

UNIVERSITY OF TECHNOLOGY, SYDNEY

Advanced Data Analysis

Likelihood Theory and Methods

©M.P. Wand, 2011

1 Preliminary Definitions

These notes use notation and definitions that are common, although not universal, in the Statistics literature. In this section we give some definitions used in these notes.

Definition

For $0 < p < 1$ the p th quantile of the $N(0, 1)$ distribution is denoted by z_p . That is,

$$P(Z \leq z_p) = p \quad \text{where} \quad Z \sim N(0, 1).$$

Examples include:

$$z_{0.95} = 1.645, \quad z_{0.975} = 1.96 \quad \text{and} \quad z_{0.025} = -1.96.$$

Definition

Let X_1, X_2, \dots be a sequence of random variables. We say that X_n **converges in distribution** to X if

$$\lim_{n \rightarrow \infty} F_{X_n}(x) = F_X(x)$$

for all x where F_X is continuous. A common shorthand is

$$X_n \xrightarrow{D} X.$$

We say that F_X is the **limiting distribution** of X_n .

Definition

Let $\hat{\theta}$ be an estimator of a parameter θ . The **standard error** of $\hat{\theta}$ is given by

$$\text{se}(\hat{\theta}) = \sqrt{\text{Var}_{\theta}(\hat{\theta})}.$$

Note that $\text{se}(\hat{\theta})$ is simply the standard deviation of the random variable $\hat{\theta}$.

An **estimated standard error** of $\hat{\theta}$ is one where parameters in $\text{se}(\hat{\theta})$ are replaced by estimators, and is denoted by

$$\widehat{\text{se}}(\hat{\theta}).$$

Definition

The **Gamma function** at $x \in \mathbb{R}$ is given by

$$\Gamma(x) = \int_0^{\infty} t^{x-1} e^{-t} dt.$$

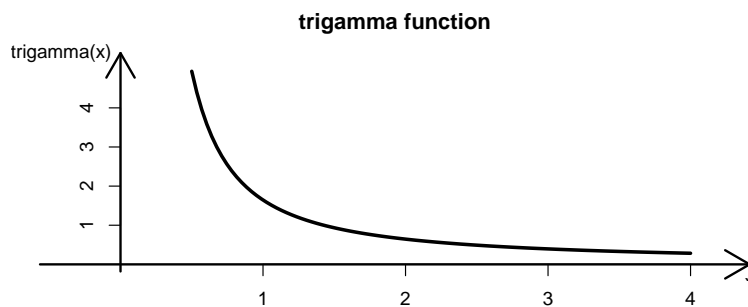
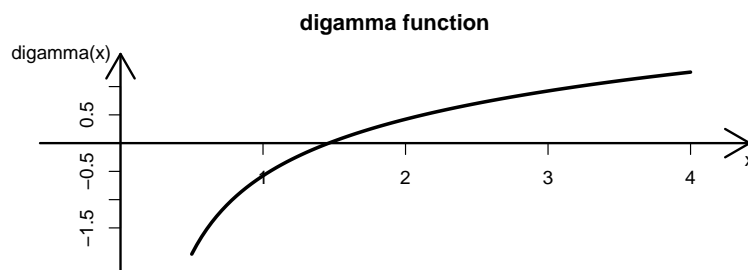
Definition

For all $x \in \mathbb{R}$,

$$\text{digamma}(x) = \frac{d}{dx} \ln\{\Gamma(x)\},$$

$$\text{trigamma}(x) = \frac{d^2}{dx^2} \ln\{\Gamma(x)\}.$$

The digamma and trigamma functions are now readily available in modern computing environments such as Maple, Matlab and R. The following figure displays them graphically.

**Definition**

The **Beta function** at $x, y \in \mathbb{R}$ is given by

$$B(x, y) = \int_0^1 t^{x-1}(1-t)^{y-1} dt.$$

Result

For all $x, y \in \mathbb{R}$,

$$B(x, y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}.$$

Definition

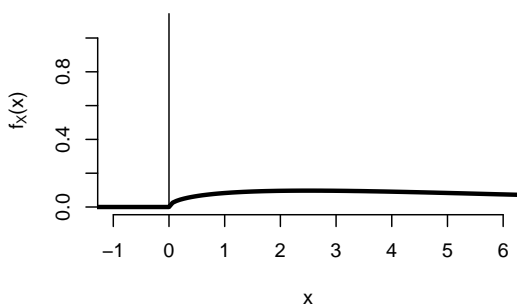
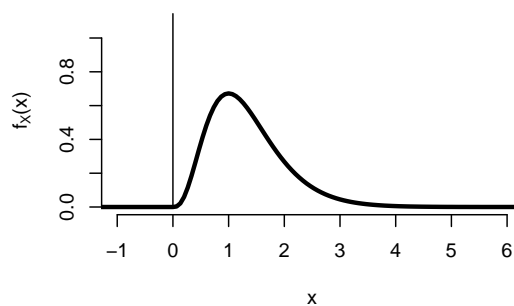
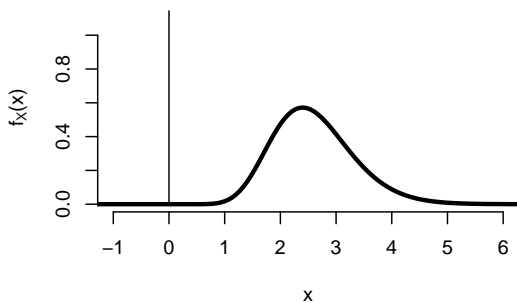
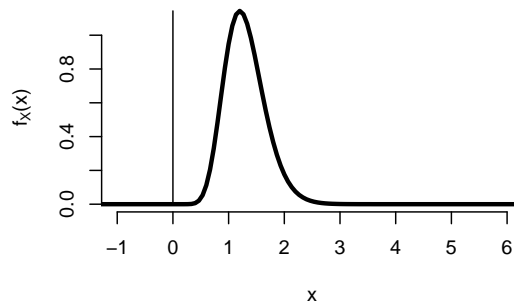
A random variable X is said to have a **Gamma distribution** with parameters α and β (where $\alpha, \beta > 0$) if X has density function:

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

A common shorthand is:

$$X \sim \text{Gamma}(\alpha, \beta).$$

Gamma density functions are skewed curves on the positive half-line. The following figure shows four different Gamma density functions.

Gamma(1.5,0.2)**Gamma(4,3)****Gamma(13,5)****Gamma(13,10)****Definition**

A random variable X is said to have a **Beta distribution** with parameters α and β (where $\alpha, \beta > 0$) if X has density function:

$$f_X(x; \alpha, \beta) = \frac{x^{\alpha-1} (1-x)^{\beta-1}}{B(\alpha, \beta)}, \quad 0 < x < 1.$$

A common shorthand is:

$$X \sim \text{Beta}(\alpha, \beta).$$

2 Maximum Likelihood Estimation

We first define the *likelihood function*:

Definition

Let X_1, \dots, X_n be a random sample with model

$$f_X(x; \theta)$$

containing the parameter $\theta \in \Theta$. The **likelihood function** of θ is

$$\mathcal{L}(\theta) = f_X(X_1; \theta) \cdots f_X(X_n; \theta) = \prod_{i=1}^n f_X(X_i; \theta), \quad \theta \in \Theta,$$

and the **log-likelihood function** of θ is

$$\ell(\theta) = \ln\{\mathcal{L}(\theta)\} = \sum_{i=1}^n \ln\{f_X(X_i; \theta)\}.$$

Note that $\mathcal{L}(\theta)$ and $\ell(\theta)$ are each “random functions” in the sense that, for any fixed value $\theta_0 \in \Theta$, $\mathcal{L}(\theta_0)$ and $\ell(\theta_0)$ are random variables (functions of X_1, \dots, X_n).

