# Advanced dataanalysis and statistical modelling, week 2

## The likelihood principle

Jan K. Møller

DTU Compute
Section for Dynamical Systems
Technical University of Denmark
jkmo@dtu.dk

February 2018

# An Abstract

- Descriptive statistics, look at the data
- Estimator (A random variable)
- Estimate (a number, the realisation of the estimator)
- The likelihood function

$$L(\theta; x_1, ..., x_n) = f_\theta(x_1, ..., x_n)$$

- log-likelihood function ($l(\theta; \cdot) = \log L(\theta; \cdot)$)
- Maximum Likelihood Esitmate

$$\hat{\theta} = \arg\max_\theta l(\theta)$$

- Score function $S(\theta) = \frac{\partial l}{\partial \theta}$

# An Abstract

- Properties of estimators
  - Bias ($E[\hat{\theta} - \theta)]$)
  - Variance ($E[(\hat{\theta} - E(\hat{\theta}))^2]$)
  - Mean square error ($E[(\hat{\theta} - \theta)^2]$)
  - Consistency $\hat{\theta}_n \to \theta$ for $n \to \infty$
  - MLE's consistent and assymptotically normal
- Hypothesis testing (comparing nested models) and Test statistics
  - Likelihood ratio test
  - Wald test

# Overview

1. The likelihood principle
2. Point estimation theory
3. The likelihood function
4. The information matrix
5. The maximum likelihood estimate (MLE)
6. Model selection
7. Profile likelihood

