# MATH 3423: Statistical Inference

HU-HTAKM

Website: **https://htakm.github.io/htakm_test/**

Last major change: November 28, 2024
Last small update: November 28, 2024

This is MATH 3423 lecture note made by me. Note that some of the notations are slightly different with what is used in the course for better clarity. For example:

| Name | My notation | Ycw's notation |
|---|---|---|
| Transpose | $\mathbf{A}^T$ | $\mathbf{A}'$ |
| Set of real numbers | $\mathbb{R}$ | $R$ |
| Fisher Information | $\mathcal{I}_X(\theta)$ | $I_X(\theta)$ |
| Convergence in distribution | $X_n \xrightarrow{D} X$ | $X_n \xrightarrow{d} X$ |
| Probability | $\mathbb{P}$ | $P$ |
| Expected value | $\mathbb{E}$ | $E$ or $E_X$ |

# Contents

# Chapter 1

# Preliminary

Statistical inference is a statistical process which investigates how to use the information from the data to make an inference about the distribution of the random variable of our interest. In MATH 3423, we will focus on two core concepts of statistical inference: point estimation and hypothesis testing.

## 1.1    Random sample and parametric distribution

To make a statistical inference about the distribution of the random variable $X$ of our interest, we need to draw a sample of data.

**Definition 1.1.** Denote the first observation by $X_1$, the second by $X_2$, and so on. A set of random variables $\{X_1, \cdots, X_n\}$ are called a **random sample** of size $n$ from the common distribution of $X$ with a PMF $p_X(x)$ or PDF $f_X(x)$ if they are independent and identically distributed (i.i.d.).

**Remark 1.1.1.** Random variables $X_1, \cdots, X_n$ are assumed to be observable with known actual values $x_1, \cdots, x_n$ respectively.

Under the random sampling setting, it is easy to get the following lemma.

**Lemma 1.2.** Given a random sample $\{X_1, \cdots, X_n\}$ of a common distribution $X$.

1. If the random sample is discrete with a common PMF $p_X(x)$, then the joint PMF of random sample can be obtained by:
$$p_{X_1,\cdots,X_n}(x_1, \cdots, x_n) = \prod_{i=1}^{n} p_{X_i}(x_i)$$

2. If the random sample is continuous with a common PDF $f_X(x)$, then the joint PDF of random sample can be obtained by:
$$f_{X_1,\cdots,X_n}(x_1, \cdots, x_n) = \prod_{i=1}^{n} f_{X_i}(x_i)$$

The underlying distribution of $X$ is assumed to be unknown or partially known in practice. In most situation, it is reasonable to assume that the form of the PMF $p_X$ or PDF $f_X$ of the distribution is known but it contains some unknown parameters $\theta$.

**Definition 1.3. Parametric distribution** is a distribution which the PMF $p_X$ and PDF $f_X$ contains some unknown parameters $\theta$. The PMF or PDF is said to be **parametric**.

**Remark 1.3.1.** Instead of using parametric distributions of the data, we may assume that the form of the distribution is unknown but that the distribution has some properties. E.g. A distribution is continuous. This distribution is called a **non-parametric distribution** and the corresponding statistical method is called **non-parametric statistical approach**. If parameters are involved but the form of the distribution is unknown, then such a distribution is called **semi-parametric distribution** and the corresponding method is called **semi-parametric statistical approach**.

**Example 1.1.** Data are usually assumed from the normal distribution with mean $\mu$ and variance $\sigma^2$, where the parameter:
$$\theta = \begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix}$$
is unknown but fixed.

**Lemma 1.4.** Given a random sample $\{X_1, \cdots, X_n\}$ of a common distribution $X$.

1. If the random sample is discrete with a common parametric PMF $p_X(x|\theta)$, then the joint PMF of random sample can be obtained by:
$$p_{X_1, \cdots, X_n}(x_1, \cdots, x_n|\theta) = \prod_{i=1}^{n} p_X(x_i|\theta)$$

2. If the random sample is continuous with a common parametric PDF $p_X(x|\theta)$, then the joint PDF of random sample can be obtained by:
$$f_{X_1, \cdots, X_n}(x_1, \cdots, x_n|\theta) = \prod_{i=1}^{n} f_X(x_i|\theta)$$

Under parametric setting, we can see that uncertainty of the distribution is the uncertainty of its parameters. One of the central problem in statistics is how to determine which function of data is the best one to estimate $\theta$.

**Definition 1.5.** If $\mathbf{X} = (X_1 \ \cdots \ X_n)^T$ is a random vector and $T(\cdot)$ is a real-valued or vector-valued function such that for all $\mathbf{X} \in \Omega$, $T(\mathbf{X})$ does not contain any unknown parameters, then $T(\mathbf{X})$ is called a **statistic**.

1. If we use the statistic $T(\mathbf{X})$ to estimate an unknown parameter $\theta$, then we call $T(\mathbf{X})$ and $T(\mathbf{x})$ an **estimator** and an **estimate** of $\theta$ respectively, where $\mathbf{x}$ is an observed value of $\mathbf{X}$.

**Remark 1.5.1.** We usually denote an estimator of $\theta$ by $\hat{\theta}(\mathbf{X})$ or simply $\hat{\theta}$.

**Remark 1.5.2.** Since $T(\mathbf{X})$ is also random, we have a distribution of $T(\mathbf{X})$ called **sampling distribution**.

## 1.2   Moments

The population moments of a distribution play an important role in theoretical and applied statistics.

**Definition 1.6.** For each positive integer $k$, the $k$-**th population moment** of $X$ about 0, denoted by $\mu'_k$, is defined as:
$$\mu'_k = \mathbb{E}(X^k)$$
if the expectation exists.
When $k = 1$, $\mu'_1$ is the population mean $\mu = E(X)$ of $X$.

**Definition 1.7.** The $k$-**th population central moment** of $X$, denoted by $\mu_k$, is defined as:
$$\mu_k = \mathbb{E}(X - \mu)^k$$
if the expectation exists.
When $k = 2$, $\mu_2$ is the **population variance** $\sigma^2$ of $X$.

**Remark 1.7.1.** Don't confuse the population mean $\mu$ and the $k$-th population central moment $\mu_k$!

**Example 1.2.** We have terminologies for some useful population moments.

1. **Skewness**: $\mu_3$, a measure of symmetry or skewness.

    (a) If $\mu_3 < 0$, then the curve is left-skewed (tail is on the left).

    (b) If $\mu_3 > 0$, then the curve is right-skewed (tail is on the right).

    (c) If $\mu_3 = 0$, the the curve is symmetrical.

    The ratio $\frac{\mu_3}{\sigma^3}$ is called the **coefficient of skewness**.

2. **Kurtosis**: $\mu_4$, a measure of excess or kurtosis, which is the degree of flatness of a density near its center. We called $\frac{\mu_4}{\sigma^4} - 3$ the **coefficient of kurtosis**.

    (a) If $\frac{\mu_4}{\sigma^4} - 3 > 0$, then the density has a sharper peak than the density of the normal curve.

    (b) If $\frac{\mu_4}{\sigma^4} - 3 < 0$, then the density has a flatter peak than the density of the normal curve.

We usually use sample moments to estimate the population moments.

**Definition 1.8.** Let $X_1, \cdots, X_n$ be a random sample of size $n$. For each positive integer $k$,

1. the $k$-**th sample moment** about 0, denoted by $\overline{X^k}$, is defined as:

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

2. the $k$-**th sample moment** about $\overline{X}$, denoted by $S_n^k$, is defined as:

$$S_n^k = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X})^k$$

**Example 1.3.** When $k = 1$, $\overline{X}$ is called the **sample mean** of $X$. When $k = 2$, $S_n^2$ is called the **sample variance**. However, we usually don't use this version of sample variance. Instead, we use another sample variance, denoted by $S_{n-1}^2$, which is defined by:

$$S_{n-1}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$$

We do this because $S_{n-1}^2$ is unbiased while $S_n^2$ is not unbiased.

**Lemma 1.9.** Let $X_1, \cdots, X_n$ be a random sample of size $n$. We have:

$$\mathbb{E}(\overline{X^k}) = \mu_k'$$

if $\mu_k'$ exists. We also have:

$$\mathrm{Var}(\overline{X^k}) = \frac{1}{n}[\mu_{2k}' - (\mu_k')^2]$$

*Proof.*
Since $X_1, \cdots, X_n$ have the same distribution, we have:

$$\mathbb{E}(X_1^k) = \cdots = \mathbb{E}(X_n^k) = \mathbb{E}(X^k) = \mu_k'$$

Therefore, we have:

$$\mathbb{E}(\overline{X^k}) = \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^{n} X_i^k\right] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}(X_i^k) = \frac{n\mu_k'}{n} = \mu_k'$$

Since $X_1^k, \cdots, X_n^k$ are independent,

$$\mathrm{Var}(\overline{X^k}) = \mathrm{Var}\left(\frac{1}{n} \sum_{i=1}^{n} X_i^k\right) = \frac{1}{n^2} \mathrm{Var}\left(\sum_{i=1}^{n} X_i^k\right) = \frac{1}{n^2} \sum_{i=1}^{n} \mathrm{Var}(X_i^k) = \frac{1}{n^2} \sum_{i=1}^{n} \left[\mathbb{E}(X_i^{2k}) - (\mathbb{E}(X_i^k)^2)\right]$$

$$= \frac{1}{n}[\mu_{2k}' - (\mu_k')^2]$$

$\square$

## 1.3    Conditional distribution

Sometimes, we would deal with cases when a certain information is given.

**Definition 1.10.** Suppose that $X$ and $Y$ are two random variables. The **conditional distribution function** of $Y$ given $X = x$ for any $x$ such that PMF $p_X(x) > 0$ or PDF $f_X(x) > 0$ is defined by:

$$F_{Y|X}(y|x) = \mathbb{P}(Y \le y | X = x)$$

The **conditional PDF/PMF** of $Y$ given $X = x$ for any $x$ such that PMF $p_X(x) > 0$ or PDF $f_X(x) > 0$ is defined by:

$$\begin{cases} p_{Y|X}(y|x) = \frac{p_{X,Y}(x,y)}{p_X(x)}, & \text{Discrete case} \\ f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)}, & \text{Continuous case} \end{cases}$$

The conditional distribution has its corresponding expectation.

**Definition 1.11.** Suppose that $X$ and $Y$ are two random variables. The **conditional expectation** of $Y$ given $X = x$ for any $x$ such that PMF $p_X(x) > 0$ or PDF $f_X(x) > 0$ is defined by:

$$\mathbb{E}(Y|X = x) = \begin{cases} \sum_y y p_{Y|X}(y|x), & \text{Discrete case} \\ \int_{-\infty}^{\infty} y f_{Y|X}(y|x)\,dy, & \text{Continuous case} \end{cases}$$

**Remark 1.11.1.** $\mathbb{E}(Y|X = x)$ is a function of $x$. Similarly, $\mathbb{E}(Y|X)$ is a function of $X$.

**Example 1.4.** Suppose that the joint PDF of $X$ and $Y$ is given by:

$$f_{X,Y}(x, y) = \begin{cases} \frac{1}{y}e^x, & x > 0, y > 0 \\ 0, & \text{Otherwise} \end{cases}$$

We want to compute $\mathbb{E}(X|Y = y)$. We may find that:

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\,dx = e^{-y} \int_{-\infty}^{\infty} \frac{1}{y} e^{-\frac{x}{y}}\,dx = e^{-y} \qquad f_{X|Y}(x|y) = \frac{f_{X,Y}(x, y)}{f_Y(y)} = \frac{1}{y} e^{-\frac{x}{y}}$$

We may see that $(X|Y = y) \sim \text{Exp}(\frac{1}{y})$. Therefore, $\mathbb{E}(X|Y = y) = y$.

Conditional expectation has the following properties.

**Lemma 1.12.** Suppose that $X$, $Y$ and $Z$ are three random variables. The conditional expectation has the following properties:

1. $\mathbb{E}(aY + bZ|X) = a\,\mathbb{E}(Y|X) + b\,\mathbb{E}(Z|X)$ for $a, b \in \mathbb{R}$.

2. $\mathbb{E}(Y|X) \ge 0$ if $Y \ge 0$.

3. If $X$ and $Y$ are independent, then $\mathbb{E}(Y|X) = \mathbb{E}(Y)$.

*Proof.*
Proof for discrete case is similar to continuous case.

1.

$$\mathbb{E}(aY + bZ|X) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (ay + bz) f_{Y,Z|X}(y, z|X)\,dy\,dz$$

$$= a \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f_{Y,Z|X}(y, z|X)\,dy\,dz + b \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} z f_{Y,Z|X}(y, z|X)\,dy\,dz$$

$$= a \int_{-\infty}^{\infty} y f_{Y|X}(y|X)\,dy + b \int_{-\infty}^{\infty} z f_{Z|X}(z|X)\,dz = a\,\mathbb{E}(Y|X) + b\,\mathbb{E}(Z|X)$$

2. If $Y \ge 0$, then since $f_{Y|X}(y|x) \ge 0$ for any $x$ such that $f_X(x) > 0$:

$$\mathbb{E}(Y|X) = \int_0^{\infty} y f_{Y|X}(y|X)\,dy \ge 0$$

3. If $X$ and $Y$ are independent, then:

$$\mathbb{E}(Y|X) = \int_{-\infty}^{\infty} y f_{Y|X}(y|X)\,dy = \int_{-\infty}^{\infty} y f_Y(y)\,dy = \mathbb{E}(Y)$$

$\square$

If $\mathbb{E}(Y|X)$ is a function of $X$, then what is its expectation?

**Theorem 1.13.** (**Law of total expectation**) Given two random variables $X$ and $Y$. We have:

$$\mathbb{E}(Y) = \mathbb{E}(\mathbb{E}(Y|X))$$

if both expectations exist.

*Proof.*
We may proof in continuous case. Discrete case works similarly.

$$\mathbb{E}(\mathbb{E}(Y|X)) = \int_{-\infty}^{\infty} \mathbb{E}(Y|x) f_X(x)\,dx = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f_{Y|X}(y|x) f_X(x)\,dy\,dx$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{Y|X}(y|x) f_X(x)\,dx\,dy$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx\,dy$$

$$= \int_{-\infty}^{\infty} f_Y(y)\,dy = \mathbb{E}(Y)$$

$\square$

The following theorem is the generalization of Law of total expectation. We will omit the proof.

**Lemma 1.14.** Given two random variables $X$ and $Y$. For any function $g$, we have:

$$\mathbb{E}(\mathbb{E}(Y|X)g(X)) = \mathbb{E}(Yg(X))$$

if both expectations exist.

Similarly, we have a conditional variance.

**Definition 1.15.** Given two random variables $X$ and $Y$. The **conditional variance** of $Y$ given $X$ is defined by:
$$\text{Var}(Y|X) = \mathbb{E}\left[(Y - \mathbb{E}(Y|X))^2|X\right]$$

**Lemma 1.16.** Given two random variables $X$ and $Y$. We have:

$$\text{Var}(Y|X) = \mathbb{E}(Y^2|X) - [\mathbb{E}(Y|X)]^2$$

*Proof.*
By Lemma 1.12 and Lemma 1.14,

$$\text{Var}(Y|X) = \mathbb{E}\left[(Y - \mathbb{E}(Y|X))^2|X\right]$$
$$= \mathbb{E}(Y^2|X) - 2\mathbb{E}(Y\mathbb{E}(Y|X)|X) + \mathbb{E}((\mathbb{E}(Y|X))^2|X)$$
$$= \mathbb{E}(Y^2|X) - [\mathbb{E}(Y|X)]^2 \qquad\qquad (\mathbb{E}(Y|X) \text{ is a function of } X)$$

$\square$

We have the law of total variance.

**Theorem 1.17.** (**Law of total variance**) Given two random variables $X$ and $Y$. We have:

$$\text{Var}(Y) = \mathbb{E}[\text{Var}(Y|X)] + \text{Var}[\mathbb{E}(Y|X)]$$

if the expectation and variances exist.

*Proof.*
By Lemma 1.16 and Law of total expectation,

$$\mathbb{E}[\text{Var}(Y|X)] + \text{Var}[\mathbb{E}(Y|X)] = \mathbb{E}[\mathbb{E}(Y^2|X)] - \mathbb{E}[\mathbb{E}(Y|X)]^2 + \mathbb{E}[\mathbb{E}(Y|X)]^2 - [\mathbb{E}[\mathbb{E}(Y|X)]]^2$$
$$= \mathbb{E}(Y^2) - [\mathbb{E}(Y)]^2 = \text{Var}(Y)$$

$\square$

## 1.4   Commonly used distribution

Let us recall some useful distributions.

---

**Example 1.5. (Binomial distribution)** $X \sim \text{Bin}(n, p)$
Random variable $X$ is a Binomial random variable with parameters $n \in \mathbb{N}$ and $p \in [0, 1]$ if it has a PMF for $x = 0, \cdots, n$:

$$p_X(x) = \binom{n}{x} p^x (1-p)^{n-x} \qquad \mathbb{E}(X) = np \qquad \text{Var}(X) = np(1-p)$$

---

**Example 1.6. (Geometric distribution)** $X \sim \text{Geom}(p)$
Random variable $X$ is geometric with parameter $p \in [0, 1]$ if it has a PMF for $x = 1, 2, \cdots$:

$$p_X(x) = p(1-p)^{x-1} \qquad \mathbb{E}(X) = \frac{1}{p} \qquad \text{Var}(X) = \frac{1-p}{p^2}$$

---

**Example 1.7. (Poisson distribution)** $X \sim \text{Poisson}(\lambda)$
Random variable $X$ is a Poisson random variable with parameter $\lambda$ if it has a PMF for $x = 0, 1, \cdots$:

$$p_X(x) = \frac{\lambda^x}{x!} e^{-\lambda} \qquad \mathbb{E}(X) = \lambda \qquad \text{Var}(X) = \lambda$$

---

**Example 1.8. (Negative Binomial distribution)** $X \sim \text{NBin}(r, p)$
Assume that $X_1, \cdots, X_r$ are independent and $X_i \sim \text{Geom}(p)$ for $i = 1, \cdots, r$. Let $Y = \sum_{i=1}^{n} X_i$. Random variable $Y$ is negative Binomial with parameters $r > 0$ and $p \in [0, 1]$ if for $x > r$:

$$p_X(x) = \binom{x-1}{r-1} (1-p)^{x-r} p^r \qquad \mathbb{E}(X) = \frac{r}{p} \qquad \text{Var}(X) = \frac{r(1-p)}{p^2}$$

---

**Example 1.9. (Cauchy distribution)** $X \sim \text{Cauchy}(\theta)$
Random variable $X$ is a Cauchy random variable with parameter $\theta$ if it has a PDF:

$$f_X(x) = \frac{1}{\pi(1 + (x-\theta)^2)} \qquad \mathbb{E}(X) \text{ DNE} \qquad \text{Var}(X) \text{ DNE}$$

---

**Example 1.10. (Uniform distribution)** $X \sim \text{U}[a, b]$
Random variable $X$ is uniform if given $a < b$, it has a PDF:

$$f_X(x) = \begin{cases} \frac{1}{b-a}, & x \in [a, b] \\ 0, & \text{Otherwise} \end{cases} \qquad \mathbb{E}(X) = \frac{a+b}{2} \qquad \text{Var}(X) = \frac{(b-a)^2}{12}$$

---

**Example 1.11. (Exponential distribution)** $X \sim \text{Exp}(\lambda)$
Random variable $X$ is exponential with parameter $\lambda$ if it has a PDF:

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases} \qquad \mathbb{E}(X) = \frac{1}{\lambda} \qquad \text{Var}(X) = \frac{1}{\lambda^2}$$

---

**Example 1.12. (Normal distribution / Gaussian distribution)** $X \sim \text{N}(\mu, \sigma^2)$
Random variable $X$ is normal if it has two parameters $\mu$ and $\sigma^2$, and its PDF and CDF is:

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \qquad F_X(x) = \int_{-\infty}^{x} f_X(u)\, du \qquad \mathbb{E}(X) = \mu \qquad \text{Var}(X) = \sigma^2$$

Random variable $X$ is standard normal if $\mu = 0$ and $\sigma^2 = 1$. $(X \sim \text{N}(0, 1))$

$$f_X(x) = \phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \qquad F_X(x) = \Phi(x) = \int_{-\infty}^{x} \phi(u)\, du \qquad \mathbb{E}(X) = 0 \qquad \text{Var}(X) = 1$$

**Example 1.13.** (**Gamma distribution**) $X \sim \text{Gamma}(\alpha, \beta)$
Random variable $X$ is a gamma random variable with parameters $\alpha > 0$ and $\beta > 0$ if its PDF is:

$$f_X(x) = \begin{cases} \frac{1}{\Gamma(\alpha)} \beta^\alpha x^{\alpha-1} e^{-\beta x}, & x \geq 0 \\ 0, & \text{Otherwise} \end{cases} \qquad \mathbb{E}(X) = \frac{\alpha}{\beta} \qquad \text{Var}(X) = \frac{\alpha}{\beta^2}$$

**Remark 1.17.1.** Gamma function $\Gamma(z)$ has the following properties:

1. If $z$ is a positive integer, then $\Gamma(z) = (z-1)!$.

2. For all $z$, $\Gamma(z+1) = z\Gamma(z)$.

3. If $\Re(z) > 0$, then $\Gamma(z) = \int_0^\infty t^{z-1} e^{-t}\, dt$.

4. $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.

