# Proability and Statistics

## MA20205

Bibhas Adhikari

Autumn 2022-23, IIT Kharagpur

Lecture 18
November 7, 2022

# Estimation

## Observations

- The special distributions are defined by certain parameters. For instance, Bernoulli depends on $p$ (probability of success), Poisson distribution depends on the parameter $\lambda$ (the expected number of arrivals); Gaussian distribution with parameters $\mu, \sigma$ etc.

# Estimation

## Observations

- The special distributions are defined by certain parameters. For instance, Bernoulli depends on $p$ (probability of success), Poisson distribution depends on the parameter $\lambda$ (the expected number of arrivals); Gaussian distribution with parameters $\mu, \sigma$ etc.
- In the real world, for a given data set, we do not know the distribution of the source random variable. We need to estimate and learn the 'true' parameters from the observed data, assuming or speculating the underlying distribution

# Estimation

Except the simple 'method of moments' there are two school of thoughts to estimate the value of parameters.

# Estimation

Except the simple 'method of moments' there are two school of thoughts to estimate the value of parameters.

1. Maximum Likelihood Estimation (MLE) and

# Estimation

Except the simple 'method of moments' there are two school of thoughts to estimate the value of parameters.

1. Maximum Likelihood Estimation (MLE) and
2. Maximum A Posteriori (MAP)

# Estimation

Except the simple 'method of moments' there are two school of thoughts to estimate the value of parameters.

1. Maximum Likelihood Estimation (MLE) and
2. Maximum A Posteriori (MAP)

First we assume that our data are (IID) samples: $X_1, X_2, \ldots, X_n$. As usual, we denote $X$ as the underlying population random variable.

# Estimation

## Maximum Likelihood Estimation (MLE)

# Estimation

## Maximum Likelihood Estimation (MLE)

We denote the pdf of the underlying or shared distribution by the sample random variables as

$$f(X \mid \boldsymbol{\theta}).$$

Here $\boldsymbol{\theta}$ could be a scalar or vector. For instance, if the random sample $X_i \sim \mathcal{N}(\mu, \sigma)$ then $\boldsymbol{\theta} = (\mu, \sigma)$.

# Estimation

Maximum Likelihood Estimation (MLE)

We denote the pdf of the underlying or shared distribution by the sample random variables as

$$f(X \mid \boldsymbol{\theta}).$$

Here $\boldsymbol{\theta}$ could be a scalar or vector. For instance, if the random sample $X_i \sim \mathcal{N}(\mu, \sigma)$ then $\boldsymbol{\theta} = (\mu, \sigma)$.

Note that we are using this notation to emphasize: conditional means that the likelihood of different values of $X$ depends on the values of the parameters. We use the same notation for both discrete and continuous distributions.

# Estimation

Question What do we mean by 'Likelihood'?

# Estimation

Question What do we mean by 'Likelihood'?

- if $X$ is discrete: likelihood can be thought as a synonym for the probability mass, or joint probability mass, of the given data

# Estimation

Question What do we mean by 'Likelihood'?

- if $X$ is discrete: likelihood can be thought as a synonym for the probability mass, or joint probability mass, of the given data
- if $X$ is continuous: likelihood refers to the probability density of the data

# Estimation

Question What do we mean by 'Likelihood'?

- if $X$ is discrete: likelihood can be thought as a synonym for the probability mass, or joint probability mass, of the given data
- if $X$ is continuous: likelihood refers to the probability density of the data

# Estimation

Question What do we mean by 'Likelihood'?

- if $X$ is discrete: likelihood can be thought as a synonym for the probability mass, or joint probability mass, of the given data
- if $X$ is continuous: likelihood refers to the probability density of the data

Since each of the data point is independent, the likelihood of the entire observed data is the product of the likelihood of each data point. Mathematically, the likelihood of the given data give parameters $\boldsymbol{\theta}$ is:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(X_i \mid \boldsymbol{\theta}).$$

# Estimation

Question What do we mean by 'Likelihood'?

- if $X$ is discrete: likelihood can be thought as a synonym for the probability mass, or joint probability mass, of the given data
- if $X$ is continuous: likelihood refers to the probability density of the data

Since each of the data point is independent, the likelihood of the entire observed data is the product of the likelihood of each data point. Mathematically, the likelihood of the given data give parameters $\boldsymbol{\theta}$ is:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(X_i \mid \boldsymbol{\theta}).$$

Note: The likelihood function depends on $\boldsymbol{\theta}$

# Estimation

## Maximization

Our goal is now to determine values of our parameters ($\boldsymbol{\theta}$) that maximizes the likelihood function. Thus MLE is concerned with solving the maximization problem:

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} L(\boldsymbol{\theta})$$

The argmax of a function is the value of the domain at which the function is maximized.

# Estimation

## Maximization

Our goal is now to determine values of our parameters ($\boldsymbol{\theta}$) that maximizes the likelihood function. Thus MLE is concerned with solving the maximization problem:

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} L(\boldsymbol{\theta})$$

The argmax of a function is the value of the domain at which the function is maximized.

## Observation

- property of argmax: since log is a monotone function, the argmax of a function is the same as the argmax of the log of the function

# Estimation

## Maximization

Our goal is now to determine values of our parameters ($\boldsymbol{\theta}$) that maximizes the likelihood function. Thus MLE is concerned with solving the maximization problem:

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} L(\boldsymbol{\theta})$$

The argmax of a function is the value of the domain at which the function is maximized.

