Statistics 2: Statistical Learning with Likelihood and Bayes

MATH27720 Semester 2

Korbinian Strimmer

3 October 2025

Table of contents

Welcome

These are the lecture notes for MATH27720 Statistics 2, a course for second year mathematics students at the Department of Mathematics of the University of Manchester.

The course text was written by Korbinian Strimmer from 2023–2025. This version is from 3 October 2025.

If you have any questions, comments, or corrections please get in touch! [1]

Updates

The lecture notes will be updated from time to time.

The most current version is found at the web page for the

- online MATH27720 Statistics 2 lecture notes.

There you can also download the PMATH27720 Statistics 2 lecture notes as

- PDF in A4 format for printing (double page layout), or as
- 6x9 inch PDF for use on tablets (single page layout).

License

[1] Email address: korbinian.strimmer@manchester.ac.uk

Preface

About the author

Hello! My name is Korbinian Strimmer and I am a Professor in Statistics. I am a member of the Statistics group at the Department of Mathematics of the University of Manchester. You can find more information about me on my home page.

About the module

The notes are intended for the MATH27720 Statistics 2 course being taught in spring 2026 at the University of Manchester.

The MATH27720 Statistics 2 module is designed to run over the course of 10 weeks. It has the following two part structure:

1. Entropy and likelihood (W1–W5)
2. Bayesian statistics (W6–W10)

This module emphasises conceptual understanding and practical methods rather than theoretical aspects. You will specifically explore the foundations of statistical learning through likelihood and Bayesian approaches, as well as the role of entropy in supporting these concepts.

The presentation in this course is non-technical and most sections and examples will be easily accessible for a year 2 mathematics student. Sections and examples marked ▶ are slightly more complex, either conceptually or technically (e.g. involving more complicated matrix operations). These examples can be skipped during the initial reading.

If you are a student at University of Manchesterenrolled in this module, you will find additional support material on Canvas:

- a weekly learning plan,
- worksheets with examples and solutions (and R code), and
- exam papers of previous years.

Furthermore, a MATH27720 online reading list is hosted by the University of Manchester library.

> **Note 1: Aims of this module**
>
> In a nutshell, the two key aims of the MATH27720 Statistics 2 module are
>
> i) to provide a principled introduction to maximum likelihood and Bayesian statistical analysis and
> ii) to demonstrate that statistics offers a well founded and coherent theory of information, rather than just seemingly unrelated collections of "recipes" for data analysis.
>
> The first part of the module (Weeks 1–5) we will explore the **method of maximum likelihood** both practically and more theoretically in terms of its foundations.
> The second part of this module (Weeks 6–10) focuses on the **Bayesian approach to statistical estimation and inference** that can be viewed as a natural extension of likelihood-based statistical analysis that overcomes some of the limitations of maximum likelihood.

Acknowledgements

These notes are based in part on my earlier notes for MATH20802 Statistical Methods which was last run in Spring 2023. Many thanks to Beatriz Costa Gomes for her help in creating the 2019 version of the lecture notes when I was teaching the MATH20802 module for the first time and to Kristijonas Raudys for his extensive feedback on the 2020 version.

Prerequisites

Statistics is a mathematical science that requires practical working knowledge of a number of subfields in mathematics, most importantly in probability theory but also in vector and matrix algebra as well as in the calculus of functions of several variables.

The MATH27720 Statistics 2 module builds on the earlier mandatory probability and statistics modules offered by the University of Manchester such as MATH11712 Statistics 1 and MATH11711 Probability 1 as well as on MATH27720 Probability 2. Many students will also have attended the modules MATH20811 Practical Statistics and the MATH27711 Linear Regression Models.

Below you find a number of resources to help you to refresh your knowledge in these areas.

Matrices and calculus

For a refresher of the essentials in matrix algebra and calculus please refer to the supplementary Matrix and Calculus Refresher notes.

Below you find topics that are particularly relevant.

Vectors and matrices

- Vector and matrix notation
- Vector algebra
- Eigenvectors and eigenvalues, eigendecomposition
- Properties of special matrices (e.g. a real symmetric matrix has real eigenvalues)
- Positive and negative definiteness of a matrix (i.e. matrices with only positive or only negative eigenvalues)
- Matrix inverse

Functions

- Gradient vector
- Hessian matrix
- Conditions for a local extremum of a function
- Convex and concave functions
- Linear and quadratic approximation

Logarithms

In these notes the logarithmic function $\log(x)$ will always refer to the *natural logarithm* when the base is not explicitly stated. Logarithms with base 2 and based 10 are denoted as $\log_2(x)$ and $\log_{10}(x)$, respectively.

Combinatorics, probability and distributions

For a detailed introduction of concepts in combinatorics and probability see the lecture notes for MATH11711 Probability 1 and MATH27720 Probability 2.

A review of the essentials can also be found in the supplementary Probability and Distribution Refresher notes. These supplementary notes also provide details of probability distributions frequently employed in statistical analysis.

Chapter 2 briefly revisits the distributions relevant for this module.

Statistics

For a refresher of statistical concepts discussed in earlier statistics courses and required in this module see Appendix A.

Part I

Entropy and likelihood

1 Statistical learning

1.1 What is statistics?

The following fundamental questions typically arise in any scientific data analysis:

- **Optimality**: How do we extract information from data as efficiently and accurately as possible?
- **Model fit**: How can we build models that accurately reflect the observed data?
- **Interpretability**: How can we construct models that reveal underlying mechanisms and remain understandable?
- **Prediction**: How do we use these models and information to make the best possible predictions?

Statistics is a mathematical science for reasoning about data and uncertainty. It employs **probabilistic models** to address the questions above, offering a principled framework for learning from data and extracting and for processing information under uncertainty in an optimal way.

1.2 Randomness, probability and uncertainty

Random and **randomness** refer to unpredictable, non-deterministic outcomes or events. Equivalent, more technical terms are **stochastic** and **stochasticity**.

The degree of randomness (or uncertainty, see below) is quantified by the **probability**, or equivalently, by the **chance** of particular outcomes.

An interesting question is the source of the randomness. On a fundamental level, some phenomena are *intrinsically* random (e.g. radioactive decay, measurement outcomes in quantum theory). However, much *apparent* randomness arises from our ignorance of the underlying mechanisms.

The process may be deterministic in principle but we treat it as random for convenience. For example, a coin flip is often considered random. However, in reality the outcome of a coin flip is fully determined by classical physics.

Randomness that is not intrinsic but stems from a lack of knowledge or understanding, is called **uncertainty**, and corresponding events are **uncertain**. Uncertainty generally decreases if more data and information or a better model is available.

1.3 Probability theory versus statistics

When studying statistics (or any other field related to information) we should recognise the key differences between the fields of probability theory versus statistics.

On the one hand, **probability theory** provides the **mathematical underpinnings** of probability and chance (e.g. probability axioms, measure theory) and corresponding models for randomness and uncertainty (e.g. probability distributions, stochastic processes). Crucially, it is neutral about the sources of randomness and interpretations of probability, and may simply be viewed as **pure mathematics**.

On the other hand, **statistics** uses probabilistic approaches to **learn from observations**, thus linking real-world phenomena with mathematical models. Thus, it is a branch of **applied mathematics**. Importantly, statistics is concerned with **uncertainty** (e.g. about events, predictions, outcomes, model parameters) without assuming that the underlying process is actually random, and to make decisions and predictions under that uncertainty.

The link of statistics with the real world also leads to different interpretations of probability, with the two most common being the **frequentist interpretation** (**ontological**, "every probability is a long running frequency and exists independently from an observer") and the **Bayesian interpretation** (**epistemological**, "probability is a degree of belief and represent the state of knowledge"), both to be discussed later.

1.4 Sketch of statistical learning

The aim of statistical learning is to use observed data in an optimal way to learn about the underlying mechanism of the data-generating

process. Since data is typically finite but models can be in principle arbitrarily complex there may be issues of over-fitting (insufficient data for the complexity of the model) but also under-fitting (model is too simplistic).

We observe data $D = \{x_1, \ldots, x_n\}$ assumed to result from an underlying probabilistic model F, the distribution for x:

Real world		**Data**
True model (unknown)	$\longrightarrow$	Samples from true model
F		$D = \{x_1, \ldots, x_n\}$
		$x_i \sim F$

The true model underlying the data-generating process is unknown and cannot be observed. However, what we can observe is data D arising from the true model F by measuring properties of interest (our observations from experiments). Sometimes we can also perturb the model and see what the effect is (interventional study).

To explain the observed data D, and also to predict future data, we will make hypotheses in the form of candidate models $P_1, P_2, \ldots$. Often these candidate models form a model family $P_{\boldsymbol{\theta}}$ indexed by a parameter vector $\boldsymbol{\theta}$, with specific values for each model so that we can also write $P_{\boldsymbol{\theta}_1}, P_{\boldsymbol{\theta}_2}, \ldots$ for the various models.

Frequently parameters are chosen such that they allow some interpretation, such as moments or other properties of the distribution. However, intrinsically parameters are just labels and may be changed by any one-to-one transformation. For statistical learning it is necessary that models are **identifiable** within a family, i.e. each distinct model is identified by a unique parameter so that $P_{\boldsymbol{\theta}_1} = P_{\boldsymbol{\theta}_2}$ implies $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$, and conversely if $P_{\boldsymbol{\theta}_1} \neq P_{\boldsymbol{\theta}_2}$ then $\boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2$.

The various candidate models $P_{\boldsymbol{\theta}}$ in the **model world** will at best be good approximations to the true underlying data-generating model F. In some cases the true model will be part of the model family, i.e. there exists a parameter $\boldsymbol{\theta}_{\text{true}}$ so that $F = P_{\boldsymbol{\theta}_{\text{true}}}$. However, more typically we cannot assume that the true underlying model is contained in the family. Nonetheless, even an imperfect candidate model will often provide a useful mathematical approximation and capture some important characteristics of the true model and thus will help to interpret the observed data.

Statistical Learning		**Model world**
Find model(s) and parameters approximating the true model and best explaining both observed and future data $F \approx P_{\hat{\theta}}$	$\longleftarrow$	Hypotheses about data-generating process: Model P_{θ} with parameter(s) θ Ideally, true model is $F = P_{\theta_{\text{true}}}$

The aim of statistical learning is to identify the model(s) that explain the current data and also predict future data (i.e. predict outcome of experiments that have not been conducted yet).

Thus a good model provides a good fit to the current data (i.e. it explains current observations well) and also to the future data (i.e. it generalises well).

A large proportion of statistical theory is devoted to finding these "good" models that avoid both *over-fitting* (models being too complex and not generalising well) or *under-fitting* (models being too simplistic and hence also not predicting well).

Typically the aim is to find an approximating model whose **model complexity** is well matched with the complexity of the unknown true model and also with the complexity of the observed data.

1.5 Finding the best models

A core task in statistical learning is to identify those distributions that explain the existing data well and that also generalise well to future yet unseen observations.

In a **non-parametric setting** we may simply rely on the law of large numbers that implies that the empirical distribution $\hat{F}_n$ constructed from the observed data D converges to the true distribution F if the sample size is large. We can therefore obtain an **empirical estimator** $\hat{\theta}$ of the functional $\theta = g(F)$ by $\hat{\theta} = g(\hat{F}_n)$, i.e. by substituting the true distribution with the empirical distribution. This allows us, e.g., to get the empirical estimate of the mean $\mathrm{E}_F(x) = \mu$ by

$$\hat{\mathrm{E}}(x) = \hat{\mu} = \mathrm{E}_{\hat{F}_n}(x) = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$$

and of the variance $\mathrm{Var}(x) = \sigma^2 = \mathrm{E}_F((x-\mu)^2)$ by

$$\widehat{\mathrm{Var}}(x) = \widehat{\sigma^2} = \mathrm{E}_{\hat{F}_n}((x-\hat{\mu})^2) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

simply by replacing the expectation with the sample average.

For parametric models we need to find estimates of the parameters that correspond to the distributions that best approximate the unknown true data-generating model. One such approach is provided by the **method of maximum likelihood**. More precisely, given a probability distribution P_{θ} with density or mass function $p(x|\theta)$ where θ is a parameter vector, and $D = \{x_1, \ldots, x_n\}$ are the observed iid data (i.e. independent and identically distributed), the **likelihood function** is defined as

$$L_n(\theta) = L(\theta|D) = \prod_{i=1}^{n} p(x_i|\theta)$$

The parameter value $\hat{\theta}_{ML}$ that maximises the likelihood function for fixed data D is the **maximum likelihood estimate**:

$$\hat{\theta}_{ML} = \arg\max_{\theta} L_n(\theta)$$

Historically, the likelihood was introduced as the probability to observe the data given the model with specified parameters θ. However, this view is incorrect as this interpretation of the likelihood breaks down for continuous random variables which use densities rather than probabilities in the likelihood. Furthermore even for discrete random variables an additional factor accounting for the possible permutations of samples is needed to obtain the actual probability of the data. Instead, as will soon become evident, the basis of the method of maximum likelihood is fundamentally linked to entropy.

Specifically, we will see that the likelihood is closely linked to the cross-entropy between the unknown true distribution F and the model P_{θ}. As a consequence the method of **maximum likelihood extends empirical estimation to parametric models**[1]. This insight illuminates both the optimality characteristics as well as the limitations of the maximum likelihood approach to statistical learning.

[1] Conversely, empirical estimators are, in fact, also likelihood estimators based on an empirical likelihood function constructed from the empirical distribution.

1.6 Models and decomposition of uncertainty

When constructing statistical models we will often choose to explain some aspects of the uncertainty while intentionally ignoring others, to create simple yet effective models.

As a result, for any given model, uncertainty decomposes into the sum of

i) **reducible uncertainty** (epistemic / **explained by model**) and
ii) **irreducible uncertainty** (aleatoric / residual / intrinsic / **unexplained by model**).

Crucially, even with arbitrarily large amounts of data, the total uncertainty cannot always be eliminated fully, since the residual uncertainty depends on the employed model. Consequently, to reduce the unexplained uncertainty the model itself must be changed, but whether this is at all desirable is a different matter (taking into account model complexity, interpretability, etc.).

For example, in linear regression or classification, the decomposition of uncertainty is expressed by the law of total variance with decomposes **total variance** into **explained variance** ("signal", between-group) and **unexplained variance** ("noise", within-group). Additional data improves the accuracy of estimates of the residual and the explained variance but does not eliminate the model's unexplained variance. To reduce the residual error you need to change the model, e.g. by adding further covariates.

Importantly, intrinsic uncertainty should not be confused with intrinsic randomness: the latter is a claim about nature (fundamental nondeterminism), whereas the first is a property of the model (residual uncertainty). Hence, unexplained uncertainty does not imply intrinsic randomness.

1.7 A bit of history

Statistics is the oldest formal science for reasoning about data and uncertainty and relies on probabilistic models. Other theories of information emphasise approximation, algorithmic approaches, or non-probabilistic methods.

