MATH 4752/6752 – Mathematical Statistics II

Point Estimation

Xiaojing Ye
Department of Mathematics & Statistics
Georgia State University

Point estimation is to use the value of a sample statistic to estimate the value of a population parameter.

This sample statistic is called a **point estimator** and its value is called a **point estimate**.

Example. Let $X_1, \dots, X_n$ be a random sample of Bernoulli(p), and we use the sample mean $\bar{X}$ to estimate p. Then $\bar{X}$ is called a point estimator, and $\bar{x}$ is called a point estimate.

Note that a point estimator is a statistic (hence a random variable), thus it has probability distribution. We want to design "good estimators" that have highest accuracy, lowest risk, etc.

Unbiased estimator

Definition. Let $f(\cdot;\theta)$ be a distribution with parameter θ. Then a statistic $\widehat{\Theta}$ is called an **unbiased estimator** of θ if $\mathbb{E}[\widehat{\Theta}] = \theta$ for all possible values of θ.

Example. Let $X_1, \dots, X_n$ be a random sample of Bernoulli(p). Show that $\bar{X}$ is an unbiased estimator of p.

Solution. We notice that

$$\mathbb{E}[\bar{X}] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[X_i] = \frac{1}{n} \cdot np = p.$$

Example. Let $X_1, \dots, X_n \sim f(\cdot; \theta)$ be a random sample where

$$f(x;\theta) = \begin{cases} e^{-(x-\theta)}, & \text{if } x \geq \theta \\ 0, & \text{elsewhere.} \end{cases}$$

Show that $\bar{X}$ is a biased estimator of θ.

Solution. We notice that

$$\mathbb{E}[\bar{X}] = \int_\theta^\infty x e^{-(x-\theta)}\, dx = 1 + \theta \neq \theta.$$

Hence $\bar{X}$ is a biased estimator of θ.

Definition. Suppose $\widehat{\Theta}$ is a point estimator of the parameter θ of $f(\cdot;\theta)$ based on a random sample of size n, then

$$b_n(\widehat{\Theta}) = \mathbb{E}[\widehat{\Theta}] - \theta$$

is called the **bias** of $\widehat{\Theta}$. If

$$\lim_{n\to\infty} b_n(\widehat{\Theta}) = 0,$$

then we call $\widehat{\Theta}$ an **asymptotically unbiased estimator** of θ.

Example. Let $X_1, \ldots, X_n \sim$ Uniform$(0, \beta)$ be a random sample where β is a parameter. Show that Y_n (the n-th order statistic) is a biased estimator but an asymptotically unbiased estimator of β.

Solution. Notice that the pdf f of Uniform$(0, \beta)$ is

$$f(x; \beta) = \begin{cases} \frac{1}{\beta}, & \text{if } 0 < x < \beta, \\ 0, & \text{elsewhere.} \end{cases}$$

Hence the order statistic Y_n has pdf:

$$g_n(y_n) = n \cdot f(y_n) \cdot \Big(\int_{-\infty}^{y_n} f(x)\, dx \Big)^{n-1} = n \cdot \frac{1}{\beta} \cdot \Big(\frac{y_n}{\beta}\Big)^{n-1} = \frac{n y_n^{n-1}}{\beta^n}.$$

Solution (cont). Therefore

$$\mathbb{E}[Y_n] = \int_0^\beta y_n g_n(y_n)\, dy_n = \frac{n}{\beta^n} \int_0^\beta y_n^n \, dy_n = \frac{n}{n+1}\beta \neq \beta.$$

Hence Y_n is a biased estimator of β. (This result also shows that $\frac{n+1}{n}Y_n$ is an unbiased estimator of β.)

We further obtain

$$b_n(Y_n) = \mathbb{E}[Y_n] - \beta = \frac{n}{n+1}\beta - \beta = -\frac{\beta}{n+1} \to 0$$

as $n \to \infty$. Hence Y_n is an asymptotically unbiased estimator of β.

Remark.

- There may exist more than one unbiased estimator of θ. For example, $2\bar{X}$ and $\frac{n+1}{n}Y_n$ are both unbiased estimators of β in Uniform(0,β).

- Even if $\widehat{\Theta}$ is an unbiased estimator of θ, $w(\widehat{\Theta})$ may not be an unbiased estimator of $w(\theta)$ in general. For example, S^2 is an unbiased estimator of σ^2, but S may not be an unbiased estimator of σ.

Efficiency

Definition. We call $\widehat{\Theta}$ a **minimum variance unbiased estimator** (MVUE) of θ if $\widehat{\Theta}$ has the smallest variance among all unbiased estimators.