## Observation

- property of argmax: since log is a monotone function, the argmax of a function is the same as the argmax of the log of the function
- if we find the argmax of the log of likelihood it will be equal to the argmax of the likelihood

# Estimation

## Maximization

Our goal is now to determine values of our parameters ($\boldsymbol{\theta}$) that maximizes the likelihood function. Thus MLE is concerned with solving the maximization problem:

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} L(\boldsymbol{\theta})$$

The argmax of a function is the value of the domain at which the function is maximized.

## Observation

- property of argmax: since log is a monotone function, the argmax of a function is the same as the argmax of the log of the function
- if we find the argmax of the log of likelihood it will be equal to the argmax of the likelihood
- chose the value of parameters that maximize the log likelihood function

# Estimation

**Bernoulli MLE Estimation** We are going to estimate the value of the parameter $p$ based on $n$ given data points corresponding to as IID random variables $X_1, X_2, \ldots, X_n$ such that $X_i \sim Ber(p)$.

The pdf of $X_i$ is $f(x) = p^x(1-p)^{1-x}$.

# Estimation

**Bernoulli MLE Estimation** We are going to estimate the value of the parameter $p$ based on $n$ given data points corresponding to as IID random variables $X_1, X_2, \ldots, X_n$ such that $X_i \sim Ber(p)$.

The pdf of $X_i$ is $f(x) = p^x(1-p)^{1-x}$.Then

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

# Estimation

**Bernoulli MLE Estimation** We are going to estimate the value of the parameter $p$ based on $n$ given data points corresponding to as IID random variables $X_1, X_2, \ldots, X_n$ such that $X_i \sim Ber(p)$.

The pdf of $X_i$ is $f(x) = p^x(1-p)^{1-x}$.Then

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

Then the log likelihood function is given by:

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

# Estimation

**Bernoulli MLE Estimation** We are going to estimate the value of the parameter $p$ based on $n$ given data points corresponding to as IID random variables $X_1, X_2, \ldots, X_n$ such that $X_i \sim Ber(p)$.

The pdf of $X_i$ is $f(x) = p^x(1-p)^{1-x}$.Then

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

Then the log likelihood function is given by:

$$\begin{aligned} LL(p) &= \sum_{i=1}^{n} \log p^{x_i}(1-p)^{1-x_i} \\ &= y \log p + (n-y)\log(1-p), y = \sum_{i=1}^{n} x_i \end{aligned}$$

# Estimation

**Bernoulli MLE Estimation** We are going to estimate the value of the parameter $p$ based on $n$ given data points corresponding to as IID random variables $X_1, X_2, \ldots, X_n$ such that $X_i \sim Ber(p)$.

The pdf of $X_i$ is $f(x) = p^x(1-p)^{1-x}$.Then

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

Then the log likelihood function is given by:

$$\begin{aligned} LL(p) &= \sum_{i=1}^{n} \log p^{x_i}(1-p)^{1-x_i} \\ &= y \log p + (n-y)\log(1-p), y = \sum_{i=1}^{n} x_i \end{aligned}$$

Then setting $\frac{\partial LL(p)}{\partial p} = 0$ implies $p = \frac{y}{n}$ i.e. $p$ equals the sample mean.

# Estimation

Normal MLE Estimation Assume that we have $n$ data points corresponding to a random sample $X_1, \ldots, X_n$ with $X_i \sim \mathcal{N}(\theta_0 = \mu, \theta_1 = \sigma^2)$.

# Estimation

Normal MLE Estimation Assume that we have $n$ data points corresponding to a random sample $X_1, \ldots, X_n$ with $X_i \sim \mathcal{N}(\theta_0 = \mu, \theta_1 = \sigma^2)$. Then

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(X_i \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\theta_1}} e^{-\frac{(x_i-\theta_0)^2}{\theta_1}}$$

# Estimation

Normal MLE Estimation Assume that we have $n$ data points corresponding to a random sample $X_1, \ldots, X_n$ with $X_i \sim \mathcal{N}(\theta_0 = \mu, \theta_1 = \sigma^2)$. Then

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(X_i \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\theta_1}} e^{-\frac{(x_i-\theta_0)^2}{\theta_1}}$$

$$LL(\boldsymbol{\theta}) = \sum_{i=1}^{n} \left[ -\log(\sqrt{2\pi\theta_1}) - \frac{1}{2\theta_1}(x_i - \theta_0)^2 \right].$$

# Estimation

Normal MLE Estimation Assume that we have $n$ data points corresponding to a random sample $X_1, \ldots, X_n$ with $X_i \sim \mathcal{N}(\theta_0 = \mu, \theta_1 = \sigma^2)$. Then

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(X_i \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\theta_1}} e^{-\frac{(x_i-\theta_0)^2}{\theta_1}}$$

$$LL(\boldsymbol{\theta}) = \sum_{i=1}^{n} \left[ -\log(\sqrt{2\pi\theta_1}) - \frac{1}{2\theta_1}(x_i - \theta_0)^2 \right].$$

Then setting the partial derivative of $LL(\boldsymbol{\theta})$ with respect to both $\theta_0$ and $\theta_1$ equal 0,

$$\widehat{\theta}_0 = \widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i, \ \widehat{\theta}_1 = \widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \widehat{\mu})^2$$