**Example 1.14.** (**Beta distribution**) $X \sim \text{Beta}(\alpha, \beta)$
Random variable $X$ is a beta random variable with parameters $\alpha > 0$ and $\beta > 0$ if its PDF is:

$$f_X(x) = \begin{cases} \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}, & x \in (0,1) \\ 0, & \text{Otherwise} \end{cases} \qquad \mathbb{E}(X) = \frac{\alpha}{\alpha+\beta} \qquad \text{Var}(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

**Remark 1.17.2.** Beta function $B(z_1, z_2)$ has the following properties:

1. $B(z_1, z_2) = \frac{\Gamma(z_1)\Gamma(z_2)}{\Gamma(z_1+z_2)}$

2. $B(z_1, z_2) = \int_0^1 t^{z_1-1}(1-t)^{z_2-1}\, dt$

We have some more distributions that are associated with normal distribution. For example, Chi-squared distribution, which is a special case of gamma distribution.

**Example 1.15.** (**Chi-squared distribution**) $Y \sim \chi^2(n)$
Assume that $X_1, X_2, \cdots, X_n$ are independent and $X_i \sim \text{N}(0,1)$ for $i = 1, \cdots, n$. Let $Y = \sum_{i=1}^n X_i^2$. Random variable $Y$ has a $\chi^2$-distribution with $n$ degree of freedom if:

$$f_Y(x) = \begin{cases} \frac{1}{\Gamma(\frac{n}{2})} 2^{-\frac{n}{2}} x^{\frac{n}{2}-1} e^{-\frac{x}{2}}, & x > 0 \\ 0, & \text{Otherwise} \end{cases} \qquad \mathbb{E}\, Y = n \qquad \text{Var}(Y) = 2n$$

**Theorem 1.18.** If the random variable $X \sim \text{N}(\mu, \sigma^2)$, where $\sigma^2 > 0$, then the random variable $V = \frac{(X-\mu)^2}{\sigma^2} \sim \chi^2(1)$.

*Proof.*
By the properties of normal distribution, we get that:

$$\frac{X-\mu}{\sigma} \sim \text{N}(0,1)$$

Therefore, by the definition of chi-squared distribution,

$$V = \left(\frac{X-\mu}{\sigma}\right)^2 \sim \chi^2(1)$$

$\square$

**Theorem 1.19.** Given a set of random variables $\{X_1, \cdots, X_k\}$. Let $Y = \sum_{i=1}^k X_i$ and $X_i \sim \chi^2(r_i)$ for all $i = 1, \cdots, k$. If they are independent, then $Y \sim \chi^2(r_1 + \cdots + r_k)$.

*Proof.*
It suffices to prove the relation for two random variables $Z_1$ and $Z_2$. If $Z_1 \sim \chi^2(n_1)$ and $Z_2 \sim \chi^2(n_2)$, then $Z_1 + Z_2 \sim \chi^2(n_1 + n_2)$. By repeating applying the relation for two random variables, one can easily have the desired relation for $k$ random variables.
From definition, $Z_1 = X_{11}^2 + \cdots + X_{1n_1}^2$ and $Z_2 = X_{21}^2 + \cdots + X_{2n_2}^2$, where $X_{1i} \sim \text{N}(0,1)$ and $X_{2j} \sim \text{N}(0,1)$ for $i = 1, \cdots, n_1$ and $j = 1, \cdots, n_2$. Therefore,

$$Z_1 + Z_2 = X_{11}^2 + \cdots + X_{1n_1}^2 + X_{21}^2 + \cdots + X_{2n_2}^2 \sim \chi^2(n_1 + n_2)$$

$\square$

**Theorem 1.20.** If $\{X_1, \cdots, X_n\}$ is a random sample of size $n > 1$ of a random variable $X \sim N(\mu, \sigma^2)$, then we have:

1. Sample mean $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$

2. Sample mean $\overline{X}$ and sample variance $S_{n-1}^2$ are independent.

3.
$$\frac{(n-1)S_{n-1}^2}{\sigma^2} = \frac{nS_n^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^n (X_i - \overline{X})^2 \sim \chi^2(n-1)$$

*Proof.*

1. From definition,
$$\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$$

   Since $X_i \sim N(\mu, \sigma^2)$ for $i = 1, \cdots, n$, we can find that $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$.

2. Let $\mathbf{X} = (X_1 \ \cdots \ X_n)^T$. We may find that:

$$\begin{pmatrix} \overline{X} \\ X_1 - \overline{X} \\ X_2 - \overline{X} \\ \vdots \\ X_n - \overline{X} \end{pmatrix} = \begin{pmatrix} \frac{1}{n}X_1 + \frac{1}{n}X_2 + \cdots + \frac{1}{n}X_n \\ \left(1 - \frac{1}{n}\right)X_1 - \frac{1}{n}X_2 - \cdots - \frac{1}{n}X_n \\ -\frac{1}{n}X_1 + \left(1 - \frac{1}{n}\right)X_2 - \cdots - \frac{1}{n}X_n \\ \vdots \\ -\frac{1}{n}X_1 - \frac{1}{n}X_2 - \cdots + \left(1 - \frac{1}{n}\right)X_n \end{pmatrix} = \mathbf{AX} \qquad \mathbf{A} = \begin{pmatrix} \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} \\ 1 - \frac{1}{n} & -\frac{1}{n} & \cdots & -\frac{1}{n} \\ -\frac{1}{n} & 1 - \frac{1}{n} & \ddots & \vdots \\ \vdots & \ddots & \ddots & -\frac{1}{n} \\ -\frac{1}{n} & \cdots & -\frac{1}{n} & 1 - \frac{1}{n} \end{pmatrix}$$

   By Lemma 1.37, we have $\mathbf{AX} \sim N_{n+1}(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\sigma^2\mathbf{I_{n\times n}}\mathbf{A}^T)$, where $\boldsymbol{\mu} = (\mu \ \cdots \ \mu)^T$.
   Let $\mathbf{X}^* = (X_1 - \overline{X} \ \cdots \ X_n - \overline{X})^T$ and $\boldsymbol{\Sigma}^*$ be the variance-covariance matrix of $\widetilde{\mathbf{X}}$. We may notice that:

$$\mathbf{A}\sigma^2\mathbf{I_{n\times n}}\mathbf{A}^T = \left( \begin{array}{c|c} \text{Var}(\overline{X}) & \text{cov}(\mathbf{X}^*, \overline{X}) \\ \hline \text{cov}(\mathbf{X}^*, \overline{X}) & \boldsymbol{\Sigma}^* \end{array} \right)$$

   We prove that $\text{cov}(X_i - \overline{X}, \overline{X}) = 0$ for $i = 1, \cdots, n$. Since $X_i$ are independent for all $i$,

$$\text{cov}(X_i - \overline{X}, \overline{X}) = \text{cov}(X_i, \overline{X}) - \text{Var}(\overline{X}) = \frac{1}{n}\text{Var}(X_i) - \frac{\sigma^2}{n} = 0$$

   Therefore, we can find that $\text{cov}(\mathbf{X}^*, \overline{X}) = 0$. By Lemma 1.38, $\overline{X}$ and $\mathbf{X}^*$ are independent.
   Since $S_n^2$ is a function of $\mathbf{X}^*$, we can conclude that $\overline{X}$ and $S_{n-1}^2$ are independent.

3. We have:

$$\frac{1}{\sigma^2}\sum_{i=1}^n (X_i - \overline{X})^2 = \frac{1}{\sigma^2}\sum_{i=1}^n (X_i - \mu + \mu - \overline{X})^2 = \frac{1}{\sigma^2}\sum_{i=1}^n (X_i - \mu)^2 - \frac{n(\overline{X} - \mu)^2}{\sigma^2} = \sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma}\right) - \left(\frac{\sqrt{n}(\overline{X} - \mu)}{\sigma}\right)^2$$

   Let $U = \sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma}\right)$ and $V = \left(\frac{\sqrt{n}(\overline{X} - \mu)}{\sigma}\right)^2$. The distribution we are finding is $U - V$.
   From definition, we know that $U \sim \chi^2(n)$. From Theorem 1.18, we find the $V \sim \chi^2(1)$. From Part 2, since functions of $\mathbf{X}^*$ and $\overline{X}$ are independent,

$$M_U(t) = M_V(t)M_{U-V}(t)$$
$$(1 - 2t)^{-\frac{n}{2}} = (1 - 2t)^{-\frac{1}{2}}M_{U-V}(t)$$
$$M_{U-V}(t) = (1 - 2t)^{-\frac{n-1}{2}}$$

   Therefore, we can conclude that:

$$\frac{1}{\sigma^2}\sum_{i=1}^n (X_i - \overline{X})^2 \sim \chi^2(n-1)$$

$\square$

**Remark 1.20.1.** From the same proof of above theorem part 2, we can also find that $\overline{X}$ and $S_n^2$ are independent.

**Example 1.16.** (**Student's t-distribution**) $T \sim t(r)$
Assume that $X \sim N(0,1)$ and $Y \sim \chi^2(r)$. Let:
$$T = \frac{X}{\sqrt{\frac{Y}{r}}}$$

Then $T$ has a t-distribution with $t$ degree of freedom and:

$$f_T(t) = \frac{\Gamma\left(\frac{r+1}{2}\right)}{\sqrt{r\pi}\Gamma\left(\frac{r}{2}\right)}\left(1 + \frac{t^2}{r}\right)^{-\frac{r+1}{2}} \qquad \mathbb{E}(T) = \begin{cases} \text{Undefined}, & r \leq 1 \\ 0, & n > 1 \end{cases} \qquad \text{Var}(T) = \begin{cases} \text{Undefined}, & r \leq 1 \\ \infty, & 1 < r \leq 2 \\ \frac{r}{r-2}, & r > 2 \end{cases}$$

**Remark 1.20.2.** As $r \to \infty$, $T \to N(0,1)$ by CLT.

**Remark 1.20.3.** If we fix $Y = y$, then we can find that $T \sim N(0, \frac{r}{y})$.

The t-distribution has the following properties.

**Theorem 1.21.** If $\{X_1, \cdots, X_n\}$ is a random sample of size $n > 1$ of random variable $X \sim N(\mu, \sigma^2)$, then we have:
$$\frac{\sqrt{n}(\overline{X} - \mu)}{S_{n-1}} \sim t(n-1)$$

*Proof.*
From Theorem 1.20, $\overline{X}$ and $S_{n-1}^2$ are independent, and:

$$\frac{\sqrt{n}(\overline{X} - \mu)}{\sigma} \sim N(0,1) \qquad\qquad \frac{(n-1)S_{n-1}^2}{\sigma^2} \sim \chi^2(n-1)$$

Therefore, from definition,
$$\frac{\sqrt{n}(\overline{X} - \mu)}{S_{n-1}} = \frac{\frac{\sqrt{n}(\overline{X}-\mu)}{\sigma}}{\sqrt{\frac{1}{n-1}\left(\frac{(n-1)S_{n-1}^2}{\sigma^2}\right)}} \sim t(n-1)$$

$\square$

**Example 1.17.** Assume that we want to find the 95% confidence interval of $\mu$ without knowing the population variance $\sigma^2$. Then we can find:

$$0.95 = \mathbb{P}\left(-t_{0.025,n-1} \leq \frac{\sqrt{n}(\overline{X} - \mu)}{S_{n-1}} \leq t_{0.025,n-1}\right)$$
$$= \mathbb{P}\left(\overline{X} - t_{0.025,n-1}\frac{S_{n-1}}{\sqrt{n}} \leq \mu \leq \overline{X} + t_{0.025,n-1}\frac{S_{n-1}}{\sqrt{n}}\right)$$

Therefore, we can find the 95% confidence interval:

$$\left(\overline{x} - t_{0.025,n-1}\frac{S_{n-1}}{\sqrt{n}}, \overline{x} + t_{0.025,n-1}\frac{S_{n-1}}{\sqrt{n}}\right)$$

Usually when $n > 30$, $t_{0.025,n-1} \approx z_{0.025}$. Therefore, we may find the 95% confidence interval:

$$\left(\overline{x} - z_{0.025}\frac{S_{n-1}}{\sqrt{n}}, \overline{x} + z_{0.025}\frac{S_{n-1}}{\sqrt{n}}\right)$$

**Example 1.18.** (**F distribution**) $F \sim F(r_1, r_2)$

Assume that $X$ and $Y$ are independent random variables with $X \sim \chi^2(r_1)$ and $Y \sim \chi^2(r_2)$. Let:

$$F = \frac{\frac{X}{r_1}}{\frac{Y}{r_2}}$$

Then $F$ has a F-distribution with $r_1$ and $r_2$ degrees of freedom with:

$$f_F(w) = \frac{\Gamma\left(\frac{r_1+r_2}{2}\right)}{\Gamma\left(\frac{r_1}{2}\right)\Gamma\left(\frac{r_2}{2}\right)} \left(\frac{r_1}{r_2}\right)^{\frac{r_1}{2}} w^{\frac{r_1}{2}-1} \left(1 + \frac{r_1 w}{r_2}\right)^{-\frac{r_1+r_2}{2}}$$

where $0 < w < \infty$. We define $F_\alpha(r_1, r_2)$ by:

$$\mathbb{P}(F \geq F_\alpha(r_1, r_2)) = \alpha$$

---

**Lemma 1.22.** Let $U \sim F(r_1, r_2)$. The F-distribution has the following properties:

1. $\frac{1}{U} \sim F(r_2, r_1)$

2. If $F_\alpha(r_1, r_2)$ is defined by $\mathbb{P}(U \geq F_\alpha(r_1, r_2)) = \alpha$, then:

$$\frac{1}{F_\alpha(r_1, r_2)} = F_{1-\alpha}(r_2, r_1)$$

---

*Proof.*

1. By definition,

$$U = \frac{\frac{X}{r_1}}{\frac{Y}{r_2}}$$

where $X \sim \chi^2(r_1)$ and $Y \sim \chi^2(r_2)$. Therefore,

$$\frac{1}{U} = \frac{\frac{Y}{r_2}}{\frac{X}{r_1}} \sim F(r_2, r_1)$$

2. With $\mathbb{P}(U \geq F_\alpha(r_1, r_2)) = \alpha$, since $f_U(w)$ is only defined for $w > 0$, we can get that:

$$\mathbb{P}\left(\frac{1}{U} \leq \frac{1}{F_\alpha(r_1, r_2)}\right) = \alpha$$

$$\mathbb{P}\left(\frac{1}{U} > \frac{1}{F_\alpha(r_1, r_2)}\right) = 1 - \alpha$$

From Part 1, we find that $\frac{1}{U} \sim F(r_2, r_1)$. Therefore,

$$\mathbb{P}\left(\frac{1}{U} \geq F_{1-\alpha}(r_2, r_1)\right) = 1 - \alpha$$

We find that $\frac{1}{F_\alpha(r_1, r_2)} = F_{1-\alpha}(r_2, r_1)$

$\square$

**Example 1.19.** Assume that we try to compare two populations. Let $X_1 \sim \mathrm{N}(\mu_1, \sigma_1^2)$ represents the random variable of the first population and $X_2 \sim \mathrm{N}(\mu_2, \sigma_2^2)$ represents the random variable of the second population. We want to find an interval guess of their ratio of variance $\frac{\sigma_1^2}{\sigma_2^2}$.

Let $\{X_{11}, \cdots, X_{1n}\}$ be a random sample of size $n$ from $X_1$ and $\{X_{21}, \cdots, X_{2m}\}$ be a random sample of size $m$ from $X_2$. We can find that:

$$\frac{(n-1)S_{n-1,1}^2}{\sigma_1^2} \sim \chi^2(n-1) \qquad\qquad \frac{(m-1)S_{m-1,2}^2}{\sigma_2^2} \sim \chi^2(m-1)$$

We can also find that $S_{n-1}, 1$ and $S_{m-1}, 2$ are independent since they are from different population. Therefore, we have:

$$\frac{\sigma_1^2}{\sigma_2^2}\left(\frac{S_{m-1,2}^2}{S_{n-1,1}^2}\right) = \frac{\frac{1}{m-1}\left(\frac{(m-1)S_{m-1,2}^2}{\sigma_2^2}\right)}{\frac{1}{n-1}\left(\frac{(n-1)S_{n-1,1}^2}{\sigma_1^2}\right)} \sim F(m-1, n-1)$$

Then we can find the 95% confidence interval by:

$$0.95 = \mathbb{P}\left(f_{0.975, m-1, n-1} \le \frac{\sigma_1^2}{\sigma_2^2}\left(\frac{S_{m-1,2}^2}{S_{n-1,1}^2}\right) \le f_{0.025, m-1, n-1}\right)$$

$$= \mathbb{P}\left(\frac{S_{n-1,1}^2}{S_{m-1,2}^2} f_{0.975, m-1, n-1} \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{S_{n-1,1}^2}{S_{m-1,2}^2} f_{0.025, m-1, n-1}\right)$$

## 1.5 Moment generating function

It would be useful if we have a function that could generate all moments.

**Definition 1.23.** The **moment generating function** (MGF) of a random variable $X$, denoted by $M_X(t)$, is:

$$M_X(t) = \mathbb{E}(e^{tX})$$

if the expectation exists for $t$ in some neighbourhood of 0.

**Remark 1.23.1.** More precisely, there exists $h > 0$ such that for all $t$ in $(-h, h)$, $\mathbb{E}(e^{tX})$ exists.

**Remark 1.23.2.** MGF of $X$ may not always exist. However, if it exists, then $M_X(t)$ is continuously differentiable in some neighbourhood of the origin.

**Remark 1.23.3.** If we replace $e^{tX}$ by its taylor series, then we would get:

$$M_X(t) = \mathbb{E}\left[\sum_{i=0}^{\infty} \frac{(tX)^i}{i!}\right] = \sum_{i=0}^{\infty} \frac{t^i}{i!} \mathbb{E}(X^i) = \sum_{i=0}^{\infty} \frac{t^i}{i!} \mu_i'$$

**Lemma 1.24.** If $M_X(t)$ is the MGF of a random variable $X$, then:

$$\left.\frac{\partial^k}{\partial t^k} M_X(t)\right|_{t=0} = \mathbb{E}(X^k) = \mu_k'$$

*Proof.*
From the taylor series of MGF, we would see that:

$$\frac{\partial^k}{\partial t^k} M_X(t) = \sum_{i=k}^{\infty} \frac{t^{i-k}}{(i-k)!} \mathbb{E}(X^i)$$

Therefore, by having $t = 0$, we would have:

$$\left.\frac{\partial^k}{\partial t^k} M_X(t)\right|_{t=0} = \mathbb{E}(X^k)$$

$\square$

**Example 1.20.** What is the MGF of $X \sim \text{Bern}(p)$? We have:

$$M_X(t) = \mathbb{E}(e^{tX}) = e^{t(0)}(1-p) + e^{t(1)}(p) = pe^t + 1 - p$$

**Lemma 1.25.** If random variables $X$ and $Y$ are independent, then:

$$M_{X+Y}(t) = M_X(t)M_Y(t)$$

*Proof.*
Since $X$ and $Y$ are independent,

$$M_{X+Y}(t) = \mathbb{E}(e^{t(X+Y)}) = \mathbb{E}(e^{tX})\,\mathbb{E}(e^{tY}) = M_X(t)M_Y(t)$$

$\square$

**Example 1.21.** By definition, if $Y = \text{Bin}(n,p)$, then $Y = X_1 + \cdots + X_n$ where $X_i \sim \text{Bern}(p)$ for all $i$ and they are independent. Therefore,

$$M_Y(t) = \prod_{i=1}^n M_{X_i}(t) = (pe^t + 1 - p)^n$$

We may also solve it without using the definition.

$$M_Y(t) = \mathbb{E}(e^{tY}) = \sum_{i=0}^n \binom{n}{i}(pe^t)^i(1-p)^{n-i} = (pe^t + 1 - p)^n$$

**Example 1.22.** Consider $X \sim \text{Poisson}(\lambda)$, The MGF of $X$ can be obtained by:

$$M_X(t) = \sum_{k=0}^\infty e^{tk}\frac{\lambda^k e^{-\lambda}}{k!} = e^{-\lambda}\sum_{k=0}^\infty \frac{(\lambda e^t)^k}{k!} = e^{\lambda(e^t - 1)}$$

**Example 1.23.** Consider $X \sim \text{Exp}(\lambda)$. If $t < \lambda$, we have:

$$M_X(t) = \mathbb{E}(e^{tX}) = \int_0^\infty e^{tx}\lambda e^{-\lambda x}\,dx = \lambda \int_0^\infty e^{-(\lambda - t)x}\,dx = \frac{\lambda}{\lambda - t}$$

**Example 1.24.** What is the MGF of $X \sim \text{N}(\mu, \sigma^2)$? We may first find the MGF of $Z \sim \text{N}(0,1)$.

$$
\begin{aligned}
M_Z(t) = \mathbb{E}(e^{tZ}) &= \int_{-\infty}^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(z^2 - 2tz)}\,dz \\
&= \int_{-\infty}^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}((z-t)^2 - t^2)}\,dz \\
&= e^{\frac{t^2}{2}} \int_{-\infty}^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(z-t)^2}\,dz \\
&= e^{\frac{t^2}{2}}
\end{aligned}
$$

Therefore, by having $X = \sigma Z + \mu$, we have:

$$M_X(t) = \mathbb{E}(e^{tX}) = e^{\mu t}\,\mathbb{E}(e^{t\sigma Z}) = e^{\mu t + \frac{1}{2}\sigma^2 t^2}$$

**Example 1.25.** Consider $X \sim \text{U}[a,b]$ where $a < b$, we have:

$$M_X(t) = \mathbb{E}(e^{tX}) = \int_a^b \frac{e^{tx}}{b-a}\,dt = \left[\frac{e^{tx}}{t(b-a)}\right]_a^b = \frac{e^{bt} - e^{at}}{t(b-a)}$$

**Example 1.26.** If $X \sim \text{NBin}(r, p)$, then for $t < -\ln(1-p)$:

$$M_X(t) = \left( \frac{pe^t}{1 - (1-p)e^t} \right)^r$$

If $X \sim \text{Gamma}(\alpha, \beta)$, then for $t < \beta$:

$$M_X(t) = \left( \frac{\beta}{\beta - t} \right)^\alpha$$

**Example 1.27.** Given $Y \sim \chi^2(r)$. How do we find the MGF of $Y$?
Note that chi-squared distribution is actually a special case of gamma distribution. We have $\chi^2(r) = \Gamma(\frac{r}{2}, \frac{1}{2})$. Therefore, by substituting, for $t < \frac{1}{2}$, we may get:

$$M_Y(t) = \left( \frac{\frac{1}{2}}{\frac{1}{2} - t} \right)^{\frac{r}{2}} = (1 - 2t)^{-\frac{r}{2}}$$

**Example 1.28.** Given that $Y \sim \chi^2(r)$. How do we find $\mathbb{E}(Y)$ without using the MGF of $Y$?
By definition, we may let $Y = \sum_{i=1}^{r} X_i^2$, where $X_i \sim \text{N}(0,1)$. Therefore,

$$\mathbb{E}(Y) = \sum_{i=1}^{r} \mathbb{E}(X_i^2) = \sum_{i=1}^{r} \frac{d^2}{dt^2} e^{\frac{1}{2}t^2} \bigg|_{t=0} = r(1 + t^2) e^{\frac{1}{2}t^2} \bigg|_{t=0} = r$$

Ultimately, the reason why we use moment generating function is the following fact.

