例 2： $X_1, X_n, \cdots, X_n$ は互いに独立で，それぞれパラメータ p を持ったベルヌイ分布に従うものとする。すなわち，X_i の確率関数は,

$$f(x;p) = p^x(1-p)^{1-x} \qquad\qquad x = 0, 1$$

となる。

このとき尤度関数は,

$$l(p) = \prod_{i=1}^{n} f(x_i;p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}$$

となり，対数尤度関数は,

$$\begin{aligned}\log l(p) &= \sum_{i=1}^{n} \log f(x_i;p)\\ &= \log(p)\sum_{i=1}^{n} x_i + \log(1-p)\sum_{i=1}^{n}(1-x_i)\end{aligned}$$

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

となる。

$\log l(p)$ を最大にする p を求める。

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

したがって，p について解くと，p の最尤推定値 $\hat{p}$ は,

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

となる。

さらに，x_i を X_i で置き換えて，p の最尤推定量 $\hat{p}$ は,

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

となる。

$\hat{p}$ の分布を求める。確率関数の対数を取って，

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

となる。p について微分すると，

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

となる。さらに，両辺を二乗して，期待値を取ると，

$$\mathrm{E}\Big[\Big(\frac{\mathrm{d} \log f(X_i; p)}{\mathrm{d}p}\Big)^2\Big] = \frac{\mathrm{E}[(X_i - p)^2]}{p^2(1 - p)^2}$$

となる。期待値を計算すると，

$$\begin{aligned}\mathrm{E}[(X_i - p)^2] &= \sum_{x_i=0}^{1} (x_i - p)^2 f(x_i; p) = \sum_{x_i=0}^{1} (x_i - p)^2 p^{x_i} (1 - p)^{1-x_i} \\ &= p^2(1 - p) + (1 - p)^2 p = p(1 - p)\end{aligned}$$

となるので，

$$\sigma_\theta^2 = \frac{1}{\sum_{i=1}^n \mathrm{E}\Big[\Big(\frac{\mathrm{d}\log f(X_i;p)}{\mathrm{d}p}\Big)^2\Big]} = \frac{p(1-p)}{n}$$

が得られる。したがって，$\hat{p}$ の分布は，n が大きいとき，

$$\hat{p} \sim N(p,\ \frac{p(1-p)}{n})$$

となる。

例 3： $X_1, X_n, \cdots, X_n$ は互いに独立で，それぞれパラメータ λ を持ったポアソン分布に従うものとする。すなわち，X_i の確率関数は,

$$f(x;\lambda) = \frac{\lambda^x e^{-\lambda}}{x!} \qquad\qquad x = 0, 1, 2, \cdots$$

となる。

このとき尤度関数は,

$$l(\lambda) = \prod_{i=1}^{n} f(x_i;\lambda) = \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}$$

となり，対数尤度関数は,

$$\log l(\lambda) = \sum_{i=1}^{n} \log f(x_i;\lambda) = \log(\lambda) \sum_{i=1}^{n} x_i - n\lambda - \sum_{i=1}^{n} \log(x_i!)$$

となる。

$\log l(\lambda)$ を最大にする p を求める。

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

したがって，λ について解くと，λ の最尤推定値 $\hat{\lambda}$ は,

$$\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

となる。

さらに，x_i を X_i で置き換えて，λ の最尤推定量 $\hat{\lambda}$ は,

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

となる。

$\hat{\lambda}$ の分布を求める。確率関数に対数を取って,

$$\log f(X_i;\lambda) = X_i \log(\lambda) - \lambda - \log(X_i!)$$

となる。λ に関して，微分すると,

$$\frac{\mathrm{d}\log f(X_i;\lambda)}{\mathrm{d}\lambda} = \frac{X_i}{\lambda} - 1$$

を得る。再度，λ に関して，微分すると，

$$\frac{\mathrm{d}^2 \log f(X_i;\lambda)}{\mathrm{d}\lambda^2} = -\frac{X_i}{\lambda^2}$$

となる。期待値をとって，

$$\mathrm{E}\Big(\frac{\mathrm{d}^2 \log f(X_i;\lambda)}{\mathrm{d}\lambda^2}\Big) = -\frac{\mathrm{E}(X_i)}{\lambda^2}$$