Cramér-Rao inequality. If $\widehat{\Theta}$ is an unbiased estimator of θ of $f(\cdot;\theta)$ for a random sample of size n, then

$$\text{var}[\widehat{\Theta}] \geq \frac{1}{nI(\theta)},$$

where $I(\theta)$ is the **Fisher information** defined by

$$I(\theta) = \mathbb{E}_{X\sim f(\cdot;\theta)}\Big[\Big(\frac{\partial \ln f(X;\theta)}{\partial\theta}\Big)^2\Big] = \int_{-\infty}^{\infty}\Big(\frac{\partial \ln f(x;\theta)}{\partial\theta}\Big)^2 f(x;\theta)\,dx.$$

Cramér-Rao inequality immediately implies the following result:

Theorem. If $\hat{\Theta}$ is an unbiased estimator of θ and

$$\text{var}[\hat{\Theta}] = \frac{1}{nI(\theta)},$$

then $\hat{\Theta}$ is an MVUE of θ.

Example. Let $X_1, \ldots, X_n$ be a random sample of $N(\mu, \sigma^2)$ where μ is to be estimated. Then $\bar{X}$ is an MVUE of μ.

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

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

Therefore

$$I(\mu) = \mathbb{E}_{X \sim f(\cdot;\mu)} \Big[\Big(\frac{\partial \ln f(X; \mu)}{\partial \mu} \Big)^2 \Big] = \frac{1}{\sigma^4} \mathbb{E}[(X - \mu)^2] = \frac{1}{\sigma^2}.$$

On the other hand, we have

$$\mathrm{var}[\bar{X}] = \frac{\sigma^2}{n} = \frac{1}{nI(\mu)}.$$

Therefore $\bar{X}$ is an MVUE of μ.

Remark. $\bar{X}$ may not be an MVUE of μ for other distributions.

Definition. The **efficiency** of an unbiased estimator $\widehat{\Theta}$ of θ based on a random sample of size n is defined by

$$e(\widehat{\Theta}) = \frac{1}{nI(\theta)\,\mathrm{var}[\widehat{\Theta}]}$$

Obviously $e(\widehat{\Theta}) \le 1$ due to the Cramér-Rao inequality.

Suppose $\widehat{\Theta}_1$ and $\widehat{\Theta}_2$ are two unbiased estimators of θ based on a random sample of size n, then

$$\frac{e(\widehat{\Theta}_2)}{e(\widehat{\Theta}_1)} = \frac{\mathrm{var}[\widehat{\Theta}_1]}{\mathrm{var}[\widehat{\Theta}_2]}$$

is called the **efficiency of $\widehat{\Theta}_2$ relative to $\widehat{\Theta}_1$.**

Example. Let $X_1, \ldots, X_n \sim$ Uniform$(0, \beta)$ be a random sample where β is a parameter. Show that both $2\bar{X}$ and $\frac{n+1}{n} Y_n$ are unbiased estimators of β. Compare their efficiency.

Solution. We have

$$\mathbb{E}[\bar{X}] = \mathbb{E}[X_i] = \frac{\beta}{2}.$$

Hence $\mathbb{E}[2\bar{X}] = \beta$ which implies that $2\bar{X}$ is an unbiased estimator of β.

We have showed that $\mathbb{E}[Y_n] = \frac{n}{n+1} \cdot \beta$. Hence $\frac{n+1}{n} Y_n$ is also an unbiased estimator of β.

Solution (cont). To compare their efficiency, we first notice that

$$\text{var}[2\bar{X}] = 4\,\text{var}[\bar{X}] = \frac{4}{n}\text{var}[X_i] = \frac{4}{n} \cdot \frac{\beta^2}{12} = \frac{\beta^2}{3n}.$$

On the other hand, we have

$$E[Y_n^2] = \frac{n}{\beta^n} \cdot \int_0^\beta y_n^{n+1} dy_n = \frac{n}{n+2} \cdot \beta^2$$

Therefore

$$\text{var}[Y_n] = \frac{n}{n+2} \cdot \beta^2 - \left(\frac{n}{n+1} \cdot \beta\right)^2$$

Thus we have

$$\text{var}\left[\frac{n+1}{n} Y_n\right] = \left(\frac{n+1}{n}\right)^2 \text{var}[Y_n] = \frac{\beta^2}{n(n+2)}.$$

Hence the efficiency ratio of $2\bar{X}$ against $\frac{n+1}{n} Y_n$ is

$$\frac{e(2\bar{X})}{e(\frac{n+1}{n} Y_n)} = \frac{\text{var}\left[\frac{n+1}{n} \cdot Y_n\right]}{\text{var}[2\bar{X}]} = \frac{\frac{\beta^2}{n(n+2)}}{\frac{\beta^2}{3n}} = \frac{3}{n+2}.$$