# Estimation

Method of moments (MoM) Recall that the $k$th moment of a random variable is

$$\mu_k = E(X^k)$$

# Estimation

Method of moments (MoM) Recall that the $k$th moment of a random variable is

$$\mu_k = E(X^k)$$

For a random sample $X_1, \ldots, X_n$, the $k$-th sample moment is defined as

$$M_k = \frac{1}{n} \sum_{i=1}^{n} X_i^k$$

# Estimation

Method of moments (MoM) Recall that the $k$th moment of a random variable is

$$\mu_k = E(X^k)$$

For a random sample $X_1, \ldots, X_n$, the $k$-th sample moment is defined as

$$M_k = \frac{1}{n} \sum_{i=1}^{n} X_i^k$$

Then we estimate parameters $\boldsymbol{\theta}$ by equating the first $k$ population moments (if they exist) to the first $m$ sample moments, i.e. setting

$$\mu_k = M_k$$

# Estimation

Method of moments (MoM) Recall that the $k$th moment of a random variable is

$$\mu_k = E(X^k)$$

For a random sample $X_1, \ldots, X_n$, the $k$-th sample moment is defined as

$$M_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

Then we estimate parameters $\boldsymbol{\theta}$ by equating the first $k$ population moments (if they exist) to the first $m$ sample moments, i.e. setting

$$\mu_k = M_k$$

For example, it can be verified that if the population random variable $X \sim \mathcal{N}(\mu, \sigma^2)$ then

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

for a random sample of size $n$

# Estimation

Observation The estimates are functions of the observed data points $x_1, \ldots, x_n$ which correspond to the random sample $X_1, \ldots, X_n$ respectively.

# Estimation

Observation The estimates are functions of the observed data points $x_1, \ldots, x_n$ which correspond to the random sample $X_1, \ldots, X_n$ respectively.

## Estimator

Let $X \sim f(x; \boldsymbol{\theta})$ and $X_1, \ldots, X_n$ be a random sample from the population $X$. Any statistic which can be employed to guess a population parameter $\theta$ is called an estimator of $\theta$. The numerical value of the statistic is called an estimate and the estimator is denoted by $\widehat{\theta}$.

# Estimation

Observation The estimates are functions of the observed data points $x_1, \ldots, x_n$ which correspond to the random sample $X_1, \ldots, X_n$ respectively.

## Estimator

Let $X \sim f(x; \boldsymbol{\theta})$ and $X_1, \ldots, X_n$ be a random sample from the population $X$. Any statistic which can be employed to guess a population parameter $\theta$ is called an estimator of $\theta$. The numerical value of the statistic is called an estimate and the estimator is denoted by $\widehat{\theta}$.

For example, using both MLE and MoM, the estimators for $\mu, \sigma^2$ of the normal population are:

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i \text{ and } \widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

# Estimation

## Order statistics

Let $X_1, \ldots, X_n$ be observations from a random sample of size $n$ from a distribution $f(x.)$ Let $X_{(1)}$ denote the smallest of $\{X_1, X_2, \ldots, X_n\}$, $X_{(2)}$ denote the second smallest of $\{X_1, X_2, \ldots, X_n\}$, and in general, the $r$th smallest of $\{X_1, X_2, \ldots, X_n\}$ is denoted by $X_{(r)}$. Then the random variables $X_{(1)}, \ldots, X_{(r)}$ are called the order statistics of the sample $X_1, \ldots, X_n$. In particular, $X_{(r)}$ is called the $r$th-order statistics of $X_1, X_2, \ldots, X_n$.

# Estimation

## Order statistics

Let $X_1, \ldots, X_n$ be observations from a random sample of size $n$ from a distribution $f(x.)$ Let $X_{(1)}$ denote the smallest of $\{X_1, X_2, \ldots, X_n\}$, $X_{(2)}$ denote the second smallest of $\{X_1, X_2, \ldots, X_n\}$, and in general, the $r$th smallest of $\{X_1, X_2, \ldots, X_n\}$ is denoted by $X_{(r)}$. Then the random variables $X_{(1)}, \ldots, X_{(r)}$ are called the order statistics of the sample $X_1, \ldots, X_n$. In particular, $X_{(r)}$ is called the $r$th-order statistics of $X_1, X_2, \ldots, X_n$.

Homework The estimator of $\theta$ for the uniform distribution over the interval $(0, \theta)$ are given by

$$\widehat{\theta} = X_{(n)} \text{ and } \widehat{\theta} = 2\overline{X}$$

corresponding to the MLE and Mom.

# Estimation

## Order statistics

Let $X_1, \ldots, X_n$ be observations from a random sample of size $n$ from a distribution $f(x.)$ Let $X_{(1)}$ denote the smallest of $\{X_1, X_2, \ldots, X_n\}$, $X_{(2)}$ denote the second smallest of $\{X_1, X_2, \ldots, X_n\}$, and in general, the $r$th smallest of $\{X_1, X_2, \ldots, X_n\}$ is denoted by $X_{(r)}$. Then the random variables $X_{(1)}, \ldots, X_{(r)}$ are called the order statistics of the sample $X_1, \ldots, X_n$. In particular, $X_{(r)}$ is called the $r$th-order statistics of $X_1, X_2, \ldots, X_n$.

Homework The estimator of $\theta$ for the uniform distribution over the interval $(0, \theta)$ are given by

$$\widehat{\theta} = X_{(n)} \text{ and } \widehat{\theta} = 2\overline{X}$$

corresponding to the MLE and Mom.

Question Which one is a better estimator?

# Estimation

Question How to evaluate the goodness of an estimator?

# Estimation

Question How to evaluate the goodness of an estimator?

1. Unbiasedness
2. Effieiency and relative efficiency
3. Uniform minumum variance unbiasedness
4. Sufficiency
5. Consistency

# Estimation

Question How to evaluate the goodness of an estimator?

