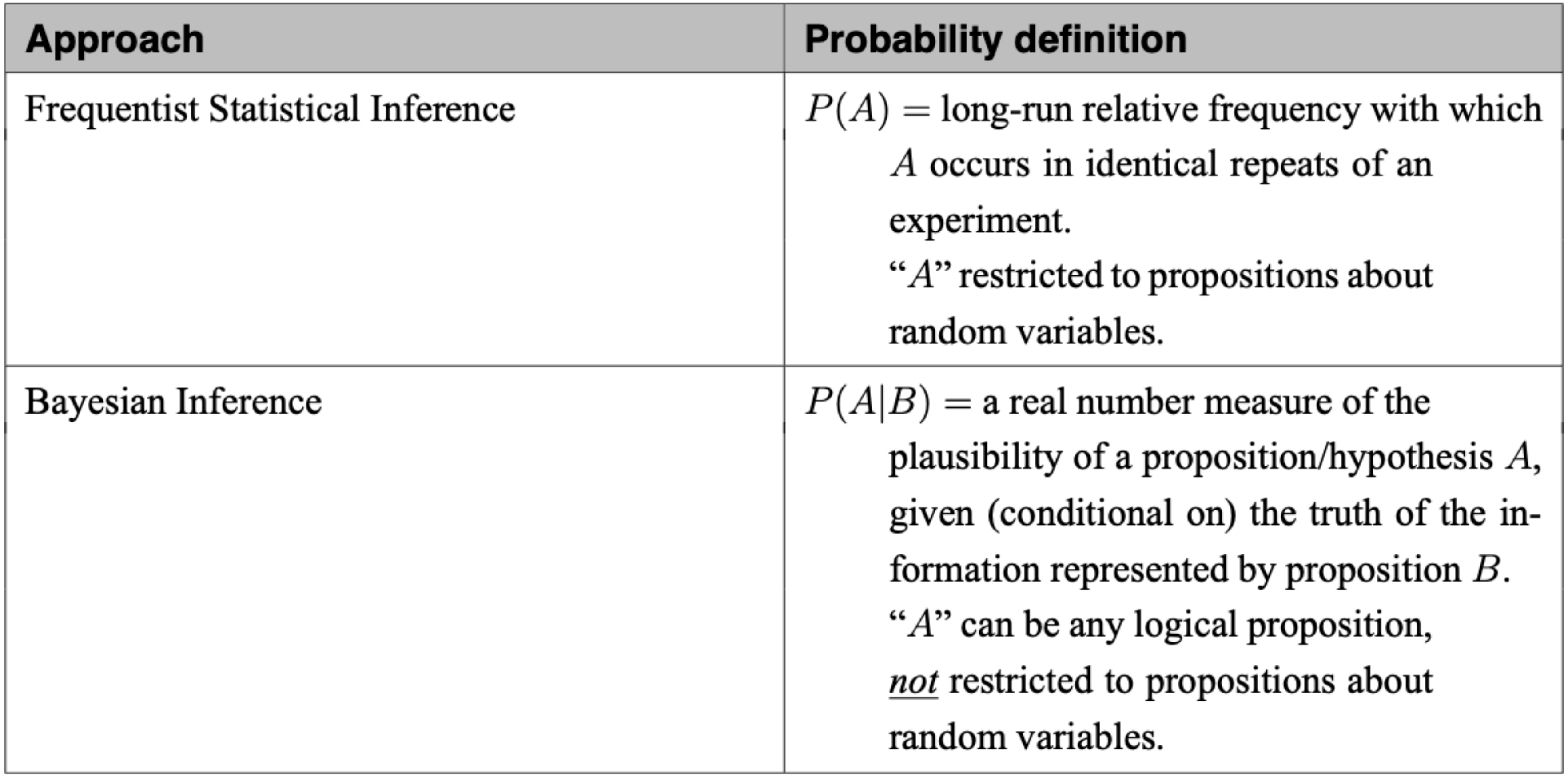
Bayesian Statistics

Teeraparb Chantavat
IF, Naresuan University

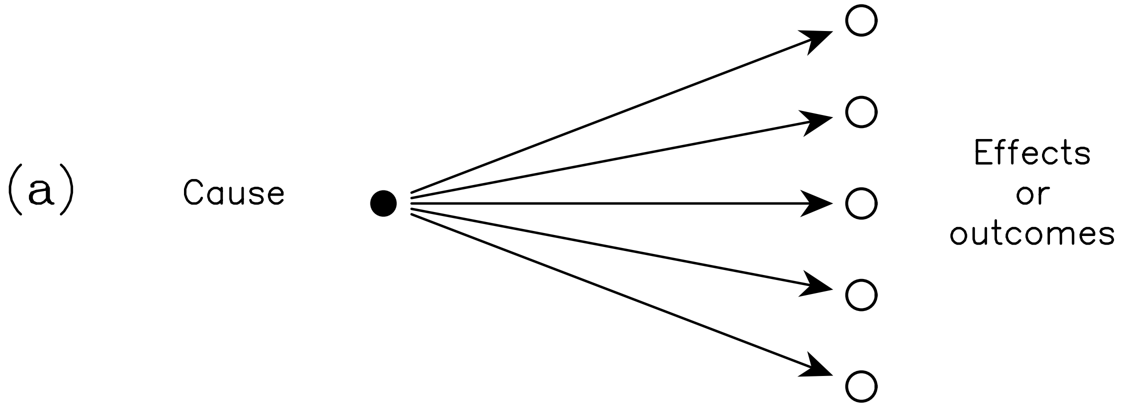
Probability Theory



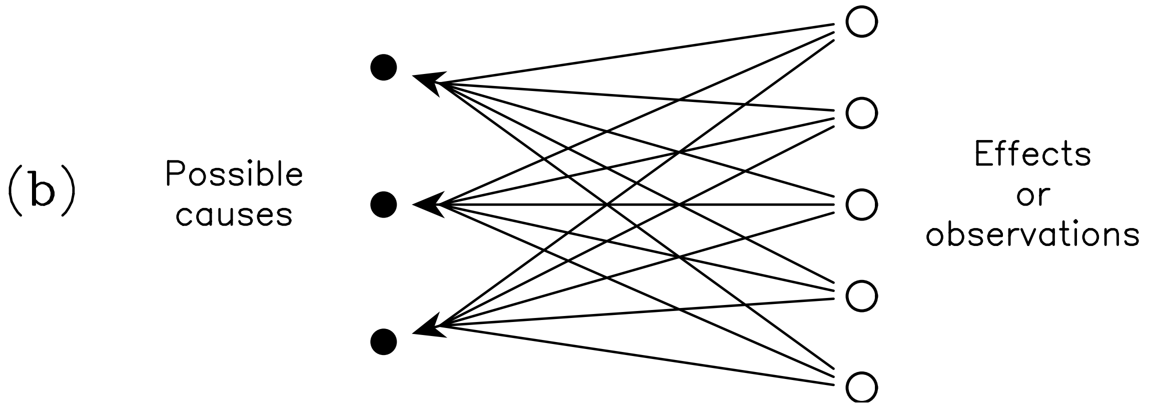
Approach	Probability definition
Frequentist Statistical Inference	$P(A) =$ long-run relative frequency with which A occurs in identical repeats of an experiment. "A" restricted to propositions about random variables.
Bayesian Inference	$P(A\|B) =$ a real number measure of the plausibility of a proposition/hypothesis A, given (conditional on) the truth of the information represented by proposition B. "A" can be any logical proposition, ***not*** restricted to propositions about random variables.

Bayesian Statistics

Deductive Logic



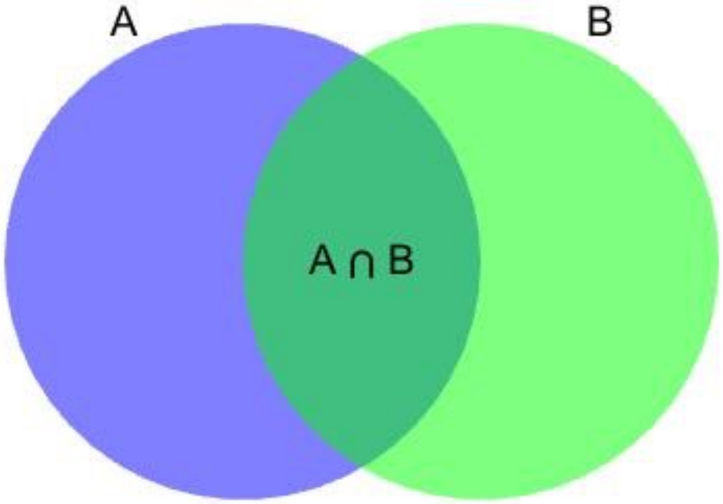
Inductive Logic



Bayesian Statistics

Conditional Probability

$$P(A\,|\,B) = \frac{P(A \cap B)}{P(B)}$$



$P(A)$	Observing the data.
$P(B)$	The theory is true.
$P(A\,\|\,B)$	The data is observed given that the theory is true

