Introduction
OOOOOO

Examples
OOOO

Properties of MLEs
OOOOOOOO

**Imperial College London**

# Lecture 06: Maximum Likelihood Estimation
## Statistical Modelling I

Dr.Riccardo Passeggeri

Introduction
OOOOOO

Examples
OOOO

Properties of MLEs
OOOOOOOO

# Outline

1. Introduction

2. Examples

3. Properties of MLEs

Introduction
●00000

Examples
0000

Properties of MLEs
00000000

# Introduction

Introduction
○●○○○○○

Examples
○○○○

Properties of MLEs
○○○○○○○○

# Motivation

▶ One of the most common methods of defining a new estimator is maximum likelihood

▶ The approach is widely applicable and the intuition behind it is useful even in more general semiparametric or nonparametric models

▶ You have seen MLE before – the intuition is to find $\theta$ that maximises the probability of the observed data

Introduction
○○●○○○

Examples
○○○○

Properties of MLEs
○○○○○○○○

# Example: Binomial distribution (discrete $\Theta$)

$X \sim \text{Binomial}(10, \theta)$, $\theta \in \Theta = \{0, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1\}$.
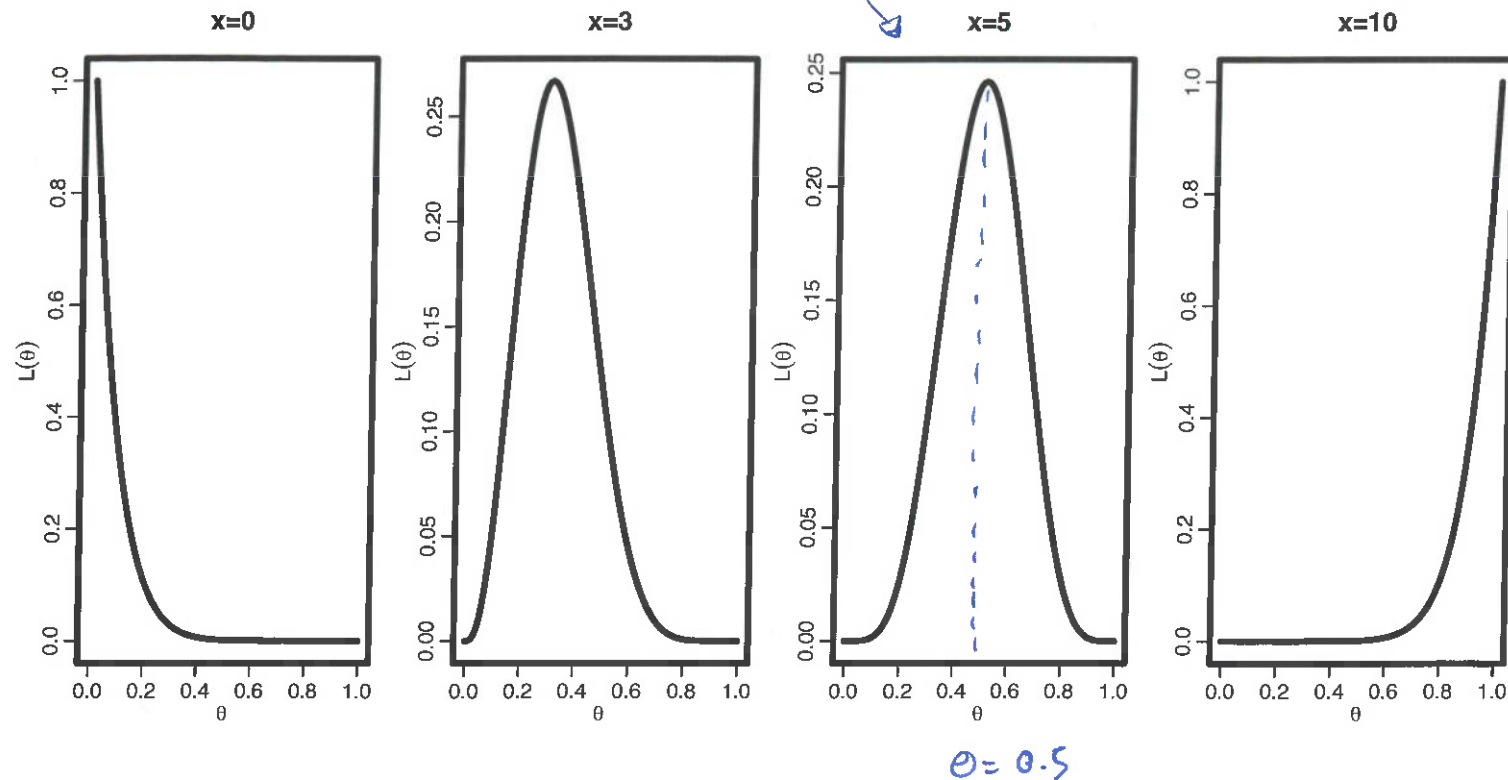
The probability of a given outcome is

