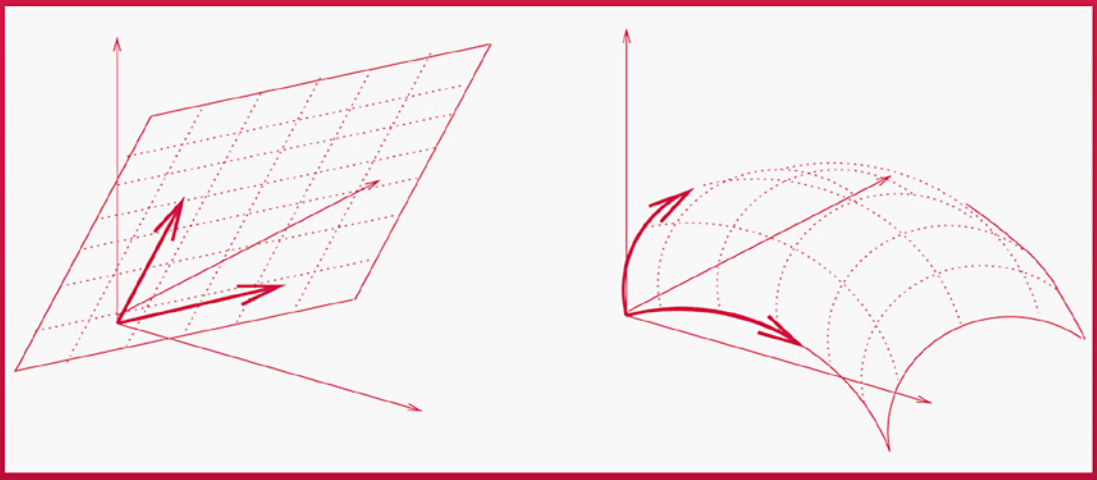
Texts in Statistical Science

Introduction to General and Generalized Linear Models



Henrik Madsen
Poul Thyregod

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

Introduction to General and Generalized Linear Models

CHAPMAN & HALL/CRC
Texts in Statistical Science Series

Analysis of Failure and Survival Data
P. J. Smith

The Analysis of Time Series — An Introduction, Sixth Edition
C. Chatfield

Applied Bayesian Forecasting and Time Series Analysis
A. Pole, M. West and J. Harrison

Applied Nonparametric Statistical Methods, Fourth Edition
P. Sprent and N.C. Smeeton

Applied Statistics — Handbook of GENSTAT Analysis
E.J. Snell and H. Simpson

Applied Statistics — Principles and Examples
D.R. Cox and E.J. Snell

Applied Stochastic Modelling, Second Edition
B.J.T. Morgan

Bayesian Data Analysis, Second Edition
A. Gelman, J.B. Carlin, H.S. Stern and D.B. Rubin

Bayesian Ideas and Data Analysis: An Introduction for Scientists and Statisticians
R. Christensen, W. Johnson, A. Branscum, and T.E. Hanson

Bayesian Methods for Data Analysis, Third Edition
B.P. Carlin and T.A. Louis

Beyond ANOVA — Basics of Applied Statistics
R.G. Miller, Jr.

Computer-Aided Multivariate Analysis, Fourth Edition
A.A. Afifi and V.A. Clark

A Course in Categorical Data Analysis
T. Leonard

A Course in Large Sample Theory
T.S. Ferguson

Data Driven Statistical Methods
P. Sprent

Decision Analysis — A Bayesian Approach
J.Q. Smith

Design and Analysis of Experiment with SAS
J. Lawson

Elementary Applications of Probability Theory, Second Edition
H.C. Tuckwell

Elements of Simulation
B.J.T. Morgan

Epidemiology — Study Design and Data Analysis, Second Edition
M. Woodward

Essential Statistics, Fourth Edition
D.A.G. Rees

Exercises and Solutions in Biostatistical Theory
L.L. Kupper, B.H. Neelon, and S.M. O'Brien

Extending the Linear Model with R — Generalized Linear, Mixed Effects and Nonparametric Regression Models
J.J. Faraway

A First Course in Linear Model Theory
N. Ravishanker and D.K. Dey

Generalized Additive Models: An Introduction with R
S. Wood

Graphics for Statistics and Data Analysis with R
K.J. Keen

Interpreting Data — A First Course in Statistics
A.J.B. Anderson

Introduction to General and Generalized Linear Models
H. Madsen and P. Thyregod

An Introduction to Generalized Linear Models, Third Edition
A.J. Dobson and A.G. Barnett

Introduction to Multivariate Analysis
C. Chatfield and A.J. Collins

Introduction to Optimization Methods and Their Applications in Statistics
B.S. Everitt

Introduction to Probability with R
K. Baclawski

Introduction to Randomized Controlled Clinical Trials, Second Edition
J.N.S. Matthews

Introduction to Statistical Inference and Its Applications with R
M.W. Trosset

Introduction to Statistical Methods for Clinical Trials

T.D. Cook and D.L. DeMets

Large Sample Methods in Statistics
P.K. Sen and J. da Motta Singer

Linear Models with R
J.J. Faraway

Logistic Regression Models
J.M. Hilbe

Markov Chain Monte Carlo — Stochastic Simulation for Bayesian Inference, Second Edition
D. Gamerman and H.F. Lopes

Mathematical Statistics
K. Knight

Modeling and Analysis of Stochastic Systems, Second Edition
V.G. Kulkarni

Modelling Binary Data, Second Edition
D. Collett

Modelling Survival Data in Medical Research, Second Edition
D. Collett

Multivariate Analysis of Variance and Repeated Measures — A Practical Approach for Behavioural Scientists
D.J. Hand and C.C. Taylor

Multivariate Statistics — A Practical Approach
B. Flury and H. Riedwyl

Pólya Urn Models
H. Mahmoud

Practical Data Analysis for Designed Experiments
B.S. Yandell

Practical Longitudinal Data Analysis
D.J. Hand and M. Crowder

Practical Statistics for Medical Research
D.G. Altman

A Primer on Linear Models
J.F. Monahan

Probability — Methods and Measurement
A. O'Hagan

Problem Solving — A Statistician's Guide, Second Edition
C. Chatfield

Randomization, Bootstrap and Monte Carlo Methods in Biology, Third Edition
B.F.J. Manly

