# ST5201: Basic Statistical Theory
# Chapter 8: Estimation of Parameters and Fitting of Probability Distributions

CHOI Yunjin
stachoiy@nus.edu.sg

Department of Statistics and Applied Probability
National University of Singapore (NUS)

10th October, 2017

# Announcement

- Homework 3 released

# Outline

- Review
- Introduction
- Parameter Estimation
- The Method of Moments
- The Method of Maximum Likelihood
  - Large Sample Theory of MLE
- Cramer-Rao Lower bound

# Review: Convergence

- Three types of convergence
  - Convergence in distribution: point-wise convergence of CDF
  - Convergence in probability: relevant with sample space
  - Almost sure convergence: relevant with sample space; point-wise convergence for r.v.'s $X_n$ ("points" means "outcomes")
  - How to prove convergence can be found in tutorial
  - Property: a.s. convergence $\Rightarrow$ Convergence in Prob $\Rightarrow$ Convergence in Dist.
  - Property on cont. functions

# Review

- Law of Large Number (LLN)
  - Conditions: independent, share $\mu$ and $\sigma$
  - Results: $\bar{X}_n \overset{a.s./P}{\Longrightarrow} X$
  - Application: Monte Carlo method for integration
- Central Limit Theorem (CLT)
  - Conditions: i.i.d, $\mu$, $\sigma$, and mgf exists in a neighborhood of 0
  - Results: $\sqrt{n}\frac{\bar{X_n}-\mu}{\sigma} \xrightarrow{d} Z \Leftrightarrow \frac{S-n\mu}{\sqrt{n}\sigma} \xrightarrow{d} Z$, $Z \sim N(0,1)$
  - Applications in many fields, especially for unknown distribution.
  - Normal approximations of Poisson distribution

# Introduction

## Learning Outcomes

- Questions to Address: What is parametric model ⋆ How to estimate the parameters ⋆ How to evaluate an estimator ⋆ What is method of moments estimator ⋆ What is MLE ⋆ Asymptotic properties of MLE ⋆ Cramer-Rao lower bound ⋆ Confidence interval

# Introduction – Cont'd

## Concepts & Terminology

- Parametric model ⋆ Estimator ⋆ Method of Moments ⋆ Consistency
- Likelihood function ⋆ Maximum Likelihood Estimator (MLE)⋆ Log-likelihood function ⋆ Score function ⋆ Fisher Information ⋆ Asymptotic normality ⋆ Limiting distribution
- Bias ⋆ Unbiased Estimator ⋆ Variance ⋆ Estimated standard error ⋆ Mean Squared Error
- Cramer-Rao lower bound ⋆ Efficiency ⋆ Efficient Estimator
- Confidence interval ⋆ confidence level

## Mandatory Reading

- Section 8.1 - Section 8.5, Section 8.7 (Cramer-Rao Lower Bound)

# Introduction

In last lecture,

- We find $\bar{X}_n$ is a good estimate for $E(X)$ when $X_1, \cdots, X_n$ are i.i.d. samples, according to LLN.
- Further, CLT gives the limiting distribution of $\bar{X}_n$

For today,

- Can we estimate the distribution? How?
- Any theoretical results for these estimation?
- What is a good estimator? How to compare different estimators?

# Introduction - cont'd

- In genetics, we want to study the gene mutations, and wonder the distribution of mutations on the genes
- A doctor wants to choose the treatment with highest recovery rate for a specific patient
- One wants to split his/her money on several stocks to maximize the revenue
- Company makes a market survey to study the distribution of people interested in the goods
- Airlines want to know distribution of possible compensation for overselling tickets

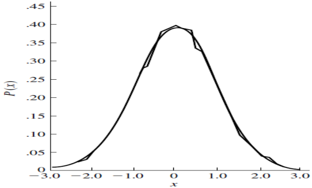
Distributions are of interest

# Parameter Estimation

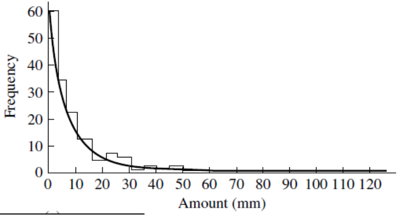
- The underlying distribution is in a specific distribution family, but the parameter is unknown
  - Example 1: Flip a coin, $Appearance\ of\ Head \sim Ber(p)$, with unknown parameter $p$
  - Example 2: $N = \#$ customer entering a shop. $N \sim Pois(\lambda)$, with unknown parameter $\lambda$
  - Specify the distribution $\Leftrightarrow$ specify the parameter (e.g., $p$, $\theta$)
  - The quantity of interest (e.g., mean, variance, ...) can be found
- The underlying distribution is totally unknown (More common)
  - Assume a form for the density according to the sample data, so only the parameters need to be estimated
  - Examples in the next slide
  - Estimate the parameter from the sample
  - Or, assume NO knowledge about the density – Non-parametric density estimation (not covered)
- In both cases, Parameter Estimation is the key

# Example: Parameter Estimation

- Random fluctuations of current across a muscle cell membrane[1]. Assume it comes from normal distribution family.



- The amounts of rainfall from different storms[2]. Assume it comes from Gamma distribution family.



