STAT2006 Basic Concepts in Statistics and Probability II

Tutorial 5 Point Estimation

Benjamin Chun Ho Chan*†‡§

January 4, 2019

Abstract

It aims to introduce concepts of point estimation. Some exercises are provided for students to practice. Some materials do credit to former TAs while some are extracted from textbooks *Probability and Statistical Inference* written by Hogg and Tanis and used in STAT2001/2006, and *Statistical Inference* by Casella and Berger and used in STAT4003.

Notations and Definitions

- Set membership: $x \in A$ means "x is an element of the set A".
- Evaluate: $f(x)\big|_{x=a}$ means "f evaluated at $x = a$"; Set of real numbers: $\mathbb{R} = (-\infty, \infty)$
- Interval: $[a, b] = \{x \in \mathbb{R} \mid a \leq x \leq b\}$; Natural logarithm: $\ln X$; Is defined to be: $:=$
- Indicator function: $\mathbb{1}_{\{x \in A\}} = 1$ if $x \in A$ while $\mathbb{1}_{\{x \in A\}} = 0$ if $x \notin A$
- Differentiation: $\frac{d}{dx} f(x)$; Integration: $\int f(x)dx$; Partial derivative: $\frac{\partial y}{\partial x}$;
- Expectation: $E(\cdot)$; Variance: $Var(\cdot)$; Standard deviation: $SD(\cdot)$;
- Probability density function (pdf): $f(\cdot)$; Cumulative distribution function (cdf): $F(\cdot)$
- Likelihood function: $L(\cdot)$; Log-likelihood function: $l(\cdot)$; Product: $\prod_{i=1}^{n} x_i = x_1 \cdot x_2 \cdots x_n$
- Probability distribution: A tilde ($\sim$) means "has the probability distribution of".
- Independent and identically distributed (i.i.d.): Each random variable has the same probability distribution and all are mutually independent.
- Parameter(s): θ denotes population characteristic(s) that can be set to different values to produce different probability distributions.
- Parameter space: Ω denotes the set of all possible values for all the different parameters in a distribution.
- Estimator: $\hat{\theta}$ denotes estimator of θ; Mean: $\bar{X}$ denotes mean of $X_1, \ldots, X_n$.

*For enquiry, please email to 1155049861@link.cuhk.edu.hk.
†Personal profile: www.linkedin.com/in/benjamin-chan-chun-ho
‡GitHub repository: https://github.com/BenjaminChanChunHo/CUHK-STAT-or-RMSC-Tutorial-Note
§RPubs: http://rpubs.com/Benjamin_Chan_Chun_Ho

1 Overview of Statistics

1.1 Two Subjects of Statistics

- **Probability theory** is to study the properties of random variables and distributions. Topics: measure theory, convergence, law of large numbers, central limit theorem, etc.
- **Statistical inference** is to draw conclusions about a population based on sample data. Topics: point estimation, confidence interval, hypothesis testing, etc.

1.2 Statistical inference

- We assumed, so far, the **parameter** θ of a distribution is given in order to calculate the probability, density, distribution function, etc. However, in practice, the parameter θ is unknown.
- On the other hand, we observe the outcomes, which are the random samples $X_1, X_2, \ldots, X_n$. Therefore, we should use these observations to make some "**inference**" on the parameter of interest θ.

1.3 Point Estimation

- The rationale behind point estimation is simple. When sampling is from a population described by a pdf or pmf $f(x;\theta)$, knowledge of θ yields knowledge of the entire population. Hence, it is natural to seek a method of finding a good estimator of the point θ, that is, a good point estimator.
- Moreover, it is often the case that the parameter itself has a meaningful physical interpretation (as in the case of a population mean) so there is direct interest in obtaining a good point estimate of θ.
- A **point estimator** is any function $u(X_1, X_2, \ldots, X_n)$ of a sample; that is, any statistic is a point estimator.
- An **estimator** is a function of the sample, while an **estimate** is the realized value of an estimator (that is, a number) that is obtained when a sample is actually taken. Notationally, when a sample is taken, an estimator is a function of the random variables $X_1, \ldots, X_n$, while an estimate is a function of the realized values $x_1, \ldots, x_n$.
- **Maximum likelihood estimator** (Section 2) and **unbiased estimator** (Section 3) are two point estimators with some well-established and nice properties.
- Sometimes some function of θ, say $\tau(\theta)$, is of interest. Invariance property of maximum likelihood estimator is useful in this case. See Section 2.3.[1]

