Statistical Inference (MA862)

Lecture Slides

Topic 2: Point Estimation

Statistical Inference

- In a typical statistical problem, our aim is to find information regarding numerical characteristic(s) of a collection of items/persons/products. This collection is called population.
- Suppose that we want to know the average height of Indian citizens.
  - Measure heights of all citizens
  - Find the average.
- However, it is a very costly (in terms of money and time) procedure.

Sample

- One approach to address these issues is to take a subset of the population based on which we try to find out the value of the numerical characteristic.
- Obviously, it will not be exact, and hence, it is an estimate.
- This subset is called a sample.
- The sample must be chosen such that it is a good representative of the population.
- There are different ways of selecting sample from a population.
- We will consider one such sample which is called *random sample*.

Modelling a Statistical Problem

- Different elements of a population may have different values of the numerical characteristic under study.
- Therefore, we will model it with a random variable and the uncertainty using a probability distribution.
- Let X be a random variable (either discrete or continuous random variable), which denotes the numerical characteristic under consideration.
- Our job is to find the probability distribution of X.
- Note that once the probability distribution is determined, the numerical summary (for example, mean, variance, median, etc.) of the distribution can be found.

Parametric and Non-parametric Inference

- There are two possibilities:
  - X has a CDF F with known functional form except perhaps some parameters. Here our aim is to (educated) guess value of the parameters. For example, in some case we may have $X \sim N(\mu, \sigma^2)$, where the functional form of the PDF is known, but the parameters μ and/or σ^2 may be unknown. In this case, we need to find value of the unknown parameters based on a sample. This is known as parametric inference.
  - X has a CDF F who's functional form is unknown. This is known as non-parametric inference.

Random Sample

Definition 1: The random variables $X_1, X_2, \ldots, X_n$ is said to be a random sample (RS) of size n from the population F if $X_1, X_2, \ldots, X_n$ are *i.i.d.* random variables with marginal CDF F. If F has a PMF/PDF f, we will write that $X_1, \ldots, X_n$ is a RS from the PMF/PDF f.

- The JCDF of a RS $X_1, \ldots, X_n$ from CDF F is

$$F(x_1, \ldots, x_n) = \prod_{i=1}^{n} F(x_i).$$

- The JPMF/JPDF of a RS $X_1, \ldots, X_n$ from PMF/PDF f is

$$f(x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i).$$

Random Sample

- In the standard framework of parametric inference, we start with a data, say $(x_1, x_2, \ldots, x_n)$. Each x_i is an observation on the numerical characteristic under study.
- There are n observations and n is fixed, pre-assigned, and known positive integer.
- Our job is to identify (based on a data) the CDF (or equivalently PMF/PDF) of the RV X, which denote the numerical characteristic in the population.

Random Sample

- In practice, we have a data.
- How to model a data using RS?
- Notice that the first observation in the sample can be one of the member of the population.
- Thus, a particular observation is one of the realizations from the whole population.
- Therefore, it can be seen as a realization of a random variable X.
- Let X_i denote the ith observation for $i = 1, 2, \ldots, n$, where n is the sample size.
- Then, a meaningful assumption is that each X_i has same CDF F, as X_i is a copy of X.
- Now, if we can ensure that the observation are taken such a way that the value of one does not effect the others, then we can assume that $X_1, X_2, \ldots, X_n$ are independent.

Parametric Inference

- The functional form of the CDF/PMF/PDF of RV X is known.
- However, the CDF/PMF/PDF involves unknown but fixed real or vector valued parameter $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_m)$.
- If the value of $\boldsymbol{\theta}$ is known, the stochastic properties of the numerical characteristic is completely known.
- Therefore, our aim is to find the value of $\boldsymbol{\theta}$ or a function of $\boldsymbol{\theta}$.
- We assume that the possible values of $\boldsymbol{\theta}$ belong to a set Θ, which is called parametric space.
- $\boldsymbol{\theta}$ is a subset of $\mathbb{R}^n$.
- Here, $\boldsymbol{\theta}$ is an indexing or a labelling parameter. We say that $\boldsymbol{\theta}$ is an indexing parameter or a labelling parameter if the CDF/PMF/PDF is uniquely specified by $\boldsymbol{\theta}$, i.e., $F(x, \boldsymbol{\theta}_1) = F(x, \boldsymbol{\theta}_2)$ for all $x \in \mathbb{R}$ implies $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$, where $F(\cdot, \boldsymbol{\theta})$ is the CDF of X.