[1] Bevan, Kullberg, and Rice (1979)

[2] Le Cam and Neyman (1967)

# General Set-up: Parametric Model

## Set-up of Parametric Model

1. Sample data with sample size $n$: $X_1$, $X_2$, $\cdots$, $X_n$.
2. Assumptions:
   - $X_1, X_2, \cdots, X_n$ are $i.i.d$ r.v.'s
   - The distribution of $X_i$ belongs to a family of distributions $\{\mathbb{P}_\theta, \theta \in \Theta\}$, indexed by parameter $\theta$
3. Goal: estimate $\theta$

Notations:

- $\theta$: the unknown parameter. $\theta$ can be a vector, e.g., $\theta = (\mu, \sigma)$ for normal distribution
- $\Theta$: The set of all possible values of $\theta$ (also called parameter space)
- $\{\mathbb{P}_\theta, \theta \in \Theta\}$: the family of distributions indexed by $\theta$. For each specified $\theta$, the corresponding PDF/PMF is clear, e.g., Poisson distribution indexed by $\theta = \lambda$.

# Example: Parametric Model

A coin is tossed 100 times. Let $X_i$ denotes the result of $i$-th toss with $X_i = 1$ for a head and $X_i = 0$ for a tail, $1 \leq i \leq 100$.

The set-up for this example:

- Sample size $n = 100$
- Sample data $X_1, X_2, \cdots, X_{100}$, each is either 0 or 1.
- Assumptions
  - $X_1, X_2, \cdots, X_{100}$ are i.i.d r.v.'s (checked)
  - The distribution of $X_i$ belongs to family of Bernoulli densities, where
  $$\mathbb{P}_\theta = \theta^x (1-\theta)^{1-x}, \qquad x = 0, 1,$$
  for $0 < \theta < 1$ ($\Theta = (0, 1)$).
- Goal: estimate $\theta$.

# Example: Parametric Model

Shoemaker (1996) collected the body temperature readings (degree Fahrenheit) of 65 males and 65 females. Assume the population distributions are normal, estimate the mean and standard deviations of the normal distribution.

The set-up for this example:

- Sample size $n = 130$
- Sample data $X_1, X_2, \cdots, X_{130}$, each is a real number.
- Assumptions
  - $X_1, X_2, \cdots, X_{130}$ are i.i.d r.v.'s (assumed)
  - The distribution of $X_i$ belongs to family of normal densities, where

    $$\mathbb{P}_\theta = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \qquad \theta = (\mu, \sigma)$$

    for $\sigma > 0$ (Here $\theta$ is a vector, and $\Theta = \mathbb{R} \times (0, \infty)$).
- Goal: estimate $\theta = (\mu, \sigma)$.

# Estimators

## Estimator

Suppose that the sample data $X_1, X_2, \cdots, X_n$ are *i.i.d* r.v.'s with the same distribution from $\mathbb{P}_\theta$. An *estimator, $\hat{\theta}$*, where

$$\hat{\theta} = \hat{\theta}_n = w(X_1, X_2, \cdots, X_n),$$

is a function of the data. We use $\hat{\theta}$ as our estimate of $\theta$.

- As a function of r.v.'s,. $\hat{\theta}$ is also a r.v.
- Usually, the performance of $\hat{\theta}$ is related to the sample size $n$.
- Good estimator
  - Small bias: $E(\hat{\theta}_n) \to \theta$, as $n \to \infty$
  - Small variance: $\text{Var}(\hat{\theta}_n) \to 0$, as $n \to \infty$
- Limiting dist. of $\hat{\theta}_n$ is important (limit theorems in Lecture 5)

# Methods of Parameter Estimation

## Two Approaches for Estimators

1. Assume the parameters are constants, and estimate these constants (Point Estimation)
   - Method of Moments (required)
   - Maximum Likelihood (required)
2. Assume the parameters follows a distribution according to prior knowledge, and estimate the PDF of the parameters with data
   - Bayesian Estimation (not covered)

# Motivation: The Method of Moments

A coin is tossed 100 times. Let $X_i$ denotes the result of $i$-th toss with $X_i = 1$ for a head and $X_i = 0$ for a tail, $1 \leq i \leq 100$. Assume that $X_i \sim Ber(\theta)$ for some $\theta$, what is a good estimate of $\theta$?

**Solution**. Note that

$$E(X_i) = \theta, \qquad 1 \leq i \leq n.$$

According to LLN, the average

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \overset{a.s.}{\to} E(X).$$

Hence, the sample average is a good estimator of $\theta$, where

$$\hat{\theta}_n = \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

# Motivation: The Method of Moments

Shoemaker (1996) collected the body temperature readings (degree Fahrenheit) of 65 males and 65 females. Assume the population distributions are normal, estimate the mean and standard deviations of the normal distribution.

**Solution**. As $E(X_i) = \mu$, according to LLN, a good estimator for $\mu$ is

$$\hat{\mu} = \bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

To estimate $\sigma^2$, note that

$$\sigma^2 = \mathrm{Var}(X_i) = E(X_i^2) - [E(X_i)]^2.$$

According to LLN, $\frac{1}{n} \sum_{i=1}^{n} X_i^2$ is a good estimator for $E(X_i^2)$, and so we have an estimator

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} X_i^2 - (\bar{X}_n)^2.$$

According to the theorem we introduced later, $\hat{\sigma}^2 \xrightarrow{a.s.} \sigma^2$.

