Estimation Methods

Chi-Chao Hung [*]

February 19, 2025

Contents

1 Models and Random Samples

In statistics, data are viewed as realizations of random variables, which we model mathematically (Bickel and Doksum, 2015).

Definition 1.1 (Parametric Model)**.** A **parametric model** for X is a complete probability function depending on an unknown parameter θ. The parameter θ belongs to a set Θ which is called the **parameter space**.

In the discrete case, we can write a parametric model as a probability mass function $\pi(X \mid \theta)$. In the continuous case, we can write it as a density function $f(x \mid \theta)$.

Example 1.1 (Normal Distribution)**.** Consider a student's preliminary exam score X. Suppose we model the random variable X by assuming

$$X \sim N(\mu, \sigma^2).$$

This model depends on the parameters $\mu \in \mathbb{R}$ and $\sigma^2 \in \mathbb{R}^+$. We call the collection of all such distributions a **parametric family**. Our goal in estimation is to determine the true distribution of X within this family.

[*] This note borrows heavily from Hansen (2022). The examples and code snippets are my own (supported by ChatGPT.) All errors are mine. If you find any errors or have any suggestions, please contact me via email at r13323021@ntu.edu.tw.

Remark 1. It's important to note that these assumptions are merely a simplification of reality. The scores are discrete not continuous, and can not be negative, so the model is technically *wrong*. Still, such assumptions are often used because they provide a reasonable approximation.

Often, we focus on **random samples**. That is, we assume our collected sample is independent and identically distributed (IID) according to $f(x \mid \theta)$. For instance, if we observe 100 exam scores $X_1, X_2, \ldots, X_{100}$, we might assume

$$X_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2).$$

Let's simulate a random sample from normal distribution. Here we generate 10 IID observations. Researchers only observe the random sample but not the parameters. Our goal is to guess the underlying parameters that generate the observed random sample.

```
# Set random seed for replication
set.seed(123)

# Data generating Process
n = 10 # Number of observation
mu <- 2 # Mean
sd <- 1 # standard deviation
nsample <- rnorm(n, mu, sd) # IID random sample
# check data
print(nsample)
```

```
##  [1] 1.4395244 1.7698225 3.5587083 2.0705084 2.1292877 3.7150650 2.4609162
##  [8] 0.7349388 1.3131471 1.5543380
```

2 Estimation Methods

First, we introduce some terms related to the estimation process.

- **Estimand:** The parameters of interest in the population.
- **Estimator:** A function of the sample that provides a guess for the estimand.
- **Estimate:** The realized numerical value of an estimator.

The problem of estimation is to construct an estimator that is a well-educated guess of the estimand. In this section, we provide three methods for constructing an estimator.

2.1 Plug-in Principle

The **plug-in principle** is one of the most intuitive methods for choosing an estimator. It states that we can replace the *true* (and unknown) distribution of a random variable with the *empirical* distribution derived from our sample and then compute the parameter of interest. The principle is also referred to as the "Sample Analogue Principle" (Stachurski, 2016).

Definition 2.1 (Empirical Distribution)**.** Let $X_1, X_2, \ldots, X_n$ be an IID sample from an unknown distribution with cumulative distribution function (CDF) F. The *empirical distribution* $\widehat{F}_n$ is defined by

$$\widehat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{X_i \leq x\},$$

where $\mathbf{1}\{\cdot\}$ is the indicator function.

Example 2.1 (Estimator of the Mean)**.** Suppose we wish to estimate $\mathrm{E}[X]$, the expectation of a random variable X. Under the plug-in principle, we replace the true distribution with $\widehat{F}_n$ and calculate its mean. As a result,

$$\widehat{\mathrm{E}}[X] = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

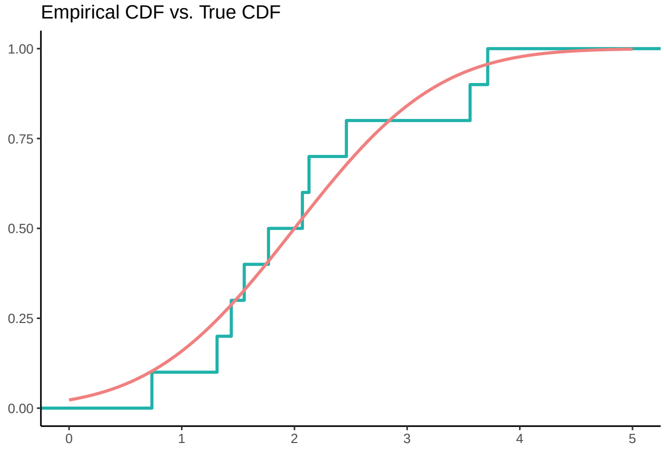
which is simply the *sample mean.*

Let's plot the empirical distribution of our sample. The empirical CDF is itself a plug-in estimator of the ture CDF.

```r
ggplot(data.frame(x = nsample), aes(x = x)) +
# Empirical CDF (step function)
stat_ecdf(geom = "step", color = "lightseagreen", size = 1) +
# Theoretical CDF for N(mu, sd^2)
stat_function(fun = function(x) pnorm(x, mean = mu, sd = sd),
                color = "lightcoral", size = 1) +
xlim(0, 5) + # Adjust x-axis range
labs(
    title = "Empirical CDF vs. Theoretical CDF",
    x = "",
    y = ""
) +
theme_classic()
```

2.2 Method of Moment

The method of moments (MM) is a straightforward application of the plug-in principle to parameter estimation. Suppose a random variable X has a parametric distribution with unknown parameters. Some (or all)



of its moments can be expressed as functions of these parameters. The key idea is to replace the **population moments** by the corresponding **sample moments** and solves for the parameters.

