Emory University

# MATH 362 Mathematical Statistics II

Learning Notes

Jiuru Lyu

March 21, 2024

## Contents

# 1 Estimation

## 1.1 Introduction

**Definition 1.1.1 (Model).** A *model* is a distribution with certain parameters.

**Example 1.1.2** The normal distribution: $N(\mu, \sigma^2)$.

**Definition 1.1.3 (Population).** The *population* is all the objects in the experiment.
**Definition 1.1.4 (Data, Sample, and Random Sample).** *Data* refers to observed value from sample. The *sample* is a subset of the population. A *random sample* is a sequence of independent, identical ($i.i.d.$) random variables.
**Definition 1.1.5 (Statistics).** *Statistics* refers to a function of the random sample.

**Example 1.1.6** The sample mean is a function of the sample:

$$\overline{Y} = \frac{1}{n}(Y_1 + \cdots + Y_n).$$

**Example 1.1.7** Central Limit Theorem

We randomly toss $n = 200$ fair coins on the table. Calculate, using the central limit theorem, the probability that at least $110$ coins have turned on the same side.

$$\overline{X} = \frac{X_1 + \cdots + X_{200}}{200} \quad \overset{\text{CLT}}{\sim} \quad N\left(\mu, \sigma^2\right),$$

where

$$\mu = \mathbf{E}\left(\overline{X}\right) = \frac{\sum_{i=1}^{200} \mathbf{E}(X_i)}{200},$$

$$\sigma^2 = \mathbf{Var}\left(\overline{X}\right) = \mathbf{Var}\left(\frac{X_1 + \cdots + X_{200}}{200}\right) = \frac{\sum_{i=1}^{200} \mathbf{Var}(X_i)}{200^2}.$$

**Definition 1.1.8 (Statistical Inference).** The process of *statistical inference* is defined to be the process of using data from a sample to gain information about the population.

**Example 1.1.9** Goals in statistical inference

1. **Definition 1.1.10 (Estimation).** To obtain values of the parameters from the data.
2. **Definition 1.1.11 (Hypothesis Testing).** To test a conjecture about the parameters.
3. **Definition 1.1.12 (Goodness of Fit).** How well does the data fit a given distribution.
4. Linear Regression

## 1.2 The Method of Maximum Likelihood and the Method of Moments

**Example 1.2.1** Given an unfair coin, or $p$-coin, such that

$$X = \begin{cases} 1 & \text{head with probability } p, \\ 0 & \text{tail with probability } 1-p. \end{cases}$$

How can we determine the value $p$?

***Solution 1.***

1. Try to flip the coin several times, say, three times. Suppose we get HHT.
2. Draw a conclusion from the experiment.

**Key idea: The choice of the parameter $p$ should be the value that maximizes the probability of the sample.**

$$\mathbf{P}(X_1 = 1, X_2 = 1, X_3 = 0) = \mathbf{P}(X_1 = 1)\mathbf{P}(X_2 = 1)\mathbf{P}(X_3 = 0) = p^2(1-p) \coloneqq f(p).$$

Solving the optimization problem $\max_{p>0} f(p)$, we find it is most likely that $p = \frac{2}{3}$. This method is called the *likelihood maximization method.* $\square$

**Definition 1.2.2 (Likelihood Function).** For a random sample of size $n$ from the discrete (or continuous) pdf $p_X(k;\theta)$ (or $f_Y(y;\theta)$), the *likelihood function*, $\mathbf{L}(\theta)$, is the product of the pdf evaluated at $X_i = k_i$ (or $Y_i = y_i$). That is,

$$\mathbf{L}(\theta) \coloneqq \prod_{i=1}^{n} p_X(k_i;\theta) \quad \text{or} \quad \mathbf{L}(\theta) \coloneqq \prod_{i=1}^{n} f_Y(y_i;\theta).$$

**Definition 1.2.3 (Maximum Likelihood Estimate).** Let $\mathbf{L}(\theta)$ be as defined in Definition 1.2.2. If $\theta_e$ is a value of the parameter such that $\mathbf{L}(\theta_e) \geq \mathbf{L}(\theta)$ for all possible values of $\theta$, then we call $\theta_e$ the *maximum likelihood estimate* for $\theta$.

**Theorem 1.2.4 The Method of Maximum Likelihood**
Given random samples $X_1, \ldots, X_N$ and a density function $p_X(x)$ (or $f_X(x)$), then we have the likelihood function defined as

$$\begin{aligned}\mathbf{L}(\theta) = p_X(X;\theta) &= \mathbf{P}(X_1, X_2, \ldots, X_N) \\ &= \mathbf{P}(X_1)\mathbf{P}(X_2)\cdots\mathbf{P}(X_N) && [independent] \\ &= \prod_{i=1}^{N} p_X(X_i;\theta) && [identical]\end{aligned}$$

Then, the maximum likelihood estimate for $\theta$ is given by

$$\theta^* = \arg\max_\theta L(\theta),$$

where

$$\mathbf{L}\Big(\arg\max_\theta L(\theta)\Big) = \mathbf{L}^*(\theta) = \max_\theta \mathbf{L}(\theta).$$

**Example 1.2.5** Consider the Poisson distribution $X = 0, 1, \ldots,$ with $\lambda > 0$. Then, the pdf is given by

$$p_X(k,\lambda) = e^{-\lambda}\frac{\lambda^k}{k!}, \quad k = 0, 1, \ldots$$

Given data $k_1, \ldots, k_n$, we have the likelihood function

$$\mathbf{L}(\lambda) = \prod_{i=1}^{n} p_X(X = k;\lambda) = \prod_{i=1}^{n} e^{-\lambda}\frac{\lambda^{k_i}}{k_i!} = e^{-n\lambda}\frac{\lambda^{\sum k_i}}{k_1!\cdots k_n!}$$

Then, to find the maximum likelihood estimate of $\lambda$, we need to $\max_\lambda \mathbf{L}(\lambda)$. That is to solve $\frac{\partial \mathbf{L}(\lambda)}{\partial \lambda} = 0$ and $\frac{\partial^2 \mathbf{L}(\lambda)}{\partial \lambda^2} < 0$.

**Example 1.2.6** Waiting Time.
Consider the exponential distribution $f_Y(y) = \lambda e^{-\lambda y}$ for $y \geq 0$. Find the MLE $\lambda_e$ of $\lambda$.
***Solution 2.***

The likelihood function of the exponential distribution is given by

$$\mathbf{L}(\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda y_i} = \lambda^n \exp\left(-\lambda \sum_{i=1}^{n} y_i\right).$$

Now, define

$$\ell(\lambda) = \ln \mathbf{L}(\lambda) = n \ln \lambda - \lambda \sum_{i=1}^{n} y_i.$$

To optimize $\ell(\lambda)$, we compute

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}\ell(\lambda) = \frac{n}{\lambda} - \sum_{i=1}^{n} y_i \stackrel{set}{=} 0$$

So,

$$\frac{n}{\lambda} = \sum_{i=1}^{n} y_i \implies \lambda_e = \frac{n}{\sum_{i=1}^{n} y_i} =: \frac{1}{\overline{y}},$$

where $\overline{y}$ is the sample mean. $\square$

**Example 1.2.7** Given the exponential distribution $f_Y(y) = \lambda e^{-\lambda y}$ for $y \geq 0$. Find the MLE of $\lambda^2$.

***Solution 3.***

Define $\tau = \lambda^2$. Then, $\lambda = \sqrt{\tau}$, and so

$$f_Y(y) = \sqrt{\tau} e^{-\sqrt{\tau} y}, \quad y \geq 0.$$

Then, the likelihood function becomes

$$\mathbf{L}(\tau) = \prod_{i=1}^{n} f_Y(y) = \tau^{\frac{n}{2}} \exp\left(-\sqrt{\tau} \sum_{i=1}^{n} y_i\right).$$

Similarly, after maximization, we find

$$\tau_e = \frac{1}{(\overline{y})^2}.$$

$\square$

**Theorem 1.2.8 Invariant Property for MLE**
Suppose $\lambda_e$ is the MLE of $\lambda$. Define $\tau := h(\lambda)$. Then, $\tau_e = h(\lambda_e)$.

***Proof 4.*** In this proof, we will prove the case when $h$ is a one-to-one function. The case of $h$ being a many-to-one function is beyond the scope of this course.

Suppose $h(\cdot)$ is a one-to-one function. Then, $\lambda = h^{-1}(\tau)$ is well-defined. Then,

$$\max_{\lambda} \mathbf{L}(\lambda; y_1, \ldots, y_n) = \max_{\tau} \mathbf{L}\big(h^{-1}(\tau); y_1, \ldots, y_n\big) = \max_{\tau} \mathbf{L}(\tau; y_1, \ldots, y_n).$$

■

**Example 1.2.9** Waiting Time with an unknown Threshold.

Let $\lambda = 1$ in exponential but there is an unknown threshold $\theta$, that, is $f_Y(y) = e^{-(y-\theta)}$ for $y \geq \theta,\ \theta > 0$.

***Solution 5.***

Note that the likelihood function is given by

$$\begin{aligned}\mathbf{L}(\theta; y_1, \ldots, y_n) = \prod_{i=1}^{n} f_Y(y_1) &= \exp\left(-\sum_{i=1}^{n}(y_i - \theta)\right), \quad y_i \geq \theta,\ \theta > 0 \\ &= \exp\left(-\sum_{i=1}^{n}(y_i - \theta)\right) \cdot \mathbb{1}_{[y_i \geq 0,\ \theta > 0]},\end{aligned}$$

where

$$\mathbb{1}_{x \in A} = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A. \end{cases}$$

Using order statistics,

$$\begin{aligned}\mathbf{L}(\theta) &= \exp\left(-\sum_{i=1}^{n}(y_i - \theta)\right) \cdot \mathbb{1}_{\left[y_{(n)} \geq y_{(n-1)} \geq \cdots \geq y_{(1)} \geq \theta,\ \theta > 0\right]} \\ &= \exp\left(-\sum_{i=1}^{n} y_i + n\theta\right) \mathbb{1}_{\left[y_{(n)} \geq \cdots \geq y_{(1)} \geq \theta,\ \theta > 0\right]}.\end{aligned}$$

So, we know $\theta \leq y_{(1)} = y_{\min}$.