が得られる。期待値を計算すると，

$$\begin{aligned}\mathrm{E}(X_i) &= \sum_{x=0}^{\infty} x f(x;\lambda) = \sum_{x=0}^{\infty} x \frac{\lambda^x e^{-\lambda}}{x!} \\ &= \sum_{x=1}^{\infty} x \frac{\lambda^x e^{-\lambda}}{x!} = \sum_{x=1}^{\infty} \lambda \frac{\lambda^{x-1} e^{-\lambda}}{(x-1)!} = \sum_{x=0}^{\infty} \lambda \frac{\lambda^x e^{-\lambda}}{x!} = \lambda\end{aligned}$$

となる。したがって，

$$\sigma_\theta^2 = -\frac{1}{\sum_{i=1}^n \mathrm{E}\Big(\frac{\mathrm{d}^2 \log f(X_i;\lambda)}{\mathrm{d}\lambda^2}\Big)} = \frac{\lambda}{n}$$

となる。よって，$\hat{\lambda}$ の分布は，n が大きいとき，

$$\hat{\lambda} \sim N(\lambda,\ \frac{\lambda}{n})$$

を得る。

例 4： $X_1, X_n, \cdots, X_n$ は互いに独立で，それぞれパラメータ λ を持った指数分布に従うものとする。すなわち，X_i の密度関数は,

$$f(x;\lambda) = \lambda e^{-\lambda x} \qquad\qquad x > 0$$

となる。

このとき尤度関数は,

$$l(\lambda) = \prod_{i=1}^{n} f(x_i;\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i}$$

となり，対数尤度関数は,

$$\log l(\lambda) = \sum_{i=1}^{n} \log f(x_i;\lambda) = n \log \lambda - \lambda \sum_{i=1}^{n} x_i$$

となる。

$\log l(\lambda)$ を最大にする p を求める。

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

したがって，λ について解くと，λ の最尤推定値 $\hat{\lambda}$ は，

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} x_i}$$

となる。

さらに，x_i を X_i で置き換えて，λ の最尤推定量 $\hat{\lambda}$ は，

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

となる。

$\hat{\lambda}$ の分布を求める。密度関数の対数を取って，

$$\log f(X_i; \lambda) = \log \lambda - \lambda X_i$$

となる。λ について微分すると，

$$\frac{\mathrm{d} \log f(X_i; \lambda)}{\mathrm{d}\lambda} = \frac{1}{\lambda} - X_i$$

を得る。再度，微分して，

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

が得られる。

$$\sigma_\theta^2 = -\frac{1}{\sum_{i=1}^n \mathrm{E}\Bigl(\frac{\mathrm{d}^2 \log f(X_i;\lambda)}{\mathrm{d}\lambda^2}\Bigr)} = \frac{\lambda^2}{n}$$

となるので，n が大きいとき，

$$\hat{\lambda} \sim N(\lambda,\ \frac{\lambda^2}{n})$$

を得る。

2 Maximum Likelihood Estimation (MLE, 最尤法) — More Formally Review

1. We have random variables X_1, X_2, $\cdots$, X_n, which are assumed to be mutually independently and identically distributed.

2. The distribution function of $\{X_i\}_{i=1}^n$ is $f(x;\theta)$, where $x = (x_1, x_2, \cdots, x_n)$ and $\theta = (\mu, \Sigma)$.

   Note that X is a vector of random variables and x is a vector of their realizations (i.e., observed data).

   Likelihood function $L(\cdot)$ is defined as $L(\theta; x) = f(x;\theta)$.

   Note that $f(x;\theta) = \prod_{i=1}^n f(x_i;\theta)$ when X_1, X_2, $\cdots$, X_n are mutually independently and identically distributed.

The maximum likelihood estimator (MLE) of θ is θ such that:

$$\max_{\theta} \ L(\theta; X). \qquad \Longleftrightarrow \qquad \max_{\theta} \ \log L(\theta; X).$$

MLE satisfies the following two conditions:

(a) $\dfrac{\partial \log L(\theta; X)}{\partial \theta} = 0.$

(b) $\dfrac{\partial^2 \log L(\theta; X)}{\partial \theta \partial \theta'}$ is a negative definite matrix.