**Theorem 1.26.** (**Uniqueness of MGF**) Let $X$ and $Y$ be two random variables. Suppose that their MGFs exist and are equal for all $t \in (-h, h)$ for some $h > 0$, then the distribution functions $F_X$ and $F_Y$ are equal.

This means that by knowing the MGF of a particular random variable $X$, we can know its distribution.

**Example 1.29.** Assume that $X_1, \cdots, X_n$ are independent and $X_i \sim \text{Bin}(m_i, p)$ for all $i = 1, \cdots, n$. Then we have:

$$M_{X_1 + \cdots + X_n}(t) = \prod_{i=1}^{n} M_{X_i}(t) = \prod_{i=1}^{n} (pe^t + 1 - p)^{m_i} = (pe^t + 1 - p)^{\sum_{i=1}^{n} m_i}$$

Therefore, we have $X_1 + \cdots + X_n \sim \text{Bin}(\sum_{i=1}^{n} m_i, p)$.

**Example 1.30.** Assume that $X_1, \cdots, X_n$ are independent and $X_i \sim \text{Poisson}(\lambda_i)$ for all $i = 1, \cdots, n$. Then we have:

$$M_{X_1 + \cdots + X_n}(t) = \prod_{i=1}^{n} M_{X_i}(t) = \prod_{i=1}^{n} e^{\lambda_i(e^t - 1)} = e^{\sum_{i=1}^{n} \lambda_i(e^t - 1)}$$

Therefore, we have $X_1 + \cdots + X_n \sim \text{Poisson}(\sum_{i=1}^{n} \lambda_i)$.

**Example 1.31.** Similarly, given a set of independent random variables $\{X_1, \cdots, X_n\}$.

1. If $X_i \sim \text{NBin}(r_i, p)$, then $X_1 + \cdots + X_n \sim \text{NBin}(\sum_{i=1}^{n} r_i, p)$.

2. If $X_i \sim \text{N}(\mu_i, \sigma_i^2)$, then $X_1 + \cdots + X_n \sim \text{N}(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2)$.

3. If $X_i \sim \text{Gamma}(\alpha_i, \beta)$, then $X_1 + \cdots + X_n \sim \text{Gamma}(\sum_{i=1}^{n} \alpha_i, \beta)$ and $cX_i \sim \text{Gamma}(\alpha_i, \frac{\beta}{c})$ for $c \neq 0$.

**Remark 1.26.1.** Not all sum of distributions will result in the same type of distribution.

More generally, we would deal with problems of limiting distribution.

**Theorem 1.27.** Suppose the $\{X_n\}$ is a sequence of random variables, each with MGF $M_{X_n}(t)$. If we have:

$$\lim_{n\to\infty} M_{X_n}(t) = M_Y(t)$$

for all $t$ in a neighbourhood of 0, where $M_Y(t)$ is a MGF for some random variables $Y$, then there is an unique distribution function $F_Y$ with corresponding $M_Y(t)$ such that:

$$\lim_{n\to\infty} F_{X_n}(y) = F_Y(y)$$

for all $y$ where $F_Y(y)$ is continuous. We denote by $X_n \to Y$ or $X_n \xrightarrow{D} X$.

**Remark 1.27.1.** Simply, the limiting distribution of $X_n$ is equal to the distribution of $Y$.

We may define the limiting convergence in truly theoretical way.

**Definition 1.28.** A sequence of random variables $\{X_n\}$ **converges in distribution** to a random variable $X$, denoted by $X_n \xrightarrow{D} X$, if for all continuity point $x$ of $F_X$, as $n \to \infty$,

$$F_{X_n}(x) \to F_X(x)$$

We also have a more strict convergence.

**Definition 1.29.** A sequence of random variables $\{X_n\}$ **converges in probability** to a random variable $X$, denoted by $X_n \xrightarrow{\mathbb{P}} X$, if for any $\varepsilon > 0$, as $n \to \infty$,

$$\mathbb{P}(|X_n - X| < \varepsilon) \to 1 \qquad\qquad \mathbb{P}(|X_n - X| \geq \varepsilon) \to 0$$

**Remark 1.29.1.** If $X_n \xrightarrow{P} X$, then $X_n \xrightarrow{D} X$. The converse is not necessarily true.

## 1.6   Limit Theorems

Using the last two theorems, the next two theorems are very useful in both statistics and probability theory by giving us approximate distribution of an average without a lot of distributional assumption.

**Theorem 1.30.** (**Weak Law of Large Numbers (WLLN)**) Let $\{X_n\}$ be a sequence of i.i.d. random variables. Let $\mathbb{E}(X_i) = \mu$ for all $i = 1, 2, \cdots$. Define by $\overline{X}$ the sample mean of the random variables. Then as $n \to \infty$

$$\overline{X} \xrightarrow{D} \mu$$

**Theorem 1.31.** (**Classical Central Limit Theorem (CLT)**) Let $\{X_n\}$ be a sequence of i.i.d. random variables whose MGFs exist in a neighbourhood of 0. Let $\mathbb{E}(X_i) = \mu$ and $\text{Var}(X_i) = \sigma^2 > 0$ for all $i = 1, 2, \cdots$. Define by $\overline{X}$ the sample mean of the random variables. Then as $n \to \infty$,

$$\frac{\sqrt{n}(\overline{X} - \mu)}{\sigma} = \frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n}\sigma} \to N(0, 1)$$

**Remark 1.31.1.** This is just an abuse of notations.

This works generally to most of the distribution. However, it is probably very tedious to find the MGF. We cam apply the following version of CLT instead.

**Theorem 1.32.** (**Lévy-Linderberg Central Limit Theorem**) Let $\{X_n\}$ be a sequence of i.i.d. random variables with common population means $\mu$ and common population variance $\sigma^2$. Assume that $0 < \sigma^2 < \infty$. Define by $\overline{X}$ the sample mean of the random variables. Then as $n \to \infty$,

$$\frac{\sqrt{n}(\overline{X} - \mu)}{\sigma} = \frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n}\sigma} \to N(0, 1)$$

Sometimes, we will deal with a function of multiple random variables. We must establish how they converges.

**Theorem 1.33. (Slutsky's Theorem)** If $X_n \xrightarrow{D} X$ and $Y_n \xrightarrow{\mathbb{P}} c$, then:

1. $X_n + Y_n \xrightarrow{D} X + c$

2. $X_n Y_n \xrightarrow{D} cX$

3. $\frac{X_n}{Y_n} \xrightarrow{D} \frac{X}{c}$ if $c \neq 0$.

**Example 1.32.** Assume that $X_i \sim \text{Bern}(p)$ for all $i$. We want to estimate the unknown $p$. We have common mean $\mu = p$ and common variance $\sigma^2 = p(1-p)$. By applying CLT, as $n \to \infty$,

$$\overline{X} \to \text{N}\left(p, \frac{p(1-p)}{n}\right)$$

Therefore, we can use normal distribution to approximate the unknown parameter. We want an estimation that we can confident about, and commonly we use probability 0.95.

$$0.95 = \mathbb{P}\left(-z_{0.025} \leq \frac{\overline{X} - p}{\sqrt{\frac{p(1-p)}{n}}} \leq z_{0.025}\right) = \mathbb{P}\left((\overline{X} - p)^2 \leq z_{0.025}^2 \frac{p(1-p)}{n}\right)$$

Solving the inequality, we would find an interval that estimates the parameter $p$. However, this is highly inconvenient. We may use another method. Let us replace $\sqrt{\frac{p(1-p)}{n}}$ with $\sqrt{\frac{\overline{X}(1-\overline{X})}{n}}$. As $n \to \infty$,

$$\sqrt{\frac{\overline{X}(1-\overline{X})}{n}} = \sqrt{\frac{p(1-p)}{n}}\sqrt{\frac{p(1-p)}{\overline{X}(1-\overline{X})}} \to \sqrt{\frac{p(1-p)}{n}}$$

since by Slutsky's Theorem, $\sqrt{\frac{p(1-p)}{\overline{X}(1-\overline{X})}} \to 1$ as $\overline{X} \to p$ by WLLN. We have:

$$0.95 = \mathbb{P}\left(-z_{0.025} \leq \frac{\overline{X} - p}{\sqrt{\frac{\overline{X}(1-\overline{X})}{n}}} \leq z_{0.025}\right) = \mathbb{P}\left(\overline{X} - z_{0.025}\sqrt{\frac{\overline{X}(1-\overline{X})}{n}} \leq p \leq \overline{X} + z_{0.025}\sqrt{\frac{\overline{X}(1-\overline{X})}{n}}\right)$$

**Example 1.33.** In a survey before an election, a poll was taken of 300 potential voters. Among them, 120 said that they would vote for candidate A. Determine a 95% confidence interval for the population proportion $p_A$ of voters who would vote for candidate A in the election.
From the poll, we have a point estimate $\overline{x} = \hat{p}_A = \frac{120}{300} = 0.4$. From the last example, we have found that the 95% confidence interval is:

$$\left(\overline{x} - z_{0.025}\sqrt{\frac{\overline{x}(1-\overline{x})}{n}}, \overline{x} + z_{0.025}\sqrt{\frac{\overline{x}(1-\overline{x})}{n}}\right) \approx (0.3446, 0.4554)$$

Equivalently, the percentage of voters of candidate A would be from 34.46% to 45.54%, with a margin of error 5.54%.

**Example 1.34.** Following the previous example. Assume that we have been given a margin of error $D$. How many data should we collect in order to to have the margin of error?

From how we find the margin of error:

$$z_{0.025}\sqrt{\frac{\overline{x}(1-\overline{x})}{n}} = D \implies n = p(1-p)\frac{z_{0.025}^2}{D^2}$$

Since $p(1-p) \leq \frac{1}{4}$, if we specify that $D = 0.05$, we have:

$$n \leq \frac{z_{0.025}^2}{4D^2} = \frac{1.96^2}{4(0.05)^2} \leq \frac{2^2}{4(0.05)^2} = 400$$

We may use this to determine whether we have obtained enough data.

Assume that we have $n^*$ respondents. Is it enough? The number of required respondents is obtained by:

$$n_{\text{required}} = \frac{\overline{x^*}(1-\overline{x^*})z_{0.025}^2}{D^2}$$

If $n^* < n_{\text{required}}$, then the current number of data is not enough. We would need to find more respondents.
If $n^* \geq n_{\text{required}}$, then it is enough.

---

**Example 1.35.** We try to use Poisson random variables to prove that as $n \to \infty$,

$$e^{-n}\sum_{k=0}^{n}\frac{n^k}{k!} \to \frac{1}{2}$$

Let $\{X_n\}$ be a sequence of i.i.d. random variables and $X_i \sim \text{Poisson}(1)$ for $i = 1, 2, \cdots$. Let $Y_n = \sum_{i=1}^{n} X_i$. By CLT, we have:

$$\frac{Y_n - n}{\sqrt{n}} \to \text{N}(0, 1)$$

Therefore, since $Y_n \sim \text{Poisson}(n)$, we have:

$$e^{-n}\sum_{k=0}^{n}\frac{n^k}{k!} = \mathbb{P}(Y_n \leq n) = \mathbb{P}(\frac{Y_n - n}{\sqrt{n}} \leq 0) \to \frac{1}{2}$$

---

**Example 1.36.** Given a sequence of i.i.d. random variables $\{X_n\}$. We want to find the asymptotic distribution for the $k$-th sample moment $\overline{X^k}$ as $n \to \infty$. Notice that $X_i^k$ are independent for $i = 1, 2, \cdots$. By CLT,

$$\frac{\sqrt{n}(\overline{X^k} - \mu_k')}{\sqrt{\mu_{2k}' - (\mu_k')^2}} \to \text{N}(0, 1)$$

Therefore, the asymptotic distribution for $\overline{X^k}$ when $n \to \infty$ is $\text{N}(\mu_k', \frac{1}{n}(\mu_{2k}' - (\mu_k')^2))$.

Central Limit Theorem can provide us a limiting standard normal distribution of sample mean. However, we usually deal with some functions of the sample mean.

---

**Theorem 1.34.** (**Continuous mapping theorem**) Let $\{X_n\}$ be a sequence of random variables and $X$ be a random variable. Suppose there is a function $g$ has a set of discontinuity points $D_g$ such that $\mathbb{P}(X \in D_g) = 0$, then:

1. If $X_n \xrightarrow{D} X$, then $g(X_n) \xrightarrow{D} g(X)$.

2. If $X_n \xrightarrow{\mathbb{P}} X$, then $g(X_n) \xrightarrow{\mathbb{P}} g(X)$.

---

**Theorem 1.35.** (**Delta method**) Let $\{X_n\}$ be a sequence of random variables such that for constants $a$ and $b > 0$, as $n \to \infty$,

$$\sqrt{n}(X_n - a) \to \text{N}(0, b^2)$$

Then for a given function $g$, suppose that $g'(a)$ exists and not 0, as $n \to \infty$:

$$\sqrt{n}(g(X_n) - g(a)) \to \text{N}(0, [g'(a)b]^2)$$

**Corollary 1.36.** If $\overline{X}$ is the sample mean of a random sample $X_1, \cdots, X_n$ of size $n$ from a distribution with a finite mean $\mu$ and finite variance $\sigma^2 > 0$. For a given function $g$, suppose that $g'(\mu)$ exists and is not 0, as $n \to \infty$:
$$\sqrt{n}(g(\overline{X}) - g(\mu)) \to N(0, [g'(\mu)\sigma]^2)$$

**Example 1.37.** Assume that there are 70 respondents, 68 of which would vote for one candidate.

If we use the same process from previous examples, we find that the 95% confidence interval is $(0.9324, 1.0105)$, which is out of the range. In fact, if the point estimate $\hat{p}$ is quite close to 0 or 1, the resulting interval guess may include values that is out of the range of $p$. This is a poor interval guess.

We take a transformation, say $g(p)$, such that $g(p) \in (-\infty, \infty)$. We may find that since $0 < p < 1$, $\ln p < 0$. Therefore, we can find that:
$$g(p) = \ln(-\ln p) \in (-\infty, \infty)$$

By Delta method,
$$\frac{g(\overline{X}) - g(p)}{g'(p)\sqrt{\frac{\overline{x}(1-\overline{x})}{n}}} \to N(0, 1)$$

By WLLN and continuous mapping theorem, we can replace $g'(p)$ with $g'(\overline{X})$. Therefore, we have:
$$0.95 = \mathbb{P}\left(-z_{0.025} \leq \frac{g(\overline{X}) - g(p)}{g'(p)\sqrt{\frac{\overline{x}(1-\overline{x})}{n}}} \leq z_{0.025}\right)$$

Solving the formula and we would get the a good 95% confidence interval of $p$.

**Example 1.38.** Let $\{X_n\}$ be a sequence of i.i.d. random variables and $X_i \sim \text{Bern}(\theta)$ for $i = 1, 2, \cdots$. Show that:
$$Z_n = 2\sqrt{n}\left(\sin^{-1}\sqrt{\overline{X}} - \sin^{-1}\sqrt{\theta}\right) \to N(0, 1)$$

Let $g(x) = \sin^{-1}\sqrt{x}$. We may obtain that:
$$g'(x) = \frac{1}{2\sqrt{x}\sqrt{1-x}}$$

We can find that the derivative is well-defined and non-zero for $0 < \theta < 1$ by substituting $x = \theta$. Note that $\mathbb{E}(X_i) = \theta$ and $\text{Var}(X_i) = \theta(1 - \theta)$ for $i = 1, \cdots, n$. By Corollary 1.36,
$$\sqrt{n}(g(\overline{X}) - g(\theta)) \to N\left(0, \frac{1}{4}\right)$$

Since $Z_n = 2\sqrt{n}(g(\overline{X}) - g(\theta))$, we may find that as $n \to \infty$,
$$Z_n \to N(0, 1)$$

**Example 1.39.** Let $\{X_n\}$ be a sequence of i.i.d. random variables and $X_i \sim \text{Exp}(\theta)$ for $i = 1, 2, \cdots$. We want to find a variance-stabilizing transformation, which is to find a function $g(x)$ such that the limiting distribution of:
$$Y_n = \sqrt{n}[g(\overline{X_n}) - g(\theta)]$$

does not depend on $\theta$.

We may find that $\mathbb{E}(X_i) = \frac{1}{\theta}$ and $\text{Var}(X_i) = \frac{1}{\theta^2}$ for $i = 1, 2, \cdots$. We claim that $g(x) = \ln x$ is what we want. We can find that:
$$g'(x) = \frac{1}{x}$$

By substituting $x = \frac{1}{\theta}$, we can see that the derivative is non-zero. Applying Corollary 1.36,
$$\sqrt{n}\left(g(\overline{X}) - g\left(\frac{1}{\theta}\right)\right) \to N(0, 1)$$

Therefore, $g(x) = \ln x$ is the variance-stabilizing transformation.

However, we usually deal with more than 1 variables. Before we extend the theorems into multivariate case, we must first introduce the multivariate normal distribution.

**Example 1.40. (Multivariate Normal Distribution)** $\mathbf{X} \sim \mathrm{N}_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
Given a random vector $\mathbf{X}$. Let the $k \times 1$ vector $\boldsymbol{\mu}$ be the expected value of $\mathbf{X}$ and the $k \times k$ matrix $\boldsymbol{\Sigma}$ be its variance-covariance matrix. Assume that $\boldsymbol{\Sigma}$ is positive-definite. (For all non-zero vectors $\mathbf{z}$ with real entries, we have $\mathbf{z}^T \boldsymbol{\Sigma} \mathbf{z} > 0$) The random vector $\mathbf{X}$ is $k$-dimensional normal if its PDF is:

$$f_{\mathbf{X}}(\mathbf{x}) = (2\pi)^{-\frac{k}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

**Remark 1.36.1.** The i-th row and j-th column of the $k \times k$ variance-covariance matrix $\boldsymbol{\Sigma}$ is the element $a_{ij}$ found by:

$$a_{ij} = \mathrm{cov}(X_i, X_j)$$

Note that if $i = j$, then $\mathrm{cov}(X_i, X_i) = \mathrm{Var}(X_i)$.

**Example 1.41.** If $k = 2$, then $X \sim \mathrm{N}_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is bivariate normal.

**Lemma 1.37.** If $\mathbf{X} \sim \mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then for any $q \times p$ matrix $\mathbf{A}$, we have:

$$\mathbf{A}\mathbf{X} \sim \mathrm{N}_q(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T)$$

**Example 1.42.** Using this lemma, one can isolate some of the random variables that made up the random vector $\mathbf{X} = (X_1 \ \cdots \ X_p)^T \sim \mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. For example, setting $(p-1) \times p$ matrix $\mathbf{A}$ as:

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \dots & \dots & 0 & 1 \end{pmatrix} = \left( \begin{array}{c|c} 0 & \\ \vdots & \mathbf{I}_{(p-1)\times(p-1)} \\ 0 & \end{array} \right)$$

We may find that:

$$\mathbf{A}\mathbf{X} = \begin{pmatrix} X_2 \\ \vdots \\ X_p \end{pmatrix} \sim \mathrm{N}_{p-1}(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T)$$

where $\mathbf{A}\boldsymbol{\mu}$ is the mean vector of $(X_2 \ \cdots \ X_p)^T$ and $\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T$ is the variance-covariance matrix of $(X_2 \ \cdots \ X_p)^T$.

**Lemma 1.38.** If we have:

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathrm{N}_2 \left( \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix} \right),$$

then $X_1$ and $X_2$ are independent if and only if $\sigma_{12} = \sigma_{21} = 0$.

