6 More about likelihood

6.1 Invariance property of m.l.e.'s

Theorem

If $\widehat{\theta}$ is an m.l.e. of θ and if g is a function, then $g\left(\widehat{\theta}\right)$ is an m.l.e. of $g(\theta)$.

Proof

If g is one-to-one, then

$$L(\theta) = L\left(g^{-1}\left(g(\theta)\right)\right)$$

are both maximised by $\widehat{\theta}$, so

$$\widehat{\theta} = g^{-1}\left(\widehat{g(\theta)}\right)$$

or

$$g\left(\widehat{\theta}\right) = \widehat{g(\theta)}.$$

If g is many-to-one, then $\widehat{\theta}$ which maximises $L(\theta)$ still corresponds to $g\left(\widehat{\theta}\right)$, so $g\left(\widehat{\theta}\right)$ still corresponds to the maximum of $L(\theta)$.

Example

Suppose $X_1, X_2, \ldots, X_n$ is a random sample from a Bernoulli distribution $B(1, \theta)$. Consider m.l.e.'s of the mean, θ, and variance, $\theta(1-\theta)$.

Note, by the way, that $\theta(1-\theta)$ is not a 1-1 function of θ.

The log-likelihood is

$$l(\theta) = \textstyle\sum_i x_i \log\theta + (n - \sum_i x_i)\log(1-\theta)$$

and

$$\tfrac{dl(\theta)}{d\theta} = \textstyle\sum_i x_i / \theta - (n - \sum_i x_i)/(1-\theta)$$

so it is easily shown that the m.l.e. of θ is $\widehat{\theta} = \overline{X}$.

Putting $\nu = \theta(1-\theta)$,

$$\frac{dl(\nu)}{d\nu} = \frac{dl\left(\nu(\theta)\right)}{d\theta} \cdot \frac{d\theta}{d\nu}$$

so it is easily seen that, since $\frac{d\theta}{d\nu}$ is not, in general, equal to zero,

$$\widehat{\nu} = \nu\left(\widehat{\theta}\right) = \overline{X}\left(1 - \overline{X}\right).$$

6.2 Relative likelihood

If $\sup_\theta L(\theta) < \infty$, the *relative likelihood* is

$$RL(\theta) = \frac{L(\theta)}{\sup_\theta L(\theta)}; \quad 0 \leq RL(\theta) \leq 1.$$

Relative likelihood is invariant to known 1-1 transformations of x, for if y is a 1-1 function of x,

$$f_Y(y;\theta) = f_X\left(x(y);\theta\right)\left|\frac{dx}{dy}\right|.$$

$\left|\frac{dx}{dy}\right|$ is independent of θ, so

$$RL_X(\theta) = RL_Y(\theta).$$

6.3 Likelihood summaries

Realistic statistical problems often have many parameters. These cause problems because it can be hard to visualise $L(\theta)$, and it becomes necessary to use summaries.

Key idea

In large samples, log-likelihoods are often approximately quadratic near the maximum.

Example

Suppose $X_1, X_2, \ldots, X_n$ is a random sample from an exponential distribution with parameter λ.
i.e.

$$f_X(x) = \lambda e^{-\lambda x}, \quad x \geq 0.$$

Then

$$l(\lambda) = n \log \lambda - \lambda \textstyle\sum_i x_i, \qquad \frac{dl(\lambda)}{d\lambda} = n/\lambda - \textstyle\sum_i x_i,$$

$$\frac{d^2 l(\lambda)}{d\lambda^2} = -n/\lambda^2, \qquad \frac{d^3 l(\lambda)}{d\lambda^3} = 2n/\lambda^3.$$

The log-likelihood has a maximum at $\widehat{\lambda} = n/\sum_i x_i$, so

$$\begin{aligned} RL(\lambda) &= \left(\frac{\lambda}{\widehat{\lambda}}\right)^n e^{n-\lambda \sum_i x_i} \\ &= \left(\frac{\lambda}{\widehat{\lambda}} e^{1-\lambda/\widehat{\lambda}}\right)^n, \qquad \lambda > 0. \\ &\to 1 \quad \text{as} \quad \lambda \to \widehat{\lambda}. \end{aligned}$$

Now, what happens as, for $\widehat{\lambda}$ fixed, $n \to \infty$?

$$\begin{aligned} \log RL(\lambda) &= l(\lambda) - l(\widehat{\lambda}) \\ &= l(\widehat{\lambda}) + l'(\widehat{\lambda})\left(\lambda - \widehat{\lambda}\right) + \frac{1}{2} l''(\widehat{\lambda})\left(\lambda_1 - \widehat{\lambda}\right)^2 - l(\widehat{\lambda}) \end{aligned}$$

using Taylor series, where $\left|\lambda_1 - \widehat{\lambda}\right| < \left|\lambda - \widehat{\lambda}\right|$.

