6 Point Estimation of Parameters and Sampling Distributions

6.1 Point Estimation

Definition. Estimation(估計) is using sample data to estimate the parameter of probability distribution function.

A method of estimation is building a estimator(估測器). It is known there are point estimation and interval estimation. In this chapter, we will discuss point estimation.

Definition. A **point estimate** of some population parameter(母體參數) θ is a single numerical value $\hat{\theta}$ of a statistic $\hat{\Theta}$. The statistic $\hat{\Theta}$ is called the **point estimator**.

Example. Let $\theta = \begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix}$ Given the sample data, how to find $\hat{\theta} = \begin{bmatrix} \hat{\mu} \\ \hat{\sigma}^2 \end{bmatrix}$?

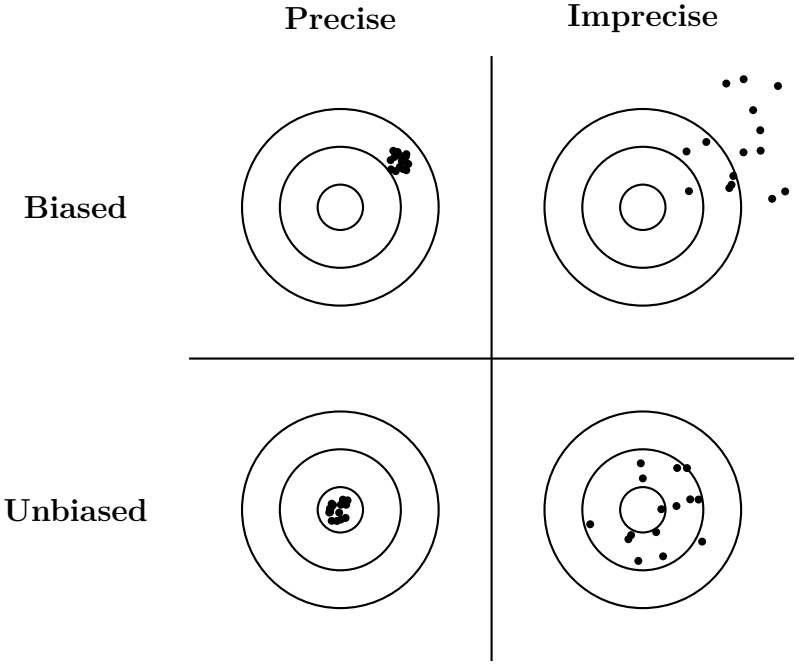
$$\text{Sample data} \longrightarrow \text{Point Estimator} \longrightarrow \hat{\theta}, \text{the point estimator of } \theta$$

6.2 General Concepts of Point Estimation (1)

6.2.1 Unbiased Estimators

How good is this estimator? There are two indexes to evaluate a effectiveness of estimator.

1. Unbiasedness(無偏差): Let θ and $\hat{\theta}$ be parameter and estimator. If $E(\hat{\theta}) = \theta$, then we called it is unbiased estimation.
2. Consistency(一致性): That means $\hat{\theta}$ has a small variance. For large sample size n, n is increasing $\implies$ $Var(\hat{\theta})$ is decreasing. i.e. $Var(\hat{\theta}) \to 0$ as $n \to \infty$.



6.2.2 Variance of a Point Estimator

Definition. If we consider all unbiased estimators of θ, the one with the smallest variance is called the **minimum variance unbiased estimator**(MVUE)[1].

Definition. If $X_1, X_2, \ldots, X_n$ is a random sample of size n from a normal distribution with mean μ and variance σ^2, the sample mean $\bar{X}$ is the MVUE for μ.

Theorem. Let $X_1, X_2, \ldots, X_n$ be a random variables of size n for a distribution with mean μ. Then the sample mean $\bar{X} = \dfrac{X_1 + X_2 + \cdots + X_n}{n}$ is an unbiased estimator for μ.

Proof. Let $\theta = \mu$ be the parameter of sample data $X_1, X_2, \ldots, X_n$, and denote estimator

$$\hat{\Theta} = \bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

Thus

$$\begin{aligned} E(\bar{X}) &= E\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right) \\ &= \frac{1}{n}E(X_1 + X_2 + \cdots + X_n) \\ &= \frac{1}{n}\sum_{k=1}^{n} E(X_k) \\ &= \frac{1}{n} \cdot n\mu = \mu \end{aligned}$$

Hence $\bar{X} = \dfrac{X_1 + X_2 + \cdots + X_n}{n}$ is an unbiased estimator for μ. □

Theorem. Let $\bar{X}$ be the sample mean of $X_1, X_2, \ldots, X_n$ from a distribution with mean μ and variance σ^2. Then $Var(\bar{X}) = \dfrac{\sigma^2}{n}$ implies $Var(\bar{X}) \to 0$ as $n \to \infty$.

6.2.3 Standard Error: Reporting a Point Estimate

Definition. The **standard error** of an estimator $\hat{\Theta}$ is its standard deviation given by $\sigma_{\hat{\Theta}} = \sqrt{V\left(\hat{\Theta}\right)}$. If the standard error involves unknown parameters that can be estimated, substitution of those values into $\sigma_{\hat{\Theta}}$ produces an estimated standard error, denoted by $\hat{\sigma}_{\hat{\Theta}}$.