*Proof.*
From the properties of covariance,

$$\sigma_{12} = \mathrm{cov}(X_1, X_2) = \mathrm{cov}(X_2, X_1) = \sigma_{21}$$

Assume that $X_1$ and $X_2$ are independent. We have:

$$\mathrm{cov}(X_1, X_2) = \mathbb{E}[(X_1 - \mathbb{E}(X_1))(X_2 - \mathbb{E}(X_2))] = \mathbb{E}(X_1 X_2) - \mathbb{E}(X_1)\mathbb{E}(X_2) = 0$$

Therefore, $\sigma_{12} = \sigma_{21} = 0$.
Assume that $\sigma_{12} = \sigma_{21} = 0$. Then $\mathbb{E}(X_1 X_2) = \mathbb{E}(X_1)\mathbb{E}(X_2)$. We check joint PDF to see if they are independent. Since $\mathrm{cov}(X_1, X_2) = 0$, we have that:

$$f_{X_1, X_2}(x_1, x_2) = \frac{1}{2\pi\sigma_{11}\sigma_{22}} \exp\left( -\frac{1}{2}\left( \frac{(x_1 - \mu_1)^2}{2\sigma_{11}^2} + \frac{(x_2 - \mu_2)^2}{2\sigma_{22}^2} \right) \right)$$

$\square$

**Remark 1.38.1.** Two random variables are uncorrelated does not mean they are independent. It is only true if they are bivariate normal.

We may extend the CLT to multivariate case.

---

**Theorem 1.39.** (**Multivariate Central Limit Theorem**) Let $\{\mathbf{X}_n = (X_{n1} \; \cdots \; X_{nk})^T \in \mathbb{R}^k\}$ be a sequence of i.i.d. random vectors with variance-covariance matrix $\mathbf{\Sigma}$. We assume that $\mathbb{E}(X_{ij}^2) < \infty$ for $i = 1, 2, \cdots$ and $j = 1, \cdots, k$. Define by $\overline{\mathbf{X}}$ the sample mean of the random vectors. Then as $n \to \infty$,

$$\sqrt{n}(\overline{\mathbf{X}} - \boldsymbol{\mu}) \xrightarrow{D} \mathrm{N}_k(\mathbf{0}, \mathbf{\Sigma})$$

---

We may extend the Delta method to multivariate cases.

---

**Theorem 1.40.** (**Multivariate 1st-order Delta Method**) Let $\{\mathbf{X}_n \in \mathbb{R}^k\}$ be a sequence of random vectors such that for constant vector $\mathbf{a} \in \mathbb{R}^k$, as $n \to \infty$,

$$\sqrt{n}(\mathbf{X}_n - \mathbf{a}) \xrightarrow{D} \mathbf{U}$$

where $\mathbf{U}$ is a random vector in $\mathbb{R}^k$. If a function $h : \mathbb{R}^k \to \mathbb{R}$ has a derivative $\nabla h(\mathbf{a}) \neq \mathbf{0}$, then as $n \to \infty$:

$$\sqrt{n}(h(\mathbf{X}_n) - h(\mathbf{a})) \xrightarrow{D} \nabla h(\mathbf{a})\mathbf{U}$$

where

$$\nabla h = \left( \frac{\partial}{\partial t_1} h(t_1, \cdots, t_k), \cdots, \frac{\partial}{\partial t_k} h(t_1, \cdots, t_k) \right)$$

---

# Chapter 2

# Point Estimation

In this chapter, we will study two different general approaches to estimate an unknown parameters of any given parametric distribution.

The basic idea of point estimation is that we use a statistic $T$, an estimate $T(\mathbf{x})$ or an estimator $T(\mathbf{X})$ to estimate the unknown parameter $g(\theta)$, where $\mathbf{x} = (x_1 \ \cdots \ x_n)^T$ is a realization of random vector $\mathbf{X} = (X_1 \ \cdots \ X_n)^T$ with a PDF $f(x|\theta)$ or PMF $p(x|\theta)$ and $\theta$ in parameter space $\Theta$.

---

**Remark 2.0.1.** Most often, the parameters of our interest to be estimated (**estimand**) is a function of the unknown distribution parameters $\theta$. E.g. $\mu^2, \frac{\sigma}{\mu}$.

---

**Remark 2.0.2.** We only estimate an unknown parameters. There is no point to estimate an already known parameters.

---

**Definition 2.1.** An estimator or estimate $\hat{\theta}$ is **unbiased** or **mean-unbiased** for $\theta$ if $\mathbb{E}(\hat{\theta}) = \theta$.

---

## 2.1 Methods of Moments Estimation

Methods of moments estimation is one of the most popular methods in statistics to estimate an unknown parameters. As the name suggests, it is related to moments. The motivation is that in some situations, the parameter of interest can be written as a function of population moments about 0.

---

**Definition 2.2.** Suppose that there are $k$ unknown parameters $\theta_1, \cdots, \theta_k$. If we can write them in terms of $k$ or more moments, i.e.:
$$\begin{cases} \theta_1 = g_1(\mu'_1, \mu'_2, \cdots, \mu'_k, \cdots) \\ \theta_2 = g_2(\mu'_1, \mu'_2, \cdots, \mu'_k, \cdots) \\ \vdots \\ \theta_k = g_k(\mu'_1, \mu'_2, \cdots, \mu'_k, \cdots) \end{cases}$$
then the **method of moments estimator** (MME), denoted by $(\widetilde{\theta}_1, \widetilde{\theta}_2, \cdots, \widetilde{\theta}_k)$, of $(\theta_1, \theta_2, \cdots, \theta_k)$ is:
$$\begin{cases} \widetilde{\theta}_1 = g_1(\overline{X}, \overline{X^2}, \cdots, \overline{X^k}, \cdots) \\ \widetilde{\theta}_2 = g_2(\overline{X}, \overline{X^2}, \cdots, \overline{X^k}, \cdots) \\ \vdots \\ \widetilde{\theta}_k = g_k(\overline{X}, \overline{X^2}, \cdots, \overline{X^k}, \cdots) \end{cases}$$

---

**Remark 2.2.1.** This estimation is quick and easy, but the MME obtained are often biased and heavily relies on the existence of the required population moments.

---

**Remark 2.2.2.** Do not mix up method of moment estimator or method of moment estimate.

---

**Remark 2.2.3.** Do not write the MME as $(\theta_1, \theta_2, \cdots, \theta_k)$. This is wrong.

**Example 2.1.** Consider a random sample of size $n$ from $X \sim \mathrm{N}(10, \sigma^2)$. We want to estimate $\sigma^2$. We have $k = 1$, $\theta_1 = \sigma^2$. We can write it in terms of moments:

$$\sigma^2 = \mathbb{E}(X^2) - 100$$

Therefore, the MME of $\sigma^2$ is:

$$\widetilde{\sigma}^2 = \overline{X^2} - 100$$

**Example 2.2.** Consider a random sample of size $n$ from $X \sim \mathrm{N}(\mu, \sigma^2)$. We want to estimate $\mu$ and $\sigma^2$. We have $k = 2$, $(\theta_1, \theta_2) = (\mu, \sigma^2)$. We can write them in terms of moments:

$$\begin{cases} \mu = E(X) \\ \sigma^2 = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2 \end{cases}$$

Therefore, the MME of $\mu$ and $\sigma^2$ are:

$$\begin{cases} \widetilde{\mu} = \overline{X} \\ \widetilde{\sigma}^2 = \overline{X^2} - (\overline{X})^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 \end{cases}$$

**Remark 2.2.4.** MME may not be unique because the parameter can be written as different functions of moments. To fix this problem, we usually prefer using fewer or lower moments to get MME.

**Example 2.3.** Consider a random sample of size $n$ from $X \sim \mathrm{Poisson}(\lambda)$. We want to estimate $\lambda$. We have $k = 1$, $\theta_1 = \lambda$. We have multiple ways to write it in terms of moments. It can be $\lambda = \mathbb{E}(X)$, $\lambda = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2$ or any other combinations. From the remark, we would choose the one with fewer or lower moments, which is:

$$\lambda = \mathbb{E}(X)$$

Therefore, the MME of $\lambda$ is:

$$\lambda = \overline{X}$$

**Example 2.4.** Consider a random sample of size $n$ from $X \sim \mathrm{Gamma}(\alpha, \beta)$. Assume that we know that $\mathbb{E}(X) = 3423$. We have $k = 2$, $(\theta_1, \theta_2) = (\alpha, \beta)$. We can write them in terms of moments:

$$\begin{cases} 3423 = \frac{\alpha}{\beta} \\ \mathbb{E}(X^2) = \frac{\alpha}{\beta^2} + 3423^2 \end{cases} \implies \begin{cases} \alpha = \frac{3423^2}{\mathbb{E}(X^2) - 3423^2} \\ \beta = \frac{3423}{\mathbb{E}(X^2) - 3423^2} \end{cases}$$

Therefore, the MME of $\alpha$ and $\beta$ is:

$$\begin{cases} \widetilde{\alpha} = \frac{3423^2}{\overline{X^2} - 3423^2} \\ \widetilde{\beta} = \frac{3423}{\overline{X^2} - 3423^2} \end{cases}$$

**Lemma 2.3.** (**Invariance property of MME**) If $\widetilde{\theta}_i$ is the MME for $\theta_i$ for $i = 1, \cdots, k$, then $h(\widetilde{\theta}_1, \cdots, \widetilde{\theta}_k)$ is the MME for $h(\theta_1, \cdots, \theta_k)$, where $h$ is a known function.

**Theorem 2.4.** A sequence of MME $\{\widetilde{\theta}_n \in \mathbb{R}^k\}$ is consistent and asymptotically unbiased for $\theta$. It is also asymptotically normally distributed. More precisely, under certain assumption like $\mathbb{E}|X|^{2k} < \infty$, as $n \to \infty$, we have:

$$\sqrt{n}(\widetilde{\theta}_n - \theta) \to \mathrm{N}_k(\mathbf{0}, \mathbf{GHG}^T)$$

where $\mathbf{G}$ is a $k \times k$ matrix with $\frac{\partial g_i}{\partial \mu'_j}$ as its $(i, j)$-th entry and $\mathbf{H}$ is a $k \times k$ matrix with $\mu'_{i+j} - \mu'_i \mu'_j$ as its $(i, j)$-th entry, for $i = 1, \cdots, k$ and $j = 1, \cdots, k$.

**Remark 2.4.1.** In the theorem, "consistent" means convergence in probability. For any $\varepsilon > 0$, as $n \to \infty$,

$$\mathbb{P}(|\widetilde{\theta}_n - \theta| > \varepsilon) \to 0$$

**Remark 2.4.2.** Also in the theorem, "asymptotically unbiased" means that we have:

$$\lim_{n \to \infty} \mathbb{E}(\widetilde{\theta}_n) = \theta$$

Note that $\mathbb{E}(\widetilde{\theta}_n) \neq \theta$ for some $n$.

---

**Example 2.5.** Consider a random sample of size $n$ from a random variable $X$ with $\mathbb{E}|X|^4 < \infty$. We take:

$$\theta = \begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix} = \begin{pmatrix} \mu_1' \\ \mu_2' - (\mu_1')^2 \end{pmatrix}$$

We have:

$$\mathbf{G} = \begin{pmatrix} 1 & 0 \\ -2\mu_1' & 1 \end{pmatrix} \qquad\qquad \mathbf{H} = \begin{pmatrix} \mu_2' - (\mu_1')^2 & \mu_3' - \mu_1'\mu_2' \\ \mu_3' - \mu_2'\mu_1' & \mu_4' - (\mu_2')^2 \end{pmatrix}$$

Therefore,

$$\begin{aligned}
\mathbf{GHG}^T &= \begin{pmatrix} 1 & 0 \\ -2\mu_1' & 1 \end{pmatrix} \begin{pmatrix} \mu_2' - (\mu_1')^2 & \mu_3' - \mu_1'\mu_2' \\ \mu_3' - \mu_2'\mu_1' & \mu_4' - (\mu_2')^2 \end{pmatrix} \mathbf{G}^T \\
&= \begin{pmatrix} \mu_2' - (\mu_1')^2 & \mu_3' - \mu_1'\mu_2' \\ \mu_3' - 3\mu_1'\mu_2' + 2(\mu_1')^3 & \mu_4' - 2\mu_1'\mu_3' - (\mu_2')^2 + 2(\mu_1')^2\mu_2' \end{pmatrix} \begin{pmatrix} 1 & -2\mu_1' \\ 0 & 1 \end{pmatrix} \\
&= \begin{pmatrix} \mu_2' - (\mu_1')^2 & \mu_3' - 3\mu_1'\mu_2' + 2(\mu_1')^3 \\ \mu_3' - 3\mu_1'\mu_2' + 2(\mu_1')^3 & \mu_4' - 4\mu_1'\mu_3' - (\mu_2')^2 + 8\mu_2'(\mu_1')^2 - 4(\mu_1')^4 \end{pmatrix}
\end{aligned}$$

Using the fact that:

$$\begin{aligned}
\mu_3 &= \mu_3' - 3\mu_2'\mu_1' + 2(\mu_1')^3 \\
\mu_4 &= \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'(\mu_1')^2 - 3(\mu_1')^4 \\
\sigma^4 &= (\mu_2')^2 - 2\mu_2'(\mu_1')^2 + (\mu_1')^4
\end{aligned}$$

We can find the resultant matrix:

$$\mathbf{GHG}^T = \begin{pmatrix} \sigma^2 & \mu_3 \\ \mu_3 & \mu_4 - \sigma^4 \end{pmatrix}$$

Using Theorem 2.4, denote:

$$\widetilde{\theta}_n = \begin{pmatrix} \overline{X}_n \\ S_n^2 \end{pmatrix}$$

As $n \to \infty$,

$$\sqrt{n}\left( \begin{pmatrix} \overline{X}_n \\ S_n^2 \end{pmatrix} - \begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix} \right) \to N_2\left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & \mu_3 \\ \mu_3 & \mu_4 - \sigma^4 \end{pmatrix} \right)$$

Based on the properties of variance-covariance matrix, we may find that as $n \to \infty$:

$$\sqrt{n}(S_n^2 - \sigma^2) \to N(0, \mu_4 - \sigma^4)$$

By Delta Method, in condition that $\sigma^2 > 0$,

$$\sqrt{n}(S_n - \sigma) \to N\left( 0, \frac{\mu_4 - \sigma^4}{4\sigma^2} \right)$$

## 2.2   Maximum Likelihood Estimation

The method of maximum likelihood is by far the most popular technique for deriving estimators, popularized by Ronald Aylmer Fisher in 1922. Currently, there are still a lot of research studying the properties of this estimation method.

**Definition 2.5.** Consider a random sample of size $n$ from a population with a PDF $f(\mathbf{x}|\theta)$ or a PMF $p(\mathbf{x}|\theta)$. Given a realization $\mathbf{x} = (x_1 \ \cdots \ x_n)^T$. The **likelihood function** is defined by:

$$L(\theta) = L(\theta_1, \cdots, \theta_k | \mathbf{x}) = \begin{cases} \prod_{i=1}^{n} f(x_i|\theta), & \text{Continuous case} \\ \prod_{i=1}^{n} p(x_i|\theta), & \text{Discrete case} \end{cases}$$

The likelihood function can be used to quantify how the observed data is likely to occur.

**Remark 2.5.1.** The likelihood function $L(\theta)$ is a function of $\theta$ with fixed $\mathbf{x}$.

**Remark 2.5.2.** Do not replace $x_i$ with $x$.

$$L(\theta) = \begin{cases} \prod_{i=1}^{n} f(x_i|\theta) \neq \prod_{i=1}^{n} f(x|\theta), & \text{Continuous case} \\ \prod_{i=1}^{n} p(x_i|\theta) \neq \prod_{i=1}^{n} p(x|\theta), & \text{Discrete case} \end{cases}$$

The idea is that for each realization of $\mathbf{x}$, we want to estimate a value of $\theta \in \Theta$ at which $L(\theta)$ attains its maximum.

**Definition 2.6.** The **maximum likelihood estimate** (MLE), denoted by $\hat{\theta}$, is obtained by:

$$\hat{\theta} = \arg\max_{\theta \in \Theta} L(\theta)$$

**Remark 2.6.1.** In some cases, especially when differentiation is used, it is easier to work with **log likelihood** defined by:
$$l(\theta) = \log L(\theta)$$
We can do this because $l(\theta)$ and $L(\theta)$ are strictly increasing and they have the same maxima.

**Example 2.6.** Consider a random sample of size $n = 10$ from $\text{Bern}(\theta)$, where $\theta$ is unknown. Therefore,

$$L(\theta) = \prod_{i=1}^{n} p(x_i|\theta) = \theta^{n\overline{x}}(1-\theta)^{n-n\overline{x}}$$

Suppose that there are only two possible values of $\theta$. We have either $\theta = 0.1$ or $\theta = 0.5$.
From the data observed, assumed that we have found that $\overline{x} = 0.4$. Substituting gets:

$$L(0.1) = (0.1)^4(0.9)^6 = 0.0000531441 \qquad\qquad L(0.5) = (0.5)^4(0.5)^6 = 0.0009765625$$

Therefore, the MLE of $\theta$ is $\hat{\theta} = 0.5$.

**Example 2.7.** In the case when $L(\theta)$ is differentiable on the interior of $\theta$. One possible way of finding an MLE of $\theta = (\theta_1 \ \cdots \ \theta_k)^T$ is to solve the first order equation for $i = 1, \cdots, k$:

$$\frac{\partial}{\partial \theta_i} L(\theta) = 0 \text{ or } \frac{\partial}{\partial \theta_i} l(\theta) = 0$$

and check all the extrema.

**Remark 2.6.2.** Solving the first-order likelihood only gives you the maximum at critical points. You need to also check the extreme values too.

**Example 2.8.** Consider a random sample of size $n$ from $N(\theta, 1)$, where $\theta$ is unknown. We may obtain the log likelihood:

$$l(\theta) = \ln\left(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x_i - \theta)^2}\right) = -\frac{1}{2}\sum_{i=1}^{n}(x_i - \theta)^2 - \frac{n}{2}\ln(2\pi)$$

We find the critical points by solving:

$$0 = \frac{\partial}{\partial \theta}l(\theta) = \sum_{i=1}^{n}(x_i - \theta)$$

This has the solution $\hat{\theta} = \bar{x}$. To check that the solution is in fact a global maximum, we can check that:

$$\frac{\partial^2}{\partial \theta^2}l(\theta) = -n < 0$$

Therefore, the MLE of $\theta$ is $\hat{\theta} = \bar{x}$.

---

**Example 2.9.** Continues the previous example. Alternatively, we may find that for any $\theta \in \Theta$,

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

Thus, for any $\theta \in \Theta$,

$$L(\theta) \leq L(\bar{x})$$

Therefore, the MLE of $\theta$ is $\hat{\theta} = \bar{x}$.

---

**Example 2.10.** Consider a random sample of size $n$ from $N(\theta, 1)$, where $\theta$ is unknown. Previously, we have found that $\hat{\theta} = \bar{x}$, which maximize the log likelihood. Let us restrict that $\theta \geq 0$.
If $\bar{x} \geq 0$, then it satisfies the constraint $\theta \geq 0$. Therefore, the MLE would be:

$$\hat{\theta} = \bar{x}$$

If $\bar{x} < 0$, then it does not satisfy the constraint $\mu \geq 0$. We analyse the log likelihood again:

$$l(\theta) = -\frac{1}{2}\sum_{i=1}^{n}(x_i - \theta)^2 - \frac{n}{2}\ln(2\pi) = -\frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x})^2 - \frac{n}{2}(\bar{x} - \theta) - \frac{n}{2}\ln(2\pi)$$

The term $(\bar{x} - \theta)^2$ is minimized while satisfying the constraint is when $\theta = 0$.
Therefore, if we restrict $\theta \geq 0$, the MLE of $\theta$ would be:

$$\hat{\theta} = \max(\bar{x}, 0)$$

---

**Remark 2.6.3.** Remember, when we estimate a parameter, we must use the data we have obtained.

---

**Example 2.11.** Consider a random sample of size $n$ from $U[0, \theta]$, where $\theta \in (0, \infty)$ is unknown. The likelihood function is:

$$L(\theta) = \frac{1}{\theta^n}\mathbf{1}_{0 \leq x_{(1)} \leq \cdots \leq x_{(n)} \leq \theta}$$

where $x_{(i)}$ represents the $i$-th smallest data for $i = 1, \cdots, n$. Therefore, the MLE is:

$$\hat{\theta} = x_{(n)}$$

---

**Remark 2.6.4.** MLE may be biased and it may not exist in $\Theta$, especially when $\Theta$ is an open set.

**Remark 2.6.5.** MLE defined may not be unique.

**Example 2.12.** Consider a random sample of size $n$ from $U[\theta - 1, \theta + 1]$, where $\theta$ is unknown. The likelihood function is:

$$L(\theta) = \frac{1}{2^n} \mathbf{1}_{\theta - 1 \leq x_{(1)} \leq \cdots \leq x_{(n)} \leq \theta + 1} = \frac{1}{2^n} \mathbf{1}_{x_{(n)} - 1 \leq \theta \leq x_{(1)} + 1}$$

where $x_{(i)}$ represents the $i$-th smallest data for $i = 1, \cdots, n$. We may find that any estimates in $[x_{(n)} - 1, x_{(1)} + 1]$ maximize $L(\theta)$. Therefore, there are infinitely many MLEs of $\theta$.