# Overview

# 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, $\theta$?

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

# 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

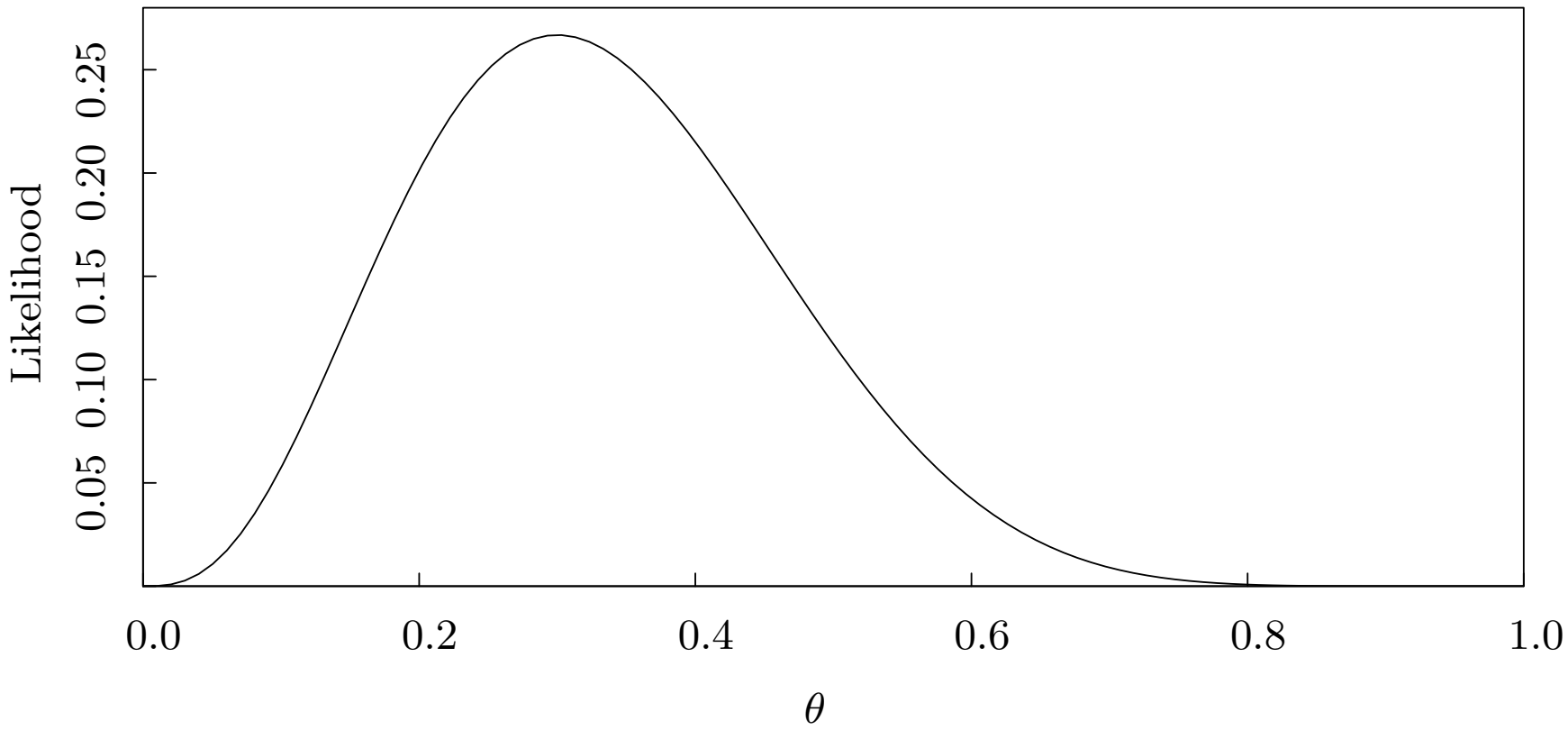



Figure: Likelihood function of the success probability $\theta$ in a binomial experiment with $n = 10$ and $y = 3$.

# 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 $\theta$.
By solving

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

it is readily seen that the maximum likelihood *estimate* (MLE) for $\theta$ 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.

# Overview

# Point estimation theory

We will assume that the statistical model for $\boldsymbol{y}$ is given by a 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

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

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

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

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

# Point estimation theory

## Dispersion matrix

The matrix $\mathrm{Var}\left[\widehat{\boldsymbol{\theta}}(\boldsymbol{Y})\right]$ 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*.

# Overview

# 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.

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

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

# 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 $y_1 = 4.6$; $y_2 = 6.3$; and $y_3 = 5.0$.

# 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
- $Y_i \sim \mathrm{N}(\mu, \sigma^2)$
- $\sigma^2$ 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 [1]

$$f_{Y_1,Y_2,Y_3}(y_1, y_2, y_3; \mu) = \frac{1}{(\sqrt{2\pi})^3} \exp\left[-\sum_{i=1}^{3} \frac{(y_1 - \mu)^2}{2}\right]$$

which for every value of $\mu$ is a function of the three variables $y_1, y_2, y_3$. Now, we can establish the likelihood function

$$\begin{aligned} L_{4.6,6.3,5.0}(\mu) &= f_{Y_1,Y_2,Y_3}(4.6, 6.3, 5.0; \mu) \\ &= \frac{1}{(\sqrt{2\pi})^3} \exp\left[-\frac{(4.6 - \mu)^2}{2} - \frac{(6.3 - \mu)^2}{2} - \frac{(5.0 - \mu)^2}{2}\right] \end{aligned}$$

The function depends only on $\mu$.

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

# Example: Likelihood function for mean of normal distribution

Reducing the expression one finds

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

which shows that (except for a factor not depending on $\mu$), 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