Some Examples

Example 3:

- Suppose we want to find the probability of germination of seeds produced by a particular brand.
  - 100 seeds of a brand were planted one in each pot.
  - Let X_i equals one or zero according as the seed in the ith pot germinates or not.
  - The data consists of $(x_1, x_2, \ldots, x_{100})$, where each x_i is either one or zero.
  - The data is regarded as a realization of $(X_1, X_2, \ldots, X_{100})$, where the RVs are *i.i.d.* with $P(X_i = 1) = \theta = 1 - P(X_i = 0)$.
  - θ is the probability that a seed germinates.
  - The natural parametric space is $\Theta = [0, 1]$.
  - θ is an indexing parameter.

Some Examples

Example 4:

- Consider determination of gravitational constant g.
  - A standard way to estimate g is to use the pendulum experiment and use the formula

    $$g = \frac{2\pi^2 l}{T^2},$$

    where l is the length of the pendulum and T is the time required for a fixed number of oscillations.
  - A variation is observed in the calculated values of g.
  - Let the repeated experiments are performed and the calculated values of g are $X_1, X_2, \ldots, X_n$.
  - Use the model $X_i = g + \epsilon_i$, where ϵ_i is the random error.
  - Assume $\epsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$.
  - Then $X_i \overset{i.i.d.}{\sim} N(g, \sigma^2)$, and the parameter is $\boldsymbol{\theta} = \left(g, \sigma^2\right)$ with parametric space $\Theta = \mathbb{R} \times \mathbb{R}^+$.
  - $\boldsymbol{\theta}$ is an indexing parameter.

Some Examples

Example 5:

- Interested in estimating the average height of a large community of people.
  - Assume that $N(\mu, \sigma^2)$ is a plausible distribution.
  - As the average of heights of persons is always a positive real number, it is realistic to assume that $\mu > 0$.
  - Hence, a better choice of Θ is $\mathbb{R}^+ \times \mathbb{R}^+$.
  - Thus, we may need to choose the parametric space based on the background of the problem.

Some Examples

Example 6:

- Consider a series system with two components. A series system works if all its components work.
- Z: lifetimes of the first component.
- Y: lifetimes of the second component.
- $Z \sim Exp(\theta)$ and $Y \sim Exp(\lambda)$ (rates θ and λ)
- Y and Z are independent RVs.
- Z and Y are not observed.
- We observe $X = \min\{Z, Y\}$.
- $X \sim Exp(\theta + \lambda)$.
- $\alpha = \theta + \lambda$ is an indexing parameter.
- However, (θ, λ) is not an indexing parameter.

Exams and Grading Policy

Exam	Weight	Date
Project-I (Group of max. 5)	10%	Will be declared
Quiz-I	10%	Feb 02, 2024
Mid-semester	25%	Feb 26, 2024
Project-II (Group of max. 5)	10%	Will be declared
Quiz-II	10%	Apr 05, 2024
End-semester	35%	May 01, 2024

- Below 25% implies a F grade.

Statistic

Definition 2: Let $X_1, \ldots, X_n$ be a RS. Let $T(x_1, \ldots, x_n)$ be a real-valued function having domain that includes the sample space, χ^n, of $X_1, X_2, \ldots, X_n$. Then, the RV $\boldsymbol{Y} = T(X_1, \ldots, X_n)$ is called a statistic if it is not a function of unknown parameters.

Definition 3: In the context of estimation, a statistic is called a point estimator (or simply estimator). A realization of a point estimator is called an estimate.

Example 7: Let $X_1, \ldots, X_n$ be a RS from a $N(\mu, \sigma^2)$ distribution, where $\mu \in \mathbb{R}$ and $\sigma > 0$ are both unknown. Then $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, $S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$ are examples of statistics. However, $\frac{\overline{X}-\mu}{\sigma}$ is not a statistic. Note that $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$.