**Lemma 2.7. (Invariance property of MLE)** If $\hat{\theta}_i$ is the MLE of $\theta_i$ for $i = 1, \cdots, k$, then $h(\hat{\theta}_1, \cdots, \hat{\theta}_k)$ is the MLE for $h(\theta_1, \cdots, \theta_k)$, where $h$ is a known function.

**Theorem 2.8.** A sequence of MLE $\{\hat{\theta}_n \in \mathbb{R}^k\}$ is consistent, asymptotically unbiased for $\theta$, asymptotically efficient and asymptotically normally distributed. More precisely, under regularity assumption, as $n \to \infty$, we have:

$$\sqrt{n}(\hat{\theta}_n - \theta) \to N_k(\mathbf{0}, \mathcal{I}_X^{-1}(\theta))$$

where $\mathcal{I}_X(\theta)$ is known as the **Fisher Information matrix** and is a $k \times k$ matrix with the $(i, j)$-th entry defined as:

$$\begin{cases} \mathbb{E}\left[\left(\frac{\partial}{\partial \theta_i} \ln f_X(X|\theta)\right)\left(\frac{\partial}{\partial \theta_j} \ln f_X(X|\theta)\right)\right], & \text{Continuous case} \\ \mathbb{E}\left[\left(\frac{\partial}{\partial \theta_i} \ln p_X(X|\theta)\right)\left(\frac{\partial}{\partial \theta_j} \ln p_X(X|\theta)\right)\right], & \text{Discrete case} \end{cases}$$

for $i = 1, \cdots, k$ and $j = 1, \cdots, k$.

**Remark 2.8.1.** In the theorem, "asymptotically efficient" means that the limiting variance is the smallest possible. This will be further discussed in Chapter 3.

Notice that we have used a special matrix called "Fisher Information Matrix". What is Fisher Information?

**Definition 2.9.** Given a set of random variables $\{X_1, \cdots, X_n\}$. The **Fisher Information**, or **Fisher Information matrix** if more than one unknown parameter is considered, of the set is defined by:

$$\mathcal{I}_{X_1, \cdots, X_n}(\theta) = \begin{cases} \mathbb{E}\left[\frac{d}{d\theta} \ln f_{X_1, \cdots, X_n}(X_1, \cdots, X_n|\theta)\right]^2, & \text{Continuous case} \\ \mathbb{E}\left[\frac{d}{d\theta} \ln p_{X_1, \cdots, X_n}(X_1, \cdots, X_n|\theta)\right]^2, & \text{Discrete case} \end{cases}$$

**Remark 2.9.1.** Fisher Information is a measure of amount of information about an unknown parameter $\theta$ that a random variable or data carries. It is very important because we do want to know how to quantify this amount appropriately.

**Example 2.13.** If $X \sim N(\mu, \sigma^2)$, where $\sigma^2$ is known but $\mu \in (-\infty, \infty)$ is unknown, then the Fisher Information about $\mu$ contained in $X$ is:

$$\mathcal{I}_X(\mu) = \mathbb{E}\left[\frac{d}{d\mu} \ln f_X(X|\mu)\right]^2 = \mathbb{E}\left[\frac{d}{d\mu}\left(-\frac{1}{2\sigma^2}(X - \mu)^2 - \frac{1}{2}\ln(2\pi\sigma^2)\right)\right]^2 = \mathbb{E}\left[\frac{1}{\sigma^2}(X - \mu)\right]^2 = \frac{1}{\sigma^2}$$

**Example 2.14.** If $X \sim \text{Bern}(p)$, where $p \in (0, 1)$ is unknown, then the Fisher Information about $p$ contained in $X$ is:

$$\mathcal{I}_X(p) = \mathbb{E}\left[\frac{d}{dp} \ln f_X(X|p)\right]^2 = \mathbb{E}\left[\frac{d}{dp}\left(X \ln p + (1 - X)\ln(1 - p)\right)\right]^2$$

$$= \mathbb{E}\left[\frac{X}{p} - \frac{1 - X}{1 - p}\right]^2$$

$$= \mathbb{E}\left[\frac{X - p}{p(1 - p)}\right]^2$$

$$= \frac{p(1 - p)}{p^2(1 - p)^2} = \frac{1}{p(1 - p)}$$

We will see some properties of the Fisher Information. *For simplicity, we will only discuss continuous random variables.* Notice that we used something called "regularity assumption"? The following are the regularity conditions that we need.

1. $\frac{d}{d\theta} \ln f_{X_1,\cdots,X_n}(x_1,\cdots,x_n|\theta)$ exists for all $x_1,\cdots,x_n$ and all $\theta \in \Theta$

2. For any statistic $T(x_1,\cdots,x_n)$,

$$\frac{d}{d\theta} \int \cdots \int T(x_1,\cdots,x_n) f_{X_1,\cdots,X_n}(x_1,\cdots,x_n|\theta)\, dx_1 \cdots dx_n$$
$$= \int \cdots \int T(x_1,\cdots,x_n)\frac{d}{d\theta} f_{X_1,\cdots,X_n}(x_1,\cdots,x_n|\theta)\, dx_1 \cdots dx_n$$

3. $0 < \mathcal{I}_{X_1,\cdots,X_n}(\theta) < \infty$ for all $\theta \in \Theta$

Condition 2 can be satisfied when the support of $X$ does not depend on $\theta$, where the support of $X$ is defined in below:

**Definition 2.10.** Suppose $X$ is a random variable with a PMF $p(x)$ or a PDF $f(x)$. The **support** of $X$ is defined as:
$$\text{supp}(X) = \begin{cases} \{x : p_X(x) > 0\}, & \text{Discrete case} \\ \{x : f_X(x) > 0\}, & \text{Continuous case} \end{cases}$$

**Lemma 2.11.** Suppose the $X$ is a random variable with PDF $f_X$. Under the regularity condition, we have:
$$\mathbb{E}\left[\frac{d}{d\theta} \ln f_X(X|\theta)\right] = 0$$

*Proof.*

$$0 = \frac{d}{d\theta} \int_{-\infty}^{\infty} f_X(x|\theta)\, dx \int_{-\infty}^{\infty} \frac{d}{d\theta} f_X(x|\theta)\, dx = \int_{-\infty}^{\infty} \left(\frac{d}{d\theta} \ln f_X(x|\theta)\right) f_X(x|\theta)\, dx = \mathbb{E}\left[\frac{d}{d\theta} \ln f_X(X|\theta)\right]$$

□

**Remark 2.11.1.** Using this lemma, we can find that:
$$\mathcal{I}_X(\theta) = \text{Var}\left(\frac{d}{d\theta} \ln f_X(X|\theta)\right)$$

**Lemma 2.12.** Suppose that $\{X_1,\cdots,X_n\}$ is a set of random variables. Under the regularity conditions and the assumption that $\frac{d^2}{d\theta^2} \ln f_{X_1,\cdots,X_n}(x_1,\cdots,x_n|\theta)$ exists for all $x_1,\cdots,x_n$ and all $\theta \in \Theta$, we have:
$$\mathbb{E}\left[\frac{d}{d\theta} \ln f_X(X|\theta)\right]^2 = -\mathbb{E}\left[\frac{d^2}{d\theta^2} \ln f_X(X|\theta)\right]$$

*Proof.*
From the proof of last Lemma,

$$0 = \frac{d}{d\theta} \int_{-\infty}^{\infty} \left(\frac{d}{d\theta} \ln f_X(x|\theta)\right) f_X(x|\theta)\, dx$$
$$= \int_{-\infty}^{\infty} \frac{d}{d\theta} \left[\left(\frac{d}{d\theta} \ln f_X(x|\theta)\right) f_X(x|\theta)\right] dx$$
$$= \int_{-\infty}^{\infty} \left(\frac{d^2}{d\theta^2} \ln f_X(x|\theta)\right) f_X(x|\theta)\, dx + \int_{-\infty}^{\infty} \left(\frac{d}{d\theta} \ln f_X(x|\theta)\right) \frac{d}{d\theta} f_X(x|\theta)\, dx$$
$$= \int_{-\infty}^{\infty} \left(\frac{d^2}{d\theta^2} \ln f_X(x|\theta)\right) f_X(x|\theta)\, dx + \int_{-\infty}^{\infty} \left(\frac{d}{d\theta} \ln f_X(x|\theta)\right)^2 f_X(x|\theta)\, dx$$
$$= \mathbb{E}\left[\frac{d^2}{d\theta^2} \ln f_X(X|\theta)\right] + \mathbb{E}\left[\frac{d}{d\theta} \ln f_X(X|\theta)\right]^2$$
$$-\mathbb{E}\left[\frac{d^2}{d\theta^2} \ln f_X(X|\theta)\right] = \mathbb{E}\left[\frac{d}{d\theta} \ln f_X(X|\theta)\right]^2$$

□

Assume that we consider two independent random variables $X$ and $Y$. We can find the Fisher Information about $\theta$ contained in $(X, Y)$ by finding the Fisher Information about $\theta$ contained in each of them.

> **Lemma 2.13.** If $X$ and $Y$ are independent and their PDFs satisfy the regularity conditions, then:
> $$\mathcal{I}_{X,Y}(\theta) = \mathcal{I}_X(\theta) + \mathcal{I}_Y(\theta)$$

*Proof.*
Since $X$ and $Y$ are independent,

$$
\begin{aligned}
\mathcal{I}_{X,Y}(\theta) &= \mathbb{E}\left[\frac{d}{d\theta}\ln f_{X,Y}(X,Y|\theta)\right]^2 \\
&= \mathbb{E}\left[\frac{d}{d\theta}\ln f_X(X|\theta) + \frac{d}{d\theta}\ln f_Y(Y|\theta)\right]^2 \\
&= \mathbb{E}\left[\frac{d}{d\theta}\ln f_X(X|\theta)\right]^2 + 2\,\mathbb{E}\left[\left(\frac{d}{d\theta}\ln f_X(X|\theta)\right)\left(\frac{d}{d\theta}\ln f_Y(Y|\theta)\right)\right] + \mathbb{E}\left[\frac{d}{d\theta}\ln f_Y(Y|\theta)\right]^2 \\
&= \mathbb{E}\left[\frac{d}{d\theta}\ln f_X(X|\theta)\right]^2 + \mathbb{E}\left[\frac{d}{d\theta}\ln f_Y(Y|\theta)\right]^2 \qquad\qquad \text{(Lemma 2.11)} \\
&= \mathcal{I}_X(\theta) + \mathcal{I}_Y(\theta)
\end{aligned}
$$

$\square$

By applying the same result to a random sample of size $n$, we can obtain the following property.

> **Lemma 2.14.** Suppose the $\{X_1, \cdots, X_n\}$ is a random sample of size $n$ from a distribution. Then,
> $$\mathcal{I}_{X_1,\cdots,X_n}(\theta) = \sum_{i=1}^{n}\mathcal{I}_{X_i}(\theta) = n\mathcal{I}_{X_1}(\theta)$$

> **Remark 2.14.1.** For any $i \neq j$, $\mathcal{I}_{X_i}(\theta) = \mathcal{I}_{X_j}(\theta)$ only means that $X_i$ and $X_j$ carries the same amount of the information about $\theta$. It does not mean they carry identical information.

> **Example 2.15.** Consider a set of i.i.d. random variables $\{X_1, \cdots, X_n\}$ where for all $i = 1, \cdots, n$, $X_i \sim$ Cauchy$(\theta)$ and has a PDF:
> $$f_{X_i}(x|\theta) = \frac{1}{\pi(1 + (x-\theta)^2)}$$
>
> We may find that:
>
> $$
> \begin{aligned}
> \mathcal{I}_{X_i}(\theta) = \mathbb{E}\left[\frac{d}{d\theta}\ln f_{X_i}(X_i|\theta)\right]^2 &= \mathbb{E}\left(\frac{\frac{2(X_i-\theta)}{\pi(1+(X_i-\theta)^2)^2}}{\frac{1}{\pi(1+(X_i-\theta)^2)}}\right)^2 \\
> &= \mathbb{E}\left(\frac{2(X_i-\theta)}{1+(X_i-\theta)^2}\right)^2 \\
> &= \int_{-\infty}^{\infty}\left(\frac{2(x-\theta)}{1+(x-\theta)^2}\right)^2 \frac{1}{\pi(1+(x-\theta)^2)}\,dx \\
> &= \frac{4}{\pi}\int_{-\infty}^{\infty}\frac{u^2}{(1+u^2)^3}\,du && (u = x-\theta,\; du = dx) \\
> &= \frac{8}{\pi}\int_{0}^{\infty}\frac{u^2}{(1+u^2)^3}\,du \\
> &= \frac{4}{\pi}\int_{0}^{1}\sqrt{y}\sqrt{1-y}\,dy && (y = \tfrac{1}{1+u^2},\; dy = -\tfrac{2u}{(1+u^2)^2}\,du) \\
> &= \frac{4}{\pi}\int_{0}^{1}(y)^{\frac{3}{2}-1}(1-y)^{\frac{3}{2}-1}\,dy && \text{(Beta integral)} \\
> &= \frac{4\Gamma(\frac{3}{2})\Gamma(\frac{3}{2})}{\pi\Gamma(3)} = \frac{4(0.5\sqrt{\pi})^2}{2!} = \frac{1}{2} && \left(\tfrac{\Gamma(z_1)\Gamma(z_2)}{\Gamma(z_1+z_2)} = \int_0^1 t^{z_1-1}(1-t)^{z_2-1}\,dt\right)
> \end{aligned}
> $$
>
> Therefore, $I_{X_1,\cdots,X_n}(\theta) = nI_{X_1}(\theta) = \frac{n}{2}$.

Note that a statistic or an estimator can be considered as a function for data condensation because we condense a random sample into a lower-dimensional quantity.

> **Lemma 2.15.** Suppose that $\mathbf{X}$ is a random vector. Under the regularity conditions, for any statistic $T(\mathbf{X})$ for $\theta$, we have:
> $$\mathcal{I}_{T(\mathbf{X})}(\theta) \leq \mathcal{I}_{\mathbf{X}}(\theta)$$

> **Remark 2.15.1.** The Fisher Information of $T(\mathbf{X})$ is defined by:
> $$\mathcal{I}_{T(\mathbf{X})}(\theta) = \mathbb{E}\left[\frac{\mathrm{d}}{\mathrm{d}\theta}\ln f_{T(\mathbf{X})}(T(\mathbf{X})|\theta)\right]^2$$

We may prove Theorem 2.8 in one-parameter case:

> **Theorem 2.16.** Consider a random sample $\{X_1, \cdots, X_n\}$ of size $n$ from a parametric distribution with a PDF $f_X$. Then under the regularity and some other conditions, for $\theta \in \mathbb{R}$, a sequence of MLE $\{\hat{\theta}_n \in \mathbb{R}\}$ allows,
> $$\sqrt{n}(\hat{\theta}_n - \theta) \to \mathrm{N}\left(0, \frac{1}{\mathcal{I}_X(\theta)}\right)$$

*Proof.*
Since the MLE $\hat{\theta}_n$ is the solution to $l'(\theta) = 0$, we can apply a Taylor expansion of $l'(\hat{\theta}_n)$ at $\theta$ to find that:

$$0 = l'(\theta) + l''(\theta)(\hat{\theta}_n - \theta) + o((\hat{\theta}_n - \theta))$$

$$\sqrt{n}(\hat{\theta}_n - \theta) = \frac{\frac{1}{\sqrt{n}}l'(\theta)}{-\frac{1}{n}l''(\theta)} - o((\hat{\theta}_n - \theta))$$

We first consider the numerator. Note that $\frac{\mathrm{d}}{\mathrm{d}\theta}\ln f_X(X_1|\theta), \cdots, \frac{\mathrm{d}}{\mathrm{d}\theta}\ln f_X(X_1|\theta)$ are i.i.d.. By CLT, we have:

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n \frac{\mathrm{d}}{\mathrm{d}\theta}\ln f_X(X_i|\theta) - \mathbb{E}\left[\frac{\mathrm{d}}{\mathrm{d}\theta}\ln f_X(X_1|\theta)\right]\right) \to \mathrm{N}\left(0, \mathrm{Var}\left[\frac{\mathrm{d}}{\mathrm{d}\theta}\ln f_X(X_1|\theta)\right]\right)$$

By Lemma 2.11, we have:

$$\frac{1}{\sqrt{n}}l'(\theta) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{\mathrm{d}}{\mathrm{d}\theta}\ln f_X(X_i|\theta) \to \mathrm{N}(0, \mathcal{I}_X(\theta))$$

Now we consider the denominator. By WLLN and Lemma 2.12, since $\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\ln f_X(X_1|\theta), \cdots, \frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\ln f_X(X_1|\theta)$ are i.i.d.,

$$-\frac{1}{n}l''(\theta) = -\frac{1}{n}\sum_{i=1}^n \frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\ln f_X(X_i|\theta) \to -\mathbb{E}\left[\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\ln f_X(X|\theta)\right] = \mathbb{E}\left[\frac{\mathrm{d}}{\mathrm{d}\theta}\ln f_X(X|\theta)\right]^2 = \mathcal{I}_X(\theta)$$

Consequently, we have:

$$\sqrt{n}(\hat{\theta}_n - \theta) = \frac{\frac{1}{\sqrt{n}}l'(\theta)}{-\frac{1}{n}l''(\theta)} - o((\hat{\theta}_n - \theta)) \to \mathrm{N}\left(0, \frac{1}{\mathcal{I}_X(\theta)}\right)$$

$\square$

> **Remark 2.16.1.** Sometime, $\mathcal{I}_X(\theta)$ cannot be determined easily. We will replace it by an observed Fisher Information defined by $-\frac{1}{n}l''(\hat{\theta}_n)$. Since $\hat{\theta}_n$ is consistent for $\theta$, $-\frac{1}{n}l''(\hat{\theta}_n)$ is also consistent for $\mathcal{I}_X(\theta)$ by continuous mapping theorem. Therefore,
> $$\sqrt{-l''(\hat{\theta}_n)}(\hat{\theta}_n - \theta) = \sqrt{n}\sqrt{-\frac{1}{n}l''(\hat{\theta}_n)}(\hat{\theta}_n - \theta) \to \mathrm{N}(0, 1)$$

**Example 2.16.** (**Principal of Numerical solution to likelihood equations**) Consider a random sample of size $n$ from $X \sim \text{Cauchy}(\theta)$, similar to the previous example. We want to find the MLE of $\theta$. We have:

$$l(\theta) = -n \ln \pi - \sum_{i=1}^{n} \ln(1 + (x_i - \theta)^2)$$

We want to find the solution of $l'(\theta) = 0$, which is the MLE. Setting:

$$\sum_{i=1}^{n} \frac{2(x_i - \theta)}{1 + (x_i - \theta)^2}$$

However, this cannot be solved explicitly in this case.

---

**Example 2.17.** (**Newton-Raphson Algorithm**) By Taylor expansion, we can get:

$$0 = \frac{1}{n} l'(\hat{\theta}) \approx \frac{1}{n} l'(\theta) + \frac{1}{n}(\hat{\theta} - \theta) l''(\theta)$$

We may modify it to find that:

$$\hat{\theta} \approx \theta - \frac{l'(\theta)}{l''(\theta)}$$

We may initially guess a number, say $\theta_0$. By iteratively applying the procedure for $j = 0, 1, \cdots$:

$$\theta_{j+1} = \theta_j - \frac{l'(\theta_j)}{l''(\theta_j)}$$

and stop it at a certain stopping criterion, say $|\theta_{j+1} - \theta_j| < K$ for some chosen constant $K$. E.g. $K = 10^{-5}$. Using this algorithm, we can obtain or approximate the MLE of $\theta$.

---

**Example 2.18.** Consider a random sample of size $n$ from $X \sim \text{Gamma}(\alpha, \beta)$, where $\beta = 3423$ and $\alpha$ is unknown. The PDF is defined by:

$$f_X(x|\theta) = \begin{cases} \frac{3423^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-3423x}, & x > 0 \\ 0, & \text{Otherwise} \end{cases}$$

We can find the log likelihood:

$$l(\alpha) = \sum_{i=1}^{n} \ln f_X(x_i|\alpha) = n\alpha \ln 3423 - n \ln(\Gamma(\alpha)) + (\alpha - 1) \sum_{i=1}^{n} \ln x_i - 3423 \sum_{i=1}^{n} x_i$$

To find MLE, we try to solve the equation:

$$0 = \frac{\mathrm{d}}{\mathrm{d}\alpha} l(\alpha) = n \ln 3423 - n \frac{\mathrm{d}}{\mathrm{d}\alpha} \ln(\Gamma(\alpha)) + \sum_{i=1}^{n} \ln x_i$$

However, we have no idea how to find $\frac{\mathrm{d}}{\mathrm{d}\alpha} \ln(\Gamma(\alpha))$. Therefore, instead, we would use the numerical methods to approximate the MLE.

# Chapter 3

# Uniformly Minimum Variance Unbiased Estimator

We usually want to find the best estimator that can approximate some parameters. However, there are a lot of estimators we can provide based on the information given. In this chapter, we would try to find the best out of them.

## 3.1 Introduction to UMVUE

Consider a class $M$ defined as all the estimators for $\theta$. If there exists an estimator $\hat{\theta}^* \in M$ that is uniformly better than any other estimator in $M$, then we say $\hat{\theta}^*$ is the best estimator of $\theta$ in $M$. However, in general, this estimator does not exist partly because there are too many estimators to consider and some of them are poor or not reasonable. To avoid this problem, we only consider a particular class of estimators, which is the mean-unbiased estimators.