Bayesian Statistics

Symmetry Rule

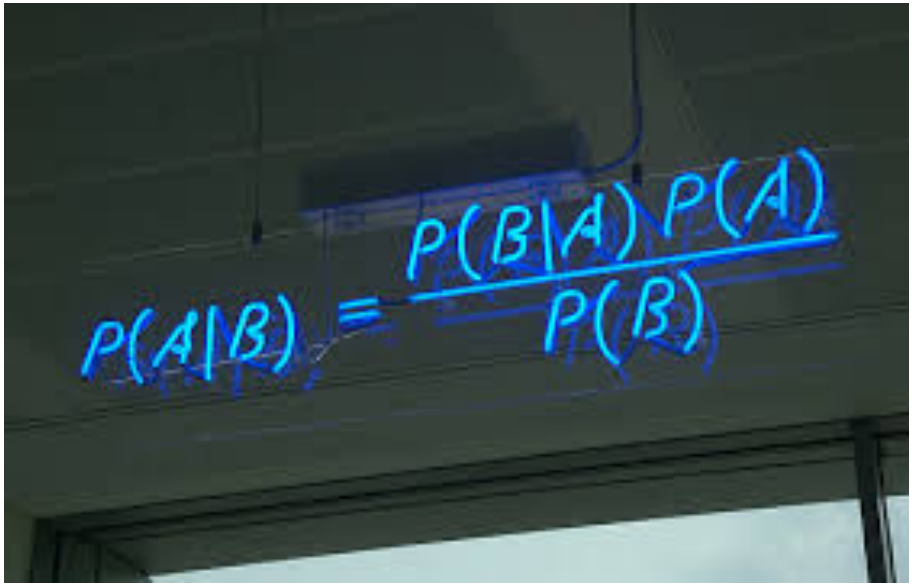
$$P(B \cap A) = P(A \cap B)$$

$$P(B \mid A)P(A) = P(A \mid B)P(B)$$

Bayesian Rule

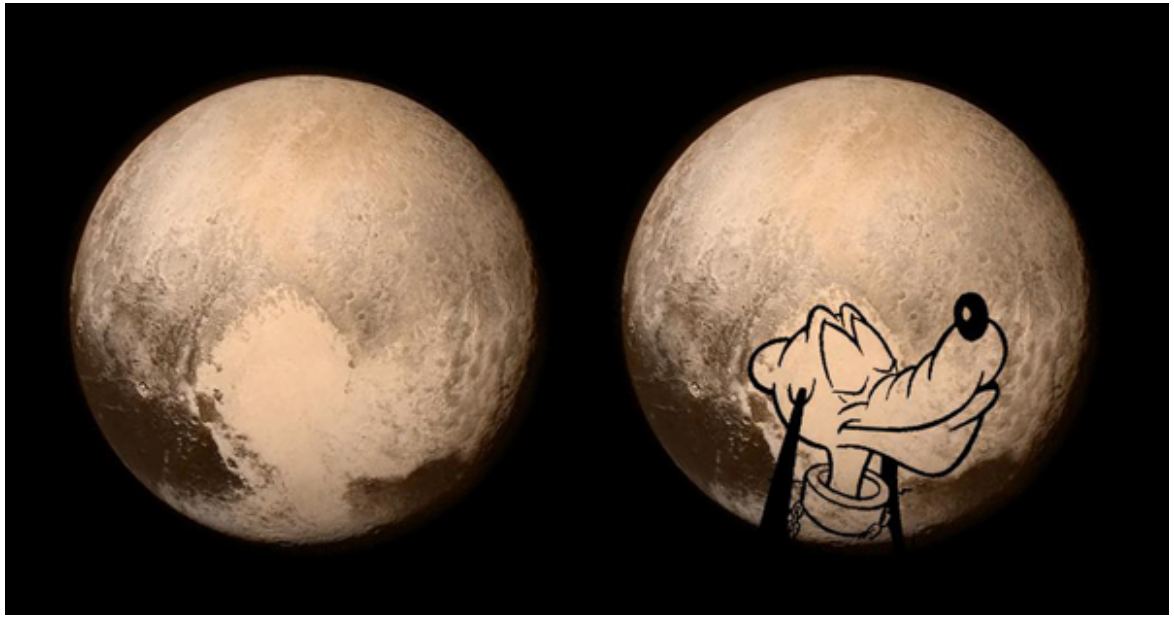
$$P(B \mid A) = \frac{P(A \mid B)P(B)}{P(A)}$$

Bayesian Statistics



"There are no problems left in statistics except the assessment of probability"

Bayesian Statistics



$$P(A\,|\,B) \neq P(B\,|\,A)$$

Bayesian Statistics

$$P(H\,|\,D) = \frac{P(D\,|\,H)P(H)}{P(D)}$$

$P(H)$	Probability that the hypothesis is true.	**Prior**
$P(D\,|\,H)$	Probability that the data is observed given that the hypothesis is true.	**Likelihood**
$P(D)$	Probability that the collections of data is liable.	**Evidence**
$P(H\,|\,D)$	Probability that the hypothesis is true given that the data is true.	**Posterior**

Bayesian Statistics

- A theory usually have many parameters, for example, a two-parameter model

$$\boldsymbol{\Theta} = \{\Theta_1, \Theta_2\}$$

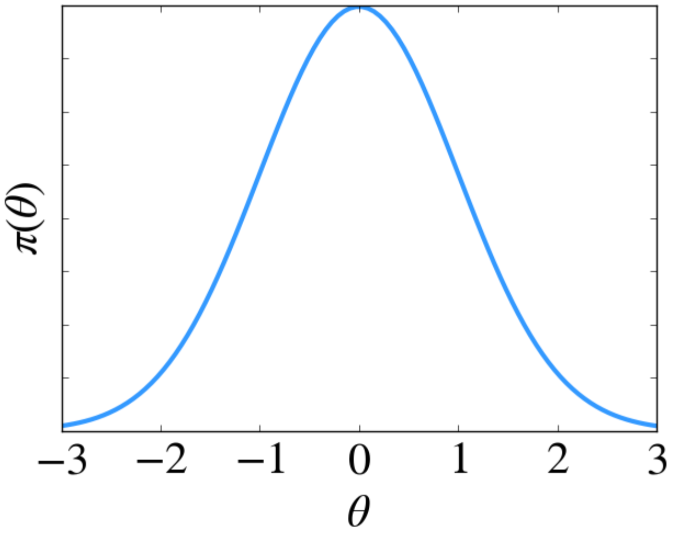
- The **hypothesis** is the assumption that the parameter have a particular value
for example

$$\mathrm{H}_1 : \quad \theta_1 = 1.0 \quad \text{and} \quad \theta_2 = 2.0$$

$$\mathrm{H}_1 \equiv \boldsymbol{\theta}_1 = (\theta_1, \theta_2)$$

Bayesian Statistics

- The prior probability is the distribution of the parameters we know before the experiment **(degree of believe)**.
- We can have a **uniform distribution** for total ignorance or a **normal distribution** if **mean** and **standard deviation** are given.



Bayesian Statistics

- The **evidence** is usually considered as a normalization constants — nothing to do with **parameter estimations**.

$$P(\boldsymbol{\theta}\,|\,\boldsymbol{x}) \propto \mathcal{L}(\boldsymbol{x}\,|\,\boldsymbol{\theta})\; \pi(\boldsymbol{\theta})$$

- However, the **evidence** is important **model comparison**.

Bayesian Statistics

- In most cases, we are working the logarithm of the likelihood function called **log-likelihood**

$$L(\boldsymbol{x}\,|\,\boldsymbol{\theta}) = \log_{\mathrm{e}} \mathcal{L}(\boldsymbol{x}\,|\,\boldsymbol{\theta})$$

- Expanding around the maximum of the **log-likelihood** at θ_0 i.e.

$$\left.\frac{\partial L}{\partial \theta_\alpha}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = 0$$

$$L(\boldsymbol{x}\,|\,\boldsymbol{\theta}) = L(\boldsymbol{x}\,|\,\boldsymbol{\theta}_0) + \frac{1}{2}\sum_{\alpha,\beta} \left.\frac{\partial L}{\partial\,\theta_\alpha}\frac{\partial L}{\partial\,\theta_\beta}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} (\theta_\alpha - \theta_{\alpha 0})\,(\theta_\beta - \theta_{\beta 0})$$

Bayesian Statistics

- We define the **precision matrix** $\mathbf{P}$ as

$$L(\boldsymbol{x}\,|\,\boldsymbol{\theta}) = L(\boldsymbol{x}\,|\,\boldsymbol{\theta}_0) - \frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^{\mathrm{T}}\cdot\mathbf{P}\cdot(\boldsymbol{\theta}-\boldsymbol{\theta}_0)$$

where

$$\mathrm{P}_{\alpha\beta} \equiv \left.\frac{\partial^2 L}{\partial\,\theta_\alpha\,\partial\,\theta_\beta}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$$

- The **log likelihood** is given by

$$\mathcal{L}(\boldsymbol{x}\,|\,\boldsymbol{\theta}) \propto \exp\left(-\frac{1}{2}\,(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^{\mathrm{T}}\cdot\mathbf{P}\cdot(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\right)$$