Remark. The standard error of the sample mean $\bar{X}$ with standard deviation σ and size n, we have the standard error of $\bar{X}$ is

$$\sqrt{Var(\bar{X})} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}.$$

[1]MVUE: 最小變異數不偏估計

Definition. The **sample variance**[2] from data X_1, X_2, ..., X_n, denotes S^2 and is defined as

$$S^2 = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1}.$$

Theorem. Let S^2 be a sample variance with sample data X_1, X_2, ..., X_n form a distribution with mean μ and variance σ^2. Then S^2 is an unbiased estimator for σ^2, that is

$$E(S^2) = \sigma^2.$$

Proof. Since

$$S^2 = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1}.$$

This implies

$$\begin{aligned}
E(S^2) &= E\left(\sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1}\right) \\
&= \frac{1}{n-1} E\left[\sum_{i=1}^{n} (X_i - \bar{X})^2\right] \\
&= \frac{1}{n-1} E\left[\sum_{i=1}^{n} X_1^2 - 2\bar{X}X_i + \bar{X}^2\right] \\
&= \frac{1}{n-1} E\left[\sum_{i=1}^{n} X_1^2 - 2\bar{X}\sum_{i=1}^{n} X_i + \sum_{i=1}^{n} \bar{X}^2\right] \\
&= \frac{1}{n-1} E\left[\sum_{i=1}^{n} X_1^2 - 2n\bar{X}^2 + n\bar{X}^2\right] \\
&= \frac{1}{n-1} E\left[\sum_{i=1}^{n} X_1^2 - n\bar{X}^2\right] \\
&= \frac{1}{n-1} \left[\sum_{i=1}^{n} E(X_i) - nE(\bar{X})\right] && E(\bar{X}) = \frac{\sigma^2}{n} + \mu^2 \\
&= \frac{1}{n-1} \left[\sum_{i=1}^{n} (\sigma^2 + \mu^2) - n\left(\frac{\sigma^2}{n} + \mu^2\right)\right] && E(X) = \sigma^2 + \mu^2 \\
&= \frac{1}{n-1} (n\sigma^2 + n\mu^2 - \sigma^2 - n\mu^2) \\
&= \frac{1}{n-1} (n\sigma^2 - \sigma^2) \\
&= \frac{1}{n-1} (n-1)\sigma^2 \\
&= \sigma^2.
\end{aligned}$$

□

[2]Sample variance: 樣本變異數

6.3 Sampling Distributions and the Central Limit Theorem

Definition. Let $X_1, X_2, \ldots, X_n$ be a **random sample** with size n. **Statistic** is called a function of the observations. The probability distribution of a statistic is called **sampling distribution**.

Theorem. (Central Limit Theorem) If $X_1, X_2, \ldots, X_n$ is a random sample of size n taken from a population (either finite or infinite) with mean μ and finite variance σ^2 and if $\bar{X}$ is the sample mean, the limiting form of the distribution of

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

as $n \to \infty$, is the **standard normal distribution**.

Example. Suppose that a random variable **X** has a continuous uniform distribution

$$f(x) = \begin{cases} 1/2, & 4 \le x \le 6 \\ 0, & \text{otherwise} \end{cases}$$

Find the distribution of the sample mean of a random sample of size $n = 40$.

Sol. Since $X \sim \text{Uniform}(4, 6)$, we obtain $\mu = \dfrac{4+6}{2} = 5$ and $\sigma^2 = \dfrac{(6-4)^2}{12} = \dfrac{1}{3}$. Let $\bar{X} = \dfrac{X_1 + X_2 + \cdots + X_{40}}{40}$, by the central limit theorem, the distribution of $\bar{X}$ is normal such that $\bar{X} \sim \mathcal{N}(\mu_{\bar{X}}, \sigma^2_{\bar{X}})$. We know $\mu_{\bar{X}} = E(\bar{X}) = \mu = 5$ and $\sigma^2_{\bar{X}} = \dfrac{\sigma^2}{n} = \dfrac{1/3}{40} = \dfrac{1}{120}$. Hence

$$\bar{X} \sim \mathcal{N}\left(5, \frac{1}{120}\right).$$

Definition. Let $\bar{X}_1$ and $\bar{X}_2$ be distinct sample means, the difference of them is defined

$$\bar{X}_1 - \bar{X}_2$$

Definition. If we have two independent populations with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2 and if $\bar{X}_1$ and $\bar{X}_2$ are the sample means of two independent random samples of sizes n_1 and n_2 from these populations, then the sampling distribution of

$$Z = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

is approximately standard normal if the conditions of the central limit theorem apply. If the two populations are normal, the sampling distribution of Z is exactly standard normal.

Example. (Aircraft Engine Life) The effective life of a component used in jet-turbine(渦輪) aircraft engine is a random variable with mean 5000 and SD 40 hours and is close to a normal distribution. The engine manufacturer introduces an improvement into the manufacturing process for this component that changes the parameters to 5050 and 30. Random samples of size 16 and 25 are selected. What is the probability that the difference in the two sample means is at least 25 hours?

