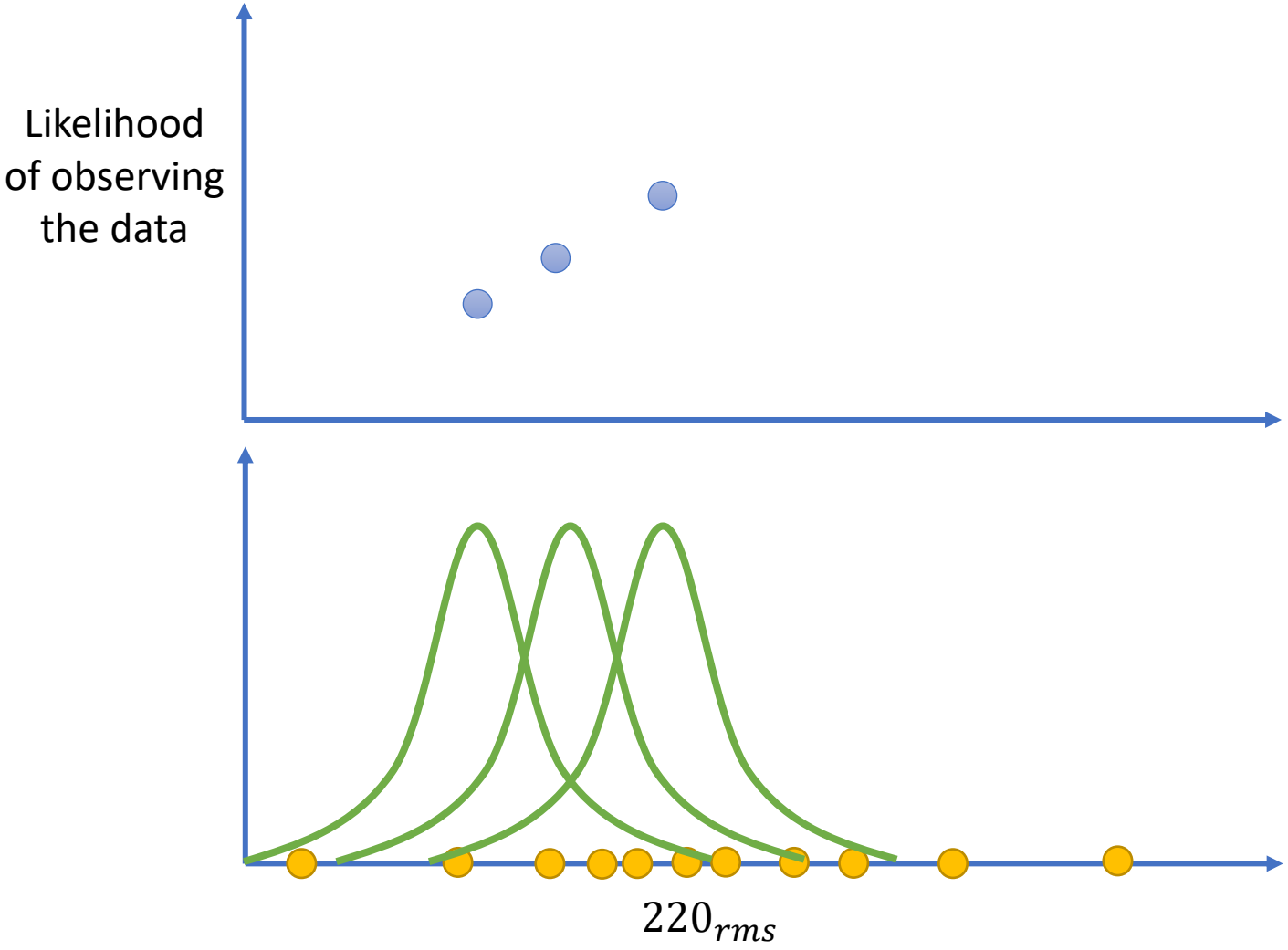
Steps:

