Heretical introduction to Bayesian inference

Luca Martino

Heretical = unorthodox = unconventional = alternative = you cannot find it in books or papers....

Viva Galileo and Giordano Bruno

2024

First Bayesian **in my opinion**: Ronald Fisher, in 1922-1924

- **Likelihood principle** $==>$ $p(\mathbf{y}|\boldsymbol{\theta})$
- **Maximum likelihood estimators**: general procedure for building consistent estimators!
- Need of a model to build the likelihood function.

Likelihood function

- $\mathbf{y}$ ==> observed DATA
- $\boldsymbol{\theta}$ ==> vector or variable to infer
- $p(\mathbf{y}|\boldsymbol{\theta})$ ==> (conditional) DENSITY with respect to $\mathbf{y}$ (normalized or normalizable)
- $p(\mathbf{y}|\boldsymbol{\theta})$ ==> IS NOT A DENSITY with respect to $\boldsymbol{\theta}$ — JUST A FUNCTION with respect to $\boldsymbol{\theta}$

EXAMPLE 1

(*independent data - observations - measurements*)

- **observation model:**

$$y_i = \theta + \epsilon_i, \quad i = 1, ..., N,$$

  $\epsilon_i \sim \mathcal{N}(\epsilon|0, \sigma^2)$.

- a piece of likelihood function is $\ell_i(\theta) = \mathcal{N}(y_i|\theta, \sigma^2)$:

$$\ell_i(\theta) = p(y_i|\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \theta)^2}{2\sigma^2}\right),$$

Fixing the data y_i, and "moving" θ.

Note that $\theta \in \mathbb{R} = (-\infty, +\infty)$.

Example 1

- **Complete likelihood function:**

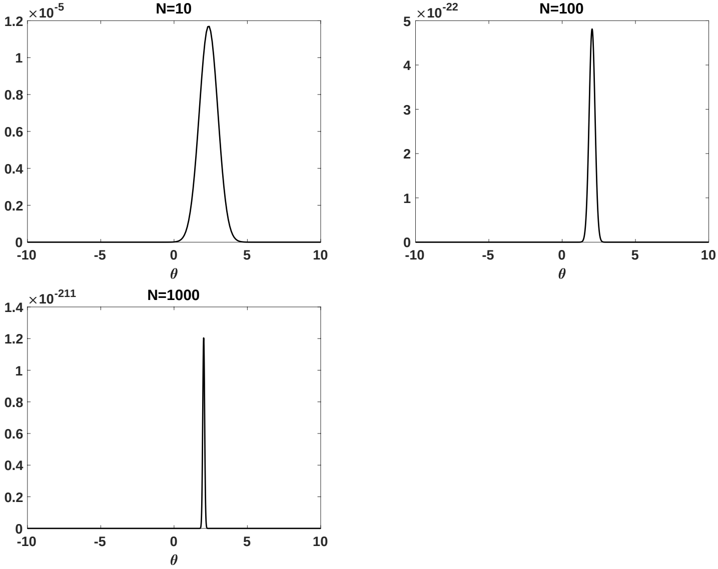
$$\ell(\theta) = p(y_1, \ldots, y_N|\theta) = \prod_{i=1}^{N} p(y_i|\theta), \tag{1}$$

$$= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i-\theta)^2}{2\sigma^2}\right), \tag{2}$$

$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N \exp\left(-\sum_{i=1}^{N}\frac{(y_i-\theta)^2}{2\sigma^2}\right). \tag{3}$$

Fixing the data y_i, and "moving" θ.

Example 1


N=10
×10^-5
1.2
1
0.8
0.6
0.4
0.2
0
-10
-5
0
5
10
θ
N=100
×10^-22
5
4
3
2
1
0
-10
-5
0
5
10
θ
N=1000
×10^-211
1.4
1.2
1
0.8
0.6
0.4
0.2
0
-10
-5
0
5
10
θ

Example 2

- **observation model:** *again independent data*

$$y_i = 1 + \theta\epsilon_i, \quad i = 1, ..., N,$$

  $\epsilon_i \sim \mathcal{N}(\epsilon|0, 1)$.

- a piece of likelihood function is $\ell_i(\theta) = \mathcal{N}(y_i|1, \theta^2)$:

$$\ell_i(\theta) = p(y_i|\theta) = \frac{1}{\sqrt{2\pi\theta^2}} \exp\left(-\frac{(y_i - 1)^2}{2\theta^2}\right),$$

Fixing the data y_i, and "moving" θ.