Definition. Let $\widehat{\Theta}$ be a point estimator (not necessarily unbiased) of θ. Then the **mean square error** (MSE) of $\widehat{\Theta}$ is defined as

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

Notice that there is

$$\begin{aligned} \text{MSE}(\widehat{\Theta}) &= \mathbb{E}[(\widehat{\Theta} - \theta)^2] \\ &= \mathbb{E}[(\widehat{\Theta} - \mathbb{E}[\widehat{\Theta}] + \mathbb{E}[\widehat{\Theta}] - \theta)^2] \\ &= \mathbb{E}[(\widehat{\Theta} - \mathbb{E}[\widehat{\Theta}])^2] + (\mathbb{E}[\widehat{\Theta}] - \theta)^2 \\ &= \text{var}[\widehat{\Theta}] + b_n(\widehat{\Theta})^2 \end{aligned}$$

That is, MSE($\widehat{\Theta}$) is the sum of the variance of $\widehat{\Theta}$ and the square of its bias.

Example. Compare the MSE of S^2 and $\frac{n-1}{n}S^2$ for a normal population $N(\mu, \sigma^2)$.

Solution. First notice that $\mathbb{E}[S^2] = \sigma^2$ which means S^2 is unbiased. We recall that $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$. Hence

$$\frac{(n-1)^2}{\sigma^4}\text{var}[S^2] = \text{var}\left[\frac{(n-1)S^2}{\sigma^2}\right] = \frac{n-1}{2} \cdot 2^2 = 2(n-1).$$

This implies that $\text{MSE}(S^2) = \text{var}[S^2] = \frac{2\sigma^4}{n-1}$.

Solution (cont). On the other hand, we have $\mathbb{E}[\frac{n-1}{n}S^2] = \frac{n-1}{n}\sigma^2$ which implies that the bias is $-\frac{\sigma^2}{n}$. Furthermore,

$$\text{var}\left[\frac{n-1}{n}S^2\right] = \frac{(n-1)^2}{n^2}\text{var}[S^2] = \frac{2(n-1)\sigma^4}{n^2}.$$

Therefore we have

$$\text{MSE}\left(\frac{n-1}{n}S^2\right) = \frac{2(n-1)\sigma^4}{n^2} + \left(-\frac{\sigma^2}{n}\right)^2 = \frac{(2n-1)\sigma^4}{n^2}.$$

Now we have that

$$\frac{\text{MSE}(S^2)}{\text{MSE}(\frac{n-1}{n}S^2)} = \frac{\frac{2\sigma^4}{n-1}}{\frac{(2n-1)\sigma^4}{n^2}} = \frac{2n^2}{2n^2-3n+1} > 1.$$

Therefore $\frac{n-1}{n}S^2$ has smaller MSE than the unbiased estimator S^2 does.

Consistency

Definition. We call $\widehat{\Theta}$ a **consisent estimator** of θ based on a random sample of size n if for any $c > 0$, there is

$$\lim_{n\to\infty} \mathsf{P}(|\widehat{\Theta} - \theta| > c) = 0.$$

The following theorem provides a sufficient condition of consistency.

Theorem. If $\widehat{\Theta}$ is an unbiased estimator of θ and var$[\widehat{\Theta}] \to 0$ as $n \to \infty$, then $\widehat{\Theta}$ is a consistent estimator of θ.

Proof. By Chebyshev's inequality, we have

$$\mathsf{P}(|\widehat{\Theta} - \theta| > c) \le \frac{\mathrm{var}[\widehat{\Theta}]}{c^2} \to 0$$

as $n \to \infty$. This implies that $\widehat{\Theta}$ is consistent.

Example. Suppose S^2 is the sample variance of the random sample from a normal population $N(\mu, \sigma^2)$, then S^2 is a consistent estimator of σ^2.

Proof. We have $\mathbb{E}[S^2] = \sigma^2$ and var$[S^2] = \frac{\sigma^4}{n-1} \to 0$ as $n \to \infty$. By the previous theorem, we know S^2 is consistent.

Remark. It is not difficult to show that we can replace the requirement "unbiased" with "asymptotically unbiased" in the previous theorem since

$$\mathsf{P}(|\widehat{\Theta} - \theta| > c) \le \mathsf{P}(|\widehat{\Theta} - \mathbb{E}[\widehat{\Theta}]| + |\,\mathbb{E}[\widehat{\Theta}] - \theta| > c) \to 0$$

since $|\,\mathbb{E}[\widehat{\Theta}] - \theta| \to 0$ as $n \to \infty$.