Now $l'(\widehat{\lambda}) = 0$ and $l''(\widehat{\lambda}) = -n \Big/ \widehat{\lambda}^2$, so

$$\log RL(\lambda) = -\frac{n\left(\lambda_1 - \widehat{\lambda}\right)^2}{\widehat{\lambda}^2} \to -\infty \quad \text{as} \quad n \to \infty$$

unless $\lambda = \widehat{\lambda}$.

Thus, as $n \to \infty$,

$$RL(\lambda) \to \begin{cases} 1, & \lambda = \widehat{\lambda}, \\ 0, & \text{otherwise.} \end{cases}$$

Conclusion

Likelihood becomes more concentrated about the maximum as $n \to \infty$, and values far from the maximum become less and less plausible.

In general

We call the value $\widehat{\theta}$ which maximises $L(\theta)$ or, equivalently, $l(\theta) = \log L(\theta)$ the *maximum likelihood estimate*, and

$$J(\theta) = -\frac{\partial^2 l(\theta)}{\partial \theta^2}$$

is called the *observed information*.

Usually $J(\theta) > 0$ and $J(\widehat{\theta})$ measures the concentration of $l(\theta)$ at $\widehat{\theta}$. Close to $\widehat{\theta}$, we summarise

$$l(\theta) \simeq l(\widehat{\theta}) - \frac{1}{2}\left(\theta - \widehat{\theta}\right)^2 J(\widehat{\theta}).$$

6.4 Information

In a model with log-likelihood $l(\theta)$, the *observed information* is

$$J(\theta) = -\frac{\partial^2 l(\theta)}{\partial \theta^2}.$$

When observations are independent, $L(\theta)$ is a product of densities so

$$l(\theta) = \sum_i \log f\left(x_i; \theta\right)$$

and

$$J(\theta) = -\sum_i \frac{\partial^2}{\partial \theta^2} \log f\left(x_i; \theta\right).$$

Since

$$l(\theta) \simeq l(\widehat{\theta}) - \frac{1}{2}\left(\theta - \widehat{\theta}\right)^2 J(\widehat{\theta}),$$

for θ near to $\widehat{\theta}$, we see that large $J(\widehat{\theta})$ implies that $l(\theta)$ is more concentrated about $\widehat{\theta}$.

This means that the data are less ambiguous about possible values of θ, *i.e.* we have more information about θ.

6.5 Expected information

6.5.1 Univariate distributions

Before an experiment is conducted, we have no data so that we cannot evaluate $J(\theta)$.

But we can find its expected value

$$I(\theta) = E\left(-\frac{\partial^2 l(\theta)}{\partial \theta^2}\right).$$

This is called the *expected information* or *Fisher's information.*

If the observations are a random sample, then the whole sample expected information is

$$I(\theta) = ni(\theta)$$

where

$$i(\theta) = E\left(-\frac{\partial^2}{\partial \theta^2} \log f\left(X_i; \theta\right)\right),$$

the single observation Fisher information.

Example

Suppose $X_1, X_2, \ldots, X_n$ is a random sample from a Poisson distribution with parameter θ.

$$L(\theta) = \prod_{i=1}^{n} \frac{\theta^{x_i} e^{-\theta}}{x_i!},$$

giving

$$l(\theta) = \log L(\theta) = \sum_i x_i \log \theta - n\theta - \sum_i \log x_i!$$

Thus

$$J(\theta) = -\frac{\partial^2 l(\theta)}{\partial \theta^2} = \sum\nolimits_i x_i \Big/ \theta^2.$$

To find $I(\theta)$, we need $E\left(X_i\right) = \theta$ and

$$I(\theta) = \frac{1}{\theta^2} \sum_i E\left(X_i\right) = \frac{n}{\theta}.$$

6.5.2 Multivariate distributions

If $\boldsymbol{\theta}$ is a $(p \times 1)$ vector of parameters, then $\mathbf{I}(\boldsymbol{\theta})$ and $\mathbf{J}(\boldsymbol{\theta})$ are $(p \times p)$ matrices.

$$\{\mathbf{J}(\boldsymbol{\theta})\}_{rs} = -\frac{\partial^2 l(\boldsymbol{\theta})}{\partial\theta_r \partial\theta_s} \quad \text{and} \quad \{\mathbf{I}(\boldsymbol{\theta})\}_{rs} = E\left(-\frac{\partial^2 l(\boldsymbol{\theta})}{\partial\theta_r \partial\theta_s}\right).$$

These matrices are obviously symmetric.