Finding Point Estimator

- There are several methods to find an estimator.
- We will mainly consider three of them:
    - Method of moment estimator
    - Maximum likelihood estimator
    - Least square estimator

Sufficient Statistics

Definition 4: A statistic $\boldsymbol{T} = \boldsymbol{T}(\boldsymbol{X})$ is called a sufficient statistic for unknown parameter $\boldsymbol{\theta}$ if the conditional distribution of $\boldsymbol{X}$ given $\boldsymbol{T} = \boldsymbol{t}$ does not include $\boldsymbol{\theta}$ for all $\boldsymbol{t}$ in the support of $\boldsymbol{T}$.

Example 8: $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} Bernoulli(p)$, $p \in (0, 1)$. Then $T = \sum_{i=1}^{n} X_i$ is sufficient statistic for θ.

Neyman-Fisher Factorization Theorem

Theorem 3: Let $X_1, \ldots, X_n$ be RS with JPMF/JPDF $f_{\boldsymbol{X}}(\boldsymbol{x}, \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$. Then $\boldsymbol{T} = \boldsymbol{T}(X_1, \ldots, X_n)$ is sufficient for $\boldsymbol{\theta}$ if and only if

$$f_{\boldsymbol{X}}(\boldsymbol{x}, \boldsymbol{\theta}) = h(\boldsymbol{x}) g_{\boldsymbol{\theta}}\left(\boldsymbol{T}(\boldsymbol{x})\right),$$

where $h(\boldsymbol{x})$ does not involve $\boldsymbol{\theta}$, $g_{\boldsymbol{\theta}}(\cdot)$ depends on $\boldsymbol{\theta}$ and $\boldsymbol{x}$ only through $\boldsymbol{T}(\boldsymbol{x})$.

Examples

Example 9: Let $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} P(\lambda)$, $\lambda > 0$. Then $\overline{X}$ is a sufficient for λ.

Example 10: Let $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$, $\mu \in \mathbb{R}$ and $\sigma > 0$. A sufficient statistic for (μ, σ^2) is $(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$.

Example 11: Let $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} U(0, \theta)$, $\theta > 0$. Then $X_{(n)} = \max\{X_1, X_2, \ldots, X_n\}$ is a sufficient for θ.

Example 12: Let $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} U(\theta - 1/2, \theta + 1/2)$, $\theta \in \mathbb{R}$. Then, $\boldsymbol{T} = \left(X_{(1)}, X_{(n)}\right)$ is a sufficient statistic for θ, where $X_{(1)} = \min\{X_1, X_2, \ldots, X_n\}$.

Example 13: Let $X_1, X_2 \overset{i.i.d.}{\sim} N(\mu, 1)$. Is $T = X_1 + 2X_2$ a sufficient statistics for μ?

Remarks

- Note that we will be able to use the definition of sufficient statistic if we can guess one. However the theorem gives necessary and sufficient conditions, which can be used to find a sufficient statistic.
- Note that the RS is always sufficient for unknown parameters. However, most of the cases we will not talk about this trivial sufficient statistic, as it does not provide any dimension reduction.

Remarks

- If $\boldsymbol{T}$ is sufficient for $\boldsymbol{\theta}$, then for any one-to-one function of $\boldsymbol{T}$ is also sufficient for $\boldsymbol{\theta}$. (Can be proved easily using Factorization theorem.) For example $\left(\overline{X}, S^2\right)$ is sufficient for parameters of $N(\mu, \sigma^2)$, where $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$.
- Any function of sufficient statistic is not sufficient. (If so, then any statistic will be sufficient.)
- One-dimensional parameter may have multidimensional sufficient statistic. (Consider the last example.)
- $\boldsymbol{T}$ and $\boldsymbol{\theta}$ are of same dimension and $\boldsymbol{T}$ is sufficient for $\boldsymbol{\theta}$ do not imply that the jth component of $\boldsymbol{T}$ is sufficient for the jth component of $\boldsymbol{\theta}$. It only tells that $\boldsymbol{T}$ is jointly sufficient for $\boldsymbol{\theta}$.

Information

- X: a RV with PMF or PDF $f(\cdot, \theta)$, which depends on a real valued parameter $\theta \in \Theta$.
- The variation in the PMF or PDF $f(x, \theta)$ with respect to $\theta \in \Theta$ for fixed value of x provides us information about θ.
- For example, suppose that $X \sim Bin(10, \theta)$.