# The Method of Moments

## The Method of Moments

1. Suppose $\theta = (\theta_1, \theta_2, \cdots, \theta_K)$, i.e., there are $K$ unknown parameters.
2. Calculate $K$ lower order moments in terms of $\theta$.

$$E(X) = h_1(\theta), \quad E(X^2) = h_2(\theta), \ \cdots, \ E(X^K) = h_K(\theta)$$

3. Find the inverse function of $h$'s to express the parameters.

$$\begin{aligned} \theta_1 &= f_1(E(X), E(X^2), \cdots, E(X^K)) \\ \cdots & \\ \theta_K &= f_K(E(X), E(X^2), \cdots, E(X^K)) \end{aligned}$$

4. Insert the sample moments into the expressions, thus obtaining the estimators $\hat{\theta}$.

$$\begin{aligned} \hat{\theta}_1 &= f_1(\frac{1}{n}\sum_{i=1}^n X_i, \frac{1}{n}\sum_{i=1}^n X_i^2, \cdots, \frac{1}{n}\sum_{i=1}^n X_i^K), \\ \cdots & \\ \hat{\theta}_K &= f_K(\frac{1}{n}\sum_{i=1}^n X_i, \frac{1}{n}\sum_{i=1}^n X_i^2, \cdots, \frac{1}{n}\sum_{i=1}^n X_i^K) \end{aligned}$$

# Consistent Estimator

**Definition: Consistency**

Let $\hat{\theta}_n$ be an estimate of a parameter $\theta$ based on a sample of size $n$. Then $\hat{\theta}_n$ is said to be *consistent in probability* if

$$\hat{\theta}_n \xrightarrow{P} \theta, \qquad n \to \infty,$$

or equivalently, for any $\epsilon > 0$,

$$P(|\hat{\theta}_n - \theta| > \epsilon) \to 0, \qquad n \to \infty.$$

- When we have enough observations (large $n$), it is guaranteed that the estimate $\hat{\theta}_n$ is arbitrarily close to $\theta$ with high probability $\Leftrightarrow$ we can always obtain more accurate estimate by increasing the sample size
- The Method of Moments estimator is a consistent estimator, as long as the functions $f_i$ is continuous at $(E(X), E(X^2), \cdots, E(X^K))$.

# Remark: The Method of Moments

- $\mathbb{P}_\theta$ can be any distribution indexed by $\theta$. It is not required to belong to the known density functions.
- Generally, $K$ lower order moments are enough to solve for the inverse functions of parameters. Otherwise, we can always calculate more moments to have more functions in Step 2.
- Advantages
  - Generally, this estimator is easy to calculate
  - The estimator is consistent
- Disadvantages:
  - Existence of moments is required
  - Sometimes, it is hard to find the limiting distribution of $\hat{\theta}_i$.
  - It does not consider the parameter space $\Theta$.

# Example: The Method of Moments

For the coin toss problem on page 16 where each sample follows Bernoulli distribution, the parameter of interest is the success rate $\theta$. What is the method of moment estimator of $\theta$? **Solution**. Here

$K = 1$. The first order moment is $E(X) = \theta$, so $\theta = E(X)$. Introduce the sample moments into the function, and the estimator is

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

# Example: The Method of Moments–cont'd

For the body temperature problem on page 17 where each sample follows Normal distribution, the parameter of interest is $\theta = (\mu, \sigma^2)$. What are the method of moments estimators of $\mu$ and $\sigma$?

**Solution**. The number of unknown parameters $K = 2$. The first order moment and the second order moment are

$$E(X) = \mu, \qquad E(X^2) = \mu^2 + \sigma^2.$$

The inverse function can be found as

$$\mu = E(X), \qquad \sigma^2 = E(X^2) - [E(X)]^2.$$

Introduce the sample moments into the function and get the estimators

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - (\frac{1}{n}\sum_{i=1}^{n} X_i)^2.$$

Note: For $\sigma$, the inverse function is $\sigma = \sqrt{E(X^2) - [E(X)]^2}$, and the estimator is $\hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} X_i^2 - (\frac{1}{n}\sum_{i=1}^{n} X_i)^2}$.

# Example: The Method of Moments - cont'd

Let $X_1, X_2, \cdots, X_n$ be gamma random variables with parameters $\alpha$ and $\lambda$, so that the probability density function is:

$$f(x) = \begin{cases} \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}, & x \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

What are the method of moments estimators of $\alpha$ and $\lambda$?

**Solution**. For this problem, the parameter of interest is $\theta = (\alpha, \lambda)$, so the number of unknown parameters $K = 2$. The first order moment and the second order moment are

$$E(X) = \alpha/\lambda, \qquad E(X^2) = \alpha^2/\lambda^2 + \alpha/\lambda^2.$$

The inverse function can be found as

$$\alpha = \frac{[E(X)]^2}{E(X^2) - [E(X)]^2}, \qquad \lambda = \frac{E(X)}{E(X^2) - [E(X)]^2}.$$

Introduce the sample moments into the function and get the estimators

$$\hat{\alpha} = \frac{[\frac{1}{n}\sum_{i=1}^n X_i]^2}{\frac{1}{n}\sum_{i=1}^n X_i^2 - [\frac{1}{n}\sum_{i=1}^n X_i]^2}, \qquad \hat{\lambda} = \frac{\frac{1}{n}\sum_{i=1}^n X_i}{\frac{1}{n}\sum_{i=1}^n X_i^2 - [\frac{1}{n}\sum_{i=1}^n X_i]^2}.$$

# Example: The Method of Moments - cont'd

Suppose that $X$ is a discrete r.v. with

$P(X=0)=\frac{2}{3}\theta$, $P(X=1)=\frac{1}{3}\theta$, $P(X=2)=\frac{2}{3}(1-\theta)$, $P(X=3)=\frac{1}{3}(1-\theta)$, and $P(X=x)=0$ for $x \notin \{0,1,2,3\}$, where $0 \leq \theta \leq 1$ is a parameter. Here are 10 indept. observations taken from such a distribution: (3, 0, 2, 1, 3, 2, 1, 0, 2, 1). Find the method of moments estimate of $\theta$.