Readings in Decision Analysis
S. French

Sampling Methodologies with Applications
P.S.R.S. Rao

Statistical Analysis of Reliability Data
M.J. Crowder, A.C. Kimber, T.J. Sweeting, and R.L. Smith

Statistical Methods for Spatial Data Analysis
O. Schabenberger and C.A. Gotway

Statistical Methods for SPC and TQM
D. Bissell

Statistical Methods in Agriculture and Experimental Biology, Second Edition
R. Mead, R.N. Curnow, and A.M. Hasted

Statistical Process Control — Theory and Practice, Third Edition
G.B. Wetherill and D.W. Brown

Statistical Theory, Fourth Edition
B.W. Lindgren

Statistics for Accountants
S. Letchford

Statistics for Epidemiology
N.P. Jewell

Statistics for Technology — A Course in Applied Statistics, Third Edition
C. Chatfield

Statistics in Engineering — A Practical Approach
A.V. Metcalfe

Statistics in Research and Development, Second Edition
R. Caulcutt

Stochastic Processes: An Introduction, Second Edition

P.W. Jones and P. Smith

Survival Analysis Using S — Analysis of Time-to-Event Data
M. Tableman and J.S. Kim

The Theory of Linear Models
B. Jørgensen

Time Series Analysis
H. Madsen

Time Series: Modeling, Computation, and Inference
R. Prado and M. West

Texts in Statistical Science

Introduction to General and Generalized Linear Models

Henrik Madsen
Technical University of Denmark
Lyngby, Denmark

Poul Thyregod
Technical University of Denmark
Lyngby, Denmark

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group an **informa** business
A CHAPMAN & HALL BOOK

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Version Date: 20141009

International Standard Book Number-13: 978-1-4398-9114-8 (eBook - PDF)

Contents

Preface

This book contains an introduction to general and generalized linear models using the popular and powerful likelihood techniques. The aim is to provide a flexible framework for the analysis and model building using data of almost any type. This implies that the more well-known analyses based on Gaussian data like regression analysis, analysis of variance and analysis of covariance are generalized to a much broader family of problems that are linked to, for instance, binary, positive, integer, ordinal and qualitative data.

By using parallel descriptions in two separate chapters of general and generalized linear models, the book facilitates a unique comparison between these two important classes of models and, furthermore, presents an easily accessible introduction to the more advanced concepts related to generalized linear models.

Likewise, the concept of hierarchical models is illustrated separately – in one chapter a description of Gaussian based hierarchical models, such as the mixed effects linear models is outlined, and in another chapter an introduction is presented to the generalized concept of those hierarchical models that are linked to a much broader class of problems connected to various types of data and related densities. The book also introduces new concepts for mixed effects models thereby enabling more flexibility in the model building and in the allowed data structures.

Throughout the book the statistical software R is used. Examples show how the problems are solved using R, and for each of the chapters individual guidelines are provided in order to facilitate the use of R when solving the relevant type of problems.

Theorems are used to emphasize the most important results. Proofs are provided, only, if they clarify the results. Problems on a smaller scale are dealt with at the end of most of the chapters, and a separate chapter with real life inspired problems is included as the final chapter of the book.

During the sequence of chapters, more advanced models are gradually introduced. With such an approach, the relationship between general and generalized linear models and methods becomes more apparent.

The last chapter of this book is devoted to problems inspired by real life situations. At the home page `http://www.imm.dtu.dk/~hm/GLM` solutions to the problems are found. The homepage also contains additional exercises – called assignments – and a complete set of data for the examples used in the

book. Furthermore, a collection of slides for an introductory course on general, generalized and mixed effects models can be found on the homepage.

The contents of this book are mostly based on a comprehensive set of material developed by Professor Poul Thyregod during his series of lectures at the Section of Mathematical Statistics at the Technical University of Denmark (DTU). Poul was the first person in Denmark who received a PhD in mathematical statistics. Poul was also one of the few highly skilled in mathematical statistics who was fully capable of bridging the gap between theory and practice within statistics and data analysis. He possessed the capability to link statistics to real problems and to focus on the real added value of statistics — in order to help us understand the real world a bit better. The ability to work with engineers and scientists, to be part of the discovery process, and to be able to communicate so distinctly what statistics is all about is clearly a gift. Poul possessed that gift. I am grateful to be one of a long list of students who had the privilege of learning from his unique capabilities. Sadly, Poul passed away in the summer of 2008, which was much too early in his life. I hope, however, that this book will reflect his unique talent to establish an easily accessible introduction to theory and practice of modern statistical modeling.

I am grateful to all who have contributed with useful comments, suggestions and contributions. First I would like to thank my colleagues Gilles Guillot, Martin Wæver Pedersen, Stig Mortensen and Anders Nielsen for their helpful and very useful assistance and comments.

In particular, I am grateful to Anna Helga Jónsdóttir for her assistance with text, proofreading, figures, exercises and examples. Without her insistent support this book would never had been completed. Finally, I would like to thank Helle Welling for proofreading, and Morten Høgholm for both proofreading and for proposing and creating a new layout in LaTeX.

Henrik Madsen
Lyngby, Denmark

Notation

All vectors are column vectors. Vectors and matrices are emphasized using a bold font. Lowercase letters are used for vectors and uppercase letters are used for matrices. Transposing is denoted with the upper index ^{T}.

Random variables are always written using uppercase letters. Thus, it is not possible to distinguish between a multivariate random variable (random vector) and a matrix. However, variables and random variables are assigned to letters from the last part of the alphabet (X, Y, Z, U, V, ...), while constants are assigned to letters from the first part of the alphabet (A, B, C, D, ...). From the context it should be possible to distinguish between a matrix and a random vector.

CHAPTER 1

Introduction

This book provides an introduction to methods for statistical modeling using essentially all kind of data. The principles for modeling are based on likelihood techniques. These techniques facilitate our aim of bridging the gap between theory and practice for modern statistical model building.