Sol. Let $\mu_1 = 5000$, $\mu_2 = 5050$, $\sigma_1 = 40$, $\sigma_2 = 30$, $n_1 = 16$, $n_2 = 25$. Compute

$$\frac{\sigma_1}{\sqrt{n_1}} = \frac{40}{4} = 10, \quad \frac{\sigma_2}{\sqrt{n_2}} = \frac{30}{5} = 6$$
$$\mu_2 - \mu_1 = 50$$

Thus

$$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{136} \approx 11.66,$$

we obtain $X \sim \mathcal{N}(50,\, 136)$. Set $X = 25$, we have

$$Z = \frac{25 - 50}{\sqrt{136}} = \frac{-25}{\sqrt{136}} \approx -2.14.$$

Using excel `1-NORMSDIST(Z)`, hence $P(Z \leq -2.14) = 0.9838$.

6.4 General Concepts of Point Estimation (2)

6.4.1 Mean Squared Error of an Estimator

Definition. The **mean squared error of an estimator**[3] $\hat{\Theta}$ of the parameter θ is defined as

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

Definition. Suppose there are two estimators $\hat{\Theta}_1$ and $\hat{\Theta}_2$ with parameter θ. Then their **relative efficiency**[4] is defined as

$$\frac{\text{MSE}(\hat{\Theta}_1)}{\text{MSE}(\hat{\Theta}_2)}.$$

Theorem. If $\dfrac{\text{MSE}(\hat{\Theta}_1)}{\text{MSE}(\hat{\Theta}_2)} < 1$, then $\hat{\Theta}_1$ is superior to[5] $\hat{\Theta}_2$.

Remark.

$$\begin{aligned}\text{MSE}(\hat{\Theta}) &= E(\hat{\Theta} - \theta)^2 \\ &= E(\Theta - E(\hat{\Theta}))^2 + (\theta - E(\hat{\Theta}))^2\end{aligned}$$

$E(\Theta - E(\hat{\Theta}))^2$ is the variance of $\hat{\Theta}$. $(\theta - E(\hat{\Theta}))^2$ is the bias of $\hat{\Theta}$; if $E(\hat{\Theta}) = \theta$, then it is the unbiased, that is $(\theta - E(\hat{\Theta}))^2 = 0$.

Example. Let X be a random sample size $2n$ form a population with $E(X) = \mu$ and $Var(X) = \sigma^2$. Let

$$\hat{\Theta}_1 : E(\bar{X}_1) = \frac{1}{2n}\sum_{i=1}^{2n} X_i$$

$$\hat{\Theta}_2 : E(\bar{X}_2) = \frac{1}{n}\sum_{i=1}^{n} X_i$$

be two estimators for $\theta = \mu$. Which is better?

Sol. For $\hat{\Theta}_1 = \bar{X}_1$, compute

$$\begin{aligned}E(\bar{X}_1) &= E\left(\frac{1}{2n}\sum_{i=1}^{2n} X_i\right) \\ &= \frac{1}{2n}E\left(\sum_{i=1}^{2n} X_i\right) \\ &= \frac{1}{2n}\sum_{i=1}^{2n} E\left(X_i\right) \\ &= \frac{1}{2n}2n\mu = \mu\end{aligned}$$

[3]MSE: 均方差，用來比較兩個 estimators 的 variance 大小
[4]Relative efficiency: 相對效能
[5]superior to: 優於

Thus $\hat{\Theta}_1$ is unbiased, that is $(\theta - E(\hat{\Theta}_1))^2 = 0$. Hence

$$\begin{aligned}\text{MSE}(\hat{\Theta}_1) &= E(\Theta - E(\hat{\Theta}))^2 + (\theta - E(\hat{\Theta}))^2 \\ &= \frac{\sigma^2}{2n} + 0 = \frac{\sigma^2}{2n}\end{aligned}$$

For $\hat{\Theta}_2 = \bar{X}_2$, compute

$$\begin{aligned}E(\bar{X}_2) &= E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\ &= \frac{1}{n}E\left(\sum_{i=1}^{n} X_i\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} E\left(X_i\right) \\ &= \frac{1}{n}n\mu = \mu\end{aligned}$$

Then $\hat{\Theta}_2$ is unbiased, that is $(\theta - E(\hat{\Theta}_2))^2 = 0$. Hence

$$\begin{aligned}\text{MSE}(\hat{\Theta}_1) &= E(\Theta - E(\hat{\Theta}))^2 + (\theta - E(\hat{\Theta}))^2 \\ &= \frac{\sigma^2}{n} + 0 = \frac{\sigma^2}{n}\end{aligned}$$

Therefore, the relative efficiency is

$$\frac{\text{MSE}(\hat{\Theta}_1)}{\text{MSE}(\hat{\Theta}_2)} = \frac{\sigma^2/2n}{\sigma/n} = \frac{1}{2} < 1.$$

So $\hat{\Theta}_1$ is superior to $\hat{\Theta}_2$, $\bar{X}_1$ is better than $\bar{X}_2$.

Example. Let X_1, X_2, ..., X_n be a random sample from a population that is $\mathcal{N}(\mu, \sigma^2)$. We plan to use the following estimator:

$$\hat{\Theta} = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{c}$$

to estimate σ^2, that is $\theta = \sigma^2$. Compute the bias in $\hat{\Theta}$ as an estimator of σ^2 and a function of the constant c.