**Solution**. For this problem, the parameter of interest is $\theta$, so the number of unknown parameters $K=1$. The first order moment is

$$E(X)=(0)\frac{2}{3}\theta+(1)\frac{1}{3}\theta+(2)\frac{2}{3}(1-\theta)+(3)\frac{1}{3}(1-\theta)=-2\theta+\frac{7}{3}.$$

The inverse function can be found as

$$\theta=(E(X)-7/3)/(-2).$$

Introduce the sample moments into the function and get the estimators

$$\hat{\theta}=(\frac{1}{n}\sum_{i=1}^{n} X_i-7/3)/(-2).$$

According to the data, $\frac{1}{n}\sum_{i=1}^{n} X_i=15/10=3/2$, so the estimate is

$$\hat{\theta}=(3/2-7/3)/(-2)=5/12.$$

# Maximum Likelihood Method

## Maximum Likelihood Estimator (MLE)

Suppose $X_1, \cdots, X_n$ are $n$ measurements of $X$ with PDF $f(x|\theta)$ or PMF $p(x|\theta)$, and the joint PDF/PMF is $f(x_1, x_2, \cdots, x_n|\theta)$ or $p(x_1, x_2, \cdots, x_n|\theta)$. Given that we have observed $X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n$, the *Maximum Likelihood Estimator (MLE)* for $\theta$ is

$$\hat{\theta}_{MLE} = \begin{cases} \arg\max_{\theta\in\Theta} f(x_1, x_2, \cdots, x_n|\theta), & \text{cont. r.v.'s;} \\ \arg\max_{\theta\in\Theta} p(x_1, x_2, \cdots, x_n|\theta), & \text{discrete r.v.'s.} \end{cases}$$

- It is not required that $X_1, X_2, \cdots, X_n$ are independent, yet the joint PDF/PMF is required.
- On the other hand, in general $X_1, \cdots, X_n$ are assumed to be i.i.d
- The MLE says that we pick $\theta$ such as to maximize the probability of getting the measurements (the $X_i$'s) that we obtained!
- The parameter space is concerned to obtain MLE
- However, maximizing the PDF/PMF might be complicated/impossible

# Maximum Likelihood Method

## A General Set Up

Let $L_n(\theta) = f(x_1, x_2, \cdots, x_n|\theta)$ be the *Likelihood function* for the data. If $X_i$ are i.i.d, $1 \leq i \leq n$, the likelihood simplifies to

$$L_n(\theta) = \prod_{i=1}^{n} f(x_i|\theta).$$

- $L_n(\theta)$ is the probability of observing the given data as a function of parameter $\theta$
- $\hat{\theta}_{MLE} = \arg\max_{\theta \in \Theta} L_n(\theta)$

In practice, it is easier to maximize $\ln L_n(\theta)$ (namely, $l_n(\theta)$) rather than the likelihood itself, since $l_n(\theta)$ turns the product into a summation.

## Log-likelihood Function

The *log-likelihood function* for i.i.d observations $X_1, X_2, \cdots, X_n$ with pdf $f$ is defined as

$$l_n(\theta) = \sum_{i=1}^{n} \ln(f(x_i|\theta)),$$

and the MLE estimator can be achieved by maximizing $l_n(\theta)$:

$$\hat{\theta}_{MLE} = \arg\max_{\theta \in \Theta} l_n(\theta).$$

# Intuition: Maximum Likelihood Method

1. Given the values of $x_1, \cdots, x_n$, by getting a value of $\theta$ as $\hat{\theta}_{MLE}$ via MLE, we establish

$$f(x_1, \cdots, x_n|\hat{\theta}_{MLE}) \geq f(x_1, \cdots, x_n|\theta^*) \text{ for any } \theta^* \in \Theta$$

2. Among all parameters $\theta \in \Theta$, $\hat{\theta}_{MLE}$ makes it most likely to observed the data $X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n$.

3. However, compare to another set of values $x_1^*, x_2^*, \cdots, x_n^*$, it is possible that

$$f(x_1^*, \cdots, x_n^*|\hat{\theta}_{MLE}) \geq f(x_1, \cdots, x_n|\hat{\theta}_{MLE}),$$

which means that $(x_1, \cdots, x_n)$ is not the most possible set of values even when $\theta = \hat{\theta}_{MLE}$

# Examples of MLE's – Bernoulli

For the coin toss problem, $X_i \overset{i.i.d}{\sim} Ber(\theta)$, $0 < \theta < 1$. For $X_1, X_2, \cdots, X_n$, the joint PMF is

$$L_n(\theta) = p(x_1, \cdots, x_n|\theta) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}$$

Taking the log with base $e$,

$$l_n(\theta) = \left(\sum_{i=1}^{n} x_i\right)\ln(\theta) + \left(n - \sum_{i=1}^{n} x_i\right)\ln(1-\theta)$$

Taking the derivative of $l_n(\theta)$, and setting to 0, we get

$$\frac{d}{d\theta} l_n(\theta) = \frac{\sum_{i=1}^n x_i}{\theta} - \frac{n - \sum_{i=1}^n x_i}{1-\theta} = 0$$

$$\hat{\theta}_{MLE} = \frac{\sum_{i=1}^n X_i}{n} = \bar{X}_n$$

Note that $\frac{d}{d\theta} l_n(\theta) > 0$ when $\theta < \bar{X}_n$, and $\frac{d}{d\theta} l_n(\theta) < 0$ when $\theta > \bar{X}_n$, so $\hat{\theta}_n = \bar{X}_n$ achieves the maximum.