Each chapter of the book contains examples and guidelines for solving the problems using the statistical software package R, which can be freely downloaded and installed on almost any computer system. We do, however, refer to other software packages as well.

In general the focus is on establishing models that explain the variation in data in such a way that the obtained models are well suited for predicting the outcome for given values of some explanatory variables. More specifically we will focus on *formulating, estimating, validating and testing models* for predicting the *mean value* of the random variables. However, by the considered approach we will consider the complete stochastic model for the data which includes an appropriate choice of the *density* describing the variation of the data. It will be demonstrated that this approach facilitates adequate methods for describing also the uncertainty of the predictions.

By the approach taken, the theory and practice in relation to widely applied methods for modeling using *regression analysis*, *analysis of variance* and the *analysis of covariance*, that are all related to Gaussian distributed data, are established in a way which facilitates an easily accessible extension to similar methods applied in the case of, e.g., Poisson, Gamma and Binomial distributed data. This is obtained by using the likelihood approach in both cases, and becomes clear that the *general linear models* are relevant for *Gaussian distributed samples* whereas the *generalized linear models* facilitate a modeling of the variation in a much broader context, namely for all data originating from the so-called *exponential family of densities* including Poisson, Binomial, Exponential, Gaussian, and Gamma distributions.

The presentation of the general and generalized linear models is provided using essentially the same methods related to the likelihood principles, but described in two separate chapters. By a parallel presentation of the methods and models in two chapters, a clear comparison between the two model types is recognized. This parallel presentation is also aiming at providing an easily accessible description of the theory for generalized linear models. This is due to the fact that the book first provides the corresponding or parallel results

for the general linear models, which is easier to understand, and in many cases well-known.

The book also contains a first introduction to both mixed effects models (also called mixed models) and hierarchical models. Again, a parallel setup in two separate chapters is provided. The first chapter concentrates on introducing the random effects and, consequently, also the mixed effects in a Gaussian context. The subsequent chapter provides an introduction to non-Gaussian hierarchical models where the considered models again are members of the exponential family of distributions.

To the readers with a theoretical interest it will be obvious that virtually all the results are based on about a handful of results from the likelihood theory, and that the results that are valid for finite samples for the general linear models are valid asymptotically in the case of generalized linear models. The necessary likelihood theory is described in the chapter following the Introduction.

1.1 Examples of types of data

Let us first illustrate the power of the methods considered in this book by listing some of the types of data which can be modelled using the described techniques. In practice several types of response variables are seen as indicated by the examples listed below:

i) *Continuous data* (e.g., $y_1 = 2.3$, $y_2 = -0.2$, $y_3 = 1.8$, ..., $y_n = 0.8$). Normal (Gaussian) distributed. Used, e.g., for air temperatures in degrees Celsius. An example is found in Example 2.18 on page 14.

ii) *Continuous positive data* (e.g., $y_1 = 0.0238$, $y_2 = 1.0322$, $y_3 = 0.0012$, ..., $y_n = 0.8993$). Log-normally distributed. Often used for concentrations.

iii) *Count data* (e.g., $y_1 = 57$, $y_2 = 67$, $y_3 = 54$, ..., $y_n = 59$). Poisson distributed. Used, e.g., for number of accidents — see Example 4.7 on page 123 on page 123.

iv) *Binary (or quantal) data* (e.g., $y_1 = 0$, $y_2 = 0$, $y_3 = 1$, ..., $y_n = 0$), or proportion of counts (e.g. $y_1 = 15/297$, $y_2 = 17/242$, $y_3 = 2/312$, ..., $y_n = 144/285$). Binomial distribution — see Example 4.6 on page 118 or Example 4.14 on page 140.

v) *Nominal data* (e.g., "Very unsatisfied", "Unsatisfied", "Neutral", "Satisfied", "Very satisfied"). Multinomial distribution — see Example 4.12 on page 133.

The reader will also become aware that the data of a given type might look alike, but the (appropriate) statistical treatment is different!

1.2 Motivating examples

The Challenger disaster

On January 28, 1986, Space Shuttle Challenger broke apart 73 seconds into its flight and the seven crew members died. The disaster was due to a disintegration of an O-ring seal in the right rocket booster. The forecast for January 28, 1986 indicated an unusually cold morning with air temperatures around 28 degrees F (−1 degrees C).

During a teleconference on January 27, one of the engineers, Morton Thiokol, responsible for the shuttle's rocket booster, expressed concern due to the low temperature.

The planned launch on January 28, 1986 was launch number 25. During the previous 24 launches problems with the O-ring were observed in 6 cases. Figure 1.1 shows the relationship between observed sealing problems and the air temperature. A model of the probability for O-ring failure as a function of the air temperature would clearly have shown that given the forecasted air temperature, problems with the O-rings were very likely to occur.

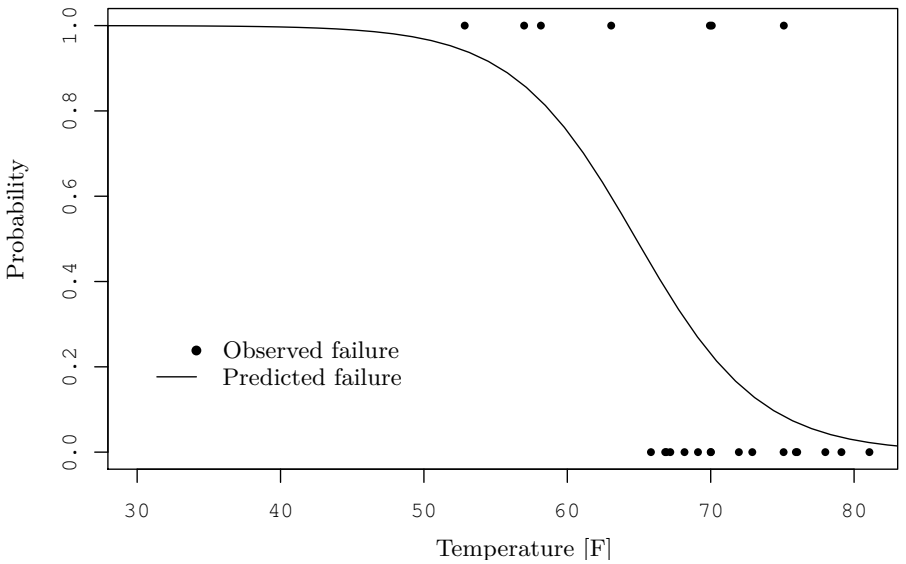


Figure 1.1: *Observed failure of O-rings in 6 out of 24 launches along with predicted probability for O-ring failure.*