Sol. Since

$$S^2 = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1},$$

we have

$$(n-1)S^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Thus we compute

$$\begin{aligned} E(\hat{\Theta}) &= E\left[\frac{(n-1)S^2}{c}\right] \\ &= \frac{(n-1)}{c}E\left[S^2\right] \end{aligned}$$

We know that the sample variance S^2 is an unbiased estimator of the population variance σ^2, which means:

$$E[S^2] = \sigma^2$$

Substituting $E[S^2] = \sigma^2$ into our equation gives:

$$E(\hat{\Theta}) = \frac{(n-1)}{c}\sigma^2.$$

Compute the bias

$$\begin{aligned} E(\hat{\Theta}) - \sigma^2 &= \frac{(n-1)}{c}\sigma^2 - \sigma^2 \\ &= \sigma^2\left(\frac{n-1}{c} - 1\right) \end{aligned}$$

Hence the bias is

$$E(\hat{\Theta}) - \theta = \sigma^2 \cdot \left(\frac{n-1}{c} - 1\right).$$

6.4.2 Bootstrap Standard Error

The bootstrap sampling is a **resampling method**[6] by independently sampling with replacement from an existing sample data with sample size n, and performing inference[7] among these resampled data.

Definition. A **bootstrap sampling** is a random sample drawn with replacement from the observed sample $\mathcal{S}$ of the same size as $\mathcal{S}$.

Definition. The distribution of a statistic across bootstrap samples is called a **bootstrap distribution**.

Definition. An estimator that is computed on the basis of bootstrap samples is a **bootstrap estimator**, denoted as $\hat{\eta}^*$. For example, if we want to estimate the variance of $\hat{\theta}$ ($\eta = Var(\hat{\theta})$), the bootstrap estimator is:

$$\hat{\eta}^* = Var_*(\hat{\theta}^*).$$

[6]bootstrap method 在中文很常說自力更生法、拔靴法或自助法
[7]Inference: 推論

To **estimate** parameter η of the distribution of $\hat{\theta}$ via the Bootstrap Method:

1. Consider all possible bootstrap samples drawn with replacement[8] from the given sample $\mathcal{S}$ as well as statistics $\hat{\theta}^*$ computed from them.
2. Derive the bootstrap distribution of $\hat{\theta}^*$.
3. Compute the parameter of this bootstrap distribution that has the same meaning as η.

Example. (Variance of a Sample Median) Suppose that we observed a small sample

$$\mathcal{S} = \{2, 5, 7\}$$

and estimated the population median M with the sample median $\hat{M} = 5$. How can we estimate its variance $Var(\hat{M})$?

Sol. Let's consider all possible bootstrap samples:

i	$\mathcal{B}_i$	$\hat{M}_i$
1	(2, 2, 2)	2
2	(2, 2, 5)	2
3	(2, 2, 7)	2
4	(2, 5, 2)	2
5	(2, 5, 5)	5
6	(2, 5, 7)	5
7	(2, 7, 2)	2
8	(2, 7, 5)	5
9	(2, 7, 7)	7

i	$\mathcal{B}_i$	$\hat{M}_i$
10	(5, 2, 2)	2
11	(5, 2, 5)	5
12	(5, 2, 7)	5
13	(5, 5, 2)	5
14	(5, 5, 5)	5
15	(5, 5, 7)	5
16	(5, 7, 2)	5
17	(5, 7, 5)	5
18	(5, 7, 7)	7

i	$\mathcal{B}_i$	$\hat{M}_i$
19	(7, 2, 2)	2
20	(7, 2, 5)	5
21	(7, 2, 7)	7
22	(7, 5, 2)	5
23	(7, 5, 5)	5
24	(7, 5, 7)	7
25	(7, 7, 2)	7
26	(7, 7, 5)	7
27	(7, 7, 7)	7

Since there are 7 samples for $\hat{M}_i^* = 2$, 13 samples for $\hat{M}_i^* = 5$, and 7 samples for $\hat{M}_i^* = 7$, the bootstrap distribution of the sample median is:

$$f(2) = \frac{7}{27}, \quad f(5) = \frac{13}{27}, \quad f(7) = \frac{7}{27}.$$

We use it to estimate $Var(\hat{M})$ with the bootstrap estimator:

$$\begin{aligned}
\hat{Var}(\hat{M}) &= \left(\sum x^2 f(x)\right) - \left(\sum x f(x)\right)^2 \\
&= \left(2^2 \cdot \frac{7}{27} + 5^2 \cdot \frac{13}{27} + 7^2 \cdot \frac{7}{27}\right) - \left(2 \cdot \frac{7}{27} + 5 \cdot \frac{13}{27} + 7 \cdot \frac{7}{27}\right)^2 \\
&= 3.303
\end{aligned}$$

[8] drawn with replacement: 取後放回

6.5 Methods of Point Estimation

6.5.1 Method of Moments

The **method of moments**[9] is a technique for constructing estimators of target parameters by equating sample moments to population moments.