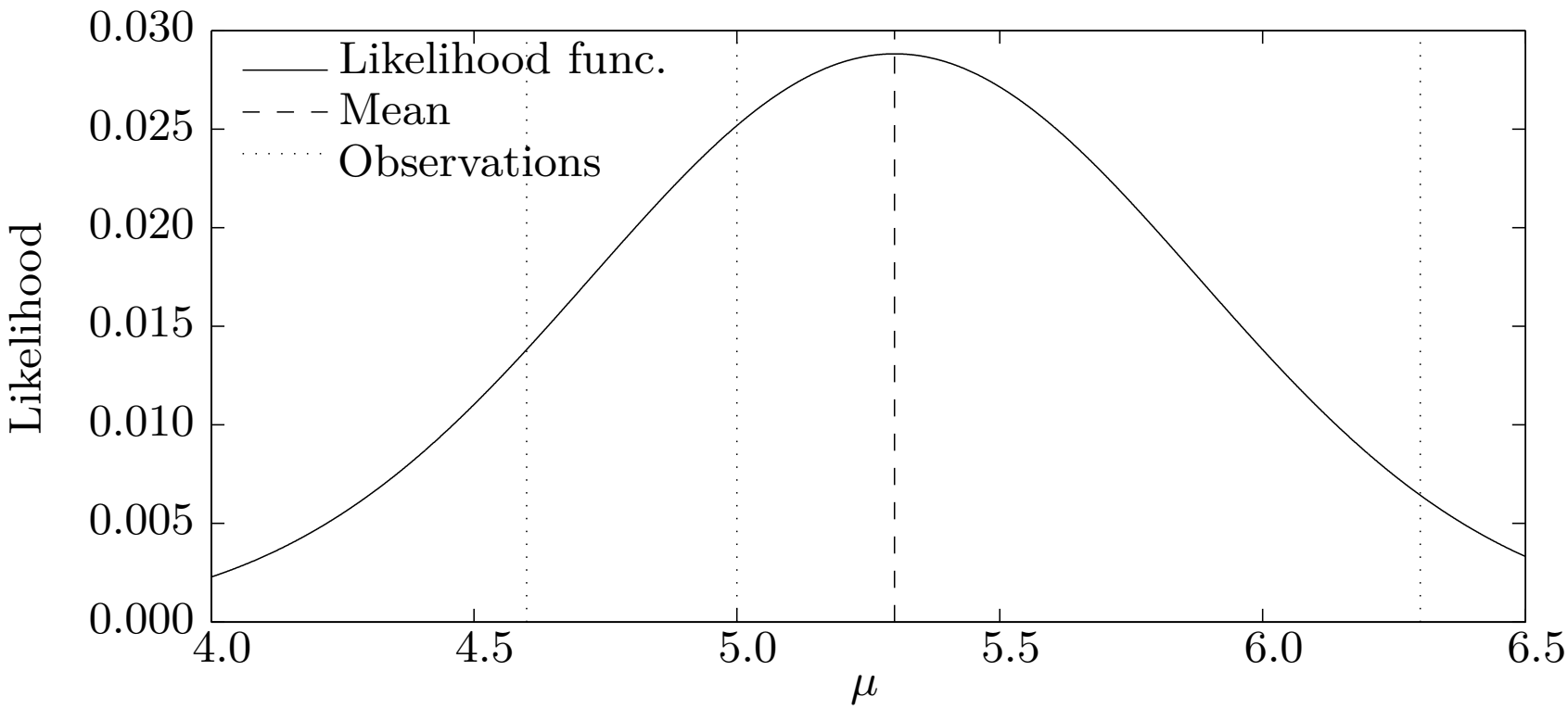



Figure: The likelihood function for $\mu$ given the observations $y_1 = 4.6$; $y_2 = 6.3$ and $y_3 = 5.0$.

# 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 that contains all the information in the observations. Formally

$$f_{\boldsymbol{Y}}(\boldsymbol{Y}; \theta) = h(\boldsymbol{y}) g(t(\boldsymbol{y}); \theta)$$

with the factor $h(\boldsymbol{y})$ not depending on the parameter $\theta$, and the factor $g(t(\boldsymbol{y}); \theta)$ only depending on $\boldsymbol{y}$ through the (vector valued) function $t()$.