Table 1.1: *Incidence of Torsade de Pointes by dose for high risk patients.*

Index i	Daily dose [mg] x_i	Number of subjects n_i	Number showing TdP z_i	Fraction showing TdP p_i
1	80	69	0	0
2	160	832	4	0.5
3	320	835	13	1.6
4	480	459	20	4.4
5	640	324	12	3.7
6	800	103	6	5.8

QT prolongation for drugs

In the process of drug development it is required to perform a study of potential prolongation of a particular interval of the electrocardiogram (ECG), the QT interval. The QT interval is defined as the time required for completion of both ventricular depolarization and repolarization. The interval has gained clinical importance since a prolongation has been shown to induce potentially fatal ventricular arrhythmia such as Torsade de Pointes (TdP). The arrhythmia causes the QRS complexes, another part of the ECG, to swing up and down around the baseline of the ECG in a chaotic fashion. This probably caused the name which means "twisting of the points" in French. A number of drugs have been reported to prolong the QT interval, both cardiac and non-cardiac drugs. Recently, both previously approved as well as newly developed drugs have been withdrawn from the market or have had their labeling restricted because of indication of QT prolongation. Table 1.1 shows results from a clinical trial where a QT prolonging drug was given to high risk patients. The patients were given the drug in six different doses and the number of incidents of Torsade de Points counted.

It is reasonable to consider the *fraction*, $Y_i = \frac{Z_i}{n_i}$, of incidences of Torsade de Points as the interesting variable. A natural distributional assumption is the binomial distribution, $Y_i \sim \mathrm{B}(n_i, p_i)/n_i$, where n_i is the number of subjects given the actual dosage and p_i is the fraction showing Torsade de Pointes.

▸ **Remark 1.1 – A bad model**
Obviously the fraction, p_i is higher for a higher daily dosage of the drug. However, a linear model of the form $Y_i = p_i + \epsilon_i$ where $p_i = \beta_0 + \beta_1 x_i$ does not reflect that, p_i is between zero and one, and the model for the fraction, Y_i (as 'mean plus noise') is clearly not adequate, since the observations are between zero and one.

It is, thus, clear that the distribution of ϵ_i and then the variance of observations must be dependent on p_i. Also, the problem with the homogeneity of the variance indicates that a traditional "mean plus noise" model is not adequate here. ◂

▸ **Remark 1.2 – A correct model**
Instead we will now formulate a model for transformed values of the observed fractions p_i.

Given that $Y_i \sim \mathrm{B}(n_i, p_i)/n_i$ we have that

$$\mathrm{E}[Y_i] = p_i \tag{1.1}$$

$$\mathrm{Var}[Y_i] = \frac{p_i(1-p_i)}{n_i} \tag{1.2}$$

i.e., the variance is now a function of the mean value. Later on the so-called mean value function $V(\mathrm{E}[Y_i])$ will be introduced which relates the variance to the mean value.

A successful construction is to consider a function, the so-called *link function* of the mean value $\mathrm{E}[Y]$. In this case we will use the *logit*-transformation

$$g(p_i) = \log\Big(\frac{p_i}{1-p_i}\Big) \tag{1.3}$$

and we will formulate a *linear model* for the transformed values. A plot of the observed logits, $g(p_i)$ as a function of the concentration indicates a linear relation of the form

$$g(p_i) = \beta_0 + \beta_1 x_i \tag{1.4}$$

After having estimated the parameters, i.e., we have obtained $(\widehat{\beta}_0, \widehat{\beta}_1)$, it is now possible to use the inverse transformation, which gives the predicted fraction $\widehat{p}$ of subjects showing Torsade de Pointes as a function of a daily dose, x using the *logistic function*:

$$\widehat{p} = \frac{\exp\left(\widehat{\beta}_0 + \widehat{\beta}_1 x\right)}{1 + \exp(\widehat{\beta}_0 + \widehat{\beta}_1 x)}. \tag{1.5}$$

This approach is called *logistic regression*.

It is easily seen that this model will ensure that the fraction is between zero and one, and we also see that we have established a reasonable description of the relation between the mean and the variance of the observations. ◂

1.3 A first view on the models

As mentioned previously, we will focus on statistical methods to formulate models for predicting the expected value of the *outcome*, *dependent*, or *response variable*, Y_i as a function of the known *independent variables*, $x_{i1}, x_{i2}, \ldots, x_{ik}$. These k variables are also called *explanatory*, or *predictor variables* or *covariates*. This means that we shall focus on models for the expectation $\mathrm{E}[Y_i]$.

Previously we have listed examples of types of response variables. Also the explanatory variables might be labeled as *continuous*, *discrete*, *categorical*, *binary*, *nominal*, or *ordinal*. To predict the response, a typical model often includes a combination of such types of variables.

Since we are going to use a likelihood approach, a specification of the probability distribution of Y_i is a very important part when specifying the model.

General linear models

In Chapter 3, which considers *general linear models*, the expected value of the response variable Y is linked linearly to the explanatory variables by an equation of the form

$$\mathrm{E}[Y_i] = \beta_1 x_{i1} + \cdots + \beta_k x_{ik} \; . \tag{1.6}$$

It will be shown that for Gaussian data it is reasonable to build a model directly for the expectation as shown in (1.6), and this relates to the fact that for Gaussian distributed random variables, all conditional expectations are linear (see e.g., Madsen (2008)).