Definition. (Population Moments) Let $X_1, X_2, \ldots, X_n$ be a random sample from the probability distribution $f(x)$ where $f(x)$ can be a discrete probability mass function or a continuous probability density function. The kth population moment[10] (or distribution moment) is $E(X^k)$, $k = 1, 2, \ldots$. The corresponding kth sample moment is

$$\frac{1}{n}\sum_{i=1}^{n} X_i^k, k = 1, 2, \ldots.$$

Example. We know that the first moment is $E(X) = \mu$, the second moment is $E(X^2)$, so that we obtain $Var(X) = E(X^2) - \mu^2$, where

$$\begin{cases} E(X) \text{ estimated by } \bar{X} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i \\ E(X^2) \text{ estimated by } \bar{X}^2 = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i^2 \end{cases}$$

Definition. (Sample Moments) Let $X_1, X_2, \ldots, X_n$ be a random sample of size n. The k-th sample moment[11], denoted as M_k (or $\overline{X^k}$), is defined as:

$$M_k = \overline{X^k} = \frac{1}{n}\sum_{i=1}^{n} X_i^k, \quad k = 1, 2, \ldots$$

- The 1st sample moment is the sample mean: $M_1 = \overline{X} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i$.
- The 2nd sample moment is the average of squares: $M_2 = \overline{X^2} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i^2$.

Example. The sample mean of $X_1, X_2, \ldots, X_n$ is $\bar{X} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i$, then $E(\overline{X}) = \mu = E(X)$.

[9] Method of Moments: 動差法
[10] Population Moments: 母體動差
[11] Sample moment: 樣本動差

To estimate m unknown population parameters $\theta_1, \theta_2, \ldots, \theta_m$:

1. Express the first m population moments $\mu_1', \mu_2', \ldots, \mu_m'$ as functions of the parameters $\theta_1, \ldots, \theta_m$.

2. Equate the population moments to their corresponding sample moments:

$$\mu_k' = M_k \implies E(X^k) = \frac{1}{n}\sum_{i=1}^{n} X_i^k, \quad \text{for } k = 1, 2, \ldots, m.$$

3. Solve the resulting system of equations for $\theta_1, \ldots, \theta_m$. The solutions are the moment estimators, denoted as $\hat{\theta}_1, \ldots, \hat{\theta}_m$.

Example. Suppose X follows a Binomial distribution, $X \sim \text{Binomial}(n_0, p)$, where n_0 is known ($n_0 = 70$) and p is the unknown parameter to be estimated ($\theta = p$). Given a sample data of size 5:

$$X_1 = 18, \quad X_2 = 19, \quad X_3 = 15, \quad X_4 = 19, \quad X_5 = 17$$

Find the method of moments estimator for p and compute its estimate.

Sol. Since there is only 1 unknown parameter (p), we only need the 1st population moment:

$$E(X) = n_0 \cdot p = 70p.$$

By equating the 1st population moment to the 1st sample moment ($E(X) = \overline{X}$), we get:

$$70p = \overline{X} \implies \hat{p} = \frac{\overline{X}}{70}.$$

Now, using the provided sample data, calculate the sample mean $\overline{X}$:

$$\overline{X} = \frac{18 + 19 + 15 + 19 + 17}{5} = \frac{88}{5} = 17.6.$$

Substitute $\overline{X}$ into the estimator formula to get the point estimate:

$$\hat{p} = \frac{17.6}{70} \approx 0.2514.$$

Example. Let $X_1, X_2, \ldots, X_n$ be a random sample drawn from a Gamma distribution with parameters α and β. The probability density function is given by:

$$f(x) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}, \quad x > 0.$$

Derive the method of moments estimators for α and β.

Sol. There are 2 unknown parameters (α, β), so we need the first two population moments. For a Gamma distribution, we know:

$$E(X) = \alpha\beta, \quad Var(X) = \alpha\beta^2.$$

Therefore, the 2nd population moment is:

$$E(X^2) = Var(X) + [E(X)]^2 = \alpha\beta^2 + (\alpha\beta)^2 = \alpha\beta^2(1+\alpha).$$

Set the population moments equal to the sample moments ($M_1 = \overline{X}$ and $M_2 = \overline{X^2}$):

$$\alpha\beta = \overline{X} \tag{1}$$
$$\alpha\beta^2 + (\alpha\beta)^2 = \overline{X^2} \tag{2}$$

From Equation (1), we can express β as $\beta = \frac{\overline{X}}{\alpha}$. Substitute this into Equation (2):

$$\alpha\left(\frac{\overline{X}}{\alpha}\right)^2 + (\overline{X})^2 = \overline{X^2} \implies \frac{\overline{X}^2}{\alpha} + \overline{X}^2 = \overline{X^2}.$$

Rearranging the terms to solve for α:

$$\frac{\overline{X}^2}{\alpha} = \overline{X^2} - \overline{X}^2 \implies \hat{\alpha} = \frac{\overline{X}^2}{\overline{X^2} - \overline{X}^2}.$$

Note that the denominator $\overline{X^2} - \overline{X}^2$ is the biased sample variance (often denoted as S_n^2).