Bayesian Statistics

- The inverse of the **precision matrix** is called **covariance matrix**

$$\mathbf{P} \equiv \mathbf{C}^{-1}$$

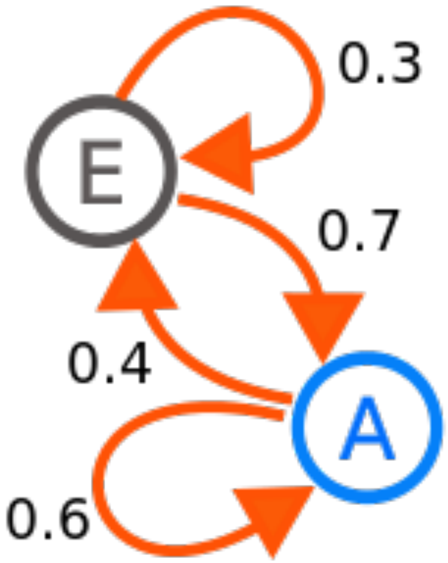
$$\mathcal{L}(\boldsymbol{x}\,|\,\boldsymbol{\theta}) \propto \exp\left(-\frac{1}{2}\,(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^{\mathrm{T}}\cdot\mathbf{C}^{-1}\cdot(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\right)$$

- The variance of the parameter can be estimated as

$$\mathrm{Var}\,(\theta_\alpha) = \mathrm{C}_{\alpha\,\alpha}$$

Markov Chain Monte Carlo (MCMC)

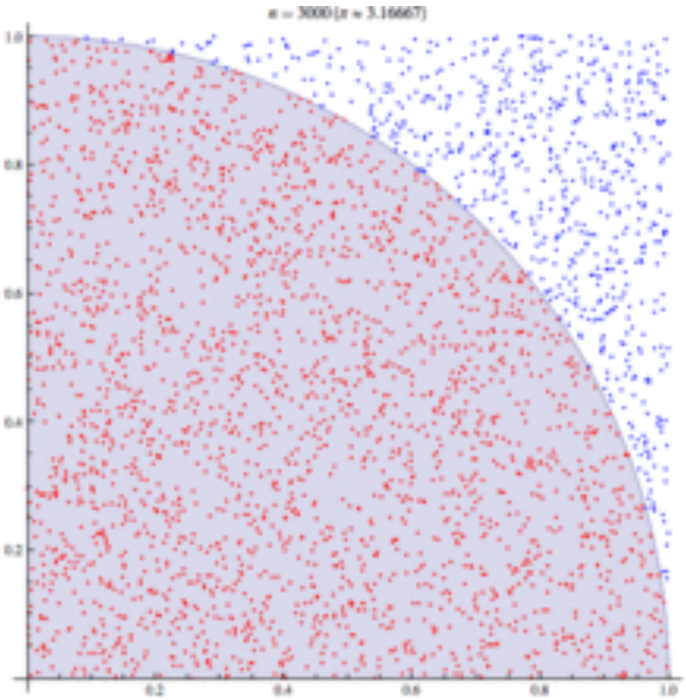
A **Markov chain** is a chain of states in a parameter space that is "memoryless" (Markov property).

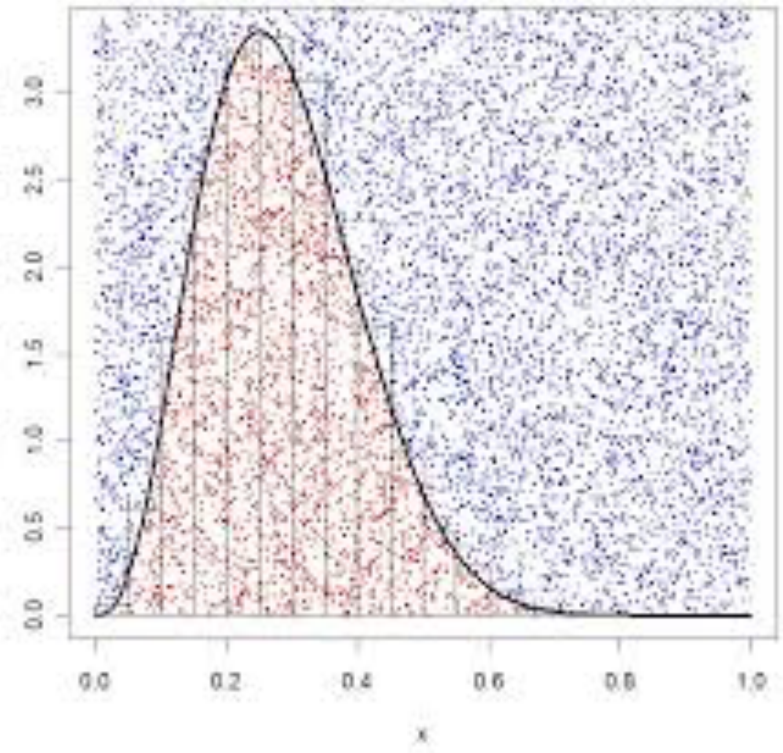


How the state change depends only on the **current state**.

Markov Chain Monte Carlo (MCMC)

A **Monte Carlo** is a method using random walk to generate the output.

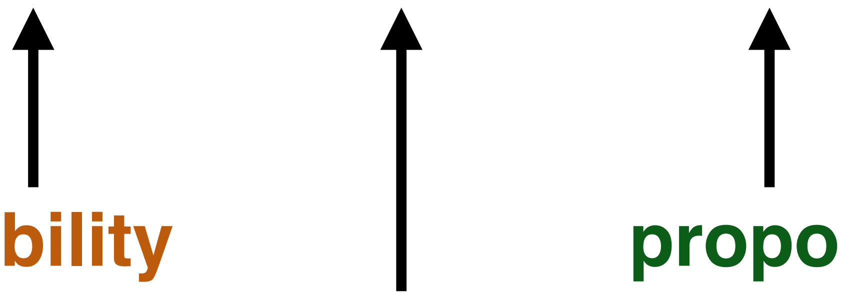




Metropolis-Hastings Algorithm

The **Metropolis-Hastings algorithm** is an algorithm for random walks that will eventually converge to a true distribution of the parameter space.

$$P(\boldsymbol{\theta}_1 \rightarrow \boldsymbol{\theta}_2) \propto \pi(\boldsymbol{\theta}_1)\; q(\boldsymbol{\theta}_1 \rightarrow \boldsymbol{\theta}_2)$$



transitional probability

proposal distribution

prior probability

Metropolis-Hastings Algorithm

The change of state from $\boldsymbol{\theta}_1$ to $\boldsymbol{\theta}_2$ is governed by the **acceptance rate**

$$\alpha(\boldsymbol{\theta}_1 \rightarrow \boldsymbol{\theta}_2) = \min\left\{1, \frac{\pi(\boldsymbol{\theta}_2)\, q(\boldsymbol{\theta}_2 \rightarrow \boldsymbol{\theta}_1)}{\pi(\boldsymbol{\theta}_1)\, q(\boldsymbol{\theta}_1 \rightarrow \boldsymbol{\theta}_2)}\right\}$$

We are assumed an equilibrium state; hence,

$$q(\boldsymbol{\theta}_1 \rightarrow \boldsymbol{\theta}_2) = q(\boldsymbol{\theta}_2 \rightarrow \boldsymbol{\theta}_1)$$

Therefore,

$$\alpha(\boldsymbol{\theta}_1 \rightarrow \boldsymbol{\theta}_2) = \min\left\{1, \frac{\pi(\boldsymbol{\theta}_2)}{\pi(\boldsymbol{\theta}_1)}\right\}$$

