Maximum Likelihood Estimation

Given "data" $X_1, X_2, \ldots, X_n$, a random sample (iid) from a distribution with unknown parameter θ, we want to find the value of θ in the parameter space that maximizes our "probability" of observing that data.

- If $X_1, X_2, \ldots, X_n$ are discrete, we can look at

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$$

as a function of θ, and find the θ that maximizes it.

- If $X_1, X_2, \ldots, X_n$ are discrete, we can look at

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$$

as a function of θ, and find the θ that maximizes it.

This is the joint pmf for $X_1, X_2, \ldots, X_n$.

- The analogue for continuous $X_1, X_2, \ldots, X_n$ is to maximize the joint pdf with respect to θ.

The pmf/pdf for any one of $X_1, X_2, \ldots, X_n$ is denoted by $\underbrace{f(x).}$

No 1,2, ..., n here.

We will emphasize the dependence of f on a parameter θ by writing it as

$$f(x;\theta)$$

The joint pmf/pdf for all n of them is

$$\underbrace{f(x_1, x_2, \ldots, x_n; \theta)}_{f(\vec{x};\theta)} = \prod_{i=1}^{n} f(x_i; \theta)$$

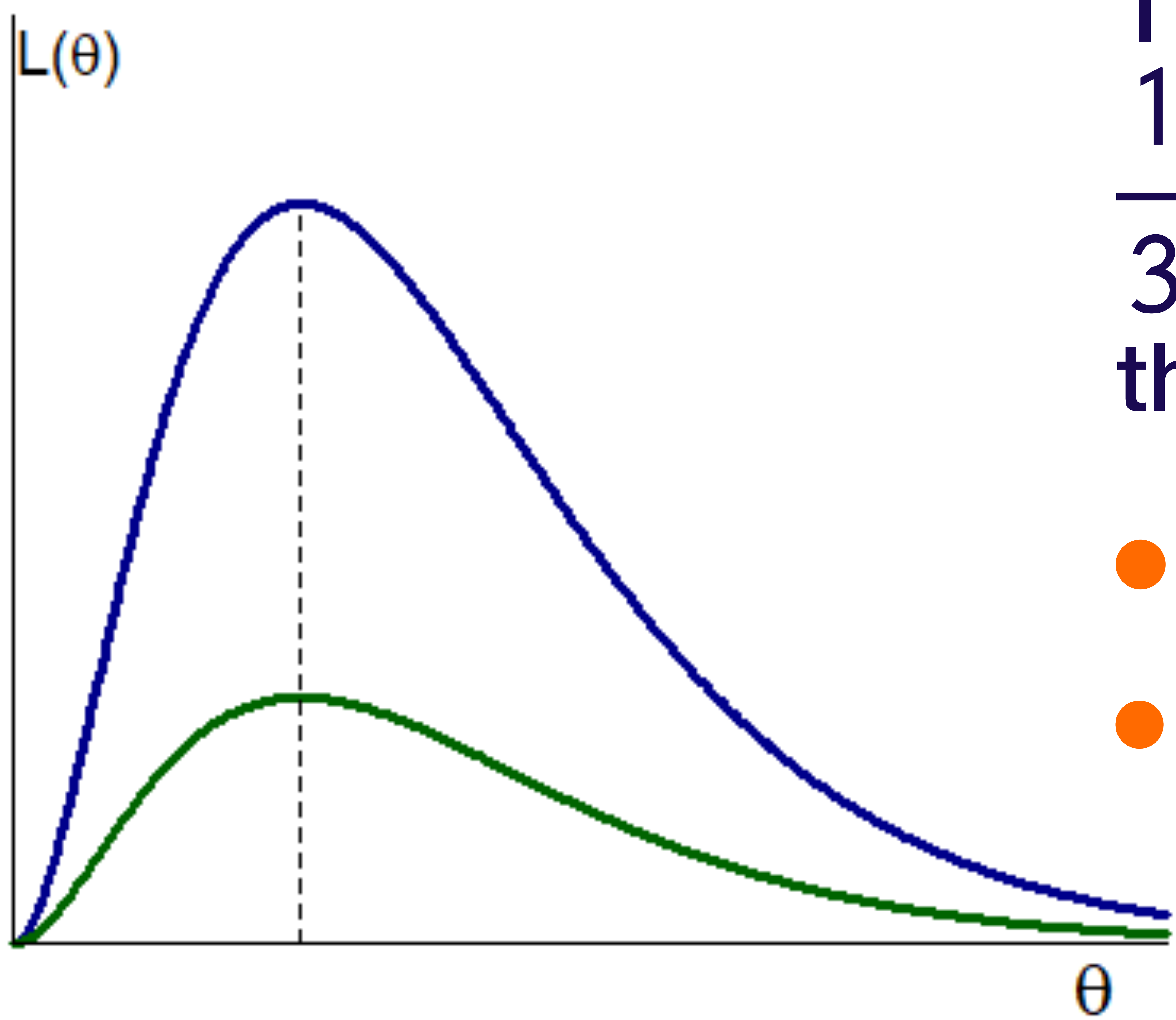
The joint pmf/pdf for all n of them is

$$f(\vec{x};\theta) = \prod_{i=1}^{n} f(x_i;\theta)$$

- The data (the x's) are fixed.
- Think of the x's as fixed and the joint pdf as a function of θ.