We can also write the above as

$$\mathbf{J}(\theta) = -\frac{\partial^2 l(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T} \quad \text{and} \quad \mathbf{I}(\theta) = E\left(-\frac{\partial^2 l(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\right).$$

Example

$X_1, X_2, \ldots, X_n$ is a random sample from a normal distribution with parameters μ and σ^2. We have already seen that

$$L\left(\mu, \sigma^2\right) = \left(2\pi\sigma^2\right)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\sum\nolimits_i \left(x_i - \mu\right)^2\right],$$

so

$$l\left(\mu, \sigma^2\right) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum\nolimits_i \left(x_i - \mu\right)^2.$$

and

$$\begin{aligned}
\tfrac{\partial l}{\partial \mu} &= \tfrac{1}{\sigma^2}\textstyle\sum_i \left(x_i - \mu\right), \\
\tfrac{\partial l}{\partial \sigma^2} &= -\tfrac{n}{2\sigma^2} + \tfrac{1}{2\sigma^4}\textstyle\sum_i \left(x_i - \mu\right)^2, \\
\tfrac{\partial^2 l}{\partial \mu^2} &= -\tfrac{n}{\sigma^2}, \\
\tfrac{\partial^2 l}{\partial \mu \partial \sigma^2} &= -\tfrac{1}{\sigma^4}\textstyle\sum_i \left(x_i - \mu\right). \\
\tfrac{\partial^2 l}{\partial (\sigma^2)^2} &= \tfrac{n}{2\sigma^4} - \tfrac{1}{\sigma^6}\textstyle\sum_i \left(x_i - \mu\right)^2.
\end{aligned}$$

$$\mathbf{J}(\mu, \sigma^2) = \begin{pmatrix} \frac{n}{\sigma^2} & \frac{1}{\sigma^4}\sum_i \left(x_i - \mu\right) \\ \frac{1}{\sigma^4}\sum_i \left(x_i - \mu\right) & \frac{1}{\sigma^6}\sum_i \left(x_i - \mu\right)^2 - \frac{n}{2\sigma^4} \end{pmatrix}.$$

To find $\mathbf{I}(\mu, \sigma^2)$, use

$$\begin{aligned}
E\left(X_i\right) &= \mu, \\
V\left(X_i\right) &= E\left[\left(X_i - \mu\right)^2\right] = \sigma^2,
\end{aligned}$$

so that

$$\mathbf{I}(\mu, \sigma^2) = E\left(\mathbf{J}(\mu, \sigma^2)\right) = \begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix}.$$

Example Censored exponential data

Lifetimes of n components, safety devices, etc. are observed for a time c, when r have failed and $(n-r)$ are still OK.

We have two kinds of observation:

1. Exact failure times x_i observed if $x_i \leq c$, so that

$$f(x;\lambda) = \lambda e^{-\lambda x}, \quad x \geq 0;$$

2. x_i unobserved if $x_i > c$,

$$P(X > c) = e^{-\lambda c}.$$

Data are therefore $x_1, \ldots, x_r, \underbrace{c, \ldots, c}_{n-r \text{ times}}$

The $(n-r)$ components, safety devices, etc. which have not failed are said to be *censored*.

The likelihood is

$$\begin{aligned} L(\lambda) &= \prod_{i=1}^{r} \lambda e^{-\lambda x_i} \prod_{i=r+1}^{n} e^{-\lambda c} \\ &= \lambda^r \exp\left[-\lambda\left(\sum_{i=1}^{r} x_i + (n-r)c\right)\right]. \\ l(\lambda) &= r\log\lambda - \lambda\left(\textstyle\sum_{i=1}^{r} x_i + (n-r)c\right) \\ l'(\lambda) &= r/\lambda - \left(\textstyle\sum_{i=1}^{r} x_i + (n-r)c\right) \\ l''(\lambda) &= -r/\lambda^2 . \end{aligned}$$

Thus $J(\lambda) = r/\lambda^2 > 0$ if $r > 0$ so we must observe *at least one* exact failure time.

$$I(\lambda) = E\left(r/\lambda^2\right) = \frac{1}{\lambda^2} E\left(\#X_i \text{ observed exactly.}\right)$$

Now $P(X_i \text{ observed exactly}) = P(X_i \leq c) = 1 - e^{-\lambda c}$, so

$$I_c(\lambda) = \frac{n\left(1 - e^{-\lambda c}\right)}{\lambda^2}.$$

No censoring if $c \to \infty$, giving

$$I_\infty(\lambda) = \frac{n}{\lambda^2} > I_c(\lambda)$$

as one might expect.

The asymptotic efficiency when there is censoring at c relative to no censoring is

$$I_c(\lambda) / I_\infty(\lambda) = 1 - e^{-\lambda c}.$$

Example Events in a Poisson process

Events are observed for period $(0, T)$.

n events occur at times $0 < t_1 < t_2 < \ldots < t_n < T$

Two observers A and B. A records exact times, B uses an automatic counter and goes to the pub (*i.e.* B merely records how many events there are).