▸ **Remark 1.3**
In model building, models for the mean value are generally considered. However, for some applications, models for, say, the 95% quantile might be of interest. Such models can be established by, e.g., *quantile regression*; see Koenker (2005). ◂

Generalized linear models

As indicated by the motivating example above it is, however, often more reasonable to build a linear model for a transformation of the expected value of the response. This approach is more formally described in connection with the *generalized linear models* in Chapter 4, where a link between the expected value of response and the explanatory variables is of the form

$$g(\mathrm{E}[Y_i]) = \beta_1 x_{i1} + \ldots + \beta_k x_{ik} \; . \tag{1.7}$$

The function $g(.)$ is called the *link function* and the right hand side of (1.7) is called the *linear component* of the model.

Thus, a full specification of the model contains a specification of

1. The *probability density* of Y. In Chapter 3 this will be the Gaussian density, i.e., $Y \sim N(\mu, \sigma^2)$, whereas in Chapter 4 the probability density will belong to the *exponential family of densities*, which includes the Gaussian, Poisson, Binomial, Gamma, and other distributions.

2. The smooth monotonic *link function* $g(.)$. Here we have some freedom, but the so-called *canonical link* function is directly linked to the used density. As indicated in the discussion related to (1.6) no link function is needed for Gaussian data – or the link is the identity.

3. The *linear component*. See the discussion above.

In statistical modeling it is very useful to formulate the model for all n observations $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)^T$.

Let us introduce the known *model vector* $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ik})^T$ for the i^{th} observation, and unknown *parameter vector* $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_k)^T$. Then the model for all n observations can be written as

$$\begin{pmatrix} g(\mathrm{E}[Y_1]) \\ \vdots \\ g(\mathrm{E}[Y_n]) \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_n^T \end{pmatrix} \boldsymbol{\beta} \tag{1.8}$$

or

$$g(\mathrm{E}[\boldsymbol{Y}]) = \boldsymbol{X}\boldsymbol{\beta} \tag{1.9}$$

where the matrix $\boldsymbol{X}$ of known coefficients is called the *design matrix*.

As indicated in the formulation above, the parameter vector $\boldsymbol{\beta}$ is *fixed*, but *unknown*, and the typical goal is to obtain an *estimate* $\widehat{\boldsymbol{\beta}}$ of ***beta***. Models with fixed parameters are called *fixed effects models*.

Suppose that we are not interested in the individual (fixed) parameter estimates, but rather in the variation of the underlying true parameter. This leads to an introduction of the *random effects models* which will be briefly introduced in the following section.

Hierarchical models

In Chapters 5 and 6 the important concept of *hierarchical models* is introduced. The Gaussian case is introduced in Chapter 5, and this includes the so-called linear mixed effects models. This Gaussian and linear case is a natural extension of the general linear models. An extension of the generalized linear models are found in Chapter 6 which briefly introduces the generalized hierarchical models.

Let us first look at the Gaussian case. Consider for instance the test of ready made concrete. The concrete are delivered by large trucks. From a number of randomly picked trucks a small sample is taken, and these samples are analyzed with respect to the strength of concrete. A reasonable model for the variation of the strength is

$$Y_{ij} = \mu + U_i + \epsilon_{ij} \tag{1.10}$$

where μ is the overall strength of the concrete and U_i is the deviation of the average for the strength of concrete delivered by the i'th truck, and $\epsilon_{ij} \sim \mathrm{N}(0, \sigma^2)$ the deviation between concrete samples from the same truck.

Here we are typically not interested in the individual values of U_i but rather in the variation of U_i, and we will assume that $U_i \sim \mathrm{N}(0, \sigma_u^2)$.

The model (1.10) is a *one-way random effects model*. The parameters are now μ, σ_u^2 and σ^2.

Putting $\mu_i = \mu + U_i$ we may formulate (1.10) as a *hierarchical model*, where we shall assume that

$$Y_{ij}|\mu_i \sim \mathrm{N}(\mu_i, \sigma^2) \ , \tag{1.11}$$

and in contrast to the *fixed effects model*, the level μ_i is modeled as a realization of a random variable,

$$\mu_i \sim \mathrm{N}(\mu, \sigma_u^2), \tag{1.12}$$

where the μ_i's are assumed to be mutually independent, and Y_{ij} are *conditionally independent*, i.e., Y_{ij} are mutually independent in the conditional distribution of Y_{ij} for given μ_i.

Let us again consider a model for all n observations and let us further extend the discussion to the vector case of the random effects. The discussion above can now be generalized to the *linear mixed effects model* where

$$\mathrm{E}[\boldsymbol{Y}|\boldsymbol{U}] = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{U} \tag{1.13}$$

with $\boldsymbol{X}$ and $\boldsymbol{Z}$ denoting known matrices. Note how the mixed effect linear model in (1.13) is a linear combination of *fixed effects*, $\boldsymbol{X}\boldsymbol{\beta}$ and *random effects*, $\boldsymbol{Z}\boldsymbol{U}$. These types of models will be described in Chapter 5.

The non-Gaussian case of the hierarchical models, where

$$g(\mathrm{E}[\boldsymbol{Y}|\boldsymbol{U}]) = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{U} \tag{1.14}$$

and where $g(.)$ is an appropriate link function will be treated in Chapter 6.

CHAPTER 2

The likelihood principle

2.1 Introduction

Fisher (1922) identified the likelihood function as the key inferential quantity conveying all inferential information in statistical modeling including the uncertainty. In particular, Fisher suggested the method of maximum likelihood to provide a *point estimate* for the parameters of interest, the so-called *maximum likelihood estimate (MLE)*.

Example 2.1 – Likelihood function
Suppose we toss a thumbtack 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? It is clear that θ cannot be zero and the probability is unlikely to be very high. However, the probability for success $\theta = 0.3$ or $\theta = 0.4$ is likely, since in a binomial experiment with $n = 10$ and $Y = 3$, the number of successes, we get the probabilities $\mathrm{P}(Y = 3) = 0.27$ or 0.21 for $\theta = 0.3$ or $\theta = 0.4$, respectively. We have thus found a non-subjective way to compare different values of θ. By considering $\mathrm{P}_\theta(Y = 3)$ to be a function of the unknown parameter we have the *likelihood function*:

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

In a general case with n trials and y successes, the likelihood function is:

$$L(\theta) = \mathrm{P}_\theta(Y = y) = \binom{n}{y}\theta^y(1-\theta)^{n-y}.$$

A sketch of the likelihood function for $n = 10$ and $y = 3$ is shown in Figure 2.1 on the following page. As will be discussed later in the chapter, it is often more convenient to consider the log-likelihood function. The log-likelihood function is:

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

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 $\widehat{\theta}(y) = \frac{y}{n}$. In the thumbtack case where we observed $Y = y = 3$ we obtain $\widehat{\theta}(y) = 0.3$. The random variable $\widehat{\theta}(Y) = \frac{Y}{n}$ is called a maximum likelihood *estimator* for θ. Notice the difference between $\widehat{\theta}(y)$ and $\widehat{\theta}(Y)$.