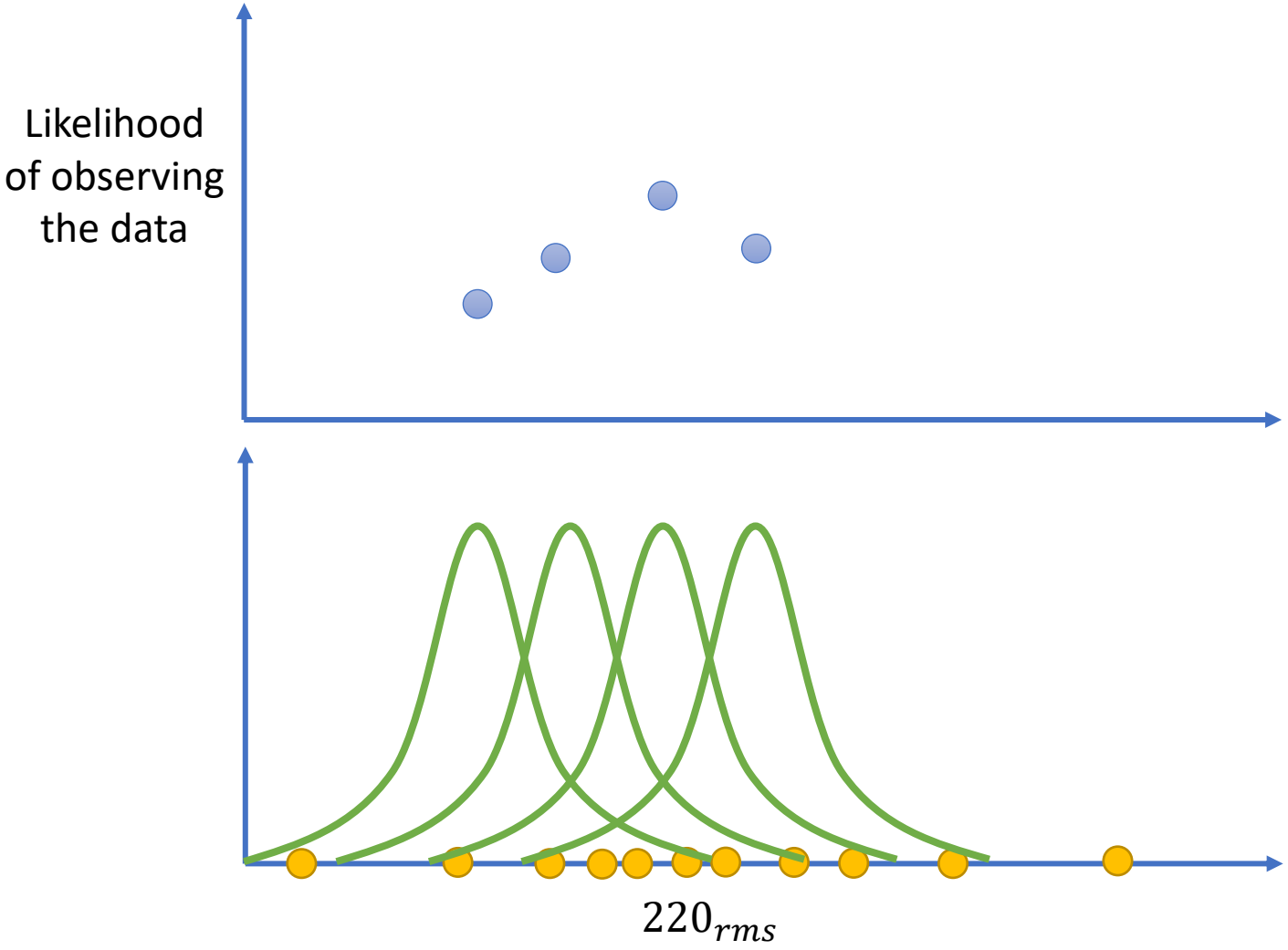
1. Make INDEPENDENT experiments.
2. Estimate the parameters of a supposedly given distribution.