To maximize the likelihood function, we want to maximize $-\sum y_i + n\theta$. That is, to maximize $\theta$, as $\theta \leq y_{\min}$, it must be that $\theta_{\max} = y_{\min}$. Therefore, the MLE is $\theta^* = y_{\min}$. □

**Example 1.2.10** Suppose $Y_1, \dots, Y_n \sim$ Uniform$[0, a]$. That is, $f_Y(y; a) = \dfrac{1}{a}$ for $y \in [0, a]$. Find MLE $a_e$ of $a$.

***Solution 6.***

Note that

$$\begin{aligned} f_Y(y; a) &= \frac{1}{a} \cdot \mathbb{1}_{\{y \in [0,a]\}} \\ &= \frac{1}{a} \cdot \mathbb{1}_{\left\{0 \le y_{(1)} \le \cdots \le y_{(n)} \le a\right\}} \qquad \text{where } y_{(1)} = \min y_i \text{ and } y_{(n)} = \max y_i \end{aligned}$$

Then,

$$\mathbf{L}(a) = \frac{1}{a^n} \mathbb{1}_{\left\{0 \le y_{(1)} \le \cdots \le y_{(n)} \le a\right\}}$$

To maximize $\mathbf{L}(a)$, we want to minimize $a^n$. Since $a \ge y_{(n)}$, it must be that $a_e = y_{(n)}$. Here, we call $a_e = y_{(n)}$ an *estimate*, and $\widehat{a_{\text{MLE}}} = Y_{(n)}$ an *estimator*. □

**Example 1.2.11 MLE that Does Not Esist**

Suppose $f_Y(y; a) = \dfrac{1}{a}, \quad y \in [0, a)$. Find the MLE.

***Solution 7.***

The likelihood function is the same:

$$\mathbf{L}(a) = \frac{1}{a^n} \mathbb{1}_{\left\{0 \le y_{(1)} \le \cdots \le y_{(n)} < a\right\}}.$$

However, since $[0, a)$ is not a closed set, the optimization problem $\max\limits_{a \in [0,a)} \mathbf{L}(a)$ does not have a solution. Hence, the estimate does not exist. □

**Remark 1.1** *MLE may not be unique all the time.*

**Example 1.2.12 Multiple MLE Values**

Suppose $X_1, \dots, X_n \sim$ Uniform$\left[a - \frac{1}{2}, a + \frac{1}{2}\right]$, where $f_X(x; a) = 1,\ x \in \left[a - \frac{1}{2}, a + \frac{1}{2}\right]$. Find the MLE.

***Solution 8.***

In the indicator function notation, we can rewrite the pdf to be

$$f_X(x; a) = \mathbb{1}_{\left\{a - \frac{1}{2} \le x \le a + \frac{1}{2}\right\}} = \mathbb{1}_{\left\{a - \frac{1}{2} \le x_{(1)} \le \cdots \le x_{(n)} \le a + \frac{1}{2}\right\}}.$$

So, the likelihood function will be

$$\mathbf{L}(a) = \prod_{i=1}^{n} f_x(x_i; a) = \begin{cases} 1, & a \in \left[x_{(n)} - \frac{1}{2}, x_{(1)} + \frac{1}{2}\right] \\ 0, & \text{otherwise}. \end{cases}$$

So, the $\mathbf{L}(a)$ will be maximized whenever $a \in \left[x_{(n)} - \frac{1}{2}, x_{(1)} + \frac{1}{2}\right]$. Therefore, MLE can be any value in the range $\left[x_{(n)} - \frac{1}{2}, x_{(1)} + \frac{1}{2}\right]$. Say,

$$a_e = x_{(n)} - \frac{1}{2} \quad \text{or} \quad a_e = x_{(1)} - \frac{1}{2} \quad \text{or} \quad a_e = \frac{x_{(n)} - \frac{1}{2} + x_{(1)} + \frac{1}{2}}{2} = \frac{x_{(n)} + x_{(1)}}{2}.$$

□

**Theorem 1.2.13 MLE for Multiple Parameters**
In general, we have the likelihood function $\mathbf{L}(\theta)$, where $\theta = (\theta_1, \ldots, \theta_p)$. To find the MLE, we need

$$\frac{\partial \mathbf{L}(\theta)}{\partial \theta_i} = 0 \quad i = 1, \ldots, p,$$

and the Hessian matrix

$$\left(\frac{\partial^2 \mathbf{L}(\theta)}{\partial \theta_i \partial \theta_j}\right)_{i,j=1,\ldots,p} := \begin{pmatrix} \frac{\partial^2 \mathbf{L}(\theta)}{\partial \theta_1^2} & \cdots & \frac{\partial^2 \mathbf{L}(\theta)}{\partial \theta_1 \partial \theta_p} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 \mathbf{L}(\theta)}{\partial \theta_p \partial \theta_1} & \cdots & \frac{\partial^2 \mathbf{L}(\theta)}{\partial \theta_p^2} \end{pmatrix}$$

should be negative dfinite.

**Example 1.2.14 MLE for Multiple Parameters: Normal Distribution**
Suppose $Y_1, \ldots, Y_n \sim N(\mu, \sigma)$. Then,

$$f_{Y_i}(u; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(y_i-\mu)^2/\left(2\sigma^2\right)}.$$

Find the MLE for $\mu$ and $\sigma$.

***Solution 9.***

The likelihood function will be

$$\mathbf{L}(\mu,\sigma)=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}e^{-(y_i-\mu)^2/\left(2\sigma^2\right)}.$$

Then, we define

$$\ell(\mu,\sigma)=\ln\mathbf{L}(\mu,\sigma)=-\frac{n}{2}\ln 2\pi-\frac{n}{2}\ln\sigma^2-\frac{1}{2}\left(\sigma^2\right)^-1\sum_{i=1}^{n}(y_i-\mu)^2.$$

Set

$$\begin{cases}\dfrac{\partial\ell(\mu,\sigma)}{\partial\mu}=0 & \text{①}\\ \dfrac{\partial\ell(\mu,\sigma)}{\partial\sigma}=0 & \text{②}\end{cases}$$

From ①, we have

$$\begin{aligned}\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_1-\mu)&=0\\ \sum_{i=1}^{n}y_i=n\mu\implies&\boxed{\mu_e=\frac{\sum y_i}{n}=\overline{y}}\end{aligned}$$

From ②, by the invariant property of MLE, we instead set

$$\begin{aligned}\frac{\partial\ell(\mu,\sigma)}{\partial\sigma^2}&=0\\ -\frac{n}{2}\cdot\frac{1}{\sigma^2}+\frac{1}{2}\left(\frac{1}{\sigma^2}\right)^2\sum_{i=1}^{n}(y_i-\mu)^2&=0\\ \frac{1}{2\sigma^2}\left(-n+\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\right)&=0\\ -n\sigma^2+\sum_{i=1}^{n}(y_i-\mu)^2&=0 \qquad (\mu_e=\overline{y})\\ \sum_{i=1}^{n}(y_i-\overline{y})^2&=n\sigma^2\\ \sigma_e^2=\frac{1}{n}\sum_{i=1}^{n}(y_i-\overline{y})^2&\implies\boxed{\sigma_e=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\overline{y})^2}}\end{aligned}$$

□

## 1.3 The Method of Moment

**Definition 1.3.1 (Moment Generating Function).** The *Moment Generating Function (MGF)* is defined as

$$\mathbf{M}_X(t) = \mathbf{E}\big[e^{tX}\big],$$

and it uniquely determines a probability distribution.

**Definition 1.3.2 (Moment).** The $k$*-th order moment* of $X$ is $\mathbf{E}\big[X^k\big]$.

**Example 1.3.3**Meaning of Different Moments

- $\mathbf{E}[X]$: location of a distribution
- $\mathbf{E}[X^2] = \mathbf{Var}(X) - \mathbf{E}[X]^2$: width of a distribution
- $\mathbf{E}[X^3]$: skewness – positively skewed / negatively skewed
- $\mathbf{E}[X^4]$: kurtosis / tailedness – speed decaying to $0$.

**Example 1.3.4 Moment Estimate: Moments of Population and Sample**

| **Population** | **Sample,** $X_1, \ldots, X_n$ |
|---|---|
| $\mathbf{E}[X] = \mu$ | $\widehat{\mu} = \overline{X} = \dfrac{X_1 + \cdots + X_n}{n}$ |
| $\mathbf{E}[X^2] = \mu^2 + \sigma^2$ | $\widehat{\mu}^2 + \widehat{\sigma}^2 = \dfrac{X_1^2 + \cdots + X_n^2}{n}$ |
| $\vdots$ | $\vdots$ |
| $\mathbf{E}\big[X^k\big]$ | $\dfrac{X_1^k + \cdots + X_n^k}{n}$ |

**Rationale**: The population moments should be close to the sample moments.

**Example 1.3.5**

- Consider $N(\mu, \sigma^2)$, where $\sigma$ is given. Estimate $\mu$.

  By the method of moment estimate, we have $\mu_e = \overline{X}$.

- Consider $N(\mu, \sigma^2)$. Estimate $\mu$ and $\sigma$.

  We have $\mu_e = \overline{X}$ and $\mu_e^2 + \sigma_e^2 = \dfrac{X_1^2 + \cdots + X_n^2}{n}$.