A knows exact times, and times between events are independent and exponentially distributed, so

$$\begin{aligned} L_A(\lambda) &= \lambda e^{-\lambda t_1} \times \lambda e^{-\lambda(t_2 - t_1)} \times \cdots \times \lambda e^{-\lambda(t_n - t_{n-1})} \times \lambda e^{-(\lambda T - t_n)} \\ &= \lambda^n e^{-\lambda T}. \end{aligned}$$

B merely observes the event $[N = n]$, where $N \sim Poi(\lambda T)$, so

$$L_B(\lambda) = \frac{(\lambda T)^n e^{-\lambda T}}{n!}.$$

Log-likelihoods are

$$\begin{aligned} l_A(\lambda) &= n \log \lambda - \lambda T, \\ l_B(\lambda) &= n \log \lambda + n \log T - \lambda T - \log n! \end{aligned}$$

and

$$J_A(\lambda) = J_B(\lambda) = n / \lambda^2 .$$

$E(N) = \lambda T$, so $I_A(\lambda) = I_B(\lambda) = T / \lambda$, and both observers get the same information. As usual, the one who went to the pub did the right thing.

6.6 Maximum likelihood estimates

The maximum likelihood estimate $\widehat{\theta}$ of θ maximises $L(\theta)$ and often (but not always) satisfies the *likelihood equation*

$$\frac{\partial l}{\partial \theta}\left(\widehat{\theta}\right) = 0,$$

with

$$J(\widehat{\theta}) = -\frac{\partial^2 l}{\partial \theta^2}\left(\widehat{\theta}\right) > 0$$

for a maximum.

In the vector case, $\widehat{\boldsymbol{\theta}}$ solves simultaneously

$$\frac{\partial l}{\partial \theta_r}\left(\widehat{\boldsymbol{\theta}}\right) = 0, \quad r = 1, \ldots, p,$$

with

$$\det \mathbf{J}(\widehat{\boldsymbol{\theta}}) > 0$$

(*i.e.* $\mathbf{J}(\widehat{\boldsymbol{\theta}})$ positive definite).

If the likelihood equation has many solutions, we find them all and check $L(\theta)$ for each.

Usually, the equation has to be solved numerically. One way is by Newton-Raphson.

Suppose we have a starting value $\theta_{\mathbf{0}}$. Then

$$0 = \frac{\partial l}{\partial \theta}\left(\widehat{\theta}\right) \simeq \frac{\partial l}{\partial \theta}\left(\theta_{\mathbf{0}}\right) + \frac{\partial^2 l}{\partial \theta^2}\left(\theta_{\mathbf{0}}\right)\left(\widehat{\theta} - \theta_{\mathbf{0}}\right)$$

which may be re-arranged to

$$\widehat{\theta} = \theta_{\mathbf{0}} + \frac{U(\theta_{\mathbf{0}})}{J(\theta_{\mathbf{0}})},$$

where

$$\begin{aligned} U(\theta) &= \frac{\partial l}{\partial \theta} \text{ is the } \textit{score function}, \\ J(\theta) &= -\frac{\partial^2 l}{\partial \theta^2} \text{ is the observed information.} \end{aligned}$$

Now we iterate using θ_0 as a starting value and

$$\theta_{n+1} = \theta_n + \frac{U(\theta_n)}{J(\theta_n)}.$$

Example Extreme value (Gumbel) distribution

This distribution is used to model such things as annual maximum temperature. Data due to Bliss on numbers of beetles killed by exposure to carbon disulphide are fitted by this model. The c.d.f. is

$$F(x) = \exp\left(-e^{-(x-\eta)}\right), \quad x \in \mathbb{R},\ \eta \in \mathbb{R},$$

and the density is

$$f(x) = \exp\left[-(x-\eta) - e^{-(x-\eta)}\right], \quad x \in \mathbb{R},\ \eta \in \mathbb{R}.$$

The sample log-likelihood is

$$l(\eta) = -\sum_i (x_i - \eta) - \sum_i e^{-(x_i-\eta)},$$

so that

$$\begin{aligned} U(\eta) &= n - \sum_i e^{-(x_i-\eta)}, \\ J(\eta) &= \sum_i e^{-(x_i-\eta)}. \end{aligned}$$

Starting at $\eta_0 = \overline{x}$, iterate using

$$\eta_{n+1} = \eta_n + \frac{n - \sum_i e^{-(x_i-\eta_n)}}{\sum_i e^{-(x_i-\eta_n)}.}$$

6.6.1 Fisher scoring

This simply involves replacing $J(\theta)$ with $I(\theta)$.

Example Extreme value distribution

We need

$$\begin{aligned} I(\eta) &= E\left[J(\eta)\right] = \sum_i E\left[e^{-(X_i-\eta)}\right] \\ &= n\int_{-\infty}^{\infty} e^{-(x-\eta)} \exp\left[-(x-\eta) - e^{-(x-\eta)}\right] dx. \end{aligned}$$

Put $u = e^{-(x-\eta)}$ and the integral becomes

$$I(\eta) = n \int_0^\infty u e^{-u} du = n,$$

so Fisher scoring gives the iteration

$$\eta_{n+1} = \eta_n + 1 - \frac{1}{n} \sum_i e^{-(x_i - \eta_n)}.$$

6.7 Sufficient statistics

You have already seen a likelihood which cannot be summarised by a quadratic.