1. Unbiasedness
2. Effieiency and relative efficiency
3. Uniform minumum variance unbiasedness
4. Sufficiency
5. Consistency

The notions of unbiasedness, efficiency and sufficiency were introduced by Ronald Fisher.

# Estimation

## Unbiasedness

An estimator $\widehat{\theta}$ of $\theta$ is called an unbiased estimator if

$$E\left(\widehat{\theta}\right) = \theta,$$

otherwise it is called a biased estimator.

# Estimation

## Unbiasedness

An estimator $\widehat{\theta}$ of $\theta$ is called an unbiased estimator if

$$E\left(\widehat{\theta}\right) = \theta,$$

otherwise it is called a biased estimator.

Meaning An estimator may not equal to the true value of the parameter for every realization of the random sample $X_1, X_2, \ldots, X_n$ but if it unbiased then on an average it will be equal to the true value

# Estimation

## Examples

- The sample mean $\overline{X} = \widehat{\mu}$ (as obtained by MoM and MLE) is an unbiased estimator for a normal population

# Estimation

## Examples

- The sample mean $\overline{X} = \widehat{\mu}$ (as obtained by MoM and MLE) is an unbiased estimator for a normal population
- The MLE estimator $\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})$ for normal population is a biased estimator

# Estimation

## Examples

- The sample mean $\overline{X} = \widehat{\mu}$ (as obtained by MoM and MLE) is an unbiased estimator for a normal population
- The MLE estimator $\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})$ for normal population is a biased estimator
- (Homework) The sample variance $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ is an unbiased estimator for $\sigma^2$ in the normal population

# Estimation

## Examples

- The sample mean $\overline{X} = \widehat{\mu}$ (as obtained by MoM and MLE) is an unbiased estimator for a normal population
- The MLE estimator $\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})$ for normal population is a biased estimator
- (Homework) The sample variance $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ is an unbiased estimator for $\sigma^2$ in the normal population

## Limitations of unbiased estimators

- an unbiased estimator may not exist

# Estimation

## Examples

- The sample mean $\overline{X} = \widehat{\mu}$ (as obtained by MoM and MLE) is an unbiased estimator for a normal population
- The MLE estimator $\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})$ for normal population is a biased estimator
- (Homework) The sample variance $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ is an unbiased estimator for $\sigma^2$ in the normal population

## Limitations of unbiased estimators

- an unbiased estimator may not exist
- unbiasedness is not invariant under functional transformation i.e. if $\widehat{\theta}$ is an unbiased estimator of $\theta$ and $g$ is a function, then $g(\widehat{\theta})$ may not be an unbiased estimator of $g(\theta)$

# Estimation

Question Between two unbiased estimators, which one to prefer?

# Estimation

Question Between two unbiased estimators, which one to prefer?

Example (Homework) Let $X_1, \ldots, X_n$ be a sample from a distribution with unknown mean $-\infty < \mu < \infty$, and unknown variance $\sigma^2 > 0$. Then the statistics

$$\overline{X} \text{ and } Y = \frac{1}{\frac{n(n+1)}{2}}(X_1 + 2X_2 + \ldots + nX_n)$$

are both unbiased estimators of $\mu$. Besides, $Var(\overline{X}) < Var(Y)$.

# Estimation

Question Between two unbiased estimators, which one to prefer?

Example (Homework) Let $X_1, \ldots, X_n$ be a sample from a distribution with unknown mean $-\infty < \mu < \infty$, and unknown variance $\sigma^2 > 0$. Then the statistics

$$\overline{X} \text{ and } Y = \frac{1}{\frac{n(n+1)}{2}}(X_1 + 2X_2 + \ldots + nX_n)$$

are both unbiased estimators of $\mu$. Besides, $Var(\overline{X}) < Var(Y)$.

# Estimation

Let $\widehat{\theta}_1$ and $\widehat{\theta}_2$ be two unbiased estimators of $\theta$. Then the estimator $\widehat{\theta}_1$ is said to be more efficient than $\widehat{\theta}_2$ if

$$Var\left(\widehat{\theta}_1\right) < Var\left(\widehat{\theta}_2\right).$$

The ratio $\eta$ given by

$$\eta\left(\widehat{\theta}_1, \widehat{\theta}_2\right) = \frac{Var\left(\widehat{\theta}_2\right)}{Var\left(\widehat{\theta}_1\right)}$$

is called the relative efficiency of $\widehat{\theta}_1$ with respect to $\widehat{\theta}_2$.

# Estimation

Let $\widehat{\theta}_1$ and $\widehat{\theta}_2$ be two unbiased estimators of $\theta$. Then the estimator $\widehat{\theta}_1$ is said to be more efficient than $\widehat{\theta}_2$ if

$$Var\left(\widehat{\theta}_1\right) < Var\left(\widehat{\theta}_2\right).$$

The ratio $\eta$ given by

$$\eta\left(\widehat{\theta}_1, \widehat{\theta}_2\right) = \frac{Var\left(\widehat{\theta}_2\right)}{Var\left(\widehat{\theta}_1\right)}$$

is called the relative efficiency of $\widehat{\theta}_1$ with respect to $\widehat{\theta}_2$.

Meaning If the unbiased estimator has smaller variance then it is more likely that the estimator will be close to the true parameter

# Estimation

From above, we know that sample mean is always an unbiased estimator for the population mean irrespective of the distribution of the population.

# Estimation

From above, we know that sample mean is always an unbiased estimator for the population mean irrespective of the distribution of the population.