Example

Let

p = proportion of university students who regularly eat breakfast.

be a parameter of interest.

Suppose that a nutritional survey is to be conducted on 8 randomly chosen university students, that includes the question

Do you regularly eat breakfast?

Let X_1, \dots, X_8 be such that

$$X_i = \begin{cases} 1, & \text{if } i\text{th surveyed student regularly eats breakfast} \\ 0, & \text{otherwise.} \end{cases}$$

for $i = 1, \dots, 8$. An appropriate model is

$$X_1, \dots, X_8 \sim f_X(x; p), \quad 0 < p < 1.$$

where

$$\begin{aligned} f_X(x; p) &= \begin{cases} p, & x = 1 \\ 1 - p, & x = 0 \end{cases} \\ &= p^x (1 - p)^{1-x} \end{aligned}$$

The likelihood function for p is

$$\mathcal{L}(p) = \prod_{i=1}^8 \{p^{X_i} (1 - p)^{1-X_i}\} = p^{\sum_{i=1}^8 X_i} (1 - p)^{8 - \sum_{i=1}^8 X_i}$$

and the log-likelihood function for p is

$$\ell(p) = \ln\{\mathcal{L}(p)\} = \left(\sum_{i=1}^8 X_i \right) \ln(p) + \left(8 - \sum_{i=1}^8 X_i \right) \ln(1 - p).$$

If we take $p = 0.3$, say, then

$$\mathcal{L}(0.3) = (0.3)^{\sum_{i=1}^8 X_i} (0.7)^{8 - \sum_{i=1}^8 X_i}$$

and

$$\ell(0.3) = \left(\sum_{i=1}^8 X_i \right) \ln(0.3) + \left(8 - \sum_{i=1}^8 X_i \right) \ln(0.7)$$

are seen to be random variables based on transformation of X_1, \dots, X_8 .

Definition

Let X_1, \dots, X_n be a random sample with model

$$f_X(x; \theta)$$

containing the parameter $\theta \in \Theta$. The **maximum likelihood estimator** for θ is

$$\hat{\theta} = \theta \text{ that maximises } \mathcal{L}(\theta) \text{ over } \theta \in \Theta.$$

Example

In the breakfast eating example ($n = 8$) the likelihood function has been shown to be

$$\mathcal{L}(p) = \prod_{i=1}^8 \{p^{X_i} (1-p)^{1-X_i}\} = p^{\sum_{i=1}^8 X_i} (1-p)^{8 - \sum_{i=1}^8 X_i} = e^{(\sum_{i=1}^8 X_i) \ln(p) + (8 - \sum_{i=1}^8 X_i) \ln(1-p)}.$$

The first derivative of $\mathcal{L}(p)$ with respect to p is then

$$\begin{aligned} \frac{d}{dp} \mathcal{L}(p) &= e^{(\sum_{i=1}^8 X_i) \ln(p) + (8 - \sum_{i=1}^8 X_i) \ln(1-p)} \left[\left(\frac{\sum_{i=1}^8 X_i}{p} \right) - \left(\frac{8 - \sum_{i=1}^8 X_i}{1-p} \right) \right] \\ &= \mathcal{L}(p) \left(\frac{\sum_{i=1}^8 X_i}{p} - \frac{8 - \sum_{i=1}^8 X_i}{1-p} \right). \end{aligned}$$

and is zero if and only if

$$\frac{\sum_{i=1}^8 X_i}{p} - \frac{8 - \sum_{i=1}^8 X_i}{1-p} = 0 \iff p = \frac{\sum_{i=1}^8 X_i}{8}.$$

Further analysis (see next example) shows that this is the unique maximiser of $\mathcal{L}(p)$ over $0 < p < 1$ so

$$\hat{p} = \frac{\sum_{i=1}^8 X_i}{8} = \text{proportion of breakfast eaters in survey}$$

is the maximum likelihood estimator of p .

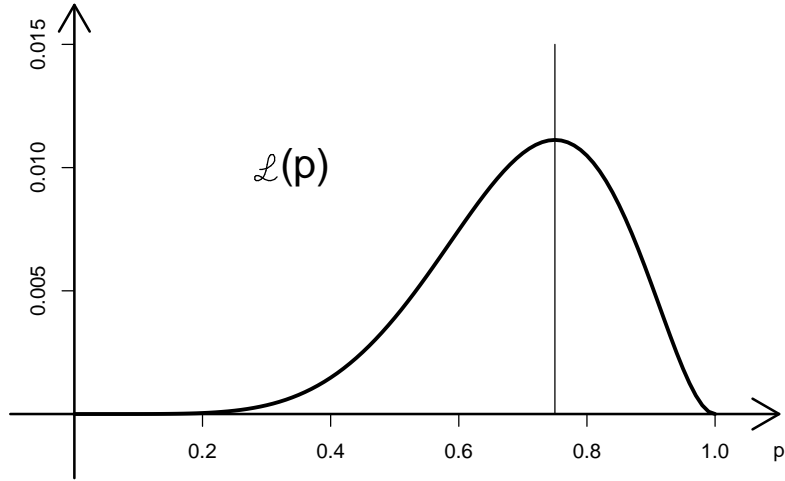
Suppose that the observed data are:

$$x_1 = 0, x_2 = 1, x_3 = 1, x_4 = 1, x_5 = 1, x_6 = 0, x_7 = 1, x_8 = 1.$$

Then the observed value of the likelihood function is

$$\mathcal{L}(p) = p^6 (1-p)^2, \quad 0 < p < 1.$$

The following figure is a graphical illustration of the maximum likelihood procedure in this case. Note that $\hat{p} = 6/8 = 0.75$ is the value of $p \in (0, 1)$ that maximises $\mathcal{L}(p)$.



2.1 Obtaining maximum likelihood estimators

As the previous definition shows, maximum likelihood estimation boils down to the problem of determining where a function reaches its maximum. The mechanics of determination of the maximum differ depending on the smoothness of $\mathcal{L}(\theta)$.