θ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
$f(2, \theta)$	0.19	0.30	0.23	0.12	0.04	0.01	$\sim$0	$\sim$0	$\sim$0

- We measure the change in a function with respect to a variable using derivative of the function with respect to the variable.
- Consider the variance of the partial derivative, *i.e.*, $Var\left(\frac{\partial}{\partial\theta}\ln f(X, \theta)\right)$.

Information: Regularity Conditions

1. Let $S_\theta = \{x \in \mathbb{R} : f(x, \theta) > 0\}$ denote the support of the PMF or PDF $f(\cdot, \theta)$ and $S = \cup_{\theta \in \Theta} S_\theta$. Here, we assume that S_θ does not depend on θ, *i.e.*, $S_\theta = S$ for all $\theta \in \Theta$.
2. We also assume that the PDF (or PMF) $f(\cdot, \theta)$ is such that differentiation (with respect to θ) and integration (or sum) (with respect to x) are interchangeable.

Fisher Information

Definition 5: The Fisher information (or simply information) about parameter θ contained in X is defined by

$$\mathcal{I}_X(\theta) = E_\theta\left[\left(\frac{\partial \ln f(X,\, \theta)}{\partial \theta}\right)^2\right].$$

- Note that $\mathcal{I}_X(\theta) = 0$ if and only if $\frac{\partial}{\partial \theta} \ln f(x,\, \theta) = 0$ with probability one, which means that the PMF or PDF of X does not involve θ.
- An alternative form of Fisher information can be obtained as follows.

$$\mathcal{I}_X(\theta) = -E_\theta\left(\frac{\partial^2 \ln f(X,\, \theta)}{\partial \theta^2}\right).$$

Fisher Information

Example 14: Let $X \sim Poi(\lambda)$, where $\lambda > 0$. Then $\mathcal{I}_X(\lambda) = \frac{1}{\lambda}$.

Example 15: Let $X \sim N(\mu, \sigma^2)$, where σ is known and $\mu \in \mathbb{R}$ is unknown parameters. Then, $\mathcal{I}_X(\mu) = \frac{1}{\sigma^2}$.

Fisher Information

Definition 6: The Fisher information contained in a collection of RVs, say $\boldsymbol{X}$, is defined by

$$\mathcal{I}_{\boldsymbol{X}}(\theta) = E_\theta\left[\left(\frac{\partial}{\partial\theta}\ln f_{\boldsymbol{X}}(\boldsymbol{X}, \theta)\right)^2\right] = -E_\theta\left[\frac{\partial^2}{\partial\theta^2}\ln f_{\boldsymbol{X}}(\boldsymbol{X}, \theta)\right],$$

where $f_{\boldsymbol{X}}(\cdot, \theta)$ is the JPDF of $\boldsymbol{X}$ under θ.

Theorem 4: Let $X_1, X_2, \ldots, X_n$ be a RS from a population with PMF or PDF $f(\cdot, \theta)$, where $\theta \in \Theta$. Let $\mathcal{I}_{\boldsymbol{X}}(\theta)$ denote the Fisher information contained in the RS, then

$$\mathcal{I}_{\boldsymbol{X}}(\theta) = n\mathcal{I}_{X_1}(\theta) \quad \text{for all } \theta \in \Theta.$$

Fisher Information

Example 16: Let $X_1, \ldots X_n \overset{i.i.d.}{\sim} Poi(\lambda)$, where $\lambda > 0$. Then $\mathcal{I}_{\boldsymbol{X}}(\lambda) = \frac{n}{\lambda}$.

Example 17: Let $X_1, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$, where σ is known and $\mu \in \mathbb{R}$ is unknown parameters. Then, $\mathcal{I}_{\boldsymbol{X}}(\mu) = \frac{n}{\sigma^2}$.

Theorem 5: Let $\boldsymbol{X}$ be a RS and $\boldsymbol{T}$ be a statistic. Then $\mathcal{I}_{\boldsymbol{X}}(\theta) \geq \mathcal{I}_{\boldsymbol{T}}(\theta)$ for all $\theta \in \Theta$. The equality holds for all $\theta \in \Theta$ if and only if $\boldsymbol{T}$ is a sufficient statistic for θ.