Example 2.2 (Uniform Distribution)**.** Let $X_i \overset{\text{iid}}{\sim} \text{Uniform}[0, \beta]$. We want to find the MM estimator for β.

1. **Identify the moment in terms of β.** For the Uniform$[0, \beta]$ distribution, the first (population) moment is

$$\mathrm{E}[X] = \frac{\beta}{2}.$$

2. **Replace the population moment by the sample moment.** The sample moment (the sample mean) is

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

According to the MM, we set

$$\bar{X} = \frac{\beta}{2}.$$

This is the **moment equation**.

3. **Solve for β.** From the above equation, the MM estimator for β is

$$\widehat{\beta}_{\text{MM}} = 2\,\bar{X} = \frac{2}{n}\sum_{i=1}^{n} X_i.$$

Example 2.3 (Normal Distribution Mean)**.** Let $X_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2)$. We want to estimate the unknown mean μ using the MM. The first population moment (expectation) is

$$\mathrm{E}[X] = \mu.$$

Setting the population moment equal to the sample moment gives

$$\bar{X} = \mu.$$

The MM estimator for μ is thus

$$\widehat{\mu}_{\mathrm{MM}} = \bar{X}.$$

Let's find the moment of method estimate of μ in the previous example.

```r
# Estimator is a function of Data (Random Variables)
mm_estimator <- function(x) { mean(x) }

# Estimator is the result after plugging samples in an estimator
mm_estimate  <- mm_estimator(nsample)

print(mm_estimate)
```

```
## [1] 2.074626
```

2.3 Maximum Likelihood

Maximum likelihood estimation is another popular method for constructing estimators. The key idea is to choose the parameter value that makes the observed sample "most likely" under the assumed model.

Definition 2.2 (Likelihood)**.** Let $X_1, X_2, \ldots, X_n$ be an IID sample from a parametric family of distributions with probability density function (PDF) $p(x \mid \theta)$. The *likelihood function* $L_n(\theta)$ is defined as the product of the individual PDF evaluated at the observed sample:

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

Example 2.4 (Coin Flips)**.** Suppose $X \sim \mathrm{Ber}(p)$. We flip this coin twice and observe the outcomes $\{X_1, X_2\} = \{0, 1\}$. The likelihood for p is

$$L_n(p) = p^{\#\text{ of 1's}}(1-p)^{\#\text{ of 0's}} = p^1(1-p)^1.$$

Hence, for $p = 0.5$, the likelihood is $0.5 \times 0.5 = 0.25$, while for $p = 0.3$, it is $0.3 \times 0.7 = 0.21$. Since $0.25 > 0.21$, the sample $\{0, 1\}$ is more likely to have been generated by $p = 0.5$ than by $p = 0.3$.

Definition 2.3 (Maximum Likelihood Estimator)**.** The *maximum likelihood estimator (MLE)* $\widehat{\theta}_{MLE}$ is the value of θ that maximizes the likelihood function $L_n(\theta)$:

$$\widehat{\theta} = \arg\max_{\theta \in \Theta} \; L_n(\theta).$$

To construct an MLE estimator we take the following steps:

1. Find the PDF $f(x \mid \theta)$
2. Construct Likelihood function $L_n(\theta)$
3. Take logarithm $\log L_n(\theta)$ for convenience's sake
4. If possible, solve the fisrt order condition to find maximum
5. If impossible to solve analytically, use numerical methods

Example 2.5 (MLE of Normal Mean with Known Variance)**.** Suppose we have an IID sample $X_1, X_2, \ldots, X_n$ from a normal distribution with unknown mean μ and **known** variance $\sigma^2 = 1$. That is,

$$X_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu, 1).$$

We wish to find the MLE of μ.