# Overview

# The Score function

## Definition (Score function)

Consider $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_k) \in \Theta^k$, and assume that $\Theta^k$ 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:

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

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

# The 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 $\theta$, 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*

# 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

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

# Overview

# The Maximum Likelihood Estimate (MLE)

## Definition (Maximum Likelihood Estimate (MLE))

Given the observation $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$ the *Maximum Likelihood Estimate (MLE)* is a function $\widehat{\boldsymbol{\theta}}(\boldsymbol{y})$ such that

$$L(\widehat{\boldsymbol{\theta}}; \boldsymbol{y}) = \sup_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta}; \boldsymbol{y})$$

The function $\widehat{\boldsymbol{\theta}}(\boldsymbol{Y})$ over the sample space of observations $\boldsymbol{Y}$ is called an *ML estimator*.

In practice it is convenient to work with the log-likelihood function $l(\boldsymbol{\theta}; \boldsymbol{y})$.

# 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 $\boldsymbol{\psi}(\boldsymbol{\theta})$.*

# Example: Invariance property

Consider the binomial example

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

The optimisation needs to be bounded ($p \in [0, 1]$), an alternative parametrisation would be

$$p(\theta) = \frac{1}{1 + e^{-\theta}}$$

here $p(-\infty) = 0$ and $p(\infty) = 1$, hence $\theta$ need not to be bounded.

# 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.
- 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 are 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 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.
The $(1-\alpha)$ Wald confidence interval is therefore

$$CI_{\theta_i} = \hat{\theta}_i \pm z_{1-\alpha/2}\hat{\sigma}_{\theta_i}$$

where $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile in the standard normal distribution.

# 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:

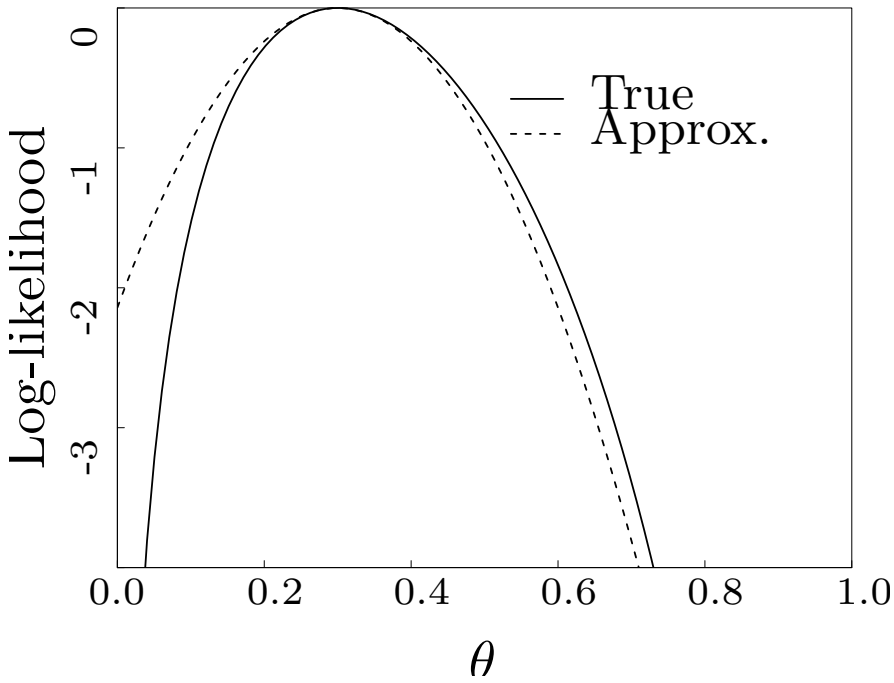
$$j(\theta) = \frac{y}{\theta^2} + \frac{n - y}{(1 - \theta)^2}.$$