Example (Homework) Let $X_1, X_2, X_3$ be a random sample of size 3 from a population with density

$$f(x;\lambda) = \begin{cases} \frac{\lambda^x e^{-\lambda}}{x!} \text{ if } x = 0, 1, 2, \ldots \\ 0, \text{ otherwise} \end{cases}$$

Then $\overline{X}$, $\widehat{\lambda}_1 = \frac{1}{4}(X_1 + 2X_2 + X_3)$ and $\widehat{\lambda}_2 = \frac{1}{9}(4X_1 + 3X_2 + 2X_3)$ are unbiased estimators of $\lambda$, and

$$Var(\overline{X}) < Var\left(\widehat{\lambda}_2\right) < Var\left(\widehat{\lambda}_1\right).$$

# Estimation

Question How to find an unbiased estimator which is smallest variance among all unbiased estimators of a given parameter?

# Estimation

Question How to find an unbiased estimator which is smallest variance among all unbiased estimators of a given parameter?

An unbiased estimator $\widehat{\theta}$ of $\theta$ id said to be a uniform minimum variance unbiased estimator of $\theta$ if

$$Var\left(\widehat{\theta}\right) \leq Var\left(\widehat{T}\right)$$

for any unbiased estimator $\widehat{T}$ of $\theta$

# Estimation

Question How to find an unbiased estimator which is smallest variance among all unbiased estimators of a given parameter?

An unbiased estimator $\widehat{\theta}$ of $\theta$ id said to be a uniform minimum variance unbiased estimator of $\theta$ if

$$Var\left(\widehat{\theta}\right) \leq Var\left(\widehat{T}\right)$$

for any unbiased estimator $\widehat{T}$ of $\theta$

However, for an unbiased estimator $\widehat{\theta}$,

$$Var(\widehat{\theta}) = E\left[(\widehat{\theta}) - E(\widehat{\theta})^2\right] = E\left[(\widehat{\theta} - \theta)^2\right]$$

# Estimation

An interesting result Given two unbiased estimators $\widehat{\theta}_1$ and $\widehat{\theta}_2$ the set

$$\mathcal{E} = \left\{\widehat{\theta} : \widehat{\theta} = c\widehat{\theta}_1 + (1-c)\widehat{\theta}_2, c \in \mathbb{R}\right\}$$

forms an uncountable set of unbiased estimators of $\theta$. However, if the variances of the estimators $\widehat{\theta}_1$ and $\widehat{\theta}_2$ are not known then it is very difficult to find the minimum variance estimator in $\mathcal{E}$.

# Estimation

An interesting result Given two unbiased estimators $\widehat{\theta}_1$ and $\widehat{\theta}_2$ the set

$$\mathcal{E} = \left\{\widehat{\theta} : \widehat{\theta} = c\widehat{\theta}_1 + (1-c)\widehat{\theta}_2, c \in \mathbb{R}\right\}$$

forms an uncountable set of unbiased estimators of $\theta$. However, if the variances of the estimators $\widehat{\theta}_1$ and $\widehat{\theta}_2$ are not known then it is very difficult to find the minimum variance estimator in $\mathcal{E}$.

Note that it is only one class.

# Estimation

One way to find a uniform variance estimator for a parameter is to employ Cramer-Rao lower bound as follows:

# Estimation

One way to find a uniform variance estimator for a parameter is to employ Cramer-Rao lower bound as follows:

Let $X_1, \ldots, X_n$ be a random sample of size $n$ from a population $X$ with density $f(x;\theta)$, where $\theta$ is a scalar. Let $\widehat{\theta}$ be any unbiased estimator of $\theta$. Suppose the likelihood function $L(\theta)$ is a differentiable function of $\theta$ and satisfies

$$\frac{d}{d\theta}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} h(x_1,\ldots,x_n)L(\theta)dx_1\ldots dx_n$$
$$= \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} h(x_1,\ldots,x_n)\frac{d}{d\theta}L(\theta)dx_1\ldots dx_n$$

for any $h(x_1,\ldots,x_n)$ with $E(h(X_1,\ldots,X_n)) < \infty$. Then

$$Var(\widehat{\theta}) \geq \frac{1}{E\left[\left(\frac{\partial \ln L(\theta)}{\partial\theta}\right)^2\right]}$$

# Estimation

Further, if the estimator $\widehat{\theta}$ is minimum variance is addition to being unbiased then the equality of the above theorem holds. However, the converse need not be true.

# Estimation

Further, if the estimator $\widehat{\theta}$ is minimum variance is addition to being unbiased then the equality of the above theorem holds. However, the converse need not be true.

An unbiased estimator $\widehat{\theta}$ is called an efficient estimator if it satisfies Cramer-Rao lower bound, that is

$$Var(\widehat{\theta}) = \frac{1}{E\left[\left(\frac{\partial \ln L(\theta)}{\partial \theta}\right)^2\right]}$$

# Estimation

Further, if the estimator $\widehat{\theta}$ is minimum variance is addition to being unbiased then the equality of the above theorem holds. However, the converse need not be true.

An unbiased estimator $\widehat{\theta}$ is called an efficient estimator if it satisfies Cramer-Rao lower bound, that is

$$Var(\widehat{\theta}) = \frac{1}{E\left[\left(\frac{\partial \ln L(\theta)}{\partial \theta}\right)^2\right]}$$

An interesting result Let $X_1, \ldots, X_n$ be a random sample of size $n$ from a normal population with mean $\mu$ and $\sigma^2 > 0$. Then $S^2$, the sample variance estimator of $\sigma^2$ but it cannot attain the Cramer-Rao lower bound.