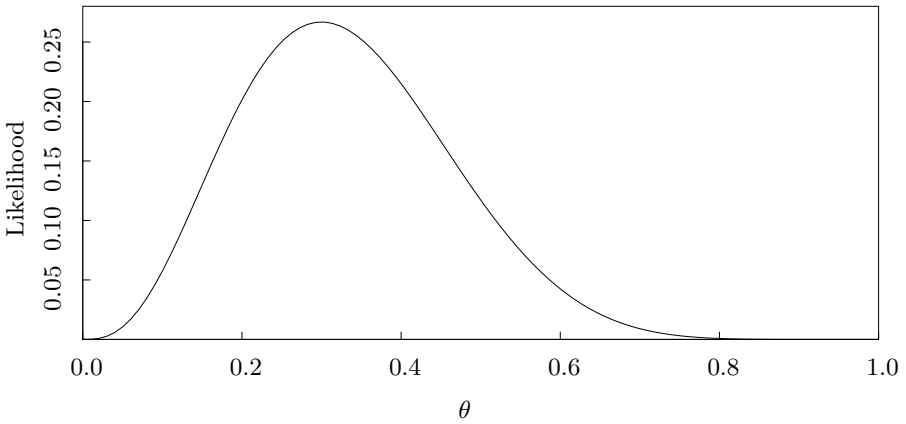


Figure 2.1: *Likelihood function of the success probability θ in a binomial experiment with $n = 10$ and $y = 3$.*

The likelihood principle is not just a method for obtaining a point estimate of parameters; it is a method for an objective reasoning with data. It is the entire likelihood function that captures all the information in the data about a certain parameter, not just its maximizer. The likelihood principle also provides the basis for a rich family of methods for selecting the most appropriate model.

Today the likelihood principles play a central role in statistical modeling and inference. Likelihood based methods are inherently computational. In general, numerical methods are needed to find the MLE.

We could view the MLE as a single number representing the likelihood function; but generally, a single number is not enough for summarising the variations of a function. If the (log-)likelihood function is well approximated by a quadratic function it is said to be *regular* and then we need at least two quantities: the location of its maximum and the curvature at the maximum. When our sample becomes large the likelihood function generally becomes regular. The curvature delivers important information about the uncertainty of the parameter estimate.

Before considering the likelihood principles in detail we shall briefly consider some theory related to point estimation.

2.2 Point estimation theory

Assume that the statistical model for the multivariate random variable, $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)^T$ is given by the parametric family of joint densities

$$\{f_Y(y_1, y_2, \ldots, y_n; \boldsymbol{\theta})\}_{\boldsymbol{\theta} \in \Theta^k} \tag{2.1}$$

with respect to some measure ν_n on $\mathcal{Y}^n$. In the following, the random variable $\boldsymbol{Y}$ will sometimes denote the observations. Assume also that we are given a realization of $\boldsymbol{Y}$ which we shall call the *observation set*, $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^T$.

We define an *estimator* as a function $\widehat{\boldsymbol{\theta}}(\boldsymbol{Y})$ of the random variable $\boldsymbol{Y}$. For given observations, $\widehat{\boldsymbol{\theta}}(\boldsymbol{y})$ is called an *estimate*. Note that an estimator is a random variable whereas an estimate is a specific number.

Example 2.2 – Estimate and Estimator
In Example 2.1 on page 9 $\widehat{\theta}(y) = y \,/\, n$ is an estimate whereas the random variable $\widehat{\theta}(Y) = Y \,/\, n$ is an estimator. In both cases they are of the maximum likelihood type.

Let us now briefly introduce some properties that are often used to describe point estimators.

Definition 2.1 – 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 $\boldsymbol{\theta} \in \Theta^k$.

Example 2.3 – Unbiased estimator
Consider again the binomial experiment from Example 2.1 where we derived the maximum likelihood estimator

$$\widehat{\theta}(Y) = \frac{Y}{n}. \tag{2.2}$$

Since

$$\mathrm{E}\left[\widehat{\theta}(Y)\right] = \frac{\mathrm{E}\,[Y]}{n} = \frac{n \cdot \theta}{n} = \theta \tag{2.3}$$

it is seen that the estimator is unbiased cf. Definition 2.1.

Another important property is *consistency*.

Definition 2.2 – Consistent estimator
An estimator is *consistent* if the sequence $\boldsymbol{\theta}_n(\boldsymbol{Y})$ of estimators for the parameter $\boldsymbol{\theta}$ converges in probability to the true value $\boldsymbol{\theta}$. Otherwise the estimator is said to be inconsistent.

For more details and more precise definitions see, e.g., Lehmann and Casella (1998) p. 332.

DEFINITION 2.3 – MINIMUM MEAN SQUARE ERROR
An estimator $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(\boldsymbol{Y})$ is said to be *uniformly minimum mean square error* if[1]

$$\mathrm{E}\left[(\widehat{\boldsymbol{\theta}}(\boldsymbol{Y}) - \boldsymbol{\theta})(\widehat{\boldsymbol{\theta}}(\boldsymbol{Y}) - \boldsymbol{\theta})^T\right] \leq \mathrm{E}\left[(\tilde{\boldsymbol{\theta}}(\boldsymbol{Y}) - \boldsymbol{\theta})(\tilde{\boldsymbol{\theta}}(\boldsymbol{Y}) - \boldsymbol{\theta})^T\right] \tag{2.4}$$

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

▸ **Remark 2.1**
In the class of unbiased estimators the minimum mean square estimator is said to be a *minimum variance unbiased estimator* (MVUE) and, furthermore, if the estimators considered are linear functions of the data, the estimator is a *best linear unbiased estimator* (BLUE). ◂

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.