The previous theorem only provides a sufficient condition on consistency. The following example shows that it is not a necessary condition.

Example. Suppose f is a pdf with mean μ and variance $\sigma^2 < \infty$. For any $n \in \mathbb{N}$, let $X_1, \ldots, X_n$ be a random sample from f, and $Y_n \sim$ Bernoulli($\frac{1}{n}$) be independent of the random sample. Define $\widehat{\Theta}_n = n^2 Y_n + (1 - Y_n)\bar{X}_n$ where $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ is the sample mean. Show that $\widehat{\Theta}_n$ is neither unbiased nor asymptotically unbiased, but it is a consistent estimator of μ.

Proof. Since Y_n is independent of $\bar{X}_n$, we have that

$$\mathbb{E}[\widehat{\Theta}_n] = n^2\,\mathbb{E}[Y_n] + \mathbb{E}[1 - Y_n]\,\mathbb{E}[\bar{X}_n] = n + \frac{n-1}{n}\mu.$$

Therefore the bias is $b(\widehat{\Theta}_n) = \mathbb{E}[\widehat{\Theta}_n] - \mu = n - \frac{\mu}{n} \neq 0$, and $\lim_{n\to\infty} b(\widehat{\Theta}_n) = \infty \neq 0$. Hence $\widehat{\Theta}_n$ is not unbiased nor asymptotically unbiased.

Proof (cont). For any $c > 0$, we have

$$\begin{aligned}
\mathsf{P}(|\widehat{\Theta}_n - \mu| > c) &= \mathsf{P}(|\widehat{\Theta}_n - \mu| > c \mid Y_n = 1)\,\mathsf{P}(Y_n = 1) \\
&\quad + \mathsf{P}(|\widehat{\Theta}_n - \mu| > c \mid Y_n = 0)\,\mathsf{P}(Y_n = 0) \\
&\le \mathsf{P}(Y_n = 1) + \mathsf{P}(|\widehat{\Theta}_n - \mu| > c \mid Y_n = 0)\,\mathsf{P}(Y_n = 0) \\
&\le \frac{1}{n} + \frac{\sigma^2}{nc^2} \cdot \frac{n-1}{n} \to 0
\end{aligned}$$

as $n \to \infty$. Therefore $\widehat{\Theta}_n$ is a consistent estimator of μ.

Sufficiency

Definition. We call $\widehat{\Theta}$ a **sufficient estimator** of θ if the conditional probability of the random sample $X_1, \dots, X_n$ given $\widehat{\Theta} = \widehat{\theta}$ is independent of θ, i.e., $f_{X_1,\dots,X_n|\widehat{\Theta}}(x_1, \dots, x_n|\widehat{\theta})$ is independent of θ.

Remark. $\widehat{\Theta}$ being a sufficient estimator of θ means that $X_1, \dots, X_n$ do not contain more information than $\widehat{\Theta}$ alone in terms of estimating θ. Moreover, since $X_1, \dots, X_n$ completely determines $\widehat{\Theta}$, we know from the definition that $\widehat{\Theta}$ is sufficient if

$$\begin{aligned} f_{X_1,\dots,X_n|\widehat{\Theta}}(x_1, \dots, x_n|\widehat{\theta}) &= \frac{f_{X_1,\dots,X_n,\widehat{\Theta}}(x_1, \dots, x_n, \widehat{\theta})}{f_{\widehat{\Theta}}(\widehat{\theta})} \\ &= \frac{f_{X_1,\dots,X_n}(x_1, \dots, x_n)}{f_{\widehat{\Theta}}(\widehat{\theta})} \end{aligned}$$

is independent of θ (note that both numerator and denominator depend on θ, so they need to be nicely canceled out using the relation between $x_1, \dots, x_n$ and $\widehat{\theta}$ for $\widehat{\Theta}$ to be sufficient).

Example. If $X_1, \ldots, X_n$ is a random sample of Bernoulli(p), then $\widehat{\Theta} = \bar{X}$ is a sufficient estimator of p.

Proof. We have

$$f(x_1, \ldots, x_n) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}.$$

On the other hand, we know $n\widehat{\Theta} = X_1 + \cdots + X_n \sim$ Binomial(n, p) and hence the pmf of $\widehat{\Theta}$ is

$$g(\widehat{\theta}) = \binom{n}{n\widehat{\theta}} p^{n\widehat{\theta}}(1-p)^{n-n\widehat{\theta}}.$$

Since $n\theta = \sum_{i=1}^n x_i$, we have

$$\frac{f(x_1, \ldots, x_n)}{g(\widehat{\theta})} = \frac{p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}}{\binom{n}{n\widehat{\theta}} p^{n\widehat{\theta}}(1-p)^{n-n\widehat{\theta}}} = \frac{1}{\binom{n}{\sum_{i=1}^n x_i}}$$

(notice that p is canceled out), which implies that $\widehat{\Theta}$ is a sufficient estimator.