Finally, substitute $\hat{\alpha}$ back into the expression for β:

$$\hat{\beta} = \frac{\overline{X}}{\hat{\alpha}} = \frac{\overline{X}\left(\overline{X^2} - \overline{X}^2\right)}{\overline{X}^2} = \frac{\overline{X^2} - \overline{X}^2}{\overline{X}}.$$

Thus, the method of moments estimators are:

$$\hat{\alpha} = \frac{\overline{X}^2}{\overline{X^2} - \overline{X}^2} = \frac{\overline{X}}{\hat{\beta}}, \quad \hat{\beta} = \frac{\overline{X^2} - \overline{X}^2}{\overline{X}}.$$

6.5.2 Method of Maximum Likelihood

Definition. Suppose that X is a random variable with probability distribution $f(x;\, \theta)$, where θ is a single unknown parameter. Let $x_1,\, x_2,\, \ldots,\, x_n$ be the observed values in a random sample of size n. Then the **likelihood function**[12] of the sample is

$$L(\theta) = f(x_1;\, \theta) \cdot f(x_2;\, \theta) \cdot \cdots \cdot f(x_2;\, \theta) = L(x_1,\, x_2,\, \ldots,\, x_n;\, \theta).$$

Note that the likelihood function is now a function of only the unknown parameter θ. The **maximum likelihood estimator** (MLE)[13] of θ is

$$\max_{\theta} L(\theta).$$

[12]Likelihood function: 可能性函數

[13]Maximum likelihood estimator: 最大可能性估測

Remark. 有關於 MLE：

(1) 你從定義可以看到，MLE 限定在所以變數 X_i 互相獨立。一旦變數沒有獨立，則 joint pdf 需要直接在題目表示，無法利用計算求出答案。

(2) MLE 的方法流程：

(i) Define likelihood function

$$L(x_1, x_2, \ldots, x_n; \theta) \text{ or } L(\theta).$$

(ii) Find θ such that $L(\theta)$ is maximum. i.e.

$$\hat{\theta} = \max_{\theta} L(\theta)$$

Example. Let $X_1, X_2, \ldots, X_n$ be random variances and X be normal random variable. Define $X \sim \mathcal{N}(\mu, \sigma^2)$ with parameter μ and σ^2

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

(a) Find $L(x_1, x_2, \ldots, x_n; \theta)$.

(b) Find the estimator $\hat{\theta} = \max_{\theta} L(\theta)$

Sol. Let $\theta = \begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix}$.

(a) Suppose $X_1, X_2, \ldots, X_n$ are identically independent distributed random variables, then we have

$$f(x_1, x_2, \ldots, x_n) = f(x_1)f(x_2)\cdots f(x_n).$$

Since for $i = 1, 2, \ldots, n$,

$$f(x_i) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

Compute

$$\begin{aligned} \ln\left(f(x_1)f(x_2)\cdots f(x_n)\right) &= \ln f(x_1) + \ln f(x_2) + \cdots + \ln f(x_n) \\ &= \sum_{i=1}^{n} \ln f(x_i) = L(\theta). \end{aligned}$$

Hence

$$L(\theta) = \sum_{i=1}^{n} \ln f(x_i),$$

it is called **log likelihood function**.

(b) We want to find $\hat{\theta} = \max_{\theta} L(\theta)$, that is

$$\frac{\partial L}{\partial \mu} = 0 \text{ and } \frac{\partial L}{\partial \sigma} = \frac{\partial L}{\partial \sigma^2} = 0.$$

Compute

$$\begin{aligned} L(\theta) &= \sum_{i=1}^{n} \ln \left(\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \right) \\ &= \sum_{i=1}^{n} \left(\ln \frac{1}{\sqrt{2\pi}\sigma} + \ln e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \right) \\ &= \sum_{i=1}^{n} \left(\ln 1 - \ln \sqrt{2\pi}\sigma - \frac{(x_i-\mu)^2}{2\sigma^2} \right) \\ &= -n \ln \sqrt{2\pi}\sigma - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i-\mu)^2 \\ &= -n(\ln \sqrt{2\pi} + \ln \sigma) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i-\mu)^2 \end{aligned}$$

Thus we have

$$\frac{\partial L}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu) = 0$$

and

$$\frac{\partial L}{\partial \sigma} = -\frac{n}{\sigma} + \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{\sigma^3} = 0.$$

Therefore,

$$\hat{\mu} = \left(\sum_{i=1}^{n} X_i \right) / n = \overline{X} \text{ and } \hat{\sigma}^2 = \left(\sum_{i=1}^{n} (X_i - \hat{\mu})^2 \right) / n.$$

Example. It is known that a sample consisting of the values 12, 11.2, 13.5, 12.3, 13.8, and 11.9 comes from a population with pdf

$$f(x;\, \theta) = \begin{cases} \dfrac{\theta}{x^{\theta+1}} & , x > 1 \\ 0 & , \text{otherwise} \end{cases}$$

where $\theta > 0$. Find the maximum likelihood estimate for θ.

Sol. Exercise. But you can follow the step:

1. Find log likelihood function
2. Find MLE for θ, that is $\hat{\theta} =$?.
3. Substitute data, find maximum likelihood estimate.

Theorem. Let f_i be a differentiable functions on an open interval for $i = 1, 2, \ldots, n$. Then the derivative of $\prod_{k=1}^{n} f_k(x)$ is

$$\left(\prod_{k=1}^{n} f_k(x)\right)' = \left(\prod_{k=1}^{n} f_k(x)\right) \cdot \left(\sum_{k=1}^{n} \frac{f_k'(x)}{f_k(x)}\right).$$