Example 18: Let $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} Poi(\lambda)$ with $\lambda > 0$. Then, Fisher information contained in $T = \sum_{i=1}^{n} X_i$ is $\mathcal{I}_T(\lambda) = \frac{n}{\lambda}$. Hence, Fisher information contained in the RS is same as that contained in T. Therefore, T is a sufficient statistic for λ.

Method of Moment Estimator (MME)

- Introduced by Karl Pearson in the year 1902.
- The method is as follows:
  1. Suppose that we have a RS of size n form a population with PMF/PDF $f(x; \boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ is the unknown parameter.
  2. Calculate first k (no. of unknown parameters) moments $\mu_1', \ldots, \mu_k'$ of $f(x; \boldsymbol{\theta})$.
  3. Calculate first k sample moments $m_1', \ldots, m_k'$. Here m_r' is define by $m_r' = \frac{1}{n} \sum_{i=1}^{n} X_i^r$.
  4. Equate $\mu_r' = m_r'$ for $r = 1, 2, \ldots, k$.
  5. Solve the system of k equations (if they are consistent) for θ_i's. The solutions are the MMEs of the unknown parameters.

Examples

Example 19: Let $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} Bernoulli(\theta)$, $\theta \in [0, 1] = \Theta$. Then, the MME of θ is $\widehat{\theta} = \overline{X}$.

Example 20: Let $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$, $\boldsymbol{\theta} = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}^+ = \Theta$. Then the MMEs of μ and σ^2 are $\widehat{\mu} = \overline{X}$ and $\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$, respectively.

Example 21: Let $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} N(0, \sigma^2)$, $\sigma > 0$. Then, the MME of σ^2 is $\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n} X_i^2$.

Example 22: Let $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} N(\theta, \theta^2)$, $\theta > 0$. Then, the MME of θ is $\widehat{\theta} = \overline{X}$. However, this may not be a meaningful estimator as $\overline{X}$ can be negative with positive probability, while $\theta > 0$.

Previous two examples show that there are some degree of arbitrariness in this method.

Maximum Likelihood Estimator (MLE)

- Proposed by R. A. Fisher in 1912.
- One of the most popular method of estimation.
- Let us start with an example (next slide).

Example

Example 23: Let a box has some red balls and some black balls. It is known that number of black balls to red balls is in 1:1 or 1:2 ratio. We want to estimate whether it is 1:1 or 1:2. We may proceed as follows:

- Randomly draw two balls with replacement from the box.
- Let X be the number of black balls out of two drawn balls.
- $X \sim Bin(2, p)$, where $p \in \left\{\frac{1}{2}, \frac{1}{3}\right\}$.
- Problem boils down to estimate the value of p.

Example (cont.)

Now consider the following table, where the entries are $P_p(X = x)$ for each possible values of x and p.

	$x = 0$	$x = 1$	$x = 2$
$p = 1/2$	$1/4$	$1/2$	$1/4$
$p = 1/3$	$4/9$	$4/9$	$1/9$

- From first column, we see that for $x = 0$, the $P(X = 0)$ is maximum if $p = 1/3$. Hence if we observe $x = 0$ (that is no black balls in the sample), it is plausible to take $p = 1/3$ and the maximum likelihood estimate (MLE) of p is $1/3$.
- From second column, we see that for $x = 1$, the $P(X = 1)$ is maximum if $p = 1/2$.
- From third column, we see that for $x = 2$, the $P(X = 2)$ is maximum if $p = 1/2$.

Example (cont.)

Hence the maximum likelihood estimator of p is

$$\widehat{p} = \begin{cases} \frac{1}{3} & \text{if } x = 0 \\ \frac{1}{2} & \text{if } x = 1,\, 2. \end{cases}$$

If $x = 0$ occur, it is more likely that there are lesser number of black balls and hence the estimate turns out to be 1:2. For other values of x, it is 1:1.

MLE

Definition 7: Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a RS from a population with PMF/PDF $f(x; \boldsymbol{\theta})$. The function

$$L(\boldsymbol{\theta}, \boldsymbol{x}) = f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \prod_{i=1}^{n} f(x_i, \boldsymbol{\theta})$$

considered as a function of $\boldsymbol{\theta} \in \Theta$ for any fixed $\boldsymbol{x} \in \mathscr{X}$ ($\mathscr{X}$ is support of the RS), is called the likelihood function.