Example. Let X_1, X_2, X_3 be a random sample of Bernoulli (θ), then $\hat{\Theta} = \frac{1}{6}(X_1 + 2X_2 + 3X_3)$ is not a sufficient estimator of θ.

Solution. We just need to show that $\frac{f(x_1,x_2,x_3)}{g(\hat{\theta})}$ depends on θ for certain value of $x_1, x_2, x_3, \hat{\theta}$. For example, consider $x_1 = x_2 = 1$, $x_3 = 0$, and $\hat{\theta} = \frac{1}{2}$, we have

$$f(1,1,0) = \theta \cdot \theta \cdot (1-\theta) = \theta^2(1-\theta),$$

and

$$g\Big(\frac{1}{2}\Big) = f(1,1,0) + f(0,0,1) = \theta^2(1-\theta) + (1-\theta)^2\theta.$$

Therefore we have

$$\frac{f(1,1,0)}{g(\frac{1}{2})} = \frac{\theta^2(1-\theta)}{\theta^2(1-\theta) + (1-\theta)^2\theta} = \theta$$

which depends on θ.

Theorem. $\widehat{\Theta}$ is a sufficient estimator of θ if and only if the joint pmf/pdf of $X_1, \dots, X_n$ can be factorized as

$$f(x_1, \dots, x_n; \theta) = \phi(\widehat{\theta}; \theta) \cdot h(x_1, \dots, x_n)$$

for some function h not involving θ.

Proof (informal). Necessity is obvious. To show sufficiency, suppose the "joint pdf" of $(X_1, \dots, X_n)$ and $\widehat{\Theta}$ satisfies

$$f(x_1, \cdots, x_n, \widehat{\theta}; \theta) = f(x_1, \cdots, x_n; \theta) = \phi(\widehat{\theta}; \theta) h(x_1, \dots, x_n),$$

then we obtain the marginal distribution of $\widehat{\Theta}$ as

$$g(\widehat{\theta}; \theta) = \int f(x_1, \cdots, x_n, \widehat{\theta}; \theta)\, dx_1 \cdots dx_n = C\phi(\widehat{\theta}; \theta)$$

for some constant C independent of θ. Therefore

$$\frac{f(x_1, \dots, x_n, \widehat{\theta}; \theta)}{g(\widehat{\theta}; \theta)} = C^{-1} h(x_1, \dots, x_n)$$

is independent of θ.

Example. Consider a normal population $N(\mu, \sigma^2)$ for known σ^2. Show that $\bar{X}$ is a sufficient estimator of μ.

Proof. Notice that

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

Hence we have

$$\begin{aligned} f(x_1,\ldots,x_n;\mu) &= (2\pi\sigma^2)^{-n/2} e^{-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}} \\ &= \underbrace{(2\pi\sigma^2)^{-n/2} e^{-\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{2\sigma^2}}}_{h(x_1,\ldots,x_n)} \underbrace{e^{-\frac{n(\bar{x}-\mu)^2}{2\sigma^2}}}_{\phi(\bar{x};\mu)}. \end{aligned}$$

Hence $\bar{X}$ is a sufficient estimator of μ.

Remark. If $\widehat{\Theta}$ is a sufficient estimator of θ, then $Y = u(\widehat{\Theta})$ is also a sufficient estimator of θ for any one-to-one correspondence u. This is because that

$$f(x_1, \ldots, x_n; \theta) = \phi(\widehat{\theta}; \theta) h(x_1, \ldots, x_n) = \underbrace{\phi(w(y); \theta)}_{=: \tilde{\phi}(y;\theta)} h(x_1, \ldots, x_n)$$

where w is the inverse of u.

Now we consider methods to construct point estimators. There are three typical methods:

- Method of moments
- Maximum likelihood estimation
- Bayesian estimation

Method of moments

Definition. Let X be a random variable. Then the k-th **moment** of X is defined by

$$\mu_k' = \mathbb{E}[X^k], \quad \text{for } k = 0, 1, 2, \dots .$$

Remarks.

- Moments are functions of the distribution parameter $\boldsymbol{\theta}$, i.e.

  $$\mu_k' = \mu_k'(\boldsymbol{\theta}).$$

  In particular, for any X, there are $\mu_0' = 1$, $\mu_1' = \mu$, $\mu_2' = \mu^2 + \sigma^2$ (if mean and variance exist).

- The k-th **central moment** of X is defined as $\mu_k = \mathbb{E}[(X - \mathbb{E}[X])^k]$.

