Example 1.17b: Suppose that $X_1, X_2, \cdots, X_n$ are mutually independently and identically distributed as Bernoulli random variables with parameter p.

We derive the maximum likelihood estimators of p.

The joint density (or the likelihood function) of $X_1, X_2, \cdots, X_n$ is:

$$f(x_1, x_2, \cdots, x_n; p) = \prod_{i=1}^{n} f(x_i; p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\sum_{i=1}^{n} x_i}(1-p)^{n-\sum_{i=1}^{n} x_i} = l(p).$$

The log-likelihood function is given by:

$$\log l(p) = (\sum_{i=1}^{n} x_i)\log(p) + (n - \sum_{i=1}^{n} x_i)\log(1-p).$$

For maximization of the likelihood function, differentiating the log-likelihood function $\log l(p)$ with respect to p , the first derivatives should be equal to zero, i.e.,

$$\frac{\mathrm{d}\log l(p)}{\mathrm{d}p} = \frac{1}{p}\sum_{i=1}^{n} x_i - \frac{1}{1-p}(n - \sum_{i=1}^{n} x_i) = \frac{n}{p}\overline{x} - \frac{n}{1-p}(1-\overline{x}) = 0$$

Let $\hat{p}$ be the solution which satisfies the above equation.

We obtain the maximum likelihood estimates as follows:

$$\hat{p} = \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

Replacing x_i by X_i for $i = 1, 2, \cdots, n$, the maximum likelihood estimator of p is given by $\hat{p} = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

● We check whether $\hat{p}$ is unbiased.

$$\mathrm{E}(\hat{p}) = \mathrm{E}(\overline{X}) = \mathrm{E}(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}(X_i) = p$$

Remember that $\mathrm{E}(X_i) = \sum_{x_i=0}^{1} x_i p^{x_i}(1-p)^{1-x_i} = p$, where x_i takes 0 or 1.

Thus, $\hat{p}$ is an unbiased estimator of p.

● Next, we check whether $\hat{p}$ is efficient.

From Cramer-Rao inequality,

$$\mathrm{V}(\hat{p}) \geq -\frac{1}{n\mathrm{E}\Big(\dfrac{\mathrm{d}^2\log f(X;p)}{\mathrm{d}p^2}\Big)}.$$

$$f(X;p) = p^X(1-p)^{1-X}$$

$$\log f(X;p) = X\log(p) + (1-X)\log(1-p)$$

$$\frac{\mathrm{d}\log f(X;p)}{\mathrm{d}p} = \frac{X}{p} - \frac{1-X}{1-p}$$

$$\frac{\mathrm{d}^2\log f(X;p)}{\mathrm{d}p^2} = -\frac{X}{p^2} - \frac{1-X}{(1-p)^2}$$

We need to check whether the equality holds.

$$\mathrm{V}(\hat{p}) = \mathrm{V}(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n^2}\mathrm{V}(\sum_{i=1}^{n} X_i) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{V}(X_i) = \frac{1}{n^2}\sum_{i=1}^{n} p(1-p) = \frac{p(1-p)}{n},$$

Note as follows:

$$\mathrm{V}(X_i) = \mathrm{E}((X_i - p)^2) = \sum_{x_i=0}^{1}(x_i - p)^2 p^{x_i}(1-p)^{1-x_i} = p(1-p).$$

The Cramer-Rao lower bound is:

$$-\frac{1}{n\mathrm{E}\Big(\dfrac{\mathrm{d}^2\log f(X;p)}{\mathrm{d}p^2}\Big)} = -\frac{1}{n\mathrm{E}\Big(-\dfrac{X}{p^2}-\dfrac{1-X}{(1-p)^2}\Big)}$$
$$= -\frac{1}{n\Big(-\dfrac{\mathrm{E}(X)}{p^2}-\dfrac{1-\mathrm{E}(X)}{(1-p)^2}\Big)} = \frac{1}{n\Big(\dfrac{1}{p}+\dfrac{1}{1-p}\Big)} = \frac{p(1-p)}{n},$$

which is equal to $\mathrm{V}(\hat{p})$.

Thus, $\hat{p}$ is an efficient estimator of p.

● We check whether $\hat{p}$ is consistent.

From Chebyshev's inequality,

$$P(|\hat{p}-p|\geq\epsilon)\leq\frac{\mathrm{E}((\hat{p}-p)^2)}{\epsilon^2}=\frac{p(1-p)}{n\epsilon^2}.$$

As $n\longrightarrow\infty$, $P(|\hat{p}-p|\geq\epsilon)\longrightarrow 0$.

That is, $\hat{p}$ converges in probability to p.

Thus, $\hat{p}$ is a consistent estimator of p.