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

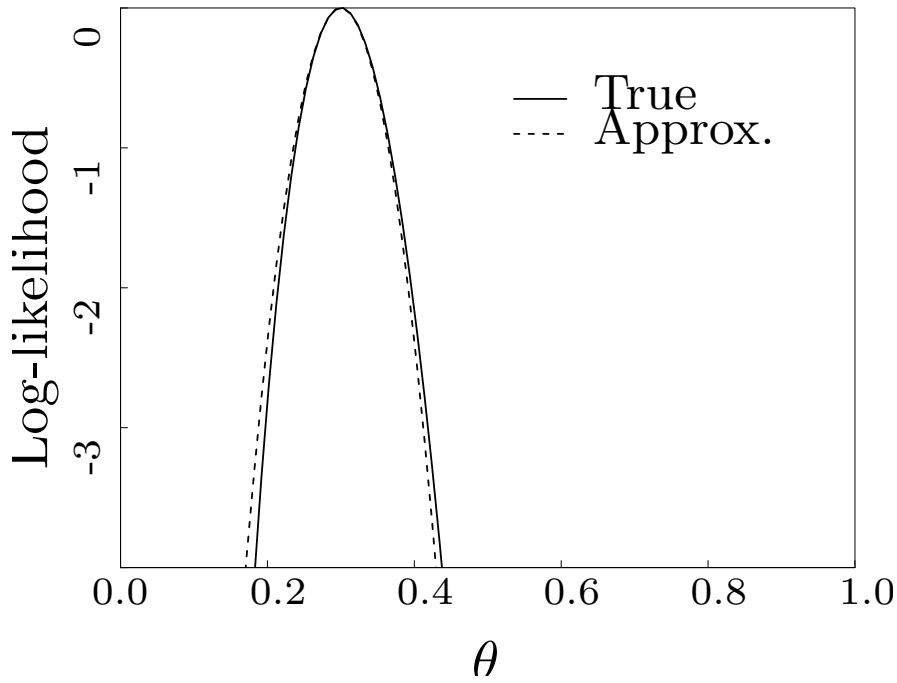
$$j(\widehat{\theta}) = 47.6$$

# Example: Quadratic approximation of the log-likelihood

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.




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




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

# Overview

# Example: Poisson regression

Consider the data

| x | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| y | 3 | 0 | 4 | 5 | 6 | 4 | 9 | 7 | 4 | 10 |

the following regression model i proposed

$$Y_i \sim Pois(\lambda_i); \quad \lambda_i = \exp\{\theta_0 + \theta_1 x_i\}$$

Fit the parameters and estimate the unceartainty of the parameters.

# Likelihood ratio tests

- Methods for testing hypotheses using the likelihood function.
- The basic idea: determine the maximum likelihood estimates under both a null and alternative hypothesis.
- It is assumed that a sufficient model with $\boldsymbol{\theta} \in \Omega$ exists.
- Then consider some theory or assumption about the parameters $\mathcal{H}_0 : \boldsymbol{\theta} \in \Omega_0$ where $\Omega_0 \subset \Omega \quad ; \dim(\Omega_0) = r$ and $\dim(\Omega) = k$
- The purpose of the testing is to analyze whether the observations provide sufficient evidence to reject this theory or assumption. If not we accept the null hypothesis.

# Likelihood ratio tests

## Definition (Likelihood ratio)

Consider the hypothesis $\mathcal{H}_0 : \boldsymbol{\theta} \in \Omega_0$ against the alternative $\mathcal{H}_1 : \boldsymbol{\theta} \in \Omega \setminus \Omega_0$ ($\Omega_0 \subseteq \Omega$), where $\dim(\Omega_0) = r$ and $\dim(\Omega) = k$.
For given observations $y_1, y_2, ..., y_n$ the *likelihood ratio* is defined as

$$\lambda(\boldsymbol{y}) = \frac{\sup_{\boldsymbol{\theta} \in \Omega_0} L(\boldsymbol{\theta}; \boldsymbol{y})}{\sup_{\boldsymbol{\theta} \in \Omega} L(\boldsymbol{\theta}; \boldsymbol{y})}$$

- If $\lambda$ is small, then the data are seen to be more plausible under the alternative hypothesis than under the null hypothesis.
- Hence the hypothesis ($\mathcal{H}_0$) is rejected for small values of $\lambda$.