Machine learning overlaps substantially with statistics. Rooted in computer science rather than mathematics, it frequently adopts a more

engineering-centric perspective. **Artificial intelligence** (AI) is a branch of computer science that makes substantial use of statistical and machine learning techniques. The emerging field of **data science** today comprises both statistics and machine learning and brings together mathematics, computer science and area-specific applications, such as **biomedical data science**.

Some important milestones in the development of learning from data are highlighted below:

- **Bayesian statistics** dates back to Thomas Bayes's 1763 essay and was further developed by Laplace in the early 19th century.
- **Maximum likelihood** was developed by R. A. Fisher in the early 20th century, with a seminar paper published in 1922.
- Links of statistical learning with **entropy** were established in the 1940s with roots going back to discoveries in statistical physics in the 1870s. The close link of physics and statistical learning has recently been underlined by the 2024 Nobel Prize in Physics awarded to J. J. Hopfield and G. Hinton for advances in artificial neural networks.
- The first artificial **neural network** model appeared in the 1950s as a nonlinear input-output mapping, without an underlying probabilistic support. The field advanced through the 1980s and accelerated after 2010 with the rise of **deep learning**. Today neural networks are among the most popular and powerful tools for analysing complex data and form the basis of many generative AI systems. Although they began as non-probabilistic constructs, contemporary perspectives treat them as high-dimensional nonlinear statistical models.
- In the 1960s the theoretical foundations for **algorithmic or computational learning** were developed under the umbrella the Vapnik-Chernov theory, with the most prominent example of the "support vector machine" (another non-probabilistic model) devised in the 1990s. Other important advances include "ensemble learning" and corresponding algorithmic approaches to classification such as "random forests".
- Classical statistics focused on data sets with a low number of variables and a large sample size. Over the past two decades, the advent of large-scale genomics and the availability of other high-dimensional datasets has driven rapid advances in statistics and machine learning to develop new methods for analysing

high-dimensional data (many variables, possibly moderate sample sizes) and **big data** (many variables as well as large sample size).

1.8 Further reading

The book "Ten Great Ideas About Chance" by Diaconis and Skyrms (2018) offers a gentle introduction to the history and and philosophical foundations of probability as well as to the frequentist and Bayesian interpretations.

The book "Probability Theory: The Logic of Science" by Jaynes (2003) advocates that probability theory, as an extension of logic, is the natural framework for scientific reasoning.

The popular science book "The Theory That Would Not Die" McGrayne (2011) focuses on the history of Bayes' theorem and its importance in statistics. Similarly, "The Master Algorithm" by Domingos (2015) provides an informal overview over the various schools of information science.

For a quick recap of essential statistical concepts introduced in earlier statistical modules in year 1 and 2 see Appendix A.

2 Distributions for statistical models

Choosing the appropriate distributions for statistical modeling is a crucial aspect of probabilistic data analysis. This chapter explores various factors to consider when selecting suitable distributions and also reviews the key distributions covered in this module.

2.1 Common chararacterics of distributions

Distributions can be differentiated by a number of characteristics.

Firstly, by the **type of random variable**:

- discrete versus continuous
- univariate versus multivariate

Secondly, by the **support** of the random variable, with typical ranges such as:

- finite discrete support, e.g. $\{1, 2, \ldots, n\}$
- infinite discrete support, e.g. $\{1, 2, \ldots\}$
- $[0, 1]$
- $[-\infty, \infty]$
- $[0, \infty]$

The choice of support will depend on the intended use of the random variable in the model. Common applications include

- proportion
- location
- scale
- mean
- variance
- spread
- concentration

- shape
- rate
- (squared) correlation

These interpretations apply not only to the random variable itself but also to the parameter of a distribution family. For instance, we might select a distribution that allows the samples to be interpreted as proportions (such as the beta distribution). Alternatively, we may wish to choose a distribution family in which a parameter represents a proportion (such as the Bernoulli distribution).

A third consideration may be the general **shape** of the distribution:

- symmetric or asymmetric
- left or right skewed
- short tails or long tails
- unimodal or multimodal

A further characteristic of a distribution family is the **number of parameters**, with choices such as

- single parameter
- multiple parameters
- multiple types of parameters (e.g. location+scale)

A distribution **family** consists of a finite or infinite set of distributions that correspond to specific instances of parameter values.

In data analysis, our goal is to employ and develop models that are sufficiently complex to capture the essential features of the observations, while also avoiding overfitting the data. As a result, models with fewer parameters are generally preferred over those with more parameters, particularly if both exhibit similar explanatory power, i.e. have similar capacity to explain the observed data and predict future outcomes.

Lastly, it is important to take into account the general **structure** of the distribution:

- parametric versus nonparametric models
- exponential family versus non-exponential family
- special exponential families, e.g. Gibbs family, natural exponential family (NEF)

Models with simpler structure can be preferable when the sample size is small and there are fewer observations, and conversely nonparametric approaches with fewer assumptions about the data generating process may be more appropriate when there is an abundance of data.

2.2 Commonly used basic distributions

In this module we will often make use of the following common univariate distributions:

1) Binomial distribution $\text{Bin}(n, \theta)$, with support $\{0, 1, \ldots, n\}$.

   As special case ($n = 1$) is:

   - Bernoulli distribution $\text{Ber}(\theta)$, with support $\{0, 1\}$.

2) Beta distribution $\text{Beta}(\alpha, \beta)$, with support $[0, 1]$.

3) Normal distribution $N(\mu, \sigma^2)$, with support $[-\infty, \infty]$.

4) Gamma distribution $\text{Gam}(\alpha, \theta)$, with support $[0, \infty]$. It is also known as univariate Wishart distribution $\text{Wis}\left(s^2, k\right)$.

   Special cases of the gamma/Wishart distribution are:

   - scaled chi-squared distribution $s^2\chi^2_k$ (discrete k)
   - chi-squared distribution χ^2_k (discrete k, $s^2 = 1$)
   - exponential distribution $\text{Exp}(\theta)$ ($\alpha = 1$)

5) Inverse gamma distribution $\text{IG}(\alpha, \beta)$, with support $[0, \infty]$. Also know as univariate inverse Wishart distribution $\text{IW}(\psi, k)$.

All the above distributions are so-called **exponential families**. As such they can be written in a particular structural form. Exponential families have many useful properties that facilitate statistical analysis.

6) Location-scale t-distribution $t_\nu(\mu, \tau^2)$, with support $[-\infty, \infty]$.

   Special cases of the location-scale t-distribution are:

   - Student's t-distribution t_ν
   - Cauchy distribution $\text{Cau}(\mu, \tau)$

   The location-scale t-distribution is generalisation of the normal distribution but with more probability mass in the tails. Depending on the choice of the degrees of freedom ν, not all moments of the distribution may exist. Furthermore, it's not an exponential family.

For all of the above univariate distribution there exist corresponding multivariate variants. In this module we will make use of the following multivariate distributions:

1) Multinomial distribution Mult($n, \boldsymbol{\pi}$), generalising the binomial distribution.

   Special case ($n = 1$):

   - Categorical distribution Cat($\boldsymbol{\pi}$), generalising the Bernoulli distribution.

2) Multivariate normal distribution $N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, generalising the univariate normal distribution.

A distribution family can be parametrised in multiple equivalent ways. Typically, there is a standard parametrisation, and also a mean parametrisation, where one of the parameters can be interpreted as the mean. Sometimes, the same distribution is referred to by different names and there are various default parametrisations.

Importantly, any parametrisation is a matter of choice and simply provides as an alternative means to index the elementary distributions within the family. However, certain parametrisations may be more interpretable or offer computational advantages.

2.3 Choosing the right distribution

When choosing a distribution we typically aim to to align the characteristics of the distribution with those of the observations. For instance, if the data exhibit long tails, we will need to use a long-tailed model. Additionally, there may be a mechanistic rationale, such as a physical law, suggesting that the underlying process follows a particular model.

In many cases, the central limit theorem justifies using a normal distribution.

Another approach to selecting a distribution family is to fix certain properties of the distribution, such as its mean and variance, and then selecting the family that maximises the spread of the probability mass. This method is closely linked to the principle of maximum entropy, which will be discussed in more detail later. It is also helps to explain why exponential families are often preferred in statistical modelling.

2.4 Building complex statistical models

Statistical analysis often utilises models that consist of numerous random variables. In practice, these can be quite intricate, featuring hierarchical or network-like structures that connect observed and latent variables, and may also display nonlinear functional relationships. Despite their complexity, even the most sophisticated statistical models are constructed from more fundamental components.

Specifically, the large class of *graphical models* provide a principled means to form complex joint distributions for observed and unobserved random variables built from more elementary components. This include *regression models*, *mixture models* and *compound models* (continuous version of mixture models) as well as more general network-like and hierarchically structured models.

In these complex models some of the underlying elementary distributions will serve to model the observed output while others represent internal variables or account for the uncertainty regarding a parameter (in a Bayesian context).

In statistical course units in year 3 and year 4 you will discuss and learn about many types of advanced models, related for instance to

- multivariate statistics and machine learning
- temporal and spatial modelling, and
- generalised linear and nonparametric models.

Much of statistics is concerned with methods to quantify how well a model fits to data and how well it predicts future observations, and this allows to build successive models and compare them in a systematic and principled fashion.

Finally, it is worth recalling that all distributions (and models in general) are best considered as approximations of the true unknown data-generating process. Hence, the focus of any data analysis will be to find the model that captures the essential properties at an appropriate level of detail[1].

[1] The fact that it's possible to model the world at one length scale independently from what's happening at other length scales is a general phenomenon in nature known in physics as *decoupling of scales*.

2.5 Further reading

For details about the above-mentioned distributions, and their parametrisations, see the supplementary Probability and Distribution Refresher notes.

There you can also find a definition of exponential families.

3 Entropy

Entropy, a fundamental concept that originated in physics, plays a crucial role in information theory and statistical learning. This chapter introduces entropy via the route of scoring rules.

3.1 Information storage and scoring rules

Information storage units

When we record information on paper or on a computer, we need an alphabet, whose symbols act as units of stored information:

- For an alphabet of size $A = 2$ (e.g. symbols 0 and 1) the storage unit is called **bit** (binary digit). To represent $K = 256$ possible states requires $\log_2 256 = 8$ bits (or 1 byte) of storage.
- For an alphabet of size $A = 10$ (e.g. arabic numerals) the unit is called **dit** (decimal digit). To represent $K = 100$ possible states requires $\log_{10} 100 = 2$ dits.

Bit and dit are **units of information** to describe information content. One dit is equal to $\log_2 10 \approx 3.32$ bits.

If the natural logarithm is used ($A = e$) the unit of information is called **nat**. This is the standard unit of information used in statistics. It is equal to $\log_2 e \approx 1.44$ bits.

Information storage with constant code length

We now assume a discrete variable x with K possible outcomes $\Omega = \{\omega_1, \ldots, \omega_K\}$. In order to record a realised state of x using an alphabet of size A we require

$$S = \log_A K$$

units of information. The base A can be larger or smaller than K. We simply set $A = e$ and use the natural logarithm and nats as units of information throughout. S is called the **code length** or the **cost** to describe the state of x.

In the above we have tacitly assumed that storage size, code length, and cost requirements are fixed and identical for all possible K states. This can be made more explicit by writing

$$S = -\log\left(\frac{1}{K}\right)$$

where $1/K$ is the equal probability of each of the K states.

Variable code length and scoring rules

However, in practice, the K states are often not equally probable. For example x might be describing the letters in a text message, where each letter has a different frequency of occurrence (e.g. the most common letter in the English language is "e"). Hence, rather than assuming equal probabilities we use a discrete distribution P with probability mass function $p(x)$ to describe the different probabilities.

In this case, instead of assuming the same code length to describe each state, we utilise *variable code lengths*, with more probable states having shorter codes and less probable states assigned longer codes. Specifically, generalising from the previous we may wish to employ the negative logarithm to map the probability of a state x to a corresponding cost and code length:

$$S(x, P) = -\log p(x)$$

This is called **logarithmic loss** or **logarithmic scoring rule**[1] — see also Note 2 and Note 3.

As we will see below (Example 3.7, Example 3.8 and Example 3.9) using variable code lengths allows for expected code lengths that are potentially much smaller than using a fixed length $\log K$ for all states, and hence leads to a more space saving representation.

We will apply the logarithmic scoring rule $S(x, P) = -\log p(x)$ to both discrete and continuous variables x and corresponding distributions P.

[1] We follow the **convention** that **scoring rules are negatively oriented** (e.g. Dawid 2007) with the **aim to minimise the score** (cost, code length). However, some authors prefer the positively oriented convention with a reversed sign in the definition of $S(x, P)$ so the score represents a reward that is maximised (e.g. Gneiting and Raftery 2007).

As densities can take on values larger than 1 the logarithmic loss may become negative when P is a continuous distribution.

> **Note 2:** ▶ General scoring rules
>
> The function $S(x, P)$ is a **scoring rule**, a loss function that assesses a probabilistic forecast P given the observed outcome x.
> A desirable property is that the scoring rule is **proper**, i.e. its **risk** the expected score $\mathrm{E}_Q\,(S(x, P))$ is minimised when the quoted model P equals the true data-generating model Q. If the minimum is unique then the scoring rule is **strictly proper**.
> Proper scoring rules are widely used in statistics and machine learning because they permit identification of best-approximating models by risk minimisation.
> The theory of proper scoring rules is closely related to that of **Bregman divergences**.

> **Note 3:** ▶ Logarithmic scoring rule
>
> The logarithmic scoring rule $S(x, P) = -\log p(x)$ is uniquely characterised (e.g. Hartley 1928, Shannon 1948, Good 1952, Bernardo 1979). In particular, it is the only scoring rule that is both
>
> - **strictly proper** and
> - **local** with the score depending only on the value of the pdmf at x.
>
> While there are other choices of suitable scoring rules (and some common in machine learning) we will exclusively use the logarithmic scoring rule because it underpins classical likelihood inference and Bayesian learning.
> See also Example 4.2.

Example 3.1. Surprise:

The negative logarithm of a probability is called the **surprise** or **surprisal**. The surprise $-\log p$ to observe a certain outcome (with probability $p = 1$) is zero, and conversely the surprise to observe an outcome that cannot happen (with probability $p = 0$) is infinite.

Example 3.2. Logit and log-odds:

The **odds** of an event with probability p are given by the ratio $\frac{p}{1-p}$.

The **log-odds** are therefore the difference of the surprise of the complementary event (with probability $1 - p$) and the surprise of the event (with probability p):

$$\begin{aligned}\operatorname{logit}(p) &= \log\left(\frac{p}{1-p}\right) \\ &= -\log(1-p) - (-\log p)\end{aligned}$$

The log-odds function is also know as **logit** function. It maps the interval $[0, 1]$ to the interval $[-\infty, +\infty]$. Its inverse is the **logistic** function $\exp(x)/(1 + \exp(x))$ mapping the interval $[-\infty, +\infty]$ to the interval $[0, 1]$.

Example 3.3. Logarithmic score and normal distribution:

If we quote in the logarithmic scoring rule the normal distribution $P = N(\mu, \sigma^2)$ with density $p(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$ we get as score

$$S\left(x, N(\mu, \sigma^2)\right) = \frac{1}{2}\left(\log(2\pi\sigma^2) + \frac{(x-\mu)^2}{\sigma^2}\right)$$

For fixed variance σ^2 the logarithmic score is thus equivalent to the squared distance between x and μ.