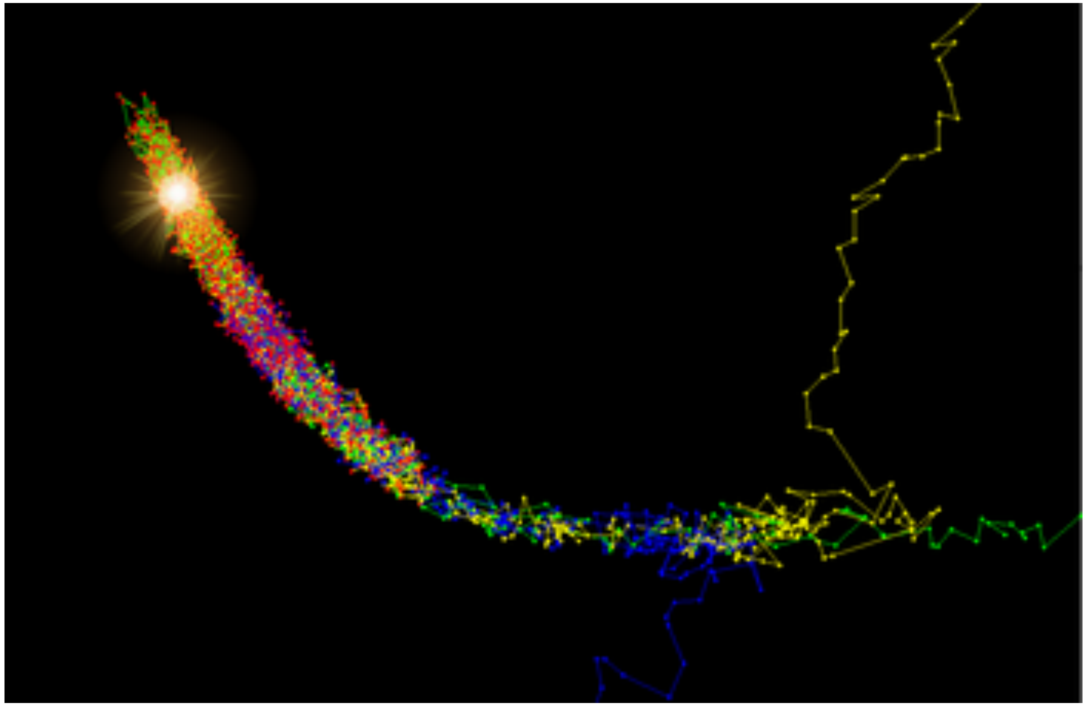
Metropolis-Hastings Algorithm

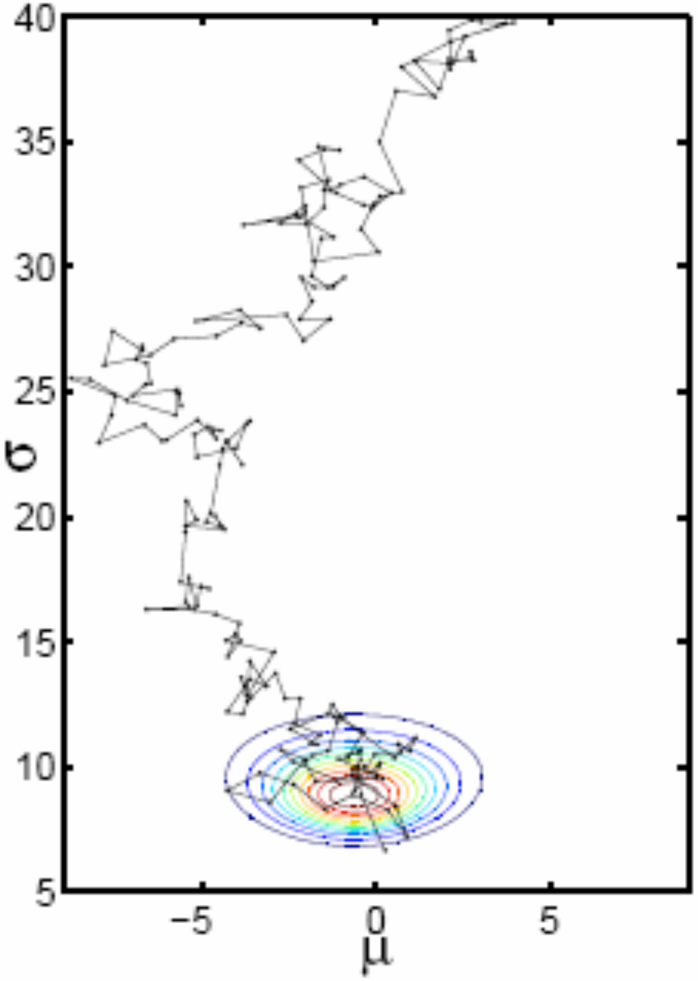
Pseudo code for Metropolis-Hastings Algorithm

```
alpha = likelihood2 / likelihood1;
if alpha > 1:
    jump to the new state;
else:
    if alpha > rand();
        jump to the new state;
    else:
        remain in the same state;
```

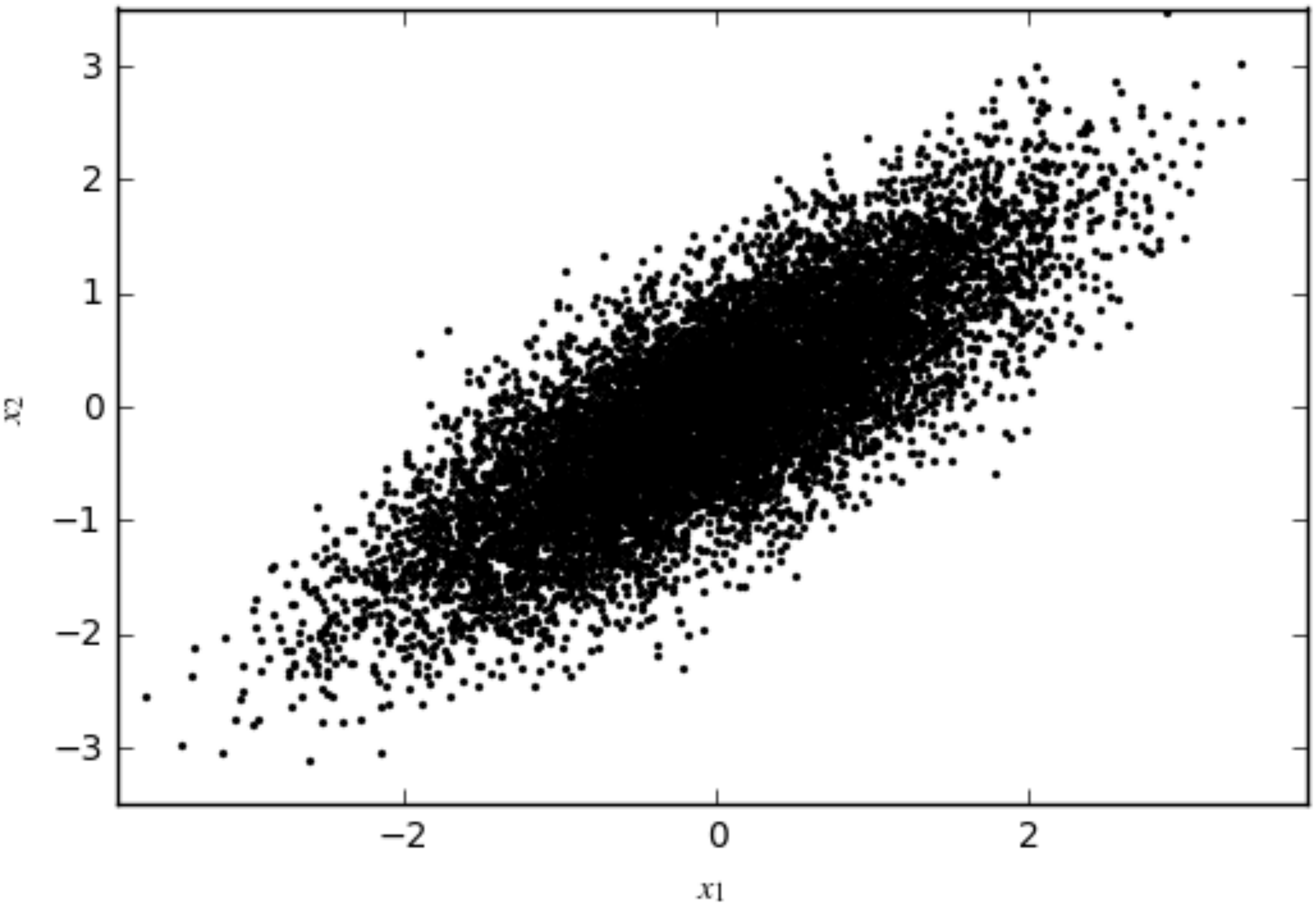
Metropolis-Hastings Algorithm

The chain will take some time to stabilize this is called the **burn-in phase**





Metropolis-Hastings Algorithm



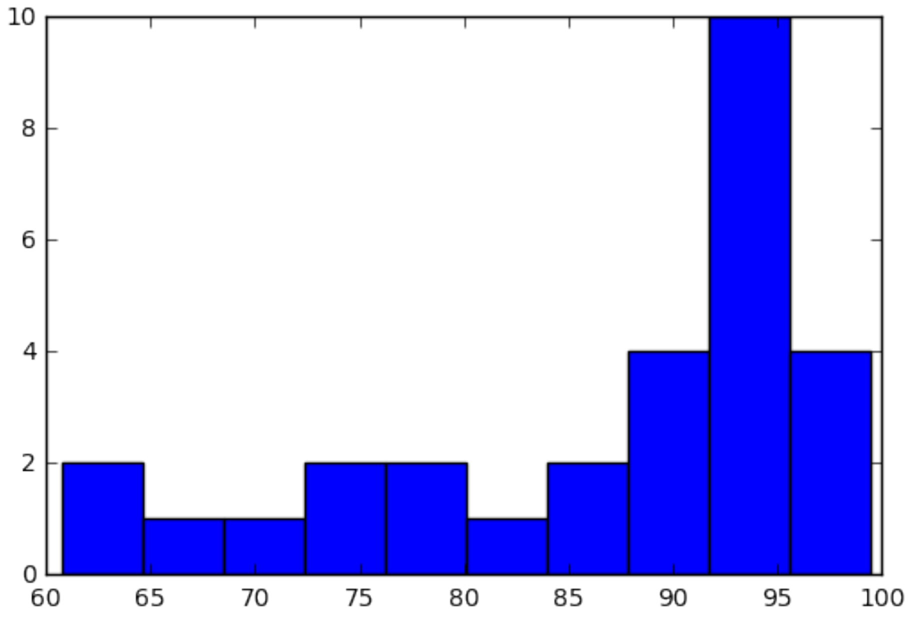
Kernel Density Estimator

- Histogram is a common way to make sense of **discrete data**.