# Likelihood ratio tests

- It is sometimes possible to transform the likelihood ratio into a statistic, the exact distribution of which is known under $\mathcal{H}_0$. This is for instance the case for the General Linear Model for Gaussian data.
- In most cases, however, we must use the following important result regarding the asymptotic behavior:

## Theorem (Wilk's Likelihood Ratio test)

*For $\lambda(\boldsymbol{y})$ as above, then under the null hypothesis $\mathcal{H}_0$, the random variable $-2\log\lambda(\boldsymbol{Y})$ converges in law to a $\chi^2$ random variable with $(k-r)$ degrees of freedom, i.e.,*

$$-2\log\lambda(\boldsymbol{Y}) \rightarrow \chi^2(k-r)$$

*under $\mathcal{H}_0$.*

$-2\log\lambda(\boldsymbol{Y})$ is also called the deviance and is denoted $D$

# Example: Linear regression

Assume we want to the the following hypothesis

$$\mathcal{H}_0: \quad Y_i = \beta_0 + \epsilon_i; \quad \epsilon_i \sim N(0, \sigma^2)$$

against the alternative

$$\mathcal{H}_1: \quad Y_i = \beta_0 + \beta_1 x_i + \epsilon_i; \quad \epsilon_i \sim N(0, \sigma^2)$$

Also assume that $\sigma^2$ is known. The log-likelihood in each of the two cases is

$$\begin{aligned} l_0(\hat{\beta}_0) &= -\frac{1}{2\sigma^2}\sum(y_i - \hat{\beta}_0)^2 &&= -\frac{1}{2\sigma^2}\sum e_{0,i}^2 \\ l_1(\hat{\beta}_0, \hat{\beta}_1) &= -\frac{1}{2\sigma^2}\sum(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 &&= -\frac{1}{2\sigma^2}\sum e_{1,i}^2 \end{aligned}$$

# Example: Linear regression, cont.

Assuming $\mathcal{H}_0$ is true, then

$$2l_0 \sim \chi^2(n-1); \quad 2l_1 \sim \chi^2(n-2)$$

and the deviance between the models is

$$D = 2(l_1 - l_0) \sim \chi^2(1)$$

Note that

- In this case the test is exact (normal errors and known variance)
- $\mathcal{H}_0$ is rejected if $D$ is large, i.e. the error in $\mathcal{H}_0$ are small compared to $\mathcal{H}_1$.
- When $\sigma^2$ is not known, exact distribution are also available (see chap. 3)

# Hypothesis chains

Consider a *chain* of hypotheses specified by a sequence of parameter spaces

$$\mathbb{R} \subseteq \Omega_M \ldots \subset \Omega_2 \subset \Omega_1 \subset \mathbb{R}^n.$$

For each parameter space $\Omega_i$ we define a hypothesis

$$\mathcal{H}_i : \boldsymbol{\theta} \in \Omega_i$$

with $\dim(\Omega_i) < \dim(\Omega_{i-1})$.

# Partial likelihood ratio test

## Definition (Partial likelihood ratio test)

Assume that the hypothesis $\mathcal{H}_i$ allows the sub hypothesis $\mathcal{H}_{i+1} \subset \mathcal{H}_i$. The *partial likelihood ratio test* for $\mathcal{H}_{i+1}$ under $\mathcal{H}_i$ is the likelihood ratio test for the hypothesis $\mathcal{H}_{i+1}$ under the assumption that the hypothesis $\mathcal{H}_i$ holds. The likelihood ratio test statistic for this partial test is

$$\lambda_{\mathcal{H}_{i+1}|\mathcal{H}_i}(y) = \frac{\sup_{\boldsymbol{\theta}\in\Omega_{i+1}} L(\boldsymbol{\theta};\boldsymbol{y})}{\sup_{\boldsymbol{\theta}\in\Omega_i} L(\boldsymbol{\theta};\boldsymbol{y})}$$

When $\mathcal{H}_{i+1}$ holds, the distribution of $\lambda_{\mathcal{H}_{i+1}|\mathcal{H}_i}(Y)$ approaches a $\chi^2(f)$ distribution with $f = \dim(\Omega_i) - \dim(\Omega_{i+1})$.