3.2 Expected score and entropy

Entropy of a distribution

Given the scoring rule $S(x, P)$ we can compute its expectation assuming $x \sim P$, i.e. that the quoted model and the model for x are identical:

$$\begin{aligned}H(P) &= \mathrm{E}_P\left(S(x, P)\right) \\ &= -\mathrm{E}_P\left(\log p(x)\right)\end{aligned}$$

For the logarithmic scoring rule this is the **entropy** of the distribution P.[2]

As note, in the above the distribution P serves two distinct purposes. First, it acts as data-generating model (as indicated by the expectation E_P with regard to P). Second, it is the model that is evaluated on the observations (through $-\log p(x)$). Therefore, entropy can be viewed as a functional of the distribution P.

[2]For other proper scoring rules it is called the **generalised entropy** or the **minimum risk**.

Shannon-Gibbs entropy

The entropy of a discrete probability distribution P with probability mass function $p(x)$ with $x \in \Omega$ is called **Shannon entropy** (1948)[3]. In statistical physics, the Shannon entropy is known as the Gibbs entropy (1878):

$$H(P) = -\sum_{x\in\Omega} \log p(x)\, p(x)$$

The entropy of a discrete distribution is the **expected surprise**. We can also interpret it as the **expected cost** or **expected code length** when the data are generated according to model P ("sender", "encoder") and we are using the same model P to describe the data ("receiver", "decoder").

Furthermore, the Shannon-Gibbs entropy also has a combinatorial interpretation (see Example 3.11).

As $p(x) \in [0, 1]$ and hence $-\log p(x) \geq 0$, so the Shannon-Gibbs entropy is bounded below and is non-negative.

Differential entropy

Applying the definition of entropy to a continuous probability distribution P with density $p(x)$ yields the **differential entropy**:

$$H(P) = -\int_x \log p(x)\, p(x)\, dx$$

Differential entropy can be negative because the logarithm is applied to a density, which, unlike a probability, can take on values greater than one.

Moreover, for continuous random variables, the shape of the density typically changes under **variable transformation**, such as from x to y, the **differential entropy will change** as well and is **not invariant** under such a transformation so that $H(P_y) \neq H(P_x)$.

[3]Shannon, C. E. 1948. A mathematical theory of communication. Bell System Technical Journal **27**:379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

> **Note 4:** ▶ Interpretation of entropy — spread versus disorder
>
> Traditionally, entropy has been considered as a measure of order, with large entropy corresponding to disorder and low entropy to order. However, this interpretation is now viewed as outdated and in fact misleading as many apparently ordered systems have large entropy and some disordered-looking systems have low entropy. A better and more useful intuition is that entropy measures **spread** or **dispersal**. This notion is now preferred in Physics[4] and it also aligns with the interpretation of entropy in Statistics and Machine Learning.
> As will become clear from the examples, entropy quantifies the **spread of the probability mass**. When the probability mass is concentrated within a specific area, the entropy is low; conversely, when the probability mass is more broadly distributed, the entropy is high.

3.3 Entropy examples

Models with single parameter

Example 3.4. The Shannon-Gibbs entropy of the geometric distribution $F_x = \text{Geom}(\theta)$ with probability mass function $p(x|\theta) = \theta(1-\theta)^{x-1}$, $\theta \in [0,1]$, support $x \in \{1,2,\ldots\}$ and $\text{E}(x) = 1/\theta$ is

$$\begin{aligned} H(F_x) &= -\text{E}\left(\log\theta + (x-1)\log(1-\theta)\right) \\ &= -\log\theta + \left(\frac{1}{\theta} - 1\right)\log(1-\theta) \\ &= -\frac{\theta\log\theta + (1-\theta)\log(1-\theta)}{\theta} \end{aligned}$$

Using the identity $0 \times \log(0) = 0$ we see that the entropy of the geometric distribution for $\theta = 1$ equals 0, i.e. it achieves the minimum possible Shannon-Gibbs entropy. Conversely, as $\theta \to 0$ it diverges to infinity.

[4] For example, see: H. S. Leff. 2007. *Entropy, its language, and interpretation*. Found. Phys. **37**: 1744–1766, D. S. Lemons. 2013. *A student's guide to entropy*. CUP, and https://en.wikipedia.org/wiki/Entropy_(energy_dispersal)

Example 3.5. Consider the uniform distribution $F_x = \text{Unif}(0,a)$ with $a > 0$, support $x \in [0,a]$ and density $p(x) = 1/a$. The corresponding differential entropy is

$$\begin{aligned} H(F_x) &= -\int_0^a \log\left(\frac{1}{a}\right)\frac{1}{a}dx \\ &= \log a \int_0^a \frac{1}{a}dx \\ &= \log a\,. \end{aligned}$$

Note that for $0 < a < 1$ the differential entropy is negative.

Example 3.6. Starting with the uniform distribution $F_x = \text{Unif}(0,a)$ from Example 3.5 the variable x is changed to $y = x^2$ yielding the distribution F_y with support from 0 to a^2 and density $p(y) = 1/\left(2a\sqrt{y}\right)$.

The corresponding differential entropy is

$$\begin{aligned} H(F_y) &= \int_0^{a^2} \log\left(2a\sqrt{y}\right)\, 1/\left(2a\sqrt{y}\right)\, dy \\ &= \left[\sqrt{y}/a\,\left(\log\left(2a\sqrt{y}\right) - 1\right)\right]_{y=0}^{y=a^2} \\ &= \log\left(2a^2\right) - 1\,. \end{aligned}$$

This is negative for $0 < a < \sqrt{e/2} \approx 1.1658$. As expected $H(F_y) \neq H(F_x)$ as differential entropy is not invariant against variable transformations.

Models with multiple parameters

Example 3.7. Entropy of the categorical distribution P with K categories.

Assuming class probabilities $p_1, \ldots, p_K$ the Shannon-Gibbs entropy is

$$H(P) = -\sum_{k=1}^{K} \log(p_k)\, p_k$$

As P is discrete $H(P)$ is bounded below by 0. Furthermore, it is also bounded above by $\log K$ (cf. Example 6.1). Hence for a categorical distribution P with K categories we have

$$0 \leq H(P) \leq \log K$$

The maximum ($\log K$) is achieved for the discrete uniform distribution (Example 3.8) and the minimum (0) for a concentrated categorical distribution (Example 3.9).

Example 3.8. Entropy of the discrete uniform distribution U_K:

Let $p_1 = p_2 = \ldots = p_K = \frac{1}{K}$. Then

$$H(U_K) = -\sum_{k=1}^{K} \log\left(\frac{1}{K}\right) \frac{1}{K} = \log K$$

Note that $\log K$ is the largest value the Shannon-Gibbs entropy can assume with K classes (cf. Example 6.1) and indicates maximum spread of probability mass.

Example 3.9. Entropy of a categorical distribution with concentrated probability mass:

Let $p_1 = 1$ and $p_2 = p_3 = \ldots = p_K = 0$. Using $0 \times \log(0) = 0$ we obtain for the Shannon-Gibbs entropy

$$H(P) = \log(1) \times 1 + \log(0) \times 0 + \cdots = 0$$

Note that 0 is the smallest value that Shannon-Gibbs entropy can assume and that it corresponds to maximum concentration of probability mass.

Example 3.10. Differential entropy of the normal distribution:

The log density of the univariate normal $N(\mu, \sigma^2)$ distribution is $\log p(x|\mu, \sigma^2) = -\frac{1}{2}\left(\log(2\pi\sigma^2) + \frac{(x-\mu)^2}{\sigma^2}\right)$ with $\sigma^2 > 0$. The corresponding differential entropy is with $\mathrm{E}((x-\mu)^2) = \sigma^2$

$$\begin{aligned} H(P) &= -\mathrm{E}\left(\log p(x|\mu, \sigma^2)\right) \\ &= \frac{1}{2}\left(\log(2\pi\sigma^2) + 1\right) . \end{aligned}$$

Crucially, the entropy of a normal distribution depends only on its variance σ^2, not on its mean μ. This is intuitively clear as the variance controls the concentration of the probability mass. A large variance means that the probability mass is more spread out and thus less concentrated around the mean.

For $\sigma^2 < 1/(2\pi e) \approx 0.0585$ the differential entropy is negative.

Example 3.11. ▶ Entropy and the multinomial coefficient:

Let $\hat{Q}$ be the empirical categorical distribution with $\hat{q}_k = n_k/n$ the observed frequencies with n_k counts in class k and $n = \sum_{k=1}^{K}$ total counts.

The number of possible permutation of n items of K distinct types is given by the multinomial coefficient

$$W = \binom{n}{n_1, \ldots, n_K} = \frac{n!}{n_1! \times n_2! \times \ldots \times n_K!}$$

It turns out that for large n both quantities are directly linked:

$$H(\hat{Q}) \approx \frac{1}{n} \log W$$

Recall the Moivre-Sterling formula which for large n allow to approximate the factorial by

$$\log n! \approx n \log n - n$$

With this

$$\begin{aligned} \log W &= \log n! - \sum_{k=1}^{K} \log n_k! \\ &\approx n \log n - n - \sum_{k=1}^{K} (n_k \log n_k - n_k) \\ &= \sum_{k=1}^{K} n_k \log n - \sum_{k=1}^{K} n_k \log n_k \\ &= -n \sum_{k=1}^{K} \frac{n_k}{n} \log\left(\frac{n_k}{n}\right) \\ &= -n \sum_{k=1}^{K} \log(\hat{q}_k)\, \hat{q}_k \\ &= nH(\hat{Q}) \end{aligned}$$

The above combinatorial derivation of entropy is one of the cornerstones of statistical mechanics and is credited to Boltzmann (1877) and Gibbs (1878). The number of elements $n_1, \ldots, n_K$ in each of the K classes corresponds to the macrostate and any of the W different allocations

of the n elements to the K classes to an underlying microstate. The multinomial coefficient, and hence entropy, is largest when there are only small differences (or none) among the n_i, i.e. when the individual elements are equally spread across the K bins.

In statistics the above derivation of entropy was rediscovered by Wallis (1962).

A bit of history

The concept of entropy was first introduced in 1865 by Rudolph Clausius (1822-1888) in the context of thermodynamics. In physics entropy measures the dispersal of energy in a system. If energy is concentrated (and capacity for work is high) then the entropy is low, and conversely if energy is spread out (and capacity for work is low) the entropy is large. The total energy is conserved[5] (first law of thermodynamics) but with time it will diffuse and thus entropy will increase with time (and capacity for work will decrease) (second law of thermodynamics).

The modern probabilistic definition of entropy was discovered in the 1870s by Ludwig Boltzmann (1844–1906) and Josiah W. Gibbs (1839–1903). In statistical mechanics entropy is proportional to the logarithm of the number of microstates (i.e. particular configurations of the system) compatible with the observed macrostate. Typically, in systems where the energy is spread out there are very large numbers of compatible configurations hence this corresponds to large entropy, and conversely, if the energy is concentrated there are only few such configurations, and thus it corresponds to low entropy.

In the 1940–1950's the notion of entropy turned out to be central also in information theory, a field pioneered by mathematicians such as Ralph Hartley (1888–1970), Solomon Kullback (1907–1994), Alan Turing (1912–1954), Richard Leibler (1914–2003), Irving J. Good (1916–2009), Claude Shannon (1916–2001), and Edwin T. Jaynes (1922–1998), and later further explored by Shun'ichi Amari (1936–), Imre Ciszár (1938–), Bradley Efron (1938–), Philip Dawid (1946–) and many others.

Of the above, Turing and Good were affiliated with the University of Manchester in the 1940–50s.

[5] Energy conservation itself arises as a consequence of the time-translation symmetry of physical laws, see Noether's theorem.

4 Relative entropy

4.1 Cross-entropy

Definition of cross-entropy

Given the scoring rule $S(x, P)$ we now compute its expectation assuming $x \sim Q$, i.e. with regard to another distribution Q:

$$\begin{aligned} H(Q, P) &= \mathrm{E}_Q\left(S(x, P)\right) \\ &= -\mathrm{E}_Q\left(\log p(x)\right) \end{aligned}$$

For the logarithmic scoring rule this is called the **cross-entropy**[1] [2].

In the above the distribution Q represent the data-generating process (note the expectation E_Q with regard to Q) and the distribution P is the the model that is evaluated on the observations (via $-\log p(x)$). Thus, cross-entropy is a functional of two distributions Q and P.

For two discrete distributions Q and P with probability mass functions $q(x)$ and $p(x)$ with $x \in \Omega$ the cross-entropy is computed as the weighted sum

$$H(Q, P) = -\sum_{x \in \Omega} \log p(x)\, q(x)$$

It can be interpreted as the expected cost or expected code length when the data are generated according to model Q ("sender", "encoder") and but we use model P to describe the data ("receiver", "decoder").

For two continuous distributions Q and P with densities $q(x)$ and $p(x)$ we compute the integral

$$H(Q, P) = -\int_x \log p(x)\, q(x)\, dx$$

[1] For other scoring rules it is called the **risk** of P under the true model Q.

[2] This follows the current accepted usage. However, in some (typically older) literature the term cross-entropy may refer instead to the KL divergence.

Example 4.1. Cross-entropy between two normals:

Assume $F_{\text{ref}} = N(\mu_{\text{ref}}, \sigma^2_{\text{ref}})$ and $F = N(\mu, \sigma^2)$. The cross-entropy $H(F_{\text{ref}}, F)$ is

$$\begin{aligned} H(F_{\text{ref}}, F) &= -\mathrm{E}_{F_{\text{ref}}}\left(\log p(x|\mu, \sigma^2)\right) \\ &= \frac{1}{2}\mathrm{E}_{F_{\text{ref}}}\left(\log(2\pi\sigma^2) + \frac{(x-\mu)^2}{\sigma^2}\right) \\ &= \frac{1}{2}\left(\frac{(\mu-\mu_{\text{ref}})^2}{\sigma^2} + \frac{\sigma^2_{\text{ref}}}{\sigma^2} + \log(2\pi\sigma^2)\right) \end{aligned}$$

using $\mathrm{E}_{F_{\text{ref}}}((x-\mu)^2) = (\mu_{\text{ref}} - \mu)^2 + \sigma^2_{\text{ref}}$.

Properties of cross-entropy

- Cross-entropy is not symmetric with regard to Q and P, because the expectation is taken with reference to Q.
- If both distributions Q and P are identical, cross-entropy reduces to entropy, i.e. $H(Q, Q) = H(Q)$.
- Like differential entropy **cross-entropy changes under variable transformation** for continuous random variables, say from x to y, hence $H(Q_y, P_y) \neq H(Q_x, P_x)$.

Gibbs' inequality

A crucial further property of the cross-entropy $H(Q, P)$ is that it is bounded below by the entropy of Q, therefore

$$H(Q, P) \geq H(Q)$$

with equality only if $Q = P$. This is known as **Gibbs' inequality**.

This follows from **Jensen's inequality**. For details see Worksheet E1.

