3.6 Some Applications of Delta Rule

3.6.1 Variance stabilizing transformation

Problem: In statistics, we want to draw inferences on some parameter θ based on Y_n where

$$\sqrt{n}\left(Y_n-\theta\right) \xrightarrow{d} \mathcal{N}\left(0,\sigma^2\left(\theta\right)\right)$$

as $n \to \infty$, i.e.

$$Yn \overset{d}{\approx} \mathcal{N}\left(\theta, \frac{\sigma^2\left(\theta\right)}{n}\right) \quad \text{when n is large}$$

But the variance of the approximated limiting distribution is a known function of the mean θ. The purpose of variance stabilizing transformation is to eliminate the dependence of the variance on the mean, in order to make the simple regression-based or analysis of variance (ANOVA) techniques more valid.

Is there some function g to apply on Y_n such that the variance of $g(Y_n)$ will be nearly constant?

$$g(Y_n) \overset{d}{\approx} \mathcal{N}\left(g(\theta), \frac{1}{n}\right) \quad ?$$

I.e. we want to find g such that

$$\sqrt{n}\left(g(Y_n)-g(\theta)\right) \xrightarrow{d} \mathcal{N}\left(0,1\right)$$

as $n \to \infty$. By delta rule

$$\sqrt{n}\left(g(Y_n)-g(\theta)\right) \xrightarrow{d} \mathcal{N}\left(0,\sigma^2\left(\theta\right)\left(g'(\theta)\right)^2\right)$$

as $n \to \infty$. So need g to satisfy $\sigma^2\left(\theta\right)\left(g'(\theta)\right)^2 = 1$, i.e.

$$g'(\theta) = \frac{1}{\sigma\left(\theta\right)}$$

So can take

$$g(\theta) = \int \frac{d\theta}{\sigma\left(\theta\right)}$$

$\boxed{\text{Summary 3.6.1.}}$ **(Variance stabilizing transformation)**. If

$$\sqrt{n}\left(Y_n - \theta\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2\left(\theta\right)\right)$$

as $n \to \infty$, then take

$$g(\theta) = \int \frac{d\theta}{\sigma\left(\theta\right)}$$

one has

$$\sqrt{n}\left(g(Y_n) - g(\theta)\right) \xrightarrow{d} \mathcal{N}\left(0, 1\right)$$

as $n \to \infty$.

$\boxed{\textit{Example 3.6.1.}}$. Let $X_1, X_2, \ldots$ be i.i.d. $\mathcal{P}oi(\theta)$, $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. By CLT

$$\sqrt{n}\left(\overline{X}_n - \theta\right) \xrightarrow{d} \mathcal{N}(0, \theta)$$

as $n \to \infty$. By the variance stabilizing transformation, $\sigma^2(\theta) = \theta$

$$g(\theta) = \int \frac{d\theta}{\sqrt{\theta}} = 2\sqrt{\theta}$$

Hence

$$\sqrt{n}\left(2\sqrt{\overline{X}_n} - 2\sqrt{\theta}\right) \xrightarrow{d} \mathcal{N}(0, 1)$$

as $n \to \infty$, i.e. when n is large

$$2\sqrt{\overline{X}_n} \overset{d}{\approx} \mathcal{N}(2\sqrt{\theta}, \frac{1}{n})$$

$\boxed{\textit{Example 3.6.2.}}$. Let $X_1, X_2, \ldots$ be i.i.d. $\mathcal{E}xp(\theta)$, $\mathbb{E}(X_i) = \theta$, $\mathbb{V}ar(X_i) = \theta^2$, $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. By CLT

$$\sqrt{n}\left(\overline{X}_n - \theta\right) \xrightarrow{d} \mathcal{N}(0, \theta^2)$$

as $n \to \infty$. By the variance stabilizing transformation, $\sigma^2(\theta) = \theta^2$

$$g(\theta) = \int \frac{d\theta}{\theta} = \ln \theta$$

Hence

$$\sqrt{n}\left(\ln \overline{X}_n - \ln \theta\right) \xrightarrow{d} \mathcal{N}(0,1)$$