# Partial tests

**Theorem (Partitioning into a sequence of partial tests)**

*Consider a chain of hypotheses.*
*Now, assume that $\mathcal{H}_1$ holds, and consider the minimal hypotheses $\mathcal{H}_M : \boldsymbol{\theta} \in \Omega_M$ with the alternative $\mathcal{H}_1 : \boldsymbol{\theta} \in \Omega_1 \setminus \Omega_M$. The likelihood ratio test statistic $\lambda_{\mathcal{H}_M|\mathcal{H}_1}(y)$ for this hypothesis may be factorized into a chain of partial likelihood ratio test statistics $\lambda_{\mathcal{H}_{i+1}|\mathcal{H}_i}(y)$ for $\mathcal{H}_{i+1}$ given $\mathcal{H}_i$, $i = 1, \ldots, M$.*

- The partial tests "corrects" for the effect of the parameters that are in the model at that particular stage
- When interpreting the test statistic corresponding to a particular stage in the hierarchy of models, one often says that there is "controlled for", or "corrected for" the effect of the parameters that are in the model at that stage.

# Example: likelihood ratio test

Consider the data

| x | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| y | 3 | 0 | 4 | 5 | 6 | 4 | 9 | 7 | 4 | 10 |

the following regression model i proposed

$$Y_i \sim Pois(\lambda_i); \quad \lambda_i = \theta_0 + \theta_1 x_i$$

Test the hypothesis $\mathcal{H}_1 : \quad \theta_1 \neq 0$

# Variable selection in hypothesis chains

## *In-sample* methods for model selection

The model complexity is evaluated using the same observations as those used for estimating the parameters of the model.

- The *training data* is used for evaluating the performance of the model.
- Any extra parameter will lead to a reduction of the loss function.
- In the in-sample case statistical tests are used to access the significance of extra parameters, and when the improvement is small in some sense the parameters are considered to be non-significant.
- The classical statistical approach.
- *test set*: used for assessing the generalized performance, i.e. the performance on new data