Essentially this means that when data are generated (encoded) under model Q and described (decoded) using model P there is always an extra cost, or penalty, to employ the approximating model P rather than the correct model Q .

Example 4.2. ▶ Logarithmic scoring rule is strictly proper:

The cross-entropy $H(Q, P)$ is the risk associated with the logarithmic scoring rule $S(x, P) = -\log p(x)$. Gibbs' inequality states that the cross-entropy $H(Q, P)$ is uniquely minimised at $P = Q$, with the minimum risk being the entropy $H(Q)$ of the data-generating model Q. Therefore, the logarithmic scoring rule is *strictly proper* (see Note 2).

Example 4.3. Normal differential entropy as lower bound of normal cross-entropy:

Revisit the cross-entropy $H(F_{\text{ref}}, F)$ in Example 4.1. Setting $\mu_{\text{ref}} = \mu$ and $\sigma^2_{\text{ref}} = \sigma^2$ the normal cross-entropy degenerates to the normal differential entropy $H(F_{\text{ref}}) = \frac{1}{2}\left(\log(2\pi\sigma^2_{\text{ref}}) + 1\right)$ as obtained in Example 3.10.

4.2 Boltzmann relative entropy and KL divergence

Boltzmann entropy aka relative entropy

The **Boltzmann entropy** of a distribution Q relative to a distribution P is given by

$$\begin{aligned} B(Q, P) &= H(Q) - H(Q, P) \\ &= -\mathrm{E}_Q \log\left(\frac{q(x)}{p(x)}\right) \end{aligned}$$

The Boltzmann entropy is also known as **relative entropy**.

As a consequence of the Gibbs's inequality we see that the Boltzmann entropy is always non-positive, $B(Q, P) \leq 0$. In Example 4.9 it is shown that $B(Q, P)$ can be interpreted as a log-probability.

By construction, the Boltzmann entropy of Q relative to a uniform distribution U (with constant probability density mass function) is

$$B(Q, U) = H(Q) + \text{const.}$$

i.e. it is equal to the entropy of Q apart from a constant.

Definition of KL divergence

The **KL divergence** is defined as the negative of the Boltzmann relative entropy as

$$\begin{aligned} D_{\mathrm{KL}}(Q,P) &= H(Q,P) - H(Q) \\ &= \mathrm{E}_Q \log\left(\frac{q(x)}{p(x)}\right) \end{aligned}$$

As a consequence of the Gibbs inequality the KL divergence is always non-negative: $D_{\mathrm{KL}}(Q,H) \geq 0$.[3]

The KL divergence $D_{\mathrm{KL}}(Q,P)$ can be interpreted as the additional cost if model P is used instead Q to describe data from Q. If Q and P are identical there is no extra cost and $D_{\mathrm{KL}}(Q,P) = 0$. However, if they are not identical then there is an additional cost and $D_{\mathrm{KL}}(Q,P) > 0$.

$D_{\mathrm{KL}}(Q,P)$ thus measures the **divergence**[4] between the two distributions P and Q. The use of the term "divergence" rather than "distance" is a reminder that Q and P are not interchangeable in $D_{\mathrm{KL}}(Q,P)$.

Properties of KL divergence and Boltzmann relative entropy

Boltzmann relative entropy and KL divergence differ only by sign and thus share a number of key properties inherited from cross-entropy:

1. $D_{\mathrm{KL}}(Q,P) \neq D_{\mathrm{KL}}(Q,P)$, i.e. the KL divergence is not symmetric, Q and P cannot be interchanged. This follows from the same property of cross-entropy.
2. $D_{\mathrm{KL}}(Q,P) \geq 0$, follows from Gibbs' inequality and proof via **Jensen's inequality**.
3. $D_{\mathrm{KL}}(Q,P) = 0$ if and only if $P = Q$, i.e., the KL divergence is zero if and only if Q and P are identical. Also follows from Gibbs' inequality.

[3] For any proper scoring rule $S(x,P)$ the associated **divergence** or **discrepancy** is defined in the same fashion: $D(Q,P) = \mathrm{E}_Q\left(S(x,P)\right) - \mathrm{E}_Q\left(S(x,Q)\right) \geq 0$, i.e. the difference between the risk and the minimum risk. Such divergences closely correspond to **Bregman divergences**.

[4] Note that divergence between distributions is unrelated to the divergence vector operator from vector calculus.

For more details and proofs of properties 2 and 3 see Worksheet E1.

Typically, we wish to minimise KL divergence $D_{\mathrm{KL}}(Q,P)$ and maximise Boltzmann relative entropy $B(Q,P)$.

Invariance and data processing properties

A further crucial property of KL divergence is its **invariance property**:

4. $D_{\mathrm{KL}}(Q,P)$ is **invariant under general invertible variable transformations** , so that $D_{\mathrm{KL}}(Q_y,P_y) = D_{\mathrm{KL}}(Q_x,P_x)$ under a change of random variable from x to y. Hence, KL divergence does not change when the sample space is reparametrised.

This is a remarkable property[5] [6] as it holds not just for discrete but also for continuous random variables. For comparison, recall that both differential entropy as well as cross-entropy for continuous random variables are not transformation invariant.

In the definition of KL divergence the expectation is taken over a *ratio of two densities*. The invariance is created because the Jacobian determinant changes both densities in the same way under variable transformation and thus cancel out. For more details and more formal proof of the invariance property see Worksheet E1.

More broadly, the KL divergence satisfies the **data processing inequality**[7] i.e. applying a stochastic or deterministic transformation to the underlying random variables cannot increase the KL divergence $D_{\mathrm{KL}}(Q,P)$ between Q and P. Thus, by processing data you cannot increase information about which distribution generated the data.

Coordinate transformations can be viewed as a special case of data processing, and for $D_{\mathrm{KL}}(Q,P)$ the data-processing inequality under general invertible transformations becomes an identity.

[5] Even more striking, this property only holds for the KL divergence, not for any other divergence induced by a proper scoring rule, making the KL divergence the sole invariant Bregman divergence.

[6] Furthermore, the KL divergence is also the only f-divergence (of which the KL divergence is a principal example) that is invariant against coordinate transformations.

[7] The data processing inequality also holds for all f-divergences but is notably *not* satisfied by divergences of other proper scoring rules (and thus other Bregman divergences).

Further properties

The KL divergence and cross-entropy inherit a number of further useful properties from proper scoring rules. We will not cover these in this text. For example, there are various **decompositions** for the risk and the divergence satisfies a **generalised Pythagorean theorem**.

In summary, KL divergence stands out among divergences between distributions due to many valuable and in some cases unique properties. It is therefore not surprising that it plays a central role in statistics and machine learning.

Origin of Boltzmann relative entropy and KL divergence and naming conventions

Boltzmann relative entropy was first discovered by Boltzmann (1878)[8] in physics in a discrete setting in the context of statistical mechanics (see Example 4.9). In statistics and information theory KL divergence was formally introduced by Kullback and Leibler (1951)[9]. Good (1979)[10] credits Turing with the first statistical application in 1940/1941 in the field of cryptography.

The KL divergence is also known as **KL information** or **KL information number** named after two of the original authors (Kullback and Leibler) who themselves referred to this quantity as **discrimination information**. Another common name is **information divergence** or short ***I*-divergence**. Some authors (e.g. Efron) call twice the KL divergence $2D_{\mathrm{KL}}(Q,P) = D(Q,P)$ the **deviance** of P from Q.

There also exist various notations for KL divergence in the literature. Here we use $D_{\mathrm{KL}}(Q,P)$ but you will often find both $\mathrm{KL}(Q||P)$ and $I^{KL}(Q;P)$.

Especially in older literature the KL divergence is also referred to as "cross-entropy". This use is outdated and only leads to confusion with the related but different definition of cross-entropy above.

[8] Boltzmann, L. 1878. Weitere Bemerkungen über einige Probleme der mechanischen Wärmetheorie. Wien Ber. **78**:7–46. https://doi.org/10.1017/CBO9781139381437.013

[9] Kullback, S., and R. A. Leibler. 1951. On information and sufficiency. Ann. Math. Statist. **22** 79–86. https://doi.org/10.1214/aoms/1177729694

[10] Good, I. J. 1979. Studies in the history of probability. XXXVII. A. M. Turing's statistical work in world war II. Biometrika, 66:393–396. https://doi.org/10.1093/biomet/66.2.393

Furthermore, KL divergence is also frequently referred to as "relative entropy" however this use also leads to confusion as KL divergence is normally minimised whereas entropy and relative entropy (i.e. Boltzmann entropy) is normally maximised. Shannon (1948) defined "relative entropy" yet differently again as the ratio of Shannon-Gibbs entropy relative to its maximum value, i.e. as standardised entropy.

In this text relative entropy always refers to **Boltzmann entropy** with the opposite orientation compared to **KL divergence**.

4.3 KL divergence examples

Models with a single parameter

Example 4.4. KL divergence between two Bernoulli distributions $\mathrm{Ber}(\theta_1)$ and $\mathrm{Ber}(\theta_2)$:

The "success" probabilities for the two distributions are θ_1 and θ_2, respectively, and the complementary "failure" probabilities are $1-\theta_1$ and $1-\theta_2$. With this we get for the KL divergence

$$D_{\mathrm{KL}}(\mathrm{Ber}(\theta_1), \mathrm{Ber}(\theta_2)) = \theta_1 \log\left(\frac{\theta_1}{\theta_2}\right) + (1-\theta_1)\log\left(\frac{1-\theta_1}{1-\theta_2}\right)$$

Example 4.5. KL divergence between two univariate normals with different means and common variance:

Assume $F_{\mathrm{ref}} = N(\mu_{\mathrm{ref}}, \sigma^2)$ and $F = N(\mu, \sigma^2)$. Setting $\sigma^2_{\mathrm{ref}} = \sigma^2$ in the more general case of the KL divergence between two normals (Example 4.8) yields

$$D_{\mathrm{KL}}(F_{\mathrm{ref}}, F) = \frac{1}{2\sigma^2}(\mu - \mu_{\mathrm{ref}})^2$$

which, apart from a scale factor, is the **squared Euclidean distance** or **squared loss** between the two means μ_{ref} and μ. Note that in this case the KL divergence is symmetric with regard to the two mean parameters.

Example 4.6. KL divergence between two univariate normals with common mean and different variances:

Assume $F_{\mathrm{ref}} = N(\mu, \sigma^2_{\mathrm{ref}})$ and $F = N(\mu, \sigma^2)$. Setting $\mu_{\mathrm{ref}} = \mu$ in the more general case of the KL divergence between two normals (Example 4.8) yields

$$D_{\mathrm{KL}}(F_{\mathrm{ref}}, F) = \frac{1}{2}\left(\frac{\sigma^2_{\mathrm{ref}}}{\sigma^2} - \log\left(\frac{\sigma^2_{\mathrm{ref}}}{\sigma^2}\right) - 1\right)$$

This is a convex function of the ratio $\sigma^2_{\mathrm{ref}}/\sigma^2$ of the two variances. Apart from the scale factor this is known as **Stein's loss** between the two variances σ^2_{ref} and σ^2.

Models with multiple parameters

Example 4.7. KL divergence between two categorical distributions with K classes:

With $Q = \mathrm{Cat}(\boldsymbol{q})$ and $P = \mathrm{Cat}(\boldsymbol{p})$ and corresponding probabilities $q_1, \ldots, q_K$ and $p_1, \ldots, p_K$ satisfying $\sum_{i=1}^K q_i = 1$ and $\sum_{i=1}^K p_i = 1$ we get:

$$D_{\mathrm{KL}}(Q, P) = \sum_{i=1}^{K} q_i \log\left(\frac{q_i}{p_i}\right)$$

To be explicit that there are only $K-1$ parameters in a categorical distribution we can also write

$$D_{\mathrm{KL}}(Q, P) = \sum_{i=1}^{K-1} q_i \log\left(\frac{q_i}{p_i}\right) + q_K \log\left(\frac{q_K}{p_K}\right)$$

with $q_K = \left(1 - \sum_{i=1}^{K-1} q_i\right)$ and $p_K = \left(1 - \sum_{i=1}^{K-1} p_i\right)$.

Example 4.8. KL divergence between two univariate normals with different means and variances:

Assume $F_{\mathrm{ref}} = N(\mu_{\mathrm{ref}}, \sigma^2_{\mathrm{ref}})$ and $F = N(\mu, \sigma^2)$. Then with Example 3.10 (entropy of normal) and Example 4.1 (cross-entropy between two normals) we get

$$\begin{aligned} D_{\mathrm{KL}}(F_{\mathrm{ref}}, F) &= H(F_{\mathrm{ref}}, F) - H(F_{\mathrm{ref}}) \\ &= \frac{1}{2}\left(\frac{(\mu - \mu_{\mathrm{ref}})^2}{\sigma^2} + \frac{\sigma^2_{\mathrm{ref}}}{\sigma^2} - \log\left(\frac{\sigma^2_{\mathrm{ref}}}{\sigma^2}\right) - 1\right) \end{aligned}$$

For the two special cases of equal variances ($\sigma_{\mathrm{ref}} = \sigma^2$) and equal means ($\mu_{\mathrm{ref}} = \mu^2$) see Example 4.5 and Example 4.6.

Example 4.9. ▶ Boltzmann relative entropy as log-probability:

Assume $\hat{Q}$ is an empirical categorical distribution based on observed counts n_k (see Example 3.11) and P is a second categorical distribution.

The KL divergence is then

$$\begin{aligned} B(\hat{Q}, P) &= H(\hat{Q}) - H(\hat{Q}, P) \\ &= H(\hat{Q}) + \sum_{i=1}^{K} \log(p_i)\, \hat{q}_i \\ &= H(\hat{Q}) + \frac{1}{n}\sum_{i=1}^{K} n_i \log p_i \end{aligned}$$

For large n we may use the multinomial coefficient $W = \binom{n}{n_1, \ldots, n_K}$ to obtain the entropy of $\hat{Q}$ (see Example 3.11). This results in

$$\begin{aligned} B(\hat{Q}, P) &\approx \frac{1}{n}\left(\log W + \sum_{i=1}^{K} n_i \log p_i\right) \\ &= \frac{1}{n}\log\left(W \times \prod_{i=1}^{K} p_i^{n_i}\right) \\ &= \frac{1}{n}\log \Pr(n_1, \ldots, n_K | \boldsymbol{p}) \end{aligned}$$

Hence the Boltzmann relative entropy is directly linked to the multinomial probability of the observed counts $n_1, \ldots, n_k$ under the model P. This derivation of the Boltzmann relative entropy as log-probability of a macrostate is due to Boltzmann (1878). See also Akaike (1985) for a historical account.

5 Expected Fisher information

5.1 Expected Fisher information

Definition of expected Fisher information

KL information measures the divergence of two distributions. Previously we have seen examples of KL divergence between two distributions belonging to the same family. We now consider the KL divergence of two such distributions separated in parameter space only by some small $\boldsymbol{\varepsilon}$.