[1] "Statistical Inference" 2nd ed. (Casella and Berger) Ch7 p.311-312

2 Maximum Likelihood Estimator

2.1 Terminology

- The method of maximum likelihood can be considered to be the most popular technique for deriving estimators. It is widely popularized by famous statistician R.A. Fisher.
- Random samples $X_1, X_2, \ldots, X_n$ from distribution $F(\theta_1, \theta_2, \ldots, \theta_m)$ means $X_i \overset{\text{i.i.d.}}{\sim} F(\theta_1, \theta_2, \ldots, \theta_m)$.
- Write the density function as $f(x; \theta_1, \theta_2, \ldots, \theta_m)$, so do the distribution function F.
- **Likelihood function** is defined as

$$L(\theta_1, \theta_2, \ldots, \theta_m; x_1, x_2, \ldots, x_n) := f_{X_1, X_2, \ldots, X_n}(x_1, x_2, \ldots, x_n; \theta_1, \theta_2, \ldots, \theta_m).$$

- L as a function in $\theta_1, \theta_2, \ldots, \theta_m$ has a meaning that given the sample observations $x_1, x_2, \ldots, x_n$, how likely $\theta_1, \theta_2, \ldots, \theta_m$ would be the true parameters.
- Since $X_1, X_2, \ldots, X_n$ are random samples, they are independent, so

$$L(\theta_1, \theta_2, \ldots, \theta_m; x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f_{X_i}(x_i; \theta_1, \theta_2, \ldots, \theta_m). \tag{1}$$

- If $u_1(X_1, X_2, \ldots, X_n), \ldots, u_m(X_1, X_2, \ldots, X_n)$ maximizes the likelihood function L, then $\hat{\theta}_{\text{MLE},j} := u_j(X_1, X_2, \ldots, X_n)$ are called **maximum likelihood estimators** of θ_j for $j = 1, \ldots, m$.
- The abbreviation **MLE** stands for maximum likelihood estimator. Sometimes it stands for maximum likelihood estimate as well for the realized value of the estimator.

2.2 Methodology

- **Method:**
  1. Write down the likelihood function L by (1).
  2. Write a system of equations by $\dfrac{\partial L}{\partial \theta_j} = 0$ for $j = 1, \ldots, m$ (assumed differentiability).
  3. Solve $\hat{\theta}_j$ from the system.
  4. Check whether the $\hat{\theta}_j$ is a local maximum or not. Remember $\dfrac{\partial L}{\partial \theta_j} = 0$ only finding points with slope 0. Local maximum if either:
     (a) 1st derivative test: $\left.\dfrac{\partial L}{\partial \theta_j}\right|_{\theta_j < \hat{\theta}_j} > 0$ and $\left.\dfrac{\partial L}{\partial \theta_j}\right|_{\theta_j > \hat{\theta}_j} < 0$ for close enough θ_j to $\hat{\theta}_j$; or,
     (b) 2nd derivative test: $\left.\dfrac{\partial^2 L}{\partial \theta_j^2}\right|_{\theta_j = \hat{\theta}_j} < 0$.
  5. If yes, also compare with the boundary points in the parameter space. Then, find the **global maximum** $\hat{\theta}_{\text{MLE},j}$.