# Examples of MLE's – Poisson

Suppose $X_1, X_2, \cdots, X_n$ are *i.i.d* Poisson r.v.'s with parameter $\theta$. Find the MLE for $\theta$.

**Solution**. The PMF for $X \sim Pois(\theta)$ is $p_X(x) = \frac{\theta^x e^{-\theta}}{x!}$. So the joint PMF for $X_1, X_2, \cdots, X_n$ is

$$L_n(\theta) = p(X_1, X_2, \cdots, X_n|\theta) = \prod_{i=1}^{n} \frac{\theta^{x_i} e^{-\theta}}{x_i!}.$$

The log likelihood will thus be:

$$l_n(\theta) = \sum_{i=1}^{n} (x_i \ln\theta - \theta - \ln x_i!) = \ln\theta \sum_{i=1}^{n} x_i - n\theta - \sum_{i=1}^{n} \ln x_i!$$

Take the derivative of $l_n(\theta)$ and let it be 0:

$$\frac{d}{d\theta} l_n(\theta) = \frac{1}{\theta} \sum_{i=1}^{n} x_i - n = 0 \quad \Longrightarrow \quad \hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} X_i = \bar{X}_n$$

Note that $\frac{d}{d\theta} l_n(\theta) > 0$ when $\theta < \bar{X}_n$, and $\frac{d}{d\theta} l_n(\theta) < 0$ when $\theta > \bar{X}_n$, so $\hat{\theta}_n = \bar{X}_n$ achieves the maximum.

# Examples of MLE's – Normal

For the body temperature problem where $X_1, X_2, \cdots, X_n$ are i.i.d. $N(\mu, \sigma^2)$ r.v.'s, find the MLE for $\mu$ and $\sigma^2$.

**Solution**. The likelihood function for $(X_1, X_2, \cdots, X_n)$ is

$$L_n(\mu, \sigma^2) = \prod_{i=1}^{n} f(x_i|\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x_i - \mu)^2}{2\sigma^2}\right\}$$

The log likelihood is

$$l_n(\mu, \sigma^2) = -\frac{n}{2}\ln\sigma^2 - \frac{n}{2}\ln(2\pi) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

Find the maximum by finding the partial derivatives w.r.t $\mu$ and $\sigma^2$:

$$\frac{\partial l}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu)$$

$$\frac{\partial l}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i - \mu)^2$$

# Examples of MLE's – Normal - cont'd

Set the partials to zero, we have

$$0 = \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu)$$

$$0 = -\frac{n}{\sigma^2} + \frac{1}{\sigma^4} \sum_{i=1}^{n} (x_i - \mu)^2$$

The solutions give the MLE's as

$$\hat{\mu} = \bar{X}_n, \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2 = \frac{1}{n} \sum_{i=1}^{n} X_i^2 - (\bar{X}_n)^2$$

Verify that this extreme value is the maximum. Note that $\frac{\partial^2 l}{\partial \mu^2} = -\frac{n}{\sigma^2} < 0$, so for any $\sigma^2$, $\hat{\mu} = \bar{X}_n$ is the maximum point. For $\sigma^2$, the second partial derivative is

$$\frac{\partial^2 l}{\partial (\sigma^2)^2} |_{(\hat{\mu}, \hat{\sigma}^2)} = \frac{n}{2\sigma^4} - \frac{1}{\sigma^6} \sum_{i=1}^{n} (X_i - \mu)^2 |_{(\hat{\mu}, \hat{\sigma}^2)} = -\frac{n}{2\hat{\sigma}^4} < 0.$$

So $\hat{\sigma}^2$ is the maximum point.

Note: this method is useful only for this function, as for **any** $\sigma^2$, $\hat{\mu}$ is the maximum point.

# Examples of MLE's – Gamma

If $X_1, X_2, \cdots, X_n$ iid Gamma$(\alpha, \lambda)$ random variables, find the MLE for $\alpha$ and $\lambda$

**Solution**. The likelihood function for $(X_1, X_2, \cdots, X_n)$ is

$$L_n(\alpha, \lambda) = \prod_{i=1}^{n} \frac{\lambda^{\alpha}}{\Gamma(\alpha)} x_i^{\alpha-1} e^{-\lambda x_i}, \; x_i \geq 0$$

The log likelihood is

$$l_n(\alpha, \lambda) = n\alpha \ln \lambda + (\alpha - 1) \sum_{i=1}^{n} \ln x_i - \lambda \sum_{i=1}^{n} x_i - n \ln \Gamma(\alpha)$$

Find the maximum by finding the partial derivatives w.r.t $\alpha$ and $\lambda$:

$$\frac{\partial l_n(\alpha, \lambda)}{\partial \alpha} = n \ln \lambda + \sum_{i=1}^{n} \ln x_i - n \frac{\Gamma^{'}(\alpha)}{\Gamma(\alpha)}$$

$$\frac{\partial l_n(\alpha, \lambda)}{\partial \lambda} = \frac{n\alpha}{\lambda} - \sum_{i=1}^{n} x_i$$

# Examples of MLE's – Gamma - cont'd

Set the second partials to zero, we have

$$\hat{\lambda} = \frac{\hat{\alpha}}{\bar{X}_n}$$

Substituting it into the first partial, we obtain a nonlinear equation for the *MLE* of $\alpha$

$$\ln \hat{\alpha} - \frac{\Gamma^{'}(\hat{\alpha})}{\Gamma(\hat{\alpha})} + \frac{1}{n}\sum_{i=1}^{n} \ln X_i - \ln \bar{X}_n = 0$$

- The equation cannot be solved analytically. An iterative method for finding the roots (e.g., Newton-Raphson) has to be employed using software.
- Based on a random sample of size $n = 227$, we obtain MLE's $\hat{\alpha} = .441$ and $\hat{\lambda} = 1.96$, provided that we assume a gamma model from which the 227 observations are sampled.