Specifically, we consider the function

$$\begin{aligned} h(\boldsymbol{\theta} + \boldsymbol{\varepsilon}) &= D_{\mathrm{KL}}(F_{\boldsymbol{\theta}}, F_{\boldsymbol{\theta}+\boldsymbol{\varepsilon}}) \\ &= \mathrm{E}_{F_{\boldsymbol{\theta}}}\left(\log f(\boldsymbol{x}|\boldsymbol{\theta}) - \log f(\boldsymbol{x}|\boldsymbol{\theta} + \boldsymbol{\varepsilon})\right) \end{aligned}$$

where $\boldsymbol{\theta}$ is kept constant and $\boldsymbol{\varepsilon}$ is varying. Assuming that $f(\boldsymbol{x}|\boldsymbol{\theta})$ is twice differentiable with regard to $\boldsymbol{\theta}$ we can approximate $h(\boldsymbol{\theta}+\boldsymbol{\varepsilon})$ quadratically by

$$h(\boldsymbol{\theta} + \boldsymbol{\varepsilon}) \approx h(\boldsymbol{\theta}) + \nabla h(\boldsymbol{\theta})^T \boldsymbol{\varepsilon} + \frac{1}{2}\boldsymbol{\varepsilon}^T \nabla\nabla^T h(\boldsymbol{\theta})\, \boldsymbol{\varepsilon}$$

From the properties of the KL divergence we know that $D_{\mathrm{KL}}(F_{\boldsymbol{\theta}}, F_{\boldsymbol{\theta}+\boldsymbol{\varepsilon}}) \geq 0$ and that it becomes zero only if $\boldsymbol{\varepsilon} = 0$. Thus, by construction the function $h(\boldsymbol{\theta} + \boldsymbol{\varepsilon})$ achieves for $\boldsymbol{\varepsilon} = 0$

i) a true minimum with $h(\boldsymbol{\theta}) = 0$,
ii) a vanishing gradient with $\nabla h(\boldsymbol{\theta}) = 0$, and
iii) a positive definite Hessian matrix with $\nabla\nabla^T h(\boldsymbol{\theta}) = -\mathrm{E}_{F_{\boldsymbol{\theta}}} \nabla\nabla^T \log f(\boldsymbol{x}|\boldsymbol{\theta})$.

Therefore in the quadratic approximation of $h(\boldsymbol{\theta} + \boldsymbol{\varepsilon})$ around $\boldsymbol{\theta}$ above the first two terms (constant and linear) vanish and only the quadratic term remains. The Hessian matrix evaluated at $\boldsymbol{\theta}$

$$\boldsymbol{I}^{\mathrm{Fisher}}(\boldsymbol{\theta}) = -\mathrm{E}_{F_{\boldsymbol{\theta}}} \nabla\nabla^T \log f(\boldsymbol{x}|\boldsymbol{\theta})$$

is called **expected Fisher information** for $\boldsymbol{\theta}$, or short **Fisher information**. Hence, the KL divergence can be locally approximated by

$$D_{\mathrm{KL}}(F_{\boldsymbol{\theta}}, F_{\boldsymbol{\theta}+\boldsymbol{\varepsilon}}) \approx \frac{1}{2}\boldsymbol{\varepsilon}^T \boldsymbol{I}^{\mathrm{Fisher}}(\boldsymbol{\theta})\boldsymbol{\varepsilon}$$

We may also vary the first argument in the KL divergence. It is straightforward to show that this leads to the same approximation to second order in $\boldsymbol{\varepsilon}$:

$$D_{\mathrm{KL}}(F_{\boldsymbol{\theta}+\boldsymbol{\varepsilon}}, F_{\boldsymbol{\theta}}) \approx \frac{1}{2}\boldsymbol{\varepsilon}^T \boldsymbol{I}^{\mathrm{Fisher}}(\boldsymbol{\theta})\, \boldsymbol{\varepsilon}$$

Hence, the KL divergence, while generally not symmetric in its arguments, is still locally symmetric.

Computing the expected Fisher information involves no observed data, it is purely a property of the model family $F_{\boldsymbol{\theta}}$. In Chapter 9 we will study a related quantity, the *observed Fisher information* that in contrast to the expected Fisher information is a function of the observed data.

> **i** Note 5: ▶ Fisher information as metric tensor
>
> In the field of *information geometry*[1] sets of distributions are studied using tools from differential geometry. It turns out that distribution families are manifolds and that the expected Fisher information matrix plays the role of the (symmetric!) metric tensor on this manifold.

Additivity of Fisher information

We may wish to compute the expected Fisher information based on a set of independent identically distributed (iid) random variables.

Assume that a random variable $x \sim F_{\boldsymbol{\theta}}$ has log-density $\log f(x|\boldsymbol{\theta})$ and expected Fisher information $\boldsymbol{I}^{\mathrm{Fisher}}(\boldsymbol{\theta})$. The expected Fisher information $\boldsymbol{I}^{\mathrm{Fisher}}_{x_1,\ldots,x_n}(\boldsymbol{\theta})$ for a set of iid random variables $x_1, \ldots, x_n \sim F_{\boldsymbol{\theta}}$ is computed from the joint log-density $\log f(x_1, \ldots, x_n) = \sum_i^n \log f(x_i|\boldsymbol{\theta})$. This

[1] A recent review is given, e.g., in: Nielsen, F. 2020. *An elementary introduction to information geometry.* Entropy **22**:1100. https://doi.org/10.3390/e22101100

yields

$$\begin{aligned} \boldsymbol{I}^{\text{Fisher}}_{x_1,\dots,x_n}(\boldsymbol{\theta}) &= -\mathrm{E}_{F_{\boldsymbol{\theta}}} \nabla\nabla^T \sum_i^n \log f(x_i|\boldsymbol{\theta}) \\ &= \sum_{i=1}^n \boldsymbol{I}^{\text{Fisher}}(\boldsymbol{\theta}) = n\boldsymbol{I}^{\text{Fisher}}(\boldsymbol{\theta}) \end{aligned}$$

Hence, the expected Fisher information for a set of n iid random variables is the n times the Fisher information of a single variable.

Invariance property of the Fisher information

Like KL divergence the **expected Fisher information is invariant against change of parametrisation of the sample space**, say from variable x to y and from distribution F_x to F_y. This is easy to see as the KL divergence itself is invariant against such reparametrisation, and thus also its curvature, and hence the expected Fisher information.

More specifically, when the sample space is changed the density will gain a factor in the form of the Jacobian determinant according to this transformation. However, since this factor does not depend on the model parameters, the first and second derivatives of the log-density with regard to the model parameters are not affected by it.

See also Section 7.4 for related sample space invariance of the gradient and curvature of the log-likelihood and Chapter 9 for the sample invariance of observed Fisher information.

▶ Transformation of Fisher information when model parameters change

The Fisher information $\boldsymbol{I}^{\text{Fisher}}(\boldsymbol{\theta})$ depends on the parameter $\boldsymbol{\theta}$. If we use a different parametrisation of the underlying parametric distribution family, say $\boldsymbol{\zeta}$ with a map $\boldsymbol{\theta}(\boldsymbol{\zeta})$ from $\boldsymbol{\zeta}$ to $\boldsymbol{\theta}$, then the Fisher information changes according to the chain rule in calculus.

To find the resulting Fisher information in terms of the new parameter $\boldsymbol{\zeta}$ we need to use the Jacobian matrix $D\boldsymbol{\theta}(\boldsymbol{\zeta})$. This matrix contains the

gradients for each component of the map $\boldsymbol{\theta}(\boldsymbol{\zeta})$ in its rows:

$$D\boldsymbol{\theta}(\boldsymbol{\zeta}) = \begin{pmatrix} \nabla^T \theta_1(\boldsymbol{\zeta}) \\ \nabla^T \theta_2(\boldsymbol{\zeta}) \\ \vdots \end{pmatrix}$$

With the above the Fisher information for $\boldsymbol{\theta}$ is then transformed to the Fisher information for $\boldsymbol{\zeta}$ applying the chain rule for the Hessian matrix:

$$\boldsymbol{I}^{\text{Fisher}}(\boldsymbol{\zeta}) = (D\boldsymbol{\theta}(\boldsymbol{\zeta}))^T \, \boldsymbol{I}^{\text{Fisher}}(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}(\boldsymbol{\zeta})} \, D\boldsymbol{\theta}(\boldsymbol{\zeta})$$

This type of transformation is also known as *covariant transformation*, in this case for the Fisher information metric tensor.

5.2 Expected Fisher information examples

Models with a single parameter

Example 5.1. Expected Fisher information for the Bernoulli distribution:

The log-probability mass function of the Bernoulli $\mathrm{Ber}(\theta)$ distribution is

$$\log p(x|\theta) = x \log(\theta) + (1-x)\log(1-\theta)$$

where θ is the probability of "success". The second derivative with regard to the parameter θ is

$$\frac{d^2}{d\theta^2} \log p(x|\theta) = -\frac{x}{\theta^2} - \frac{1-x}{(1-\theta)^2}$$

Since $\mathrm{E}(x) = \theta$ we get as Fisher information

$$\begin{aligned} I^{\text{Fisher}}(\theta) &= -\mathrm{E}\left(\frac{d^2}{d\theta^2} \log p(x|\theta)\right) \\ &= \frac{\theta}{\theta^2} + \frac{1-\theta}{(1-\theta)^2} \\ &= \frac{1}{\theta(1-\theta)} \end{aligned}$$

Example 5.2. Quadratic approximations of the KL divergence between two Bernoulli distributions:

From Example 4.4 we have as KL divergence

$$D_{\mathrm{KL}}\left(\mathrm{Ber}(\theta_1), \mathrm{Ber}(\theta_2)\right) = \theta_1 \log\left(\frac{\theta_1}{\theta_2}\right) + (1-\theta_1)\log\left(\frac{1-\theta_1}{1-\theta_2}\right)$$

and from Example 5.1 the corresponding expected Fisher information.

The quadratic approximation implies that

$$D_{\mathrm{KL}}\left(\mathrm{Ber}(\theta), \mathrm{Ber}(\theta+\varepsilon)\right) \approx \frac{\varepsilon^2}{2} I^{\mathrm{Fisher}}(\theta) = \frac{\varepsilon^2}{2\theta(1-\theta)}$$

and also that

$$D_{\mathrm{KL}}\left(\mathrm{Ber}(\theta+\varepsilon), \mathrm{Ber}(\theta)\right) \approx \frac{\varepsilon^2}{2} I^{\mathrm{Fisher}}(\theta) = \frac{\varepsilon^2}{2\theta(1-\theta)}$$

In Worksheet E1 this is verified by using a second order Taylor series applied to the KL divergence.

Example 5.3. Expected Fisher information for the normal distribution $N(\mu, \sigma^2)$ with known variance.

The log-density is

$$\log f(x|\mu,\sigma^2) = -\frac{1}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(x-\mu)^2 - \frac{1}{2}\log(2\pi)$$

The second derivative with respect to μ is

$$\frac{d^2}{d\mu^2}\log f(x|\mu,\sigma^2) = -\frac{1}{\sigma^2}$$

Therefore the expected Fisher information is

$$\boldsymbol{I}^{\mathrm{Fisher}}\left(\mu\right) = \frac{1}{\sigma^2}$$

Models with multiple parameters

Example 5.4. Expected Fisher information for the normal distribution $N(\mu, \sigma^2)$.

The log-density is

$$\log f(x|\mu,\sigma^2) = -\frac{1}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(x-\mu)^2 - \frac{1}{2}\log(2\pi)$$

The gradient with respect to μ and σ^2 (!) is the vector

$$\nabla \log f(x|\mu,\sigma^2) = \begin{pmatrix} \frac{1}{\sigma^2}(x-\mu) \\ -\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4}(x-\mu)^2 \end{pmatrix}$$

Hint for calculating the gradient: replace σ^2 by v and then take the partial derivative with regard to v, then substitute back.

The corresponding Hessian matrix is

$$\nabla\nabla^T \log f(x|\mu,\sigma^2) = \begin{pmatrix} -\frac{1}{\sigma^2} & -\frac{1}{\sigma^4}(x-\mu) \\ -\frac{1}{\sigma^4}(x-\mu) & \frac{1}{2\sigma^4} - \frac{1}{\sigma^6}(x-\mu)^2 \end{pmatrix}$$

As $\mathrm{E}(x) = \mu$ we have $\mathrm{E}(x-\mu) = 0$. Furthermore, with $\mathrm{E}((x-\mu)^2) = \sigma^2$ we see that $\mathrm{E}\left(\frac{1}{\sigma^6}(x-\mu)^2\right) = \frac{1}{\sigma^4}$. Therefore the expected Fisher information matrix as the negative expected Hessian matrix is

$$\boldsymbol{I}^{\mathrm{Fisher}}\left(\mu,\sigma^2\right) = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix}$$

Example 5.5. ▶ Expected Fisher information of the categorical distribution:

The log-probability mass function for the categorical distribution with K classes and $K-1$ free parameters $\pi_1, \ldots, \pi_{K-1}$ is

$$\begin{aligned}\log p(\boldsymbol{x}|\pi_1,\ldots,\pi_{K-1}) &= \sum_{k=1}^{K-1} x_k \log \pi_k + x_K \log \pi_K \\ &= \sum_{k=1}^{K-1} x_k \log \pi_k + \left(1 - \sum_{k=1}^{K-1} x_k\right) \log\left(1 - \sum_{k=1}^{K-1} \pi_k\right)\end{aligned}$$

From the log-probability mass function we compute the Hessian matrix of second order partial derivatives $\nabla\nabla^T \log p(\boldsymbol{x}|\pi_1,\ldots,\pi_{K-1})$ with regard to $\pi_1, \ldots, \pi_{K-1}$:

- The diagonal entries of the Hessian matrix (with $i = 1, \ldots, K-1$) are
$$\frac{\partial^2}{\partial \pi_i^2} \log p(\boldsymbol{x}|\pi_1,\ldots,\pi_{K-1}) = -\frac{x_i}{\pi_i^2} - \frac{x_K}{\pi_K^2}$$
- the off-diagonal entries are (with $j = 1, \ldots, K-1$ and $j \neq i$)
$$\frac{\partial^2}{\partial \pi_i \partial \pi_j} \log p(\boldsymbol{x}|\pi_1,\ldots,\pi_{K-1}) = -\frac{x_K}{\pi_K^2}$$

Recalling that $\mathrm{E}(x_i) = \pi_i$ we obtain the expected Fisher information matrix for a categorical distribution as $K-1 \times K-1$ dimensional matrix

$$\begin{aligned}\boldsymbol{I}^{\text{Fisher}}(\pi_1, \dots, \pi_{K-1}) &= -\mathrm{E}\left(\nabla\nabla^T \log p(\boldsymbol{x}|\pi_1, \dots, \pi_{K-1})\right) \\ &= \begin{pmatrix} \frac{1}{\pi_1} + \frac{1}{\pi_K} & \cdots & \frac{1}{\pi_K} \\ \vdots & \ddots & \vdots \\ \frac{1}{\pi_K} & \cdots & \frac{1}{\pi_{K-1}} + \frac{1}{\pi_K} \end{pmatrix} \\ &= \mathrm{Diag}\left(\frac{1}{\pi_1}, \dots, \frac{1}{\pi_{K-1}}\right) + \frac{1}{\pi_K}\mathbf{1}\end{aligned}$$

For $K = 2$ and $\pi_1 = \theta$ this reduces to the expected Fisher information of a Bernoulli variable, see Example 5.1.

$$\begin{aligned}I^{\text{Fisher}}(\theta) &= \left(\frac{1}{\theta} + \frac{1}{1-\theta}\right) \\ &= \frac{1}{\theta(1-\theta)}\end{aligned}$$

Example 5.6. ▶ Quadratic approximation of KL divergence of the categorical distribution and the Neyman and Pearson divergence:

We now consider the local approximation of the KL divergence $D_{\mathrm{KL}}(Q, P)$ between the categorical distribution $Q = \mathrm{Cat}(\boldsymbol{q})$ with probabilities $\boldsymbol{q} = (q_1, \dots, q_K)^T$ with the categorical distribution $P = \mathrm{Cat}(\boldsymbol{p})$ with probabilities $\boldsymbol{p} = (p_1, \dots, p_K)^T$.