The PDF for a single observation x_i is

$$f(x_i \mid \mu) = \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{(x_i - \mu)^2}{2}\Big).$$

Because the X_i's are IID, the joint likelihood for the sample $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$ is

$$L_n(\mu) = \prod_{i=1}^{n} f(x_i \mid \mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{(x_i - \mu)^2}{2}\Big).$$

Explicitly,

$$L_n(\mu) = \Big(\frac{1}{\sqrt{2\pi}}\Big)^n \exp\Big(-\frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2\Big).$$

It is often more convenient to work with the *log-likelihood.* In this example,

$$\ell(\mu) = \log L_n(\mu) = \log\Bigg[\Big(\frac{1}{\sqrt{2\pi}}\Big)^n\Bigg] - \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2.$$

The maximization problem can be solved by numerical method. Let's construct the likelihood function of μ and find the maximum likelihood estimate numerically.

```
# Likelihood Function of Normal(mu,1)
likelihood <- function(x, mu){
    prob <- prod( dnorm(x, mean = mu, 1)) # Product of each pdf(x
    )
    return(prob)
}

# optim() performs minimization
mle_result <- optim( c(0), function(t) - likelihood( nsample, t)
    )

# Likelihood at mu=2
cat("Likelihood(mu = 2; nsample):", likelihood( nsample, mu = 2 )
    , "\n")
# Numerical Estimate
cat("Numerical MLE estimates:", mle_result$par, "\n" )
```

```
## Likelihood(mu = 2; nsample): 1.656343e-06
## Numerical MLE estimates: 2.074628
```

We can also derive the analytical solution by applying the first-order condition (FOC).

$$\frac{d}{d\mu}\ell(\mu) = \sum_{i=1}^{n}(x_i - \mu).$$

Setting this equal to 0:

$$\sum_{i=1}^{n}(x_i - \hat{\mu}) = 0 \quad \Longrightarrow \quad \hat{\mu}_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{X}.$$

In this example the MM estimator and the MLE estimator coincide ($\hat{\mu}_{\text{MM}} = \hat{\mu}_{\text{MLE}} = \bar{X}$). This is not generally true. (Try to find the MLE of β in Exercise 2.2.)

A special property of MLE is that it is invariant to transformations (Hansen, 2022).

Theorem 2.1 (MLE Invariance)**.** If $\hat{\theta}$ is the MLE of θ then for any function $h(\theta)$ the MLE of $\beta = h(\theta)$ is $\hat{\beta} = h(\hat{\theta})$.

Example 2.6 (MLE of Bernoulli variance)**.** Let $X_i \overset{\text{iid}}{\sim} \text{Ber}(p)$ and $\hat{p} = \bar{X}$ is the MLE of p. We also know that $\sigma_X^2 = \text{Var}(X) = p(1-p)$. By the invarinace property of MLE, the MLE of σ_X^2 is $\bar{X}(1-\bar{X})$.[1]

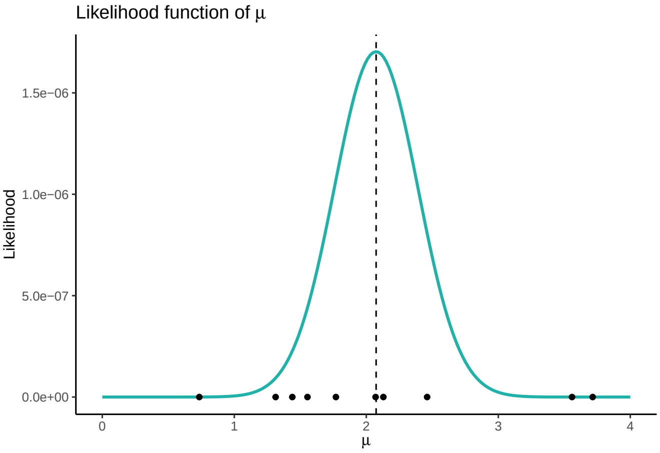
[1] This example is adapted from Chen (2023).

Here's an illustration of the likelihood function. The guess of μ that MLE provides is the value at which the likelihood function peaked.

```r
# A vector of possible mu
x_grid <- seq(0, 4, length.out = 200)

df <- data.frame(
    mu = x_grid,
    L = sapply( x_grid, function(t) likelihood( nsample,t) )
)

ggplot(df, aes(x = mu, y = L)) +
    geom_line( color = "lightseagreen", size = 1) +
    geom_vline( xintercept = mle_result$par,
                linetype = "dashed",
                linewidth = 0.5) +
    geom_point( data = data.frame(mu = nsample, L = rep( min(df$L
    ), n) ),
                aes(x = mu, y = L),
                inherit.aes = FALSE) +
    labs(title = expression("Likelihood function of " * mu),
        x = expression(mu),
        y = "Likelihood")+
    theme_classic()
```



References

Bickel, Peter J. and Kjell A. Doksum, *Mathematical Statistics. 1* Texts in Statistical Science, 2. ed ed., Boca Raton, Fla.: CRC Press, 2015.

Chen, Shiu-Sheng, *Probability and statistical inference with R*, 2 ed., Taipei: Tung Hua Book Co.,Ltd., 2023.

Hansen, Bruce E., *Probability & Statistics for Economists*, Princeton Oxford: Princeton University Press, 2022.

Stachurski, John, *A Primer in Econometric Theory*, Cambridge, Massachusetts London, England: The MIT Press, 2016.

Acronyms

CDF	cumulative distribution function. 3
FOC	first-order condition. 7
IID	independent and identically distributed. 2, 3, 5, 6
MLE	maximum likelihood estimator. 6–8
MM	method of moments. 3–5, 7
PDF	probability density function. 5, 6