**Remark 3.0.1.** In this context, $\hat{\theta}^*$ is "uniformly better" means that $\mathrm{Var}(\hat{\theta}^*) < \mathrm{Var}(\hat{\theta})$ for any other $\hat{\theta} \in M$.

Recall the definition of mean-unbiased estimator. If an estimator $\hat{\theta}$ satisfies:

$$\mathbb{E}(\hat{\theta}) = \theta$$

for all $\theta \in \Theta$, then it is mean-unbiased or simply unbiased for $\theta$. Otherwise, it is biased.
From past experiences, we may get the following remarks.

**Remark 3.0.2.** Unbiasedness means that by repeated sampling, $\hat{\theta} = \theta$ on average. The underestimation and overestimation will balance in the long run.

**Remark 3.0.3.** Sample variance $S_{n-1}^2$ is unbiased for $\sigma^2$ but $S_n^2$ is not. This is why we use $S_{n-1}^2$ to estimate $\sigma^2$ instead of $S_n^2$.

**Remark 3.0.4.** MME and MLE are usually biased, but they are asymptotically unbiased.

**Remark 3.0.5.** It is possible to have infinitely many different unbiased estimators for $\theta$.

**Example 3.1.** Consider $\{X_1, \cdots, X_n\}$ a random sample of size $n$ from a distribution with a finite mean $\theta$. Any estimators $\hat{\theta}$ in the form of:
$$\hat{\theta} = \frac{\sum_{i=1}^n a_i X_i}{\sum_{i=1}^n a_i}$$
where $a_i \in \mathbb{R}$ for $i = 1, \cdots, n$ and $\sum_{i=1}^n a_i \neq 0$, are unbiased for $\theta$.

**Remark 3.0.6.** It is possible to have no unbiased estimators for $\theta$.

**Example 3.2.** Consider a random sample of size $n$ from a random variable $X \sim \mathrm{Bin}(1, \theta)$ with $g(\theta) = \frac{\theta}{1-\theta}$ as the parameter being estimated. There does not exist an unbiased estimator for $g(\theta)$.

**Remark 3.0.7.** Unbiasedness does not have an invariance property. $\hat{\theta}$ is unbiased for $\theta$ does not mean $h(\hat{\theta})$ is unbiased for $h(\theta)$.

**Example 3.3.** We have $\overline{X}$ is unbiased for $\mu$, but $(\overline{X})^2$ is not unbiased for $\mu^2$ when $\sigma > 0$.

The best unbiased estimator is the unbiased estimator with the smallest variance.

**Definition 3.1.** The **Uniformly Minimum Variance Unbiased Estimator (UMVUE)** $\hat{\theta}^*$ for $\theta$ is an unbiased estimator such that for all other unbiased estimator $\hat{\theta}$ for $\theta$,

$$\text{Var}(\hat{\theta}^*) \leq \text{Var}(\hat{\theta})$$

for all $\theta \in \Theta$.

**Lemma 3.2.** (**Uniqueness of UMVUE**) Assume that UMVUE for $\theta$ exists. Then, it is unique.

*Proof.*
Assume that there are two distinct UMVUEs, $\hat{\theta}^*$ and $\hat{\theta}^{**}$, for $\theta$. We may find that for all $\theta \in \theta$ and any unbiased estimator $\hat{\theta}$ of $\theta$:

$$\text{Var}(\hat{\theta}^*) = \text{Var}(\hat{\theta}^{**}) \leq \text{Var}(\hat{\theta})$$

Let $\hat{\theta}' = \frac{1}{2}(\hat{\theta}^* + \hat{\theta}^{**})$. We can easily find that $\hat{\theta}'$ is unbiased for $\theta$. We have:

$$\begin{aligned}
\text{Var}(\hat{\theta}') &= \frac{1}{4}\text{Var}(\hat{\theta}^*) + \frac{1}{4}\text{Var}(\hat{\theta}^{**}) + \frac{1}{2}\text{cov}(\hat{\theta}^*, \hat{\theta}^{**}) \\
&\leq \frac{1}{2}\text{Var}(\hat{\theta}^*) + \frac{1}{2}\sqrt{\text{Var}(\hat{\theta}^*)\text{Var}(\hat{\theta}^{**})} \\
&\leq \text{Var}(\hat{\theta}^*) \leq \text{Var}(\hat{\theta}')
\end{aligned}$$

Thus $\text{Var}(\hat{\theta}') = \text{Var}(\hat{\theta}^*)$ and $\text{cov}(\hat{\theta}^*, \hat{\theta}^{**}) = \sqrt{\text{Var}(\hat{\theta}^*)\text{Var}(\hat{\theta}^{**})}$.
We recall the Pearson correlation coefficient. We can find that:

$$\rho = \frac{\text{cov}(\hat{\theta}^*, \hat{\theta}^{**})}{\sqrt{\text{Var}(\hat{\theta}^*)\text{Var}(\hat{\theta}^{**})}} = 1$$

This means $\hat{\theta}^*$ and $\hat{\theta}^{**}$ have a perfectly linear positive relationship. We say $\hat{\theta}^* = a\hat{\theta}^{**} + b$ where $a > 0$ and $b \in \mathbb{R}$. We solve the equation:

$$\begin{cases}
\text{Var}(\hat{\theta}^{**}) = \text{Var}(\hat{\theta}^*) = a^2\text{Var}(\hat{\theta}^{**}) \\
\mathbb{E}(\hat{\theta}^{**}) = \theta = \mathbb{E}(\hat{\theta}^*) = a\,\mathbb{E}(\hat{\theta}^{**}) + b
\end{cases}$$

We may find that $a = 1$ and $b = 0$. Therefore, $\hat{\theta}^* = \hat{\theta}^{**}$ and UMVUE is unique.          $\square$

## 3.2   Sufficient Statistic

It is not easy to find UMVUE for a parameter being estimated. However, we have Rao-Blackwell Theorem (we will talk about it later) which tells us that UMVUE must be a function of sufficient statistic. Let us discuss about sufficient statistic.
Note that a statistic or an estimator can be considered as a function for data condensation because we condense a random sample into a lower-dimensional quantity. However, in the process, we may lose some information about the parameter $\theta$.
We recall this lemma. Under regularity conditions, for any statistic $T = T(\mathbf{X})$ for $\theta$, we have:

$$\mathcal{I}_T(\theta) \leq \mathcal{I}_{\mathbf{X}}(\theta)$$

Most statistics would lose some information about $\theta$, but there exists some statistics that we can substantially reduce the dimension without losing any information. We call this sufficient statistic.

**Definition 3.3.** Under regularity conditions, the **sufficient statistic** for $\theta$, denoted by $S = S(\mathbf{X})$, is a statistic that satisfies:
$$\mathcal{I}_S(\theta) = \mathcal{I}_{\mathbf{X}}(\theta)$$

**Remark 3.3.1.** If the conditional distribution of the sample given a statistic $T$ depends on $\theta$, then there is still some information about $\theta$ contained in the sample that $T$ does not carry. Therefore, $T$ is not sufficient.

**Example 3.4.** Let $\{X_1, X_2\}$ be a random sample of size 2 from $X \sim \text{Bin}(m, \theta)$. We show that $T = T(X_1, X_2) = X_1 + X_2$ is sufficient.

$$
\begin{aligned}
p_{X_1, X_2 | T}(x_1, x_2 | t) &= \frac{p_{X_1, X_2, T}(x_1, x_2, t)}{p_T(t)} \\
&= \frac{p_{X_1, X_2, T}(x_1, t - x_1, t)}{p_T(t)} \\
&= \frac{p_{X_1}(x_1) p_{X_2}(t - x_1)}{p_T(t)} \\
&= \frac{\binom{m}{x_1}\binom{m}{t-x_1}\theta^t(1-\theta)^{n-t}}{\binom{2m}{t}\theta^t(1-\theta)^{n-t}} = \frac{\binom{m}{x_1}\binom{m}{t-x_1}}{\binom{2m}{t}}
\end{aligned}
$$

Therefore, since the conditional distribution of the sample given a statistic $T$ does not depend on $\theta$, we can find that $T$ is a sufficient statistic.

We may rewrite the definition of sufficient statistic as follows:

**Definition 3.4.** Let $\mathbf{X} = \{X_1, \cdots, X_n\}$ be a random vector of random sample of size $n$ from a PDF $f(x|\theta)$ or PMF $p(x|\theta)$, where $\theta \in \Theta \subset \mathbb{R}^k$ for integer $k > 1$. A set of statistics $\{S_1, S_2, \cdots, S_r\}$, where $r \geq k$ and $S_i = S_i(\mathbf{X})$ for $i = 1, \cdots, r$, is said to be **jointly sufficient**, if and only if the conditional distribution:

$$
\begin{cases}
p_{\mathbf{X}|S_1,\cdots,S_r}(\mathbf{x}|S_1 = s_1, \cdots, S_r = s_r, \theta), & \text{Discrete case} \\
f_{\mathbf{X}|S_1,\cdots,S_r}(\mathbf{x}|S_1 = s_1, \cdots, S_r = s_r, \theta), & \text{Continuous case}
\end{cases}
$$

does not depend on $\theta$, for all values $s_1$ of $S_1$, $\cdots$, $s_r$ of $S_r$.

or in one-parameter case,

**Definition 3.5.** Let $\mathbf{X} = \{X_1, \cdots, X_n\}$ be a random sample of size $n$ from a PDF $f(x|\theta)$ or PMF $p(x|\theta)$, where $\theta \in \Theta \subset \mathbb{R}$. A statistic $S = S(\mathbf{X})$ is said to be **sufficient** if and only if the conditional distribution:

$$
\begin{cases}
p_{\mathbf{X}|S}(\mathbf{x}|S = s, \theta), & \text{Discrete case} \\
f_{\mathbf{X}|S}(\mathbf{x}|S = s, \theta), & \text{Continuous case}
\end{cases}
$$

does not depend on $\theta$, for all values $s$ of $S$.

**Theorem 3.6. (Fisher-Neyman Factorization Theorem)** Let $\mathbf{X} = \{X_1, \cdots, X_n\}$ be a random sample of size $n$ from a PDF $f(x|\theta)$ or a PMF $p(x|\theta)$, where $\theta \in \Theta \subset \mathbb{R}^k$. A set of statistics $\{S_1, \cdots, S_r\}$, where $r \geq k$ and $S_i = S_i(\mathbf{X})$ for $i = 1, \cdots, n$, is jointly sufficient if and only if:

$$
\begin{cases}
p_{\mathbf{X}}(\mathbf{x}|\theta) = g(S_1(\mathbf{x}), \cdots, S_r(\mathbf{x})|\theta)h(\mathbf{x}), & \text{Discrete case} \\
f_{\mathbf{X}}(\mathbf{x}|\theta) = g(S_1(\mathbf{x}), \cdots, S_r(\mathbf{x})|\theta)h(\mathbf{x}), & \text{Continuous case}
\end{cases}
$$

where $g$ is a non-negative function of $x_1, \cdots, x_n$ only through the statistics $S_1, \cdots, S_r$ and depends on $\theta$, and $h$ is a non-negative function of $x_1, \cdots, x_n$ not depending on $\theta$.

*Proof.*
We shall prove it in discrete case with only one statistic. The proof in continuous case or more than one statistic is out of our scope.
Note that if $\mathbf{X} \in B$ for a set $B$, then $S(X) \in S(B)$. Therefore, for $i = 1, \cdots, n$,

$$
\{\mathbf{X} \in B\} \cap \{S(\mathbf{X}) \in S(B)\} = \{\mathbf{X} \in B\}
$$

$$
\begin{aligned}
\mathbb{P}(\mathbf{X} \in B|\theta) &= \mathbb{P}(\mathbf{X} \in B, S(\mathbf{X}) \in S(B)|\theta) \\
&= \mathbb{P}(\mathbf{X} \in B|S(\mathbf{X}) \in S(B), \theta)\,\mathbb{P}(S(\mathbf{X}) \in S(B)|\theta) \qquad (\mathbb{P}(A \cap B|D) = \mathbb{P}(A|B \cap D)\,\mathbb{P}(B \cap D))
\end{aligned}
$$

If $S$ is sufficient, then by definition,

$$\mathbb{P}(\mathbf{X} \in B | S(\mathbf{X}) \in S(B), \theta) = \mathbb{P}(X_i \in B | S(X_i) \in S(B))$$

Therefore, by substituting $B = \{\mathbf{x}\}$, we get:

$$p_{\mathbf{X}}(\mathbf{x}|\theta) = p_{\mathbf{X}|S}(\mathbf{x}|S(\mathbf{x}))p_S(S(\mathbf{x})|\theta)$$

We may find that $g(S(\mathbf{x})|\theta) = p_S(S(\mathbf{x})|\theta)$ and $h(\mathbf{x}) = p_{\mathbf{X}|S}(\mathbf{x}|S(\mathbf{x}))$.
If $p_X(\mathbf{x}|\theta) = g(T(\mathbf{x})|\theta)h(\mathbf{x})$, then:

$$p_{\mathbf{X}|T}(\mathbf{x}|t) = \frac{p_{\mathbf{X},T}(\mathbf{x},t|\theta)}{p_T(t|\theta)} = \begin{cases} 0, & t \neq T(\mathbf{x}) \\ \frac{p_{\mathbf{X}}(\mathbf{x}|\theta)}{p_T(t|\theta)}, & t = T(\mathbf{x}) \end{cases}$$

We only need to consider when $t = T(\mathbf{x})$, we have:

$$p_{\mathbf{X}|T}(\mathbf{x}|t) = \frac{p_{\mathbf{X}}(\mathbf{x}|\theta)}{\sum_{\mathbf{x}:T(\mathbf{x})=t} p_{\mathbf{X}}(\mathbf{x}|\theta)} = \frac{g(t|\theta)h(\mathbf{x})}{\sum_{\mathbf{x}:T(\mathbf{x})=t} g(t|\theta)h(\mathbf{x})} = \frac{h(\mathbf{x})}{\sum_{\mathbf{x}:T(\mathbf{x})=t} h(\mathbf{x})}$$

We have found that $p_{\mathbf{X}|T}(\mathbf{x}|t)$ does not depend on $\theta$. Therefore, by definition, $T$ is sufficient.  □

---

**Example 3.5.** Let $\{X_1, \cdots, X_n\}$ be a random sample from Bern$(\theta)$, where $\theta \in [0,1]$ is unknown. The joint PMF of the random sample is:

$$p_{X_1,\cdots,X_n}(x_1,\cdots,x_n|\theta) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{\sum_{i=1}^{n} x_i}(1-\theta)^{n-\sum_{i=1}^{n} x_i} \times 1$$

$$= g\left(\sum_{i=1}^{n} x_i \,\middle|\, \theta\right) \times h(x_1,\cdots,x_n)$$

Therefore, $S = \sum_{i=1}^{n} X_i$ is a sufficient statistic.

---

**Example 3.6.** Let $\{X_1, \cdots, X_n\}$ be a random sample from N$(\mu, \sigma^2)$, where $\mu$ and $\sigma^2 > 0$ are unknown. The joint PDF of the random sample is:

$$f_{X_1,\cdots,X_n}(x_1,\cdots,x_n|\mu,\sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)$$

$$= \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right)$$

$$= \sigma^{-n}\exp\left[-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n} x_i^2 - 2\mu\sum_{i=1}^{n} x_i + n\mu^2\right)\right] \times \left(\frac{1}{(2\pi)^{\frac{n}{2}}}\right)$$

$$= g\left(\sum_{i=1}^{n} x_i, \sum_{i=1}^{n} x_i^2 \,\middle|\, \mu,\sigma^2\right) \times h(x_1,\cdots,x_n)$$

Therefore, $S_1 = \sum_{i=1}^{n} X_i$ and $S_2 = \sum_{i=1}^{n} X_i^2$ are jointly sufficient.

---

**Example 3.7.** Let $\{X_1, \cdots, X_n\}$ be a random sample of size $n$ from U$[0,\theta]$, where $\theta > 0$ is unknown. The joint PDF of the random sample is:

$$f_{X_1,\cdots,X_n}(x_1,\cdots,x_n|\theta) = \frac{1}{\theta^n}\prod_{i=1}^{n} \mathbf{1}_{0 \leq x_i \leq \theta}$$

$$= \frac{1}{\theta^n}\mathbf{1}_{0 \leq x_{(1)} < x_{(n)} \leq \theta} \qquad\qquad (x_{(i)} \text{ is the } i\text{-th smallest sample})$$

$$= \frac{1}{\theta^n}\mathbf{1}_{x_{(n)} \leq \theta} \times \mathbf{1}_{x_{(1)} \geq 0}$$

$$= g(x_{(n)}|\theta) \times h(x_1,\cdots,x_n)$$

Therefore, $S = X_{(n)} = \max\{X_1,\cdots,X_n\}$ is sufficient.

**Example 3.8.** Let $\{X_1, \cdots, X_n\}$ be a random sample of size $n$ from $U[\theta - \frac{1}{2}, \theta + \frac{1}{2}]$, where $\theta$ is unknown. The joint PDF of the random sample is:

$$f_{X_1, \cdots, X_n}(x_1, \cdots, x_n | \theta) = \prod_{i=1}^{n} \mathbf{1}_{\theta - \frac{1}{2} \leq x_i \leq \theta + \frac{1}{2}}$$

$$= \mathbf{1}_{\theta - \frac{1}{2} \leq x_{(1)} < x_{(n)} \leq \theta + \frac{1}{2}} \qquad\qquad (x_i \text{ is the } i\text{-th smallest sample})$$

$$= \mathbf{1}_{x_{(n)} - \frac{1}{2} \leq \theta \leq x_{(1)} + \frac{1}{2}} \times 1$$

$$= g(x_{(1)}, x_{(n)} | \theta) \times h(x_1, \cdots, x_n)$$

Therefore, $S_1 = X_{(1)} = \min\{X_1, \cdots, X_n\}$ and $S_2 = X_{(n)} = \max\{X_1, \cdots, X_n\}$ are jointly sufficient.

---

**Example 3.9.** Let $\{X_1, \cdots, X_n\}$ be a random sample of size $n$ from $U[\theta_1, \theta_2]$, where $\theta_1, \theta_2$ are unknown with $\theta_1 < \theta_2$ and $\theta_1$ is not a function of $\theta_2$. The joint PDF of the random sample is:

$$f_{X_1, \cdots, X_n}(x_1, \cdots, x_n | \theta_1, \theta_2) = \frac{1}{(\theta_2 - \theta_1)^n} \prod_{i=1}^{n} \mathbf{1}_{\theta_1 \leq x_i \leq \theta_2}$$

$$= \frac{1}{(\theta_2 - \theta_1)^n} \mathbf{1}_{\theta_1 \leq x_{(1)} < x_{(n)} \leq \theta_2} \times 1 \qquad\qquad (x_i \text{ is the } i\text{-th smallest sample})$$

$$= g(x_{(1)}, x_{(n)} | \theta_1, \theta_2) \times h(x_1, \cdots, x_n)$$

Therefore, $S_1 = X_{(1)} = \min\{X_1, \cdots, X_n\}$ and $S_2 = X_{(n)} = \max\{X_1, \cdots, X_n\}$ are jointly sufficient.

---

**Remark 3.6.1.** Sufficient statistic may not be unique because we may have more than one factorization.

---

**Example 3.10.** Let $\{X_1, \cdots, X_n\}$ be a random sample of size $n$ from $N(\mu, 1)$, where $\mu$ is unknown. The joint PDF of the random sample is:

$$f_{X_1, \cdots, X_n}(x_1, \cdots, x_n | \mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x_i - \mu)^2\right) = \frac{1}{(2\pi)^{\frac{n}{2}}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2\right)$$

One of the ways to factorize is:

$$\frac{1}{(2\pi)^{\frac{n}{2}}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2\right) = \frac{1}{(2\pi)^{\frac{n}{2}}} \exp\left[-\frac{1}{2}\left(\sum_{i=1}^{n}x_i^2 - 2\mu\sum_{i=1}^{n}x_i + n\mu^2\right)\right]$$

$$= \frac{1}{(2\pi)^{\frac{n}{2}}} \exp\left(\mu\sum_{i=1}^{n}x_i - \frac{n}{2}\mu^2\right) \times \exp\left(-\frac{1}{2}\sum_{i=1}^{n}x_i^2\right)$$

$$= g\left(\sum_{i=1}^{n}x_i \,\middle|\, \mu\right) \times h(x_1, \cdots, x_n)$$

Therefore, we find that $S_1 = \sum_{i=1}^{n} X_i$ is sufficient.

Another way to factorize is:

$$\frac{1}{(2\pi)^{\frac{n}{2}}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2\right) = \frac{1}{(2\pi)^{\frac{n}{2}}} \exp\left[-\frac{1}{2}\left(\sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\bar{x} - \mu)^2\right)\right]$$

$$= \frac{1}{(2\pi)^{\frac{n}{2}}} \exp\left(-\frac{n}{2}(\bar{x} - \mu)^2\right) \times \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)$$

$$= g(\bar{x} | \mu) \times h(x_1, \cdots, x_n)$$

Therefore, we find that $S_2 = \overline{X}$ is sufficient.

Based on the above example, we may notice that $\overline{X}$ and $\sum_{i=1}^{n} X_i$ are functions of each other. Is any transformation of $S$ also sufficient? Yes if it is one-to-one!

