Probability and Random Process (SWE3026)

Statistical Inference

JinYeong Bak
jy.bak@skku.edu
College of Computing, SKKU

H. Pishro-Nik, "Introduction to probability, statistics, and random processes", available at https://www.probabilitycourse.com, Kappa Research LLC, 2014.

Objectives

We will:

1) **Learn statistical inference**
2) **Understand two different statistical inference methods**

Statistical Inference

Definition

A collection of methods that deal with drawing conclusions from data that are prone to random variation

Example

We want to predict the outcome of an election. But we cannot poll the entire population, so we will choose a random sample from the population.

Statistical Inference

Frequentist (classical) inference

The unknown quantity which we want to estimate from data is assumed to be a fixed (deterministic, non-random) quantity,

It is to be estimated by the observed data

Bayesian inference

The unknown quantity is assumed to be a random variable,

We have some initial guess about the distribution of the quantity and we update the distribution using Bayes Rule

Random Sampling

Definition

The collection of random variables $X_1, X_2, \ldots, X_n$ is said to be a random sample of size n if they are independent and identically distributed (i.i.d)

Random Sampling

Example

We want to identify the height distribution of people. So we define n random variables $X_1, X_2, \dots, X_n$ where X_i is the height of the i-th sampled person with replacement. (X_i is the height of the i-th person that is chosen uniformly and independently from the population.)

We can estimate the average height in the population as follows:

$$\widehat{\Theta} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

: point estimator for the average height in the population

Random Sampling

Properties

- The X_i's are independent
- $F_{X_1}(x) = F_{X_2}(x) = \cdots = F_{X_n}(x) = F_X(x)$
- $E[X_i] = E[X] = \mu < \infty$
- $0 < Var(X_i) = Var(X) = \sigma^2 < \infty$

Sample Mean

Definition

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

Properties

- $E[\overline{X}] = \mu$
- $Var(\overline{X}) = \frac{\sigma^2}{n}$
- $\lim_{n \to \infty} P(|\overline{X} - \mu| \geq \epsilon) = 0$
- $\lim_{n \to \infty} P(Z_n \leq x) = \Phi(x)$ where $Z_n = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}}$

Good Estimator?

We want to know an unknown parameter θ (i.e., expected value of a R.V)

To estimate θ, we need some random samples $X_1, X_2, \dots, X_n$

Point estimator $\widehat{\Theta} = h(X_1, X_2, \dots, X_n)$

One of the estimators for expected value of a R.V is sample mean

$$\widehat{\Theta} = \overline{X} = \frac{X_1 + X_2 + \dots + X_n}{n}$$

Here, how can we know a good estimator?

A good estimator $\widehat{\Theta}$ is able to give us values that are close to the real value of θ

Bias

Definition

The bias of a estimator $\widehat{\Theta}$:

$$B(\widehat{\Theta}) = E[\widehat{\Theta}] - \theta$$

Definition

Unbiased estimator:

$$B(\widehat{\Theta}) = 0$$

Example

Let $X_1, X_2, \ldots, X_n$ be a random sample. Show that the sample mean

$$\widehat{\Theta} = \overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

is an unbiased estimator of $\theta = E[X_i]$

Mean Squared Error

How to ensure that an estimator is a good estimator?

Definition

Mean Squared Error (MSE) of a estimator $\widehat{\Theta}$:

$$MSE(\widehat{\Theta}) = E[(\widehat{\Theta} - \theta)^2]$$

Example

Let $X_1, X_2, \dots, X_n$ be a random sample. Its mean is $E[X_i] = \theta$, and variance $Var(X_i) = \sigma^2$. Show that $\widehat{\Theta}_2 = \overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$ is better estimator than $\widehat{\Theta}_1 = X_1$

Mean Squared Error

Definition

$$MSE(\hat{\Theta}) = E[(\hat{\Theta} - \theta)^2]$$

Properties

- $MSE(\hat{\Theta}) = Var(\hat{\Theta}) + B(\hat{\Theta})^2$
- **Consistent estimator if** $\lim_{n \to \infty} P(|\widehat{\Theta_n} - \theta| \geq \epsilon) = 0$**, for all** $\epsilon > 0$

Sample Variance

Definition

$$S^2 = \frac{1}{n-1}\sum_{k=1}(X_k - \bar{X})^2 = \frac{1}{n-1}(\sum_{k=1} X_k^2 - n\bar{X}^2)$$

Example

Let $X_1, X_2, \dots, X_n$ be a random sample. Find the bias of sample variance.

$$S^2 = \frac{1}{n-1}\sum_{k=1}(X_k - \overline{X})^2 = \frac{1}{n-1}(\sum_{k=1} X_k^2 - n\overline{X}^2)$$

Likelihood Estimation

Definition

Let $X_1, X_2, \dots, X_n$ be a random sample from a distribution with a parameter θ. Suppose that we have observed $X_1 = x_1, X_2 = x_2, \dots, X_n = x_n$.