# Estimation

Limitations of Cramer-Rao lower bound:

1. Not every density function $f(x;\theta)$ satisfies the assumptions of Cramer-Rao theorem

# Estimation

Limitations of Cramer-Rao lower bound:

1. Not every density function $f(x;\theta)$ satisfies the assumptions of Cramer-Rao theorem
2. Not every allowable estimator attains the Cramer-Rao lower bound

# Estimation

Limitations of Cramer-Rao lower bound:

1. Not every density function $f(x;\theta)$ satisfies the assumptions of Cramer-Rao theorem
2. Not every allowable estimator attains the Cramer-Rao lower bound
3. Hence in any one of these situations, one does not knw an estimator is a uniform minimum variance unbiased estimator or not!

# Estimation

Limitations of Cramer-Rao lower bound:

1. Not every density function $f(x;\theta)$ satisfies the assumptions of Cramer-Rao theorem
2. Not every allowable estimator attains the Cramer-Rao lower bound
3. Hence in any one of these situations, one does not knw an estimator is a uniform minimum variance unbiased estimator or not!

Observations If the distribution of an estimator $\widehat{\theta}$ is hard to find even if the distribution of the population is known then there is no way to check whether $\widehat{\theta}$ is unbiased or biased. Thus we need other criteria to measure the quality of an estimator.

# Estimation

Let $X \sim f(x;\theta)$ be a population and let $X_1, X_2, \ldots, X_n$ be a random sample of size $n$ from $X$. Then an estimator $\widehat{\theta}$ is said to be sufficient estimator of $\theta$ if the conditional distribution of the sample given the estimator $\widehat{X}$ does not depend on the parameter $\theta$.

# Estimation

Let $X \sim f(x;\theta)$ be a population and let $X_1, X_2, \ldots, X_n$ be a random sample of size $n$ from $X$. Then an estimator $\widehat{\theta}$ is said to be sufficient estimator of $\theta$ if the conditional distribution of the sample given the estimator $\widehat{X}$ does not depend on the parameter $\theta$.

Example If $X_1, \ldots, X_n$ is a random sample of $X \sim Ber(\theta)$ then $Y = \sum_{i=1}^{n} X_i$ is a sufficient statistic of $\theta$.

# Estimation

Let $X \sim f(x;\theta)$ be a population and let $X_1, X_2, \ldots, X_n$ be a random sample of size $n$ from $X$. Then an estimator $\widehat{\theta}$ is said to be sufficient estimator of $\theta$ if the conditional distribution of the sample given the estimator $\widehat{X}$ does not depend on the parameter $\theta$.

Example If $X_1, \ldots, X_n$ is a random sample of $X \sim Ber(\theta)$ then $Y = \sum_{i=1}^{n} X_i$ is a sufficient statistic of $\theta$.

Note However the density of the estimator is not easy always. Thus verifying the sufficiency of an estimator is NOT easy!!

# Estimation

Factorization theorem of Fisher and Neyman Let $X_1, \ldots, X_n$ be a random sample with pdf $f(x_1, \ldots, x_n; \theta)$ which depends on the population parameter $\theta$. Then an estimator $\widehat{\theta}$ is sufficient for $\theta$ if

$$f(x_1, \ldots, x_n; \theta) = \phi(\widehat{\theta}, \theta) h(x_1, \ldots, x_n)$$

where $\phi$ depends on $x_1, \ldots, x_n$ only through $\widehat{\theta}$ and $h(x_1, \ldots, x_n)$ does not depend on $\theta$.

# Estimation

Factorization theorem of Fisher and Neyman Let $X_1, \ldots, X_n$ be a random sample with pdf $f(x_1, \ldots, x_n; \theta)$ which depends on the population parameter $\theta$. Then an estimator $\widehat{\theta}$ is sufficient for $\theta$ if

$$f(x_1, \ldots, x_n; \theta) = \phi(\widehat{\theta}, \theta) h(x_1, \ldots, x_n)$$

where $\phi$ depends on $x_1, \ldots, x_n$ only through $\widehat{\theta}$ and $h(x_1, \ldots, x_n)$ does not depend on $\theta$.

Example Let $X_1, \ldots, X_n$ be a random sample from normal distribution with mean $\mu$. Then $\overline{X}$ is a sufficient estimator.

# Estimation

## Two interesting results

1. If there exists a sufficient estimator for a parameter $\theta$ and if the ML estimator that parameter is unique then the ML estimator is a function of the sufficient estimator

# Estimation

## Two interesting results

1. If there exists a sufficient estimator for a parameter $\theta$ and if the ML estimator that parameter is unique then the ML estimator is a function of the sufficient estimator
2. If there exists a sufficient estimator of $\theta$ and if the uniform minimum variance unbiased estimator of that parameter is unique, then the uniform minimum variance unbiased estimator is a function of the sufficient estimator.

# Estimation

Let $X_1, \ldots, X_n$ be a random sample of population $X$ with density $f(x;\theta)$. Obviously, an estimator $\widehat{\theta}$ of $\theta$ is based on the sample of size $n$, and hence it depends on $n$. Thus we denote $\widehat{\theta}$ as $\widehat{\theta}_n$.

# Estimation

Let $X_1, \ldots, X_n$ be a random sample of population $X$ with density $f(x;\theta)$. Obviously, an estimator $\widehat{\theta}$ of $\theta$ is based on the sample of size $n$, and hence it depends on $n$. Thus we denote $\widehat{\theta}$ as $\widehat{\theta}_n$.