2.1.1 Smooth likelihood functions

Consider estimation of a general parameter θ . If $\mathcal{L}(\theta)$ is smooth then differential calculus methods can be employed to obtain the maximiser of $\mathcal{L}(\theta)$. However, it is usually simpler to work with the log-likelihood function $\ell(\theta)$. Maximising $\ell(\theta)$ rather than $\mathcal{L}(\theta)$ is justified by:

Result

The point at which $\mathcal{L}(\theta)$ attains its maximum over $\theta \in \Theta$ is also that where

$$\ell(\theta) = \ln\{\mathcal{L}(\theta)\} = \sum_{i=1}^n \ln\{f_X(X_i; \theta)\}$$

attains its maximum. Therefore, the maximum likelihood estimator for θ is

$$\hat{\theta} = \theta \text{ that maximises } \ell(\theta) \text{ over } \theta \in \Theta.$$

Example

Re-visit the breakfast eating example, but because of the previous result we now aim to maximise

$$\ell(p) = \ln\{\mathcal{L}(p)\} = \left(\sum_{i=1}^8 X_i\right) \ln(p) + \left(8 - \sum_{i=1}^8 X_i\right) \ln(1-p).$$

The first derivative is

$$\frac{d}{dp} \ell(p) = \frac{\sum_{i=1}^8 X_i}{p} - \frac{8 - \sum_{i=1}^8 X_i}{1-p}$$

and is zero if and only if

$$\frac{\sum_{i=1}^8 X_i}{p} - \frac{8 - \sum_{i=1}^8 X_i}{1-p} = 0 \iff p = \frac{\sum_{i=1}^8 X_i}{8},$$

which is the same answer obtained previously using $\mathcal{L}(p)$, but via simpler calculus.

Is this the unique maximiser of $\ell(p)$ over $p \in (0, 1)$? The second derivative is

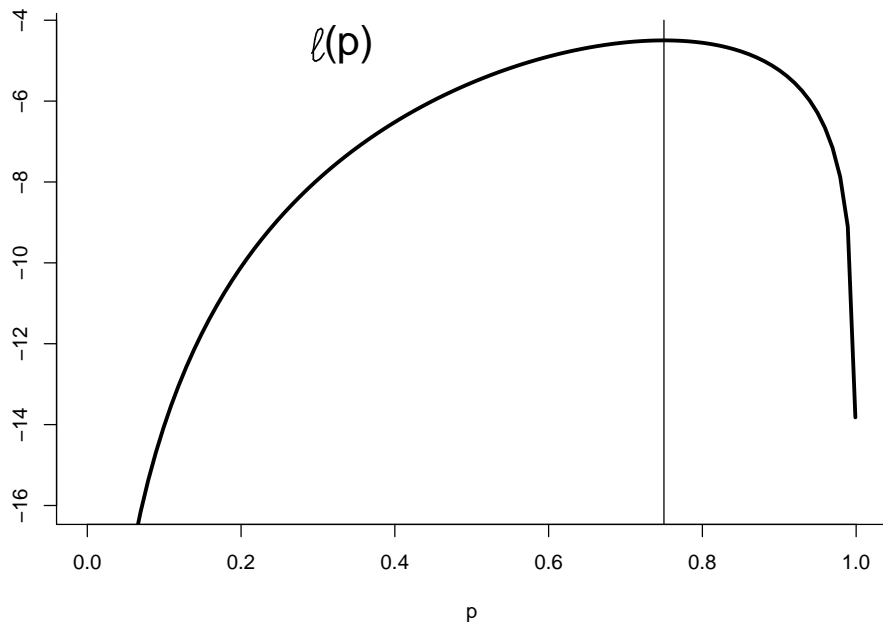
$$\frac{d^2}{dp^2} \ell(p) = -\frac{\sum_{i=1}^8 X_i}{p^2} - \frac{8 - \sum_{i=1}^8 X_i}{(1-p)^2}$$

which is negative for all $0 < p < 1$ and samples $X_i \in \{0, 1\}$, $1 \leq i \leq 8$. Hence $\ell(p)$ is concave (downwards) over $0 < p < 1$ and the point at which $\frac{d}{dp} \ell(p) = 0$ must be a maximum.

For the case where the observed data are again

$$x_1 = 0, x_2 = 1, x_3 = 1, x_4 = 1, x_5 = 1, x_6 = 0, x_7 = 1, x_8 = 1$$

the following figure shows the maximum likelihood estimation procedure via the log-likelihood function $\ell(p)$. As in the case of $\mathcal{L}(p)$ (see previous figure), the maximiser occurs at $\hat{p} = 6/8 = 0.75$.



Example

Consider a random sample X_1, \dots, X_n with common density function:

$$f_X(x; \theta) = 2\theta x e^{-\theta x^2}, \quad x \geq 0; \quad \theta > 0.$$

The log-likelihood function is

$$\ell(\theta) = \sum_{i=1}^n \ln\{f(X_i; \theta)\} = n \ln(2) + n \ln(\theta) + \sum_{i=1}^n \ln(X_i) - \theta \sum_{i=1}^n X_i^2.$$

The first derivative of $\ell(\theta)$ is

$$(\partial/\partial\theta)\ell(\theta) = n/\theta - \sum_{i=1}^n X_i^2$$

and the second derivative of $\ell(\theta)$ is

$$(\partial^2/\partial\theta^2)\ell(\theta) = -n/\theta^2 < 0 \quad \text{for all } \theta > 0.$$

Hence $\ell(\theta)$ is concave (downwards) for all $\theta > 0$ and is therefore maximised at the value for which $(\partial/\partial\theta)\ell(\theta) = 0$ for $\theta > 0$. Solving this for θ we obtain

$$\theta = n / \sum_{i=1}^n X_i^2$$

so the maximum likelihood estimator of θ is

$$\hat{\theta} = n / \sum_{i=1}^n X_i^2.$$

2.1.2 Non-smooth likelihood functions

Not all likelihood functions are differentiable, or even continuous, over $\theta \in \Theta$. In such non-smooth cases calculus methods, alone, cannot be used to locate the maximiser. Also, it is usually better to work directly with $\mathcal{L}(\theta)$ rather than $\ell(\theta)$. The following notation is useful in non-smooth likelihood situations.

Definition

Let \mathcal{P} be a logical condition. Then the **indicator function** of \mathcal{P} , $\mathcal{I}(\mathcal{P})$ is given by

$$\mathcal{I}(\mathcal{P}) = \begin{cases} 1 & \text{if } \mathcal{P} \text{ is true,} \\ 0 & \text{if } \mathcal{P} \text{ is false.} \end{cases}$$

Example

Some examples of use of \mathcal{I} are

$$\mathcal{I}(\text{Morris Iemma was the premier of New South Wales on 5th March, 2007}) = 1.$$

$$\mathcal{I}(4^2 = 16) = 1,$$

$$\mathcal{I}(e^\pi = 17) = 0,$$

$$\mathcal{I}(\text{The Earth is bigger than the Moon}) = 1,$$

$$\mathcal{I}(\text{The Earth is bigger than the Moon \& The Moon is made of blue cheese}) = 0.$$

The \mathcal{I} notation allows one to write density functions in explicit algebraic terms. For example, the Gamma(α, β) density function is usually written as

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

However, it can also be written using \mathcal{I} as

$$f_X(x; \alpha, \beta) = \frac{e^{-x/\beta} x^{\alpha-1}}{\Gamma(\alpha)\beta^\alpha} \mathcal{I}(x > 0).$$

The following result is useful for deriving maximum likelihood estimators when the likelihood function is non-smooth:

Result

For any two logical conditions \mathcal{P} and \mathcal{Q} ,

$$\mathcal{I}(\mathcal{P}\&\mathcal{Q}) = \mathcal{I}(\mathcal{P})\mathcal{I}(\mathcal{Q}).$$

Non-smooth likelihood functions arise when the range of f_X depends on θ .