3. **Fisher's information matrix (フィッシャーの情報行列)** is defined as:

$$I(\theta) = -\mathrm{E}\Big(\frac{\partial^2 \log L(\theta; X)}{\partial \theta \partial \theta'}\Big),$$

where we have the following equality:

$$-\mathrm{E}\Big(\frac{\partial^2 \log L(\theta; X)}{\partial \theta \partial \theta'}\Big) = \mathrm{E}\Big(\frac{\partial \log L(\theta; X)}{\partial \theta} \frac{\partial \log L(\theta; X)}{\partial \theta'}\Big) = \mathrm{V}\Big(\frac{\partial \log L(\theta; X)}{\partial \theta}\Big)$$

Proof of the above equality:

$$\int L(\theta; x)\mathrm{d}x = 1$$

Take a derivative with respect to θ.

$$\int \frac{\partial L(\theta; x)}{\partial \theta}\mathrm{d}x = 0$$

(We assume that (i) the domain of x does not depend on θ and (ii) the derivative $\dfrac{\partial L(\theta; x)}{\partial \theta}$ exists.)

Rewriting the above equation, we obtain:

$$\int \frac{\partial \log L(\theta; x)}{\partial \theta} L(\theta; x)\mathrm{d}x = 0,$$

i.e.,

$$\mathrm{E}\left(\frac{\partial \log L(\theta; X)}{\partial \theta}\right) = 0.$$

Again, differentiating the above with respect to θ, we obtain:

$$\begin{aligned}
&\int \frac{\partial^2 \log L(\theta; x)}{\partial\theta\partial\theta'} L(\theta; x)\mathrm{d}x + \int \frac{\partial \log L(\theta; x)}{\partial\theta} \frac{\partial L(\theta; x)}{\partial\theta'} \mathrm{d}x \\
&\qquad = \int \frac{\partial^2 \log L(\theta; x)}{\partial\theta\partial\theta'} L(\theta; x)\mathrm{d}x + \int \frac{\partial \log L(\theta; x)}{\partial\theta} \frac{\partial \log L(\theta; x)}{\partial\theta'} L(\theta; x)\mathrm{d}x \\
&\qquad = \mathrm{E}\Big(\frac{\partial^2 \log L(\theta; X)}{\partial\theta\partial\theta'}\Big) + \mathrm{E}\Big(\frac{\partial \log L(\theta; X)}{\partial\theta} \frac{\partial \log L(\theta; X)}{\partial\theta'}\Big) = 0.
\end{aligned}$$

Therefore, we can derive the following equality:

$$-\mathrm{E}\left(\frac{\partial^2 \log L(\theta; X)}{\partial\theta\partial\theta'}\right) = \mathrm{E}\left(\frac{\partial \log L(\theta; X)}{\partial\theta} \frac{\partial \log L(\theta; X)}{\partial\theta'}\right) = \mathrm{V}\left(\frac{\partial \log L(\theta; X)}{\partial\theta}\right),$$

where the second equality utilizes $\mathrm{E}\left(\frac{\partial \log L(\theta; X)}{\partial\theta}\right) = 0.$

4. **Cramer-Rao Lower Bound (クラメール・ラオの下限)**: $(I(\theta))^{-1}$

Suppose that an unbiased estimator of θ is given by $s(X)$.

Then, we have the following:

$$\mathrm{V}(s(X)) \geq (I(\theta))^{-1}$$

Proof:

The expectation of $s(X)$ is:

$$\mathrm{E}(s(X)) = \int s(x) L(\theta; x) \mathrm{d}x.$$

Differentiating the above with respect to θ,

$$\begin{aligned}
\frac{\partial \mathrm{E}(s(X))}{\partial \theta'} &= \int s(x) \frac{\partial L(\theta; x)}{\partial \theta'} \mathrm{d}x = \int s(x) \frac{\partial \log L(\theta; x)}{\partial \theta'} L(\theta; x) \mathrm{d}x \\
&= \mathrm{Cov}\left(s(X), \frac{\partial \log L(\theta; X)}{\partial \theta} \right)
\end{aligned}$$