From Example 4.7 we already know the KL divergence and from Example 5.5 the corresponding expected Fisher information.

First, we keep the first argument Q fixed and assume that P is a perturbed version of Q with $\boldsymbol{p} = \boldsymbol{q} + \boldsymbol{\varepsilon}$. Note that the perturbations $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_K)^T$ satisfy $\sum_{k=1}^K \varepsilon_k = 0$ because $\sum_{k=1}^K q_i = 1$ and $\sum_{k=1}^K p_i = 1$. Thus $\varepsilon_K = -\sum_{k=1}^{K-1} \varepsilon_k$. Then

$$\begin{aligned}D_{\mathrm{KL}}(\mathrm{Cat}(\boldsymbol{q}), \mathrm{Cat}(\boldsymbol{q}+\boldsymbol{\varepsilon})) &\approx \frac{1}{2}(\varepsilon_1, \dots, \varepsilon_{K-1})\, \boldsymbol{I}^{\text{Fisher}}(q_1, \dots, q_{K-1}) \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_{K-1} \end{pmatrix} \\ &= \frac{1}{2}\left(\sum_{k=1}^{K-1} \frac{\varepsilon_k^2}{q_k} + \frac{\left(\sum_{k=1}^{K-1} \varepsilon_k\right)^2}{q_K}\right) \\ &= \frac{1}{2}\sum_{k=1}^{K} \frac{\varepsilon_k^2}{q_k} \\ &= \frac{1}{2}\sum_{k=1}^{K} \frac{(q_k - p_k)^2}{q_k} \\ &= \frac{1}{2} D_{\text{Neyman}}(Q, P)\end{aligned}$$

Similarly, if we keep P fixed and consider Q as a perturbed version of P we get

$$\begin{aligned}D_{\mathrm{KL}}(\mathrm{Cat}(\boldsymbol{p}+\boldsymbol{\varepsilon}), \mathrm{Cat}(\boldsymbol{p})) &\approx \frac{1}{2}\sum_{k=1}^{K} \frac{(q_k - p_k)^2}{p_k} \\ &= \frac{1}{2} D_{\text{Pearson}}(Q, P)\end{aligned}$$

Note that in both approximations we divide by the probabilities of the distribution that is kept fixed.

Note the appearance of the *Pearson* χ^2 *divergence* and the *Neyman* χ^2 *divergence* in the above. Both are, like the KL divergence, part of the family of f-divergences. The Neyman χ^2 divergence is also known as the reverse Pearson divergence as $D_{\text{Neyman}}(Q, P) = D_{\text{Pearson}}(P, Q)$.

6 Principle of maximum entropy

6.1 ▶ Maximum entropy principle to characterise distributions

Both Shannon entropy and differential entropy are useful to characterise distributions.

As discussed in Chapter 3, **large entropy** implies that the **distribution is spread out** whereas **small entropy** indicates that the **distribution is concentrated**.

Correspondingly, **maximum entropy distributions** can be considered **minimally informative** about a random variable. Higher entropy entropy means a more spread-out and therefore less informative distribution. Conversely, low entropy indicates that probability mass is concentrated, making the distribution more informative about the random variable.

Examples:

1) The **discrete uniform distribution** is the **maximum entropy distribution** among all discrete distributions.
2) the maximum entropy distribution among all continuous distributions supported in $[0, \infty]$ with a specified mean is the **exponential distribution**.
3) the maximum entropy distribution of a continuous random variable with support $[-\infty, \infty]$ with a specific mean and variance is the **normal distribution**.

Using maximum entropy to characterise maximally uninformative distributions was advocated by Edwin T. Jaynes (1922–1998) (who also proposed to use maximum entropy in the context of finding Bayesian priors). The maximum entropy principle in statistical physics goes back to Boltzmann.

A list of maximum entropy distribution is given here: https://en.wikipedia.org/wiki/Maximum_entropy_probability_distribution.

Many distributions commonly used in statistical modelling are exponential families. Intriguingly, these distribution are all maximum entropy distributions, so there is a very close link between the principle of maximum entropy and common model choices in statistics and machine learning.

Example 6.1. ▶ Discrete uniform distribution as maximum entropy distribution:

Assume G is an categorical distribution with K classes and probabilities g_i. We now show that G has maximum entropy when G is the discrete uniform distribution.

Let $P = U_K$ be the discrete uniform with equal probabilities $p_i = 1/K$. The entropy of P is $H(P) = \log K$ (see also Example 3.8). The cross-entropy $H(G, P) = -\mathrm{E}_G \log p_i = \log K$. Note that both entropies are identical.

From Gibbs' inequality we know that $H(G, P) \geq H(G)$. Since in our case $H(G, P) = H(P)$ it follows directly that $H(P) \geq H(G)$, i.e. the discrete uniform distribution P achieves the maximum entropy. Furthermore, any other distribution G will have lower entropy, with equality only if $G = P$ and thus only if $g_i = 1/K$.

Example 6.2. ▶ Exponential distribution as maximum entropy distribution:

Assume G is an continuous distribution for x with support $[0, \infty]$ and with specified mean $\mathrm{E}(x) = \theta$. We now show that G has maximum entropy when G is the exponential distribution.

The log-density of the exponential distribution P with scale parameter θ is $\log p(x|\theta) = x/\theta - \log \theta$. The differential entropy of P is $H(P) = -\mathrm{E}_P \log p(x|\theta) = 1 + \log \theta$ as $\mathrm{E}_P(x) = \theta$. The cross-entropy $H(G, P) = -\mathrm{E}_G \log p(x|\theta) = 1 + \log \theta$ as $\mathrm{E}_G(x) = \theta$. Note that both entropies are identical.

From Gibbs' inequality we know that $H(G, P) \geq H(G)$. Since in our case $H(G, P) = H(P)$ it follows directly that $H(P) \geq H(G)$, i.e. the exponential distribution P achieves the maximum entropy. Furthermore, any other distribution G will have lower entropy, with equality only if $G = P$.

7 Principle of maximum likelihood

7.1 Overview

Outline of maximum likelihood estimation

Maximum likelihood is a very general method for fitting probabilistic models to data, generalising the earlier method of least-squares. It plays a very important role in statistics and was advocated and pioneered by R.A. Fisher in the early 20th century. [1]

In a nutshell, the starting points in a maximum likelihood analysis are

i) the observed data $D = \{x_1, \ldots, x_n\}$ with n independent and identically distributed (iid) samples, with the ordering irrelevant, and a
ii) a model $P_{\boldsymbol{\theta}}$ with corresponding probability density or probability mass function $p(x|\boldsymbol{\theta})$ and parameters $\boldsymbol{\theta}$

From model and data the likelihood function (note upper case "L") is constructed as

$$L_n(\boldsymbol{\theta}) = L(\boldsymbol{\theta}|D) = \prod_{i=1}^{n} p(x_i|\boldsymbol{\theta})$$

Equivalently, the log-likelihood function (note lower case "ℓ") is

$$\ell_n(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}|D) = \sum_{i=1}^{n} \log p(x_i|\boldsymbol{\theta})$$

In the above notation, we either explitly mention the data D or use an index to remind us of the sample size (here n).

The likelihood is multiplicative and the log-likelihood additive over the samples x_i because of the iid assumption.

[1] Aldrich J. 1997. *R. A. Fisher and the Making of Maximum Likelihood 1912–1922.* Statist. Sci. **12**:162–176. https://doi.org/10.1214/ss/1030037906

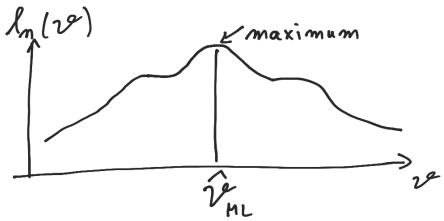


Figure 7.1: Finding the maximum likelihood estimate by maximisation of the (log)-likelihood function.

The maximum likelihood estimate (MLE) $\hat{\boldsymbol{\theta}}^{ML}$ is then found by maximising the (log)-likelihood function with regard to the parameters $\boldsymbol{\theta}$ (see Figure 7.1):

$$\hat{\boldsymbol{\theta}}_{ML} = \arg\max_{\boldsymbol{\theta}} \ell_n(\boldsymbol{\theta})$$

Hence, once the model is chosen and data are collected, finding the MLE and thus fitting the model to the data is an *optimisation problem*.

Depending on the complexity of the likelihood function and the number of the parameters finding the maximum likelihood can be very difficult. On the other hand, for likelihood functions constructed from common distribution families, such as exponential families, maximum likelihood estimation is very straightforward and can even be done analytically (this is the case for most examples we encounter in this course).

In practice in application to more complex models the optimisation required for maximum likelihood analysis is done on the computer, typically on the log-likelihood rather than on the likelihood function in order to avoid problems with the computer representation of small floating point numbers. Suitable optimisation algorithm may rely only on function values without requiring derivatives, or use in addition gradient and possibly curvature information. In recent years there has been a lot of progress in high-dimensional optimisation using combined numerical and analytical approaches (e.g. using automatic differentiation) and stochastic approximations (e.g. stochastic gradient descent).

Origin of the method of maximum likelihood

Historically, the likelihood has often interpreted and justified as the probability of the data given the model. However, while providing

an intuitive understanding this is not strictly correct. First, this interpretation only applies to discrete random variables. Second, since the samples $x_1, \ldots, x_n$ are typically **exchangeable** (i.e. permutation invariant) even in this case one would still need to add a factor accounting for the multiplicity of the possible orderings of the samples to obtain the correct probability of the data. Third, the interpretation of likelihood as probability of the data completely breaks down for continuous random variables because then $p(x|\boldsymbol{\theta})$ is a density, not a probability.

Next, we will see that maximum likelihood estimation is a well-justified method that arises naturally from an entropy perspective. More specifically, the maximum likelihood estimate corresponds to the distribution $P_{\boldsymbol{\theta}}$ that is closest in terms of KL divergence to the unknown true data-generating model as represented by the observed data and the empirical distribution.

7.2 From minimum KL divergence (or maximum Boltzmann relative entropy) to maximum likelihood

The KL divergence between true model and approximating model

Assume we have observations $D = \{x_1, \ldots, x_n\}$. The data are sampled from F, the true but unknown data-generating distribution. We also specify a family of distributions $P_{\boldsymbol{\theta}}$ indexed by $\boldsymbol{\theta}$ to approximate F.

The KL divergence $D_{\mathrm{KL}}(F, P_{\boldsymbol{\theta}})$ measures the divergence of the approximation $P_{\boldsymbol{\theta}}$ from the unknown true model F. It can be written as

$$\begin{aligned} D_{\mathrm{KL}}(F, P_{\boldsymbol{\theta}}) &= H(F, P_{\boldsymbol{\theta}}) - H(F) \\ &= \underbrace{-\mathrm{E}_F \log p_{\boldsymbol{\theta}}(x)}_{\text{cross-entropy}} - \Big(\underbrace{-\mathrm{E}_F \log f(x)}_{\text{entropy of } F, \text{ does not depend on } \boldsymbol{\theta}} \Big) \end{aligned}$$

However, since we do not know F we cannot actually compute this divergence. Nonetheless, we may use the empirical distribution $\hat{F}_n$ — a function of the observed data — as approximation for F, and in this way we arrive at an approximation for $D_{\mathrm{KL}}(F, P_{\boldsymbol{\theta}})$ that becomes more and more accurate with growing sample size.

Recall the "Law of Large Numbers" :

- The empirical distribution $\hat{F}_n$ based on observed data $D = \{x_1, \ldots, x_n\}$ converges strongly (almost surely) to the true underlying distribution F as $n \to \infty$:

$$\hat{F}_n \xrightarrow{a.s.} F$$

- Correspondingly, for $n \to \infty$ the average $\mathrm{E}_{\hat{F}_n}(h(x)) = \frac{1}{n}\sum_{i=1}^{n} h(x_i)$ converges to the expectation $\mathrm{E}_F(h(x))$.

Hence, for large sample size n we can approximate cross-entropy and as a result the KL divergence. The cross-entropy $H(F, P_{\boldsymbol{\theta}})$ is approximated by the **empirical cross-entropy** where the expectation is taken with regard to $\hat{F}_n$ rather than F:

$$\begin{aligned} H(F, P_{\boldsymbol{\theta}}) &\approx H(\hat{F}_n, P_{\boldsymbol{\theta}}) \\ &= -\mathrm{E}_{\hat{F}_n}(\log p(x|\boldsymbol{\theta})) \\ &= -\frac{1}{n}\sum_{i=1}^{n} \log p(x_i|\boldsymbol{\theta}) \\ &= -\frac{1}{n}\ell_n(\boldsymbol{\theta}) \end{aligned}$$

The empirical cross-entropy is equal to the negative log-likelihood standardised by the sample size n. Conversely, the **log-likelihood** is the **negative empirical cross-entropy multiplied by sample size** n.

Minimum KL divergence and maximum likelihood

If we knew F we would simply minimise $D_{\mathrm{KL}}(F, P_{\boldsymbol{\theta}})$ to find the particular model $P_{\boldsymbol{\theta}}$ that is closest to the true model, or equivalent, we would minimise the cross-entropy $H(F, P_{\boldsymbol{\theta}})$. However, since we actually don't know F this is not possible.

However, for large sample size n when the empirical distribution $\hat{F}_n$ is a good approximation for F, we can use the results from the previous

section. Thus, instead of minimising the KL divergence $D_{\mathrm{KL}}(F, P_{\boldsymbol{\theta}})$ we simply minimise $H(\hat{F}_n, P_{\boldsymbol{\theta}})$ which is the same as maximising the log-likelihood $\ell_n(\boldsymbol{\theta})$.

