Introduction to General and Generalized Linear Models

The Likelihood Principle - part I

Henrik Madsen
Poul Thyregod

DTU Informatics
Technical University of Denmark
DK-2800 Kgs. Lyngby

January 2012

This lecture

- The likelihood principle
- Point estimation theory
- The likelihood function
- The score function
- The information matrix

The beginning of likelihood theory

- Fisher (1922) identified the likelihood function as the key inferential quantity conveying all inferential information in statistical modelling including the uncertainty
- The Fisherian school offers a Bayesian-frequentist compromise

A motivating example

Suppose we toss a thumbtack (used to fasten up documents to a background) 10 times and observe that 3 times it lands point up. Assuming we know nothing prior to the experiment, what is the probability of landing point up, θ?

- Binomial experiment with $y = 3$ and $n = 10$.
- P(Y=3;10,3,0.2) = 0.2013
- P(Y=3;10,3,0.3) = 0.2668
- P(Y=3;10,3,0.4) = 0.2150

A motivating example

By considering $\mathsf{P}_\theta(Y = 3)$ to be a function of the unknown parameter we have the *likelihood function*:

$$L(\theta) = \mathsf{P}_\theta(Y = 3)$$

In general, in a Binomial experiment with n trials and y successes, the likelihood function is:

$$L(\theta) = \mathsf{P}_\theta(Y = y) = \left(\begin{array}{c} n \\ y \end{array} \right) \theta^y (1-\theta)^{n-y}$$

A motivating example

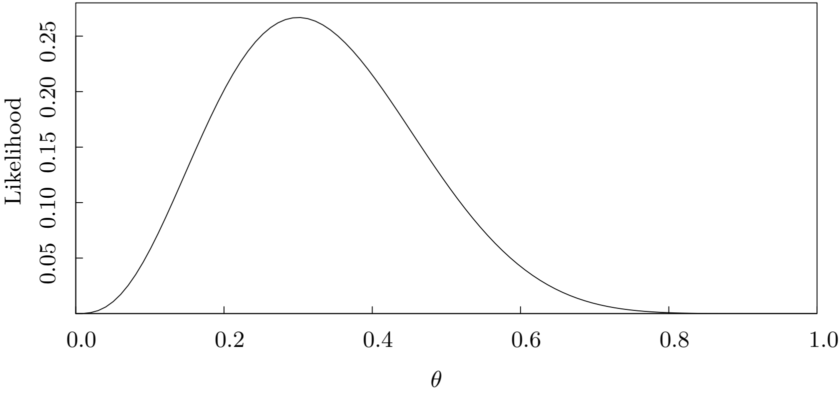


Figure: Likelihood function of the success probability θ in a binomial experiment with $n = 10$ and $y = 3$.

A motivating example

It is often more convenient to consider the log-likelihood function. The log-likelihood function is:

$$\log L(\theta) = y \log \theta + (n - y) \log(1 - \theta) + const$$

where $const$ indicates a term that does not depend on θ.
By solving

$$\frac{\partial \log L(\theta)}{\partial \theta} = 0$$

it is readily seen that the maximum likelihood *estimate* (MLE) for θ is

$$\widehat{\theta}(y) = \frac{y}{n} = \frac{3}{10} = 0.3$$

The likelihood principle

- Not just a method for obtaining a point estimate of parameters.
- It is the entire likelihood function that captures all the information in the data about a certain parameter.
- Likelihood based methods are inherently computational. In general numerical methods are needed to find the MLE.
- Today the likelihood principles play a central role in statistical modelling and inference.

Some syntax

- Multivariate random variable: $\boldsymbol{Y} = \{ Y_1, Y_2, ..., Y_n\}^T$
- Observation set: $\{\boldsymbol{y} = y_1, y_2, \ldots, y_n\}^T$
- Joint density: $\{f_{\mathbf{Y}}(y_1, y_2, \ldots, y_n; \boldsymbol{\theta})\}_{\boldsymbol{\theta} \in \Theta^k}$
- Estimator (random) $\widehat{\boldsymbol{\theta}}(\mathbf{Y})$
- Estimate (number/vector) $\widehat{\boldsymbol{\theta}}(\mathbf{y})$

Point estimation theory

We will assume that the statistical model for $\boldsymbol{y}$ is given by parametric family of joint densities:

$$\{f_{\mathbf{Y}}(y_1, y_2, \ldots, y_n; \boldsymbol{\theta})\}_{\boldsymbol{\theta} \in \Theta^k}$$