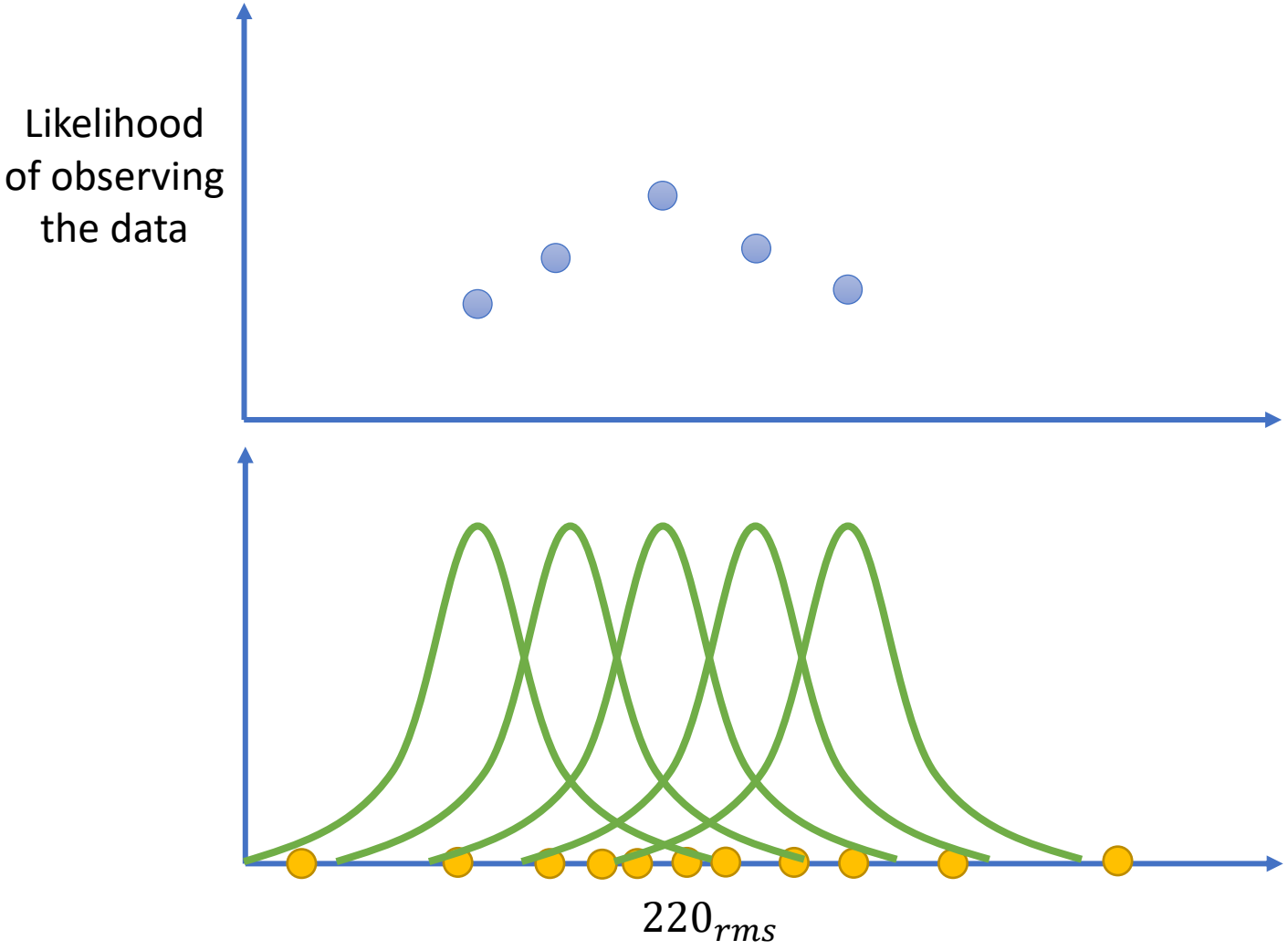

220_{rms}

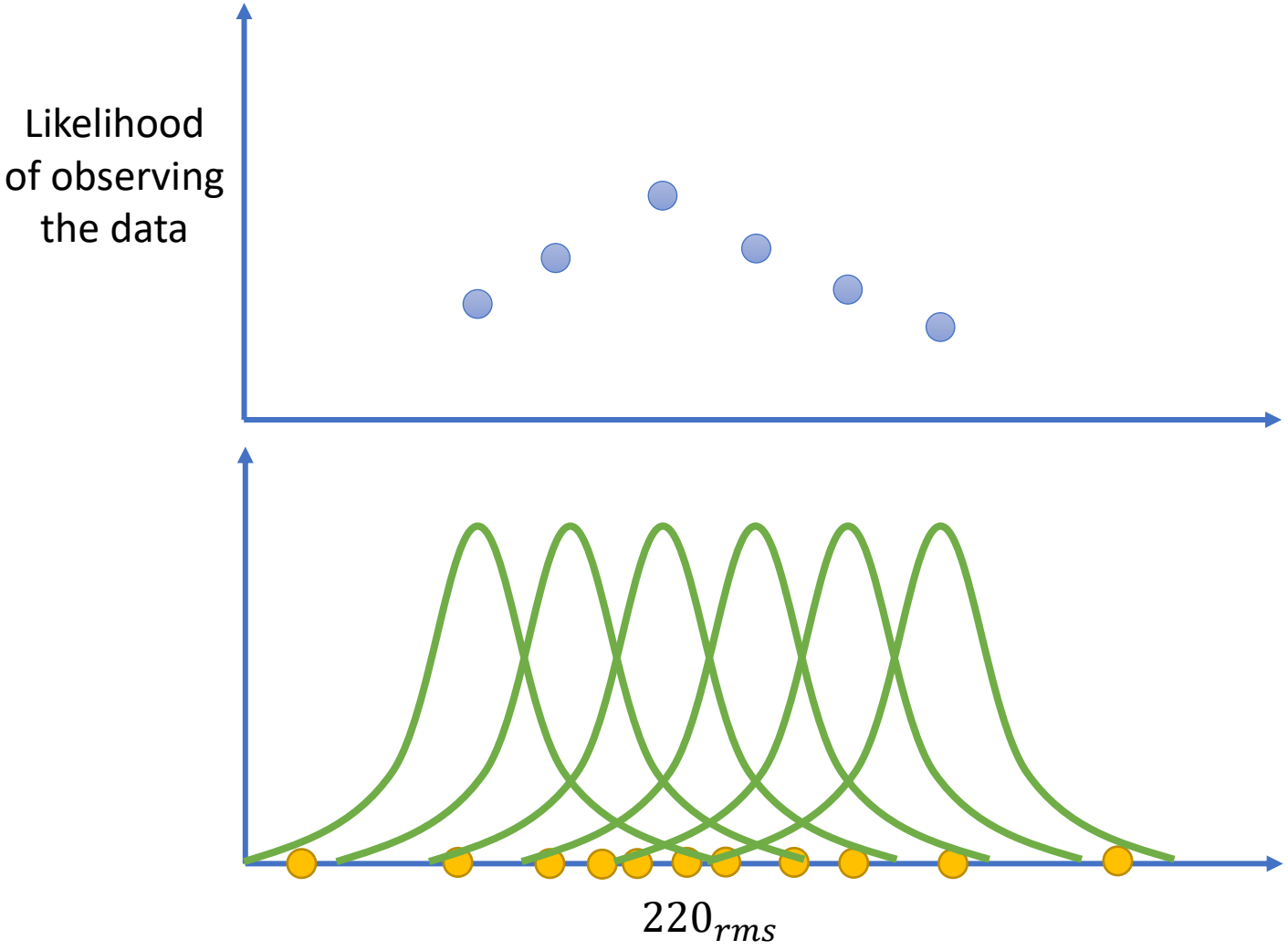

Likelihood
of observing
the data
220_{rms}

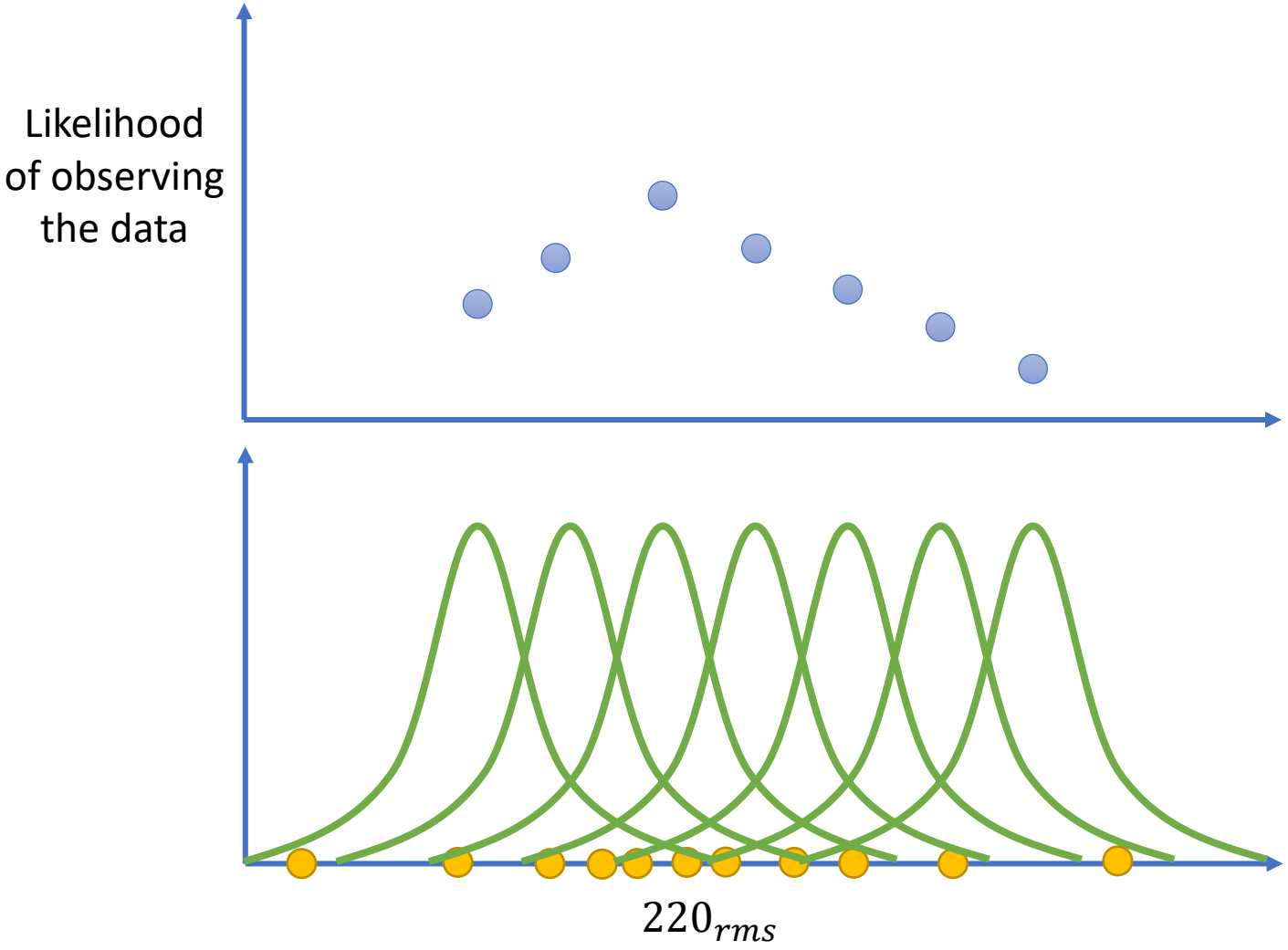

Likelihood
of observing
the data
220_{rms}

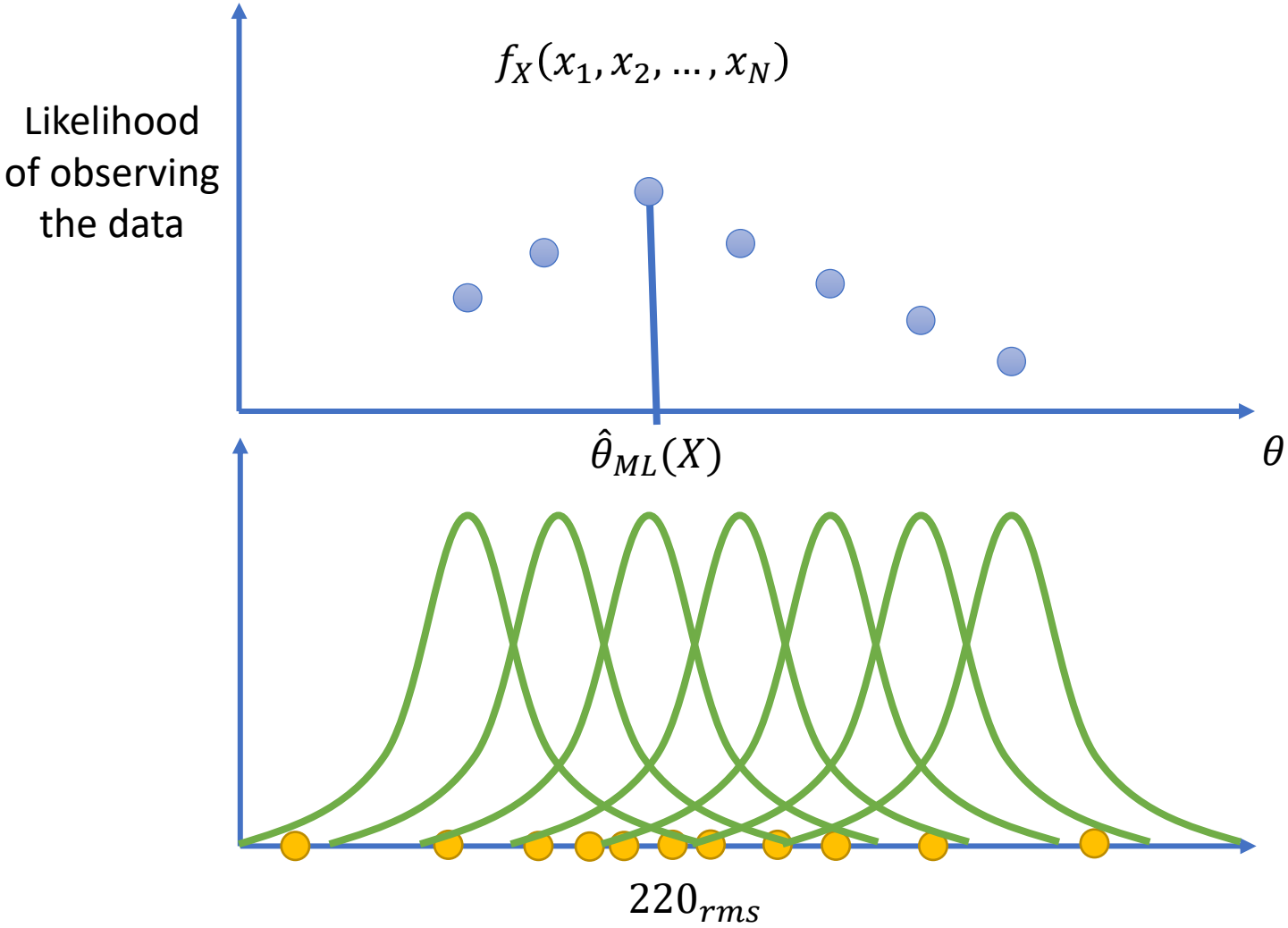

Likelihood
of observing
the data
220_{rms}

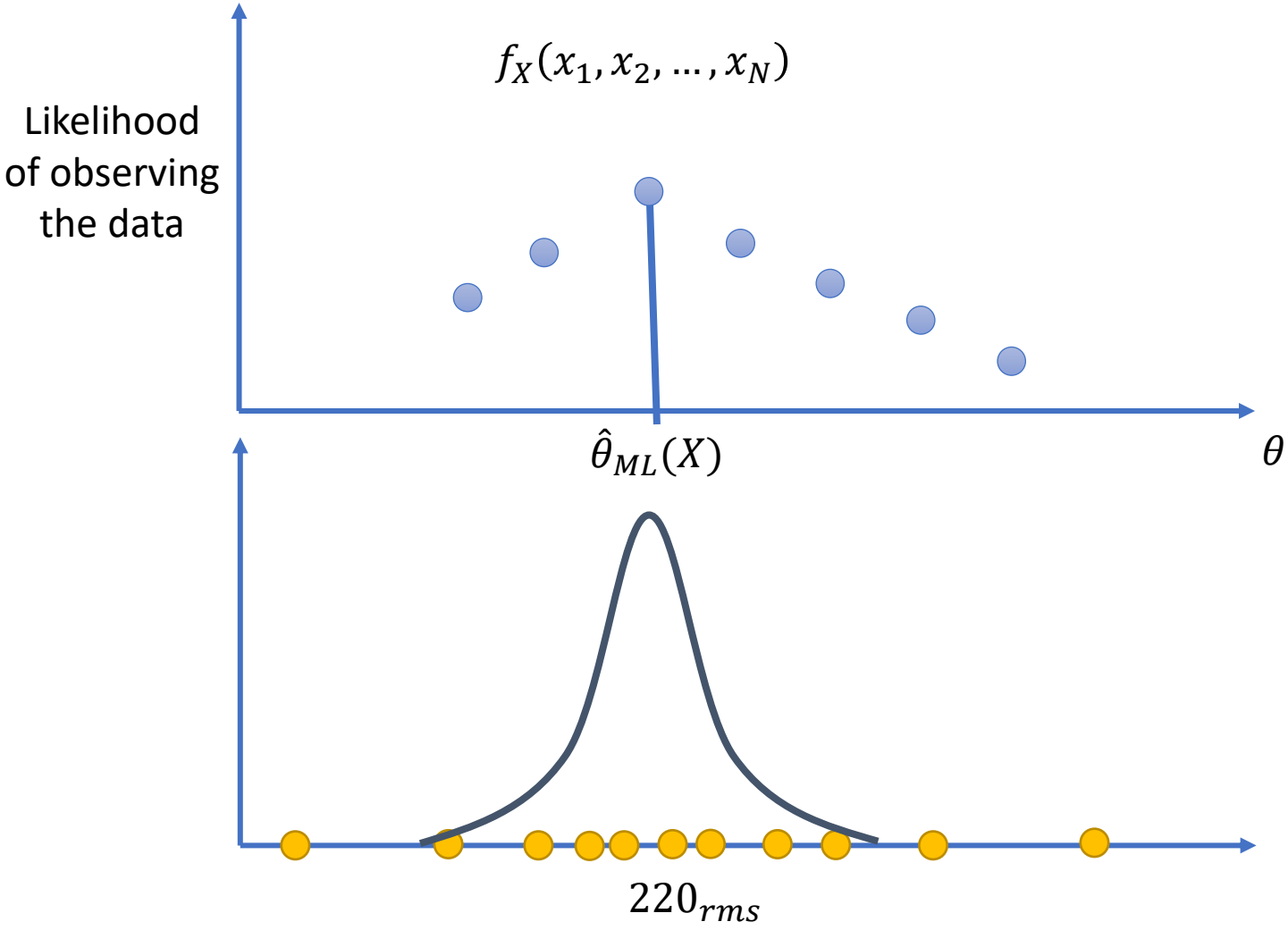

Likelihood
of observing
the data
220_{rms}

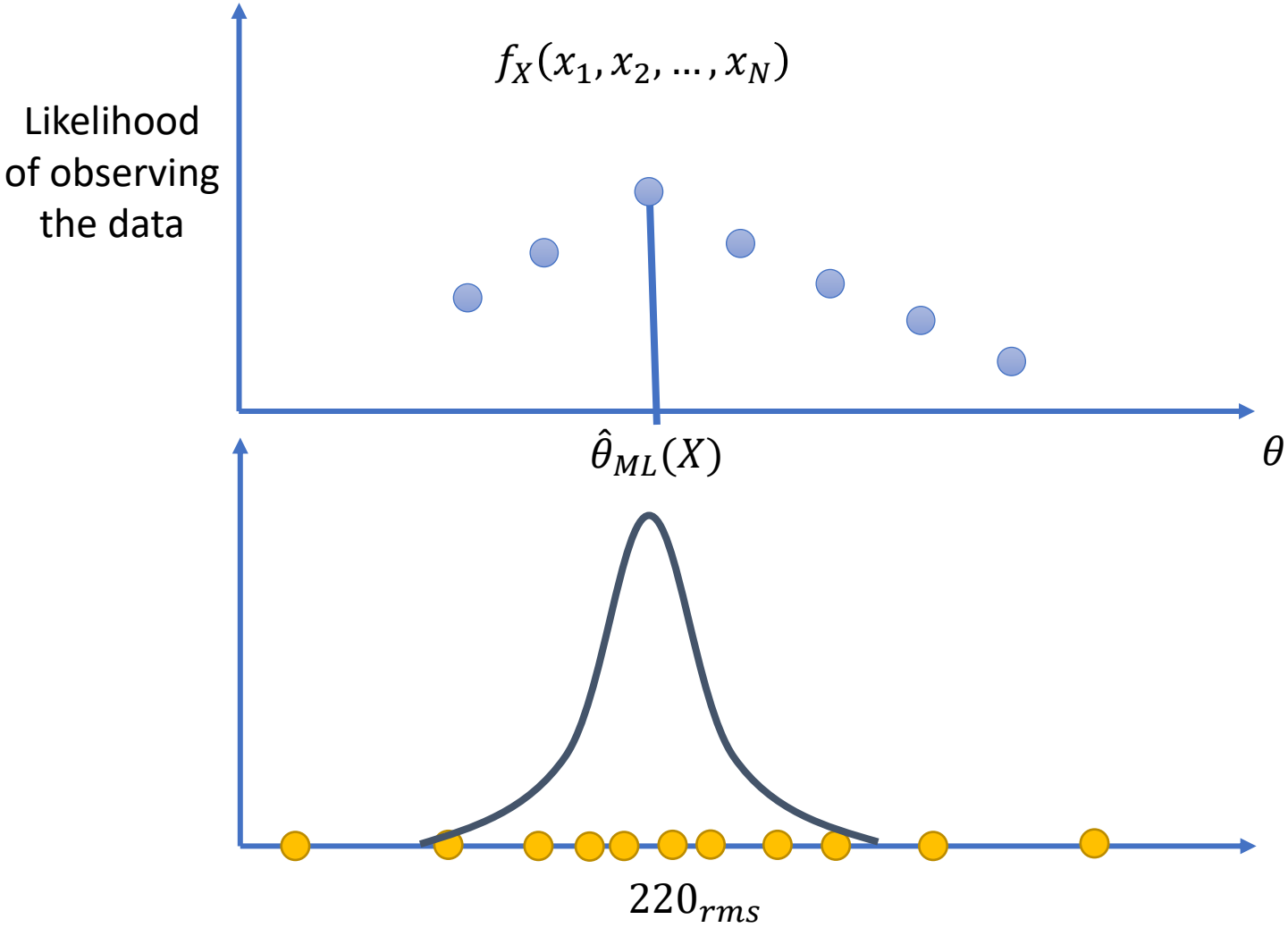

Likelihood
of observing
the data
220_{rms}


Likelihood
of observing
the data
220_{rms}


Likelihood
of observing
the data
220_{rms}


Likelihood of observing the data