Example

Suppose that the random sample X_1, \dots, X_n has model

$$f_X(x; \theta) = 5(x^4/\theta^5), \quad 0 < x < \theta.$$

Then, using the \mathcal{I} notation, and the previous result:

$$f_X(x; \theta) = 5(x^4/\theta^5)\mathcal{I}(0 < x < \theta) = 5(x^4/\theta^5)\mathcal{I}(x > 0)\mathcal{I}(\theta > x).$$

The likelihood function is then

$$\begin{aligned} \mathcal{L}(\theta) &= \prod_{i=1}^n f_X(X_i; \theta) \\ &= 5(X_1^4/\theta^5)\mathcal{I}(X_1 > 0)\mathcal{I}(\theta > X_1) \cdots 5(X_n^4/\theta^5)\mathcal{I}(X_n > 0)\mathcal{I}(\theta > X_n) \\ &= 5^n \left(\prod_{i=1}^n X_i \right)^4 \left\{ \prod_{i=1}^n \mathcal{I}(X_i > 0) \right\} \left\{ \prod_{i=1}^n \mathcal{I}(\theta > X_i) \right\} \theta^{-5n} \end{aligned}$$

Note that

$$\prod_{i=1}^n \mathcal{I}(\theta > X_i) = \mathcal{I}(\theta > X_1 \& \theta > X_2 \& \cdots \& \theta > X_n) = \mathcal{I}(\theta > \max(X_1, \dots, X_n)).$$

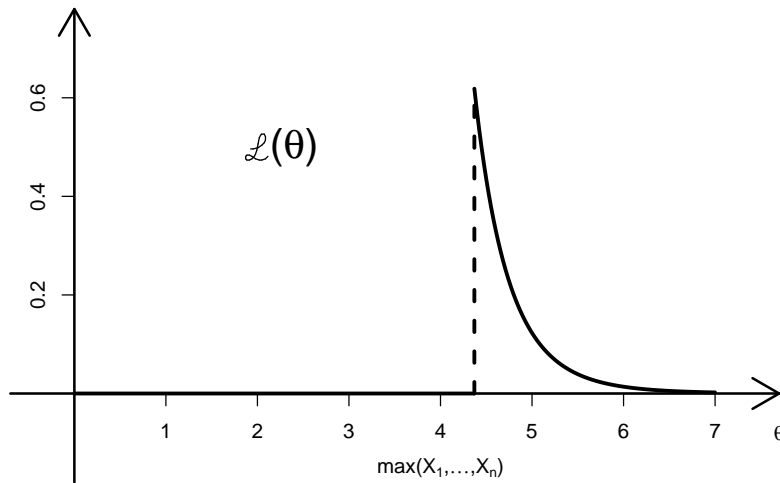
Also, since $P(X_i > 0) = 1$, $\prod_{i=1}^n \mathcal{I}(X_i > 0) = 1$ with probability 1. Hence, the likelihood function is

$$\mathcal{L}(\theta) = 5^n \left(\prod_{i=1}^n X_i \right)^4 \theta^{-5n} \mathcal{I}(\theta > \max(X_1, \dots, X_n))$$

or, even more digestibly,

$$\mathcal{L}(\theta) = \begin{cases} C_n \theta^{-5n}, & \theta > \max(X_1, \dots, X_n) \\ 0, & \text{otherwise} \end{cases}$$

where $C_n = 5^n (\prod_{i=1}^n X_i)^4$. The accompanying figure shows an example of such an $\mathcal{L}(\theta)$.



Since $C_n \theta^{-5n}$ is clearly decreasing for $\theta > \max(X_1, \dots, X_n)$ (easily verified via calculus) it is clear that $\mathcal{L}(\theta)$ attains its maximum at $\max(X_1, \dots, X_n)$. Thus, the maximum likelihood estimator of θ is

$$\hat{\theta} = \max(X_1, \dots, X_n).$$

2.2 Properties of maximum likelihood estimators

2.2.1 Consistency

It can be shown under fairly mild conditions that the maximum likelihood estimator $\hat{\theta}$ of θ is consistent; i.e.

$$\hat{\theta} \xrightarrow{P} \theta.$$

However, the derivation is relatively complicated and will be omitted from these notes. The interested reader is referred to Section 9.5 of:

Wasserman, L. (2004). *All of Statistics*, Springer-Verlag, New York.

2.2.2 Equivariance

Maximum likelihood estimators are *equivariant* under functions of the parameter of interest:

Result

Suppose that $\hat{\theta}$ is the maximum likelihood estimator of θ . Then for any function g

$$g(\hat{\theta}) \text{ is the maximum likelihood estimator of } g(\theta).$$

Example

Let X_1, \dots, X_n be a random sample with model

$$f_X(x; \theta) = 2\theta x e^{-\theta x^2}, \quad x > 0.$$

It had previously been shown that the maximum likelihood estimator of θ is

$$\hat{\theta} = \frac{n}{\sum_{i=1}^n X_i^2}.$$

From the equivariance property of maximum likelihood estimation, the maximum likelihood estimator of $\tau = 1/\theta$ is

$$\hat{\tau} = \frac{1}{n} \sum_{i=1}^n X_i^2$$

and the maximum likelihood estimator of $\omega = \ln(\theta)$ is

$$\hat{\omega} = \ln(n) - \ln\left(\sum_{i=1}^n X_i^2\right).$$

2.2.3 Asymptotic normality

For smooth likelihood functions it is possible to show that the maximum likelihood estimator is asymptotically normal. The asymptotic standard error is closely tied to the **Fisher information**:

Definition

Let X_1, \dots, X_n be a random sample with model

$$f_X(x; \theta), \quad \theta \in \Theta$$

and let

$$\ell(\theta) = \sum_{i=1}^n \ln\{f_X(X_i; \theta)\}$$

be the log-likelihood function, assumed to be smooth. Then the **Fisher information** is

$$I_n(\theta) = -E_{\theta}\{\ell''(\theta)\}.$$

Example

Recall the previous example of a random sample X_1, \dots, X_n with common density function:

$$f_X(x; \theta) = 2\theta x e^{-\theta x^2}, \quad x \geq 0; \theta > 0.$$

The log-likelihood function is

$$\ell(\theta) = n \ln(2) + n \ln(\theta) + \sum_{i=1}^n \ln(X_i) - \theta \sum_{i=1}^n X_i^2.$$

The first derivative of $\ell(\theta)$ is

$$\ell'(\theta) = n\theta^{-1} - \sum_{i=1}^n X_i^2$$

and the second derivative is

$$\ell''(\theta) = -n\theta^{-2}.$$

The Fisher information is then

$$I_n(\theta) = -E_\theta(-n\theta^{-2}) = n/\theta^2.$$

(Note that, in this example, $\ell''(\theta)$ has no random variables present so the expected value operation is redundant. This will not be the case in general.)

Result

$$I_n(\theta) = nI_1(\theta)$$

where

$$I_1(\theta) = -E_\theta \left[\frac{\partial^2}{\partial \theta^2} \ln\{f_X(X; \theta)\} \right]$$

is the Fisher information based on X , a random sample of size $n = 1$.