Data

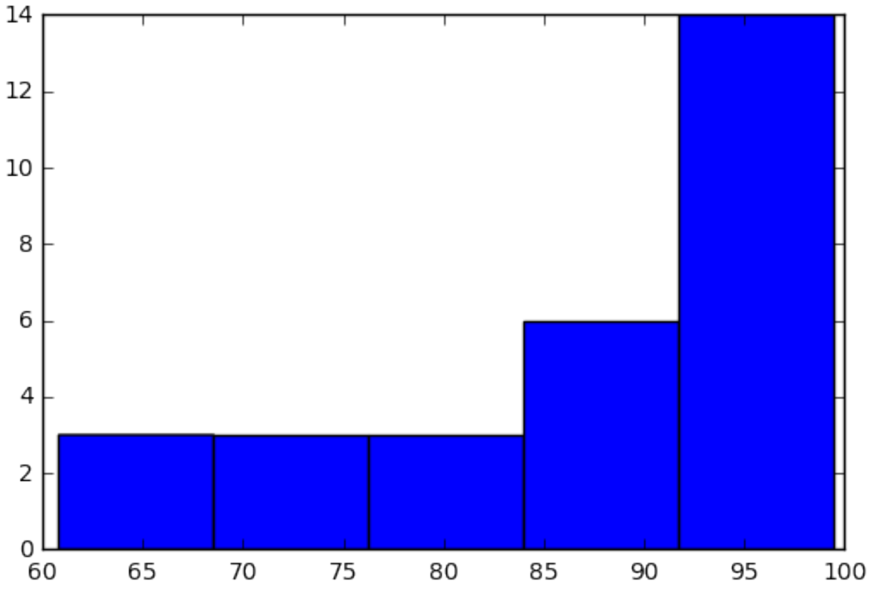
93.5, 93, 60.8, 94.5,
82, 87.5, 91.5, 99.5,
86, 93.5, 92.5, 78,
76, 69, 94.5, 89.5,
92.8, 78, 65.5, 98,
98.5, 92.3, 95.5, 76,
91, 95, 61.4, 96, 90

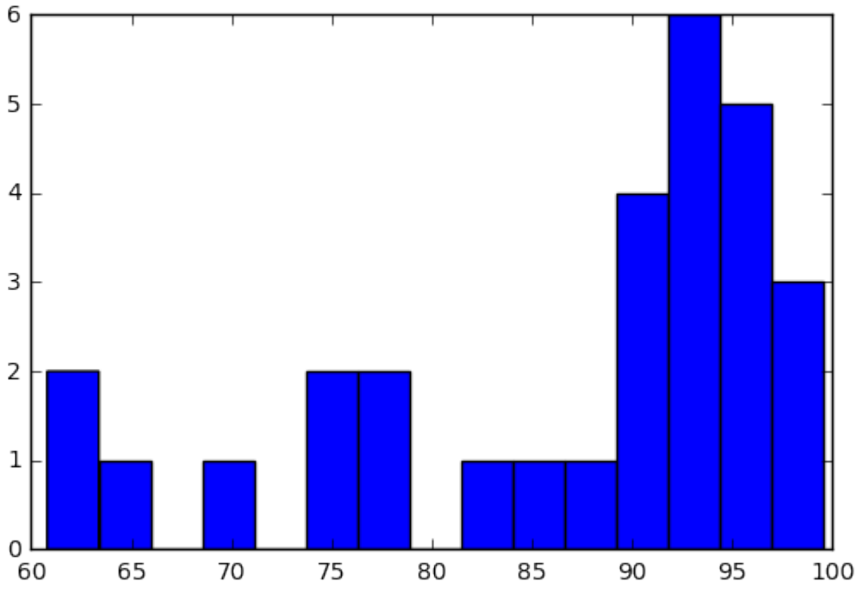
Histogram



Kernel Density Estimator

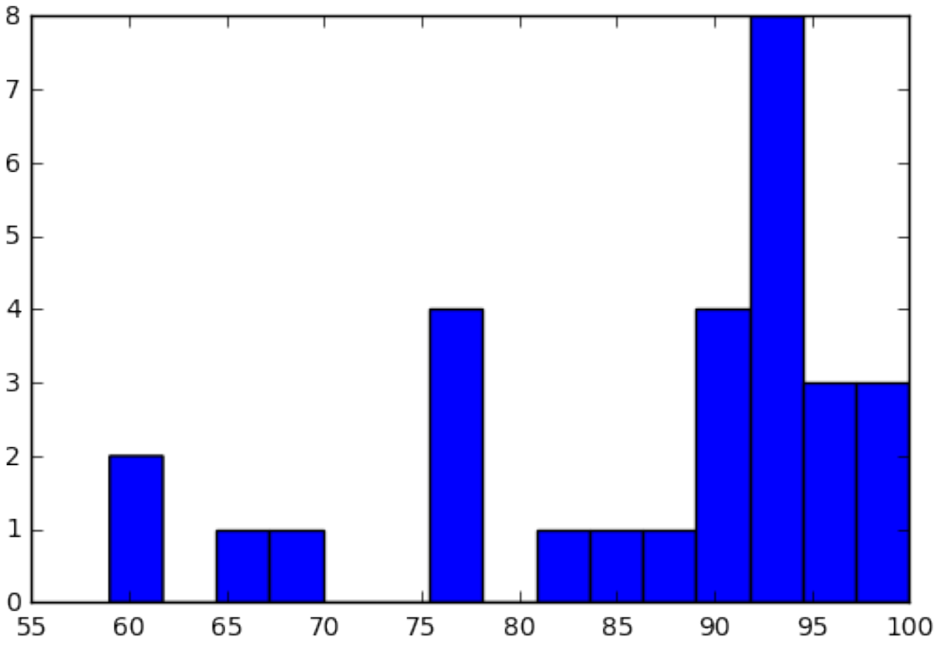
- The same data could generate **different histograms** with **different number of bins**.

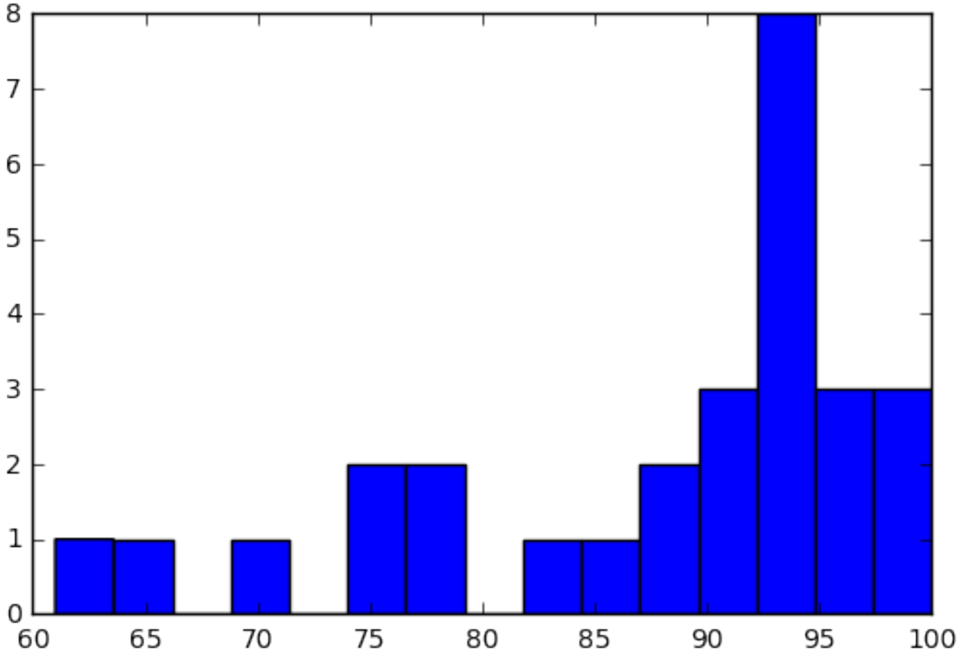




Kernel Density Estimator

- The same data could generate different histograms with different starts of left-edge of bins.