as $n \to \infty$, i.e. when n is large

$$\ln \overline{X}_n \overset{d}{\approx} \mathcal{N}(\ln \theta, \frac{1}{n})$$

3.6.2 Wilson-Hilferty approximation to χ^2 distribution

$\boxed{\text{Summary 3.6.2.}}$ **(Wilson-Hilferty approximation to χ^2).**

$$\chi^2(v) \overset{d}{\approx} v\left(1 - \frac{2}{9v} + \sqrt{\frac{2}{9v}}Z\right)^3$$

where $Z \sim \mathcal{N}(0,1)$.

- quite accurate for v around 10
- let $Y \sim \chi^2(v)$, it works backwards

$$\sqrt{\frac{9v}{2}}\left(\left(\frac{Y}{v}\right)^{1/3} - 1 + \frac{2}{9v}\right) \overset{d}{\approx} Z$$

Consider the linear combination of $X_1, \ldots, X_v$ that are i.i.d. $\chi^2(1)$, and $\mathbb{E}X_i = 1$, $\mathbb{V}arX_i = 2$

$$\frac{Y}{v} \overset{d}{=} \frac{X_1 + \cdots + X_v}{v}$$

By CLT

$$\sqrt{v}\left(\frac{Y}{v} - 1\right) \xrightarrow{d} \mathcal{N}(0,2) \quad \text{as } n \to \infty$$

By delta rule, if $g(x) = x^{1/3}$, $g'(x) = \frac{1}{3}x^{-2/3}$, $g'(1) = \frac{1}{3}$

$$\sqrt{v}\left(\left(\frac{Y}{v}\right)^{1/3} - 1\right) \xrightarrow{d} \mathcal{N}\left(0, 2\left(\frac{1}{3}\right)^2\right) = \mathcal{N}\left(0, \frac{2}{9}\right) \quad \text{as } v \to \infty$$

i.e.

$$\sqrt{\frac{9v}{2}}\left(\left(\frac{Y}{v}\right)^{1/3}-1\right)\xrightarrow{d}\mathcal{N}(0,1)\quad\text{as } v\to\infty$$

Apply corollary 3.4.1 of the continuous mapping theorem and since $\sqrt{\frac{9v}{2}}\frac{2}{9v}=\sqrt{\frac{2}{9v}}\to 0$

$$\sqrt{\frac{9v}{2}}\left(\left(\frac{Y}{v}\right)^{1/3}-1+\underbrace{\frac{2}{9v}}_{\text{why?}}\right)\xrightarrow{d}\mathcal{N}(0,1)\quad\text{as } v\to\infty$$

Where did the extra additive term $\frac{2}{9v}$ come from?

Look closer to the Taylor expansion of $g(x)=x^{1/3}$ at 1

$$g(x)\approx g(1)+g'(1)(x-1)+\frac{g''(1)}{2}(x-1)^2\quad x\text{ near }1$$

Plug in $g'(x)=\frac{1}{3}x^{-2/3}$, $g'(1)=\frac{1}{3}$, $g''(x)=-\frac{2}{9}x^{-5/3}$, $g''(1)=-\frac{2}{9}$

$$x^{1/3}\approx 1+\frac{1}{3}(x-1)-\frac{1}{9}(x-1)^2\quad x\text{ near }1$$

Substitute $x=Y/v$

$$\left(\frac{Y}{v}\right)^{1/3}\approx 1+\frac{1}{3}\left(\left(\frac{Y}{v}\right)-1\right)-\frac{1}{9}\left(\left(\frac{Y}{v}\right)-1\right)^2$$

Take expectation

$$\mathbb{E}\left(\frac{Y}{v}\right)^{\frac{1}{3}}-1\approx\frac{1}{3}\left(\mathbb{E}\left(\frac{Y}{v}\right)-1\right)-\frac{1}{9}\mathbb{E}\left(\left(\frac{Y}{v}\right)-1\right)^2=-\frac{1}{9}\mathbb{V}ar\frac{Y}{v}=-\frac{2}{9v}$$

which implies

$$\mathbb{E}\left(\left(\frac{Y}{v}\right)^{\frac{1}{3}}-1+\frac{2}{9v}\right)\approx 0=\mathbb{E}Z$$

3.6.3 Sample quantiles (central order statistics)

Let $X_1, \ldots, X_n$ be i.i.d. with CDF F and PDF $f = F'$. $X_{1:n} \le X_{2:n} \le \cdots \le X_{n:n}$ are order statistics.