Let $X_1, \ldots, X_n$ be a random sample from a population $X$ with dendity $f(x;\theta)$. Then a sequence of estimators $\widehat{\theta}_n$ of $\theta$ is said to be consistent for $\theta$ if the sequence $\{\widehat{\theta}_n\}$ converges in probability to $\theta$ i.e.

$$\lim_{n\to\infty} P\left(\left|\widehat{\theta}_n - \theta\right| \geq \epsilon\right) = 0.$$

# Estimation

Let $X_1, \ldots, X_n$ be a random sample of population $X$ with density $f(x;\theta)$. Obviously, an estimator $\widehat{\theta}$ of $\theta$ is based on the sample of size $n$, and hence it depends on $n$. Thus we denote $\widehat{\theta}$ as $\widehat{\theta}_n$.

Let $X_1, \ldots, X_n$ be a random sample from a population $X$ with dendity $f(x;\theta)$. Then a sequence of estimators $\widehat{\theta}_n$ of $\theta$ is said to be consistent for $\theta$ if the sequence $\{\widehat{\theta}_n\}$ converges in probability to $\theta$ i.e.

$$\lim_{n\to\infty} P\left(\left|\widehat{\theta}_n - \theta\right| \geq \epsilon\right) = 0.$$

Meaning Consistency is a property of a large sample for an estimator.

# Estimation

## Connection to mean squared error

Let $X_1, \ldots, X_n$ be a random sample from $X$ with density $f(x;\theta)$ and $\{\widehat{\theta}_n\}$ be a sequence of estimators of $\theta$ based on the sample. If the variance of $\widehat{\theta}_n$ exists for each $n$ and finite and

$$\lim_{n\to\infty} E\left(\left(\widehat{\theta}_n - \theta\right)^2\right) = 0$$

then for any $\epsilon > 0$,

$$\lim_{n\to\infty} P\left(\left|\widehat{\theta}_n - \theta\right| \geq \epsilon\right) = 0.$$

# Estimation

## Connection to mean squared error

Let $X_1, \ldots, X_n$ be a random sample from $X$ with density $f(x;\theta)$ and $\{\widehat{\theta}_n\}$ be a sequence of estimators of $\theta$ based on the sample. If the variance of $\widehat{\theta}_n$ exists for each $n$ and finite and

$$\lim_{n\to\infty} E\left(\left(\widehat{\theta}_n - \theta\right)^2\right) = 0$$

then for any $\epsilon > 0$,

$$\lim_{n\to\infty} P\left(\left|\widehat{\theta}_n - \theta\right| \geq \epsilon\right) = 0.$$

Example Let $X_1, \ldots, X_n$ be a random sample from a normal population $X$ with variance $\sigma^2$. Then

$$\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

is a consistent estimator of $\sigma^2$.

# Estimation

As we discussed, it is obvious (why?) that the point estimate of a parameter $\theta$ is rarely equal to the true value of $\theta$. Can we find a list of probable values for $\theta$ with certain degree of confidence?

# Estimation

As we discussed, it is obvious (why?) that the point estimate of a parameter $\theta$ is rarely equal to the true value of $\theta$. Can we find a list of probable values for $\theta$ with certain degree of confidence?

Interval estimation problem Given a random sample $X_1, \ldots, X_n$ and a probability value $1 - p$, find a pair of statistics $L = L(X_1, \ldots, X_n)$ and $U = U(X_1, \ldots, X_n)$ with $L \leq U$ such that

$$P(L \leq \theta \leq U) = 1 - p?$$

# Estimation

Let $X_1, \ldots, X_n$ be a random sample from $f(x;\theta)$. The interval estimator of $\theta$ is a pair of statistics $L = L(X_1, \ldots, X_n)$ and $U = U(X_1, \ldots, X_n)$ with $L \leq U$ such that if $x_1, \ldots, x_n$ is a sample data then
$\theta \in [L(x_1, \ldots, x_n),\ U(x_1, \ldots, x_n)]$

# Estimation

Let $X_1, \ldots, X_n$ be a random sample from $f(x;\theta)$. The interval estimator of $\theta$ is a pair of statistics $L = L(X_1, \ldots, X_n)$ and $U = U(X_1, \ldots, X_n)$ with $L \leq U$ such that if $x_1, \ldots, x_n$ is a sample data then $\theta \in [L(x_1, \ldots, x_n),\ U(x_1, \ldots, x_n)]$

The interval estimator is called a $100(1-p)\%$ confidence interval for $\theta$ if

$$P(L \leq \theta \leq U) = 1 - p.$$

The number $(1-p)$ is called the confidence coefficient or degree of confidence.

# Estimation

We will discuss two methods for finding interval estimator:

# Estimation

We will discuss two methods for finding interval estimator:

1. Pivotal quantity method
2. Maximum likelihood estimator (MLE)

# Estimation

We will discuss two methods for finding interval estimator:

1. Pivotal quantity method
2. Maximum likelihood estimator (MLE)

## Pivotal quantity method

Let $X_1, \ldots, X_n$ be a random sample from $X$ with density $f(x;\theta)$. A *pivotal quantity* is a function of $X_1, \ldots, X_n, \theta$ whose probability distribution is independent of $\theta$.

# Estimation

We will discuss two methods for finding interval estimator:

1. Pivotal quantity method
2. Maximum likelihood estimator (MLE)

## Pivotal quantity method

Let $X_1, \ldots, X_n$ be a random sample from $X$ with density $f(x;\theta)$. A *pivotal quantity* is a function of $X_1, \ldots, X_n, \theta$ whose probability distribution is independent of $\theta$.