Remember that when the n random variables are independent, the joint probability density equals the product of the corresponding marginal densities or:

$$f(y_1, y_2, ...y_n) = f_1(y_1) \cdot f_2(y_2) \cdot \ldots \cdot f_n(y_n)$$

Point estimation theory

Definition (Unbiased estimator)

Any estimator $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(\boldsymbol{Y})$ is said to be *unbiased* if

$$\mathsf{E}[\widehat{\boldsymbol{\theta}}] = \boldsymbol{\theta}$$

for all $\boldsymbol{\theta} \in \Theta^k$.

Definition (Minimum mean square error)

An estimator $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(\boldsymbol{Y})$ is said to be *uniformly minimum mean square error* if

$$\mathsf{E}\left[(\widehat{\boldsymbol{\theta}}(\boldsymbol{Y}) - \boldsymbol{\theta})(\widehat{\boldsymbol{\theta}}(\boldsymbol{Y}) - \boldsymbol{\theta})^T\right] \leq \mathsf{E}\left[(\tilde{\boldsymbol{\theta}}(\boldsymbol{Y}) - \boldsymbol{\theta})(\tilde{\boldsymbol{\theta}}(\boldsymbol{Y}) - \boldsymbol{\theta})^T\right]$$

for all $\boldsymbol{\theta} \in \Theta^k$ and all other estimators $\tilde{\boldsymbol{\theta}}(\boldsymbol{Y})$.

Point estimation theory

- By considering the class of unbiased estimators it is most often not possible to establish a suitable estimator.
- We need to add a criterion on the variance of the estimator.
- A low variance is desired, and in order to evaluate the variance a suitable lower bound is given by the Cramer-Rao inequality.

Point estimation theory

Theorem (Cramer-Rao inequality)

Given the parametric density $f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta^k$*, for the observations* $\boldsymbol{Y}$*. Subject to certain regularity conditions, the variance of any unbiased estimator* $\widehat{\boldsymbol{\theta}}(\boldsymbol{Y})$ *of* $\boldsymbol{\theta}$ *satisfies the inequality*

$$\mathrm{Var}\left[\widehat{\boldsymbol{\theta}}(\boldsymbol{Y})\right] \;\geq\; \boldsymbol{i}^{-1}(\boldsymbol{\theta})$$

where $\boldsymbol{i}(\boldsymbol{\theta})$ *is the Fisher information matrix defined by*

$$\boldsymbol{i}(\boldsymbol{\theta}) = \mathrm{E}\left[\left(\frac{\partial \log f_Y(\boldsymbol{Y};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial \log f_Y(\boldsymbol{Y};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)^T\right]$$

and $\mathrm{Var}\left[\widehat{\boldsymbol{\theta}}(\boldsymbol{Y})\right] = \mathrm{E}\left[(\widehat{\boldsymbol{\theta}}(\boldsymbol{Y}) - \boldsymbol{\theta})(\widehat{\boldsymbol{\theta}}(\boldsymbol{Y}) - \boldsymbol{\theta})^T\right].$

Point estimation theory

Definition (Efficient estimator)

An unbiased estimator is said to be *efficient* if its covariance is equal to the Cramer-Rao lower bound.

Dispersion matrix

The matrix $\operatorname{Var}\left[\widehat{\boldsymbol{\theta}}(\boldsymbol{Y})\right]$ is often called a variance covariance matrix since it contains variances in the diagonal and covariances outside the diagonal. This important matrix is often termed the *Dispersion matrix*.

The likelihood function

- The likelihood function is built on an assumed parameterized statistical model as specified by a parametric family of joint densities for the observations $\boldsymbol{Y} = (Y_1, Y_2, ..., Y_n)^T$.
- The *likelihood* of any specific value $\boldsymbol{\theta}$ of the parameters in a model is (proportional to) the probability of the actual outcome, $Y_1 = y_1, Y_2 = y_2, ..., Y_n = y_n$, calculated for the specific value $\boldsymbol{\theta}$.
- The likelihood function is simply obtained by considering the likelihood as a function of $\boldsymbol{\theta} \in \Theta^k$.