- Consider $N(\theta, \sigma^2)$. Given $\mathbf{E}(X^4) = 3\sigma^4$, estimate $\mu$ and $\sigma$.

  We have $\mu_e = \overline{X}$, $\mu_e^2 + \sigma_e^2 = \dfrac{X_1^2 + \cdots + X_n^2}{n}$, and $3\sigma^4 = \dfrac{X_1^4 + \cdots + X_n^4}{n}$. We have three equations but only two unknowns, then a solution is not guaranteed. So, we need some restrictions on this method (see Remark 1.2).

**Theorem 1.3.6 Method of Moments Estimates**
For a random sample of size $n$ from the discrete (or continuous) population/pdf $p_X(k; \theta_1, \ldots, \theta_s)$ (or $f_Y(y; \theta_1, \ldots, \theta_s)$), solutions to the system

$$\begin{cases} \mathbf{E}(Y) = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} y_i \\ \qquad \vdots \\ \mathbf{E}(Y^s) = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} y_i^s \end{cases}$$

which are denoted by $\theta_{1e}, \ldots, \theta_{se}$, are called the **method of moments estimates** of $\theta_1, \ldots, \theta_s$.

**Remark 1.2** *To estimate $k$ parameters with the method of moments estimates, we will only match the first $k$ orders of moments.*

**Example 1.3.7** Consider the Gamma distribution:

$$f_Y(y; r, \lambda) = \frac{\lambda^r}{\Gamma(r)} y^{r-1} e^{-\lambda y} \quad \text{for } y \geq 0.$$

Given $\mathbf{E}(Y) = \dfrac{r}{\lambda}$ and $\mathbf{E}(Y^2) = \dfrac{r}{\lambda^2} + \dfrac{r^2}{\lambda^2}$. Estimate $r$ and $\lambda$.

***Solution 1.***

$$\mathbf{E}(Y) = \frac{r}{\lambda} \implies \frac{r_e}{\lambda_e} = \frac{y_1 + \cdots + y_n}{n} = \overline{y} \qquad ①$$

$$\mathbf{E}\big(Y^2\big) = \frac{r}{\lambda^2} + \frac{r^2}{\lambda^2} \implies \frac{r_e}{\lambda_e^2} + \frac{r_e^2}{\lambda_e^2} = \frac{y_1^2 + \cdots + y_n^2}{n} \qquad ②$$

Substitute ① into ②, we have

$$\frac{\overline{y}}{\lambda_e} + (\overline{y})^2 = \frac{1}{n}\sum_{i=1}^{n} y_i^2 \implies \boxed{\lambda_e = \frac{\overline{y}}{\frac{1}{n}\sum y_i^2 - \overline{y}^2}} \qquad ③$$

Substitute ③ into ①, we have

$$r_e = \overline{y}\lambda_e = \boxed{\frac{\overline{y}^2}{\frac{1}{n}\sum y_i^2 - \overline{y}^2}}.$$

□

**Remark 1.3** *The* sample variance *is defined as*

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}(y_i-\overline{y})^2 &= \frac{1}{n}\sum_{i=1}^{n}\left(y_i^2 - 2y_i\overline{y} + \overline{y}^2\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} y_i^2 - 2\overline{y}\cdot\frac{\sum y_i}{n} + \frac{1}{n}\cdot n\overline{y}^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} y_i^2 - 2\overline{y}^2 + \overline{y}^2 \qquad\qquad \overline{y} = \frac{\sum y_i}{n} \\
&= \frac{1}{n}\sum_{i=1}^{n} y_i^2 - \overline{y}^2.
\end{aligned}$$

*So, in Example 1.3.7, if we define $\widehat{\sigma}^2$ to be the sample variance, we can further simply our estimate as follows:*

$$\lambda_e = \frac{\overline{y}}{\widehat{\sigma}^2}, \qquad r_e = \frac{\overline{y}^2}{\widehat{\sigma}^2}.$$

## 1.4 Interval Estimation

**Example 1.4.1** Estimate $\mu$, where $X \sim N(\mu, 1)$.

We take some samples and compute their sample means:

$$\overline{X}^1 = \frac{x_1 + \cdots + x_n}{n}, \overline{X}^2 = \frac{\widetilde{x_1} + \cdots + \widetilde{x_n}}{n}, \cdots$$

Finding the distribution of $\overline{X}$, we can find an interval $\left[\widehat{\theta}_L, \widehat{\theta}_U\right]$ such that

$$\mathbf{P}\left(\widehat{\theta}_L \le \overline{X} \le \widehat{\theta}_U\right) = 1 - \alpha.$$

**Remark 1.4** *By using the variance of the estimator, one can construct an interval such that with a high probability that the interval contains the unknown parameter.*

**Definition 1.4.2 (Confidence Interval).** The interval, $\left[\widehat{\theta}_L, \widehat{\theta}_U\right]$ is called the *confidence interval*, and the high probability is $1-\alpha$, where $\alpha$ is given.

**Remark 1.5** *Take* $\alpha = 5\%$*, then* $\left[\widehat{\theta}_L, \widehat{\theta}_U\right]$ *is the* $95\%$ *confidence interval of* $\mu$*. It does not mean that* $\mu$ *has* $95\%$ *chance to be in* $\left[\widehat{\theta}_L, \widehat{\theta}_U\right]$*. However, if we construct* $1000$ *such intervals,* $950$ *of them will contain* $\mu$*.*

**Example 1.4.3** A random sample of size $4$, $(Y_1 = 6.5,\ Y_2 = 9.2,\ Y_3 = 9.9,\ Y_4 = 12.4)$, from a normal population:

$$f_Y(y;\mu) = \frac{1}{\sqrt{2\pi}0.8} e^{-\frac{1}{2}\left(\frac{y-\mu}{0.8}\right)^2} \sim N(\mu, \sigma^2 = 0.64).$$

Both MLE and MME give $\mu_e = \overline{y} = 9.5$. The estimator $\widehat{\mu} = \overline{Y}$ follows normal distribution. Construct $95\%$-confidence interval for $\mu$.

***Solution 1.***

$\mathbf{E}(\overline{Y}) = \mu$ and $\mathbf{Var}(\overline{Y}) = \frac{\sigma^2}{n} = \frac{0.64}{4}$. By the Central Limit Theorem, $\overline{Y}$ approximately follow $N\left(\mu, \frac{\sigma^2}{n}\right)$. So, $\frac{\overline{Y}-\mu}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1)$. Then,

$$\mathbf{P}\left(z_1 \le \frac{\overline{Y}-\mu}{\sqrt{\frac{\sigma^2}{n}}} \le z_2\right) = 0.95 \implies \mathbf{P}\left(\overline{Y} - z_2\sqrt{\frac{\sigma^2}{n}} \le \mu \le \overline{Y} - z_1\sqrt{\frac{\sigma^2}{n}}\right) = 0.95$$

There are infinite many ways to construct a confidence interval by selecting different $z_1$ and $z_2$. However, since we don't have any prior knowledge on $\mu$, it is good for us to choose $z_1$ and $z_2$ symmetrically. Moreover, symmetric $z_1$ and $z_2$ will yield a smaller interval. We know the symmetric $z_1, z_2$ pair will be $z_1 = -1.96$ and $z_2 = 1.96$. Therefore,

$$\mathbf{P}\left(\overline{Y} - 1.96\sqrt{\frac{0.64}{4}} \le \mu \le \overline{Y} + 1.96\sqrt{\frac{0.64}{4}}\right) = 0.95.$$

Then, $95\%$ confidence interval is $[9.5 - 1.96 \times 0.4,\ 9.5 + 1.96 \times 0.4]$. □

**Theorem 1.4.4 Confidence Interval**

In general, for a normal population with $\sigma$ known, the $100(1-\alpha)\%$ *two-sided confidence interval* for $\mu$ is

$$\left(\overline{y} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \overline{y} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

**Theorem 1.4.5 Variation of Confidence Interval**

- One-sided interval:

$$\left(\overline{y} - z_\alpha \frac{\sigma}{\sqrt{n}},\, \overline{y}\right) \text{ or } \left(\overline{y},\, \overline{y} + z_\alpha \frac{\sigma}{\sqrt{n}}\right)$$

- $\sigma$ is unknown and sample size is small: $z$-score $\to$ $t$-score.
- $\sigma$ is unknown and sample size is large: $z$-score by CLT.
- Non Gaussian population but sample size is large: $z$-score by CLT.

**Theorem 1.4.6**
Let $k$ be the number of successes in $n$ independent trials, where $n$ is large and $p = \mathbf{P}(\text{success})$ is unknown. An approximate $100(1-\alpha)\%$ confidence interval for $p$ is the set of numbers

$$\left(\frac{k}{n} - z_{\alpha/2}\sqrt{\frac{(k/n)(1-k/n)}{n}},\, \frac{k}{n} + z_{\alpha/2}\sqrt{\frac{(k/n)(1-k/n)}{n}}\right).$$

**Definition 1.4.7 (Margin of Error).** The *margin of error*, denoted by $d$, is the quantity

$$d = z_{\alpha/2}\sqrt{\frac{(k/n)(1-k/n)}{n}}.$$

**Remark 1.6** *Stating the sample mean and the margin of error is equivalent to stating the confidence interval. Note that* $\mathrm{C.I.} = \widehat{p} \pm d$.

**Theorem 1.4.8 Estimate Margin of Error**
When $p$ is close to $\frac{1}{2}$, then $d \approx d_m = \frac{z_{\alpha/2}}{2\sqrt{n}}$, which is equivalent to $\sigma_n \approx \frac{1}{2\sqrt{n}}$. However, if $p$ is away from $\frac{1}{2}$, $d$ and $d_m$ are very different.