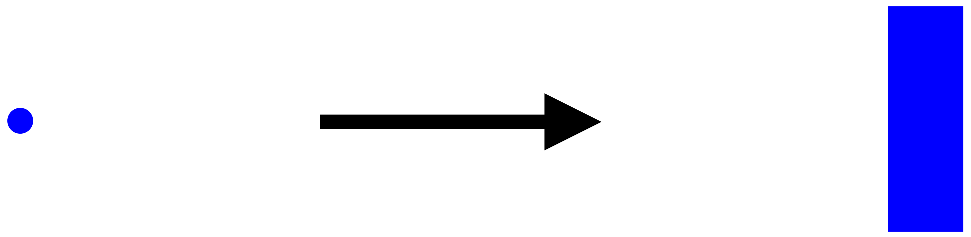
Kernel Density Estimator

Drawbacks of Histogram

- Not smooth
- Depend on width of bins
- Depends on end points of bins

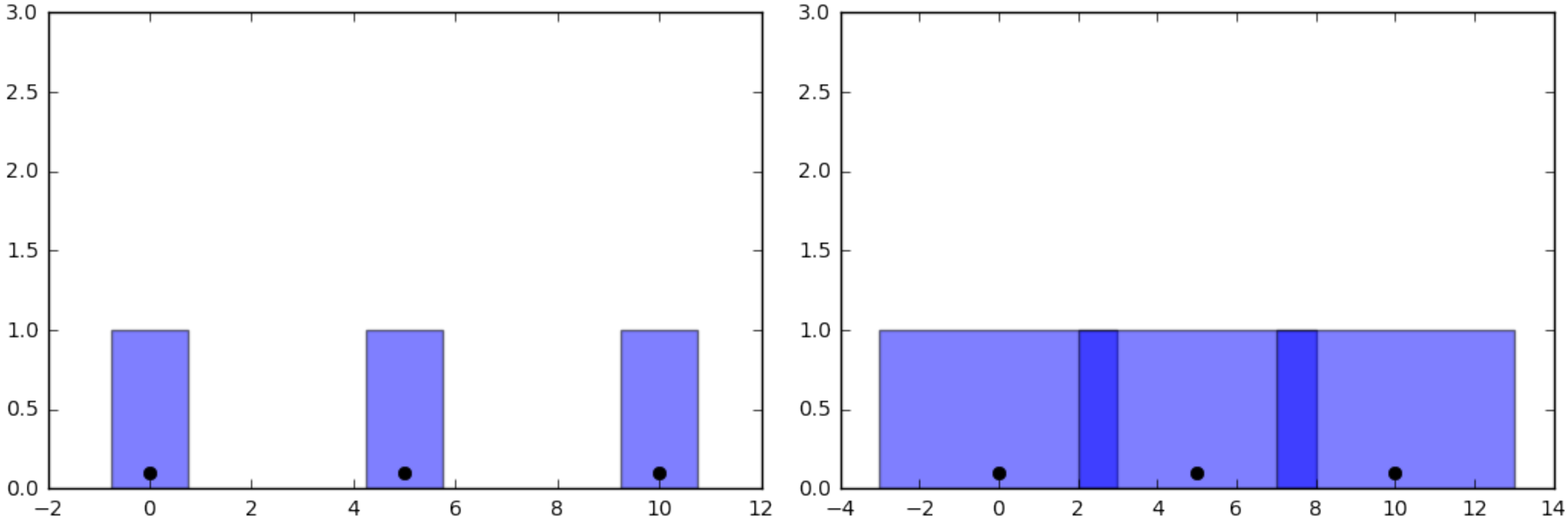
Kernel Density Estimator

- If we instead replace the data point by a kernel function



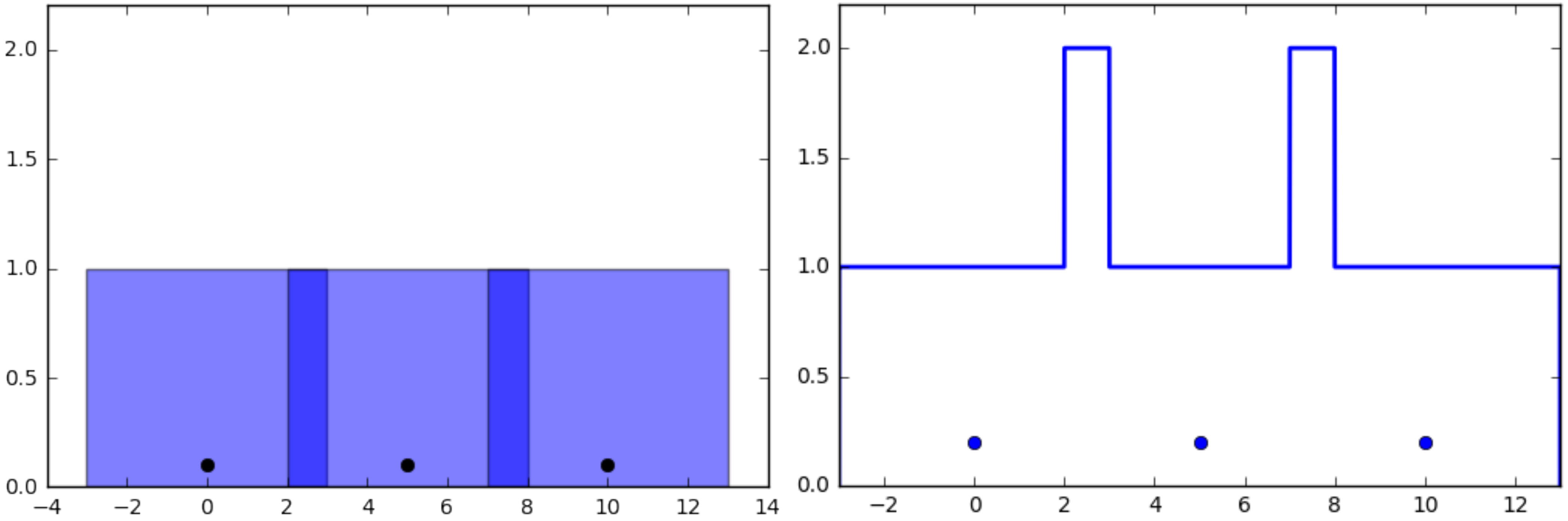
Kernel Density Estimator

- For, simplicity suppose we have only three data points **0, 5, 10**



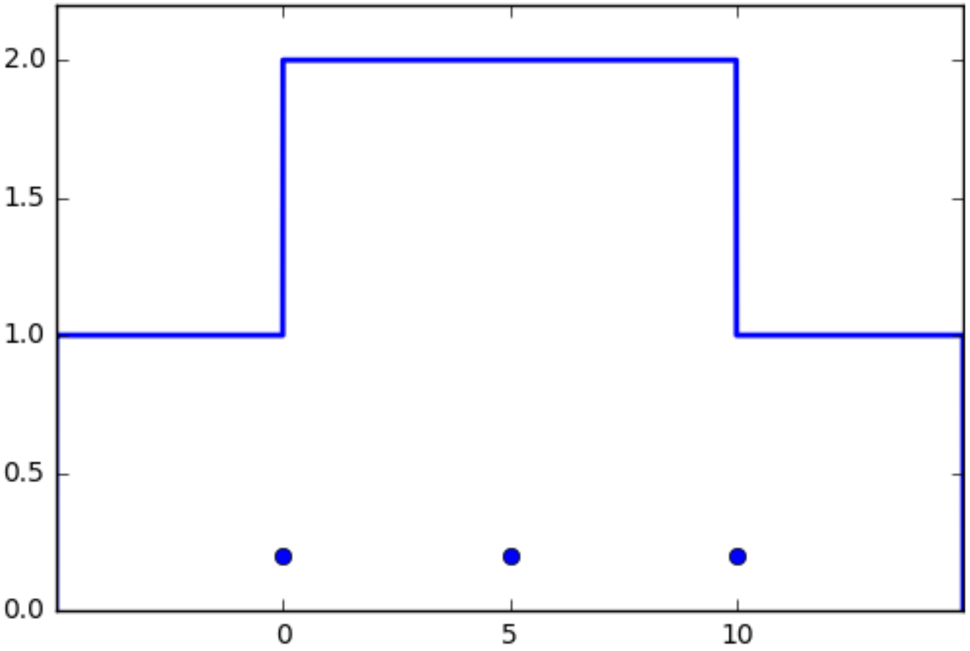
Kernel Density Estimator

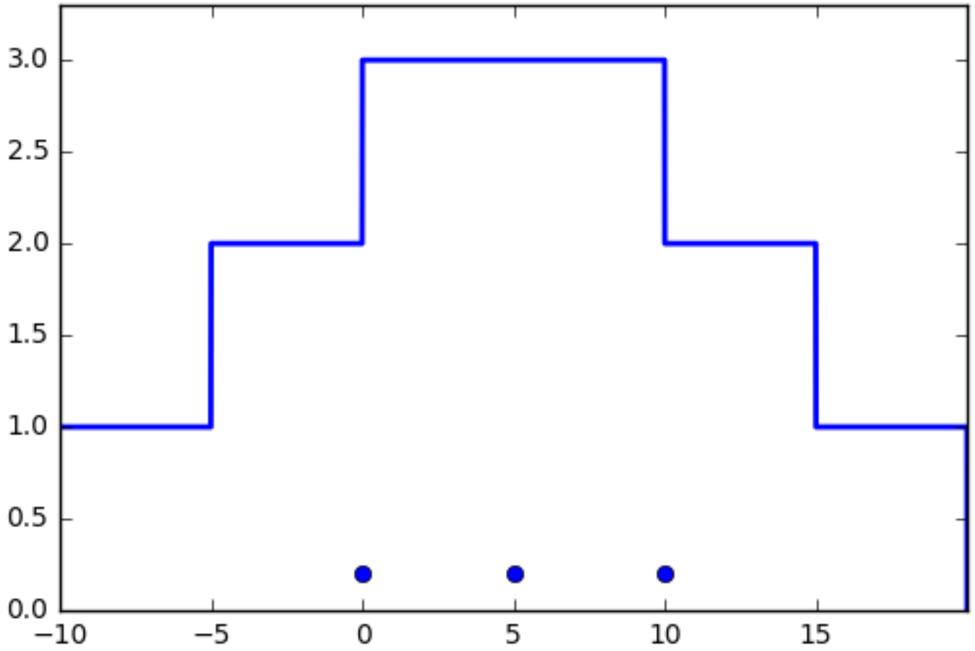
- For, simplicity suppose we have only three data points **0, 5, 10**