Conversely, this implies that maximising the likelihood with regard to the $\boldsymbol{\theta}$ is equivalent (asymptotically for large n!) to minimising the KL divergence of the approximating model and the unknown true model.

$$\begin{aligned}\hat{\boldsymbol{\theta}}_{ML} &= \arg\max_{\boldsymbol{\theta}} \ell_n(\boldsymbol{\theta}) \\ &= \arg\min_{\boldsymbol{\theta}} H(\hat{F}_n, P_{\boldsymbol{\theta}}) \\ &\approx \arg\min_{\boldsymbol{\theta}} D_{\mathrm{KL}}(F, P_{\boldsymbol{\theta}})\end{aligned}$$

Therefore, the reasoning behind the method of **maximum likelihood** is that it minimises a **large sample approximation of the KL divergence** of the candidate model $P_{\boldsymbol{\theta}}$ from the unkown true model F. In other words, **maximum likelihood estimators are minimum empirical KL divergence estimators**.

As the KL divergence is a functional of the true distribution F **maximum likelihood provides empirical estimators for parametric models**.

As a consequence of the close link of maximum likelihood and KL divergence maximum likelihood inherits for large n (and only then!) all the optimality properties from KL divergence.

7.3 Properties of maximum likelihood estimation

Consistency of maximum likelihood estimates

One important property of the method of maximum likelihood is that in general it produces **consistent estimates**. This means that estimates are well behaved so that they become more accurate with more data and in the limit of infinite data converge to the true parameters.

Specifically, if the true underlying model F is contained in the set of specified candidates models $P_{\boldsymbol{\theta}}$

$$\underbrace{F}_{\text{true model}} \subset \underbrace{P_{\boldsymbol{\theta}}}_{\text{specified models}}$$

there is a parameter $\boldsymbol{\theta}_{\text{true}}$ for which $F = P_{\boldsymbol{\theta}_{\text{true}}}$.

Maximisation of the likelihood function is, for large n, equivalent to minimising the KL divergence, and $D_{\mathrm{KL}}(F, P_{\boldsymbol{\theta}}) \to 0$ implies $P_{\boldsymbol{\theta}} \to F$, hence

$$\hat{\boldsymbol{\theta}}_{ML} \overset{\text{large } n}{\longrightarrow} \boldsymbol{\theta}_{\text{true}}$$

Thus, given sufficient data, the maximum likelihood estimate of the model parameters will converge to the true parameter values. Note that the above assumes the model family, and therefore the number of parameters, is fixed.

As a consequence of consistency, **maximum likelihood estimates are asympotically unbiased**. As we will see in the examples, they can still be biased in finite samples.

Note that even if the candidate model family $P_{\boldsymbol{\theta}}$ is **misspecified** (so that it does not contain the actual true model), the maximum likelihood estimate is still optimal, for large sample size, in the sense in that it will identify the distribution in the family that is closest in terms of KL divergence.

Finally, inconsistent maximum likelihood estimates can occur, but only in non-regular situations, for example when the MLE lies on a boundary or when there are singularities in the likelihood function. Furthermore, inconsistency can arise when the number of parameters grows with the sample size and hence the information per parameter does not sufficiently increase (as in the well-known Neyman–Scott paradox).

Invariance property of the maximum likelihood

The maximum likelihood invariance principle states that the **achieved maximum likelihood is invariant against reparametrisation of the model parameters**. This property is shared by the KL divergence minimisation procedure as the achieved minimum KL divergence is also invariant against the change of parameters.

Recall that the model parameter is just an arbitrary label to index a specific distribution within a distribution family, and changing that

label does not affect the maximum (likelihood) or the minimum (KL divergence). For example, consider a function $h_x(x)$ with a maximum at $x_{\max} = \arg\max h_x(x)$. Now we relabel the argument using $y = g(x)$ where g is an invertible function. Then the function in terms of y is $h_y(y) = h_x(g^{-1}(y))$. and clearly this function has a maximum at $y_{\max} = g(x_{\max})$ since $h_y(y_{\max}) = h_x(g^{-1}(y_{\max})) = h_x(x_{\max})$. Furthermore, the achieved maximum value is the same.

In application to maximum likelihood, assume we transform a parameter θ into another parameter ω using some invertible function $g()$ so that $\omega = g(\theta)$. Then the maximum likelihood estimate $\hat{\omega}_{ML}$ of the new parameter ω is simply the transformation of the maximum likelihood estimate $\hat{\theta}_{ML}$ of the original parameter θ with $\hat{\omega}_{ML} = g(\hat{\theta}_{ML})$. The achieved maximum likelihood is the same in both cases.

The invariance property can be very useful in practice because it is often easier (and sometimes numerically more stable) to maximise the likelihood for a different set of parameters.

See Worksheet L1 for an example application of the invariance principle.

Sufficient statistics

Another important concept are so-called sufficient statistics to summarise the information available in the data about a parameter in a model.

A statistic $\boldsymbol{t}(D)$ is called a **sufficient statistic** for the model parameters $\boldsymbol{\theta}$ if the corresponding likelihood function can be written using only $\boldsymbol{t}(D)$ in the terms that involve $\boldsymbol{\theta}$ such that

$$L(\boldsymbol{\theta}|D) = h(\boldsymbol{t}(D), \boldsymbol{\theta})\, k(D)\,,$$

where $h()$ and $k()$ are positive-valued functions. This is known as the **Fisher-Pearson factorisation**. Equivalently on log-scale this becomes

$$\ell(\boldsymbol{\theta}|D) = \log h(\boldsymbol{t}(D), \boldsymbol{\theta}) + \log k(D)\,.$$

By construction, estimation and inference about $\boldsymbol{\theta}$ based on the factorised likelihood $L(\boldsymbol{\theta})$ is mediated through the sufficient statistic $\boldsymbol{t}(D)$ and does not require knowledge of the original data D. Instead, the sufficient statistic $\boldsymbol{t}(D)$ contains all the information in D required to learn about the parameter $\boldsymbol{\theta}$.

Note that **a sufficient statistic always exists** since the data D are themselves sufficient statistics, with $\boldsymbol{t}(D) = D$. However, in practice one aims to find sufficient statistics that summarise the data D and hence provide data reduction. This will become clear in the examples below.

Furthermore, sufficient statistics are **not unique** since applying a one-to-one transformation to $\boldsymbol{t}(D)$ yields another sufficient statistic.

Therefore, if the MLE $\hat{\boldsymbol{\theta}}_{ML}$ of $\boldsymbol{\theta}$ exists and is unique then **the MLE is a unique function of the sufficient statistic** $\boldsymbol{t}(D)$. If the MLE is not unique then it can be chosen to be function of $\boldsymbol{t}(D)$.

7.4 Maximum likelihood estimation for regular models

Regular models

A regular model is one that is well-behaved and well-suited for model fitting by optimisation. In particular this requires that:

- the support does not depend on the parameters,
- the model is identifiable (in particular the model is not over-parametrised and has a minimal set of parameters),
- the density / probability mass function and hence the log-likelihood function is twice differentiable everywhere with regard to the parameters,
- the maximum (peak) of the likelihood function lies inside the parameter space and not at a boundary,
- the second derivative of the log-likelihood at the maximum is negative and not zero (for multiple parameters: the Hessian matrix at the maximum is negative definite and not singular)

Most models considered in this course are regular.

Maximum likelihood estimation in regular models

For a regular model maximum likelihood estimation and the necessary optimisation is greatly simplified by being able to using gradient and curvature information.

In order to maximise $\ell_n(\boldsymbol{\theta})$ one may use the **score function** $\boldsymbol{S}_n(\boldsymbol{\theta})$ which is the first derivative of the log-likelihood function with regard to the parameters [2].

$$S_n(\theta) = \frac{d\ell_n(\theta)}{d\theta} \quad \text{scalar parameter } \theta\text{: first derivative}$$

$$\boldsymbol{S}_n(\boldsymbol{\theta}) = \nabla \ell_n(\boldsymbol{\theta}) \quad \text{multivariate parameter } \boldsymbol{\theta}\text{: gradient}$$

In this case a necessary (but not sufficient) condition for the MLE is that

$$\boldsymbol{S}_n(\hat{\boldsymbol{\theta}}_{ML}) = 0$$

To demonstrate that the log-likelihood function actually achieves a maximum at $\hat{\boldsymbol{\theta}}_{ML}$ the curvature at the MLE must negative, i.e. that the log-likelihood must be locally concave at the MLE.

In the case of a single parameter (scalar θ) this requires to check that the second derivative of the log-likelihood function with regard to the parameter is negative:

$$H_n(\theta) = \frac{d^2\ell_n(\theta)}{d\theta^2}$$

and

$$H_n(\hat{\theta}_{ML}) < 0$$

In the case of a parameter vector (multivariate $\boldsymbol{\theta}$) one first computes the Hessian matrix (matrix of second order derivatives) of the log-likelihood function

$$\boldsymbol{H}_n(\boldsymbol{\theta}) = \nabla\nabla^T \ell_n(\boldsymbol{\theta})$$

For a multivariate parameter vector $\boldsymbol{\theta}$ of dimension d the Hessian is a matrix of size $d \times d$. Then one needs to verify that the Hessian matrix is negative definite at the MLE:

$$\boldsymbol{H}_n(\hat{\boldsymbol{\theta}}_{ML}) < 0,$$

[2] The score function $\boldsymbol{S}_n(\boldsymbol{\theta})$ as the gradient of the log-likelihood function must not be confused with the scoring rule $S(x, P)$ mentioned in the introduction to entropy and KL divergence, cf. Note 2.

i.e. all its eigenvalues must be negative.

Invariance of score function and second derivative of the log-likelihood

The score function $\boldsymbol{S}_n(\boldsymbol{\theta})$ is **invariant against transformation of the sample space**. Assume $\boldsymbol{x}$ has log-density $\log f_x(\boldsymbol{x}|\boldsymbol{\theta})$ then the log-density for $\boldsymbol{y}$ is

$$\log f_y(\boldsymbol{y}|\boldsymbol{\theta}) = \log|\det\left(D\boldsymbol{x}(\boldsymbol{y})\right)| + \log f_x\left(\boldsymbol{x}(\boldsymbol{y})|\boldsymbol{\theta}\right)$$

where $D\boldsymbol{x}(\boldsymbol{y})$ is the Jacobian matrix of the inverse transformation $\boldsymbol{x}(\boldsymbol{y})$. When taking the derivative of the log-likelihood function with regard to the parameter $\boldsymbol{\theta}$ the first term containing the Jacobian determinant vanishes. Hence the score function $\boldsymbol{S}_n(\boldsymbol{\theta})$ is not affected by a change of variables.

As a consequence, the second derivative of log-likelihood function with regard to $\boldsymbol{\theta}$ is also invariant against transformations of the sample space.

8 Maximum likelihood estimation in practice

Next, maximum likelihood estimation is illustrated on a number of examples. Among others we discuss three basic problems, namely how to estimate a proportion, the mean and the variance in the likelihood framework. We also consider an example of non-regular model (Example 8.4).

8.1 Likelihood estimation for a single parameter

In the following we illustrate likelihood estimation for models with a single parameter. In this case the score function and the second derivative of the log-likelihood are all scalar-valued like the log-likelihood function itself.

Example 8.1. Maximum likelihood estimation for the Bernoulli model:

We aim to estimate the true proportion θ in a Bernoulli experiment with binary outcomes, say the proportion of "successes" vs. "failures" or of "heads" vs. "tails" in a coin tossing experiment.

- Bernoulli model $\text{Ber}(\theta)$: Pr("success") $= \theta$ and Pr("failure") $= 1-\theta$.
- The "success" is indicated by outcome $x = 1$ and the "failure" by $x = 0$.
- We conduct n trials and record n_1 successes and $n - n_1$ failures.
- Parameter: θ probability of "success".

What is the MLE of θ?

- the observations $D = \{x_1, \ldots, x_n\}$ take on values 0 or 1.
- the average of the data points is $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{n_1}{n}$.

- the probability mass function (PMF) of the Bernoulli distribution $\text{Ber}(\theta)$ is:

$$p(x|\theta) = \theta^x(1-\theta)^{1-x} = \begin{cases} \theta & \text{if } x = 1 \\ 1-\theta & \text{if } x = 0 \end{cases}$$

- log-PMF:

$$\log p(x|\theta) = x\log(\theta) + (1-x)\log(1-\theta)$$

- log-likelihood function:

$$\begin{aligned} \ell_n(\theta) &= \sum_{i=1}^{n} \log p(x_i|\theta) \\ &= n_1 \log\theta + (n-n_1)\log(1-\theta) \\ &= n\left(\bar{x}\log\theta + (1-\bar{x})\log(1-\theta)\right) \end{aligned}$$

  Note that the log-likelihood depends on the data only via $\bar{x}$. Thus, $t(D) = \bar{x}$ is a **sufficient statistic** for the parameter θ. In fact it is also a **minimally sufficient statistic** as will be discussed in more detail later.

- Score function:

$$S_n(\theta) = \frac{d\ell_n(\theta)}{d\theta} = n\left(\frac{\bar{x}}{\theta} - \frac{1-\bar{x}}{1-\theta}\right)$$

- Maximum likelihood estimate: Setting $S_n(\hat{\theta}_{ML}) = 0$ yields as solution

$$\hat{\theta}_{ML} = \bar{x} = \frac{n_1}{n}$$

  With $H_n(\theta) = \frac{dS_n(\theta)}{d\theta} = -n\left(\frac{\bar{x}}{\theta^2} + \frac{1-\bar{x}}{(1-\theta)^2}\right) < 0$ the optimum corresponds indeed to the maximum of the (log-)likelihood function as this is negative for $\hat{\theta}_{ML}$ (and indeed for any θ).

  The maximum likelihood estimator of θ is therefore identical to the frequency of the successes among all observations.

Note that to analyse the coin tossing experiment and to estimate θ we may equally well use the binomial distribution $\text{Bin}(n, \theta)$ as model for the number of successes. This results in the same MLE for θ but the likelihood function based on the binomial PMF includes the binomial coefficient. However, as it does not depend on θ it disappears in the score function and has no influence in the derivation of the MLE.

Example 8.2. Maximum likelihood estimation for the normal distribution with unknown mean and known variance:

- $x \sim N(\mu, \sigma^2)$ with $\mathrm{E}(x) = \mu$ and $\mathrm{Var}(x) = \sigma^2$
- the parameter to be estimated is μ whereas σ^2 is known.

What's the MLE of the parameter μ?

- the data $D = \{x_1, \ldots, x_n\}$ are all real in the range $x_i \in [-\infty, \infty]$.
- the average $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is real as well.
- Density:

$$p(x|\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- Log-Density:

$$\log p(x|\mu) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{(x-\mu)^2}{2\sigma^2}$$

- Log-likelihood function:

$$\begin{aligned}
\ell_n(\mu) &= \sum_{i=1}^{n} \log p(x_i|\mu) \\
&= -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 \qquad \underbrace{-\frac{n}{2}\log(2\pi\sigma^2)}_{\text{constant term, does not depend on } \mu, \text{ can be removed}} \\
&= -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i^2 - 2x_i\mu + \mu^2) + \text{ const.} \\
&= \frac{n}{\sigma^2}(\bar{x}\mu - \frac{1}{2}\mu^2) \quad \underbrace{-\frac{1}{2\sigma^2}\sum_{i=1}^{n} x_i^2}_{\text{another constant term}} + \text{const.}
\end{aligned}$$

  Note how the non-constant terms of the log-likelihood depend on the data only through $\bar{x}$. Hence $t(D) = \bar{x}$ this is a **sufficient statistic** for μ.