**Remark 1.7** *Theorem 1.4.8 gives aconservative estimation of the margin of error, which is* $d_m$.

**Proposition 1.9 :** Given $d$, we can estimate the sample size.

***Proof 2.***

$$d = z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} \implies n \approx \widehat{p}(1-\widehat{p})/\left(\frac{d}{z_{\alpha/2}}\right)^2.$$

However, since $n$ is unknown, $\widehat{p}$ is also unknown. We, therefore, need information on the actual $p$ to conclude an estimation of the sample size.

- If $p$ is known,
$$n = \frac{p(1-p)}{\left(\frac{d}{z_{\alpha/2}}\right)^2}.$$

- If $p$ is unknown. Let $f(p) = p(1-p)$. $f$ will be maximized when $p = 0.5$. So, $f(p) = p(1-p) \leq 0.25$. Then,
$$n \leq \frac{0.25}{\left(\frac{d}{z_{\alpha/2}}\right)^2}.$$
Since we are conservative, take $n = \frac{\frac{1}{4}z^2_{\alpha/2}}{d^2} = \frac{z^2_{\alpha/2}}{4d^2}$. This estimation is a conservative estimation of the sample size.

■

## 1.5 Properties of Estimation

The main question is that estimators are not unique in general. How do we choose a good estimator?

**Definition 1.5.1 (Unbiasedness).** Given a random sample of size $n$ when whose population distribution depends on an unknown parameter $\theta$. Let $\widehat{\theta}$ be an estimator of $\theta$. Then,

- $\widehat{\theta}$ is called *unbiased* if $\mathbf{E}(\widehat{\theta}) = \theta$.
- $\widehat{\theta}$ is called *asymptotically unbiased* if $\lim\limits_{n\to\infty} \mathbf{E}(\widehat{\theta}) = \theta$.
- If $\theta$ is biased, then the *bias* is given by the quantity $\mathbf{B}(\widehat{\theta}) = \mathbf{E}(\widehat{\theta}) - \theta$.

**Example 1.5.2** Consider the exponential distribution: $f_Y(y;\lambda) = \lambda e^{-\lambda y}$ for $y \geq 0$. Determine if the estimator $\widehat{\lambda} = \frac{1}{\overline{Y}}$ is biased or not.

*Hint:* $n\overline{Y} = \sum_{i=1}^{n} Y_i \sim$ *Gamma*$(n, \lambda)$.

***Solution 1.***

Recall that $\mathbf{E}[g(x)] = \int_x g(x) f_X(x)\,\mathrm{d}x$. Define $X = \sum_{i=1}^{n} Y_i \sim \text{Gamma}(n, \lambda)$. Also, recall the following facts:

$$\Gamma(n) = (n-1)! = (n-1)\Gamma(n-1)$$

and the integration over any probability density function will yield a result of $1$ by definition.

Then,

$$\begin{aligned}
\mathbf{E}(\widehat{\lambda}) = \mathbf{E}\left(\frac{1}{\overline{\overline{Y}}}\right) = \mathbf{E}\left(\frac{n}{\sum Y_i}\right) &= n\mathbf{E}\left(\frac{1}{\sum Y_i}\right) \\
&= n\mathbf{E}\left(\frac{1}{X}\right) \\
&= n\int_x \frac{1}{x}\cdot\frac{\lambda^n}{\Gamma(n)}x^{n-1}e^{-\lambda x}\,\mathrm{d}x \\
&= n\int_x \frac{\lambda^n}{(n-1)!}x^{n-2}e^{-\lambda x}\,\mathrm{d}x \\
&= \frac{n\lambda}{(n-1)}\underbrace{\int_x \frac{\lambda^{n-1}}{\Gamma(n-1)}x^{n-2}e^{-\lambda x}\,\mathrm{d}x}_{=1} \\
&= \frac{n}{n-1}\lambda.
\end{aligned}$$

Therefore, $\mathbf{E}(\widehat{\lambda}) \neq \lambda$, and so $\widehat{\lambda}$ is biased. However, note that

$$\lim_{n\to\infty}\mathbf{E}(\widehat{\lambda}) = \lim_{n\to\infty}\frac{n}{n-1}\lambda = \lambda.$$

By definition, then $\widehat{\lambda}$ is asymptotically unbiased. □

**Example 1.5.3** Consider the exponential distribution $f(y;\theta) = \dfrac{1}{\theta}e^{-y/\theta}$ for $y \geq 0$. Then, $\widehat{\theta} = \overline{Y}$ is unbiased.

**Remark 1.8** *Suppose* $\{X_1,\ldots,X_n\}$ *are* $i.i.d.$ *random variables, and* $\mathbf{E}(X_i) = \mu$ *for* $i = 1,\ldots,n$ *. Then,* $\overline{X}$*, the sample mean, is always an unbiased estimator:*

$$\mathbf{E}(\overline{X}) = \mathbf{E}\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{E}(X_i) = \frac{1}{n}\cdot n\cdot\mu = \mu.$$

**Theorem 1.5.4 Sample Variance is Biased**
Suppose $\{X_1,\ldots,X_n\}$ are $i.i.d.$ random variables, and $\mathbf{E}(X_i) = \mu$, $\mathbf{Var}(X_i) = \sigma^2$ for $i = 1,\ldots,n$. Then, the sample variance $\widehat{\sigma}^2 = \dfrac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$ is biased.

***Proof 2.*** Note that

$$
\begin{aligned}
\mathbf{E}(\widehat{\sigma}^2) &= \mathbf{E}\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\right) \\
&= \mathbf{E}\left(\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \mu + \mu - \overline{X}\right)^2\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbf{E}\Big[(X_i-\mu)^2 + \left(\mu - \overline{X}\right)^2 + 2(X_i-\mu)(\mu-\overline{X})\Big] \\
&= \frac{1}{n}\sum_{i=1}^{n}\left\{\underbrace{\mathbf{E}(X_i-\mu)^2}_{\mathbf{Var}(X_i)} + \mathbf{E}\left(\mu-\overline{X}\right)^2 + 2\mathbf{E}\big[(\mu-\overline{X})(X_i-\mu)\big]\right\}
\end{aligned}
$$

*Hint:* $\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu) = \frac{1}{n}\sum_{i=1}^{n}X_i - \frac{1}{n}\sum_{i=1}^{n}\mu = \overline{X} - \mu$

$$
\begin{aligned}
&= \frac{1}{n}\sum_{i=1}^{n}\mathbf{Var}(X_i) + \frac{1}{n}\cdot n\mathbf{E}\left(\mu-\overline{X}\right)^2 + 2\mathbf{E}\left[(\mu-\overline{X})\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\sigma^2 + \mathbf{E}\left(\mu-\overline{X}\right)^2 + 2\mathbf{E}\big[(\mu-\overline{X})(\overline{X}-\mu)\big] \\
&= \frac{1}{n}\cdot n\cdot\sigma^2 + \mathbf{E}\left(\mu-\overline{X}\right)^2 - 2\mathbf{E}\big[(\mu-\overline{X})^2\big] \\
&= \sigma^2 - \mathbf{E}\left(\mu-\overline{X}\right)^2 \\
&= \sigma^2 - \underbrace{\mathbf{E}\left(\overline{X}-\mu\right)^2}_{=\mathbf{Var}(\overline{X})} \\
&= \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2 \neq \sigma^2
\end{aligned}
$$

Therefore, $\widehat{\sigma}^2$ is not an unbiased estimator. ■

**Theorem 1.5.5 Adjusted Sample Variance is Unbiased**

With the same set up in Theorem 1.5.4, define the adjusted sample variance to be

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

Then, $S^2$ is an unbiased estimator of $\sigma^2$.

**Definition 1.5.6 (Decision Theory).** Minimize the error of an estimator (sample statistics) relative to the true parameter (population parameter) using a loss function.

**Definition 1.5.7 (Mean Squared Error).** The *mean squared error* (MSE) is defined by

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

**Theorem 1.5.8 Decomposition of MSE**
Generally,

$$\mathbf{MSE}(\theta) = \mathbf{Var}(\widehat{\theta}) + \mathbf{B}\left(\widehat{\theta}\right)^2$$

If $\widehat{\theta}$ is unbiased, $\mathbf{MSE}(\widehat{\theta}) = \mathbf{Var}(\widehat{\theta})$. $\mathbf{Var}(\theta)$ measures the precision of the estimator.

***Proof 3.*** Note that we will the following:

$$\begin{aligned}
\mathbf{MSE}(\widehat{\theta}) &= \mathbf{E}\left[(\widehat{\theta} - \theta)^2\right] \\
&= \mathbf{E}(\widehat{\theta}^2 + \theta^2 - 2\widehat{\theta}\theta) \\
&= \mathbf{E}(\widehat{\theta}) - 2\theta\mathbf{E}(\widehat{\theta}) + \theta^2 \\
&= \underbrace{\mathbf{E}(\widehat{\theta}^2) - \mathbf{E}(\widehat{\theta})^2}_{} + \underbrace{\mathbf{E}(\widehat{\theta})^2 - 2\theta\mathbf{E}(\widehat{\theta}) + \theta^2}_{} \\
&= \mathbf{Var}(\widehat{\theta}) + \left[\mathbf{E}(\widehat{\theta}) - \theta\right]^2 \\
&= \mathbf{Var}(\theta) + \mathbf{B}(\widehat{\theta})^2
\end{aligned}$$

If $\widehat{\theta}$ is unbiased, $\mathbf{B}(\widehat{\theta}) = 0$, and so $\mathbf{MSE}(\widehat{\theta}) = \mathbf{Var}(\widehat{\theta})$. ■

**Definition 1.5.9 (Efficiency).** Let $\widehat{\theta}_1$ and $\widehat{\theta}_2$ be two unbiased estimators for a parameter $\theta$. If we have $\mathbf{Var}(\widehat{\theta}_1) < \mathbf{Var}(\widehat{\theta}_2)$, then we say that $\widehat{\theta}_1$ is *more efficient* than $\widehat{\theta}_2$. The *relative efficiency* of $\widehat{\theta}_1$ with respect to $\widehat{\theta}_2$ is the ratio $\dfrac{\mathbf{Var}(\widehat{\theta}_2)}{\mathbf{Var}(\widehat{\theta}_1)}$.