The likelihood function

Definition (Likelihood function)

Given the parametric density $f_{\boldsymbol{Y}}(\boldsymbol{y}, \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta^P$, for the observations $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$ the *likelihood function for* $\boldsymbol{\theta}$ is the function

$$L(\boldsymbol{\theta}; \boldsymbol{y}) = c(y_1, y_2, \ldots, y_n) f_Y(y_1, y_2, \ldots, y_n; \boldsymbol{\theta})$$

where $c(y_1, y_2, \ldots, y_n)$ is a constant.

The likelihood function is thus (proportional to) the joint probability density for the actual observations considered as a function of $\boldsymbol{\theta}$.

The log-likelihood function

- Very often it is more convenient to consider the *log-likelihood* function defined as

$$l(\boldsymbol{\theta}; \boldsymbol{y}) = \log(L(\boldsymbol{\theta}; \boldsymbol{y})).$$

- Sometimes the likelihood and the log-likelihood function will be written as $L(\boldsymbol{\theta})$ and $l(\boldsymbol{\theta})$, respectively, i.e. the dependency on $\boldsymbol{y}$ is suppressed.

Example: Likelihood function for mean of normal distribution

An automatic production of a bottled liquid is considered to be stable. A sample of three bottles were selected at random from the production and the volume of the content volume was measured. The deviation from the nominal volume of 700.0 ml was recorded.

The deviations (in ml) were 4.6; 6.3; and 5.0.

Example: Likelihood function for mean of normal distribution

First a *model* is formulated

i Model: C+E (center plus error) model, $Y = \mu + \epsilon$

ii Data: $Y_i = \mu + \epsilon_i$

iii Assumptions:

- Y_1, Y_2, Y_3 are independent
- $Y_i \sim \mathrm{N}(\mu, \sigma^2)$
- σ^2 is known, $\sigma^2 = 1$,

Thus, there is only one unknown model parameter, $\mu_Y = \mu$.

Example: Likelihood function for mean of normal distribution

The joint probability density function for Y_1, Y_2, Y_3 is given by

$$\begin{aligned} f_{Y_1,Y_2,Y_3}(y_1,y_2,y_3;\mu) &= \frac{1}{\sqrt{2\pi}}\ \exp\left[-\frac{(y_1-\mu)^2}{2}\right] \\ &\times \frac{1}{\sqrt{2\pi}}\ \exp\left[-\frac{(y_2-\mu)^2}{2}\right] \\ &\times \frac{1}{\sqrt{2\pi}}\ \exp\left[-\frac{(y_3-\mu)^2}{2}\right] \end{aligned}$$

which for every value of μ is a function of the three variables y_1, y_2, y_3.

Remember that the normal probability density is: $f(y;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\ \exp\left[-\frac{(y-\mu)^2}{2\sigma^2}\right]$

Example: Likelihood function for mean of normal distribution

Now, we have the observations, $y_1 = 4.6$; $y_2 = 6.3$ and $y_3 = 5.0$, and establish the likelihood function

$$\begin{aligned} L_{4.6,6.3,5.0}(\mu) &= f_{Y_1,Y_2,Y_3}(4.6, 6.3, 5.0; \mu) \\ &= \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{(4.6-\mu)^2}{2}\right] \\ &\times \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{(6.3-\mu)^2}{2}\right] \\ &\times \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{(5.0-\mu)^2}{2}\right] \end{aligned}$$

The function depends only on μ.

Note that the likelihood function expresses the infinitesimal probability of obtaining the sample result $(4.6, 6.3, 5.0)$ as a function of the unknown parameter μ.

Example: Likelihood function for mean of normal distribution

Reducing the expression one finds

$$\begin{aligned} L_{4.6,6.3,5.0}(\mu) &= \frac{1}{(\sqrt{2\pi})^3} \exp\left[-\frac{1.58}{2}\right] \exp\left[-\frac{3(5.3-\mu)^2}{2}\right] \\ &= \frac{1}{(\sqrt{2\pi})^3} \exp\left[-\frac{1.58}{2}\right] \exp\left[-\frac{3(\bar{y}-\mu)^2}{2}\right] \end{aligned}$$

which shows that (except for a factor not depending on μ), the likelihood function does only depend on the observations (y_1, y_2, y_3) through the average $\bar{y} = \sum y_i/3$.