# Examples of MLE's - Comparison with MM

Suppose that $X$ is a discrete r.v. with

$$P(X=0)=\frac{2}{3}\theta,\ P(X=1)=\frac{1}{3}\theta,\ P(X=2)=\frac{2}{3}(1-\theta),\ P(X=3)=\frac{1}{3}(1-\theta),$$

and $P(X=x)=0$ for $x\notin\{0,1,2,3\}$, where $0\le\theta\le1$ is a parameter. Here are 10 indept. observations taken from such a distribution: (3, 0, 2, 1, 3, 2, 1, 0, 2, 1). Find the MLE of $\theta$.

**Solution**. The likelihood function is

$$L_n(\theta)=\prod_{i=1}^{n}p_X(x)=(\frac{2}{3}\theta)^{N_0}(\frac{1}{3}\theta)^{N_1}[\frac{2}{3}(1-\theta)]^{N_2}[\frac{1}{3}(1-\theta)]^{N_3},$$

where $N_k=\#$ observations with $X_i=k$, $1\le i\le n$, $k=0,1,2,3$. The log-likelihood function is

$$l_n(\theta)=\ln L_n(\theta)=N_0\ln(\frac{2}{3}\theta)+N_1\ln(\frac{1}{3}\theta)+N_2\ln[\frac{2}{3}(1-\theta)]+N_3\ln[\frac{1}{3}(1-\theta)]$$

Take the derivative of $l_n(\theta)$ and set the derivative to be 0,

$$\frac{dl_n(\theta)}{d\theta}=\frac{N_0}{\theta}+\frac{N_1}{\theta}-\frac{N_2}{1-\theta}-\frac{N_3}{1-\theta}\overset{set}{\equiv}0\quad\Longrightarrow\quad\hat{\theta}_{MLE}=\frac{N_0+N_1}{N_0+N_1+N_2+N_3},$$

and verify that this is the maximum point.
According to the data, $\hat{\theta}_{MLE}=1/2$. (Previous slide: $\hat{\theta}=5/12$!)

# Large Sample/Asymptotic Theory of MLE

- For the Method of Moments estimator, we have $\hat{\theta}_n \xrightarrow{p} \theta$.
- Is MLE consistent?
- Further, is it possible to find out the limiting distribution for $\hat{\theta}_{MLE}$?

# Consistency of MLE

### Consistency of MLE

Under appropriate smoothness conditions on $f(x|\theta)$, the MLE $\hat{\theta}_n$ from an $i.i.d$ sample $X_1, X_2, \cdots, X_n$ sharing the common PDF/PMF $f(x|\theta_0)$ is consistent, i.e., for any $\epsilon > 0$

$$P(|\hat{\theta}_n - \theta_0| \geq \epsilon) \to 0, \qquad n \to \infty$$

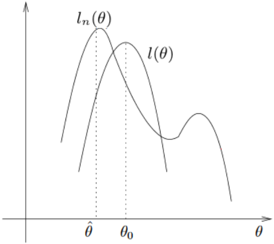
- The “appropriate smoothness conditions” will be specified in future stat. modules, such as point estimation
- MLE is consistent $\Rightarrow$ We can always increase the accuracy of MLE by increasing the sample size, and the estimation can be arbitrarily close to the truth with large probability as long as the sample size is large enough

# Proof: Consistency of MLE

**Proof**. Define

$$l(\theta) = \ln f(X|\theta), \qquad \hat{l}_n(\theta) = \frac{1}{n} l_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ln f(X_i|\theta)$$

Then $l(\theta)$ is the log-likelihood for $X$, which can be viewed as a new r.v., and $\hat{l}_n(\theta)$ is the average of $n$ i.i.d observations of $l(\theta)$.



$$\left\{ \begin{array}{l} \theta_0 = \arg\max E[l(\theta)] \\ \hat{\theta}_{MLE} = \arg\max \hat{l}_n(\theta) \\ \hat{l}_n(\theta) \to E[l(\theta)] \end{array} \right. \Longrightarrow \hat{\theta} \to \theta_0$$

# Proof: Consistency of MLE - cont'd

- According to LLN, $\hat{l}_n(\theta) \to E[l(\theta)]$ for each $\theta$.
- Re-write MLE: $\hat{\theta}_{MLE} = \arg\max_{\theta \in \Theta} \frac{1}{n} l_n(\theta) = \arg\max_{\theta \in \Theta} \hat{l}_n(\theta)$
- To maximize $E[l(\theta)]$, we consider its derivative:

$$\begin{aligned} \frac{\partial}{\partial \theta} E[l(\theta)] = \frac{\partial}{\partial \theta} E[\ln f(X|\theta)] &= \frac{\partial}{\partial \theta} \int \ln f(x|\theta) f(x|\theta_0) dx \\ &= \int \frac{\frac{\partial}{\partial \theta} f(x|\theta)}{f(x|\theta)} f(x|\theta_0) dx \end{aligned}$$

At the point $\theta = \theta_0$, the derivative becomes

$$\int \frac{\partial}{\partial \theta} f(x|\theta) dx|_{\theta=\theta_0} = \left( \frac{\partial}{\partial \theta} \int f(x|\theta) dx \right) |_{\theta=\theta_0} = \frac{\partial}{\partial \theta}(1) = 0$$

Later, we would show $E[l''(\theta)]|_{\theta=\theta_0} < 0$, which means that this local extrema is a maximum. We need strong smooth condition on $f$ to justify the interchangeability and integration above.