## 1.6 Best Unbiased Estimator

**Definition 1.6.1 (Best/Minimum-Variance Estimator).** Let $\Theta$ be the set of all estimators $\widehat{\theta}$ that are unbiased for the parameter $\theta$. We way that $\widehat{\theta}^*$ is a *best* or *minimum-variance estimator* (MVE) if $\widehat{\theta}^* \in \Theta$ and $\mathbf{Var}(\widehat{\theta}^*) \leq \mathbf{Var}(\widehat{\theta}) \quad \forall\, \widehat{\theta} \in \Theta$.

**Definition 1.6.2 (Fisher's Information).** The *Fisher's information* of a continuous random variable $Y$ with pdf $f_Y(y;\theta)$ is defined as

$$\mathbf{I}(\theta) = \mathbf{E}\left[\left(\frac{\partial \ln f_Y(y;\theta)}{\partial\theta}\right)^2\right] = -\mathbf{E}\left[\frac{\partial^2}{\partial\theta^2}\ln f_Y(y;\theta)\right].$$

**Remark 1.9** *The Fisher's information measures the amount of information that a sample $Y$ contains about the unknown parameter $\theta$. If $\mathbf{I}(\theta)$ is big, then the curvature of $f_Y(y;\theta)$ is big, and* *thus it is more likely that we can find a region where* $\widehat{\theta}$ *is concentrated.*

**Extension 1.1 (Joint Fisher's Information)** *Suppose* $Y_1, \ldots, Y_n$ *are continuous* $i.i.d.$ *random variables, each has a Fisher's information of* $\mathbf{I}(\theta)$*. Then,*

$$\mathbf{E}\left[\left(\frac{\partial}{\partial\theta}\ln f_{Y_1,\ldots,Y_n}(y_1,\ldots,y_n;\theta)\right)^2\right] = n\mathbf{I}(\theta).$$

**Theorem 1.6.3 Properties of Fisher's Information**
Define the *Fisher's Score Function* $\frac{\partial}{\partial\theta}\ln f_Y(y;\theta)$. Then,

$$\mathbf{E}_Y\left[\frac{\partial}{\partial\theta}\ln f_Y(y;\theta)\right] = 0.$$

***Proof 1.*** Note that by chain rule, we have

$$\begin{aligned}
\mathbf{E}_Y\left[\frac{\partial}{\partial\theta}\ln f_Y(y;\theta)\right] &= \int_Y \left(\frac{\partial}{\partial\theta}\ln f_Y(y;\theta)\right) f_Y(y;\theta)\,\mathrm{d}y \\
&= \int_Y \frac{1}{f_Y(y;\theta)}\left(\frac{\partial}{\partial\theta} f_Y(y;\theta)\right) f_Y(y;\theta)\,\mathrm{d}y \\
&= \int_Y \frac{\partial}{\partial\theta} f_Y(y;\theta)\,\mathrm{d}y \\
&= \frac{\partial}{\partial\theta}\int_Y f_Y(y;\theta)\,\mathrm{d}y = \frac{\partial}{\partial\theta}(1) = 0.
\end{aligned}$$

■

**Corollary 1.4 :**

$$\mathbf{I}(\theta) = \mathbf{Var}\left(\frac{\partial}{\partial\theta}\ln f_Y(y;\theta)\right).$$

***Proof 2.*** By definition, we have

$$\begin{aligned}
\mathbf{Var}\left(\frac{\partial}{\partial\theta}\ln f_Y(y;\theta)\right) &= \mathbf{E}\left[\left(\frac{\partial}{\partial\theta}\ln f_Y(y;\theta)\right)^2\right] - \left(\underbrace{\mathbf{E}\left(\frac{\partial}{\partial\theta}\ln f_Y(y;\theta)\right)}_{=0,\text{ by Theorem 1.6.3.}}\right)^2 \\
&= \mathbf{E}\left[\left(\frac{\partial}{\partial\theta}\ln f_Y(y;\theta)\right)^2\right] \\
&= \mathbf{I}(\theta).
\end{aligned}$$

■

**Theorem 1.6.5 Cramér-Rao Inequality**
Under regular condition, let $Y_1, \ldots, Y_n$ be a random sample of size $n$ form the continuous population pdf $f_Y(y;\theta)$. Let $\widehat{\theta} = \widehat{\theta}(Y_1, \ldots, Y_n)$ be any unbiased estimator for $\theta$. Then,

$$\mathbf{Var}(\widehat{\theta}) \geq \frac{1}{n\mathbf{I}(\theta)}.$$

**Remark 1.10** *A similar statement holds for the discrete case* $p_X(k;\theta)$.

**Definition 1.6.6 (Efficiency of Unbiased Estimator).** An unbiased estimator $\widehat{\theta}$ is *efficient* if $\mathbf{Var}(\widehat{\theta})$ is equal to the Cramér-Rao lower bound. That is, $\mathbf{Var}(\widehat{\theta}) = (n\mathbf{I}(\theta))^{-1}$. Such an estimator is the MVE defined in Definition 1.6.1. The *efficiency* of an unbiased estimator $\widehat{\theta}$ is defined to be the quantity

$$\left(n\mathbf{I}(\theta)\mathbf{Var}(\widehat{\theta})\right)^{-1}.$$

**Example 1.6.7** Suppose $X \sim$ Bernoulli$(p)$. Is $\widehat{p} = \overline{X}$ efficient?

***Solution 3.***

Note that we have the following

$$\begin{aligned}
f_X(x;p) &= p^x(1-p)^{1-x}, \quad x = 0, 1 \\
\ln f_X(x;p) &= x \ln p + (1-x)\ln(1-p) \\
\frac{\partial}{\partial p} \ln f_X(x;p) &= \frac{x}{p} - \frac{1-x}{1-p} \\
\frac{\partial^2}{\partial p^2} \ln f_X(x;p) &= -\frac{x}{p^2} - \frac{1-x}{(1-p)^2}
\end{aligned}$$

Therefore, the Fisher's information can be computed by

$$\begin{aligned}
\mathbf{I}(p) = -\mathbf{E}\left[\frac{\partial^2}{\partial p^2} \ln f_X(x;p)\right] &= -\mathbf{E}\left[-\frac{x}{p^2} - \frac{1-x}{(1-p)^2}\right] \\
&= \mathbf{E}\left[\frac{x}{p^2}\right] + \mathbf{E}\left[\frac{1-x}{(1-p)^2}\right] \\
&= \frac{\mathbf{E}(x)}{p^2} + \frac{1-\mathbf{E}(x)}{(1-p)^2} \\
&= \frac{p}{p^2} + \frac{1-p}{(1-p)^2} = \frac{1}{p} + \frac{1}{1-p} = \frac{1}{p(1-p)}.
\end{aligned}$$

Note that

$$\mathbf{Var}(\overline{X}) = \mathbf{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathbf{Var}(X_i) = \frac{1}{n}\mathbf{Var}(X_i) = \frac{1}{n} \cdot p(1-p).$$

So, we have

$$\mathbf{Var}(\overline{X}) = \frac{p(1-p)}{n} = \frac{1}{n\left(\frac{1}{p(1-p)}\right)} = \frac{1}{n\mathbf{I}(p)}.$$

Therefore, $\widehat{p}$ is efficient. □

**Example 1.6.8** Suppose $X \sim N(\mu, \sigma^2)$, with $\sigma^2$ is known. What is $\mathbf{I}(\mu)$?

***Solution 4.***

Note that

$$\frac{\mathrm{d}^2}{\mathrm{d}\mu^2} \ln f_X(x;\mu) = -\frac{1}{\sigma^2}.$$

Then,

$$\mathbf{I}(\mu) = -\mathbf{E}\left[\frac{\mathrm{d}^2}{\mathrm{d}\mu^2} \ln f_X(x;\mu)\right] = -\mathbf{E}\left[-\frac{1}{\sigma^2}\right] = \frac{1}{\sigma^2}.$$

□

## 1.7 Sufficiency

**Remark 1.11** *Use Likelihood Function to Define Fisher's Information*

- *We can define the* score function *as* $\frac{\partial \ln \mathbf{L}(Y_1, \ldots, y_n;\theta)}{\partial\theta} = 0 \implies$ *MLE.*
- $\mathbf{E}\left[\frac{\partial \ln \mathbf{L}(Y;\theta)}{\partial\theta}\right] = 0$
- $\mathbf{I}(\theta) = \mathbf{E}\left[\left(\frac{\partial \ln \mathbf{L}(Y;\theta)}{\partial\theta}\right)^2\right] = -\mathbf{E}_Y\left[\frac{\partial^2 \ln \mathbf{L}(Y;\theta)}{\partial\theta^2}\right]$
- $-\mathbf{E}_Y\left[\frac{\partial^2 \ln \mathbf{L}(Y_1, \ldots, Y_n;\theta)}{\partial\theta^2}\right] = n\mathbf{I}(\theta)$.