$f_X(x_1, x_2, \ldots, x_N)$
$\hat{\theta}_{ML}(X)$
θ
220_{rms}


$f_X(x_1, x_2, \ldots, x_N)$
Likelihood of observing the data
$\hat{\theta}_{ML}(X)$
θ
220_{rms}



- ML tells us that this density best "adapts" to the samples.
- Probability of being this distribution: $P_1 \cdot P_2 \cdots P_N$.

Suppositions

- Independent samples.

Summary

The method of maximum likelihood assumes that the given sample data set is representative of the population $f_X(x_1, x_2, \cdots, x_N; \theta)$ *, and chooses that value for* θ *that most likely caused the observed data to occur, i.e., once observations* $x_1, x_2, \cdots, x_N$ *are given,* $f_X(x_1, x_2, \cdots, x_N; \theta)$ *is a function of* θ *alone, and the value of* θ *that maximizes the above probability distribution function it the most likely value for* θ*, and it is chosen as the ML estimate* $\hat{\theta}$ *for* θ*.*

MLE

$$\hat{\theta} = \sup_{\widehat{\theta}_{ML}} f_X(x_1, x_2, \cdots, x_N; \theta)$$

Since we know that samples are independent, the joint probability is the product of individual probabilities, i.e.;

$$f_X(x_1, x_2, \cdots, x_N; \theta) = f_X(x_1; \theta) f_X(x_2; \theta) \cdots f_X(x_N; \theta) = \prod_{i=1}^{N} f_X(x_i; \theta)$$

$$\mathcal{L}(\theta) = \prod_{i=1}^{N} f_X(x_i; \theta)$$

x_i are samples and θ is the variable.

MLE

We can make $\log f_X(x_i;\theta)$ without altering the position $\hat{\theta}_{ML}$ in which $f_X(x_i;\theta)$ is maximized, i.e.;

$$\mathcal{L}(\theta) = \log f_X(x_i;\theta) = \log \prod_{i=1}^{N} f_X(x_i;\theta)$$

hence

$$\mathcal{L}(\theta) = \sum_{i=1}^{N} \log f_X(x_i;\theta)\,.$$

And to maximize;

$$\frac{\partial}{\partial\theta}\left[\log f_X(x_1, x_2, \cdots, x_N;\theta)\right]_{\theta=\widehat{\theta}} = 0.$$

Example

Assume a normal distribution $f_X(x_i;\mu,\sigma)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$.