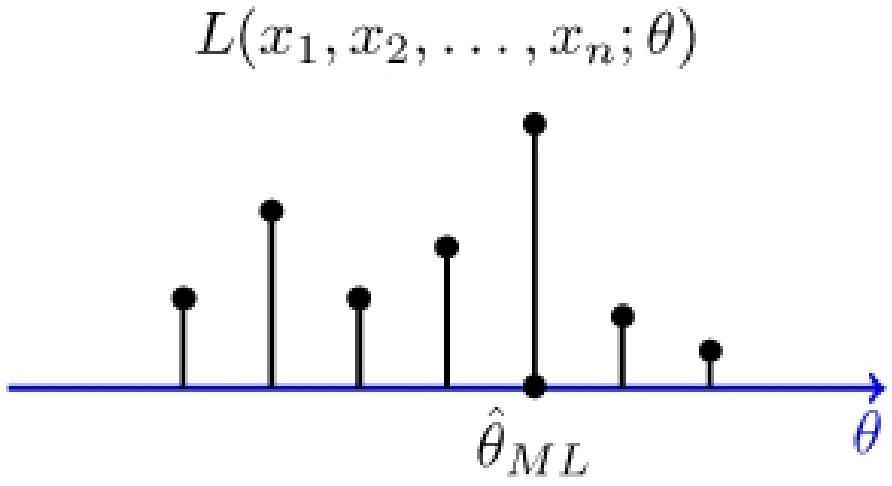
The likelihood function

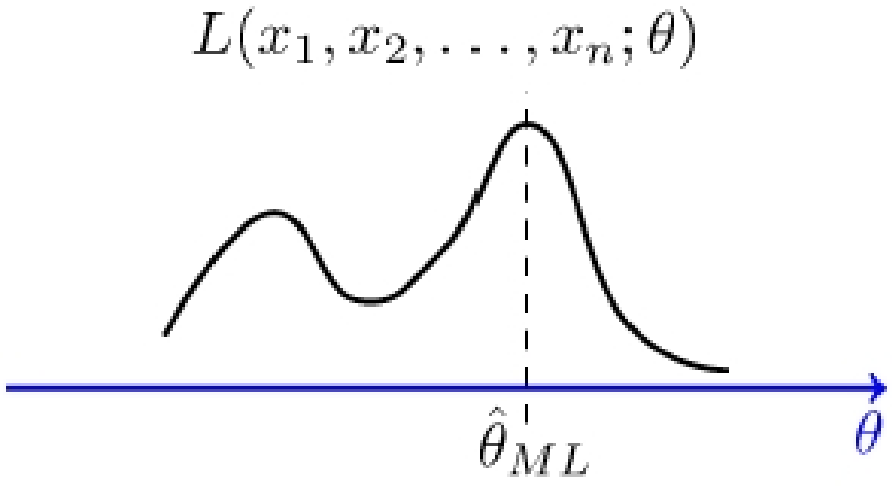
- $L(x_1, x_2, \dots, x_n; \theta) = P_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n; \theta)$ **if X_is are discrete**
- $L(x_1, x_2, \dots, x_n; \theta) = f_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n; \theta)$ **if X_is are continuous**

Maximum Likelihood Estimation

Definition

A systematic way of parameter estimation that finds the parameter value that maximizes the likelihood function





Example

I have a bag that contains 3 balls. Each ball is either red or blue, but I have no information in addition to this. Thus, the number of blue balls, call it θ, might be 0, 1, 2, or 3. I am allowed to choose 4 balls at random from the bag with replacement. We define the random variables X_1, X_2, X_3, and X_4 as follows

$$X_i = \begin{cases} 1 \; if \; the \; ith \; chosen \; ball \; is \; blue \\ 0 \; if \; the \; ith \; chosen \; ball \; is \; red \end{cases}$$

Note that $X_i \sim Bernoulli(\frac{\theta}{3})$. After doing my experiment, I observe the following values of X_is

$$x_1 = 1, x_2 = 0, x_3 = 1, x_4 = 1$$

1. Find the probability of the observed sample
2. For which value of θ is the probability of the observed sample is the largest?

I have a bag that contains 3 balls. Each ball is either red or blue, but I have no information in addition to this. Thus, the number of blue balls, call it θ, might be 0, 1, 2, or 3. I am allowed to choose 4 balls at random from the bag with replacement. We define the random variables X_1, X_2, X_3, and X_4 as follows

$$X_i = \begin{cases} 1 \; if \; the \; ith \; chosen \; ball \; is \; blue \\ 0 \; if \; the \; ith \; chosen \; ball \; is \; red \end{cases}$$

Note that $X_i \sim Bernoulli(\frac{\theta}{3})$. After doing my experiment, I observe the following values of X_is

$$x_1 = 1, x_2 = 0, x_3 = 1, x_4 = 1$$

1. Find the probability of the observed sample
2. For which value of θ is the probability of the observed sample is the largest?

$P_{X_1}(1) P_{X_2}(0) \cdot P_{X_3}(1) \cdot P_{X_4}(1)$

$= \frac{\theta}{3} \left(1 - \frac{\theta}{3}\right) \cdot \frac{\theta}{3} \cdot \frac{\theta}{3}$

$= \left(\frac{\theta}{3}\right)^3 \cdot \left(1 - \frac{\theta}{3}\right)$

Example

For the following random samples, find the maximum likelihood estimate of θ:

$X_i \sim Exponential(\theta)$ **and we have observed**

$$(x_1, x_2, x_3, x_4) = (1.23, 3.32, 1.98, 2.12)$$