  ***Proof 1.***

$$\begin{aligned}
-\mathbf{E}_Y\left[\frac{\partial^2 \ln \mathbf{L}(Y_1, \ldots, Y_n;\theta)}{\partial\theta^2}\right] &= -\mathbf{E}_Y\left[\frac{\partial^2}{\partial\theta^2} \ln \mathbf{L}(Y_1, \ldots, Y_m;\theta)\right] \\
&= -\mathbf{E}_Y\left[\frac{\partial^2}{\partial\theta^2} \ln\left(\prod_{i=1}^{n} f_Y(Y_i;\theta)\right)\right] \\
&= -\mathbf{E}_Y\left[\frac{\partial^2}{\partial\theta^2} \sum_{i=1}^{n} f_Y(y_i;\theta)\right] = \sum_{i=1}^{n}\left(-\mathbf{E}_Y\left[\frac{\partial^2}{\partial\theta^2} f_Y(y_i;\theta)\right]\right) = n\mathbf{I}(\theta)
\end{aligned}$$

■

- $\widehat{\theta_{MLE}} \xrightarrow{n\to\infty} N\left(\theta, \frac{1}{\mathbf{I}(\theta)}\right)$. *Note that* $\frac{1}{\mathbf{I}(\theta)}$ *is the C-R lower bound. We see that* $\widehat{\theta_{MLE}}$ *is asymptotically efficient.*

**Remark 1.12 (Sufficiency Intuition)** *Sufficiency tells us how much information can we get out of the data.*

**Rationale** *Let* $\widehat{\theta}$ *be an estimator to the unknown parameter* $\theta$. *Does* $\widehat{\theta}$ *contain all information about* $\theta$*?* e.g., The data itself is a sufficient estimator.

**Definition 1.7.1 (Sufficiency).** Let $(X_1, \ldots, X_n)$ be a random sample of size $n$ from a continuous population with an unknown parameter $\theta$. We call $\theta$ is *sufficient* if

$$f_{Y_1,\ldots,Y_n|\widehat{\theta}}\left(Y_1, \ldots, Y_n \mid \widehat{\theta} = \theta_e\right) = b(y_1, \ldots, y_n),$$

where $b(y_1, \ldots, y_n)$ is independent of $\theta$ ($\perp\!\!\!\perp \theta$). Also, $\widehat{\theta} = h(Y_1, \ldots, Y_n)$ and $\theta_e = h(y_1, \ldots, y_n)$. In this case, $\widehat{\theta}$ contains all the information about $\theta$ from $\{y_1, \ldots, y_n\}$.

**Example 1.7.2**

- Toss a coin $5$ times and get $3$ heads. Estimate $p =$ probability of $H$.

  ***Solution 2.***

$$\mathbf{P}\left(HHHTT \mid p_e = \frac{3}{5}\right) = \frac{1}{\binom{3}{5}} \perp\!\!\!\perp p \implies \text{sufficient}$$

$\square$

- A random sample of size $n$ from Bernoulli$(p)$. Check the sufficiency of $p = \sum_{i=1}^{n} X_i$.

  ***Solution 3.***

  Suppose the random sample is $\{X_1, \ldots, X_n\}$. Then, consider

$$\mathbf{P}(X_1 = x_1, \ldots, X_n = x_n, \sum_{i=1}^{n} X_i = C \mid \widehat{p} = C) = \frac{\mathbf{P}(X_1 = x_1, \ldots, X_n = x_n, \sum_{i=1}^{n} X_i = C)}{\mathbf{P}(\widehat{p} = C)}.$$

  What new information can $\sum_{i=1}^{n} X_i = C$ tell us? $X_n = C - \sum_{i=1}^{n-1} X_i$.

Note that $\mathbf{P}(\widehat{p}=C)=\mathbf{P}\left(\sum_{i=1}^{n}X_i=C\right)$. Since the summation of Bernoulli$(p)$ random variables is a Binomial$(n,p)$ random variable, we have $\mathbf{P}(\widehat{p}=C)=\binom{n}{C}p^C(1-p)^{n-C}$.

$\boxed{\text{Case I}}$ Suppose $\sum_{i=1}^{n}X_i=C$. Then,

$$
\begin{aligned}
&\frac{\mathbf{P}(X_1=x_1,\ldots,X_n=x_n,\sum_{i=1}^{n}X_i=C)}{\mathbf{P}(\widehat{p}=C)} \\
&=\frac{\left(\prod_{i=1}^{n-1}\right)p^{X_i}(1-p)^{1-X_i}p^{C-\sum_{i=1}^{n-1}X_i}(1-p)^{\left(1-C+\sum_{i=1}^{n-1}X_i\right)}}{\binom{n}{C}p^C(1-p)^{n-C}} \\
&=\frac{p^{\sum_{i=1}^{n-1}X_i+C-\sum_{i=1}^{n-1}X_i}(1-p)^{(n-1)-\sum_{i=1}^{n-1}X_i+1-C+\sum_{i=1}^{n-1}X_i}}{\binom{n}{C}p^C(1-p)^{n-C}} \\
&=\frac{p^C(1-p)^{n-C}}{\binom{n}{C}p^C(1-p)^{n-C}}=\frac{1}{\binom{n}{C}}\perp\!\!\!\perp p \implies \text{sufficient}
\end{aligned}
$$

$\boxed{\text{Case II}}$ Suppose $\sum_{i=1}^{n}X_i\neq C$. Then,

$$
\frac{\mathbf{P}(X_1=x_1,\ldots,X_n=x_n,\sum_{i=1}^{n}X_i=C)}{\mathbf{P}(\widehat{p}=C)}=\frac{0}{\mathbf{P}(\widehat{p}=C)}=0\perp\!\!\!\perp p \implies \text{sufficient}
$$

□

**Theorem 1.7.3 Factorization Property**
$\widehat{\theta}$ is sufficient if and only if the likelihood can be factorized as

$$\mathbf{L}(\theta) = \underbrace{g(\theta_e;\theta)}_{\theta_e=h(y_1,\ldots,y_n)\ \&\ \theta} \cdot \underbrace{u(y_1,\ldots,y_n)}_{\perp\!\!\!\perp\theta}.$$

## 1.8 Consistency

**Definition 1.8.1 (Consistency).** An estimator $\widehat{\theta}_n = h(W_1,\ldots,W_n)$ is said to be *consistent* if it converges to $\theta$ in probability; i.e., for any $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbf{P}\left(\left|\widehat{\theta}_n - \theta\right| < \varepsilon\right) = 1.$$

**Remark 1.13**
1. *Consistency is an asymptotical property (defined in a large sample limit).*
2. *$n$= sample size. $\left|\widehat{\theta}_n - \theta\right|$ is the distance between estimator and true $\theta$.*

**Lemma 1.2 Markov Inequality:** Suppose $X \geq 0$ is a random variable and $a > 0$ is a constant. Then,

$$\mathbf{P}(X \geq a) \leq \frac{\mathbf{E}(X)}{a}.$$

**Remark 1.14** *Markov inequality is good for determining extreme values. If $\mathbf{E}(X)$ is small, then it is very unlikely that $X$ will take some extremely large numbers.*

**Theorem 1.8.3 Chebyshev Inequality**
Let $W$ be some random variable with finite mean $\mu$ and variance $\sigma^2$. Then, for any $\varepsilon > 0$, we have

$$\mathbf{P}(|W-\mu| < \varepsilon) \leq 1 - \frac{\sigma^2}{\varepsilon^2}$$

or, equivalently,

$$\mathbf{P}(|W-\mu| \geq \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2}.$$

***Proof 1.*** Consider the random variable $|W-\mu|$. Then, by Markov Inequality,

$$\begin{aligned}\mathbf{P}(|X-\mu| \geq \varepsilon) &= \mathbf{P}\left(|X-\mu|^2 \geq \varepsilon^2\right) \\ &= \mathbf{P}\left((X-\mu)^2 \geq \varepsilon^2\right) \leq \frac{\mathbf{E}[(X-\mu)^2]}{\varepsilon^2} = \frac{\sigma^2}{\varepsilon^2}\end{aligned}$$

■

**Corollary 1.4 :** The sample mean $\widehat{\mu}_n = \dfrac{1}{n}\sum_{i=1}^{n} W_i$ is a consistent estimator for $\mathbf{E}(W) = \mu$, provided that the population $W$ has finite mean $\mu$ and variance $\sigma^2$.

**Proposition 1.5 :** If $\widehat{\theta}_n$ is an unbiased estimator of $\theta$, then $\widehat{\theta}_n$ is consistent if

$$\lim_{n\to\infty} \mathbf{Var}\left(\widehat{\theta}_n\right) = 0.$$

***Proof 2.*** Suppose $\widehat{\theta}_n$ is an unbiased estimator of $\theta$. Then, $\mathbf{E}\left(\widehat{\theta}_n\right) = \theta$. So, by Chebyshev Inequality, we have

$$\mathbf{P}\left(\left|\widehat{\theta}_n \theta\right| \geq \varepsilon\right) = \mathbf{P}\left(\left|\widehat{\theta}_n - \mathbf{E}\left(\widehat{\theta}_n\right)\right| \geq \varepsilon\right) \leq \frac{\mathbf{E}\left[\left(\widehat{\theta}_n - \mathbf{E}\left(\widehat{\theta}_n\right)\right)^2\right]}{\varepsilon^2} = \frac{\mathbf{Var}\left(\widehat{\theta}_n\right)}{\varepsilon^2}.$$

If we have $\mathbf{Var}\left(\widehat{\theta}_n\right) \to 0$ when $n \to \infty$, then

$$\lim_{n\to\infty} \mathbf{P}\left(\left|\widehat{\theta}_n - \theta\right| \geq \varepsilon\right) \leq \lim_{n\to\infty} \frac{\mathbf{Var}\left(\widehat{\theta}_n\right)}{\varepsilon^2} = \frac{0}{\varepsilon} = 0.$$