Definition. Let $X_1, \ldots, X_n$ be a random sample of size n from a distribution f. Then the kth **sample moment** of X is defined by

$$M'_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k.$$

If we obtain values $X_i = x_i$ for all i, then we also call the value of M_k the kth sample moment:

$$m'_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k.$$

Suppose we want to estimate the parameters $\theta_1, \ldots, \theta_r$ of the distribution f, then the **method of moments** is to solve a system of r equations:

$$\mu'_k(\theta_1, \ldots, \theta_r) = m'_k, \quad \text{for } k = 1, \ldots, r.$$

for $\theta_1, \ldots, \theta_r$. This yields estimates

$$\widehat{\theta}_k = \widehat{\theta}_k(x_1, \ldots, x_n), \quad \text{for } k = 1, \ldots, r.$$

The corresponding estimators obtained by the method of moments are:

$$\widehat{\Theta}_k = \widehat{\Theta}_k(X_1, \ldots, X_n), \quad \text{for } k = 1, \ldots, r.$$

Remark. We need the specific distribution type (e.g., exponential, normal etc) to obtain the functions $\mu'_k(\theta_1, \ldots, \theta_r)$, unless the parameters we want to estimate are the moments.

Example. Given a random sample of size n from Uniform($\alpha, 1$), use the method of moments to obtain an estimator of α.

Solution. We know that

$$\mu_1' = \mu = \frac{\widehat{\alpha} + 1}{2}, \qquad m_1' = \bar{x}.$$

Equating the two and solving for α, we obtain estimate

$$\widehat{\alpha} = 2\bar{x} - 1.$$

Hence the method of moments yields the estimator for α as $2\bar{X} - 1$.

Example. Given a random sample of size n from Gamma(α, β), use the method of moments to obtain estimators of α and β.

Solution. We know that

$$\mu_1' = \mu = \alpha\beta, \qquad \mu_2' = \mu^2 + \sigma^2 = (\alpha\beta)^2 + \alpha\beta^2 = \alpha(\alpha+1)\beta^2.$$

Equating them to m_1' and m_2' respectively yields

$$\begin{aligned} \widehat{\alpha}\widehat{\beta} &= m_1' = \bar{x}, \\ \widehat{\alpha}(\widehat{\alpha}+1)\widehat{\beta}^2 &= m_2' = \frac{1}{n}\sum_{i=1}^{n} x_i^2, \end{aligned}$$

solving which for estimates $\widehat{\alpha}$ and $\widehat{\beta}$ yields

$$\widehat{\alpha} = \frac{n\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \qquad \widehat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n\bar{x}}.$$

The estimators can be obtained accordingly.

Maximum likelihood estimation

Suppose $X_1, \dots, X_n$ is a random sample from a distribution $f(\cdot;\theta)$ and we obtain values of $x_1, \dots, x_n$ of this random sample. What is the value of θ that makes these values $x_1, \dots, x_n$ most probable?

Definition. For given values $x_1, \dots, x_n$, we define the **likelihood function**

$$L(\theta) = f_{X_1,\dots,X_n}(x_1, \dots, x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta).$$

The value $\widehat{\theta}$ that maximizes $L(\theta)$ is called a **maximum likelihood estimate (MLE)** of θ.

Remark. It is equivalent to maximizing the **log-likelihood function**

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

since ln is a strictly increasing function.

Example. Given x successes in n trials, find the MLE of θ in the corresponding Binomial(n, θ).

Solution. We first have the likelihood function

$$L(\theta) = f(x; \theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}$$

The log-likelihood function is

$$\ell(\theta) = \ln \binom{n}{x} + x \ln \theta + (n-x) \ln(1-\theta).$$

To find its maximizer, we first find the critical points such that $\ell'(\theta) = 0$:

$$\ell'(\theta) = \frac{x}{\theta} - \frac{n-x}{1-\theta} = 0$$

which yields a single solution $\theta = \frac{x}{n}$ (it is easy to check that it's a maximizer of ℓ). Hence the MLE is $\widehat{\theta} = \frac{x}{n}$, and the maximum likelihood estimator is $\widehat{\Theta} = \frac{\bar{X}}{n}$.

Example. Let $X_1, \dots, X_n$ be a random sample from Exponential(θ). Find the MLE of θ.