Now let x_p satisfy $F(x_p) = p$, $f(x_p) > 0$, f continuous at x_p. Let k_n be sequence of integers such that $\frac{k_n}{n} \to p$ as $n \to \infty$ while $|k_n - np| \le C \ \forall n$ for some C. Then $X_{k_n:n}$ are referred to as central order statistics. For example, $k_n = [np]$ satisfies this and here $[x]$ is the integer part of x.

Theorem 3.6.1..

$$\sqrt{n}\left(X_{k_n:n} - x_p\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{p(1-p)}{f^2(x_p)}\right)$$

as $n \to \infty$, i.e. when n is large

$$X_{k_n:n} \overset{d}{\approx} \mathcal{N}\left(x_p, \frac{p(1-p)}{nf^2(x_p)}\right)$$

The following lemmas are needed to proof the theorem.

Lemma 3.6.1.. Let $U_1, U_2, \ldots, U_n$ be i.i.d. $\mathcal{U}nif(0,1)$, and $U_{1:n} \le \ldots \le U_{n:n}$ be order statistics. Then

$$\sqrt{n}\left(U_{k_n:n} - p\right) \xrightarrow{d} \mathcal{N}\left(0, p(1-p)\right) \quad \text{as } n \to \infty$$

Proof.

Step 1: Recall the marginal CDF of the kth order statistics is given by

$$G_k\left(y_k\right) = \sum_{j=k}^{n} \binom{n}{j} \left(F\left(y_k\right)\right)^j \left(1 - F\left(y_k\right)\right)^{n-j} \quad [1]$$

So

$$\mathbb{P}\{U_{k_n:n} \le w\} = \sum_{j=k_n}^{n} \binom{n}{j} w^j \left(1-w\right)^{n-j} \overset{2}{=} \mathbb{P}\{B \ge k_n\}$$

where $B \sim \mathcal{B}in(n, w)$.

[1] $\binom{n}{j} = \frac{n!}{j!(n-j)!}$ is the number of k-combinations

[2] Recall PMF of binomial distribution $B \sim \mathcal{B}in(n,p)$: $\mathbb{P}\{B = j\} = \binom{n}{j} p^j \left(1-p\right)^{n-j}$

Step 2:

$$\mathbb{P}\left\{\sqrt{n}\,(U_{k_n:n} - p) \le w\right\} = \mathbb{P}\Big\{U_{k_n:n} \le \underbrace{\frac{w}{\sqrt{n}} + p}_{:=p_n}\Big\} \overset{\text{Step 1}}{=} \mathbb{P}\{B_n \ge k_n\}$$

$$= \mathbb{P}\{\frac{B_n - np_n}{\sqrt{n}} \ge \frac{k_n - np_n}{\sqrt{n}}\} = \mathbb{P}\{\frac{B_n - np_n}{\sqrt{n}} \ge \frac{k_n - n\frac{w}{\sqrt{n}} - np}{\sqrt{n}}\}$$

$$= \mathbb{P}\{\frac{B_n - np_n}{\sqrt{n}} - \frac{k_n - np}{\sqrt{n}} \ge -w\}$$

where $B_n \sim \mathcal{B}in(n, p_n)$ and $p_n = \frac{w}{\sqrt{n}} + p$.