Example: Likelihood function for mean of normal distribution

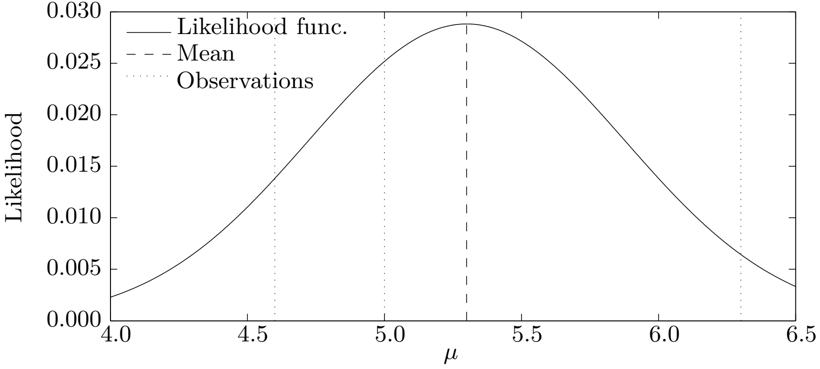


Figure: The likelihood function for μ given the observations $y_1 = 4.6$; $y_2 = 6.3$ and $y_3 = 5.0$.

Sufficient statistic

- The primary goal in analysing observations is to characterise the information in the observations by a few numbers.
- A *statistics* $t(Y_1, Y_2, \ldots, Y_n)$ is a function of the observations.
- In estimation a sufficient statistic is a statistic than contains all the information in the observations.

Sufficient statistic

Definition (Sufficient statistic)

A (possibly vector-valued) function $t(Y_1, Y_2, \ldots, Y_n)$ is said to be a *sufficient statistic* for a (possibly vector-valued) parameter, θ, if the probability density function for $t(Y_1, Y_2, \ldots, Y_n)$ can be factorized into a product

$$f_{Y_1,\ldots,Y_n}(y_1,\ldots,y_n;\theta) = h(y_1,\ldots,y_n)g(t(y_1,y_2,\ldots,y_n);\theta)$$

with the factor $h(y_1,\ldots,y_n)$ not depending on the parameter θ, and the factor $g(t(y_1,y_2,\ldots,y_n);\theta)$ only depending on $y_1,\ldots,y_n$ through the function $t(\cdot,\cdot,\ldots,\cdot)$. Thus, if we know the value of $t(y_1,y_2,\ldots,y_n)$, the individual values $y_1,\ldots,y_n$ do not contain further information about the value of θ.

Roughly speaking, a statistic is sufficient if we are able to calculate the likelihood function (apart from a factor) only knowing $t(Y_1, Y_2, \ldots, Y_n)$.

The Score function

Definition (Score function)

Consider $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_k) \in \Theta^k$, and assume that Θ^k is an open subspace of $\mathbb{R}^k$, and that the log-likelihood is continuously differentiable. Then consider the first order partial derivative (gradient) of the log-likelihood function:

$$l'_\theta(\boldsymbol{\theta}; \boldsymbol{y}) = \frac{\partial}{\partial \boldsymbol{\theta}}\, l(\boldsymbol{\theta}; \boldsymbol{y}) = \begin{pmatrix} \dfrac{\partial}{\partial \theta_1}\, l(\boldsymbol{\theta}; \boldsymbol{y}) \\ \vdots \\ \dfrac{\partial}{\partial \theta_k}\, l(\boldsymbol{\theta}; \boldsymbol{y}) \end{pmatrix}$$

The function $l'_\theta(\boldsymbol{\theta}; \boldsymbol{y})$ is called the *score function* often written as $S(\boldsymbol{\theta}; \boldsymbol{y})$.

The Score function

Theorem

Under normal regularity conditions

$$\mathrm{E}_\theta\left[\frac{\partial}{\partial\boldsymbol{\theta}} l(\boldsymbol{\theta};\,\boldsymbol{Y})\right] = \mathbf{0}$$

This follows by differentiation of

$$\int f_Y(\boldsymbol{y};\boldsymbol{\theta})\ \mu\{dy\} = 1$$

The information matrix

Definition (Observed information)

The matrix

$$\boldsymbol{j}(\boldsymbol{\theta};\boldsymbol{y}) = -\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T} l(\boldsymbol{\theta};\boldsymbol{y})$$

with the elements

$$\boldsymbol{j}(\boldsymbol{\theta};\boldsymbol{y})_{ij} = -\frac{\partial^2}{\partial\theta_i\partial\theta_j} l(\boldsymbol{\theta};\boldsymbol{y})$$

is called the *observed information* corresponding to the observation $\boldsymbol{y}$, evaluated in $\widehat{\boldsymbol{\theta}}$.