Kernel Density Estimator

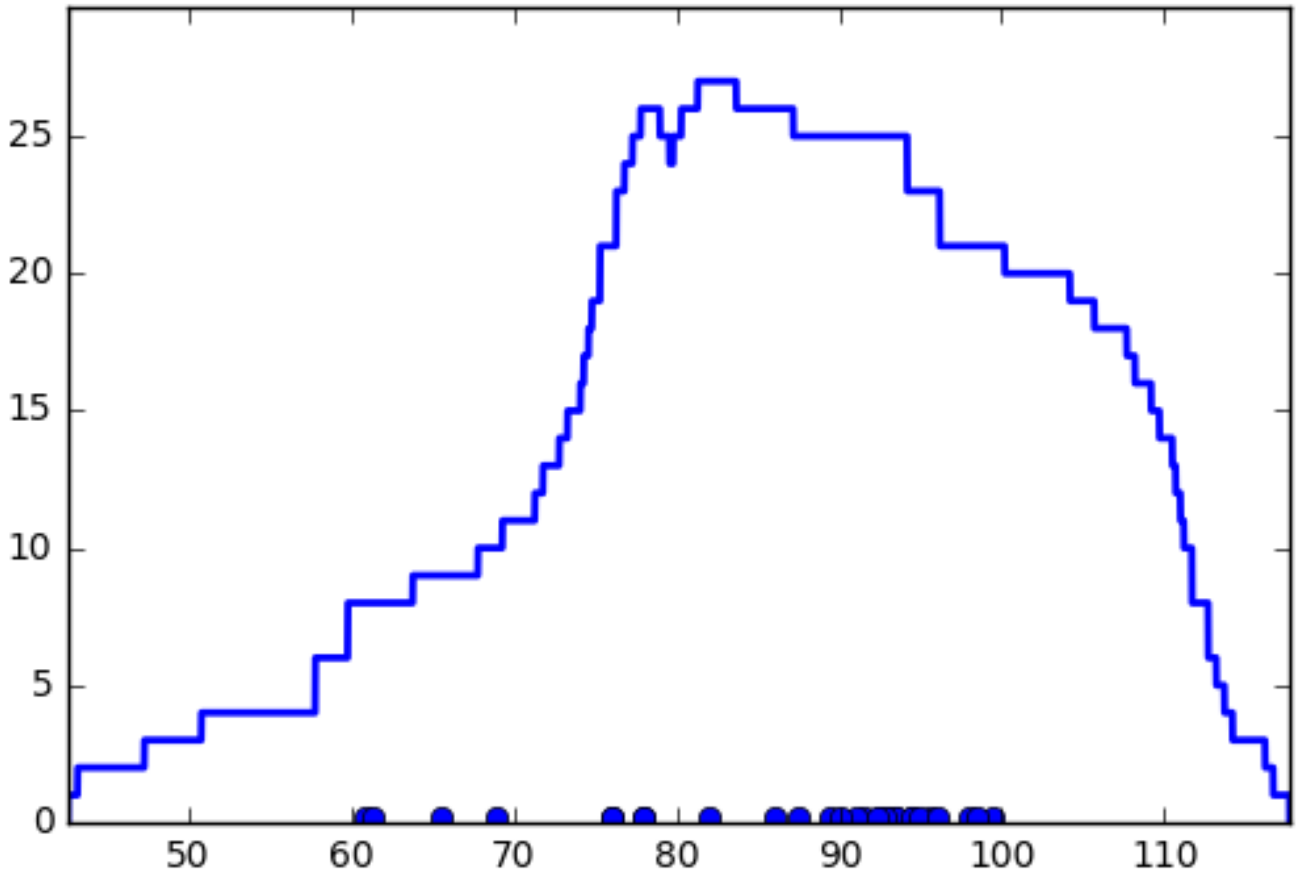
- For, simplicity suppose we have only three data points **0, 5, 10**





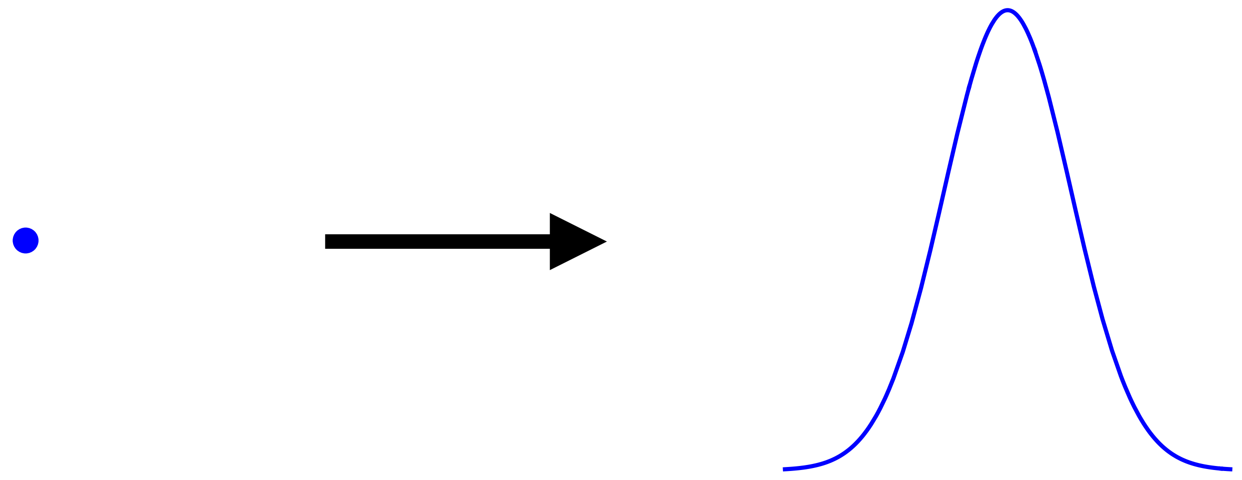
Kernel Density Estimator

- With our previous data



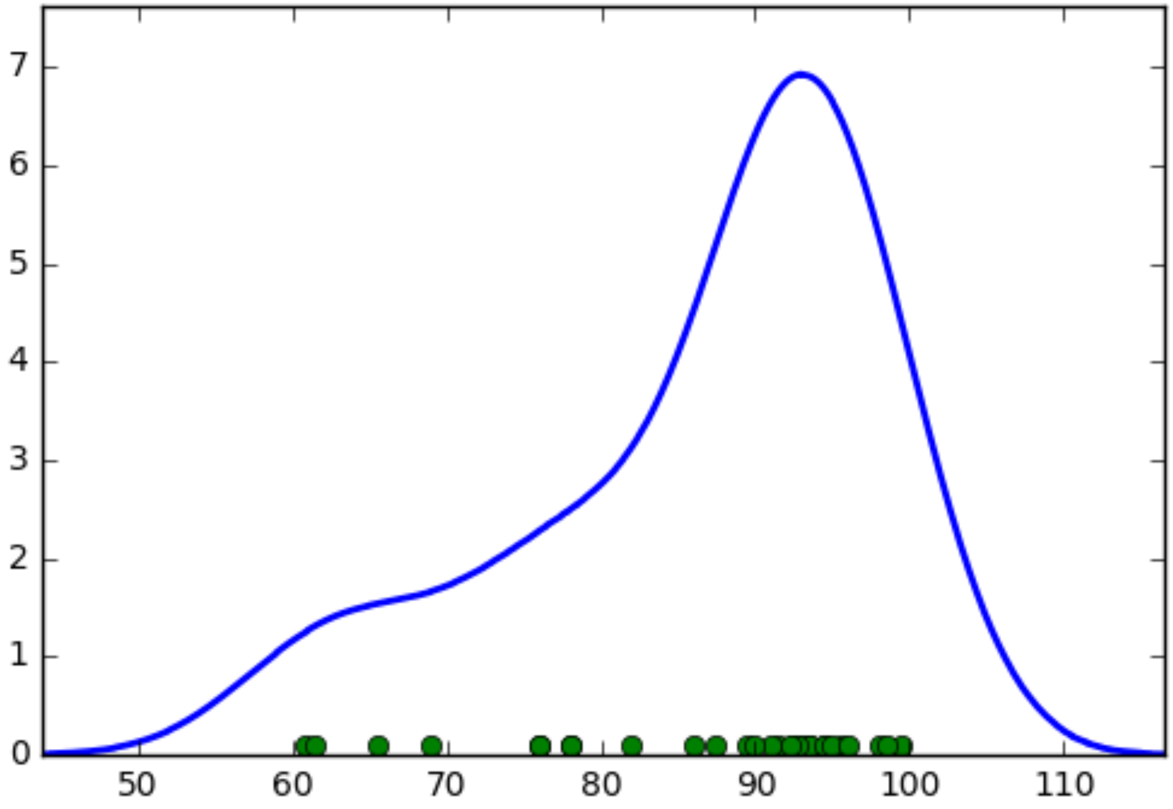
Kernel Density Estimator

- The plot is not smooth because we have a non-smooth kernel function
- We can use a smooth kernel function for example a **Gaussian function**



Kernel Density Estimator

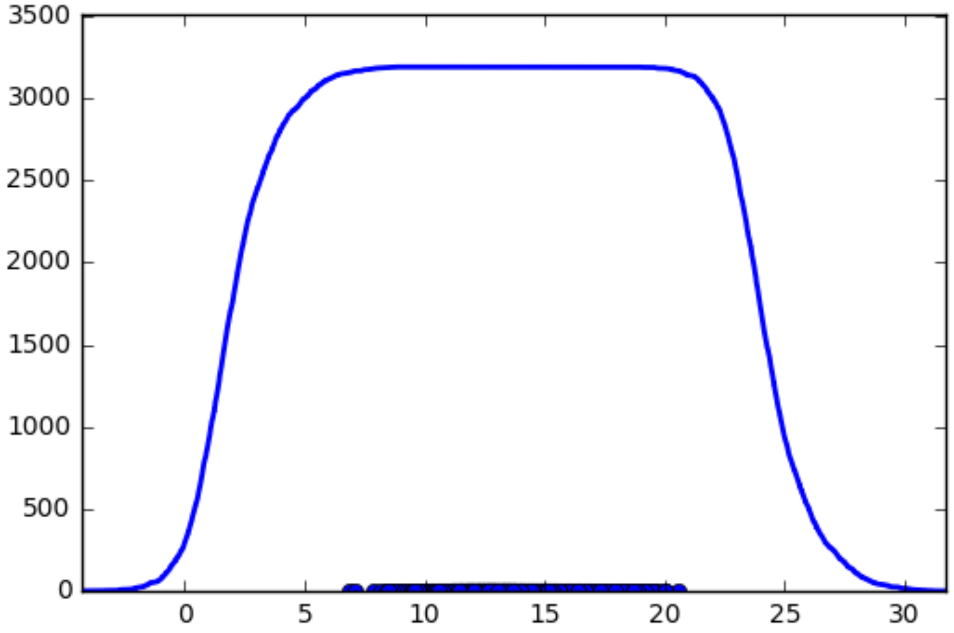
- We have a smooth distribution.