- There are inherent drawbacks associated with the general problem of finding the maximum of a function, and hence of maximum likelihood estimation. One problem is that of actually finding the global maximum and verifying that, indeed, a global maximum has been found.[2]
- **Useful technique:** If the likelihood function L is (usually) too complicated to be differentiated, consider **log-likelihood function** $l := \ln L$ and perform the steps as above.
- The product of marginal pdfs or pmfs becomes the sum of natural logarithm of them.
- Natural logarithm is strictly increasing on $(0, \infty)$, so the maximum point of l and L are the same.
- One other point to be aware of when finding a maximum likelihood estimator is that the maximization takes place only over the range of parameter values. In some cases this point plays an important part, e.g. it is known that θ must be nonnegative. In lectures, you learn the method of moments, in which chances are that the range of the estimator does not coincide with the range of the parameter it is estimating.[3]
- (Optional) If the likelihood cannot be maximized analytically (with closed-form solution), sometimes it has to be done by a computer numerically. However, numerical stability/ sensitivity and convergence to global maximum become critical issues. Interested students may explore statistical computing and/ or numerical analysis.
- (Optional) Alternative way to find an MLE is to abandon differentiation and proceed with a direct maximization. One general technique is to find a global upper bound on the likelihood function and then establish that there is a unique point for which the upper bound is attained.[4]

Example 1. Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} N(\mu, 1)$. Find the MLE $\hat{\mu}$.

To illustrate the use of direct maximization, we prove the following lemma:

Lemma 2.1. Let $x_1, \ldots, x_n$ be any numbers and

$$\bar{x} = \frac{x_1 + \cdots + x_n}{n}.$$

Then

$$\min_a \sum_{i=1}^n (x_i - a)^2 = \sum_{i=1}^n (x_i - \bar{x})^2.$$

In other words,

$$\bar{x} = \arg\min_a E(x - a)^2.$$

Proof.

$$\begin{aligned}
\sum_{i=1}^n (x_i - a)^2 &= \sum_{i=1}^n (x_i - \bar{x} + \bar{x} - a)^2 \\
&= \sum_{i=1}^n (x_i - \bar{x})^2 + 2\sum_{i=1}^n (x_i - \bar{x})(\bar{x} - a) + \sum_{i=1}^n (\bar{x} - a)^2 \\
&= \sum_{i=1}^n (x_i - \bar{x})^2 + \sum_{i=1}^n (\bar{x} - a)^2 \qquad \text{(cross term is 0)}
\end{aligned}$$

[2] "Statistical Inference" 2nd ed. (Casella and Berger) Ch7 p.315-316
[3] "Statistical Inference" 2nd ed. (Casella and Berger) Ch7 p.312-314, 318-319
[4] "Statistical Inference" 2nd ed. (Casella and Berger) Ch7 p.317-318

The right-hand side is minimized at $a = \bar{x}$. □

For the above problem, the likelihood function is

$$L(\mu; x_1, \ldots, x_n) = \prod_{i=1}^{n} \frac{1}{(2\pi)^{1/2}} e^{-(1/2)(x_i-\mu)^2} = \frac{1}{(2\pi)^{n/2}} e^{-(1/2)\sum_{i=1}^{n}(x_i-\mu)^2}.$$

By Lemma 2.1, for any number a,

$$\sum_{i=1}^{n}(x_i - a)^2 \geq \sum_{i=1}^{n}(x_i - \bar{x})^2$$

with equality if and only if $a = \bar{x}$. It implies that for any μ,

$$e^{-(1/2)\sum_{i=1}^{n}(x_i-\mu)^2} \leq e^{-(1/2)\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

with equality if and only if $\mu = \bar{x}$. Hence, $\bar{X}$ is the MLE.

2.3 Invariance Property of MLEs

- If $\hat{\theta}$ is the MLE of θ, then for any functions $\tau(\theta)$, the MLE of $\tau(\theta)$ is $\tau(\hat{\theta})$.
- The proof is complicated because of consideration of the mapping $\theta \to \tau(\theta)$ and the use of induced likelihood function. It is omitted here.
- For example, the MLE of μ^2, the square of a normal mean, is $\bar{X}^2$. In this case, $\tau(\mu) = \mu^2$.

3 Unbiased Estimator

- Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution $F(x; \theta_1, \ldots, \theta_m)$ and $\hat{\theta}$ be an estimator of θ.
- $\hat{\theta}$ is called **unbiased** if $E\left[\hat{\theta}\right] = \theta$. Otherwise, it is said to be **biased**.
- The mean squared error (MSE) of an estimator W of a parameter θ is the function of θ defined by $E_\theta(W-\theta)^2$.[5]
- Notice that the MSE measures the average squared difference between the estimator W and the parameter θ. It serves to measure the goodness of an estimator.