Example

Recall the previous example of a random sample X_1, \dots, X_n with common density function:

$$f_X(x; \theta) = 2\theta x e^{-\theta x^2}, \quad x \geq 0; \theta > 0.$$

Then for a single random variable X from this density function:

$$\ln\{f_X(X; \theta)\} = \ln(2) + \ln(\theta) + \ln(X) - \theta X^2.$$

The second partial derivative with respect to θ is

$$\frac{\partial^2}{\partial \theta^2} \ln\{f_X(X; \theta)\} = -\theta^{-2}$$

so

$$I_1(\theta) = -E_\theta(-\theta^{-2}) = 1/\theta^2.$$

Then, from the previous result,

$$I_n(\theta) = n/\theta^2$$

as we obtained before.

Theorem: Asymptotic Normality of Maximum Likelihood Estimators

Under appropriate regularity conditions, including the existence of two derivatives of $\mathcal{L}(\theta)$,

$$\frac{\hat{\theta} - \theta}{\widehat{\text{se}}(\hat{\theta})} \xrightarrow{D} N(0, 1)$$

where

$$\widehat{\text{se}}(\hat{\theta}) = \frac{1}{\sqrt{I_n(\hat{\theta})}}$$

is a consistent estimator of $\text{se}(\hat{\theta})$.

This is one of the most important theorems in Statistics. The proof is given in an appendix.

Example

In the previous example with

$$X_1, \dots, X_n \sim f_X(x; \theta) = 2\theta x e^{-\theta x^2}, \quad x \geq 0; \theta > 0$$

the Fisher information was shown to be

$$I_n(\theta) = n/\theta^2.$$

Therefore, the estimated standard error of $\hat{\theta}$ is

$$\widehat{\text{se}}(\hat{\theta}) = \frac{1}{\sqrt{I_n(\hat{\theta})}} = \hat{\theta}/\sqrt{n}$$

and

$$\frac{\hat{\theta} - \theta}{\hat{\theta}/\sqrt{n}} \xrightarrow{D} N(0, 1).$$

The Delta Method permits extension of the asymptotic normality result to a general smooth function of θ :

Result

Under appropriate regularity conditions, including the existence of two derivatives of $\mathcal{L}(\theta)$, if $\tau = g(\theta)$ where g is differentiable and $g'(\theta) \neq 0$ then

$$\frac{\hat{\tau} - \tau}{\widehat{\text{se}}(\hat{\tau})} \xrightarrow{D} N(0, 1)$$

where

$$\widehat{\text{se}}(\hat{\tau}) = \frac{|g'(\hat{\theta})|}{\sqrt{I_n(\hat{\theta})}}$$

is a consistent estimator of $\text{se}(\hat{\tau})$.

Example

In the previous example with

$$X_1, \dots, X_n \sim f_X(x; \theta) = 2\theta x e^{-\theta x^2}, \quad x \geq 0; \theta > 0$$

suppose that the parameter of interest is

$$\omega = \ln(\theta).$$

Then the maximum likelihood estimator of ω is

$$\hat{\omega} = \ln(n) - \ln \left(\sum_{i=1}^n X_i \right).$$

The estimated standard error of $\hat{\omega}$ is

$$\widehat{\text{se}}(\hat{\omega}) = \frac{|1/\hat{\theta}|}{\sqrt{n/\hat{\theta}^2}} = \frac{1}{\sqrt{n}}.$$

Also,

$$\frac{\hat{\omega} - \omega}{1/\sqrt{n}} \xrightarrow{D} N(0, 1).$$

2.2.4 Asymptotic optimality

In the case of smooth likelihood functions where asymptotic normality results can be derived it is possible to argue that, asymptotically, the maximum likelihood estimator is *optimal* or *best*.

Result

Let X_1, \dots, X_n be a random sample with model

$$f_X(x; \theta), \quad \theta \in \Theta$$

and suppose that the maximum likelihood estimator $\hat{\theta}_n$ is asymptotically normal; i.e.

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{D} N(0, I_1(\theta)).$$

Let $\tilde{\theta}_n$ be any other estimator of θ for which

$$\sqrt{n}(\tilde{\theta}_n - \theta) \xrightarrow{D} N(0, \tilde{\sigma}^2).$$

Then $\tilde{\sigma}^2 \geq I_1(\theta)$. Hence, maximum likelihood estimation achieves the lowest possible asymptotic standard error.

Since the bias of maximum likelihood estimators is asymptotically negligible $\hat{\theta}$ has lowest possible asymptotic mean squared error so is asymptotically optimal, or best, among all estimators.

Example

Let

$$X_1, \dots, X_n \sim N(\mu, 1).$$

The maximum likelihood estimator is $\hat{\mu}_n = \bar{X}_n$ which satisfies

$$\sqrt{n}(\hat{\mu}_n - \mu) \xrightarrow{D} N(0, 1).$$

An alternative estimator is the median

$$\tilde{\mu}_n = \begin{cases} \text{middle value of ordered } X_i\text{'s,} & n \text{ odd} \\ \text{average of middle values of ordered } X_i\text{'s,} & n \text{ even.} \end{cases}$$

It can be shown that

$$\sqrt{n}(\tilde{\mu}_n - \mu) \xrightarrow{D} N(0, \frac{1}{2}\pi).$$

Since $\frac{1}{2}\pi \simeq 1.57 > 1$ it is clear that $\hat{\mu}_n$ is asymptotically better than $\tilde{\mu}_n$. Such is the case for *all other* competing estimators of μ .

3 Multiparameter Maximum Likelihood Estimation

In multiparameter models such as

$$X_1, \dots, X_n \sim N(\mu, \sigma^2)$$

and

$$X_1, \dots, X_n \sim \text{Gamma}(\alpha, \beta)$$

the maximum likelihood paradigm still applies. Instead of maximising over a single variable, the maximisation is performed simultaneously over several variables.

Example

Consider the model

$$X_1, \dots, X_n \sim N(\mu, \sigma^2), \quad -\infty < \mu < \infty, \sigma > 0.$$

The log-likelihood function is

$$\ell(\mu, \sigma) = \ln \prod_{i=1}^n \left[\frac{1}{\sqrt{2\pi}\sigma} e^{-(X_i - \mu)^2 / (2\sigma^2)} \right] = -\frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - \mu)^2 - \frac{n}{2} \ln(2\pi) - n \ln(\sigma).$$

Then,

$$\frac{\partial}{\partial \mu} \ell(\mu, \sigma) = \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \mu) = \frac{n}{\sigma^2} (\bar{X} - \mu) = 0$$

if and only if

$$\mu = \bar{X}$$

and regardless of the value of σ . Also

$$\frac{\partial}{\partial \sigma} \ell(\mu, \sigma) = \sigma^{-3} \sum_{i=1}^n (X_i - \mu)^2 - n\sigma^{-1} = 0$$

if and only if

$$\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2}$$

The unique stationary point of $\ell(\mu, \sigma)$ is then

$$(\hat{\mu}, \hat{\sigma}) = \left(\bar{X}, \sqrt{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2} \right).$$