Recall $B_n = \sum_{i=1}^n X_i \sim \mathcal{B}in(n, p_n)$ with $X_i \sim \mathcal{B}in(1, p_n)$ and $\mathbb{E}X_i = p_n$, $\mathbb{V}arX_i = p_n(1 - p_n)$. Apply CLT on X_i

$$\frac{B_n - np_n}{\sqrt{np_n(1 - p_n)}} \xrightarrow{d} \mathcal{N}(0, 1) \quad \text{as } n \to \infty$$

then

$$\frac{B_n - np_n}{\sqrt{np(1 - p)}} \sqrt{\frac{p(1 - p)}{p_n(1 - p_n)}} \xrightarrow{d} \mathcal{N}(0, 1) \quad \text{as } n \to \infty$$

Because $\frac{p(1-p)}{p_n(1-p_n)} = \frac{p(1-p)}{\left(\frac{w}{\sqrt{n}}+p\right)\left(1-\frac{w}{\sqrt{n}}-p\right)} \to 1$ as $n \to \infty$ and then $\frac{p(1-p)}{p_n(1-p_n)} \xrightarrow{P} 1$ as $n \to \infty$, apply Corollary 3.4.1 of the continuous mapping theorem

$$\frac{B_n - np_n}{\sqrt{n}} \xrightarrow{d} \mathcal{N}(0, p(1 - p)) \quad \text{as } n \to \infty$$

Now since $\frac{k_n - np}{\sqrt{n}} \to 0$, apply Corollary 3.4.1 of the continuous mapping theorem again

$$\frac{B_n - np_n}{\sqrt{n}} - \frac{k_n - np}{\sqrt{n}} \xrightarrow{d} \mathcal{N}(0, p(1 - p)) \quad \text{as } n \to \infty$$

Step 3: Therefore let $Y = \frac{B_n - np_n}{\sqrt{n}} - \frac{k_n - np}{\sqrt{n}}$

$$\mathbb{P}\left\{\sqrt{n}\,(U_{k_n:n} - p) \le w\right\} = \mathbb{P}\{Y \ge -w\} = \mathbb{P}\{Y \le w\}$$

where $Y \sim \mathcal{N}(0, p(1 - p))$.

$\square$

Lemma 3.6.2.. Let X have CDF F where F is continuous and strictly increasing. Let $U \sim \mathcal{U}nif(0,1)$. Then $F^{-1}(U)$ (inverse function of F) has a CDF of F.

Proof.

$$\mathbb{P}\{F^{-1}(U) \leq x\} \stackrel{1}{=} \mathbb{P}\{F\left(F^{-1}(U)\right) \leq F(x)\} = \mathbb{P}\{U \leq F(x)\} \stackrel{2}{=} F(x)$$

$\square$

Proof of Theorem 3.6.1. If $X_1, \ldots, X_n$ are i.i.d. and have CDF F and $U_1, \ldots, U_n$ are i.i.d. $\mathcal{U}nif(0,1)$. By Lemma 3.6.2,

$$X_i \stackrel{d}{=} F^{-1}(U_i)$$

Hence

$$X_{1:n} \leq \cdots \leq X_{n:n} \stackrel{d}{=} F^{-1}(U_{1:n}) \leq \cdots \leq F^{-1}(U_{n:n})$$

Since $p = F(x_p)$, $x_p = F^{-1}(p)$,

$$\sqrt{n}\left(X_{k_n:n} - x_p\right) \stackrel{d}{=} \sqrt{n}\left(F^{-1}\left(U_{k_n:n}\right) - F^{-1}(p)\right)$$

Now by Lemma 3.6.1, $\sqrt{n}\left(U_{k_n:n} - p\right) \xrightarrow{d} \mathcal{N}\left(0, p(1-p)\right)$ as $n \to \infty$. Apply delta rule with $g(y) = F^{-1}(y)$,

$$\sqrt{n}\left(g\left(U_{k_n:n}\right) - g(p)\right) = \sqrt{n}\left(F^{-1}\left(U_{k_n:n}\right) - F^{-1}(p)\right)$$

$$\xrightarrow{d} \mathcal{N}\left(0, \left(g'(p)\right)^2 p(1-p)\right) \quad \text{as } n \to \infty$$