Which we can use to compute the likelihood.

$$\mathcal{L}(\theta)=\sum_{i=1}^{N}\left\{\log\left[\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x_i-\theta)^2}{2\sigma^2}}\right]\right\}$$

$$\frac{\partial}{\partial\theta}\mathcal{L}(\theta)=\frac{\partial}{\partial\theta}\sum_{i=1}^{N}\left\{\log\left[\frac{1}{\sqrt{2\pi\sigma^2}}\right]+\log\left[e^{-\frac{(x_i-\theta)^2}{2\sigma^2}}\right]\right\}=0$$

$$\frac{\partial}{\partial\theta}\mathcal{L}(\theta)\Big|_{\theta=\widehat{\theta}}=\sum_{i=1}^{N}\left\{\frac{2(x_i-\theta)}{2\sigma^2}\right\}_{\theta=\widehat{\theta}}=0$$

$$0=\sum_{i=1}^{N}\left(x_i-\hat{\theta}\right)$$

$$0=\sum_{i=1}^{N}x_i-\sum_{i=1}^{N}\hat{\theta}$$

$$0=\sum_{i=1}^{N}x_i-N\hat{\theta}$$

$$\boldsymbol{\widehat{\theta}=\frac{1}{N}\sum_{i=1}^{N}x_i}$$

Example

Assume a normal distribution $f_X(x_i;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$.

Now let's estimate the variance:

$$\mathcal{L}(\theta) = \sum_{i=1}^{N} \left\{ \log\left[\frac{1}{\sqrt{2\pi\theta^2}} e^{-\frac{(x_i-\mu)^2}{2\theta^2}} \right] \right\}$$

$$\mathcal{L}(\theta) = \sum_{i=1}^{N} \left\{ \log\left[\frac{1}{\sqrt{2\pi\theta^2}} \right] + \log\left[e^{-\frac{(x_i-\mu)^2}{2\theta^2}} \right] \right\}$$

$$\mathcal{L}(\theta) = \sum_{i=1}^{N} \left\{ -\frac{1}{2}\log[2\pi\theta^2] - \frac{(x_i-\mu)^2}{2\theta^2} \right\}$$

$$\frac{\partial}{\partial\theta}\mathcal{L}(\theta)\Big|_{\theta=\hat{\theta}} = \sum_{i=1}^{N} \left\{ -\frac{1}{2}\left(\frac{4\pi\,\theta}{2\pi\theta^2} \right) + \frac{(x_i-\mu)^2 4\theta}{4\theta^4} \right\}\Big|_{\theta=\hat{\theta}} = 0$$

$$\sum_{i=1}^{N} \left\{ -\frac{1}{\hat{\theta}} + \frac{(x_i-\mu)^2}{\hat{\theta}^3} \right\} = \sum_{i=1}^{N} \left\{ -\hat{\theta}^2 + (x_i-\mu)^2 \right\} = 0$$

$$\boldsymbol{\widehat{\theta}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2}$$

Example

We can evaluate the dependence of the expected value of $\hat{\theta}$ w.r.t. N:

$$E\left[\hat{\theta}_{ML}\right] = E\left[\frac{1}{N}\sum_{i=1}^{N} x_i\right] = \frac{1}{N}\left[\sum_{i=1}^{N} E[x_i]\right] = \theta$$

$$E\left[\hat{\theta}_{ML}\right] = \theta$$

Unbiased **estimator**.

Example

We can do the same thing for the variance:

$$var\left[\hat{\theta}_{ML}\right] = E\left[\left(\hat{\theta}_{ML} - E\left[\hat{\theta}_{ML}\right]\right)^2\right] = E\left[\left(\frac{1}{N}\sum_{i=1}^{N} x_i - \theta\right)^2\right] = E\left[\left(\frac{1}{N}\sum_{i=1}^{N} (x_i - \theta)\right)^2\right]$$

$$var\left[\hat{\theta}_{ML}\right] = \frac{1}{N^2} \cdot E\left[\left(\sum_{i=1}^{N} (x_i - \theta)\right)^2\right]$$

$$\frac{1}{N^2} \cdot E\left[\left(\sum_{i=1}^{N}(x_i - \theta)\right)^2\right]$$

Since $\left(\sum_{i=1}^{N}\alpha_i\right)^2 = \sum_{i=1}^{N}\alpha_i^2 + \sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N}\alpha_i\alpha_j$, we have;

$$var[\hat{\theta}_{ML}] = \frac{1}{N^2} \cdot E\left[\sum_{i=1}^{N}(x_i - \theta)^2 + \sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N}(x_i - \theta)(x_j - \theta)\right]$$

0 (*independent*)

$$var[\hat{\theta}_{ML}] = \frac{1}{N^2} \cdot \left[\sum_{i=1}^{N}E[(x_i - \theta)^2] + \sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N}E[(x_i - \theta)(x_j - \theta)]\right]$$