Definition 8: For a sample point $\boldsymbol{x} \in \mathscr{X}$, let $\widehat{\boldsymbol{\theta}}(\boldsymbol{x})$ be a value in Θ at which $L(\boldsymbol{\theta}, \boldsymbol{x})$ attains its maximum as a function of $\boldsymbol{\theta}$, with $\boldsymbol{x}$ held fixed. Then maximum likelihood estimator of the parameter $\boldsymbol{\theta}$ based on a RS $\boldsymbol{X}$ is $\widehat{\boldsymbol{\theta}}(\boldsymbol{X})$.

MLE

- MLE always lies in the parametric space.
- Problem of finding MLE boils down to finding maxima of a function, the likelihood function.
- Most of the cases it is easier to work with $l(\boldsymbol{\theta}, \boldsymbol{x}) = \ln L(\boldsymbol{\theta}, \boldsymbol{x})$ instead of $L(\boldsymbol{\theta}, \boldsymbol{x})$. Note that $\ln(\cdot)$ is a strictly increasing function on the positive side of $\mathbb{R}$.

Examples

Example 24: $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} P(\lambda)$, $\lambda > 0$. Then, the MLE of λ is $\widehat{\lambda} = \overline{X}$.

Example 25: $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu, 1)$, $\mu \in \mathbb{R}$. The MLE of μ is $\widehat{\mu} = \overline{X}$.

Example 26: $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$, where $\mu \in \mathbb{R}$ and $\sigma > 0$. Then, the MLEs of μ and σ^2 are $\widehat{\mu} = \overline{X}$ and $\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$, respectively.

Example 27: $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} N(0, \sigma^2)$, where $\sigma > 0$. Then, the MLE of σ^2 is $\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n} X_i^2$.

Examples

Example 28: $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu, 1)$, $\mu \leq 0$. The MLE of μ is

$$\widehat{\mu} = \begin{cases} \overline{X} & \text{if } \overline{X} \leq 0 \\ 0 & \text{otherwise.} \end{cases}$$

Example 29: Let X_1 be a sample of size one from $Bernoulli(\frac{1}{1+e^{\theta}})$, where $\theta \geq 0$. The MLE does not exist for $x = 0$ as $L(\theta, 0)$ is a increasing function of θ. On the other hand MLE exist for $x = 1$ and it is $\widehat{\theta} = 0$.

Example 30: $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} U(0, \theta)$, $\theta > 0$. The MLE of θ is $\widehat{\theta} = X_{(n)}$.

Example 31: $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} U\left(\theta - \frac{1}{2}, \theta + \frac{1}{2}\right)$, $\theta \in \mathbb{R}$. Any point in the interval $\left[X_{(n)} - \frac{1}{2}, X_{(1)} + \frac{1}{2}\right]$ is a MLE of θ.

Invariance Property of MLE

Theorem 6: (Without Proof) If $\widehat{\boldsymbol{\theta}}$ is MLE of $\boldsymbol{\theta}$, then for any function $\tau(\cdot)$ defined on Θ, the MLE of $\tau(\boldsymbol{\theta})$ is $\tau(\widehat{\boldsymbol{\theta}})$.

Example 32: $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} P(\lambda)$, $\lambda > 0$. The MLE of $P(X_1 = 0)$ is $e^{-\overline{X}}$.

MLE and Sufficient Statistics

Theorem 7: Let $\boldsymbol{T}$ be a sufficient statistics for $\boldsymbol{\theta}$. If a unique MLE exists for $\boldsymbol{\theta}$, it is a function of $\boldsymbol{T}$. If MLE of $\boldsymbol{\theta}$ exists but is not unique, then one can find a MLE that is a function of $\boldsymbol{T}$.

Example 33: Let $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} U(0, \theta)$, $\theta > 0$. We know that the MLE is unique and $X_{(n)}$, which is also sufficient.