Example

$$f(x_i;\theta) = \theta^{-1}, \quad 0 < x_i < \theta,$$

so

$$L(\theta) = \theta^{-n}, \quad 0 < \max\{x_i\} < \theta.$$

Clearly a quadratic approximation is useless here.

Suppose there exists a statistic $s(\mathbf{x})$ such that $L(\theta)$ only depends upon data $\mathbf{x}$ through $s(\mathbf{x})$. Then $s(\mathbf{X})$ is a *sufficient statistic* for θ and obviously always exists.

The important question is:

> Does $s(\mathbf{x})$ reduce the dimensionality of the problem?

Definition

If $S = s(\mathbf{X})$ is such that the conditional density $f_{X|S}(\mathbf{x}|\mathbf{s};\boldsymbol{\theta})$ is independent of θ, then S is a sufficient statistic.

Example

Suppose $X_1, X_2 \sim B(n, \theta)$ and consider

$$\begin{aligned}
& P\left(X_1 = x | X_1 + X_2 = r\right) \\
= & \frac{P\left(X_1 = x, X_1 + X_2 = r\right)}{P\left(X_1 + X_2 = r\right)} \\
= & \frac{P\left(X_1 = x, X_2 = r - x\right)}{P\left(X_1 + X_2 = r\right)} \\
= & \frac{\binom{n}{x}\theta^x \left(1-\theta\right)^{n-x} \binom{n}{r-x}\theta^{r-x}\left(1-\theta\right)^{n-r+x}}{\binom{2n}{r}\theta^r\left(1-\theta\right)^{2n-r}} \\
= & \frac{\binom{n}{x}\binom{n}{r-x}}{\binom{2n}{r}}
\end{aligned}$$

This does not contain θ, so that $X_1 + X_2$ is a sufficient statistic for θ.

Example

$X_1, X_2, \ldots, X_n \sim U(0, \theta)$, so that

$$L(\theta) = \theta^{-n}, \quad 0 < x_1, \ldots, x_n < \theta.$$

Suppose we find the conditional density of $X_1, X_2, \ldots, X_n$ given $X_{(n)}$.

The joint density of $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ is

$$f(x_{(1)}, x_{(2)}, \ldots, x_{(n)}) = \frac{n!}{\theta^n}, \quad 0 < x_{(1)}, \ldots, x_{(n)} < \theta.$$

and the density of $X_{(n)}$ is nx^{n-1}/θ^n so that the conditional density of $X_{(1)}, \ldots, X_{(n-1)} \,|X_{(n)} = y$ is

$$\frac{n!}{\theta^n} \Big/ \frac{nx^{n-1}}{\theta^n} = \frac{(n-1)!}{x^{n-1}}, \quad 0 < x_{(1)}, \ldots, x_{(n-1)} < y.$$

Thus the density of $X_1, X_2, \ldots, X_n | X_{(n)}$ is

$$f\left(x_1, \ldots, x_n \,|x_{(n)} = y\right) = \frac{1}{x^{n-1}}, \quad 0 < x_1, \ldots, x_n < y,$$

which is free of θ, so that $X_{(n)}$ is a sufficient statistic for θ.

Factorization Theorem

$s(\mathbf{X})$ is a *sufficient statistic* for θ if and only if there exist functions g and h such that

$$f(\mathbf{x}; \theta) = g(s(\mathbf{x}); \theta)\, h(\mathbf{x})$$

for all $\mathbf{x} \in \mathbb{R}^n,\ \theta \in \Theta$.

Proof for discrete random variables

(i) Let $s(\mathbf{x}) = a$ and suppose the factorization condition to be satisfied, so that $f(\mathbf{x}; \theta) = g(s(\mathbf{x}); \theta)\, h(\mathbf{x})$.

Then

$$P(s(\mathbf{X}) = a) = \sum_{\mathbf{y} \in s^{-1}(a)} p(\mathbf{y}) = g(a; \theta) \sum_{\mathbf{y} \in s^{-1}(a)} h(\mathbf{y}).$$

Hence

$$P(\mathbf{X} = \mathbf{x} \mid s(\mathbf{X}) = a) = \frac{h(\mathbf{x})}{\sum\limits_{\mathbf{y} \in s^{-1}(a)} h(\mathbf{y})}$$

and this does not depend upon θ.

(ii) Let $s(\mathbf{X})$ be a sufficient statistic for θ. Then

$$P(\mathbf{X}=\mathbf{x}) = \mathbf{P}\left(\mathbf{X}=\mathbf{x} \mid s(\mathbf{X})=a\right)\mathbf{P}\left(s(\mathbf{X})=a\right).$$

But sufficiency $\Rightarrow P\left(\mathbf{X}=\mathbf{x} \mid s(\mathbf{X})=a\right)$ does not depend upon θ so, writing $P\left(s(\mathbf{X})=a\right) = g(a;\theta)$ and $P\left(\mathbf{X}=\mathbf{x} \mid s(\mathbf{X})=a\right) = h(\mathbf{x})$ gives the result.