THEOREM 2.1 – 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 covariance of any unbiased estimator $\widehat{\boldsymbol{\theta}}(\boldsymbol{Y})$ of $\boldsymbol{\theta}$ satisfies the inequality

$$\mathrm{Var}\left[\widehat{\boldsymbol{\theta}}(\boldsymbol{Y})\right] \geq \boldsymbol{i}^{-1}(\boldsymbol{\theta}) \tag{2.5}$$

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

$$\boldsymbol{i}(\boldsymbol{\theta}) = \mathrm{E}\left[\left(\frac{\partial \log f_Y(\boldsymbol{Y};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial \log f_Y(\boldsymbol{Y};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)^T\right] \tag{2.6}$$

and where $\mathrm{Var}\big[\widehat{\boldsymbol{\theta}}(\boldsymbol{Y})\big] = \mathrm{E}\big[(\widehat{\boldsymbol{\theta}}(\boldsymbol{Y}) - \boldsymbol{\theta})(\widehat{\boldsymbol{\theta}}(\boldsymbol{Y}) - \boldsymbol{\theta})^T\big]$. The Fisher information matrix is discussed in more detail in Section 2.5 on page 18.

Proof Since $\widehat{\boldsymbol{\theta}}(\boldsymbol{Y})$ is unbiased we have that

$$\mathrm{E}\left[\widehat{\boldsymbol{\theta}}(\boldsymbol{Y})\right] = \boldsymbol{\theta} \tag{2.7}$$

i.e.,

$$\int \widehat{\boldsymbol{\theta}}(\boldsymbol{y}) f_Y(\boldsymbol{y};\boldsymbol{\theta})\{dy\} = \boldsymbol{\theta} \tag{2.8}$$

[1]Note that the inequality should be understood in the way that the left hand side ÷ right hand side is non-negative definite.

which implies that

$$\frac{\partial}{\partial \boldsymbol{\theta}} \int \widehat{\boldsymbol{\theta}}(\boldsymbol{y}) f_Y(\boldsymbol{y};\boldsymbol{\theta})\{dy\} = \boldsymbol{I}. \tag{2.9}$$

Assuming sufficient regularity to allow for differentiation under the integral we obtain

$$\int \widehat{\boldsymbol{\theta}}(\boldsymbol{y}) \frac{\partial}{\partial \boldsymbol{\theta}} f_Y(\boldsymbol{y};\boldsymbol{\theta})\{dy\} = \boldsymbol{I} \tag{2.10}$$

or

$$\int \widehat{\boldsymbol{\theta}}(\boldsymbol{y}) \frac{\partial \log f_Y(\boldsymbol{y};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} f_Y(\boldsymbol{y};\boldsymbol{\theta})\{dy\} = \boldsymbol{I} \tag{2.11}$$

or

$$\mathrm{E}\left[\widehat{\boldsymbol{\theta}}(\boldsymbol{Y}) \frac{\partial \log f_Y(\boldsymbol{Y};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right] = \boldsymbol{I}. \tag{2.12}$$

Furthermore, we see that

$$\begin{aligned} \mathrm{E}\left[\frac{\partial \log f_Y(\boldsymbol{Y};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right] &= \int \frac{\partial \log f_Y(\boldsymbol{y};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} f_Y(\boldsymbol{y};\boldsymbol{\theta})\{dy\} \\ &= \int \frac{\partial f_Y(\boldsymbol{y};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\{dy\} = \frac{\partial}{\partial \boldsymbol{\theta}} \int f_Y(\boldsymbol{y};\boldsymbol{\theta})\{dy\} = \mathbf{0}^T. \end{aligned} \tag{2.13}$$

Using (2.12) and (2.13) we are able to find the variance (or variance covariance matrix) for $\begin{bmatrix} \widehat{\boldsymbol{\theta}}(\boldsymbol{Y}) \\ \partial \log f_Y(\boldsymbol{Y};\boldsymbol{\theta})/\partial \boldsymbol{\theta} \end{bmatrix}$.

$$\begin{aligned} \mathrm{E}\left[\begin{pmatrix} \widehat{\boldsymbol{\theta}}(\boldsymbol{Y}) - \boldsymbol{\theta} \\ (\partial \log f_Y(\boldsymbol{Y};\boldsymbol{\theta})/\partial \boldsymbol{\theta})^T \end{pmatrix} \begin{pmatrix} (\widehat{\boldsymbol{\theta}}(\boldsymbol{Y}) - \boldsymbol{\theta})^T & \partial \log f_Y(\boldsymbol{Y};\boldsymbol{\theta})/\partial \boldsymbol{\theta} \end{pmatrix}\right] \\ = \mathrm{Var}\begin{bmatrix} \widehat{\boldsymbol{\theta}}(\boldsymbol{Y}) & \boldsymbol{I} \\ \boldsymbol{I} & \boldsymbol{i}(\boldsymbol{\theta}) \end{bmatrix}. \end{aligned} \tag{2.14}$$

This variance matrix is clearly non-negative definite, and we have

$$[\boldsymbol{I} \quad \boldsymbol{i}^{-1}(\boldsymbol{\theta})] \begin{bmatrix} \mathrm{Var}\left[\widehat{\boldsymbol{\theta}}(\boldsymbol{Y})\right] & \boldsymbol{I} \\ \boldsymbol{I} & \boldsymbol{i}(\boldsymbol{\theta}) \end{bmatrix} \begin{bmatrix} \boldsymbol{I} \\ \boldsymbol{i}^{-1}(\boldsymbol{\theta}) \end{bmatrix} \geq \mathbf{0} \tag{2.15}$$

i.e.,

$$\mathrm{Var}\left[\widehat{\boldsymbol{\theta}}(\boldsymbol{Y})\right] - \boldsymbol{i}^{-1}(\boldsymbol{\theta}) \geq \mathbf{0} \tag{2.16}$$

which establishes the Cramer-Rao inequality. ∎

Definition 2.4 – Efficient estimator
An unbiased estimator is said to be *efficient* if its covariance is equal to the Cramer-Rao lower bound.