Proof. It is easy to solve $(fg)' = f'g + fg'$ by product rule. However, it is difficult to solve when more than two function. We know that

$$\ln\left(\prod_{k=1}^{n} f_k(x)\right) = \sum_{k=1}^{n} \ln f_k(x).$$

Differentiate both side, we obtain

$$\left[\ln\left(\prod_{k=1}^{n} f_k(x)\right)\right]' = \left(\sum_{k=1}^{n} \ln f_k(x)\right)' \iff \frac{1}{\prod_{k=1}^{n} f_k(x)} \left(\prod_{k=1}^{n} f_k(x)\right)' = \sum_{k=1}^{n} \frac{f_k'(x)}{f_k(x)}$$

$$\iff \left(\prod_{k=1}^{n} f_k(x)\right)' = \left(\prod_{k=1}^{n} f_k(x)\right) \cdot \left(\sum_{k=1}^{n} \frac{f_k'(x)}{f_k(x)}\right)$$

Hence

$$\left(\prod_{k=1}^{n} f_k(x)\right)' = \left(\prod_{k=1}^{n} f_k(x)\right) \cdot \left(\sum_{k=1}^{n} \frac{f_k'(x)}{f_k(x)}\right).$$

□

Moreover,

$$\begin{aligned}
\left(\prod_{k=1}^{n} f_k(x)\right)' &= \left(\prod_{k=1}^{n} f_k(x)\right) \cdot \left(\sum_{k=1}^{n} \frac{f_k'(x)}{f_k(x)}\right) \\
&= f_1 f_2 \cdots f_k \cdots f_n \cdot \left(\sum_{k=1}^{n} \frac{f_k'(x)}{f_k(x)}\right) \\
&= \sum_{k=1}^{n} \frac{f_k'(x)}{f_k(x)} f_1 f_2 \cdots f_k \cdots f_n \\
&= \sum_{k=1}^{n} f_k'(x) \prod_{j=1,\, j\neq k}^{n} f_j.
\end{aligned}$$

6.5.3 Bayesian Estimation of Parameters

Recall this:

Theorem. (Total Probability or Rule of Elimination) If the events $B_1, B_2, \ldots, B_k$ constitute a **partition** of the sample space such that $P(B_i) \neq 0$ for $i = 1, 2, \ldots, k$, then for any event A of S,

$$P(A) = \sum_{i=1}^{k} P(B_i \cap A) = \sum_{i=1}^{k} P(B_i) \cdot P(A \mid B_i).$$

Proof.

$$\begin{aligned} P(B_i) \cdot P(A \mid B_i) &= \frac{P(A \cap B_i)}{P(B_i)} \cdot P(B_i) \\ &= P(A \cap B_i) \end{aligned}$$

□

Theorem. (Bayes' Theorem) If $B_1, B_2, \ldots, B_k$ constitute a partition of S such that $P(B_i) \neq 0$ for $i = 1, 2, \ldots, k$, then for event A in S such that $P(A) \neq 0$,

$$P(B_r \mid A) = \frac{P(B_r \cap A)}{\sum_{i=1}^{k} P(B_i \cap A)} = \frac{P(B_r) \cdot P(A \mid B_r)}{\sum_{i=1}^{k} P(B_i) \cdot P(A \mid B_i)}.$$

it is called **posterior probability**[14].

$$\frac{P(B_r) \cdot P(A \mid B_r)}{\sum_{i=1}^{k} P(B_i) \cdot P(A \mid B_i)}$$

is called **prior probability**[15].

Bayesian Estimation[16]

Problem: Find a point estimator of the parameter θ for the population with distribution

$$f(x \mid \theta).$$

Denote $f(\theta)$ as the prior distribution of θ. Suppose that a random sample size $n : (X_1, X_2, \ldots, X_n) = X$ is observed. Then the distribution of θ given X is

$$f(\theta \mid x) \quad \text{(posterior distribution)}$$

is

$$f(\theta \mid x) = \frac{f(x \mid \theta) \cdot f(\theta)}{f(x)}.$$

[14]Posterior probability: 後驗機率
[15]Prior probability: 先驗機率
[16]Bayesian Estimation: 貝氏估測

Remark. To find $f(x)$:

1. If θ is discrete,
$$f(x) = \sum_{\text{all } \theta} f(x \mid \theta) f(\theta) = \sum_{\text{all } \theta} f(x, \theta) = f_X(x).$$
2. If θ is continuous,
$$f(x) = \int_{-\infty}^{\infty} f(x \mid \theta) f(\theta)\, d\theta \iff \int_{-\infty}^{\infty} f(x, \theta)\, d\theta = f_X(x).$$

Remark. In summary, given prior information $f(x \mid \theta)$ and $f(\theta)$, find $f(x)$[17], you will find $f(\theta \mid x)$.

Remark. $\hat{\theta}$ is called Bayesian estimator for θ, there are three methods:

(1) Posterior Mean:

(a) Continuous: $\hat{\theta} = \int_I \theta f(\theta \mid x)\, d\theta = E(\theta)$, for all $\theta \in I$.

(b) Discrete: $\hat{\theta} = \sum_{\text{all } \theta} \theta f(\theta \mid x) = E(\theta)$

(2) Posterior Median:
$$F(\hat{\theta}) = 1 - F(\hat{\theta}) = \frac{1}{2}$$

(3) Posterior Mode:
$$\hat{\theta} = \max_{\theta} f(\theta \mid x) \iff f'(\theta \mid x) = 0$$

Example. Assume that the prior distribution for the proportion of "defections" produced by a machine is denoted by $f(p)$, where p is the parameter of a binomial distribution. Let a RV X be the number of defectives among a random sample of size 2. The prior distribution $f(p)$ is defined as
$$f(p) = \begin{cases} 0.6 & \text{if } p = 0.1 \\ 0.4 & \text{if } p = 0.2 \end{cases}$$

RV X: binomial distribution with parameter p
$$f(x \mid p) = \binom{2}{x} p^x (1-p)^{2-x}, x = 0,\ 1,\ 2.$$

Thus
$$f(p \mid x) = \frac{f(x \mid p) f(p)}{f(x)}.$$

Find $f(x) = \sum_{p=0.1}^{0.2} f(x \mid p) f(p)$ and find the posterior distribution of p given that x is observed $f(p \mid x)$.