The proof in the continuous case requires measure theory and is beyond the scope of this course.

Example

Suppose $X_1, X_2, \ldots, X_n$ is a random sample from a Bernoulli distribution. Then

$$p(\mathbf{x};\theta) = \theta^{\sum_i x_i}(1-\theta)^{n-\sum_i x_i}.$$

Trivially this factorizes with $s(\mathbf{x}) = \sum_i x_i$ and $h(\mathbf{x}) = 1$.

Example

Suppose $X_1, X_2, \ldots, X_n$ is a random sample from a $N(\mu, \sigma^2)$ distribution, where $(\mu, \sigma^2)^T$ is a vector of unknown parameters. Then

$$\begin{aligned} f(\mathbf{x};\mu,\sigma^2) &= \left(2\pi\sigma^2\right)^{-n/2}\exp\left[-\frac{1}{2\sigma^2}\sum_i (x_i-\mu)^2\right] \\ &= \left(2\pi\sigma^2\right)^{-n/2}\exp\left[-\frac{1}{2\sigma^2}\sum_i (x_i-\overline{x})^2 + n(\overline{x}-\mu)^2\right]. \end{aligned}$$

Again this factorizes where $s(\mathbf{x}) = (\overline{x}, \sum_i (x_i-\overline{x})^2)^T$, a vector valued function.

6.8 The exponential family

The density function/probability mass function has the form

$$f(x;\varphi) = \exp\left[a(x)b(\varphi) - c(\varphi) + d(x)\right],$$

where x may be continuous or discrete and φ is in a suitable space (usually open reals).

For a random sample $X_1, X_2, \ldots, X_n$, we obtain

$$L(\varphi) = \exp\left[b(\varphi)\sum_i a(x_i) - nc(\varphi) + \sum_i d(x_i)\right],$$

and, therefore, by the factorization theorem, $\sum_i a(x_i)$ is sufficient for φ.

Example

Let $X \sim B(n,\theta)$. Then

$$\begin{aligned} p_X(x) &= \binom{n}{x}\theta^x(1-\theta)^{n-x} \\ &= \exp\left[x\log\theta + (n-x)\log(1-\theta) + \log\binom{n}{x}\right] \\ &= \exp\left[x\log\left(\theta/(1-\theta)\right) + n\log(1-\theta) + \log\binom{n}{x}\right] \end{aligned}$$

Calling $Y = a(X)$, $\theta = b(\varphi)$ the *natural parameterisation*, we can write the density function/probability mass function in the form

$$f(y;\theta) = \exp\left[y\theta - k(\theta)\right] m(y).$$

Clearly $\sum_i Y_i$ is a sufficient statistic.

Note that, in the continuous case, the moment generating function is

$$\begin{aligned} E\left(e^{tY}\right) &= \int e^{ty+\theta y-k(\theta)}m(y)dy \\ &= e^{k(\theta+t)-k(\theta)}\int e^{ty+\theta y-k(\theta+t)}m(y)dy \\ &= e^{k(\theta+t)-k(\theta)}. \end{aligned}$$

The function $k(\theta)$ is called the *cumulant generator*.

Let us see why.

The cumulant generating function

If X is a random variable with moment generating function $M(t)$, then $K(t) = \log M(t)$ is said to be the *cumulant generating function*.

Differentiating,

$$K'(t) = \frac{M'(t)}{M(t)}, \quad K'(0) = \frac{M'(0)}{M(0)} = E(X),$$

$$\begin{aligned} K''(t) &= \tfrac{M''(t)}{M(t)} - \tfrac{M'(t)^2}{M(t)^2}, \\ K''(0) &= \tfrac{M''(0)}{M(0)} - \tfrac{M'(0)^2}{M(0)^2} = V(X) \end{aligned}$$

and so on. The cumulants are generated directly.

For the exponential family,

$$K(t) = \log M(t) = k(\theta + t) - k(\theta)$$

so that

$$K'(t) = k'(\theta + t), \quad K'(0) = k'(\theta),$$

and so on. The cumulants are generated by repeated differentiation of $k(\theta)$.

Example *Poisson distribution*

The p.m.f. is

$$\begin{aligned} p(x;\mu) &= \frac{\mu^x e^{-\mu}}{x!}, \quad x = 0, 1, \ldots \\ &= \exp\left[x \log \mu - \mu - \log x!\right] \end{aligned}$$

so that

$$a(x) = x, \quad b(\mu) = \log \mu, \quad c(\mu) = \mu, \quad d(x) = -\log x!$$

Under natural parameterisation,

$$y = x, \quad \theta = \log \mu, \quad k(\theta) = e^\theta, \quad m(y) = \frac{1}{y!}.$$

Cumulants are given by derivatives of $k(\theta)$, all of which are $e^\theta = \mu$.