Kernel Density Estimator

- We oversmooth the distribution.

Oversmoothed



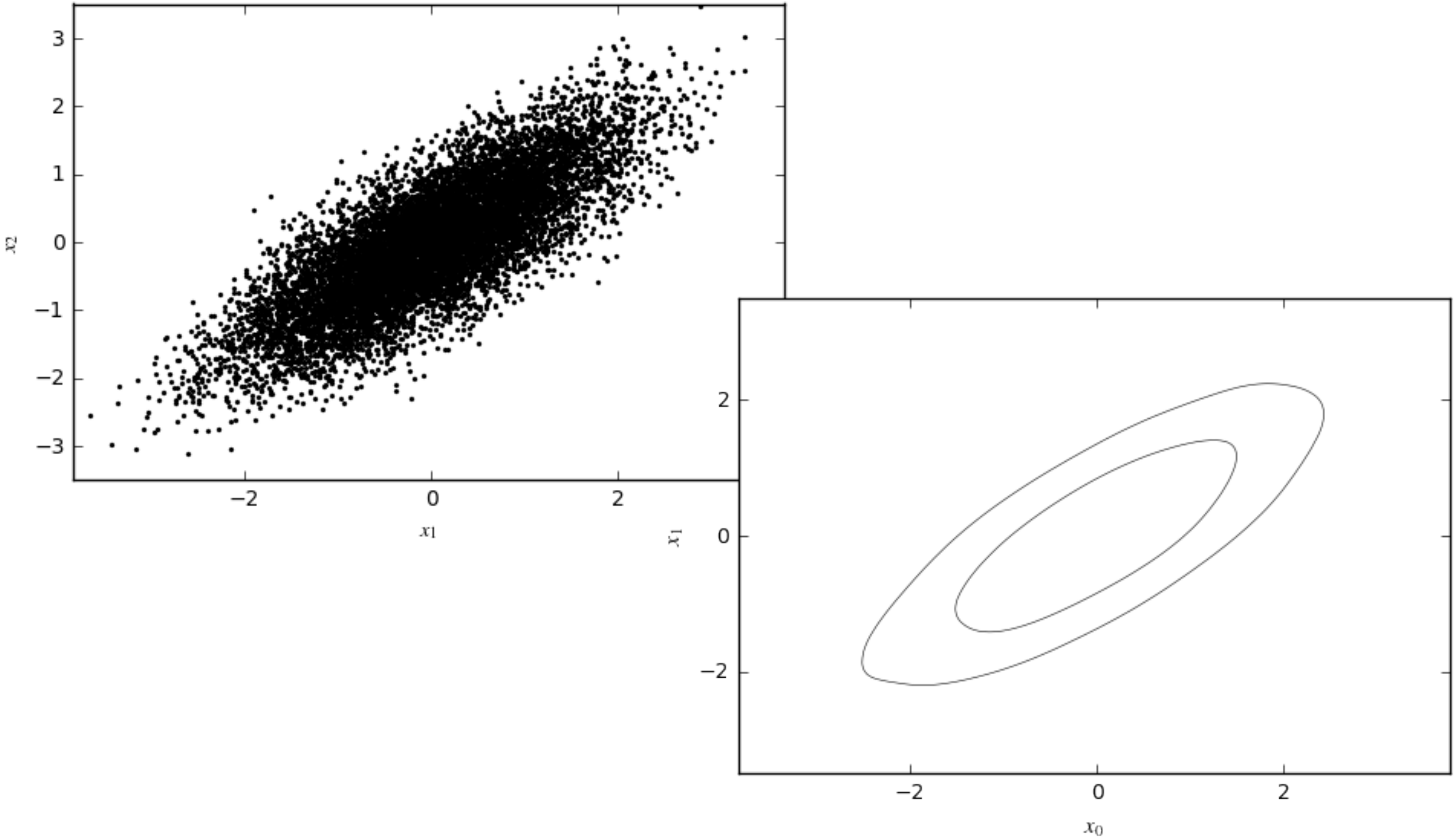
Kernel Density Estimator

- The optimal bandwidth has to be estimated.

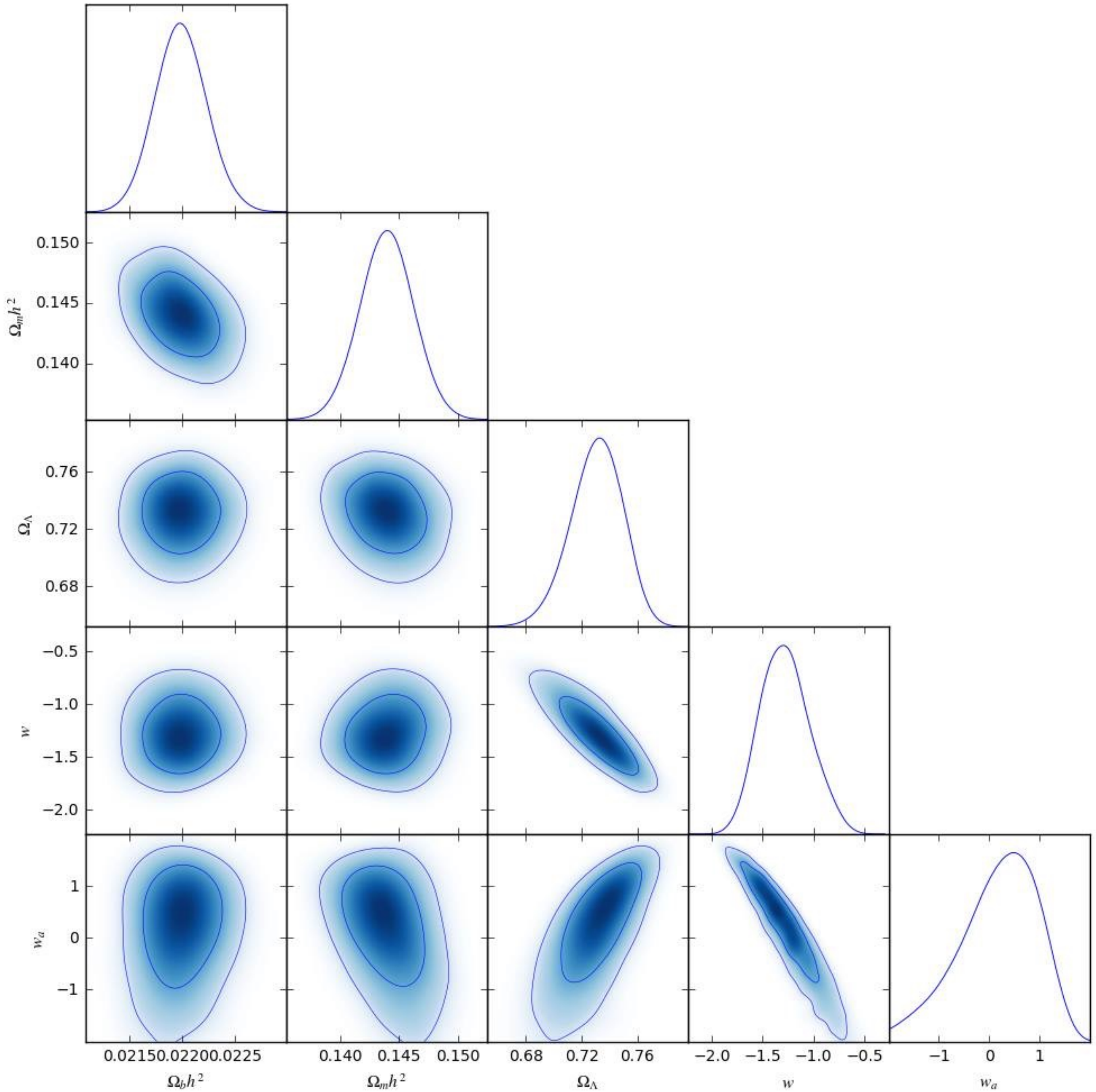
- A normal way to estimated the optimal bandwidth is to minimize the **Asymptotic Mean Integrated Squared Error (AMISE)**

$$\int (f(x) - f_n(x))^2 \mathrm{d}x$$

Kernel Density Estimator



Kernel Density Estimator


$\Omega_m h^2$
Ω_Λ
w
w_a
$\Omega_b h^2$
$\Omega_m h^2$
Ω_Λ
w
w_a