▸ **Remark 2.2 – Dispersion matrix**
The matrix $\mathrm{Var}[\widehat{\boldsymbol{\theta}}(\boldsymbol{Y})]$ is often called a variance covariance matrix since it contains variances in the diagonal and covariances outside the diagonal. This important matrix will often be termed the *Dispersion matrix* in this book. ◂

2.3 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 $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)^T$. The *likelihood* of any specific value $\boldsymbol{\theta}$ of the parameters in a model is (proportional to) the probability of the actual outcome, $Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n$, calculated for the specific value $\boldsymbol{\theta}$. The likelihood function is simply obtained by considering the likelihood as a function of $\boldsymbol{\theta} \in \Theta^k$.

DEFINITION 2.5 – LIKELIHOOD FUNCTION
Given the parametric density $f_Y(\boldsymbol{y}, \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta^k$, for the observations $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$ the *likelihood function for* $\boldsymbol{\theta}$ is the function

$$L(\boldsymbol{\theta}; \boldsymbol{y}) = c(y_1, y_2, \ldots, y_n) f_Y(y_1, y_2, \ldots, y_n; \boldsymbol{\theta}) \tag{2.17}$$

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

▸ **Remark 2.3**
The likelihood function is thus the joint probability density for the actual observations considered as a function of $\boldsymbol{\theta}$. ◂

▸ **Remark 2.4**
The likelihood function is only meaningful up to a multiplicative constant, meaning that we can ignore terms not involving the parameter. ◂

As illustrated in Example 2.1 on page 9, the likelihood function contains a measure of relative preference for various parameter values. The measure is closely linked to the assumed statistical model, but given the model the likelihood is an objective quantity that provides non-subjective measures of belief about the values of the parameter.

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

$$\ell(\boldsymbol{\theta}; \boldsymbol{y}) = \log(L(\boldsymbol{\theta}; \boldsymbol{y})) \tag{2.18}$$

where $L(\boldsymbol{\theta}; \boldsymbol{y})$ is given by (2.17). Sometimes the likelihood and the log-likelihood function will be written as $L(\boldsymbol{\theta})$ and $\ell(\boldsymbol{\theta})$, respectively, i.e., the dependency on $\boldsymbol{y}$ is not explicitly mentioned.

▸ **Remark 2.5**
It is common practice, especially when plotting, to normalize the likelihood function to have unit maximum and the log-likelihood to have zero maximum. ◂

Example 2.4 – Likelihood function for mean of normal distribution
An automatic production of a bottled liquid is considered to be stable. A sample of three bottles was selected randomly from the production and the volume of the content 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.

At 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, $\mu_Y = \mu$.

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

$$\begin{aligned} f_{Y_1,Y_2,Y_3}(y_1, y_2, y_3; \mu) = & \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{(y_1-\mu)^2}{2}\right] \\ & \times \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{(y_2-\mu)^2}{2}\right] \\ & \times \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{(y_3-\mu)^2}{2}\right] \end{aligned} \tag{2.19}$$

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

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

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

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.

Reducing the expression we find

$$L_{4.6,6.3,5.0}(\mu) = \frac{1}{(\sqrt{2\pi})^3} \exp\left[-\frac{1.58}{2}\right] \exp\left[-\frac{3(5.3-\mu)^2}{2}\right]. \tag{2.20}$$

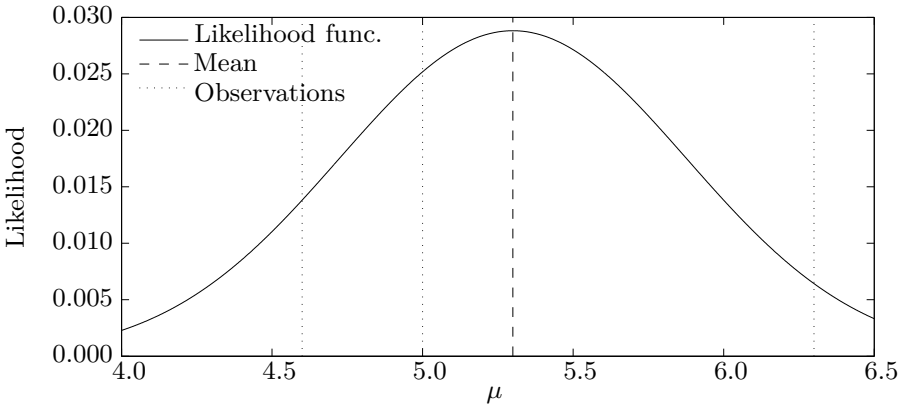


Figure 2.2: *The likelihood function for μ given the observations $y_1 = 4.6$, $y_2 = 6.3$, and $y_3 = 5.0$, as in Example 2.4.*

A sketch of the likelihood function is shown in Figure 2.2. Note that, while the probability density function (2.19) is a function of (y_1, y_2, y_3) which describes the prospective *variation* in data, the likelihood function (2.20) is a function of the unknown parameter μ, describing the relative *plausibility* or likelihood of various values of μ in light of the given data. The likelihood function indicates to which degree the various values of μ are in agreement with the given observations. Note, that the maximum value of the likelihood function (2.20) is obtained for $\hat{\mu} = 5.3$ which equals the sample mean $\overline{y} = \sum_{i=1}^{n} y_i/n$.

Sufficient statistic

The primary goal in analyzing observations is to characterize the information in the observations by a few numbers. A *statistic* $t(Y_1, Y_2, \ldots, Y_n)$ is the result of applying a function (algorithm) to the set of observations. In estimation a sufficient statistic is a statistic that contains all the information in the observations.

DEFINITION 2.6 – 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, $\boldsymbol{\theta}$, if the probability density function for $t(Y_1, Y_2, \ldots, Y_n)$ can be factorized into a product

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

with the factor $h(y_1, \ldots, y_n)$ not depending on the parameter $\boldsymbol{\theta}$, and the factor $g(t(y_1, y_2, \ldots, y_n); \boldsymbol{\theta})$ only depending on $y_1, \ldots, y_n$ through the function