Example *Binomial distribution*

The p.m.f. is

$$\begin{aligned} p(x;p) &= \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, \ldots, n \\ &= \exp\left[x \log\left(\frac{p}{1-p}\right) + n\log(1-p) + \log\binom{n}{x} \right]. \end{aligned}$$

Natural parameterisation is

$$y = x, \quad \theta = \log\left(\frac{p}{1-p}\right) \quad k(\theta) = n\log(1+e^{\theta}).$$

For the cumulants,

$$k'(\theta) = \frac{ne^{\theta}}{1+e^{\theta}} = np, \quad k''(\theta) = \frac{ne^{\theta}}{(1+e^{\theta})^2} = np(1-p),$$

and so on.

6.9 Large sample distribution of $\widehat{\theta}$

From the data summary point of view, the m.l.e. $\widehat{\theta}$ and $J\left(\widehat{\theta}\right)$have been thought of in terms of a particular set of data. We now wish to think of $\widehat{\theta}$ in terms of repeated sampling (*i.e.* as a random variable).

Main results

In many situations and subject to regularity conditions

$$\widehat{\theta} \xrightarrow{D} N(\theta, I(\theta)^{-1}),$$

and an approximate 95% confidence interval for θ is given by

$$\widehat{\theta} \pm 1.96 I(\widehat{\theta})^{-1/2}.$$

[or $\widehat{\theta} \pm 1.96 J(\widehat{\theta})^{-1/2}$, regarded by many as better, but not in the books].

In the multivariate case,

$$\widehat{\boldsymbol{\theta}} \xrightarrow{D} N(\boldsymbol{\theta}, \mathbf{I}(\boldsymbol{\theta})^{-1}).$$

Example Exponential distribution

For an exponential distribution with mean θ,

$$\begin{aligned} L(\theta) &= \theta^{-n} e^{-\sum_i x_i/\theta}, \quad \theta > 0, \\ l(\theta) &= -n \log \theta - \sum_i x_i \, / \theta \, , \end{aligned}$$

so that

$$U(\theta) = -\frac{n}{\theta} + \frac{\sum_i x_i}{\theta^2}, \quad J(\theta) = -\frac{n}{\theta^2} + \frac{2\sum_i x_i}{\theta^3}.$$

Thus

$$\widehat{\theta} = \overline{x}, \quad J(\widehat{\theta}) = \frac{n}{\overline{x}^2}$$

and an approximate 95% confidence interval is

$$\overline{x} \pm 1.96 \overline{x} \, / \sqrt{n}$$

Example Normal distribution

For a normal random sample,

$$\widehat{\mu} = \overline{x}, \quad \widehat{\sigma}^2 = n^{-1} \sum_i (x_i - \overline{x})^2 \, ,$$

$$J\left(\mu, \sigma^2\right) = \begin{pmatrix} n \, / \sigma^2 & \sigma^{-4} \sum_i (x_i - \mu) \\ \sigma^{-4} \sum_i (x_i - \mu) & \sigma^{-6} \sum_i (x_i - \mu)^2 - n \, / 2\sigma^4 \end{pmatrix}.$$

Therefore

$$I\left(\widehat{\mu}, \widehat{\sigma}^2\right) = \begin{pmatrix} n \, / \widehat{\sigma}^2 & 0 \\ 0 & n \, / 2\widehat{\sigma}^4 \end{pmatrix}.$$

An approximate 95% confidence interval for μ is

$$\overline{x} \pm 1.96 \widehat{\sigma} \, / \sqrt{n} \, ,$$

and for σ^2 is

$$\widehat{\sigma}^2 \pm 1.96 \widehat{\sigma}^2 \sqrt{\frac{2}{n}}.$$

Note that the estimators $\widehat{\mu}$ and $\widehat{\sigma}^2$ are asymptotically uncorrelated.

The exact interval for μ is

$$\overline{x} \pm \frac{S}{\sqrt{n}} t_{0.975}(n-1)$$

which is not quite the same.

Proof of asymptotic normality

Suppose $X_1, X_2, \ldots, X_n$ is a random sample from a distribution with p.d.f. $f(x;\theta)$. Then the log-likelihood, score and observed information are

$$\begin{aligned} l(\theta) &= \sum_i \log f(x_i;\theta), \\ U(\theta) &= \sum_i \frac{\partial}{\partial\theta} \log f(x_i;\theta), \\ J(\theta) &= -\sum_i \frac{\partial^2}{\partial\theta^2} \log f(x_i;\theta). \end{aligned}$$