Solution. Recall that the pdf of Exponential(θ) is $f(x;\theta) = \frac{1}{\theta}e^{-x/\theta}I_{\{x\geq 0\}}(x)$. Hence we have the likelihood function:

$$L(\theta) = \prod_{i=1}^{n} f(x_i;\theta) = \prod_{i=1}^{n} \frac{1}{\theta}e^{-x_i/\theta} = \theta^{-n}e^{-\sum_{i=1}^{n} x_i/\theta}.$$

The log-likelihood function is

$$\ell(\theta) = \ln L(\theta) = -n \ln \theta - \frac{\sum_{i=1}^{n} x_i}{\theta}.$$

Taking derivative of ℓ and equating it to 0 yield

$$\ell'(\theta) = -\frac{n}{\theta} + \frac{\sum_{i=1}^{n} x_i}{\theta^2} = 0,$$

solving which yields the MLE $\widehat{\theta} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}$. Hence the maximum likelihood estimator is $\widehat{\Theta} = \bar{X}$.

Sometimes we may need to check the boundary points if the likelihood function is not differentiable.

Example. Let $X_1, \ldots, X_n$ be a random sample of Uniform$(0, \beta)$. Find the MLE of β.

Solution. We know the likelihood function is

$$L(\beta) = \prod_{i=1}^{n} f(x_i; \beta) = \prod_{i=1}^{n} \frac{1}{\beta} I_{\{x \le \beta\}}(x_i) = \begin{cases} \beta^{-n}, & \text{if } \beta \ge \max_{1 \le i \le n} x_i \\ 0, & \text{elsewhere.} \end{cases}$$

Note that this function is strictly decrease and does not have critical point for $\beta \ge \max_{1 \le i \le n} x_i$. However the maximum is attained at $\max_{1 \le i \le n} x_i$. Hence $\widehat{\beta} = \max_{1 \le i \le n} x_i = y_n$ and the maximum likelihood estimator is Y_n (the nth order statistic).

We can also find MLE of multiple parameters simultaneously.

Example. Let $X_1, \ldots, X_n$ be a random sample of $N(\mu, \sigma^2)$ where both μ and σ^2 are unknown. Find the MLE of μ and σ^2.

Solution. We know the pdf of $N(\mu, \sigma^2)$ is $f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$. Hence the likelihood function is

$$L(\mu, \sigma^2) = \prod_{i=1}^{n} f(x_i; \mu, \sigma^2) = (2\pi\sigma^2)^{-n/2} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}.$$

The log-likelihood function is

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

Solution (cont). To find the maximizer of ℓ, we compute the partial derivatives of ℓ with respect to μ and σ^2:

$$\partial_\mu \ell(\mu, \sigma^2) = \frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \mu),$$

$$\partial_{\sigma^2} \ell(\mu, \sigma^2) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (x_i - \mu)^2.$$

Equating them to 0 and solving for μ and σ^2 jointly yield the MLE:

$$\widehat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x}, \qquad \widehat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \widehat{\mu})^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2.$$

Remark. $\widehat{\sigma} = (\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2)^{1/2}$ is the MLE of σ because of the invariance property of MLE (see next page).

Theorem. MLE has **invariance property**: if $\widehat{\Theta}$ is an MLE of θ and g is a continuous function (not necessarily one-to-one), then $g(\widehat{\Theta})$ is an MLE of $g(\theta)$.

Proof. Let the values $x_1, \dots, x_n$ of the random sample be held fixed. We first define the **induced likelihood function** L^* of $\eta = g(\theta)$ as

$$L^*(\eta) = \max_{\{\theta : g(\theta) = \eta\}} L(\theta).$$

Suppose $\widehat{\eta}$ is a maximizer of $L^*(\eta)$, then we have

$$\begin{aligned}
L^*(\widehat{\eta}) &= \max_{\eta} L^*(\eta) && (\widehat{\eta} \text{ is a maximizer}) \\
&= \max_{\eta} \max_{\{\theta : g(\theta) = \eta\}} L(\theta) && (\text{Definition of } L^*) \\
&= \max_{\theta} L(\theta) && (\text{Double max is max}) \\
&= L(\widehat{\theta}) && (\widehat{\theta} \text{ is MLE})
\end{aligned}$$

Proof (cont). On the other hand, we have

$$
\begin{aligned}
L(\widehat{\theta}) &= \max_{\{\theta : g(\theta) = g(\widehat{\theta})\}} L(\theta) && (\widehat{\theta} \text{ is MLE}) \\
&= L^*(g(\widehat{\theta})) && (\text{Definition of } L^*)
\end{aligned}
$$

Combining the two equations above yields $L^*(g(\widehat{\theta})) = L^*(\widehat{\eta})$ which is equal to $\max_\eta L^*(\eta)$ since $\widehat{\theta}$ is a maximizer of L^*. Therefore $g(\widehat{\theta})$ is an MLE of $g(\theta)$.

Bayesian estimation

Suppose we also treat the parameter θ of $f(x;\theta)$ as a random variable Θ following a **prior distribution** $p(\theta)$ based on our belief or previous experience.