The observed information is thus equal to the Hessian (with opposite sign) of the log-likelihood function evaluated at $\boldsymbol{\theta}$. The Hessian matrix is simply (with opposite sign) the *curvature* of the log-likelihood function.

The information matrix

Definition (Expected information)

The expectation of the observed information

$$\boldsymbol{i}(\boldsymbol{\theta}) = \mathsf{E}[\boldsymbol{j}(\boldsymbol{\theta}; \boldsymbol{Y})],$$

where the expectation is determined under the distribution corresponding to θ, is called the *expected information*, or the *information matrix* corresponding to the parameter $\boldsymbol{\theta}$. The expected information is also known as the *Fisher information matrix*

Fisher Information Matrix

Fisher Information Matrix

The expected information or Fisher Information Matrix is equal to the dispersion matrix for the score function, i.e.

$$\begin{aligned} \boldsymbol{i}(\boldsymbol{\theta}) &= \mathrm{E}_\theta\left[-\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T} l(\boldsymbol{\theta}; \boldsymbol{Y})\right] \\ &= \mathrm{E}_\theta\left[\frac{\partial}{\partial\boldsymbol{\theta}} l(\boldsymbol{\theta}; \boldsymbol{Y})\left(\frac{\partial}{\partial\theta} l(\boldsymbol{\theta}; \boldsymbol{Y})\right)^T\right] \\ &= D_\theta[l'_\theta(\boldsymbol{\theta}; \boldsymbol{Y})] \end{aligned}$$

where $D[\cdot]$ denotes the dispersion matrix.

In estimation the information matrix provides a measure for the accuracy obtained in determining the parameters.

Example: Score function, Observed and Expected Information

Consider again the production of a bottled liquid example from slide 18.

The log-likelihood function is:

$$l(\mu; 4.6, 6.3, 5.0) = -\frac{3(5.3-\mu)^2}{2} + C(4.6, 6.3, 5.0)$$

and hence the score function is

$$l'_\mu(\mu; 4.6, 6.3, 5.0) = 3 \cdot (5.3 - \mu),$$

with the observed information

$$\boldsymbol{j}(\mu; 4.6, 6.3, 5.0) = 3.$$

Example: Score function, Observed and Expected Information

In order to determine the expected information it is necessary to perform analogous calculations substituting the data by the corresponding random variables Y_1, Y_2, Y_3.

The likelihood function can be written as

$$L_{y_1,y_2,y_3}(\mu) = \frac{1}{(\sqrt{2\pi})^3} \exp\left[-\frac{\sum(y_i - \bar{y})^2}{2}\right] \exp\left[-\frac{3(\bar{y} - \mu)^2}{2}\right].$$

Example: Score function, Observed and Expected Information

Introducing the random variables (Y_1, Y_2, Y_3) instead of (y_1, y_2, y_3) and taking logarithms one finds

$$l(\mu; Y_1, Y_2, Y_3) = -\frac{3(\overline{Y} - \mu)^2}{2} - 3\ln(\sqrt{2\pi}) - \frac{\sum (Y_i - \overline{Y})^2}{2},$$

and hence the score function is

$$l'_\mu(\mu; Y_1, Y_2, Y_3) = 3(\overline{Y} - \mu),$$

and the observed information

$$\boldsymbol{j}(\mu; Y_1, Y_2, Y_3) = 3.$$

Example: Score function, Observed and Expected Information

It is seen in this (Gaussian) case that the observed information (curvature of log likelihood function) does not depend on the observations Y_1, Y_2, Y_3, and hence the expected information is

$$\boldsymbol{i}(\mu) = \mathsf{E}[\boldsymbol{j}(\mu; Y_1, Y_2, Y_3)] = 3.$$

Alternative parameterizations of the likelihood

Definition (The likelihood function for alternative parameterizations)

The likelihood function depends not on the actual parameterization. Let $\boldsymbol{\psi} = \boldsymbol{\psi}(\boldsymbol{\theta})$ denote a one-to-one mapping of $\Omega \subset \mathbb{R}^k$ onto $\Psi \subset \mathbb{R}^k$. The parameterization given by $\boldsymbol{\psi}$ is just an alternative parameterization of the model. The likelihood and log-likelihood function for the parameterization given by $\boldsymbol{\psi}$ is

$$L_\Psi(\boldsymbol{\psi}; \boldsymbol{y}) = L_\Omega(\boldsymbol{\theta}(\psi); \boldsymbol{y})$$
$$l_\Psi(\boldsymbol{\psi}; \boldsymbol{y}) = l_\Omega(\boldsymbol{\theta}(\psi); \boldsymbol{y})$$

This gives rise to the very useful invariance property.