Example 34: Let $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} U(\theta - 1/2, \theta + 1/2)$, $\theta \in \mathbb{R}$. Here a sufficient statistic is $\boldsymbol{T} = (X_{(1)}, X_{(n)})$. Also MLE is not unique and any point in the interval $\left[X_{(n)} - \frac{1}{2}, X_{(1)} + \frac{1}{2}\right]$ is a MLE of θ. Hence $\frac{1}{2}\left(X_{(1)} + X_{(n)}\right)$ is a MLE and it is also a function of $\boldsymbol{T}$. On the other hand $Q = \sin X_1 \left(X_{(n)} - \frac{1}{2}\right) + (1 - \sin X_1)\left(X_{(1)} - \frac{1}{2}\right)$ is a MLE but not a function of $\boldsymbol{T}$ only.

Comparison of Different Estimators

- We have considered two different methods of estimation.
- One may want to know which method provide a better estimator in a particular situation.
- Can we talk about average error? Can we talk about average squared error?
- There are some desirable properties of an estimator. Some of them are discussed here.

Unbiased Estimator

Definition 9: A statistic T is said to be an unbiased estimator (UE) of a parametric function $\tau(\theta)$ if $E_\theta(T) = \tau(\theta)$ for all $\theta \in \Theta$, the parametric space.

Remark 1:

- Unbiasedness tells us that there is no error on an average taken over all samples.
- Please note that **all** $\theta \in \Theta$ in the definition.
- An estimator which is not unbiased is called a biased estimator.
- The bias of T as an estimator of $\tau(\theta)$ is defined by $Bias(T) = E_\theta(T) - \tau(\theta)$ for all $\theta \in \Theta$.
- In general for unbiased estimator invariance property does not hold true.

Example

Example 35: Let $X_1, \ldots, X_n$ be a RS from a population with mean $\mu \in \mathbb{R}$. Then $\overline{X}$ is an unbiased estimator for μ as $E_\mu(\overline{X}) = \mu$ for all $\mu \in \mathbb{R}$.

Example 36: $X_1, \ldots, X_n \overset{i.i.d.}{\sim} U(0, \theta)$, $\theta > 0$. We saw that the MLE of θ is $X_{(n)}$. Now we want to check if $X_{(n)}$ is unbiased or not.

Example 37: Let $X_1, \ldots, X_n$ be a RS from a population with mean $\mu \in \mathbb{R}$ and finite variance σ^2. Define $T_1 = X_1$, $T_2 = \frac{1}{2}(X_1 + X_2), \ldots, T_n = \overline{X}$. It is easy to verify that $E(T_i) = \mu$ for all $\mu \in \mathbb{R}$ and for all $i = 1, 2, \ldots, n$.

Example 38: Let X be distributed as $Bin(2, p)$, where $p \in (0, 1)$. An UE of $\tau(p) = \frac{1}{p}$ does not exist.

Mean Square Error

Definition 10: The mean square error (MSE) of a statistic T as an estimator of θ is defined by $MSE(T) = E\left((T-\theta)^2\right)$.

Remark 2:

- $MSE(T) = Var(T) + (Bias(T))^2$.
- If T is UE for θ, then $MSE(T) = Var(T)$.
- An estimator with smaller value of MSE is preferred.

Example 39: Let $X_1, \ldots, X_n$ be a RS from a population with mean $\mu \in \mathbb{R}$ and finite variance σ^2. $T_1, T_2, \ldots, T_n$ are UE for μ. Which one to prefer?

Example 40: Let $X_1, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$, $n > 1$. Is the MLE of σ^2 is UE?

Consistent Estimator

Definition 11: Let T_n be an estimator based on a RS of size n. The estimator T_n is said to be consistent estimator of θ if the sequence of random variables $\{T_n : n \geq 1\}$ converges to θ in probability for all $\theta \in \Theta$.

Example 41: Let $X_1, X_2 \ldots, X_n$ be a RS from a population with mean $\mu \in \mathbb{R}$. Then using WLLN, $\overline{X}_n$ is a consistent estimator for μ.

Example 42: $X_1, \ldots, X_n \overset{i.i.d.}{\sim} U(0, \theta)$, $\theta > 0$. We saw that the MLE of θ is $X_{(n)}$. Now using the CDF of $X_{(n)}$, it can be shown that $X_{(n)}$ is a consistent estimator of θ.