$$P_\theta(X = x) = \binom{10}{x} \theta^x (1 - \theta)^{10-x}, \quad x = 0, \ldots, 10.$$

|  | x=0 | x=1 | x=2 | x=3 | x=4 | x=5 | x=6 | x=7 | x=8 | x=9 | x=10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $\theta = 0$ | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| $\theta = 0.25$ | 0.06 | 0.19 | 0.28 | 0.25 | 0.15 | 0.06 | 0.02 | 0.00 | 0.00 | 0.00 | 0.00 |
| $\theta = 0.5$ | 0.00 | 0.01 | 0.04 | 0.12 | 0.21 | 0.25 | 0.21 | 0.12 | 0.04 | 0.01 | 0.00 |
| $\theta = 0.75$ | 0.00 | 0.00 | 0.00 | 0.00 | 0.02 | 0.06 | 0.15 | 0.25 | 0.28 | 0.19 | 0.06 |
| $\theta = 1$ | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 |

Introduction
○○○●○○

Examples
○○○○

Properties of MLEs
○○○○○○○○○

# Example: Binomial ($\Theta = [0, 1]$)

The likelihood function is

$$L(\theta) = P_\theta(X = 5) = \binom{10}{5} \theta^5 (1 - \theta)^5$$



x=0      x=3      x=5      x=10

$\theta = 0.5$

Introduction
○○○○●○

Examples
○○○○

Properties of MLEs
○○○○○○○○○

# Likelihood function

Suppose we observe the random object $Y$ with realisation $y$. The likelihood function is

$$L(\theta) = L(\theta; y) = \begin{cases} P(Y = y; \theta), & \text{discrete data} \\ f_Y(y; \theta), & \text{absolutely continuous data.} \end{cases}$$

**Random sample**

If $Y_i$ has pdf $f(\cdot; \theta)$ then

$$L(\theta) = \prod_{i=1}^{n} f(y_i; \theta).$$

Introduction
○○○○○○●

Examples
○○○○

Properties of MLEs
○○○○○○○○○

# Maximum likelihood estimator

A **maximum likelihood estimator** (MLE) of $\theta$ is an estimator $\hat{\theta}$ s.t.

$$L(\hat{\theta}) = \sup_{\theta \in \Theta} L(\theta).$$

Usually, the MLE is well defined. However, one can construct situations in which it does not exist or is not unique.

Introduction
○○○○○○

Examples
●○○○

Properties of MLEs
○○○○○○○○

# Examples

Introduction
oooooo

Examples
o●oo

Properties of MLEs
oooooooo

# Example: $X_1, \ldots, X_n$ iid Poisson$(\theta)$ , $\theta \in (0, \infty)$

Find the MLE of $\theta$

$$P_\theta(X_i = x_i) = \frac{e^{-\theta} \theta^{x_i}}{x_i!} \qquad L(\theta) = \prod_{i=1}^{n} \frac{e^{-\theta} \theta^{x_i}}{x_i!}$$

$$\log L(\theta) = -n\theta + \sum_{i=1}^{n} x_i \log \theta + \log \left( \prod_{i=1}^{n} \frac{1}{x_i!} \right)$$

$$\frac{\partial}{\partial \theta} \log L(\theta) = -n + \frac{1}{\theta} \sum_{i=1}^{n} x_i = 0 \implies \boxed{\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} x_i}$$

$$\frac{\partial^2}{\partial \theta^2} \log L(\theta) = -\frac{1}{\theta^2} \sum_{i=1}^{n} x_i < 0$$

Introduction
oooooo

Examples
oo●o

Properties of MLEs
ooooooooo

# Example: $Y_1, \ldots, Y_n$ iid $N(\mu, \sigma^2)$ with $\mu$, $\sigma^2$ unknown

Then the MLEs are $\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} y_i$ and $\hat{\sigma}^2 = \frac{1}{n} \sum (y_i - \bar{y})^2$    (Check).