So the consistency is proved.

# Large Sample/Asymptotic Theory of MLE

Question: limiting distribution for MLE?

**Definition: Score Function**

Suppose $X$ is a single observation with PDF/PMF $f(x|\theta)$, we call the the partial derivative w.r.t. $\theta$ of the natural logarithm of the likelihood function for a $X$ as *score function*, which is

$$\frac{\partial}{\partial\theta} l(\theta) = \frac{\partial}{\partial\theta} \ln f(X|\theta).$$

- For given $\theta$, $l(\theta) = \ln f(X|\theta)$ is a function of the r.v. $X$, which is also a r.v.
- For given $\theta$, the score function $\frac{\partial}{\partial} l(\theta)$ is also a r.v., for which we can calculate expectation and variance.
- According to our analysis in previous slide, the expectation of the score function when $X$ has PDF/PMF $f(x|\theta)$ is

$$E\left(\frac{\partial}{\partial\theta} \ln f(X|\theta)\right) = 0$$

# Large Sample/Asymptotic Theory of MLE – cont'd

### Definition: Fisher Information

The variance of the score function, is called the *Fisher information*,

$$I(\theta) = E\left[\left(\frac{\partial}{\partial\theta} l(\theta)\right)^2\right] = E\left[\left(\frac{\partial}{\partial\theta} \ln f(X|\theta)\right)^2\right].$$

Under appropriate smoothness conditions on $f$, $I(\theta)$ can be reduced to

$$I(\theta) = -E\left[\frac{\partial^2}{\partial\theta^2} \ln f(X|\theta)\right] = -E\left[\frac{\partial^2}{\partial\theta^2} l(\theta)\right]$$

- As the expectation is 0, the variance is $E[(\frac{\partial}{\partial\theta} l(\theta) - 0)^2] = E[\left(\frac{\partial}{\partial\theta} l(\theta)\right)^2]$
- The variance equals to the product of $-1$ and expectation of the second partial derivative of $l(\theta)$ w.r.t. $\theta$, given $X \sim f(X|\theta)$. This expression is easier to calculate
- Note that $I(\theta) > 0 \Leftrightarrow$ that $E\left[\frac{\partial^2}{\partial\theta^2} l(\theta)\right]$ is always negative at $\theta$ (proof for maximum in previous slide)

# Large Sample/Asymptotic Theory of MLE – cont'd

## Fisher Information of a Random Sample

Suppose $\mathbf{X} = (X_1, \cdots, X_n)$ are i.i.d. r.v.'s from a distribution $f(x|\theta)$, where the log-likelihood function is

$$l_n(\theta) = \sum_{i=1}^{n} \ln f(X_i|\theta) \text{ (because of iid sample).}$$

The *Fisher information of the random sample* $\mathbf{X}$ is defined as

$$\begin{aligned}
I_{\mathbf{n}}(\theta) &= E\left[\left(\frac{\partial}{\partial\theta} l_n(\theta)\right)^2\right] = E\left[\left(\frac{\partial}{\partial\theta}\left(\sum_{i=1}^{n} \ln f(X_i|\theta)\right)\right)^2\right] \quad (1) \\
&= \sum_{i=1}^{n} E\left[\left(\frac{\partial}{\partial\theta} \ln f(X_i|\theta)\right)^2\right] + \sum_{i\neq j} E\left[\frac{\partial}{\partial\theta} \ln f(X_i|\theta)\right] E\left[\frac{\partial}{\partial\theta} \ln f(X_j|\theta)\right] \quad (2) \\
&= -E\left[\frac{\partial^2}{\partial\theta^2}\left(\sum_{i=1}^{n} \ln f(X_i|\theta)\right)\right] \quad (3) \\
&= nI(\theta) \quad (4)
\end{aligned}$$

- Obviously, Equations (2) and (4) are for i.i.d. random sample only
- Comparison: $l_n(\theta) \neq nl(\theta)$.

# Large Sample/Asymptotic Theory of MLE - cont'd

## Asymptotic Normality of MLE

For an i.i.d. sample $X_1, X_2, \cdots, X_n$ from $f(x|\theta_0)$, let the MLE be $\hat{\theta}_n = \arg\max_{\theta \in \Theta} \sum_{i=1}^{n} \ln f(X_i|\theta)$. Under appropriate smoothness conditions on $f(x|\theta)$, *the MLE $\hat{\theta}_n$ asymptotically normal*, i.e., as $n \to \infty$,

$$\sqrt{I_n(\theta_0)}(\hat{\theta}_n - \theta_0) = \sqrt{nI(\theta_0)}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0,1)$$

- In practice, $\sqrt{I(\theta_0)}$ can be approximated by $\sqrt{I(\hat{\theta}_n)}$
- We call the results asymptotic results as it holds when $n \to \infty$.
- The results are also valid for multivariate $\theta$; however, asymptotic results for multivariate $\theta$ are out of the scope of this class