Example Let $X_1, \ldots, X_n$ is a random sample from $\mathcal{N}(\mu, \sigma^2)$. Then $\overline{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$. Then

$$Q(X_1, \ldots, X_n, \theta) = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0, 1)$$

# Estimation

There is no general rule for finding the pivotal quantity. However there is a family of pdfs $f(x;\theta)$, called location-scale family then there is a systematic way.

# Estimation

There is no general rule for finding the pivotal quantity. However there is a family of pdfs $f(x;\theta)$, called location-scale family then there is a systematic way.

If $Q = Q(X_1, \ldots, X_n, \theta)$ is a pivot then $100(1-p)\%$ confidence interval for $\theta$ can be found by first finding two values $a, b$ such that

$$P(a \leq Q \leq b) = 1 - p$$

and then convert it the inequality $a \leq Q \leq b$ into the form $L \leq \theta \leq U$.

# Estimation

Example Let $X \sim \mathcal{N}(\mu, \sigma^2)$ with $\sigma^2$ known. Then for some $p$, we have

$$\begin{aligned} 1-2p &= P\left(-z_p \leq \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq z_p\right) \\ &= P\left(\mu - z_p\frac{\sigma}{\sqrt{n}} \leq \overline{X} \leq \mu + z_p\frac{\sigma}{\sqrt{n}}\right) \\ &= P\left(\overline{X} - z_p\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + z_p\frac{\sigma}{\sqrt{n}}\right) \end{aligned}$$

# Estimation

Example Let $X \sim \mathcal{N}(\mu, \sigma^2)$ with $\sigma^2$ known. Then for some $p$, we have

$$\begin{aligned} 1-2p &= P\left(-z_p \leq \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq z_p\right) \\ &= P\left(\mu - z_p\frac{\sigma}{\sqrt{n}} \leq \overline{X} \leq \mu + z_p\frac{\sigma}{\sqrt{n}}\right) \\ &= P\left(\overline{X} - z_p\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + z_p\frac{\sigma}{\sqrt{n}}\right) \end{aligned}$$

Then the $100(1-2p)\%$ confidence interval for $\mu$ is

$$\left[\overline{X} - z_p\frac{\sigma}{\sqrt{n}}, \overline{X} + z_p\frac{\sigma}{\sqrt{n}}\right],$$

where $z_p$ denotes the $100(1-p)$-percentile of a standard normal variable $Z$ i.e $P(Z \leq z_p) = 1-p$, where $p \leq 0.5$

# Estimation

Confidence interval using MLE Let $X_1, \ldots, X_n$ be a random sample from $X$ with density $f(x; \theta)$.

# Estimation

Confidence interval using MLE Let $X_1, \ldots, X_n$ be a random sample from $X$ with density $f(x;\theta)$. Let $\widehat{\theta}$ be the MLE of $\theta$.

# Estimation

Confidence interval using MLE Let $X_1, \ldots, X_n$ be a random sample from $X$ with density $f(x;\theta)$. Let $\widehat{\theta}$ be the MLE of $\theta$. If the sample size is large

$$\frac{\widehat{\theta} - E\left(\widehat{\theta}\right)}{\sqrt{Var\left(\widehat{\theta}\right)}} \sim \mathcal{N}(0,1) \text{ as } n \to \infty.$$

# Estimation

Confidence interval using MLE Let $X_1, \ldots, X_n$ be a random sample from $X$ with density $f(x;\theta)$. Let $\widehat{\theta}$ be the MLE of $\theta$. If the sample size is large

$$\frac{\widehat{\theta} - E\left(\widehat{\theta}\right)}{\sqrt{Var\left(\widehat{\theta}\right)}} \sim \mathcal{N}(0,1) \text{ as } n \to \infty.$$

Since for large $n$, the MLE of $\theta$ is unbiased,

$$\frac{\widehat{\theta} - \theta}{\sqrt{Var\left(\widehat{\theta}\right)}} \sim \mathcal{N}(0,1) \text{ as } n \to \infty.$$

# Estimation

Then setting $Q = \frac{\widehat{\theta}-\theta}{\sqrt{Var(\widehat{\theta})}}$ as a pivotal quantity, we have

$$\begin{aligned} 1-p &= P(-z_{\frac{p}{2}} \leq Q \leq z_{\frac{p}{2}}) \\ &= P\left(-z_{\frac{p}{2}} \leq \frac{\widehat{\theta}-\theta}{\sqrt{Var\left(\widehat{\theta}\right)}} \leq z_{\frac{p}{2}}\right) \end{aligned}$$

with $(1-p)100\%$ confidence interval.

# Estimation

Then setting $Q = \frac{\widehat{\theta}-\theta}{\sqrt{Var(\widehat{\theta})}}$ as a pivotal quantity, we have

$$\begin{aligned} 1-p &= P(-z_{\frac{p}{2}} \leq Q \leq z_{\frac{p}{2}}) \\ &= P\left(-z_{\frac{p}{2}} \leq \frac{\widehat{\theta}-\theta}{\sqrt{Var\left(\widehat{\theta}\right)}} \leq z_{\frac{p}{2}}\right) \end{aligned}$$

with $(1-p)100\%$ confidence interval. If $Var\left(\widehat{\theta}\right)$ is independent of $\theta$ then with $100(1-p)\%$ confidence interval for $\theta$ is

$$\left[\widehat{\theta} - z_{\frac{p}{2}}\sqrt{Var\left(\widehat{\theta}\right)}, \widehat{\theta} + z_{\frac{p}{2}}\sqrt{Var\left(\widehat{\theta}\right)}\right]$$