$\hookrightarrow$ SAMPLE VARIANCE

$\hat{\mu}$ is unbiased but $\hat{\sigma}^2$ is not:

$$E(\hat{\sigma}^2) = \frac{1}{n} E\left(\sum (Y_i - \bar{Y})^2\right) = \frac{n-1}{n} \sigma^2 \neq \sigma^2$$

Introduction
○○○○○○

Examples
○○○●

Properties of MLEs
○○○○○○○○

# Summary

It is straightforward to derive an MLE, but we need to understand why we would prefer to use the MLE over other estimators (especially if they may be biased). We will see in the next section that the MLE is asymptotically well behaved.

Introduction
○○○○○○

Examples
○○○○

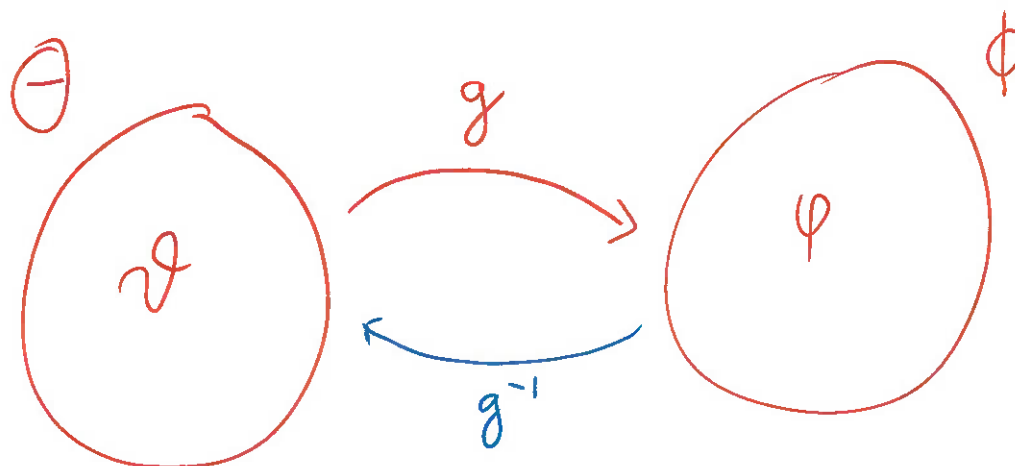Properties of MLEs
●○○○○○○○○

# Properties of MLEs

Introduction
OOOOOO

Examples
OOOO

Properties of MLEs
O●OOOOOOO

# Properties

MLEs are

- ▶ Functionally invariant
- ▶ Consistent
- ▶ Asymptotically normal

Introduction
oooooo

Examples
oooo

Properties of MLEs
ooo●oooooo

# MLEs are functionally invariant

If $g$ is a bijective function and if $\hat{\theta}$ is an MLE of $\theta$, then $\hat{\phi} = g(\hat{\theta})$ is an MLE of $\phi = g(\theta)$.



① $X \sim N(\theta, 1)$, $\theta \in [0, 1]$

② $X \sim N(\phi - 3, 1)$, $\phi \in [3, 4]$

$\phi = g(\theta) = \theta + 3$

$\widehat{\Theta} = [0, 1]$ $\quad \widehat{\Phi} = [3, 4]$

$P_\theta(X \in B) \overset{\text{NOTATION}}{=} \widetilde{P}_{g(\theta)}(X \in B)$

$\left( X \sim N(\theta, 1) \Longleftrightarrow X \sim N(g(\theta) - 3, 1) = N(\theta, 1) \right)$

$\widetilde{P}_\phi(X \in B) = P_{g^{-1}(\phi)}(X \in B)$

where $P$ refers to ① and $\widetilde{P}$ refers to ②

Introduction
○○○○○○

Examples
○○○○

Properties of MLEs
○○○●○○○○

# Proof

$\tilde{L}$ IS THE ~~LARLE~~ LIKELIHOOD FUNCTION FOR

LET $\hat{\theta}$ BE MLE FOR $\theta \in \textcircled{H}$

THE MODEL $\textcircled{2}$

To see that $\hat{\phi} = g(\hat{\theta})$ maximises $\tilde{L}$: For all $\phi \in \Phi$,

$$\tilde{L}(\hat{\phi}) = L(g^{-1}(\hat{\phi})) = L(g^{-1}(g(\hat{\theta})) = L(\hat{\theta}) \geq L(g^{-1}(\phi)) = \tilde{L}(\phi)$$

Thus $\hat{\phi}$ maximises $\tilde{L}$.