---

**Lemma 3.7. (One-to-one sufficiency)** Let $\{X_1, \cdots, X_n\}$ be a random sample of size $n$. If a set of statistics $\{S_1, \cdots, S_r\}$, where $r \geq k$ and $S_i = S_i(X_1, \cdots, X_r)$ for $i = 1, \cdots, r$, is jointly sufficient, then any set of one-to-one functions $\{h_1, \cdots, h_m\}$ for some $m$, where $m \geq r$ for $i = 1, \cdots, m$:

$$h_i = h_i(S_1, \cdots, S_r)$$

is also jointly sufficient.

---

**Example 3.11.** Let $\{X_1, \cdots, X_n\}$ be a random sample of size $n$ from a PDF $f(x|\theta)$ or a PMF $p(x|\theta)$, where $\theta \in \Theta \subset \mathbb{R}^k$. Assume that $\sum_{i=1}^{n} X_i$ and $\sum_{i=1}^{n} X_i^2$ are jointly sufficient. We may find that:

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \qquad\qquad \sum_{i=1}^{n} (X_i - \overline{X})^2 = \sum_{i=1}^{n} X_i^2 - \frac{1}{n} \left( \sum_{i=1}^{n} X_i \right)^2$$

Both are one-to-one functions of $\sum_{i=1}^{n} X_i$ and $\sum_{i=1}^{n} X_i^2$.
Therefore, by Lemma 3.7, $\overline{X}$ and $\sum_{i=1}^{n} (X_i - \overline{X})^2$ are jointly sufficient.
However,

$$(\overline{X})^2 = \frac{1}{n^2} \left( \sum_{i=1}^{n} X_i \right)$$

It is not an one-to-one function. Therefore, $(\overline{X})^2$ and $\sum_{i=1}^{n} (X_i - \overline{X})^2$ may not be jointly sufficient.

---

**Remark 3.7.1.** Number of one-to-one functions of statistics can be smaller than the number of the statistics themselves.

---

From previous examples, the number of sufficient statistics can sometimes be more than the number of unknown parameters. How much should data be condense most without losing any information about unknown parameter $\theta$?

---

**Definition 3.8.** A set of jointly sufficient statistics $\{S_1, \cdots, S_n\}$ is **minimal jointly sufficient** if and only if for any other set of jointly sufficient statistics $\{T_1, \cdots, T_m\}$, there exists a set of functions $\{f_1, \cdots, f_n\}$ such that for $i = 1, \cdots, n$,

$$S_i = f_i(T_1, \cdots, T_m)$$

---

or in one-statistic case,

---

**Definition 3.9.** A sufficient statistic $S$ is **minimal sufficient** if and only if for nay other sufficient statistic $T$, there exists a function $f$ such that:

$$S = f(T)$$

---

**Remark 3.9.1.** Minimal jointly sufficient statistics may not be unique. We can say minimal joint sufficiency is closed under any one-to-one transformation.

---

In general, it is not easy to find the minimal jointly sufficient statistics except form some special distribution. One of those special distributions are called exponential family, which we will discuss later.

## 3.3  Relationship of Sufficiency with UMVUE

Recall that we actually want to find UMVUE. Does sufficiency help us find UMVUE? Turns out, by Rao-Blackwell Theorem, it helps us find an improved unbiased estimator!

> **Theorem 3.10. (Rao-Blackwell Theorem)** Let $\mathbf{X} = \{X_1, \cdots, X_n\}$ be a random sample from a PDF $f(x|\theta)$ or a PMF $p(x|\theta)$, where $\theta \in \Theta \subset \mathbb{R}^k$, and a set of jointly sufficient statistics $\{S_1, \cdots, S_r\}$, where $r \geq k$ and $S_i = S_i(\mathbf{X})$ for $i = 1, \cdots, r$. Suppose that $T = T(\mathbf{X})$ is an unbiased estimator for $g(\theta)$ for a function $g$. Define $T'$ by $\mathbb{E}(T|S_1, \cdots, S_r)$. Then:
>
> 1. $T'$ is a statistic, and it is a function of jointly sufficient statistics.
>
> 2. $T'$ is unbiased for $g(\theta)$.
>
> 3. $\mathrm{Var}(T') \leq \mathrm{Var}(T)$.

*Proof.*

1.
$$T' = \mathbb{E}(T|S_1, \cdots, S_r) = \int_{-\infty}^{\infty} t f_{T|S_1, \cdots, S_r}(t|S_1, \cdots, S_r)\, dt$$

   By definition, $T'$ is a statistic and it is a function of jointly sufficient statistic $\{S_1, \cdots, S_r\}$.

2.
$$\mathbb{E}(T') = \mathbb{E}(\mathbb{E}(T|S_1, \cdots, S_r)) = \mathbb{E}(T) = g(\theta)$$

   Therefore, $T'$ is unbiased for $g(\theta)$.

3.
$$\begin{aligned} \mathrm{Var}(T') &= \mathrm{Var}(\mathbb{E}(T|S_1, \cdots, S_r)) \\ &= \mathrm{Var}(T) - \mathbb{E}[\mathrm{Var}(T|S_1, \cdots, S_r)] \leq \mathrm{Var}(T) \qquad (\mathrm{Var}(Y) = \mathrm{Var}(\mathbb{E}(Y|X)) + \mathbb{E}[\mathrm{Var}(Y|X)]) \end{aligned}$$

$\square$

> **Example 3.12.** Let $\{X_1, \cdots, X_n\}$ be a random sample of size $n$ from $\mathrm{Bern}(\theta)$ where $\theta$ is unknown.
>
> 1. Since $\mathbb{E}(X_1) = \theta$, $X_1$ is an unbiased estimator of $\theta$.
>
> 2. From Example 3.5, we have found that $\sum_{i=1}^{n} X_i$ is a sufficient statistic. It is also obvious that it is minimal.
>
> 3. By Rao-Blackwell Theorem, $T' = \mathbb{E}(X_1 | \sum_{i=1}^{n} X_i)$ is an unbiased estimator for $\theta$ with $\mathrm{Var}(T') \leq \mathrm{Var}(X_1)$.
>
> We want to find $T'$. Assume that we have given that $\sum_{i=1}^{n} X_i = s$. We have:
>
> $$\begin{aligned} \mathbb{P}\left(X_1 = 0 \,\middle|\, \sum_{i=1}^{n} X_i = s\right) &= \frac{\mathbb{P}\left(X_1 = 0, \sum_{i=1}^{n} X_i = s\right)}{\mathbb{P}\left(\sum_{i=1}^{n} X_i = s\right)} = \frac{\mathbb{P}(X_1 = 0)\,\mathbb{P}\left(\sum_{i=2}^{n} X_i = s\right)}{\mathbb{P}\left(\sum_{i=1}^{n} X_i = s\right)} \\ &= \frac{(1-\theta)\binom{n-1}{s}\theta^s(1-\theta)^{n-1-s}}{\binom{n}{s}\theta^s(1-\theta)^{n-s}} \\ &= \frac{n-s}{n} \end{aligned}$$
>
> Therefore, we have:
>
> $$\mathbb{E}\left(X_1 \,\middle|\, \sum_{i=1}^{n} X_i = s\right) = \mathbb{P}\left(X_1 = 1 \,\middle|\, \sum_{i=1}^{n} X_i = s\right) = 1 - \mathbb{P}\left(X_1 = 0 \,\middle|\, \sum_{i=1}^{n} X_i = s\right) = \frac{s}{n}$$
>
> We have found that $T' = \mathbb{E}\left(X_1 | \sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} X_i$. We may find that:
>
> $$\mathrm{Var}(T') = \frac{1}{n}\theta(1-\theta) \leq \theta(1-\theta) = \mathrm{Var}(X_1)$$

**Remark 3.10.1.** If $T$ is already a function of a jointly sufficient statistics, then $T'$ would be identical to $T$.

**Example 3.13.** Let $\{X_1, \cdots, X_n\}$ be a random sample of size $n$ from $\text{Bern}(\theta)$ and let $\overline{X}$ be the sample mean. We know that:
$$\mathbb{E}(\overline{X}) = \theta$$
Therefore, $\overline{X}$ is an unbiased estimator of $\theta$. We want to find $T'$. We have:
$$T' = \mathbb{E}\left(\overline{X} \,\middle|\, \sum_{i=1}^{n} X_i\right) = \overline{X} = T$$

**Remark 3.10.2.** Although Rao-Blackwell Theorem can provide us with a constructive way to improve a given unbiased estimator. it does not guarantee that the one constructed must be a UMVUE.

**Example 3.14.** Consider a random sample $\mathbf{X} = \{X_1, \cdots, X_n\}$ of size $n$ from $\text{N}(\theta, 1)$, where $\theta$ is unknown. Let $g(\theta) = \theta$. Consider $T = T(\mathbf{X}) = X_1$ and the random sample is a set of jointly sufficient statistics $\{S_1, \cdots, S_n\}$. We may find that:
$$\mathbb{E}(X_1 | X_1, \cdots, X_n) = X_1 \qquad\qquad \text{(Expectation of } X_1 \text{ given } X_1 \text{ is of course } X_1)$$
However, from Example 3.10, we have found a better statistic $S^* = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ since:
$$\text{Var}(\overline{X}) = \frac{1}{n} \leq \text{Var}(X_1)$$
Therefore, $T' = X_1$ is not a UMVUE.

## 3.4   Complete Statistics

In addition to sufficiency, we would need completeness in order to find the UMVUE.

**Definition 3.11.** Let $\mathbf{X} = \{X_1, \cdots, X_n\}$ be a random sample of size $n$ from a PDF $f(x|\theta)$ and a PMF $p(x|\theta)$, where $\theta \in \Theta \subset \mathbb{R}^k$ for an integer $k > 1$. A set of statistics $\{T_1, \cdots, T_r\}$, where $r \geq k$ and $T_1 = T_1(\mathbf{X})$ for $i = 1, \cdots, r$, is said to be **jointly complete** if and only if for any function $g$:
$$\mathbb{E}[g(T_1, \cdots, T_r)] \text{ for all } \theta \in \Theta \implies \mathbb{P}(g(T_1, \cdots, T_r) = 0) = 1 \text{ for all } \theta \in \Theta$$

or in one-parameter case,

**Definition 3.12.** Let $\mathbf{X} = \{X_1, \cdots, X_n\}$ be a random sample of size $n$ from a PDF $f(x|\theta)$ or a PMF $p(x|\theta)$, where $\theta \in \Theta \subset \mathbb{R}$. A statistic $T = T(\mathbf{X})$ is said to be **complete** if and only if for any function $g$:
$$\mathbb{E}[g(T)] = 0 \text{ for all } \theta \in \Theta \implies \mathbb{P}(g(T) = 0) = 1 \text{ for all } \theta \in \Theta$$

**Remark 3.12.1.** Function $g(T)$ or $g(T_1, \cdots, T_r)$ is not an unbiased estimator for $\theta$.

**Remark 3.12.2.** If there exists a function $g^*$ such that $\mathbb{E}[g^*(T)] = 0$ but $\mathbb{P}(g^*(T) \neq 0) > 0$, then $T$ is not complete. This is the same for joint completeness.

**Example 3.15.** Let $\{X_1, \cdots, X_n\}$ be a random sample of size $n$ from Bern$(\theta)$, where $\theta \in (0,1)$ is unknown. Let $T_1 = X_1 - X_2$. We may easily find that for all $\theta \in (0,1)$:

$$\mathbb{E}(X_1 - X_2) = 0 \text{ but } \mathbb{P}(X_1 - X_2 \neq 0) > 0$$

Therefore, $T_1 = X_1 - X_2$ is not a complete statistic.
Let $T_2 = \sum_{i=1}^n X_i$. For any function $g$,

$$\mathbb{E}[g(T_2)] = \sum_{i=0}^n g(t) \binom{n}{t} \theta^t (1-\theta)^{n-t} = (1-\theta)^n \sum_{i=1}^n g(t) \binom{n}{t} \left(\frac{\theta}{1-\theta}\right)^t$$

Thus, $\mathbb{E}(g(T_2)) = 0$ for all $\theta \in (0,1)$ implies that the equation $\sum_{i=1}^n g(t) \binom{n}{t} \left(\frac{\theta}{1-\theta}\right)^t = 0$ for all $\theta \in (0,1)$.
If not all coefficients $g(t)\binom{n}{t}$ are equal to zero, then there are at most $n$ solutions of the equation for all $\theta \in (0,1)$.
This means only $n$ values of $\theta \in \Theta$ satisfy the equation, but not all $\theta \in (0,1)$.
Therefore, $g(t)\binom{n}{t} = 0$ and thus $g(t) = 0$ for $t = 0, \cdots, n$ and for all $\theta \in (0,1)$.
Since the only possible values of $T_2 = \sum_{i=1}^n X_i$ are in $\{0, \cdots, n\}$, we find that:

$$\mathbb{P}(g(T_2) = 0) = 1$$

We conclude that $T_2 = \sum_{i=1}^n X_i$ is complete.

---

**Example 3.16.** Let $\{X_1, \cdots, X_n\}$ be a random sample of size $n$ from U$[0, \theta]$, where $\theta > 0$ is unknown. We check if the sufficient statistic $X_{(n)}$ is complete. Note that for any function $g$,

$$\mathbb{E}[g(X_{(n)})] = \int_{-\infty}^{\infty} g(y) f_{X_{(n)}}(y)\, dy = \frac{n}{\theta^n} \int_0^\theta g(y) y^{n-1}\, dy$$

Therefore, if $\mathbb{E}[g(X_{(n)})] = 0$ for all $\theta > 0$, then:

$$\int_0^\theta g(y) y^{n-1}\, dy = 0$$

Differentiating both sides with respect to $\theta$ will get $g(\theta)\theta^{n-1} = 0$ and hence $g(\theta) = 0$ for $\theta > 0$. We replace a number $y$ with the parameter $\theta$ and get $g(y) = 0$ for $y \in (0, \theta]$ for all $\theta > 0$.
Since $0 \leq x_{(n)} \leq \theta$, we may find that for all $\theta \in \Theta$:

$$\mathbb{P}(g(X_{(n)}) = 0) = 1$$

Therefore, $X_{(n)}$ is complete.

## 3.5 Exponential Family

In most of the time, it is quite difficult to check the completeness and minimal sufficiency of a statistic by definition, especially for joint completeness. However, there is one special distribution which we can check easily. It is called exponential family.

**Definition 3.13.** Suppose that a random variable $X$ has a PDF $f(x|\theta)$ or a PMF $p(x|\theta)$, where $\theta \in \Theta \subset \mathbb{R}^k$ for an integer $k > 1$. If we find that:

1. supp$(X)$ does not depend on $\theta$.

2. The PDF or PMF of $X$ can be written in the form of:

$$\exp\left(a(\theta) + b(x) + \sum_{j=1}^k c_j(\theta) d_j(x)\right)$$

where $a(\theta)$, $b(x)$, $c_i(\theta)$ and $d_i(x)$ for $i = 1, \cdots, k$, are real-valued functions.

then the distribution of $X$ is said to be an member of $k$-**parameter exponential family**.

or in one-parameter case,

**Definition 3.14.** Suppose that a random variable $X$ has a PDF $f(x|\theta)$ or a PMF $p(x|\theta)$, where $\theta \in \Theta \subset \mathbb{R}$. If we find that:

1. $\text{supp}(X)$ does not depend on $\theta$.

2. The PDF or PMF of $X$ can be written in the form of:

$$\exp[a(\theta) + b(x) + c(\theta)d(x)]$$

where $a(\theta)$, $b(x)$, $c(\theta)$ and $d(x)$ are real-valued functions.

then the distribution of $X$ is said the be an member of **one-parameter exponential family**.

**Remark 3.14.1.** The distribution whose support depends on $\theta$ does not belong to the exponential family. E.g. $U[0, \theta]$.

**Remark 3.14.2.** Most of the parametric distributions we discussed are members of an exponential family. E.g. normal distribution, gamma distribution, Poisson distribution, binomial distribution, and so on.

The following results show that we can find complete and minimal sufficient and statistic from exponential family.

**Theorem 3.15.** Let $\mathbf{X} = \{X_1, \cdots, X_n\}$ be a random sample of size $n$ from a distribution in an one-parameter exponential family with a PDF $f(x|\theta)$ or a PMF $p(x|\theta)$, where $\theta \in \Theta \subset \mathbb{R}$, that can be written in the form of:

$$\exp[a(\theta) + b(x) + c(\theta)d(x)]$$

Then, $\sum_{i=1}^n d(X_i)$ is a complete and minimal sufficient statistic.

**Theorem 3.16.** Let $\mathbf{X} = \{X_1, \cdots, X_n\}$ be a random sample of size $n$ from a distribution in a $k$-parameter exponential family with a PDF $f(x|\theta)$ or a PMF $p(x|\theta)$, where $\theta \in \Theta \subset \mathbb{R}^k$ for an integer $k > 1$, that can be written in the form of:

$$\exp\left(a(\theta) + b(x) + \sum_{j=1}^k c_j(\theta)d_j(x)\right)$$

Then the set $\{\sum_{i=1}^n d_1(X_i), \cdots, \sum_{i=1}^n d_k(X_i)\}$ is a set of jointly complete and minimal sufficient statistics.

**Example 3.17.** Consider a random sample from Poisson$(\lambda)$, where $\lambda \in (0, \infty)$ is unknown. We have:

$$p(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!} = \exp[-\lambda - \ln(x!) + x \ln \lambda]$$

Since the support $\{0, 1, \cdots\}$ does not depend on $\lambda$, by Theorem 3.15, $\sum_{i=1}^n X_i$ is a complete and minimal sufficient statistic.

**Example 3.18.** Consider a random sample from Bern$(\theta)$, where $\theta \in (0, 1)$ is unknown. We have:

$$p(x|\theta) = \theta^x (1 - \theta)^{1-x} = \exp\left[\ln(1 - \theta) + x \ln\left(\frac{\theta}{1 - \theta}\right)\right]$$

Since the support $\{0, 1\}$ does not depend on $\theta$, by Theorem 3.15, $\sum_{i=1}^n X_i$ is a complete and minimal sufficient statistic.

**Example 3.19.** Consider a random sample from N$(\mu, \sigma^2)$, where $\mu$ and $\sigma > 0$ are unknown. We have:

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right) = \exp\left(-\frac{\mu^2}{2\sigma^2} - \frac{1}{2}\ln(2\pi\sigma^2) + \frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}x^2\right)$$

Since the support does not depend on $\mu$ and $\sigma^2$, by Theorem 3.16, $\sum_{i=1}^n X_i$ and $\sum_{i=1}^n X_i^2$ are jointly complete and minimal sufficient statistics.

## 3.6 Relationship of completeness and sufficiency with UMVUE

We still haven't explained why complete and minimal sufficient statistic can lead to UMVUE. This is due to the following theorem.

---

**Theorem 3.17. (Lehmann-Scheffé Theorem)** Let $CS$ be a complete and (minimal) sufficient statistic. If there exists a function $h(CS)$ which is unbiased for $g(\theta)$, then $h(CS)$ is the unique UMVUE of $g(\theta)$.

---

**Theorem 3.18.** Let $CS$ be a complete and (minimal) sufficient statistic. If $\mathbb{E}[f(\mathbf{X})] = g(\theta)$ for all $\theta$, then $h(CS) = \mathbb{E}[f(\mathbf{X})|T]$ is the UMVUE for $g(\theta)$.

---

**Remark 3.18.1.** From this theorem, we can formulate some strategies to find the UMVUE which is a function of $CS$:

1. Guess the correct form of the function of $CS$.

2. Solve for $h(CS)$ by $\mathbb{E}[h(CS)] = g(\theta)$.

3. Use Rao-Blackwell Theorem to construct $h(CS)$ by guessing or finding any unbiased estimator $T$ for $g(\theta)$ and then evaluating $h(CS) = \mathbb{E}(T|CS)$.

---

**Example 3.20.** Let $\{X_1, \cdots, X_n\}$ be a random sample of size $n$ from $X \sim \text{Exp}(\theta)$, where $\theta \in (0, \infty)$ is unknown. Find the UMVUE of $g(\theta) = \theta$. We try Strategy 1.
We have $\theta = \frac{1}{\mathbb{E}(X)}$. Since exponential distribution of $X$ belongs to an exponential family, we can find that $CS = \sum_{i=1}^{n} X_i$. We suspect that UMVUE is related to $\frac{n}{\sum_{i=1}^{n} X_i}$. For $n > 1$, since $\text{Exp}(\theta) = \text{Gamma}(1, \theta)$,

$$\mathbb{E}\left(\frac{1}{\sum_{i=1}^{n} X_i}\right) = \int_0^\infty \frac{\theta^n}{x\Gamma(n)} x^{n-1} e^{-\theta x} \, dx = \frac{\theta}{\Gamma(n)} \int_0^\infty \theta(\theta x)^{n-2} e^{-\theta x} \, dx = \frac{\theta\Gamma(n-1)}{\Gamma(n)} = \frac{\theta}{n-1}$$

Therefore, we have found that $\frac{n-1}{\sum_{i=1}^{n} X_i}$ is the UMVUE.

---

**Example 3.21.** We continue the example above. This time we try Strategy 2, which is solving for $h(CS)$ by $\mathbb{E}[h(CS)] = g(\theta) = \theta$.

$$\int_0^\infty h(x) \frac{\theta^n}{\Gamma(n)} x^{n-1} e^{-\theta x} \, dx = \theta$$

$$\int_0^\infty h(x) \frac{\theta^{n-1}}{\Gamma(n)} x^{n-1} e^{-\theta x} \, dx = 1$$

$$\int_0^\infty \left(h(x) \frac{x}{n-1}\right) \frac{\theta^{n-1}}{\Gamma(n-1)} x^{(n-1)-1} e^{-\theta x} \, dx = 1$$

It is only true if $h(x) \frac{x}{n-1} = 1$ for all $s > 0$. Thus, $h(x) = \frac{n-1}{x}$ and therefore:

$$h\left(\sum_{i=1}^{n} X_i\right) = \frac{n-1}{\sum_{i=1}^{n} X_i}$$

Since it is unbiased for $\theta$ and is a function of $CS$, by Lehmann-Scheffé Theorem, it is the UMVUE of $\theta$.