Call it a likelihood function and denote it by L(θ).

The likelihood function $L(\theta)$:



The maximum of $\frac{1}{3}L(\theta)$ occurs in the same place.

- Same for $3L(\theta)$.
- Same for $x_1 L(\theta)$.

The likelihood $L(\theta)$ is defined to be anything proportional to the joint pmf/pdf.

Example:

$$X_1, X_2, \ldots, X_n \overset{iid}{\sim} \text{Bernoulli}(p)$$

The pmf for one of them is

$$f(x;p) = p^x(1-p)^{1-x} I_{\{0,1\}}(x)$$

The joint pmf for all of them is

$$f(\vec{x};p) = \prod_{i=1}^{n} f(x_i;p)$$

$$= \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} I_{\{0,1\}}(x_i)$$

The parameter space is [0,1].

$$f(\vec{x};p)$$

$$= p^{\sum_{i=1}^{n} x_i}(1-p)^{n-\sum_{i=1}^{n} x_i} \prod_{i=1}^{n} I_{\{0,1\}}(x_i)$$

A likelihood is

$$L(p) = p^{\sum_{i=1}^{n} x_i}(1-p)^{n-\sum_{i=1}^{n} x_i}$$

It is almost always easier to minimize the "log-likelihood":

$$\ell(p) = \ln L(p) = \sum_{i=1}^{n} x_i \ln p + \left(n - \sum_{i=1}^{n} x_i\right)\ln(1-p)$$

$$\ell(p) = \sum_{i=1}^{n} x_i \ln p + (n - \sum_{i=1}^{n} x_i) \ln(1-p)$$

$$\frac{\partial}{\partial p}\ell(p) = \frac{\sum_{i=1}^{n} x_i}{p} - \frac{n - \sum_{i=1}^{n} x_i}{1-p} \overset{\text{set}}{=} 0$$

$$p(1-p)\left[\frac{\sum_{i=1}^{n} x_i}{p} - \frac{n - \sum_{i=1}^{n} x_i}{1-p}\right] = p(1-p) \cdot 0$$

$$(1-p)\sum_{i=1}^{n} x_i - p\left(n - \sum_{i=1}^{n} x_i\right) = 0$$

$$\sum_{i=1}^{n} x_i - p \cancel{\sum_{i=1}^{n} x_i} - np + p \cancel{\sum_{i=1}^{n} x_i} = 0$$

$$\Rightarrow \quad p = \frac{\sum_{i=1}^{n} x_i}{n}$$

The maximum likelihood estimator for p is:

$$\boxed{\hat{p} = \frac{\sum_{i=1}^{n} X_i}{n} = \overline{X}}$$

Uppercase!

Continuous Example:

$$X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \exp(\text{rate} = \lambda)$$

The pdf for one of them is

$$f(x;\lambda) = \lambda e^{-\lambda x} I_{(0,\infty)}(x)$$

The joint pdf for all of them is

$$f(\vec{x};\lambda) = \prod_{i=1}^{n} f(x_i;\lambda)$$

The parameter space is $(0, \infty)$.

$$= \prod_{i=1}^{n} \lambda e^{-\lambda x_i} I_{(0,\infty)}(x_i)$$

$$f(\vec{x}; p) = \lambda^n e^{-\lambda \sum_{i=1}^n x_i} \prod_{i=1}^n I_{(0,\infty)}(x_i)$$

A likelihood is

$$L(\lambda) = \lambda^n e^{-\lambda \sum_{i=1}^n x_i}$$

The log-likelihood is

$$\ell(\lambda) = n \ln \lambda - \lambda \sum_{i=1}^n x_i$$

$$\ell(\lambda) = n \ln \lambda - \lambda \sum_{i=1}^{n} x_i$$

$$\frac{\partial}{\partial \lambda} \ell(\lambda) = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i \overset{\text{set}}{=} 0$$

$$\Rightarrow \quad \lambda = \frac{n}{\sum_{i=1}^{n} x_i}$$

Same as method of moments. Biased! 😳

The maximum likelihood estimator for p is:

$$\boxed{\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} X_i} = \frac{1}{\overline{X}}}$$

Uppercase!