# Variable selection in hypothesis chains

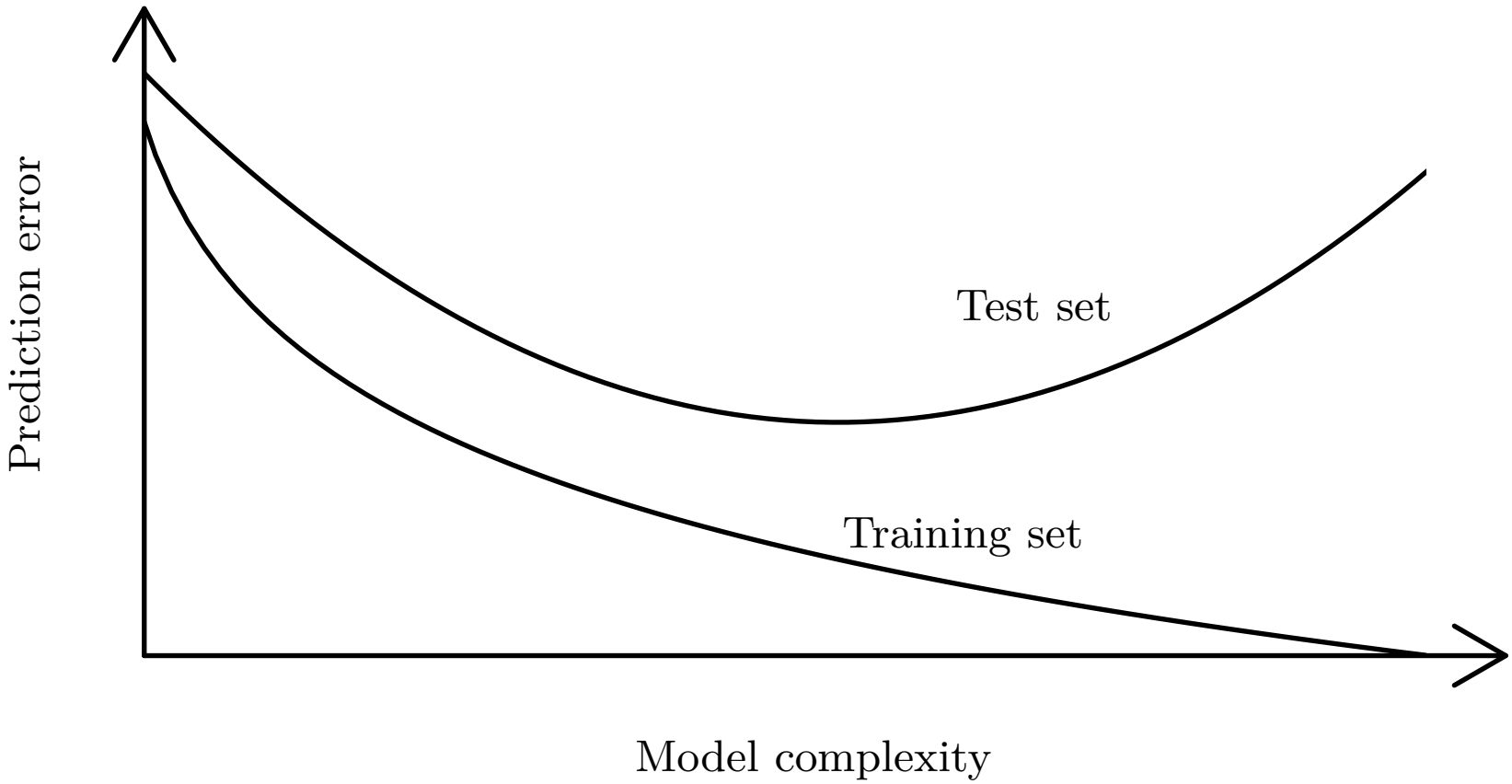



Figure: A typical behavior of the (possibly generalized) training and test prediction error as a function of the model complexity.

# Overview

# Profile likelihood

## Definition (Profile likelihood)

Assume that the statistical model for the observations $Y_1, Y_2, \ldots, Y_n$ is given by the family of joint densities, $\theta = (\tau, \zeta)$ and $\tau$ denoting the parameter of interest. Then the *profile likelihood function* for $\tau$ is the function

$$L_P(\tau; \boldsymbol{y}) = \sup_{\zeta} L((\tau, \zeta); \boldsymbol{y})$$

where the maximization is performed at a fixed value of $\tau$.

# Profile likelihood - Confidence intervals

Consider again $\theta = (\tau, \zeta)$. From the likelihood ratio test it is seen that

$$\left\{ \tau; \frac{L_P(\tau; y)}{L(\widehat{\theta}, y)} > \exp(-\frac{1}{2}\chi^2_{1-\alpha}(p)) \right\} \tag{1}$$

defines a set of values of $\tau$ ($p$-dimensional) that constitutes a $100(1-\alpha)\%$ confidence region for $\tau$.

- For normal distributions the confidence is exact and else it is an approximation.
- Likelihood based confidence intervals have the advantage of possibly being asymmetric when this is relevant.

# Overview

1. The likelihood principle
2. Point estimation theory
3. The likelihood function
4. The information matrix
5. The maximum likelihood estimate (MLE)
6. Model selection
7. Profile likelihood