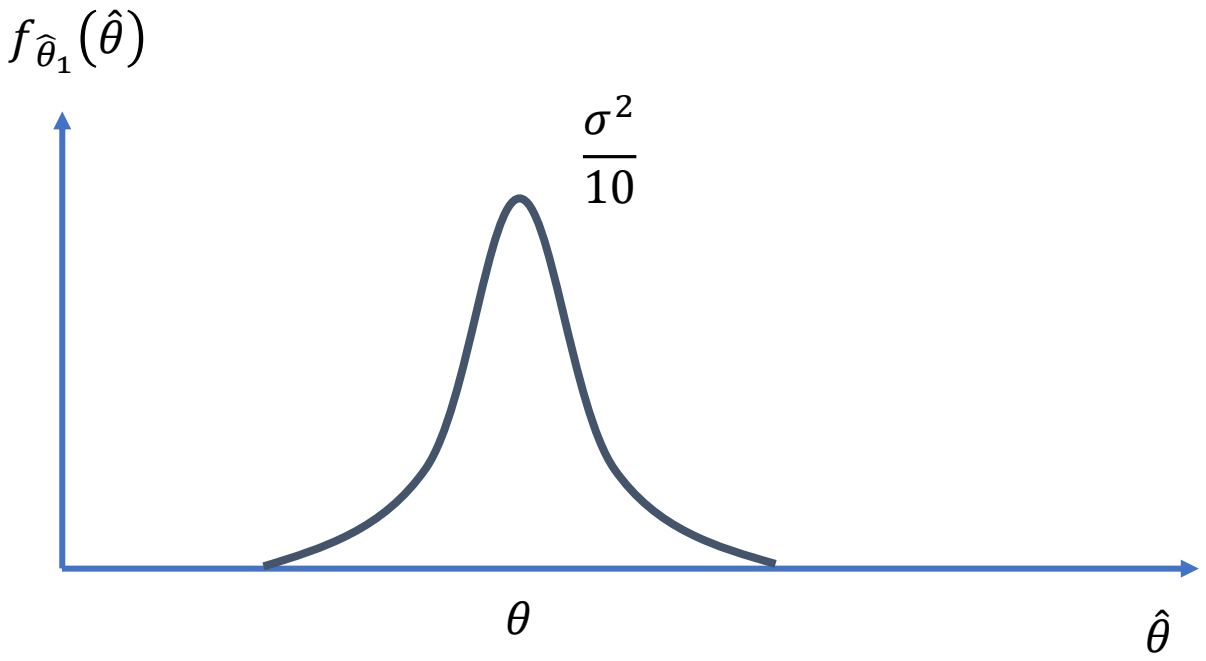
$$var[\hat{\theta}_{ML}] = \frac{1}{N^2} \cdot \sum_{i=1}^{N}E[(x_i - \theta)^2]$$

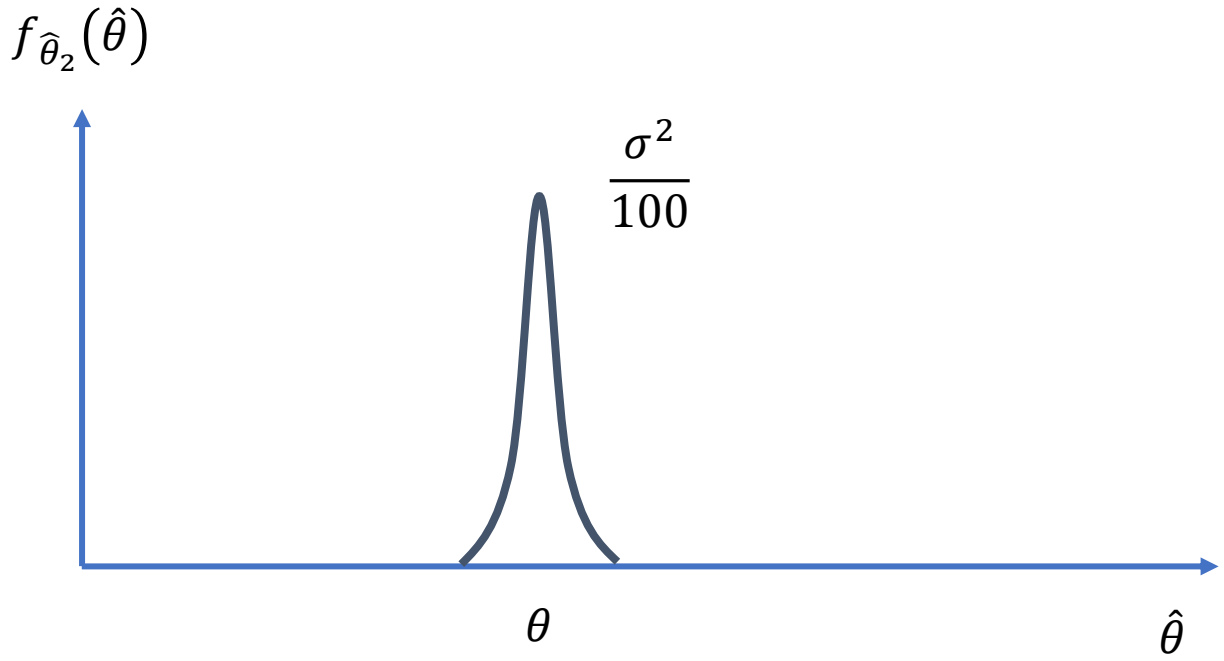
$$var[\hat{\theta}_{ML}] = \frac{1}{N^2} \cdot \sum_{i=1}^{N}E[(x_i - E[x_i])^2] = \frac{1}{N^2} \cdot \sum_{i=1}^{N}Var(x_i) = \frac{\sigma^2}{N}$$

$$\boldsymbol{var[\hat{\theta}_{ML}] = \frac{\sigma^2}{N}}$$

$$\boldsymbol{var}[\hat{\boldsymbol{\theta}}_{\boldsymbol{ML}}] = \frac{\boldsymbol{\sigma^2}}{\boldsymbol{N}}$$

- The variance of the estimator $\hat{\theta}$ depends on the number of samples N.
- Which of the following estimates is best? $N_1 = 10; N_2 = 100$?
  - The higher the number of samples, more certain we are that our measurements are close to the “real”mean.
  - For $N \mapsto \infty$ we have the mean with probability equal to 1.





Thank you