Therefore, it must be that $\lim\limits_{n\to\infty} \mathbf{P}\left(\left|\widehat{\theta}_n - \theta\right| \geq \varepsilon\right) = 0$ as probability cannot take negative values. Hence,

$$\begin{aligned}
\lim_{n\to\infty} \mathbf{P}\left(\left|\widehat{\theta}_n - \theta\right| < \varepsilon\right) &= \lim_{n\to\infty} \left(1 - \mathbf{P}\left(\left|\widehat{\theta}_n - \theta\right| \geq \varepsilon\right)\right) \\
&= 1 - \lim_{n\to\infty} \mathbf{P}\left(\left|\widehat{\theta}_n - \theta\right| \geq \varepsilon\right) \\
&= 1 - 0 = 1.
\end{aligned}$$

Then, by definition, $\widehat{\theta}_n$ is consistent. ■

## 1.9 Bayesian Estimator

**Theorem 1.9.1 Bayes' Rule**

$$\mathbf{P}(A \mid B) = \frac{\mathbf{P}(B \mid A)\mathbf{P}(A)}{\mathbf{P}(B \mid A)\mathbf{P}(A) + \mathbf{P}(B \mid A^C)\mathbf{P}(A^C)}.$$

$$\mathbf{P}(A \mid B^C) = 1 - \mathbf{P}(A \mid B) = \frac{\mathbf{P}(B^C \mid A)\mathbf{P}(A)}{\mathbf{P}(B^C \mid A)\mathbf{P}(A) + \mathbf{P}(B^C \mid A^C)\mathbf{P}(A^C)}.$$

**Rationale** Let $W$ be an estimator dependent on a parameter $\theta$.

1. Frequentists view $\theta$ as a parameter whose exact value to be estimated ($\theta$ is fixed).
2. Bayesians view $\theta$ is the value of a random variable $\Theta$. ($\theta$ *is uncertain and has its known parameter distribution*).

**Data Generation** The following procedure generates data with an additional layer of randomness.

1. $\theta$ is sampled from a distribution.
2. Under this $\theta$, we sample the data.

**Definition 1.9.2 (Prior distribution, Posterior distribution).** Our prior knowledge on $\Theta$ is called the *prior distribution*: $p_\Theta(\theta)$. The conditional distribution of the data given the parameter is the *likelihood*: $p(X \mid \Theta)$. Then, the Bayes' Rule will be

$$\underbrace{\mathbf{P}(\Theta \mid X)}_{\text{posterior distribution given the observation}} = \frac{\overbrace{\mathbf{P}(X \mid \Theta)}^{\text{likelihood}} \cdot \overbrace{\mathbf{P}(\Theta)}^{\text{prior distribution}}}{\underbrace{\mathbf{P}(X)}_{\text{margin distirbution of data}}}$$

**Theorem 1.9.3 Bayesian Estimator**

$$g_\Theta(\theta \mid W = w) = \begin{cases} \dfrac{p_W(w \mid \Theta = \theta) p_\Theta(\theta)}{p_W(w)} & \text{if } W \text{ and } \Theta \text{ are discrete} \\ \\ \dfrac{f_W(w \mid \Theta = \theta) f_\Theta(\theta)}{f_W(w)} & \text{if } W \text{ and } \Theta \text{ are constinuous,} \end{cases}$$

where

$$\begin{aligned} f_W(x) &= \int_H f_{W,\Theta}(w, \theta)\, \mathrm{d}\theta \quad \text{for } \theta \in H \\ &= \int_H f_W(w \mid \Theta = \theta) f_\Theta(\theta)\, \mathrm{d}\theta. \end{aligned}$$

Further, let $A = f_W(w) = \int_H f_W(w \mid \Theta = \theta) f_\Theta(\theta)\, \mathrm{d}\theta$. Then, $A$ normalizes $\text{likelihood} \times \text{prior}$:

$$1 = \int \frac{f_W(w \mid \Theta = \theta) f_\Theta(\theta)}{A}\, \mathrm{d}\theta.$$

So,

$$g_\Theta(\theta \mid W = w) = \text{constant} \cdot f_W(w \mid \Theta = \theta) f_\Theta(\theta) \quad \text{or} \quad \text{posterior} \propto \text{likelihood} \times \text{prior}.$$

**Example 1.9.4** A call center. Let $X$ =number of calls coming into the center. Then we know that $X \sim$ Poisson($\lambda$). This particular call center believes that $\Lambda$ is distributed with pdf

$$p_\Lambda(8) = 0.25 \quad \text{and} \quad p_\Lambda(10) = 0.75.$$

The call center believes that the number of calls coming into the center has recently changed, so they pick an hour and observe that $X = 7$ calls come in.

***Solution 1.***

We want to find: $\mathbf{P}(\Lambda = 8 \mid X = 7)$ and $\mathbf{P}(\Lambda = 10 \mid X = 7)$. By Bayes' Rule:

$$\begin{aligned}
\mathbf{P}(\Lambda = 8 \mid X = 7) &= \frac{\mathbf{P}(X = 7 \mid \Lambda = 8)\mathbf{P}(\Lambda = 8)}{\mathbf{P}(X = 7)} \\
&= \frac{\mathbf{P}(X = 7 \mid \Lambda = 8)\mathbf{P}(\Lambda = 8)}{\mathbf{P}(X = 7 \mid \Lambda = 8)\mathbf{P}(\Lambda = 8) + \mathbf{P}(X = 7 \mid \Lambda = 10)\mathbf{P}(\Lambda = 10)} \\
&= \frac{e^{-8}\left(\frac{8^7}{7!}\right)(0.25)}{e^{-8}\left(\frac{8^7}{7!}\right)(0.25) + e^{-10}\left(\frac{10^7}{7!}\right)(0.75)} \approx 0.66
\end{aligned}$$

Then, $\mathbf{P}(\Lambda = 10 \mid X = 7) = 1 - \mathbf{P}(\Lambda = 8 \mid X = 7) = 1 - 0.66 = 0.34$. Or, alternatively, we can use the Bayes' Rule again. □

Table 1: Convention of Picking a Prior Distribution

| Parameter | Prior Distribution |
|---|---|
| Bernoulli($p$) | Beta |
| Binomial($p$) | Beta |
| Poisson($\lambda$) | Gamma |
| Exponential($\lambda$) | Gamma |
| Normal($\mu$) | Normal |
| Normal($\sigma^2$) | Inverse Gamma |

**Remark 1.15** *When we have no prior knowledge on the belief, we choose a uniform distribution.*

**Example 1.9.5** Consider an unfair coin $\Theta$ (a random variable indicating the probability of getting head). Flip the coin $n$ times, $X$ = number of heads. Find the posterior distribution.

***Solution 2.***

By the Bayes' rule,

$$f_{\Theta|X}(\theta \mid X = x) = \frac{f_\Theta(\theta)\mathbf{P}(X = k \mid \theta)}{\mathbf{P}(X = k)}.$$

We know $\theta \in [0, 1]$, so $\Theta \sim$ Uniform$[0, 1]$ and $f_\Theta(\theta) = 1$. So,

$$f_{\Theta|X}(\theta \mid X = x) = \frac{1 \cdot \binom{n}{k} \cdot \theta^k (1-\theta)^{n-k}}{\mathbf{P}(X = k)} = \underbrace{\frac{1 \cdot \binom{n}{k}}{\mathbf{P}(X = k)}}_{\text{constant}} \theta^k (1-\theta)^{n-k}$$

**Definition 1.9.6 (Beta Distribution).** For a distribution Beta$(\alpha, \beta)$, the pdf is given by

$$f_Y(y; \alpha, \beta) = \frac{y^{\alpha-1}(1-y)^{\beta-1}}{\mathbf{B}(\alpha, \beta)} \quad \text{for } y \in [0,1] \text{ and } \alpha, \beta > 0,$$

where

$$\mathbf{B}(\alpha, \beta) := \int_0^1 y^{\alpha-1}(1-y)^{\beta-1}\,\mathrm{d}y = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}, \quad \alpha, \beta > 0.$$

The expectation of $X \sim$ Beta$(\alpha, beta)$ is given by

$$\mathbf{E}(X) = \frac{\alpha}{\alpha + \beta}.$$

Disregarding the constant, $\theta^k(1-\theta)^{n-k}$ is part of the Beta distribution with $\alpha = k+1$ and $\beta = n-k+1$. So, $\Theta \sim$ Beta$(k+1, n-k+1)$. To form a distribution, the constant must, therefore, be

$$\begin{aligned}
\frac{\binom{n}{k}}{\mathbf{P}(X = k)} = \frac{1}{\mathbf{B}(k+1, n-k+1)} &= \frac{\Gamma(k+1+n-k+1)}{\Gamma(k+1)\Gamma(n-k+1)} \\
&= \frac{\Gamma(n+2)}{\Gamma(k+1)\Gamma(n-k+1)} \\
&= \frac{(n+1)!}{k!(n-k)!} \qquad\qquad If\ n \in \mathbb{N},\ then\, \Gamma(n) = (n-1)!
\end{aligned}$$