# Large Sample/Asymptotic Theory of MLE - cont'd

### Another view of the limiting distribution

The asymptotic normality for MLE

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

can also be expressed (in a non-rigorous way) as

$$\hat{\theta}_n - \theta_0 \xrightarrow{d} N(0, \frac{1}{nI(\theta_0)}) \Longleftrightarrow \hat{\theta}_n \xrightarrow{d} N(\theta_0, \frac{1}{nI(\theta_0)})$$

- The asymptotic mean of $\hat{\theta}_n$ is $\theta_0$, which also shows its consistency
- The asymptotic variance of $\hat{\theta}_n$ is

$$\frac{1}{nI(\theta_0)} = -\frac{1}{nE(l''(\theta_0))}$$

# Example of Fisher Information - Bernoulli

Recall the coin toss problem, $X_i \overset{i.i.d}{\sim} Ber(\theta)$, $0 < \theta < 1$. We have figured that the MLE is $\hat{\theta}_n = \bar{X}_n$. Find the limiting distribution of $\hat{\theta}_n$ with Fisher information.

**Solution**: The point mass function of $X$ is

$$f(x|\theta) = \theta^x (1-\theta)^{1-x} \text{ for } x = 1 \text{ or } x = 0,$$

and

$$l(x|\theta) = \ln f(x|\theta) = x \ln \theta + (1-x)\ln(1-\theta)$$

$$l^{'}(x|\theta) = \frac{x}{\theta} - \frac{1-x}{1-\theta} \text{ and } l^{''}(x|\theta) = -\frac{x}{\theta^2} - \frac{1-x}{(1-\theta)^2}$$

By $E(X) = \theta$, we have

$$I(x|\theta) = -E[l^{''}(x|\theta)] = \frac{E(X)}{\theta^2} + \frac{1-E(X)}{(1-\theta)^2} = \frac{1}{\theta} + \frac{1}{1-\theta} = \frac{1}{\theta(1-\theta)}$$

Therefore, the limiting distribution for $\bar{X}_n$ is

$$\sqrt{\frac{n}{\theta(1-\theta)}}(\bar{X}_n - \theta) \sim N(0,1) \quad \text{Same as CLT}$$

# Example of Fisher Information - Poisson

Suppose we have an i.i.d. sample $X_1, \cdots, X_n$ from Poisson distribution with parameter $\lambda$. Recall the MLE for $\lambda$ is $\hat{\lambda}_n = \bar{X}_n$. Find the limiting distribution of MLE with Fisher information.

**Solution**: The point mass function of $X$ is

$$f(x|\lambda) = \frac{e^{-\lambda}\lambda^x}{x!} \text{ for } x = 0, 1, 2, 3, \cdots$$

and

$$l(x|\lambda) = \ln f(x|\lambda) = x\ln\lambda - \lambda - \ln x!$$

$$l'(x|\lambda) = \frac{x}{\lambda} - 1 \text{ and } l''(x|\lambda) = -\frac{x}{\lambda^2}$$

By $E(X) = \theta$, we have

$$I(\lambda) = -E\left[\frac{\partial^2}{\partial\lambda^2}\ln f(X|\lambda)\right] = \frac{E(X)}{\lambda^2} = \frac{1}{\lambda}$$

Therefore, the limiting distribution for $\bar{X}_n$ is

$$\sqrt{\frac{n}{\lambda}}(\bar{X}_n - \lambda) \sim N(0, 1) \quad \text{Same as CLT}$$

# Example of Fisher Information - Normal

Suppose $X \sim N(\mu, \sigma^2)$ and parameter $\mu$ is unknown but $\sigma^2$ is known. Find the Fisher information $I(\mu)$ in X and the corresponding limiting distribution for $\hat{\mu}_{MLE} = \bar{X}_n$

**Solution**: For $\infty < x < \infty$, we have

$$f(x|\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$l(x|\mu) = \ln f(x|\mu) = -\frac{1}{2}\ln(2\pi\sigma^2) - \frac{(x-\mu)^2}{2\sigma^2}$$

$$l^{'}(x|\mu) = \frac{x-\mu}{\sigma^2} \text{ and } l^{''}(x|\mu) = -\frac{1}{\sigma^2}$$

It follows that the Fisher information is

$$I(x|\mu) = -E[l^{''}(x|\mu)] = \frac{1}{\sigma^2}$$

Therefore, the limiting distribution for $\bar{X}_n$ is

$$\sqrt{\frac{n}{\sigma^2}}(\bar{X}_n - \mu) \sim N(0, 1)$$

# Conclusion on MLE

- MLE is defined as the maximizer of Likelihood function $L_n(\theta)$, equivalently, the maximizer of log-likelihood function $l_n(\theta)$
- Consistency: MLE $\hat{\theta}_n \to \theta$ in probability
- Define the fisher information as $I(\theta) = -E(l''(\theta))$, then $\sqrt{nI(\theta)}(\hat{\theta}_{MLE} - \theta) \approx N(0,1)$ when $n \to \infty$
- Advantages:
  - MLE is consistent
  - Limiting distribution is clear
  - Involve the information of $\Theta$ (e.g., when $\Theta = \{1, 2\}$ for Poisson distribution)
  - Allow relationship between samples, as long as $L_n(\theta)$ is known
- Disadvantages:
  - Calculation might be complicated (have to calculate the maximizer of function), or even unachievable