What is $g'(p) = \frac{dF^{-1}(y)}{dy}\Big|_{y=p}$? let $x = F^{-1}(y)$, then $y = F(x)$

$$1 = \frac{dy}{dy} = \frac{dF(x)}{dx}\frac{dx}{dy} = f(x)\frac{dx}{dy}$$

[1] F is strictly increasing
[2] Recall for $\mathcal{U}nif(0,1)$, $F(x) = x$

So

$$\frac{dF^{-1}(y)}{dy} = \frac{dx}{dy} = \frac{1}{f(x)} = \frac{1}{f\left(F^{-1}(y)\right)}$$

So

$$g'(p) = \frac{1}{f\left(F^{-1}(p)\right)} = \frac{1}{f(x_p)}$$

Therefore

$$\sqrt{n}\left(X_{k_n:n} - x_p\right) \stackrel{d}{=} \sqrt{n}\left(F^{-1}\left(U_{k_n:n}\right) - F^{-1}(p)\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{p(1-p)}{f^2(x_p)}\right)$$

as $n \to \infty$. $\square$

4

POINT ESTIMATION

4.1 Introduction

The objective of point estimation is to assign an appropriate value for unknown parameter based on observed data from the population by repeated trials of an experiment. It is assumed that the distribution of the population of interest can be represented by some PDF, $f(x;\theta)$, indexed by a parameter θ (could be a vector).

Definition 4.1.1. **(Parameter Space)**. The parameter space, Θ is the set of all possible values that the parameter θ could assume.

Example 4.1.1.. $X \sim \mathcal{N}(\mu, \sigma^2)$, $\Theta = \{(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 > 0\}$

Example 4.1.2.. $X \sim \mathcal{E}xp(\theta)$, $\Theta = \{\theta : \theta > 0\}$

Definition 4.1.2. **(Random sample)**. $X_1, \ldots, X_n$ is called a random sample from PDF $f(x;\theta)$ if the joint distribution of $X_1, \ldots, X_n$ is

$$f(x_1, \ldots, x_n; \theta) = f(x_1;\theta) \cdots f(x_n;\theta)$$

i.e. $X_1, \ldots, X_n$ are i.i.d. with $f(x;\theta)$.

The objective of point estimation can be also stated as follows. Based on a random sample $X_1, \ldots, X_n$ from $f(x;\theta)$, assign an appropriate value to θ.

Definition 4.1.3. **(Statistic)**. A function of $X_1, \ldots, X_n$, $T = t(X_1, \ldots, X_n)$, that does not depend on any unknown parameters, is called a statistic.

Example 4.1.3.. $X_1, \dots, X_n$ is a random sample from $\mathcal{N}(\mu, \sigma^2)$, with μ, σ^2 unknown. $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ and $S^2 = \sum_{i=1}^n \frac{(X_i - \overline{X}_n)^2}{n-1}$ are statistics. How about $\sqrt{n}\left(\frac{\overline{X}_n - \mu}{\sigma}\right)$, $\frac{(n-1)S^2}{\sigma^2}$?

Definition 4.1.4. **(Estimator; Estimate)**. A statistic $T = t(X_1, \dots, X_n)$ used to assign a value to a function $\tau(\theta)$ is called an estimator of $\tau(\theta)$. And an observed value of T, $t = t(x_1, \dots, x_n)$, where $x_1, \dots, x_n$ are the observed values of $X_1, \dots, X_n$, is called an estimate of $\tau(\theta)$.

4.2 Methods for Formulatory Estimators

4.2.1 Method of moments

Suppose $X_1, \dots, X_n$ is a random sample from $f(x; \theta_1, \dots, \theta_k)$, with $\theta_1, \dots, \theta_k$ unknown. Find $\mathbb{E}X^j = \mu_j^1(\theta_1, \dots, \theta_k)$[1], $j = 1, \dots, k$ and equate them with sample moments $m_j = \frac{1}{n}\sum_{i=1}^n X_i^j$. then solve for $\theta_1, \dots, \theta_k$. I.e. let

$$
\begin{aligned}
m_1 &= \mu_1^1(\hat{\theta}_1, \dots, \hat{\theta}_k) \\
&\vdots \\
m_k &= \mu_k^1(\hat{\theta}_1, \dots, \hat{\theta}_k)
\end{aligned}
$$

Solve for $\hat{\theta}_j = t_j(X_1, \dots, X_n)$[2], $1 \le j \le k$, as the MM estimators of $\theta_1, \dots, \theta_k$.