- The likelihood is thus *not* a joint probability density on Ω, since then the Jacobian should have been used
- However, the score function and the information matrix depends in general on the parameterization.

The Maximum Likelihood Estimate (MLE)

The *score function* can be used to obtain the estimate, since the MLE can be found as the solution to

$$l'_\theta(\boldsymbol{\theta}; \boldsymbol{y}) = 0$$

which are called the *estimation equations for the ML-estimator*, or, just the ML equations.

- It is common practice, especially when plotting, to normalize the likelihood function to have unit maximum and the log-likelihood to have zero maximum.

Invariance property

Theorem (Invariance property)

Assume that $\widehat{\boldsymbol{\theta}}$ *is a maximum likelihood estimator for* $\boldsymbol{\theta}$*, and let* $\boldsymbol{\psi} = \boldsymbol{\psi}(\boldsymbol{\theta})$ *denote a one-to-one mapping of* $\Omega \subset \mathbb{R}^k$ *onto* $\Psi \subset \mathbb{R}^k$*. Then the estimator* $\boldsymbol{\psi}(\widehat{\boldsymbol{\theta}})$ *is a maximum likelihood estimator for the parameter* $\boldsymbol{\psi}(\boldsymbol{\theta})$*.*

The principle is easily generalized to the case where the mapping is not one-to-one.

Distribution of the ML estimator

Theorem (Distribution of the ML estimator)

We assume that $\widehat{\boldsymbol{\theta}}$ is consistent. Then, under some regularity conditions,

$$\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta} \rightarrow \mathrm{N}(0, \boldsymbol{i}(\boldsymbol{\theta})^{-1})$$

where $\boldsymbol{i}(\boldsymbol{\theta})$ is the expected information or the information matrix.

- The results can be used for inference under very general conditions. As the price for the generality, the results are only asymptotically valid.
- Asymptotically the variance of the estimator is seen to be equal to the Cramer-Rao lower bound for any unbiased estimator.
- The practical significance of this result is that the MLE makes efficient use of the available data for large data sets.

Distribution of the ML estimator

In practice, we would use

$$\widehat{\theta} \sim \mathrm{N}(\theta, \boldsymbol{j}^{-1}(\widehat{\boldsymbol{\theta}}))$$

where $\boldsymbol{j}(\widehat{\boldsymbol{\theta}})$ is the observed (Fisher) information.

This means that asymptotically

i) $\mathrm{E}[\widehat{\boldsymbol{\theta}}] = \boldsymbol{\theta}$

ii) $D[\widehat{\boldsymbol{\theta}}] = \boldsymbol{j}^{-1}(\widehat{\boldsymbol{\theta}})$

Distribution of the ML estimator

- The standard error of $\widehat{\theta}_i$ is given by

  $$\widehat{\sigma}_{\hat{\theta}_i} = \sqrt{\mathrm{Var}_{ii}[\widehat{\theta}]}$$

  where $\mathrm{Var}_{ii}[\widehat{\theta}]$ is the i'th diagonal term of $\boldsymbol{j}^{-1}(\widehat{\boldsymbol{\theta}})$
- Hence we have that an estimate of the dispersion (variance-covariance matrix) of the estimator is

  $$D[\widehat{\boldsymbol{\theta}}] = \boldsymbol{j}^{-1}(\widehat{\boldsymbol{\theta}})$$

- An estimate of the uncertainty of the individual parameter estimates is obtained by decomposing the dispersion matrix as follows:

  $$D[\widehat{\boldsymbol{\theta}}] = \widehat{\boldsymbol{\sigma}}_{\hat{\boldsymbol{\theta}}} \boldsymbol{R} \widehat{\boldsymbol{\sigma}}_{\hat{\boldsymbol{\theta}}}$$

  into $\widehat{\boldsymbol{\sigma}}_{\hat{\boldsymbol{\theta}}}$, which is a diagonal matrix of the standard deviations of the individual parameter estimates, and $\boldsymbol{R}$, which is the corresponding correlation matrix. The value R_{ij} is thus the estimated correlation between $\widehat{\theta}_i$ and $\widehat{\theta}_j$.