Analysis of the second order partial derivatives can be used to show that this is the global maximiser of $\ell(\mu, \sigma)$ over $\mu \in \mathbb{R}$ and $\sigma > 0$. Hence, the maximum likelihood estimators of μ and σ are

$$\hat{\mu} = \bar{X} \quad \text{and} \quad \hat{\sigma} = \sqrt{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2}.$$

In multiparameter maximum likelihood estimation the extension of Fisher information is as follows:

Definition

Let $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)$ be the vector of parameters in a multiparameter model. The **Fisher information matrix** is given by

$$I_n(\boldsymbol{\theta}) = - \begin{bmatrix} E_{\theta}(H_{11}) & E_{\theta}(H_{12}) & \cdots & E_{\theta}(H_{1k}) \\ E_{\theta}(H_{21}) & E_{\theta}(H_{22}) & \cdots & E_{\theta}(H_{2k}) \\ \vdots & \vdots & \ddots & \vdots \\ E_{\theta}(H_{k1}) & E_{\theta}(H_{k2}) & \cdots & E_{\theta}(H_{kk}) \end{bmatrix}$$

where

$$H_{ij} = \frac{\partial^2}{\partial \theta_i \partial \theta_j} \ell(\boldsymbol{\theta}).$$

Example

Let X_1, \dots, X_n be a random sample from the Gamma(α, β) distribution:

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

The log-likelihood of (α, β) is

$$\begin{aligned} \ell(\alpha, \beta) &= \sum_{i=1}^n \ln\{f_X(X_i; \alpha, \beta)\} \\ &= (\alpha - 1) \sum_{i=1}^n \ln(X_i) - \beta^{-1} \sum_{i=1}^n X_i - n \ln\{\Gamma(\alpha)\} - n\alpha \ln(\beta) \end{aligned}$$

The first order partial derivatives are

$$\begin{aligned} \frac{\partial}{\partial \alpha} \ell(\alpha, \beta) &= \sum_{i=1}^n \ln(X_i) - n \operatorname{digamma}(\alpha) - n \ln(\beta) \\ \text{and } \frac{\partial}{\partial \beta} \ell(\alpha, \beta) &= \beta^{-2} \sum_{i=1}^n X_i - n\alpha \beta^{-1}. \end{aligned}$$

The second order partial derivatives are then

$$\begin{aligned} \frac{\partial^2}{\partial \alpha^2} \ell(\alpha, \beta) &= -n \operatorname{trigamma}(\alpha), \\ \frac{\partial^2}{\partial \alpha \partial \beta} \ell(\alpha, \beta) &= -n/\beta \\ \text{and } \frac{\partial^2}{\partial \beta^2} \ell(\alpha, \beta) &= -2\beta^{-3} \sum_{i=1}^n X_i + n\alpha/\beta^2. \end{aligned}$$

Noting that $E_{\alpha, \beta}(X_i) = \alpha\beta$ for each $1 \leq i \leq n$ we then get

$$\begin{aligned} E_{\alpha, \beta} \left[\frac{\partial^2}{\partial \alpha^2} \ell(\alpha, \beta) \right] &= -n \operatorname{trigamma}(\alpha), \\ E_{\alpha, \beta} \left[\frac{\partial^2}{\partial \alpha \partial \beta} \ell(\alpha, \beta) \right] &= -n/\beta \\ \text{and } E_{\alpha, \beta} \left[\frac{\partial^2}{\partial \beta^2} \ell(\alpha, \beta) \right] &= -2\beta^{-3} n\alpha\beta + n\alpha/\beta^2 = -n\alpha/\beta^2. \end{aligned}$$

The Fisher information matrix is then

$$I_n(\alpha, \beta) = - \begin{bmatrix} -n \operatorname{trigamma}(\alpha) & -n/\beta \\ -n/\beta & -n\alpha/\beta^2 \end{bmatrix} = n \begin{bmatrix} \operatorname{trigamma}(\alpha) & 1/\beta \\ 1/\beta & \alpha/\beta^2 \end{bmatrix}.$$

Definition

Let $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)$ be the vector of parameters in a multiparameter model and $g(\boldsymbol{\theta}) = g(\theta_1, \dots, \theta_k)$ be a real-valued function. The **derivative vector** of g is given by

$$Dg(\boldsymbol{\theta}) = \left[\frac{\partial g(\boldsymbol{\theta})}{\partial \theta_1} \quad \dots \quad \frac{\partial g(\boldsymbol{\theta})}{\partial \theta_k} \right].$$

Example

Consider again the Gamma(α, β) situation of the previous example. Functions of (α, β) are:

$$g(\alpha, \beta) = \alpha^4 \beta^7 \quad \text{and} \quad h(\alpha, \beta) = \cos(3\alpha + e^\beta).$$

Then

$$Dg(\alpha, \beta) = [4\alpha^3 \beta^7 \quad 7\alpha^4 \beta^6]$$

and

$$Dh(\alpha, \beta) = [-3 \sin(3\alpha + e^\beta) \quad -e^\beta \sin(3\alpha + e^\beta)].$$

Result

Let $\tau = g(\boldsymbol{\theta})$ be a real-valued function of $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)$. Under appropriate regularity conditions, including the existence of all second order partial derivatives of $\mathcal{L}(\boldsymbol{\theta})$ and first order partial derivatives of g :

$$\frac{\hat{\tau} - \tau}{\widehat{\text{se}}(\hat{\tau})} \xrightarrow{D} N(0, 1)$$

where

$$\widehat{\text{se}}(\hat{\tau}) = \sqrt{Dg(\hat{\boldsymbol{\theta}}) I_n(\hat{\boldsymbol{\theta}})^{-1} Dg(\hat{\boldsymbol{\theta}})^T}.$$

Example

Let

$$X_1, \dots, X_n \sim N(\mu, \sigma^2)$$

and consider the parameter $\tau = g(\mu, \sigma) = \mu/\sigma$. The maximum likelihood estimator for τ is

$$\hat{\tau} = \frac{\hat{\mu}}{\hat{\sigma}} = \frac{\bar{X}}{\sqrt{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2}}.$$

The Fisher information matrix may be shown to be

$$I_n(\mu, \sigma) = \begin{bmatrix} n/\sigma^2 & 0 \\ 0 & 2n/\sigma^2 \end{bmatrix}$$

which has inverse

$$I_n(\mu, \sigma)^{-1} = \frac{1}{n} \begin{bmatrix} \sigma^2 & 0 \\ 0 & \frac{1}{2}\sigma^2 \end{bmatrix}$$

The derivative vector is

$$Dg(\mu, \sigma) = [1/\sigma \quad -\mu/\sigma^2].$$

Since

$$\begin{aligned} Dg(\mu, \sigma) I_n(\mu, \sigma)^{-1} Dg(\mu, \sigma)^T &= \frac{1}{n} [1/\sigma \quad -\mu/\sigma^2] \begin{bmatrix} \sigma^2 & 0 \\ 0 & \frac{1}{2}\sigma^2 \end{bmatrix} \begin{bmatrix} 1/\sigma \\ -\mu/\sigma^2 \end{bmatrix} \\ &= \frac{1}{n} \left(1 + \frac{\mu^2}{2\sigma^2} \right) \end{aligned}$$

the approximate estimated standard error of $\hat{\tau}$ is

$$\widehat{\text{se}}(\hat{\tau}) = \sqrt{\frac{1}{n} \left(1 + \frac{\bar{X}^2}{\frac{2}{n} \sum_{i=1}^n (X_i - \bar{X})^2} \right)}.$$

The asymptotic normality result for $\hat{\tau}$ is then

$$\frac{\hat{\tau} - \tau}{\sqrt{\frac{1}{n} \left(1 + \frac{\bar{X}^2}{\frac{2}{n} \sum_{i=1}^n (X_i - \bar{X})^2}\right)}} \xrightarrow{D} N(0, 1).$$

4 Likelihood-based Confidence Intervals

For models with smooth likelihood functions the asymptotic normality results of the previous two sections can be used to derive approximate confidence intervals.