Note that $\theta \in \mathbb{R}^+ = (0, +\infty)$.

Example 2

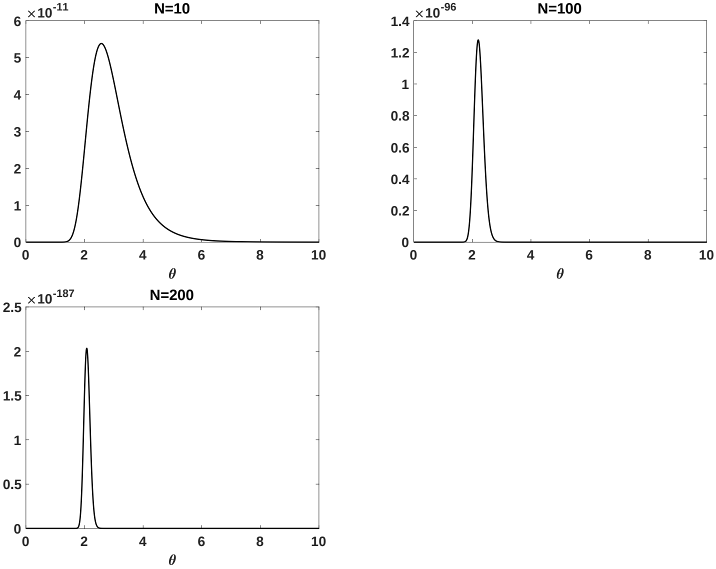
- **Complete likelihood function:**

$$\ell(\theta) = p(y_1, ..., y_N|\theta) = \prod_{i=1}^{N} p(y_i|\theta), \tag{4}$$

$$= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\theta^2}} \exp\left(-\frac{(y_i - 1)^2}{2\theta^2}\right). \tag{5}$$

Fixing the data y_i, and "moving" θ.

Example 2



×10^-11
N=10
0 1 2 3 4 5 6
0 2 4 6 8 10
θ
×10^-96
N=100
0 0.2 0.4 0.6 0.8 1 1.2 1.4
0 2 4 6 8 10
θ
×10^-187
N=200
0 0.5 1 1.5 2 2.5
0 2 4 6 8 10
θ

Example 3

- **Observation model:** flipping/tossing a coin... Heads or Tails....

$$\text{Prob}(y_i = 1) = \theta, \tag{6}$$
$$\text{Prob}(y_i = 0) = 1 - \theta, \quad \theta \in [0, 1]. \tag{7}$$

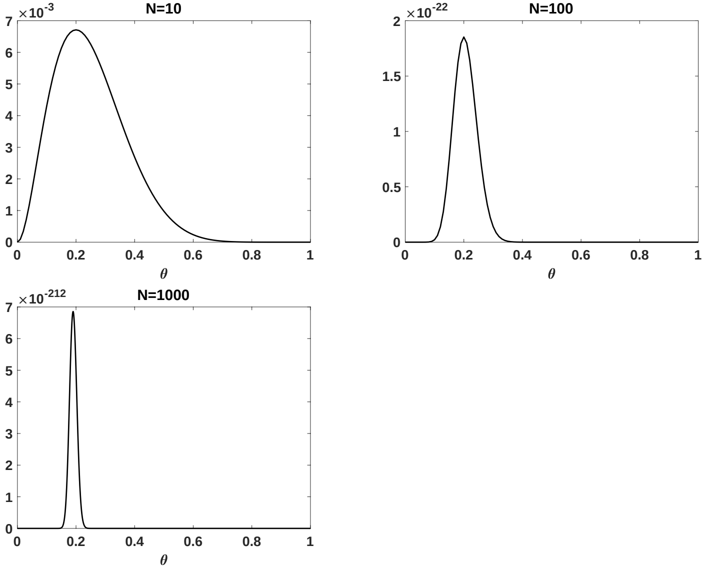
- **Complete likelihood function:**

$$\ell(\theta) = p(y_1, ..., y_N | \theta) = \prod_{i=1}^{N} \theta^{y_i} (1 - \theta)^{1 - y_i}. \tag{8}$$

Fixing the data y_i, and "moving" θ.

Note that $\theta \in \mathbb{R}^+ = [0, 1]$.