Properties of Maximum Likelihood Estimator: For **small sample (小標本)**, the MLE has the following properties.

- MLE is not necessarily unbiased in general, but we often have the case where we can construct the unbiased estimator by an appropriate transformation.

  For instance, the MLE of σ^2, S^{**2}, is not unbiased. However, $\dfrac{n}{n-1}S^{**2}=S^2$ is an unbiased estimator of σ^2.
- If the efficient estimator exists, the maximum likelihood estimator is efficient.

Efficient estimator $\Longleftrightarrow$ The variance of the estimator is equal to the Cramer-Rao lower bound.

For **large sample (大標本)**, as $n\longrightarrow\infty$, the maximum likelihood estimator of θ, $\hat{\theta}_n$, has the following property:

$$\sqrt{n}(\hat{\theta}_n-\theta)\longrightarrow N(0,\sigma^2(\theta)), \tag{23}$$

where

$$\sigma^2(\theta)=\frac{1}{\mathrm{E}\bigg(\Big(\dfrac{\partial\log f(X;\theta)}{\partial\theta}\Big)^2\bigg)}=-\frac{1}{\mathrm{E}\bigg(\dfrac{\partial^2\log f(X;\theta)}{\partial\theta^2}\bigg)}.$$

(23) indicates that the MLE has consistency, **asymptotic unbiasedness (漸近不偏性)**, **asymptotic efficiency (漸近有効性)** and **asymptotic normality (漸近正規性)**.

Asymptotic normality of the MLE comes from the central limit theorem discussed in Section 6.3.

Even though the underlying distribution is not normal, i.e., even though $f(x;\theta)$ is not normal, the MLE is asymptotically normally distributed.

Note that the properties of $n\longrightarrow\infty$ are called the asymptotic properties, which include consistency, asymptotic normality and so on.

By normalizing, as $n\longrightarrow\infty$, we obtain as follows:

$$\frac{\sqrt{n}(\hat{\theta}_n-\theta)}{\sigma(\theta)}=\frac{\hat{\theta}_n-\theta}{\sigma(\theta)/\sqrt{n}}\longrightarrow N(0,1).$$

$\sqrt{n}(\hat{\theta}_n-\theta)$ has the distribution, which does not depend on n.

$\sqrt{n}(\hat{\theta}_n-\theta)=O(1)$ is written, where $O()$ is a function n.

That is, $\hat{\theta}_n-\theta=n^{-1/2}\times O(1)=O(n^{-1/2})$.

As another representation, when n is large, we can approximate the distribution of $\hat{\theta}_n$ as follows:

$$\hat{\theta}_n \sim N\Big(\theta, \frac{\sigma^2(\theta)}{n}\Big).$$

This implies that when $n \longrightarrow \infty$, $\hat{\theta}_n$ approaches the lower bound of Cramer-Rao inequality: $\dfrac{\sigma^2(\theta)}{n}$.
This property is called an asymptotic efficiency.

Moreover, replacing θ in variance $\sigma^2(\theta)$ by $\hat{\theta}_n$, when $n \longrightarrow \infty$, we have the following property:

$$\frac{\hat{\theta}_n - \theta}{\sigma(\hat{\theta}_n)/\sqrt{n}} \longrightarrow N(0, 1). \tag{24}$$

Practically, when n is large, we approximately use:

$$\hat{\theta}_n \sim N\Big(\theta, \frac{\sigma^2(\hat{\theta}_n)}{n}\Big). \tag{25}$$

Proof of (23): By the central limit theorem (11) on p.254,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial \log f(X_i; \theta)}{\partial \theta} \longrightarrow N\Big(0, \frac{1}{\sigma^2(\theta)}\Big), \tag{26}$$

where $\sigma^2(\theta)$ is defined in (14), i.e., $\mathrm{V}\Big(\dfrac{\partial \log f(X_i; \theta)}{\partial \theta}\Big) = \dfrac{1}{\sigma^2(\theta)}$.

Note that $\mathrm{E}\Big(\dfrac{\partial \log f(X_i; \theta)}{\partial \theta}\Big) = 0$.

Apply the central limit theorem, taking $\dfrac{\partial \log f(X_i; \theta)}{\partial \theta}$ as the ith random variable.

By the Taylor series expansion around $\hat{\theta}_n = \theta$,

$$\begin{aligned} 0 &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial \log f(X_i; \hat{\theta}_n)}{\partial \theta} \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial \log f(X_i; \theta)}{\partial \theta} + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial^2 \log f(X_i; \theta)}{\partial \theta^2}(\hat{\theta}_n - \theta) \\ &\quad + \frac{1}{2!} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial^3 \log f(X_i; \theta)}{\partial \theta^3}(\hat{\theta}_n - \theta)^2 + \cdots \end{aligned}$$