$\hat{\phi} = g(\hat{\theta})$

$g$ IS BIJECTIVE

$L(\hat{\theta}) \geq L(\theta) \quad \forall \theta \in \textcircled{H}$

$L(\hat{\theta}) \geq L(g^{-1}(\phi)) \quad \forall \phi \in \Phi$

$$\tilde{L}(\phi) = \tilde{P}_{\phi}(X=x) = \tilde{P}_{g^{-1}(\phi)}(X=x) = L(g^{-1}(\phi))$$

$$\implies \tilde{L}(\hat{\phi}) \geq \tilde{L}(\phi) \quad \forall \phi \in \Phi$$

$$\implies \hat{\phi} \text{ IS THE MLE FOR } \textcircled{2}$$

16 / 20

Introduction
oooooo

Examples
oooo

Properties of MLEs
ooooo●ooo

# What if $g$ is not bijective?

If $g$ is **not surjective** then there are $\phi$s that are not in the range of $g$, implying that for these parameter values no model is defined. In this case, we should set the likelihood for these parameter values to the lowest possible value (which is 0). By doing this, we can easily argue that the invariance of the MLE under the transformation induced by $g$ is retained.

$$\tilde{L}(\hat{\phi}) \geq \tilde{L}(\phi)$$

If $g$ is **not injective** then knowing $\phi$ does not uniquely identify the parameter $\theta$ or the model. One way to define the likelihood on $\Phi$ is the "induced" likelihood function

$$\tilde{L} : \mathbb{R} \to \mathbb{R}, \tilde{L}(\phi) = \sup\{L(\theta) : g(\theta) = \phi\}.$$

This gives every $\phi \in \Phi$ the highest likelihood of all $\theta$ that $g$ maps onto it.

With this definition the invariance is retained: If $\hat{\theta}$ is the MLE then $\hat{\phi} = g(\hat{\theta})$ maximises the induced likelihood function $\tilde{L}$.

Introduction
○○○○○○

Examples
○○○○

Properties of MLEs
○○○○○●○○

# Large sample properties

Let $X_1, X_2, \ldots$ be iid observations with pdf (or pmf) $f_\theta(x)$, where $\theta \in \Theta$ and $\Theta$ is an open interval. Let $\theta_0 \in \Theta$ denote the true parameter. Under regularity conditions (e.g. $\{x : f_\theta(x) > 0\}$ does not depend on $\theta$), the following holds:

(i) There exists a **consistent** sequence $(\hat{\theta}_n)_{n \in \mathbb{N}}$ of maximum likelihood estimators. [$\hat{\theta}_n$ is an MLE based on $X_1, \ldots, X_n$].

(ii) Suppose $(\hat{\theta}_n)_{n \in \mathbb{N}}$ is a consistent sequence of MLEs. Then

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, (I_f(\theta_0))^{-1}),$$

where $I_f(\theta) = E_\theta[(\frac{\partial}{\partial \theta} \log f_\theta(X))^2]$ is the **Fisher Information** of a sample of size 1.

Introduction
oooooo

Examples
oooo

Properties of MLEs
oooooo●o

# Estimating the variance

The above theorem has a limiting distribution that depends on $I_f(\theta_0)$, which would not be known in practical situations. To use this result, we need to estimate $I_f(\theta_0)$.

In an iid sample, $I_f(\theta_0)$ can be estimated by

- $I_f(\hat{\theta})$

- $\frac{1}{n} \sum_{i=1}^{n} \left( \frac{\partial}{\partial \theta} \log(f(x_i; \theta))|_{\theta=\hat{\theta}} \right)^2$

- $-\frac{1}{n} \sum_{i=1}^{n} \left( \frac{\partial}{\partial \theta} \right)^2 \log(f(x_i; \theta))|_{\theta=\hat{\theta}}$

These estimators are often consistent, i.e. will converge to $I_f(\theta_0)$ in probability.

Introduction
○○○○○○

Examples
○○○○

Properties of MLEs
○○○○○○○●

# Standard error of the MLE

Under mild regularity conditions, the standard error of an asymptotically normal MLE $\hat{\theta}_n$ can be approximated by $SE(\hat{\theta}_n) \approx \sqrt{\hat{I}_n^{-1}}/\sqrt{n}$. Here, $\hat{I}_n$ can be any of the estimators for $I_f(\theta_0)$ listed above.