Result

Let X_1, \dots, X_n be a random sample with model

$$f_X(x; \theta), \quad \theta \in \Theta$$

and let $\hat{\theta}_n$ be the maximum likelihood estimator of θ . Under the regularity conditions for which θ_n is asymptotically normal

$$\lim_{n \rightarrow \infty} P(\hat{\theta}_n - z_{1-\alpha/2} \widehat{\text{se}}(\hat{\theta}_n) < \theta < \hat{\theta}_n + z_{1-\alpha/2} \widehat{\text{se}}(\hat{\theta}_n)) = 1 - \alpha$$

where $\widehat{\text{se}}(\hat{\theta}_n) = 1/\sqrt{I_n(\hat{\theta}_n)}$. Therefore,

$$(\hat{\theta}_n - z_{1-\alpha/2} \widehat{\text{se}}(\hat{\theta}_n), \hat{\theta}_n + z_{1-\alpha/2} \widehat{\text{se}}(\hat{\theta}_n))$$

is an approximate $1 - \alpha$ confidence interval for θ for large n .

This result follows fairly straightforwardly from the asymptotic normality result

$$\frac{\hat{\theta}_n - \theta}{\widehat{\text{se}}(\hat{\theta}_n)} \xrightarrow{D} N(0, 1).$$

Example

Recall the example of a random sample X_1, \dots, X_n with common density function:

$$f_X(x; \theta) = 2\theta x e^{-\theta x^2}, \quad x \geq 0; \theta > 0.$$

Then the maximum likelihood estimator and corresponding standard error are:

$$\hat{\theta}_n = \frac{n}{\sum_{i=1}^n X_i^2} \quad \text{and} \quad \widehat{\text{se}}(\hat{\theta}_n) = \frac{\hat{\theta}_n}{\sqrt{n}} = \frac{\sqrt{n}}{\sum_{i=1}^n X_i^2}.$$

For a 95% confidence interval the appropriate $N(0, 1)$ quantile is

$$z_{0.975} = 1.96.$$

An approximate 95% confidence interval for θ is then:

$$\left(\frac{n}{\sum_{i=1}^n X_i^2} - 1.96 \frac{\sqrt{n}}{\sum_{i=1}^n X_i^2}, \frac{n}{\sum_{i=1}^n X_i^2} + 1.96 \frac{\sqrt{n}}{\sum_{i=1}^n X_i^2} \right).$$

To illustrate the use of this result, consider the following sample of size $n = 100$ which we will assume is from a member of $f_X(x; \theta)$, $\theta > 0$:

0.366 0.568 0.300 0.115 0.204 0.128 0.277 0.391 0.328 0.451
0.412 0.190 0.207 0.147 0.116 0.326 0.256 0.524 0.217 0.485
0.265 0.375 0.267 0.360 0.250 0.258 0.583 0.413 0.481 0.468
0.406 0.336 0.305 0.321 0.268 0.361 0.632 0.283 0.258 0.466
0.276 0.232 0.133 0.316 0.468 0.496 0.573 0.523 0.256 0.491
0.127 0.054 0.440 0.228 0.249 0.754 0.430 0.111 0.459 0.233
0.257 0.640 0.147 0.273 0.112 0.389 0.126 0.356 0.273 0.296
0.433 0.253 0.234 0.514 0.177 0.221 0.534 0.509 0.510 0.269
0.262 0.625 0.183 0.541 0.705 0.078 0.847 0.149 0.031 0.453
0.299 0.226 0.069 0.211 0.195 0.381 0.317 0.467 0.289 0.593

For these data (x_1, \dots, x_{100}) we have

$$\sum_{i=1}^{100} x_i^2 = 14.018.$$

So an approximate 95% confidence interval for θ is

$$\left(\frac{100}{14.018} - 1.96 \frac{\sqrt{100}}{14.018}, \frac{100}{14.018} + 1.96 \frac{\sqrt{100}}{14.018} \right) = (5.17, 9.09).$$

Result

Under the same conditions as the previous result and with $\tau = g(\theta)$

$$\lim_{n \rightarrow \infty} P(\widehat{\tau}_n - z_{1-\alpha/2} \widehat{\text{se}}(\widehat{\tau}_n) < \tau < \widehat{\tau}_n + z_{1-\alpha/2} \widehat{\text{se}}(\widehat{\tau}_n)) = 1 - \alpha$$

where $\widehat{\text{se}}(\widehat{\tau}_n) = |g'(\widehat{\theta}_n)| / \sqrt{I_n(\widehat{\theta}_n)}$. Therefore,

$$(\widehat{\tau}_n - z_{1-\alpha/2} \widehat{\text{se}}(\widehat{\tau}_n), \widehat{\tau}_n + z_{1-\alpha/2} \widehat{\text{se}}(\widehat{\tau}_n))$$

is an approximate $1 - \alpha$ confidence interval for τ for large n .

4.1 Extension to Multiparameter Models

Suppose that the random sample X_1, \dots, X_n is modelled as coming from a multiparameter model:

$$f_X(x; \boldsymbol{\theta}) = f_X(x; \theta_1, \dots, \theta_k).$$

Then we can use a previous asymptotic normality result to obtain an approximate confidence interval for

$$\tau = g(\boldsymbol{\theta}) = g(\theta_1, \dots, \theta_k).$$

Result

Under appropriate regularity conditions, including the existence of all second order partial derivatives of $\mathcal{L}(\theta)$, if $\tau = g(\theta)$ where each component of g is differentiable then:

$$\lim_{n \rightarrow \infty} P(\hat{\tau}_n - z_{1-\alpha/2} \widehat{\text{se}}(\hat{\tau}_n) < \tau < \hat{\tau}_n + z_{1-\alpha/2} \widehat{\text{se}}(\hat{\tau}_n)) = 1 - \alpha$$

where

$$\widehat{\text{se}}(\hat{\tau}_n) = \sqrt{\text{D}g(\hat{\theta}) I_n(\hat{\theta})^{-1} \text{D}g(\hat{\theta})^T}.$$

Therefore,

$$(\hat{\tau}_n - z_{1-\alpha/2} \widehat{\text{se}}(\hat{\tau}_n), \hat{\tau}_n + z_{1-\alpha/2} \widehat{\text{se}}(\hat{\tau}_n))$$

is an approximate $1 - \alpha$ confidence interval for τ for large n .