Let $U_i(\theta)$ be the random variable $U_i(\theta) = \frac{\partial}{\partial\theta} \log f(X_i;\theta)$, and, provided that conditions are such that integration and differentiation are interchangeable,

$$\begin{aligned} E\left[U_i(\theta)\right] &= \int f(x;\theta) \frac{\partial}{\partial\theta} \log f(x;\theta) dx \\ &= \int \frac{\partial}{\partial\theta} f(x;\theta) dx \\ &= \frac{\partial}{\partial\theta} \int f(x;\theta) dx = \frac{\partial}{\partial\theta} 1 = 0 \end{aligned}$$

and

$$\begin{aligned} 0 &= \frac{\partial}{\partial\theta} \int f(x;\theta) \frac{\partial}{\partial\theta} \log f(x;\theta) dx \\ &= \int f(x;\theta) \frac{\partial^2}{\partial\theta^2} \log f(x;\theta) dx + \int \frac{\partial}{\partial\theta} f(x;\theta) \frac{\partial}{\partial\theta} \log f(x;\theta) dx \\ &= E\left[\frac{\partial^2}{\partial\theta^2} \log f(X;\theta)\right] + \int f(x;\theta) \left(\frac{\partial}{\partial\theta} \log f(x;\theta)\right)^2 dx. \end{aligned}$$

So

$$0 = -i(\theta) + E\left[U_i(\theta)^2\right]$$

and, therefore, $V\left[U_i(\theta)\right] = i(\theta)$.

It follows that $E\left[U(\theta)\right] = 0$, $V\left[U(\theta)\right] = ni(\theta) = I(\theta)$, and the CLT shows that

$$U(\theta) \xrightarrow{D} N\left(0, I(\theta)\right).$$

Now the m.l.e. is a solution of $U(\widehat{\theta}) = 0$, so that, Taylor expanding,

$$U(\theta) + U'(\theta)(\widehat{\theta} - \theta) \simeq 0$$

or

$$U(\theta) - J(\theta)(\widehat{\theta} - \theta) \simeq 0.$$

Re-arranging,

$$\sqrt{I(\theta)}(\widehat{\theta} - \theta) \simeq U(\theta)\frac{\sqrt{I(\theta)}}{J(\theta)} = \frac{U(\theta)}{\sqrt{I(\theta)}} \Big/ \frac{J(\theta)}{I(\theta)} \ .$$

From the CLT,

$$\frac{U(\theta)}{\sqrt{I(\theta)}} \xrightarrow{D} N(0,1)$$

and from WLLN

$$\frac{J(\theta)}{I(\theta)} \xrightarrow{P} 1.$$

Slutsky's Theorem therefore results in

$$\sqrt{I(\theta)}(\widehat{\theta} - \theta) \xrightarrow{D} N(0,1)$$

or

$$\widehat{\theta} \xrightarrow{D} N(\theta, I(\theta)^{-1}).$$

Requirements of this proof

1. The true value of θ is interior to the partameter space.
2. Differentiation under the integral is valid, so that $E\left[U(\theta)\right] = 0$ and $V\left[U(\theta)\right] = ni(\theta)$. This allows a central limit theorem to apply to $U(\theta)$.
3. Taylor expansions are valid for the derivatives of the log-likelihood, so that higher order terms may be neglected.
4. A weak law of large numbers applies to $J(\theta)$.

Example: *Exponential family*

In the natural parameterisation, the likelihood has the form

$$L(\theta) = m(\mathbf{y})e^{\theta \sum y_i - nk(\theta)}$$

so that

$$U(\theta) = \sum y_i - nk'(\theta)$$

$$J(\theta) = nk''(\theta).$$

$\widehat{\theta}$ solves

$$U(\widehat{\theta}) = 0 \quad \Rightarrow \quad k'(\theta) = \overline{y}.$$

Expanding,

$$k'(\theta) + (\widehat{\theta} - \theta)k''(\theta) \simeq \overline{y}$$

so that

$$\widehat{\theta} \simeq \theta + \frac{\overline{y} - k'(\theta)}{k''(\theta)}.$$

Since

$$E(\overline{Y}) = n^{-1} \sum_i E(Y_i) = k'(\theta),$$

we have

$$E(\widehat{\theta}) \simeq \theta.$$

$$V(\widehat{\theta}) = \frac{1}{k''(\theta)^2} V(\overline{Y}) = \frac{n^{-1}k''(\theta)}{k''(\theta)^2} = \frac{1}{nk''(\theta)}$$

which, of course, we could have obtained directly from $\widehat{\theta} \dot{\sim} N\left(\theta, I(\theta)^{-1}\right)$.

Example: *Exponential distribution*

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

so

$$\theta = -\lambda, \quad k(\theta) = -\log\lambda = -\log(-\theta)$$

$$k'(\theta) = -\frac{1}{\theta}, \quad k''(\theta) = \frac{1}{\theta^2}$$

so

$$\widehat{\theta} = -\frac{1}{\overline{y}}, \quad I(\widehat{\theta}) = \frac{n}{\overline{y}^2}.$$

Thus, approximately,

$$-\frac{1}{\overline{y}} \pm z_\alpha \frac{\overline{y}}{\sqrt{n}}$$

gives a confidence interval for θ, and

$$\frac{1}{\overline{y}} \pm z_\alpha \frac{\overline{y}}{\sqrt{n}}$$

gives a confidence interval for λ.