**Example 3.22.** Let $\{X_1, \cdots, X_n\}$ be a random sample of size $n$ from Poisson$(\lambda)$, where $\lambda \in (0, \infty)$ is unknown. Find the UMVUE of $g(\lambda) = e^{-\lambda}$. We try Strategy 3.

From Example 3.17, $\sum_{i=1}^n X_i$ is a complete and minimal sufficient statistic. Note that $g(\lambda) = e^{-\lambda} = \mathbb{P}(X_1 = 0) = \mathbf{1}_{X_1=0}$ and thus it is a trivial unbiased estimator of $g(\lambda)$. By Rao-Blackwell Theorem, $\mathbb{E}(\mathbf{1}_{X_1=0}|\sum_{i=1}^n X_i)$ is unbiased for $g(\lambda)$. By Lehmann-Scheffé Theorem, it is the unique UMVUE of $g(\lambda)$. We compute the UMVUE.
For $n = 1$,

$$\mathbb{E}\left(\mathbf{1}_{X_1=0}\,\bigg|\,\sum_{i=1}^n X_i\right) = \mathbb{E}(\mathbf{1}_{X_1=0}|X_1) = \mathbb{P}(X_1 = 0|X_1) = \mathbf{1}_{X_1=0}$$

For $n > 1$,

$$\mathbb{E}\left(\mathbf{1}_{X_1=0}\,\bigg|\,\sum_{i=1}^n X_i = s\right) = \mathbb{P}\left(X_1 = 0\,\bigg|\,\sum_{i=1}^n X_i = s\right) = \frac{\mathbb{P}(X_1 = 0, \sum_{i=1}^n X_i = s)}{\mathbb{P}(\sum_{i=1}^n X_i = s)}$$

$$= \frac{\mathbb{P}(X_1 = 0)\,\mathbb{P}(\sum_{i=2}^n X_i = s)}{\mathbb{P}(\sum_{i=1}^n X_i = s)}$$

$$= \frac{e^{-\lambda}e^{-(n-1)\lambda}[(n-1)\lambda]^s s!}{e^{-n\lambda}(n\lambda)^s s!}$$

$$= \left(\frac{n-1}{n}\right)^s$$

Therefore, the UMVUE for $g(\lambda) = e^{-\lambda}$ is:

$$\mathbb{E}\left(\mathbf{1}_{X_1=0}\,\bigg|\,\sum_{i=1}^n X_i\right) = \begin{cases} \mathbf{1}_{X_1=0}, & n = 1 \\ \left(\frac{n-1}{n}\right)^{\sum_{i=1}^n X_i}, & n > 1 \end{cases}$$

**Example 3.23.** Let $\{X_1, \cdots, X_n\}$ be a random sample of size $n$ from U$[0, \theta]$, where $\theta > 0$ is unknown. Since uniform distribution is not in an exponential family, we cannot use Theorem 3.15 to find a complete and minimal sufficient statistic.

From Example 3.16, we have found that $X_{(n)}$ is a complete and sufficient statistic. By checking for unbiasedness,

$$\mathbb{E}(X_{(n)}) = \frac{n}{n+1}\theta$$

Therefore, by Lehmann-Scheffé Theorem, the UMVUE of $\theta$ is $\frac{n+1}{n}X_{(n)}$.

## 3.7   Cramér-Rao Inequality

Recall Theorem 2.4, we claim that a sequence of MLE is asymptotically efficient, which means that:

$$\mathcal{I}_X^{-1}(\theta)$$

is the lowest possible bound for any unbiased estimator. The reason is because of the Cramér-Rao Inequality (C-R Inequality).

**Theorem 3.19.** (**Cramér-Rao Inequality**) Under the regularity conditions, the variance of an unbiased estimator $T(\mathbf{X}) = T(X_1, \cdots, X_n)$ for $\theta$, based on a set of random variables $\mathbf{X} = \{X_1, \cdots, X_n\}$ from their joint PDF $f_{X_1,\cdots,X_n}(x_1, \cdots, x_n|\theta)$ satisfies the following inequality:

$$\text{Var}(T(\mathbf{X})) \geq \frac{1}{\mathcal{I}_{X_1,\cdots,X_n}(\theta)} = \frac{1}{\mathbb{E}\left[\frac{d}{d\theta}\ln f_{X_1,\cdots,X_n}(X_1, \cdots, X_n|\theta)\right]^2}$$

The lower bound is called **Cramér-Rao lower bound** (CRLB).

**Remark 3.19.1.** The Cramér-Rao Inequality can be written as:

$$\text{Var}(T(\mathbf{X})) \geq \frac{1}{-\mathbb{E}\left[\frac{d^2}{d\theta^2}\ln f_{X_1,\cdots,X_n}(X_1, \cdots, X_n|\theta)\right]}$$

**Remark 3.19.2.** If $\mathbf{X}$ is a random sample of size $n$, then we have:

$$\text{Var}(T(\mathbf{X})) \geq \frac{1}{n\mathcal{I}_{X_1}(\theta)} = \frac{1}{n\,\mathbb{E}\left[\frac{\mathrm{d}}{\mathrm{d}\theta}\ln f_{X_1}(X_1|\theta)\right]^2} = \frac{1}{-n\,\mathbb{E}\left[\frac{\mathrm{d}^2}{\mathrm{d}\theta^2}\ln f_{X_1}(X_1|\theta)\right]}$$

**Example 3.24.** Let $\mathbf{X} = \{X_1, \cdots, X_n\}$ be a random sample of size $n$ from $\text{N}(\theta, \sigma^2)$., where $\sigma^2$ is known and $\theta$ is unknown. The CRLB for $\theta$ is:

$$\frac{1}{n\mathcal{I}_{X_1}(\theta)} = \frac{\sigma^2}{n}$$

**Example 3.25.** Let $\mathbf{X} = \{X_1, \cdots, X_n\}$ be a random sample of size $n$ from $\text{Bern}(p)$, where $p$ is unknown. The CRLB for $p$ is:

$$\frac{1}{n\mathcal{I}_{X_1}(p)} = \frac{p(1-p)}{n}$$

Often time, we want to estimate a function of $\theta$, $g(\theta)$, instead of $\theta$.

**Theorem 3.20.** Under the regularity conditions, if $T(\mathbf{X}) = T(X_1, \cdots, X_n)$ is an unbiased estimator for $g(\theta)$, then the Cramér-Rao Inequality for $g(\theta)$ is:

$$\text{Var}(T(\mathbf{X})) \geq \frac{\left[\frac{\mathrm{d}}{\mathrm{d}\theta}g(\theta)\right]^2}{\mathcal{I}_{X_1,\cdots,X_n}(\theta)} = \frac{\left[\frac{\mathrm{d}}{\mathrm{d}\theta}g(\theta)\right]^2}{\mathbb{E}\left[\frac{\mathrm{d}}{\mathrm{d}\theta}\ln f_{X_1,\cdots,X_n}(X_1,\cdots,X_n|\theta)\right]^2}$$

*Proof.*
Let $U = \frac{\mathrm{d}}{\mathrm{d}\theta}\ln f_{\mathbf{X}}(\mathbf{X}|\theta)$ and $V = T(\mathbf{X})$. We have:

$$-1 \leq \frac{\text{cov}(U,V)}{\sqrt{\text{Var}(U)\,\text{Var}(V)}} \leq 1 \implies \frac{(\text{cov}(U,V))^2}{\text{Var}(U)} \leq \text{Var}(V)$$

We may find that $\text{Var}(U) = \text{Var}\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\ln f_{\mathbf{X}}(\mathbf{X}|\theta)\right) = \mathcal{I}_{\mathbf{X}}(\theta)$. In addition,

$$\begin{aligned}
\text{cov}\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\ln f_{\mathbf{X}}(\mathbf{X}|\theta), T(\mathbf{X})\right) &= \mathbb{E}\left[T(\mathbf{X})\frac{\mathrm{d}}{\mathrm{d}\theta}\ln f_{\mathbf{X}}(\mathbf{X}|\theta)\right] - \mathbb{E}\left[\frac{\mathrm{d}}{\mathrm{d}\theta}\ln f_{\mathbf{X}}(\mathbf{X}|\theta)\right]\mathbb{E}[T(\mathbf{X})] \\
&= \mathbb{E}\left[T(\mathbf{X})\frac{\mathrm{d}}{\mathrm{d}\theta}\ln f_{\mathbf{X}}(\mathbf{X}|\theta)\right] \\
&= \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} T(\mathbf{x})\left(\frac{\mathrm{d}}{\mathrm{d}\theta}\ln f_{\mathbf{X}}(\mathbf{x}|\theta)\right) f_{\mathbf{X}}(\mathbf{x}|\theta)\,d\mathbf{x} \\
&= \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} T(\mathbf{x})\frac{\mathrm{d}}{\mathrm{d}\theta}f_{\mathbf{X}}(\mathbf{x}|\theta)\,d\mathbf{x} \\
&= \frac{\mathrm{d}}{\mathrm{d}\theta}\mathbb{E}[T(\mathbf{X})] \\
&= \frac{\mathrm{d}}{\mathrm{d}\theta}g(\theta)
\end{aligned}$$

Therefore, we have that:

$$\text{Var}(T(\mathbf{X})) \geq \frac{\left[\frac{\mathrm{d}}{\mathrm{d}\theta}g(\theta)\right]^2}{\mathcal{I}_{\mathbf{X}}(\theta)}$$

$\square$

**Remark 3.20.1.** Since CRLB is the lowest bound of variance for any unbiased estimator, any unbiased estimator whose variance can achieve the CRLB for $g(\theta)$ is the UMVUE for $g(\theta)$.

**Remark 3.20.2.** It is not necessary that a UMVUE has a variance equal to the CRLB.

**Example 3.26.** Let $\{X_1, \cdots, X_n\}$ be a random sample of size $n$ from $\text{Exp}(\theta)$, where $\theta \in (0, \infty)$ is unknown. The CRLB of $\theta$ is $\frac{\theta^2}{n}$. From Example 3.20, we have found that $\frac{n-1}{\sum_{i=1}^{n} X_i}$ is the UMVUE of $\theta$ when $n > 1$.

After some tedious calculation, for $n > 2$, we would find that $\text{Var}\left(\frac{n-1}{\sum_{i=1}^{n} X_i}\right) = \frac{\theta^2}{n-2} \geq \frac{\theta^2}{n}$.

When would the equality for C-R inequality holds?

---

**Theorem 3.21.** Under the regularity conditions, the C-R equality holds if and only if:

$$\frac{\mathrm{d}}{\mathrm{d}\theta} \ln f_{X_1,\cdots,X_n}(X_1,\cdots,X_n|\theta) = A(\theta,n)[T(X_1,\cdots,X_n) - g(\theta)]$$

where $A(\theta,n)$ is a non-zero function. The statistic $T(X_1,\cdots,X_n)$ is an UMVUE of $g(\theta)$.

---

**Lemma 3.22.** If $T(X_1,\cdots,X_n)$ is an UMVUE of $g(\theta)$ such that C-R equality holds, then $aT(X_1,\cdots,X_n) + b$ is an UMVUE of $ag(\theta) + b$, where $a \neq 0$.

---

*Proof.*

$$\frac{\mathrm{d}}{\mathrm{d}\theta} \ln f_{X_1,\cdots,X_n}(X_1,\cdots,X_n|\theta) = A(\theta,n)[T(X_1,\cdots,X_n) - g(\theta)] = \frac{A(\theta,n)}{a}[(aT(X_1,\cdots,X_n) + b) - (ag(\theta) + b)]$$

By having $B(\theta,n) = \frac{A(\theta,n)}{a}$, we can find that $aT(X_1,\cdots,X_n) + b$ is an UMVUE of $ag(\theta) + b$.   □

This theorem has an interesting result that if we can write $\frac{\mathrm{d}}{\mathrm{d}\theta} \ln f_{X_1,\cdots,X_n}(X_1,\cdots,X_n|\theta)$ into $A(\theta,n)[T(X_1,\cdots,X_n) - g(\theta)]$, then the statistic must be an UMVUE.

---

**Example 3.27.** Let $\{X_1,\cdots,X_n\}$ be a random sample of size $n$ from Poisson($\lambda$), where $\lambda$ is unknown.

$$\frac{\mathrm{d}}{\mathrm{d}\lambda} \ln p_{X_1,\cdots,X_n}(X_1,\cdots,X_n|\lambda) = \frac{\mathrm{d}}{\mathrm{d}\lambda} \ln \left( \prod_{i=1}^{n} \frac{\lambda^{X_i} e^{-\lambda}}{X_i!} \right)$$

$$= \frac{\mathrm{d}}{\mathrm{d}\lambda} \sum_{i=1}^{n} [X_i \ln \lambda - \lambda - \ln(X_i!)]$$

$$= \sum_{i=1}^{n} \frac{X_i}{\lambda} - 1$$

$$= \frac{n}{\lambda}(\overline{X} - \lambda)$$

Therefore, by Theorem 3.21, $\overline{X}$ is an UMVUE of $\lambda$ and:

$$\mathrm{Var}(\overline{X}) = \frac{1}{n\mathcal{I}_{X_1}(\lambda)} = \frac{\lambda}{n}$$

---

**Remark 3.22.1.** For any particular function of $\theta$ other than Euclidean transformation, Theorem 3.21 is not useful.

---

**Example 3.28.** Let $\{X_1,\cdots,X_n\}$ be a random sample of size $n$ from Exp($\theta$), where $\theta$ is unknown.

$$\frac{\mathrm{d}}{\mathrm{d}\theta} \ln f_{X_1,\cdots,X_n}(X_1,\cdots,X_n|\theta) = \frac{\mathrm{d}}{\mathrm{d}\theta} \ln \left( \prod_{i=1}^{n} \theta e^{-\theta X_i} \right)$$

$$= \frac{\mathrm{d}}{\mathrm{d}\theta} \sum_{i=1}^{n} (\ln \theta - \theta X_i)$$

$$= \sum_{i=1}^{n} \left( \frac{1}{\theta} - X_i \right)$$

$$= -n \left( \overline{X} - \frac{1}{\theta} \right)$$

Therefore, by Theorem 3.21, $\overline{X}$ is an UMVUE of $\frac{1}{\theta}$ and:

$$\mathrm{Var}(\overline{X}) = \frac{1}{n\mathcal{I}_{X_1}(\theta)} = \frac{\theta^2}{n}$$

However, since $\theta$ cannot be written as Euclidean transformation of $\frac{1}{\theta}$, we cannot use the Theorem to find the UMVUE of $\theta$.

**Remark 3.22.2.** Theorem 3.21 also can only be used under regularity conditions. For instance, for $U[0, \theta]$, we cannot use this Theorem.

# Chapter 4

# Hypothesis Testing

This chapter will be primarily focus on comparing different unbiased point estimators.

In engineering and science field, people usually hypothesize something about a system. Before they prove the conjecture using experimental data, they need to define a hypothesis.

> **Definition 4.1. Statistical hypothesis** is an assertion or conjecture of the random variable of our interest. If a parametric distribution is considered, then a statistical hypothesis can be a conjecture about the true value of the unknown parameters of the parametric distribution.

> **Example 4.1.** An engineer decides on the basis of sample data whether the true average lifetime of a certain kind of tire is at least $22,000$ miles.
> The engineer has to test the hypothesis that $\theta$ in $\text{Exp}(\theta)$ is at least $22,000$.

> **Example 4.2.** An agronomist want to decide on the basis of experiments whether one kind of fertilizer produces a higher yield of soybeans than another.
> The agronomist has to test the hypothesis that $\mu_1 > \mu_2$ from two distributions $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$.

> **Example 4.3.** A manufacturer of pharmaceutical products decides on the basis of samples whether $90\%$ of all patients given a new medication will recover from a certain disease.
> The manufacturer has to test the hypothesis that $\theta$ in $\text{Bin}(n, \theta)$ equals 0.90.

## 4.1 Null and Alternative Hypotheses

The hypothesis of our interest is related to a particular class of $\theta$, say $\Theta_0$, and its complement $\Theta_1 = \Theta \setminus \Theta_0$. This two classes are subsets of the parameter space $\Theta$ of $\theta$.

> **Definition 4.2.** The hypothesis with $\theta \in \Theta_0$ is the **null hypothesis** $H_0$, and the hypothesis with $\theta \in \Theta_1$ is the **alternative hypothesis**.

> **Remark 4.2.1.** We usually use the signs with implied equals in $H_0$.

> **Definition 4.3.** The hypothesis is **simple** if the parametric distribution would be fully specified under the hypothesis. Otherwise, the hypothesis is **composite**.

> **Example 4.4.** Using the Example 4.1, we have that:
>
> $$\begin{cases} H_0: & \theta \geq 22,000 \\ H_1: & \theta < 22,000 \end{cases}$$
>
> Both $H_0$ and $H_1$ are composite because it does not specify the parameter.

**Example 4.5.** Using the Example 4.3, we have that:

$$\begin{cases} H_0 : & \theta = 0.9 \\ H_1 : & \theta \neq 0.9 \end{cases}$$

$H_0$ is simple while $H_1$ is composite.

In hypothesis testing, we want to see whether or not we can find an evidence to say that $H_0$ is false.

**Remark 4.3.1.** Hypothesis testing usually follows these three steps:

1. Determine $H_0$ and $H_1$.

2. Under $H_0$, define a rare event, event that happens with a very small probability in one experiment with $n$ data.

3. Collect data.

   (a) If data causes the rare event to happen, it contradicts $H_0$. This mean we can say that $H_0$ is false and reject $H_0$.

   (b) If data does not cause the rare event to happen, it does not contradicts $H_0$. This mean we cannot say that $H_0$ is false and we do not reject $H_0$.

**Remark 4.3.2.** Not rejecting $H_0$ does not mean we accept $H_0$. It just mean there is no sufficient evidence to reject $H_0$. The whole idea is to try gathering enough evidence to have great confidence that $H_0$ is false and $H_1$ is true.

**Example 4.6.** We want to know whether or not a coin is fair. Consider a random experiment of flipping the coin 10 times. We may determine that:

$$\begin{cases} H_0 : & \mathbb{P}(\{H\}) = \mathbb{P}(\{T\}) = 0.5 \\ H_1 : & \mathbb{P}(\{H\}) \neq \mathbb{P}(\{T\}) \end{cases}$$

Under $H_0$, we define that the event of getting 10 tails to be the rare event under $H_0$ since the probability of getting 10 tails in one experiment is $0.5^{10} \approx 0.00098$.
We can them preform the experiment to collect data by flipping the coin 10 times. If we get 10 tails, then the collected data tells us that getting 10 tails is not a rare event, which contradicts $H_0$. Therefore, we have evidence to suspect the reliability of $H_0$ and thus reject $H_0$ and accept $H_1$.

## 4.2   Test Errors and Error Probabilities

After we decide the null and alternative hypotheses, we need to determine a test statistic, i.e. the point estimator, to construct a test of rejecting or not rejecting the null hypothesis.

**Definition 4.4. Non-rejection region** $C_0$ is a subset of $\Theta$ such that we do not reject $H_0$, i.e.

$$C_0 = \{\mathbf{x} : \text{Not reject } H_0\}$$

**Rejection region** $C_1$ is a subset of $\Theta$ such that we reject $H_0$, i.e.

$$C_1 = \{\mathbf{x} : \text{Reject } H_0\}$$

However, there is no prefect test statement due to randomness of sample of data. Each test would lead to the following two kinds of errors.

**Definition 4.5. Type I Error** is the error of rejecting $H_0$ when it is true. **Type II Error** is the error of not rejecting $H_0$ when it is false.

|                  | Not reject $H_0$ | Reject $H_0$  |
| ---------------- | ---------------- | ------------- |
| If $H_0$ is true | No error         | Type I Error  |
| If $H_0$ is false| Type II Error    | No error      |

We may define their corresponding probabilities.

**Definition 4.6. Type I error probability**, denoted by $\gamma(\theta)$, is the probability of rejecting $H_0$ for $\theta \in \Theta_0$.

$$\gamma(\theta) = \mathbb{P}(\text{Reject } H_0 | \theta) = \mathbb{P}(\mathbf{X} \in C_1 | \theta)$$

**Type II error probability**, denoted by $\beta(\theta)$, is the probability of not rejecting $H_0$ for $\theta \in \Theta_1$.

$$\beta(\theta) = \mathbb{P}(\text{Not reject } H_0 | \theta) = \mathbb{P}(\mathbf{X} \in C_0 | \theta)$$

**Remark 4.6.1.** Since we cannot control $\gamma(\theta)$ and $\beta(\theta)$ at the same time, conventionally, we assign an upper bound to $\gamma(\theta)$ over $\Theta_0$ and find a test with $\beta(\theta)$ as small as possible. If we are dealing with continuous case,

$$\sup_{\theta \in \Theta_0} \gamma(\theta) = \alpha$$

If we are dealing with discrete case,

$$\sup_{\theta \in \Theta_0} \gamma(\theta) \leq \alpha$$