The Wald Statistic

A test of an individual parameter

$$\mathcal{H}_0 : \theta_i = \theta_{i,0}$$

is given by the *Wald statistic*:

$$Z_i = \frac{\widehat{\theta}_i - \theta_{i,0}}{\widehat{\sigma}_{\hat{\theta}_i}}$$

which under $\mathcal{H}_0$ is approximately $\mathrm{N}(0,1)$-distributed.

Quadratic approximation of the log-likelihood

- A second-order Taylor expansion around $\widehat{\theta}$ provides us with a quadratic approximation of the normalized log-likelihood around the MLE.
- A second-order Taylors expansion around $\widehat{\theta}$ we get

$$l(\theta) \approx l(\widehat{\theta}) + l'(\widehat{\theta})(\theta - \widehat{\theta}) - \frac{1}{2} j(\widehat{\theta})(\theta - \widehat{\theta})^2$$

  and then

$$\log \frac{L(\theta)}{L(\widehat{\theta})} \approx -\frac{1}{2} j(\widehat{\theta})(\theta - \widehat{\theta})^2$$

- In the case of normality the approximation is exact which means that a quadratic approximation of the log-likelihood corresponds to normal approximation of the $\widehat{\theta}(\boldsymbol{Y})$ estimator.

Example: Quadratic approximation of the log-likelihood

Consider again the thumbtack example.
The log-likelihood function is:

$$l(\theta) = y \log \theta + (n - y) \log(1 - \theta) + const$$

The score function is:

$$l'(\theta) = \frac{y}{\theta} - \frac{n - y}{1 - \theta},$$

and the observed information:

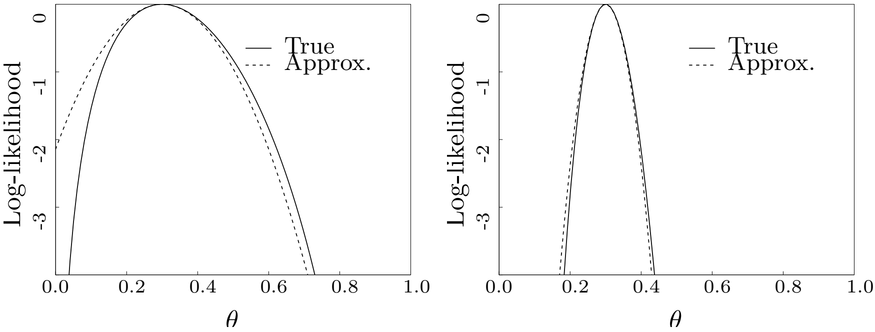
$$j(\theta) = \frac{y}{\theta^2} + \frac{n - y}{(1 - \theta)^2}.$$

For $n = 10,\ y = 3$ and $\widehat{\theta} = 0.3$ we obtain

$$j(\widehat{\theta}) = 47.6$$

The quadratic approximation is poor in this case. By increasing the sample size to $n = 100$, but still with $\widehat{\theta} = 0.3$ the approximation is much better.

Example: Quadratic approximation of the log-likelihood



(a) $n = 10, y = 3$

(b) $n = 100, y = 30$

Figure: Quadratic approximation of the log-likelihood function.

Some R hints

Function for calculating the likelihood function as a function of the mean value parameter for Gaussian data (in x) with known standard deviation:

```
> L.gaussian.data <- function(theta)  {
+     prod(dnorm(x,mean=theta,sd=standard.dev))
+     }
```

To plot the likelihood function you may use something like

```
th <- seq(mean(x) - 3*standard.dev, mean(x) + 3*standard.dev, length =200)
L <- sapply(th, L.gaussian.data)
plot(th,L/max(L), ylab="L", xlab=expression(theta))
```

To calculate the log likelihood function and estimate the parameter(s) you may use something like

```
nll.gaussian.data <- function(theta) {
    -sum(dnorm(x, mean=theta, sd=standard.dev, log=TRUE))
  }
fit <- optim(x, nll.gaussian.data, hessian = TRUE)
fit[c("convergence","par","hessian")]
```