Note that Beta$(\alpha = 1, \beta = 1) =$ Unform$(0, 1)$. So, in this example,

$$\text{Beta}(1, 1) \xrightarrow{\text{Data}} \text{Beta}(k+1, n-k+1).$$

Moreover, $\mathbf{E}(\Theta) = \dfrac{k+1}{k+1+n-k+1} = \dfrac{k+1}{n+2}$. □

**Example 1.9.7** Let $X_1, \ldots, X_n$ be a random sample form Bernoulli($\theta$): $p_X(k;\theta) = \theta^k(1-\theta)^{1-k}$ for $k = 0, 1$. Let $X = \sum_{i=1}^{n} X_i$. Then, $X$ follows Binomial($n, \theta$). Consider the prior distribution $\Theta \sim$ Beta($r, s$), i.e., $f_\Theta(\theta) = \frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)}\theta^{r-1}(1-\theta)^{s-1}$ for $\theta \in [0,1]$. Then, the posterior distribution is

$$\Theta \mid X \sim \text{Beta}(r+k, s+n-k).$$

***Proof 3.*** Note that

$$\begin{aligned}
f_{\Theta|X}(\theta \mid X = x) &= \frac{p_X(X=k \mid \theta) f_\Theta(\theta)}{\displaystyle\int_0^1 p_X(X=k \mid \theta) f_\Theta(\theta)\,\mathrm{d}\theta} \\
&= \frac{\binom{n}{k}\theta^k(1-\theta)^{n-k}\frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)}\theta^{r-1}(1-\theta)^{s-1}}{\displaystyle\int_0^1 \binom{n}{k}\theta^k(1-\theta)^{n-k}\frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)}\theta^{r-1}(1-\theta)^{s-1}\,\mathrm{d}\theta} \\
&= \frac{\binom{n}{k}\frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)}\theta^{k+r-1}(1-\theta)^{n-k+s-1}}{\binom{n}{k}\frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)}\displaystyle\int_0^1 \theta^{k+r-1}(1-\theta)^{n-k+s-1}\,\mathrm{d}\theta}
\end{aligned}$$

Note that $\theta^{k+r-1}(1-\theta)^{n-k+s-1}$ is part of Beta($k+r, n-k+s$). So,

$$\begin{aligned}
1 &= \int_0^1 \frac{\Gamma(k+r+n-k+s)}{\Gamma(k+r)\Gamma(n-k+s)}\theta^{k+r-1}(1-\theta)^{n-k+s-1}\,\mathrm{d}\theta \\
1 &= \frac{\Gamma(r+n+s)}{\Gamma(k+r)\Gamma(n-k+s)}\int_0^1 \theta^{k+r-1}(1-\theta)^{n-k+s-1}\,\mathrm{d}\theta \\
\int_0^1 \theta^{k+r-1}(1-\theta)^{n-k+s-1}\,\mathrm{d}\theta &= \frac{\Gamma(k+r)\Gamma(n-k+s)}{\Gamma(r+n+s)}.
\end{aligned}$$

Therefore,

$$f_{\Theta|X}(\theta \mid X = x) = \frac{\theta^{k+r-1}(1-\theta)^{n-k+s-1}}{\frac{\Gamma(k+r)\Gamma(n-k+s)}{\Gamma(r+n+s)}} = \frac{\Gamma(r+n+s)}{\Gamma(k+r)\Gamma(n-k+s)}\theta^{k+r-1}(1-\theta)^{n-k+s-1}.$$

This is exactly a Beta distribution with parameter $\alpha = k+r$ and $\beta = n-k+s$. ∎

**Definition 1.9.8 (Conjugate Prior).** If the posterior distributions $p(\Theta \mid X)$ are in the sample probability distribution family as the prior probability distribution $p(\Theta)$, the prior and posterior are called *conjugate distributions* and the prior is called a *conjugate prior* for the likelihood function.

**Remark 1.16** *Common Conjugate Priors*

- *Beta distributions are conjugate priors for Bernoulli, Binomial, Negative binomial, and Geometric likelihood.*
- *Gamma distributions are conjugate priors for Poisson and Exponential likelihood*

**Definition 1.9.9 (Bayesian Point Estimation).** Given the posterior $f_{\Theta|W}(\theta \mid W = w)$, how can one calculate the appropriate point estimate $\theta_e$?

**Definition 1.9.10 (Loss Function).** Let $\theta_e$ be an estimate for $\theta$ based on a statistic $W$. The *loss function* associated with $\theta_e$ is denoted $\mathbf{L}(\theta_e, \theta)$, where $\mathbf{L}(\theta_e, \theta) \geq 0$ and $\mathbf{L}(\theta, \theta) = 0$.

- The lost function is $\mathbf{E}\left[\mathbf{L}(\widehat{\theta}, \theta)\right]$.
- The MSE, mean square error, is $\mathbf{E}\left[\left(\widehat{\theta} - \theta\right)^2\right]$.

1. If we have not data, then notice that

$$\mathbf{E}\left[(\theta - c)^2\right] = \mathbf{E}\left(\theta^2\right) + \mathbf{E}\left(c^2\right) - 2c\mathbf{E}(\theta)$$

is minimized at $c = \mathbf{E}(\theta)$. Therefore,

$$\min \mathbf{E}\left[(\theta - \widehat{\theta})^2\right] = \mathbf{E}[(\theta - \mathbf{E}(\theta))]^2 = \mathbf{Var}(\theta).$$

So, $\widehat{\theta}^* = \mathbf{E}(\theta)$, the prior expectation.

2. If we have data $X = x$, then

$$\min \mathbf{E}\left[(\theta - \widehat{\theta})^2 \mid X = x\right] \implies \widehat{\theta}^* = \mathbf{E}[\theta \mid X = x].$$

This $\widehat{\theta}^*$ is called the posterior expectation.

**Theorem 1.9.11 Squared-Loss Bayesian Estimation**

**Step 1.** Solve the posterior distribution.

**Step 2.** Calculate the posterior expectation.

Generally, if we know the posterior pdf $f_\Theta(\theta \mid X = x)$, the point estimate is

$$\mathbf{E}[\theta \mid X = x] = \int_\Theta \theta f_\Theta(\theta \mid X = x)\, \mathrm{d}\theta.$$

**Theorem 1.9.12**
Let $f_\Theta(\theta \mid W = w)$ be the posterior distribution of the random variable $\Theta$.

- If $\mathbf{L}(\theta_e, \theta) = |\theta_e - \theta|$, then the Bayesian point estimate for $\theta$ is the median of the posterior distribution $f_\Theta(\theta \mid W = w)$;
- If $\mathbf{L}(\theta_e, \theta) = (\theta_e - \theta)^2$, then the Bayesian point estimate for $\theta$ is the mean of the posterior distribution $f_\Theta(\theta \mid W = w)$.

# 2 Inference Based on Normal

## 2.1 Sample Variance and Chi-Square Distribution

Recall that if $Y \sim \text{Normal}(\mu, \sigma^2)$, we have MLEs defined as

$$\widehat{\mu} = \overline{Y} \quad \text{and} \quad \widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2.$$

If $\sigma$ is known, we can do the interval estimation:

$$Z := \frac{\overline{Y} - \mathbf{E}(\overline{Y})}{\sqrt{\mathbf{Var}(\overline{Y})}} \sim N(0,1).$$

However, what if we don't know $\sigma$? We will have to estimate it with a sample variance.

**Definition 2.1.1 (Sample Variance).** To estimate $\sigma^2$, we define the following unbiased *sample variance*:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2.$$

**Remark 2.1** *We often compute $S^2$ using the fact that*

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n} y_i^2 - n\overline{y}^2 \quad \textit{i.e.,}\ S^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} y_i^2 - n\overline{y}^2\right]$$

**Definition 2.1.2 (Chi-Squared Distribution).** Suppose $W_k \sim \chi^2(k)$, the *chi-squared distribution with degree of freedom $k$*. Then,

$$W_k = Z_1^2 + Z_2^2 + \cdots + Z_k^2, \text{ where } Z_i \overset{i.i.d.}{\sim} N(0,1).$$

$k$ is called the *degree of freedom* of the chi-squared distribution and is denoted as $df = k$.

**Theorem 2.1.3 Chi-Squared Distribution and Gamma Distribution**
$\chi^2(1)$ is equivalent to Gamma$\left(\frac{1}{2}, \frac{1}{2}\right)$. Hence, $\chi^2(n)$ is equivalent to Gamma$\left(\frac{n}{2}, \frac{1}{2}\right)$.

***Proof 1.*** Recall: For $Y_1 \sim \text{Gamma}(n, \lambda)$ and $Y_2 \sim \text{Gamma}(m, \lambda)$, we have the following sum rule

$$Y_1 + Y_2 \sim \text{Gamma}(n + m, \lambda).$$

Then, as $Z_1^2 \sim \chi^2(1) = \text{Gamma}\left(\frac{1}{2}, \frac{1}{2}\right)$, we have

$$Z_1^2 + Z_2^2 + \cdots + Z_n^2 \sim \chi^2(n) = \text{Gamma}\left(\frac{1}{2} + \cdots + \frac{1}{2}, \frac{1}{2}\right) = \text{Gamma}\left(\frac{n}{2}, \frac{1}{2}\right).$$

■

**Theorem 2.1.4 Expectation and Variance of $\chi^2(n)$**
If $W_n \sim \chi^2(n)$, then

$$\mathbf{E}(W_n) = n = df \quad \text{and} \quad \mathbf{Var}(W_n) = 2n$$

***Proof 2.*** For $Y \sim \text{Gamma}(n, \lambda)$, $\mathbf{E}(Y) = \dfrac{n}{\lambda}$ and $\mathbf{Var}(Y) = \dfrac{n}{\lambda^2}$. As $W_n \sim \chi^2(n) = \text{Gamma}\left(\frac{n}{2}, \frac{1}{2}\right)$, we have

$$\mathbf{E}(W_n) = \frac{n/2}{1/2} = n \quad \text{and} \quad \mathbf{Var}(W_n) = \frac{n/2}{1/4} = 2n.$$

■

**Theorem 2.1.5**
Consider a random sample $Y_1, \ldots, Y_n$ drawn from $N(0,1)$. Let $S^2$ be the sample variance and $\overline{Y}$ be the sample mean. Then,

- $S^2$ and $\overline{Y}$ are independent;
- $\dfrac{(n-1)}{\sigma^2} S^2 \sim \chi^2(n-1)$

**Remark 2.2** *We can think of the second bullet point as the following rationale: knowing $\overline{Y}$, we only need $(n-1)$ data, and we can calculate $Y_n$ from $\overline{Y}$ and $Y_1, \ldots, Y_{n-1}$. This explains why the chi-squared distribution is of $df = n - 1$.*