[17] marginal pdf

Sol.

$$\begin{aligned} f(x) &= \sum_{p=0.1}^{0.2} f(x \mid p) f(p) \\ &= f(x \mid 0.1) \cdot f(0.1) + f(x \mid 0.2) \cdot f(0.2) \\ &= \binom{2}{x} (0.1)^x (0.9)^{2-x} \cdot 0.6 + \binom{2}{x} (0.2)^x (0.8)^{2-x} \cdot 0.4,\ x = 0,\ 1,\ 2 \end{aligned}$$

It is shown as the table:

x	0	1	2
$f(x)$	0.742	0.236	0.022

For $p = 0.1$,

$$\begin{aligned} f(0.1 \mid x) &= \frac{f(x \mid 0.1) f(0.1)}{f(x)} \\ &= \frac{\binom{2}{x} (0.1)^x (0.9)^{2-x} \cdot 0.6}{f(x)},\ x = 0,\ 1,\ 2. \end{aligned}$$

For $p = 0.2$,

$$f(0.1 \mid x) = 1 - f(0.1 \mid x),\ x = 0,\ 1,\ 2.$$

Substitute, we have

p	$x = 0$	$x = 1$	$x = 2$
$f(0.1 \mid x)$	0.6550	0.4576	0.2727
$f(0.2 \mid x)$	0.3450	0.5424	0.7273

Example. Assume that the prior distribution for the proportion of defections produced by a machine is denoted by $f(p)$, where p is the parameter of a **binomial distribution**. Let a RV X be the number of defectives among a random sample of size 2. The prior distribution $f(p)$ is defined as **uniformly distributed** over (0, 1):

$$f(p) = \begin{cases} 1 & ,\ 0 < p < 1 \\ 0 & ,\ \text{otherwise} \end{cases}$$

Find the posterior distribution of p given that x is observed $f(p \mid x)$.

Sol. Since X is binomial,

$$f(x \mid p) = \binom{2}{x} p^x (1-p)^{2-x},\ x = 0,\ 1,\ 2$$

and given

$$f(p) = \begin{cases} 1 & ,\ 0 < p < 1 \\ 0 & ,\ \text{otherwise} \end{cases}$$

Compute the marginal $f(x)$,

$$\begin{aligned} f(x) &= \int_0^1 f(x \mid p) f(p)\, dp \\ &= \int_0^1 \binom{2}{x} p^x (1-p)^{2-x} \cdot 1\, dp \\ &= \binom{2}{x} \int_0^1 p^x (1-p)^{2-x}\, dp,\ x = 0,\ 1,\ 2, \quad 0 < p < 1. \end{aligned}$$

If $x = 0$,

$$\begin{aligned} f(0) &= \binom{2}{0} \int_0^1 p^0 (1-p)^{2-0}\, dp \\ &= \int_0^1 1 - 2p + p^2\, dp \\ &= p - p^2 + \frac{p^3}{3}\bigg|_0^1 \\ &= \frac{1}{3}. \end{aligned}$$

If $x = 1$,

$$\begin{aligned} f(1) &= \binom{2}{1} \int_0^1 p^1 (1-p)^{2-1}\, dp \\ &= 2 \int_0^1 p - p^2\, dp \\ &= 2 \left[\frac{p^2}{2} - \frac{p^3}{3} \right]_0^1 \\ &= 2 \cdot \frac{1}{6} = \frac{1}{3}. \end{aligned}$$

If $x = 2$,

$$\begin{aligned} f(2) &= \binom{2}{2} \int_0^1 p^2 (1-p)^{2-2}\, dp \\ &= \int_0^1 p^2\, dp = \frac{1}{3}. \end{aligned}$$

Therefore,

$$\begin{aligned} f(p \mid x) &= \frac{f(x \mid p) f(p)}{f(x)} \\ &= \frac{\binom{2}{x} p^x (1-p)^{2-x}}{1/3} \\ &= 3\binom{2}{x} p^x (1-p)^{2-x},\ x = 0,\ 1,\ 2, \quad 0 < p < 1. \end{aligned}$$

Remark. If we want to find $\hat{p}$, there are three methods:

(1) Posterior Mean

(2) Posterior Median

(3) Posterior Mode

Remark. The error loss is $\min_{\hat{\theta}} \left|\theta - \hat{\theta}\right|$, then the **squared error loss** is $\min_{\hat{\theta}} \left(\theta - \hat{\theta}\right)^2$. We call it is loss function.

Theorem. In the following:

(1) If loss function is squared error loss, select posterior mean.

(2) If loss function is $\left|\theta - \hat{\theta}\right|$, select posterior median.

(3) If $\hat{\theta}$ is selected as posterior mode, your result will be the same as maximum likelihood estimator (MLE).