Example 4.2.1.. Let $X_1, \dots, X_n$ be a random sample from $\mathcal{N}(\mu, \sigma^2)$. We have

$$\mathbb{E}X_1 = \mu, \quad \mathbb{E}X_1^2 = \mu^2 + \sigma^2$$

$$m_1 = \overline{X}, \quad m_2 = \frac{1}{n}\sum_{i=1}^n X_i^2$$

[1] $\mu_j^1(\theta_1, \dots, \theta_k)$ is a function in terms of $\theta_1, \dots, \theta_k$

[2] $t_j(X_1, \dots, X_n)$ is a function in terms of $X_1, \dots, X_n$

Set

$$\overline{X}_n = \hat{\mu}, \quad \frac{1}{n}\sum_{i=1}^{n} X_i^2 = \hat{\mu}^2 + \hat{\sigma}^2$$

which implies

$$\hat{\mu} = \overline{X}_n \qquad \text{MM estimator of } \mu$$

and

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\overline{X}_n\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2 \qquad \text{MM estimator of } \sigma^2$$

$\boxed{\textit{Example 4.2.2.}}$. Let $X_1, \ldots, X_n$ be a random sample from

$$f(x;\theta,\eta) = \begin{cases} \frac{1}{\eta} e^{-\left(\frac{x-\theta}{\eta}\right)}, & x > \theta \\ 0, & \text{otherwise} \end{cases}$$

we have

$$\mathbb{E}X_1 = \int_\theta^\infty \frac{x}{\eta} e^{-\left(\frac{x-\theta}{\eta}\right)}\, dx \overset{1}{=} \eta \underbrace{\int_0^\infty y e^{-y}\, dy}_{=\Gamma(2)=1} + \theta \underbrace{\int_0^\infty e^{-y}\, dy}_{=1} = \eta + \theta$$

$$\mathbb{E}X_1^2 = \int_\theta^\infty \frac{x^2}{\eta} e^{-\left(\frac{x-\theta}{\eta}\right)}\, dx$$

$$= \eta^2 \underbrace{\int_0^\infty y^2 e^{-y}\, dy}_{=\Gamma(3)=2} + \theta^2 \underbrace{\int_0^\infty e^{-y}\, dy}_{=1} + 2\eta\theta \underbrace{\int_0^\infty y e^{-y}\, dy}_{=\Gamma(2)=1} = (\theta+\eta)^2 + \eta^2$$

$$m_1 = \overline{X}_n, \qquad m_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

Set

$$\overline{X}_n = \hat{\eta} + \hat{\theta}, \quad \frac{1}{n}\sum_{i=1}^{n} X_i^2 = (\hat{\theta} + \hat{\eta})^2 + \hat{\eta}^2$$

which implies

$$\hat{\eta} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}_n^2}, \quad \hat{\theta} = \overline{X}_n - \sqrt{\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}_n^2}$$

[1] Change-of-variable: $y = \frac{x-\theta}{\eta}$

4.2.2 Maximum likelihood

Definition 4.2.1. **(Likelihood function)**. Let $X_1, \ldots, X_n$ have joint PDF (or PMF) $f(x_1, \ldots, x_n; \theta)$. The likelihood function is $L(\theta; x_1, \ldots, x_n) = f(x_1, \ldots, x_n; \theta)$, which is viewed as function of θ for fixed $x_1, \ldots, x_n$.

Algebraically, the likelihood function is just the same as the joint PDF or joint PMF, but its meaning is quite different. A PDF or PMF is a function of $x_1, \ldots, x_n$ with θ fixed. A likelihood function, on the other hand, is a function of θ with fixed $x_1, \ldots, x_n$.