We treat $f(x|\theta) = f(x;\theta)$ as the conditional probability of X given Θ.

After the experiment is done and we obtain value $X = x$, we can update the prior distribution to the **posterior distribution** $\phi(x|\theta)$. By the Bayes rule, there is

$$\phi(\theta|x) = \frac{p(\theta)f(x|\theta)}{g(x)} \propto p(\theta)f(x|\theta).$$

Here θ and X have a joint distribution, and $p(\theta)$ and $g(x)$ are their marginal distributions. Note that $g(x)$ does not involve θ.

This idea can be easily extended to the case with a random sample $X_1, \dots, X_n$:

$$\phi(\theta|x_1, \dots, x_n) = \frac{p(\theta) \prod_{i=1}^n f(x_i|\theta)}{g(x_1, \dots, x_n)} \propto p(\theta) \prod_{i=1}^n f(x_i|\theta).$$

Then Bayesian estimation is to find $\widehat{\theta}$ that maximizes this posterior distribution. This method is also called **maximum-a-posteriori (MAP)**.

Maximizing $\phi(\theta|x_1, \dots, x_n)$ is equivalent to maximizing

$$\ln \phi(\theta|x_1, \dots, x_n) = \ln p(\theta) + \underbrace{\sum_{i=1}^n \ln f(x_i|\theta)}_{\text{log-likelihood } \ell(\theta)} - \underbrace{\ln g(x_1, \dots, x_n)}_{\text{not involving } \theta}.$$

The prior $p(\theta)$ serves as a “regularization” added to the likelihood function.

Example. Let X follow Binomial(n, θ) for unknown $\theta \in (0, 1)$. Suppose the prior distribution of θ is Beta(α, β) for some given $\alpha, \beta > 0$. Find the posterior distribution and Bayesian estimate of θ.

Solution. The prior distribution of Θ is

$$p(\theta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha) + \Gamma(\beta)} \theta^{\alpha - 1}(1 - \theta)^{\beta - 1}.$$

The conditional distribution (or the likelihood function) is

$$f(x|\theta) = \binom{n}{x} \theta^x (1 - \theta)^{n - x}.$$

Hence the posterior distribution is

$$\phi(\theta|x) \propto p(\theta) f(x|\theta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \binom{n}{x} \theta^{x + \alpha - 1}(1 - \theta)^{n - x + \beta - 1}.$$

This means that Θ given $X = x$ follows Beta($x + \alpha, n - x + \beta$) distribution.

Solution (cont). To find the Bayesian estimate, we take logarithm of $\phi(\theta|x)$:

$$\ln \phi(\theta|x) = (x + \alpha - 1) \ln \theta + (n - x + \beta - 1) \ln(1 - \theta) + C$$

where C is a constant independent of θ.

Taking derivative of $\ln \phi(\theta|x)$ with respect to θ, equating to 0 and solving for θ, we obtain the Bayesian estimate:

$$\widehat{\theta} = \frac{x + \alpha - 1}{n + \alpha + \beta - 1}.$$

Remark. When we have more data, i.e., large n and x, there is $\widehat{\theta} \approx \frac{x}{n}$.

Example. Suppose $X_1, \ldots, X_n$ is a random sample of $N(\mu, \sigma^2)$ where σ^2 is known. Assume the prior distribution of μ is $N(\mu_0, \sigma_0^2)$ for some known μ_0 and σ_0^2. Find the posterior distribution of μ and the Bayesian estimate.

Solution. We know the prior distribution is

$$p(\mu) = \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}}.$$

The conditional distribution of $X_1, \ldots, X_n$ given $\Theta = \theta$ is

$$f(x_1, \ldots, x_n|\mu) = (2\pi\sigma^2)^{-n/2} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2}.$$

Solution (cont). The posterior distribution is

$$\phi(\mu|x_1,\ldots,x_n) \propto p(\mu)f(x_1,\ldots,x_n|\mu) = \underbrace{\cdots}_{\text{completing squares}} \propto e^{-\frac{(\mu-\mu_1)^2}{2\sigma_1^2}}$$

where

$$\mu_1 = \frac{n\bar{x}\sigma_0^2 + \mu_0\sigma^2}{n\sigma_0^2 + \sigma^2} \quad \text{and} \quad \frac{1}{\sigma_1^2} = \frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}.$$

This means that the posterior distribution of μ given $X_1 = x_1, \ldots, X_n = x_n$ is $N(\mu_1, \sigma_1^2)$.

Remark. When $n \to \infty$, we have $\sigma_1^2 \to 0$ and $\mu_1 \to \bar{x}$.