Theorem 3.1. In general, $E_\theta(W-\theta)^2 = Var_\theta(W) + [E_\theta(W) - \theta]^2 = Var_\theta(W) + [Bias_\theta(W)]^2$. For an unbiased estimator we have $E_\theta(W-\theta)^2 = Var_\theta(W)$.

Proof. Suppress subscript θ and some brackets of $E(\cdot)$ for simplicity.

$$\begin{aligned} E(W-\theta)^2 &= E(W - EW + EW - \theta)^2 \\ &= E(W-EW)^2 + 2E[(W-EW)(EW-\theta)] + E(EW-\theta)^2 \\ &= Var(W) + (EW-\theta)^2 \end{aligned}$$

In the second term, $(EW-\theta)$ is constant and can be pulled out. Then $E(W-EW) = EW - EW = 0$. Thus the second term is 0. In the last term, $(EW-\theta)^2$ is constant and hence one expectation sign can be dropped, i.e. $E(EW-\theta)^2 = (EW-\theta)^2$. □

Example 2. From Tutorial 0: Review of Normal Distribution, if

$$X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$$

and let the sample mean be

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

and the sample variance be

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

Then

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \quad \text{and} \quad \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}.$$

Since $E(\bar{X}) = \mu$, $\bar{X}$ is an unbiased estimator of μ. On the other hand,

$$E(S^2) = E\left[\frac{\sigma^2}{n-1}\frac{(n-1)S^2}{\sigma^2}\right] = \frac{\sigma^2}{n-1}(n-1) = \sigma^2.$$

Under normal model, S^2 is an unbiased estimator of σ^2. However, S is a biased estimator of σ. From Lecture Note: Theory of Point Estimation, the MLE of μ is $\hat{\mu} = \bar{X}$ and the MLE of σ^2 is $\hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$. Note that

$$E(\hat{\sigma^2}) = \frac{n-1}{n}E(S^2) = \frac{n-1}{n}\sigma^2.$$

So $\hat{\mu}$ is unbiased while $\hat{\sigma^2}$ is biased.

---

[5] "Statistical Inference" 2nd ed. (Casella and Berger) Ch7 p.330-332

4 Exercises

Exercise 1. Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with pdf

$$f(x;\theta) = \frac{x^2 e^{-\frac{x}{\theta}}}{2\theta^3}$$

where $0 < x < \infty$ and $0 < \theta < \infty$. Find its maximum likelihood estimator $\hat{\theta}_{\text{MLE}}$. Is it an unbiased estimator?

Exercise 2. Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with pdf

$$f(x;\theta) = \frac{5\theta^5}{x^6}$$

where $0 < \theta \leq x < \infty$. Find the maximum likelihood estimator of θ. Is it unbiased estimator of θ?

Exercise 3 [13-14 Final Q1, 20%]
Assume that $\theta > 0$. Let $X_1, \ldots, X_n$ be a random sample from a distribution that has a pdf

$$f(x;\theta) = Cx^4, \quad -\theta \le x \le \theta,$$

where C is some normalization constant, and $f(x;\theta) = 0$ otherwise.

(a) Find the normalization constant C. [4%]

(b) Find the maximum likelihood estimator $\hat{\theta}$ of θ, and also the mean and variance of $\hat{\theta}$. [8%]

(c) Consider $\overline{|X|} := \frac{\sum_{i=1}^{n} |X_i|}{n}$ as the other estimator of θ. Find its mean and variance. [3%]

(d) Find constants c_1 and c_2 so that $c_1\hat{\theta}$ and $c_2\overline{|X|}$ are unbiased estimators of θ. [2%]

(e) Which one of $c_1\hat{\theta}$ and $c_2\overline{|X|}$ is more preferable? [3%]

Exercise 3

Exercise 3

Exercise 4 [13-14 Final Q3(a) Modified, $\sim 6\%$]
Let $X \sim N(0, \sigma^2)$ where σ^2 is unknown. Find the maximum likelihood estimator for σ^2.

Exercise 5 [13-14 Final Q4(b), 3%]
Let $X_1, X_2, \ldots, X_n$ be a random sample from Poisson distribution with mean $\lambda > 0$ which is unknown. Find the MLE of λ and show that it is unbiased.