Example 3



Just the maximum? or we have more info?

- From the figures: there are likelihood functions "bigger/fatter", narrow, with different symmetries etc.
- Is the only useful information in the point where the maximum is reached?

Likelihood as a density?

- **Can we interpreted the likelihood as a density?**
- only if.... the likelihood function is normalizable...

$$Z_\ell = \int_\Theta p(\mathbf{y}|\boldsymbol{\theta}) d\boldsymbol{\theta} < \infty$$

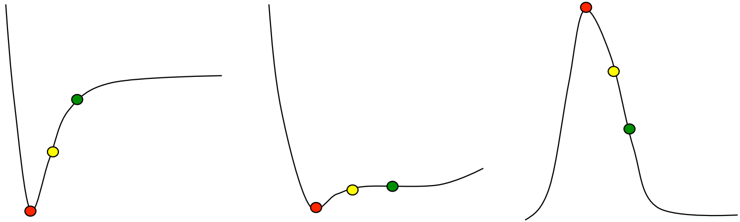
- in this case, we can define **a density** as

$$\bar{\ell}(\boldsymbol{\theta}; \mathbf{y}) = \frac{1}{Z_\ell} p(\mathbf{y}|\boldsymbol{\theta}).$$

- It looks that we are inverted the positions of $\mathbf{y}$ and $\boldsymbol{\theta}$...

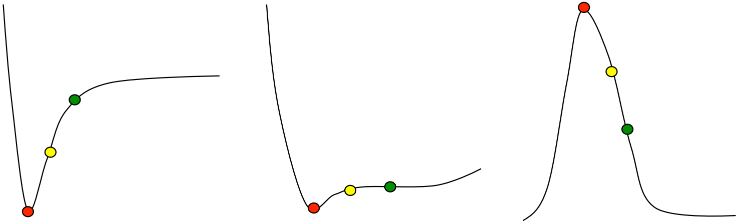
Extract more information from the likelihood function

- **Only the maximum of the likelihood is relevant?**
- we can define “more estimators” and variance etc.
- Cost function versus probabilistic approach...



Extract more information from the likelihood function

(a) different possible "estimators", **(b)** compute areas (e.g., variance, for confidence intervals, quantiles etc.) ...



- **from optimization ==> to sampling**

Likelihood as a density?

- What happens if the likelihood function is NOT normalizable? namely, $\int_\Theta p(\mathbf{y}|\boldsymbol{\theta})d\boldsymbol{\theta} = \infty$....
- and/or if we have more additional information, belief about $\boldsymbol{\theta}$?
- **We can use a prior density $g(\theta)$ and the Bayesian rule:**

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{1}{p(\mathbf{y})} p(\mathbf{y}|\boldsymbol{\theta}) g(\boldsymbol{\theta}).$$

Main actors in Bayesian inference

- The **posterior probability density function (pdf)** is

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\theta})g(\boldsymbol{\theta})}{p(\mathbf{y})} \propto p(\mathbf{y}|\boldsymbol{\theta})g(\boldsymbol{\theta}), \tag{9}$$

  where
  - $p(\mathbf{y}|\boldsymbol{\theta})$ is the likelihood function (induced by the observation model);
  - $g(\boldsymbol{\theta})$ is the prior pdf,
  - $Z = p(\mathbf{y})$: marginal likelihood/Bayesian evidence - (useful for model selection and hypothesis testing)

Main benefit of Bayesian inference

- **All the problems in statistics can be solved extracting information from the posterior density $p(\boldsymbol{\theta}|\mathbf{y})$ and computing the marginal likelihood:**

$$E[\boldsymbol{\Theta}^{\alpha}] = \int_{\boldsymbol{\Theta}} \boldsymbol{\theta}^{\alpha} p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta} \tag{10}$$

$$Z = p(\mathbf{y}) = \int_{\boldsymbol{\Theta}} p(\mathbf{y}|\boldsymbol{\theta}) g(\boldsymbol{\theta}) d\boldsymbol{\theta}. \tag{11}$$

- **All the problems in statistics becomes quadrature problems (computing integrals).**
- We a general procedure/recipe to do any statistical analysis.

Problems of Bayesian inference

- We a general procedure/recipe to do any statistical analysis.
- But to do that:
  - **We have to compute/approximate complicated integrals.**
  - **Dependence on the prior densities.**