Example

Let

$$X_1, \dots, X_n \sim N(\mu, \sigma^2)$$

and consider the parameter $\tau = g(\mu, \sigma) = \mu/\sigma$. As shown previously, the maximum likelihood estimator for τ is

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

and the approximate estimated standard error of $\hat{\tau}$ is

$$\widehat{\text{se}}(\hat{\tau}) = \sqrt{\text{D}g(\hat{\theta}) I_n(\hat{\theta})^{-1} \text{D}g(\hat{\theta})^T} = \sqrt{\frac{1}{n} \left(1 + \frac{\bar{X}^2}{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2} \right)}.$$

Suppose a 99% confidence interval for τ is sought. The appropriate quantile from the $N(0, 1)$ distribution is

$$z_{0.995} = 2.576.$$

An approximate 99% confidence interval for μ/σ is then

$$\left(\frac{\bar{X}}{\sqrt{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2}} - 2.576 \sqrt{\frac{1}{n} \left(1 + \frac{\bar{X}^2}{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2} \right)}, \right. \\ \left. \frac{\bar{X}}{\sqrt{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2}} + 2.576 \sqrt{\frac{1}{n} \left(1 + \frac{\bar{X}^2}{\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2} \right)} \right).$$

Appendix: Proof of Theorem on Asymptotic Normality of Maximum Likelihood Estimators

Lemma:

Under some regularity conditions

$$E_{\theta} \left[\frac{\partial}{\partial \theta} \ln \{f_X(X; \theta)\} \right] = 0.$$

and

$$\text{Var}_{\theta} \left[\frac{\partial}{\partial \theta} \ln f_X(X; \theta) \right] = I_1(\theta).$$

Proof of Lemma:

Since $f_X(x; \theta)$ is a density function we have

$$1 = \int_{-\infty}^{\infty} f_X(x; \theta) dx.$$

Differentiation of both sides with respect to θ leads to

$$\begin{aligned} 0 &= \frac{\partial}{\partial \theta} \int_{-\infty}^{\infty} f_X(x; \theta) dx = \int_{-\infty}^{\infty} \frac{\partial}{\partial \theta} f_X(x; \theta) dx \\ &= \int_{-\infty}^{\infty} \frac{\frac{\partial}{\partial \theta} f_X(x; \theta)}{f_X(x; \theta)} \cdot f_X(x; \theta) dx \\ &= \int_{-\infty}^{\infty} \frac{\partial \ln f_X(x; \theta)}{\partial \theta} \cdot f_X(x; \theta) dx = E_{\theta} \left[\frac{\partial}{\partial \theta} \ln f_X(X; \theta) \right]. \end{aligned}$$

Differentiation of the above equation: $0 = \int_{-\infty}^{\infty} \frac{\partial \ln f_X(x; \theta)}{\partial \theta} \cdot f_X(x; \theta) dx$ with respect to θ leads to

$$0 = \int_{-\infty}^{\infty} \frac{\partial^2 \ln\{f_X(x; \theta)\}}{\partial \theta^2} dx + \int_{-\infty}^{\infty} \frac{\partial \ln f_X(x; \theta)}{\partial \theta} \frac{\partial \ln f_X(x; \theta)}{\partial \theta} \cdot f_X(x; \theta) dx$$

which is equivalent to

$$0 = -I_1(\theta) + E_{\theta} \left[\left\{ \frac{\partial \ln f_X(x; \theta)}{\partial \theta} \right\}^2 \right]$$

But since $E_{\theta} \left[\frac{\partial}{\partial \theta} \ln f_X(X; \theta) \right] = 0$ the previous displayed equation is actually:

$$0 = -I_1(\theta) + \text{Var}_{\theta} \left[\frac{\partial}{\partial \theta} \ln f_X(X; \theta) \right]$$

and the required result follows.

Proof of Theorem:

Let

$$\ell'(\theta) = \frac{\partial}{\partial \theta} \ell(\theta) \quad \text{and} \quad \ell''(\theta) = \frac{\partial^2}{\partial \theta^2} \ell(\theta).$$

The maximum likelihood estimator $\hat{\theta}_n$ satisfies

$$0 = \ell'(\hat{\theta}_n) = \ell'(\theta + \hat{\theta}_n - \theta) = \ell'(\theta) + (\hat{\theta}_n - \theta)\ell''(\theta) + \dots$$

Since $\hat{\theta}_n \xrightarrow{P} \theta$ we are justified in ignoring lower order terms and working with the approximation

$$\hat{\theta}_n - \theta \simeq -\ell'(\theta)/\ell''(\theta).$$

With some minor rearrangement, this may be rewritten as

$$\sqrt{n}(\hat{\theta}_n - \theta) \simeq \frac{N_n}{D_n}$$

where the numerator is

$$N_n = \frac{1}{\sqrt{nI_1(\theta)}} \sum_{i=1}^n N_i \quad \text{with} \quad N_i = \frac{\partial}{\partial \theta} \ln\{f_X(X_i; \theta)\}$$

and the denominator is

$$D_n = \frac{1}{n} \sum_{i=1}^n D_i \quad \text{with} \quad D_i = -\frac{\frac{\partial^2}{\partial \theta^2} \ln\{f_X(X_i; \theta)\}}{\sqrt{I_1(\theta)}}.$$

From the Lemma,

$$E_\theta(N_i) = 0 \quad \text{and} \quad \text{Var}_\theta(N_i) = I_1(\theta).$$

Application of the Central Limit Theorem to N_n then leads to

$$N_n = \frac{\frac{1}{n} \sum_{i=1}^n N_i - E_\theta(N_i)}{\sqrt{\text{Var}(N_i)/n}} \xrightarrow{D} N(0, 1).$$

Application of the Weak Law of Large Numbers to the denominator leads to

$$D_n \xrightarrow{P} E_\theta(D_i) = -E_\theta \left[\frac{\frac{\partial^2}{\partial \theta^2} \ln\{f_X(X_i; \theta)\}}{\sqrt{I_1(\theta)}} \right] = \frac{I_1(\theta)}{\sqrt{I_1(\theta)}} = \sqrt{I_1(\theta)}.$$

Application of Slutsky's Theorem leads to

$$\frac{N_n}{D_n} \xrightarrow{D} \frac{N(0, 1)}{\sqrt{I_1(\theta)}}$$

which, in turn, leads to

$$\frac{\hat{\theta}_n - \theta}{1/\sqrt{nI_1(\theta)}} \xrightarrow{D} N(0, 1).$$

Since $I_n(\theta) = nI_1(\theta)$

$$\frac{\hat{\theta}_n - \theta}{1/\sqrt{I_n(\theta)}} \xrightarrow{D} N(0, 1).$$

Another application of Slutsky's Theorem gives

$$\frac{\hat{\theta}_n - \theta}{1/\sqrt{I_n(\hat{\theta}_n)}} \xrightarrow{D} N(0, 1).$$

That is,

$$\frac{\hat{\theta}_n - \theta}{\widehat{\text{se}}(\hat{\theta}_n)} \xrightarrow{D} N(0, 1).$$

where $\widehat{\text{se}}(\hat{\theta}_n) = 1/\sqrt{I_n(\hat{\theta}_n)}$.