If $X_1, \ldots, X_n$ is a random sample from $f(x; \theta)$, then

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

Definition 4.2.2. **(Maximum likelihood estimator (MLE))**. Let $X_1, \ldots, X_n$ have joint PDF (or PMF) $f(x_1, \ldots, x_n; \theta)$. For a given observation $(X_1 = x_1, \ldots, X_n = x_n)$, a value $\hat{\theta}$, $\hat{\theta} \in \Theta$ at which $L(\theta; x_1, \ldots, x_n)$ is a maximum is called a maximum likelihood estimate of θ. I.e. $\hat{\theta}$ satisfies

$$f(x_1, \ldots, x_n; \hat{\theta}) = \max_{\theta \in \Theta} f(x_1, \ldots, x_n; \theta)$$

If each distinct value of $x_1, \ldots, x_n$ produces one $\hat{\theta}$, then this procedure define a function $\hat{\theta}(x_1, \ldots, x_n)$. This function, when applied to the random variables X_i's, $\hat{\theta}(X_1, \ldots, X_n)$ is called the maximum likelihood estimator (MLE) of θ.

Example 4.2.3.. Let $X_1, \ldots, X_n$ be i.i.d. $\mathcal{E}xp(\theta)$. Suppose $x_1, \ldots, x_n$ are observed set of date, we have

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

For fixed $x_1, \ldots, x_n$, $L(\theta; x_1, \ldots, x_n)$ is continuous and smooth function of θ. To maximize $L(\theta; x_1, \ldots, x_n)$, we may as well maximize log-likelihood

$$\ln L(\theta; x_1, \ldots, x_n) = -n \ln \theta - \frac{1}{\theta} \sum_{i=1}^{n} x_i$$

Take derivative with respect to θ and set it to 0

$$\frac{\partial \ln L(\theta; x_1, \ldots, x_n)}{\partial \theta} = -\frac{n}{\theta} + \frac{1}{\theta^2} \sum_{i=1}^{n} x_i = 0$$

which implies

$$\theta = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Check the second derivative

$$\frac{\partial^2 \ln L(\theta; x_1, \ldots, x_n)}{\partial \theta^2}\Big|_{\theta=\frac{1}{n}\sum_{i=1}^{n} x_i} = \frac{n}{\theta^2} - \frac{2}{\theta^3} \sum_{i=1}^{n} x_i \Big|_{\theta=\frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x}} = \frac{-n}{\overline{x}^2} < 0$$

So $\theta = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the θ maximizes the likelihood function. Therefore, $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is a maximum likelihood estimate and then $\hat{\theta}(X_1, \ldots, X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i$ is the maximum likelihood estimator (MLE).

Example 4.2.4. **(Practical example of MLE for exponential distribution)**. Let $X_1, \ldots, X_{10}$ be random sample from $\mathcal{E}xp(\theta)$, θ is unknown. Assume we observed a set of data:

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
2574	1310	282	1233	1925	135	281	2254	671	495

- The ML estimate is $\hat{\theta} = \overline{x} = 1116$
- The ML estimator is $\hat{\theta} = \frac{1}{10}\sum_{i=1}^{10} X_i$

Example 4.2.5.. Let $X_1, \ldots, X_n$ be i.i.d. $\mathcal{U}nif(0, \theta)$.

1. MM: $\mathbb{E}X_i = \theta/2$, set $\hat{\theta}/2 = \overline{X}_n$ we have $\hat{\theta} = 2\overline{X}_n$ as the MM estimator of θ.

2. MLE:

$$f(x;\theta) = \begin{cases} \frac{1}{\theta}, & 0 \leq x \leq \theta \\ 0, & \text{otherwise} \end{cases}$$

$$L(\theta; x_1, \ldots, x_n) = \begin{cases} \frac{1}{\theta^n}, & 0 \leq \min\limits_{1 \leq i \leq n} x_i \leq \max\limits_{1 \leq i \leq n} x_i \leq \theta \\ 0, & \text{otherwise} \end{cases}$$

Because the maximum of $L(\theta; x_1, \ldots, x_n)$ is obtained when $\theta = \max\limits_{1 \leq i \leq n} x_i$, the ML estimator is $\hat{\theta} = \max\limits_{1 \leq i \leq n} X_i$.