- Score function:

$$S_n(\mu) = \frac{n}{\sigma^2}(\bar{x} - \mu)$$

- Maximum likelihood estimate:

$$S_n(\hat{\mu}_{ML}) = 0 \Rightarrow \hat{\mu}_{ML} = \bar{x}$$

- With $H_n(\mu) = \frac{dS_n(\mu)}{d\mu} = -\frac{n}{\sigma^2} < 0$ the optimum is indeed the maximum

The constant term in the log-likelihood function collects all terms that do not depend on the parameter of interest. After taking the first derivative with regard to the parameter the constant term disappears thus it has no influence in maximum likelihood estimation. **Therefore constant terms can be dropped from the log-likelihood function.**

Example 8.3. Maximum likelihood estimation for the normal distribution with known mean and unknown variance:

- $x \sim N(\mu, \sigma^2)$ with $\mathrm{E}(x) = \mu$ and $\mathrm{Var}(x) = \sigma^2$
- σ^2 needs to be estimated whereas the mean μ is known

What's the MLE of σ^2?

- the data $D = \{x_1, \ldots, x_n\}$ are all real in the range $x_i \in [-\infty, \infty]$.
- the average of the squared centred data $\overline{(x-\mu)^2} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2 \geq 0$ is non-negative.
- Density:

$$p(x|\sigma^2) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- Log-Density:

$$\log p(x|\sigma^2) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{(x-\mu)^2}{2\sigma^2}$$

- Log-likelihood function:

$$\begin{aligned}
\ell_n(\sigma^2) = \ell_n(\mu, \sigma^2) &= \sum_{i=1}^{n} \log p(x_i|\sigma^2) \\
&= -\frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 \qquad \underbrace{-\frac{n}{2}\log(2\pi)}_{\text{constant not depending on } \sigma^2} \\
&= -\frac{n}{2}\log(\sigma^2) - \frac{n}{2\sigma^2}\overline{(x-\mu)^2} + C
\end{aligned}$$

Note how the log-likelihood function depends on the data only through $\overline{(x-\mu)^2}$. Hence $t(D) = \overline{(x-\mu)^2}$ is a sufficient statistic for σ^2.

- Score function:

$$S_n(\sigma^2) = -\frac{n}{2\sigma^2} + \frac{n}{2\sigma^4}\overline{(x-\mu)^2}$$

Note that to obtain the score function the derivative needs to be taken with regard to the variance parameter σ^2 — not with regard to σ! As a trick, relabel $\sigma^2 = v$ in the log-likelihood function, then take the derivative with regard to v, then backsubstitute $v = \sigma^2$ in the final result.

- Maximum likelihood estimate:

$$S_n(\widehat{\sigma^2}_{ML}) = 0 \Rightarrow$$

$$\widehat{\sigma^2}_{ML} = \overline{(x-\mu)^2} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2$$

- To confirm that we actually have maximum we need to verify that the second derivative of log-likelihood at the optimum is negative. With $H_n(\sigma^2) = \frac{dS_n(\sigma^2)}{d\sigma^2} = -\frac{n}{2\sigma^4}\left(\frac{2}{\sigma^2}\overline{(x-\mu)^2} - 1\right)$ and hence $H_n(\widehat{\sigma^2}_{ML}) = -\frac{n}{2}\left(\widehat{\sigma^2}_{ML}\right)^{-2} < 0$ the optimum is indeed the maximum.

Example 8.4. Uniform distribution with upper bound θ:

This is an example of a non-regular model, as the parameter θ determines the support of the model.

- $x \sim \text{Unif}(0, \theta)$ with $\theta > 0$
- the data $D = \{x_1, \ldots, x_n\}$ are all real in the range $x_i \in [0, \theta]$.
- by $x_{[i]}$ we denote the *ordered* observations with $0 \le x_{[1]} < x_{[2]} < \ldots < x_{[n]} \le \theta$ with $x_{[n]} = \max(x_1, \ldots, x_n)$.

We would like to obtain the maximum likelihood estimator $\hat{\theta}_{ML}$.

- The probability density function of $\text{Unif}(0, \theta)$ is

$$p(x|\theta) = \begin{cases} \frac{1}{\theta} & \text{if } x \in [0, \theta] \\ 0 & \text{otherwise.} \end{cases}$$

- the corresponding the log-density is

$$\log p(x|\theta) = \begin{cases} -\log\theta & \text{if } x \in [0, \theta] \\ -\infty & \text{otherwise.} \end{cases}$$

- the log-likelihood function is

$$\ell_n(\theta) = \begin{cases} -n\log\theta & \text{for } \theta \ge x_{[n]} \\ -\infty & \text{otherwise} \end{cases}$$

since all observed data $D = \{x_1, \ldots, x_n\}$ lie in the interval $[0, \theta]$. Note that the log-likelihood is a function of $x_{[n]}$ only so this single data point is the sufficient statistic $t(D) = x_{[n]}$.

- the log-likelihood function remains at value $-\infty$ until $\theta = x_{[n]}$, where it jumps to $-n\log x_{[n]}$ and then it decreases monotonically with increasing $\theta > x_{[n]}$. Hence the log-likelihood function has a maximum at $\hat{\theta}_{ML} = x_{[n]}$.
- Due to the discontinuity at $x_{[n]}$ the log-likelihood $\ell_n(\theta)$ **is not differentiable** at $\hat{\theta}_{ML}$. and hence the maximum cannot be found by setting the score function equal to zero as in a regular model.
- In addition, **there is no quadratic approximation around** $\hat{\theta}_{ML}$ and therefore the **observed Fisher information cannot be computed** either.

8.2 Likelihood estimation for multiple parameters

If there are several parameters likelihood estimation is conceptually no different from the case of a single parameter. However, the score function is now vector-valued and the second derivative of the log-likelihood is a matrix-valued function.

Example 8.5. Normal distribution with mean and variance both unknown:

- $x \sim N(\mu, \sigma^2)$ with $\text{E}(x) = \mu$ and $\text{Var}(x) = \sigma^2$
- both μ and σ^2 need to be estimated.

What's the MLE of the parameter vector $\boldsymbol{\theta} = (\mu, \sigma^2)^T$?

- the data $D = \{x_1, \ldots, x_n\}$ are all real in the range $x_i \in [-\infty, \infty]$.
- the average $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$ is real as well.
- the average of the squared data $\overline{x^2} = \frac{1}{n}\sum_{i=1}^n x_i^2 \geq 0$ is non-negative.
- Density:

$$f(x|\mu, \sigma^2) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- Log-Density:

$$\log f(x|\mu, \sigma^2) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{(x-\mu)^2}{2\sigma^2}$$

- Log-likelihood function:

$$\begin{aligned}\ell_n(\boldsymbol{\theta}) = \ell_n(\mu, \sigma^2) &= \sum_{i=1}^n \log f(x_i|\mu, \sigma^2) \\ &= -\frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu)^2 \qquad \underbrace{-\frac{n}{2}\log(2\pi)}_{\text{constant not depending on } \mu \text{ or } \sigma^2} \\ &= -\frac{n}{2}\log(\sigma^2) - \frac{n}{2\sigma^2}(\overline{x^2} - 2\bar{x}\mu + \mu^2) + C\end{aligned}$$

  Note how the log-likelihood function depends on the data only through $\bar{x}$ and $\overline{x^2}$. Hence, $\boldsymbol{t}(D) = (\bar{x}, \overline{x^2})^T$ are sufficient statistics for $\boldsymbol{\theta}$.

- Score function $\boldsymbol{s}_n$, gradient of $\ell_n(\boldsymbol{\theta})$:

$$\begin{aligned}\boldsymbol{s}_n(\boldsymbol{\theta}) &= \boldsymbol{s}_n(\mu, \sigma^2) \\ &= \nabla \ell_n(\mu, \sigma^2) \\ &= \begin{pmatrix} \frac{n}{\sigma^2}(\bar{x} - \mu) \\ -\frac{n}{2\sigma^2} + \frac{n}{2\sigma^4}\left(\overline{x^2} - 2\bar{x}\mu + \mu^2\right) \end{pmatrix}\end{aligned}$$

- Maximum likelihood estimate:

$$\boldsymbol{s}_n(\hat{\boldsymbol{\theta}}_{ML}) = 0 \Rightarrow$$

$$\hat{\boldsymbol{\theta}}_{ML} = \begin{pmatrix} \hat{\mu}_{ML} \\ \widehat{\sigma^2}_{ML} \end{pmatrix} = \begin{pmatrix} \bar{x} \\ \overline{x^2} - \bar{x}^2 \end{pmatrix}$$

  The ML estimate of the variance can also be written $\widehat{\sigma^2}_{ML} = \overline{x^2} - \bar{x}^2 = \overline{(x - \bar{x})^2} = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2$.

- To confirm that we actually have a maximum we need to verify that the eigenvalues of the Hessian matrix at the optimum are all negative. This is indeed the case, for details see Example 9.4.

Example 8.6. ▶ Maximum likelihood estimates of the parameters of the multivariate normal distribution:

The results from Example 8.5 can be generalised to the multivariate normal distribution:

- $\boldsymbol{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\mathrm{E}(\boldsymbol{x}) = \boldsymbol{\mu}$ and $\mathrm{Var}(\boldsymbol{x}) = \boldsymbol{\Sigma}$
- both $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ need to be estimated.

With

- the data $D = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ containing real vector-valued observations,

the maximum likelihood can be written as follows:

MLE for the mean:

$$\hat{\boldsymbol{\mu}}_{ML} = \frac{1}{n}\sum_{k=1}^n \boldsymbol{x}_k = \bar{\boldsymbol{x}}$$

MLE for the covariance:

$$\underbrace{\widehat{\boldsymbol{\Sigma}}_{ML}}_{d\times d} = \frac{1}{n}\sum_{k=1}^n \underbrace{(\boldsymbol{x}_k - \bar{\boldsymbol{x}})}_{d\times 1}\underbrace{(\boldsymbol{x}_k - \bar{\boldsymbol{x}})^T}_{1\times d}$$

Note the factor $\frac{1}{n}$ in the estimator of the covariance matrix.

With $\overline{\boldsymbol{x}\boldsymbol{x}^T} = \frac{1}{n}\sum_{k=1}^n \boldsymbol{x}_k\boldsymbol{x}_k^T$ we can also write

$$\widehat{\boldsymbol{\Sigma}}_{ML} = \overline{\boldsymbol{x}\boldsymbol{x}^T} - \bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^T$$

Hence, the MLEs correspond to the well-known empirical estimates.

The derivation of the MLEs is discussed in more detail in the module MATH38161 Multivariate Statistics and Machine Learning.

Example 8.7. ▶ Maximum likelihood estimation of the parameters of the categorical distribution:

Maximum likelihood estimation of the parameters of $\text{Cat}(\boldsymbol{\pi})$ at first seems a trivial extension of the Bernoulli model (cf. Example 8.1) but this a bit more complicated because of the constraint on the allowed values of $\boldsymbol{\pi}$ so there are only $K-1$ free parameters and not K. Hence we either need to optimise with regard to a specific set of $K-1$ parameters (which is what we do below) or use a constrained optimisation procedure to enforce that $\sum_{k=1}^{K} \pi_k = 1$ (e.g using Lagrange multipliers).

- The data: We observe n samples $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$. The data matrix of dimension $n \times K$ is $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T = (x_{ik})$. It contains each $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{iK})^T$. The corresponding summary (minimal sufficient) statistics are $\boldsymbol{t}(D) = \bar{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i = (\bar{x}_1, \ldots, \bar{x}_K)^T$ with $\bar{x}_k = \frac{1}{n}\sum_{i=1}^{n} x_{ik}$. We can also write $\bar{x}_K = 1 - \sum_{k=1}^{K-1} \bar{x}_k$. The number of samples for class k is $n_k = n\bar{x}_k$ with $\sum_{k=1}^{K} n_k = n$.
- The log-likelihood is

$$\begin{aligned}
\ell_n(\pi_1, \ldots, \pi_{K-1}) &= \sum_{i=1}^{n} \log f(\boldsymbol{x}_i) \\
&= \sum_{i=1}^{n} \left(\sum_{k=1}^{K-1} x_{ik} \log \pi_k + \left(1 - \sum_{k=1}^{K-1} x_{ik}\right) \log \left(1 - \sum_{k=1}^{K-1} \pi_k\right) \right) \\
&= n \left(\sum_{k=1}^{K-1} \bar{x}_k \log \pi_k + \left(1 - \sum_{k=1}^{K-1} \bar{x}_k\right) \log \left(1 - \sum_{k=1}^{K-1} \pi_k\right) \right) \\
&= n \left(\sum_{k=1}^{K-1} \bar{x}_k \log \pi_k + \bar{x}_K \log \pi_K \right)
\end{aligned}$$

- Score function (gradient)

$$\begin{aligned}
\boldsymbol{s}_n(\pi_1, \ldots, \pi_{K-1}) &= \nabla \ell_n(\pi_1, \ldots, \pi_{K-1}) \\
&= \begin{pmatrix} \frac{\partial}{\partial \pi_1} \ell_n(\pi_1, \ldots, \pi_{K-1}) \\ \vdots \\ \frac{\partial}{\partial \pi_{K-1}} \ell_n(\pi_1, \ldots, \pi_{K-1}) \end{pmatrix} \\
&= n \begin{pmatrix} \frac{\bar{x}_1}{\pi_1} - \frac{\bar{x}_K}{\pi_K} \\ \vdots \\ \frac{\bar{x}_{K-1}}{\pi_{K-1}} - \frac{\bar{x}_K}{\pi_K} \end{pmatrix}
\end{aligned}$$

Note in particular the need for the second term that arises because π_K depends on all the $\pi_1, \ldots, \pi_{K-1}$.

- Maximum likelihood estimate: Setting $\boldsymbol{s}_n(\hat{\pi}_1^{ML}, \ldots, \hat{\pi}_{K-1}^{ML}) = 0$ yields $K-1$ equations

$$\bar{x}_i \left(1 - \sum_{k=1}^{K-1} \hat{\pi}_k^{ML}\right) = \hat{\pi}_i^{ML} \left(1 - \sum_{k=1}^{K-1} \bar{x}_k\right)$$

for $i = 1, \ldots, K-1$ and with solution

$$\hat{\pi}_i^{ML} = \bar{x}_i$$

It also follows that

$$\hat{\pi}_K^{ML} = 1 - \sum_{k=1}^{K-1} \hat{\pi}_k^{ML} = 1 - \sum_{k=1}^{K-1} \bar{x}_k = \bar{x}_K$$

The maximum likelihood estimator is therefore the frequency of the occurrence of a class among the n samples.

- To confirm that we actually have a maximum we need to verify that the eigenvalues of the Hessian matrix at the optimum are all negative. This is indeed the case, for details see Example 9.5.

8.3 Further properties of ML

Relationship of maximum likelihood with least squares estimation

In Example 8.2 the form of the log-likelihood function is a function of the sum of squared differences. Maximising $\ell_n(\mu) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$ is equivalent to *minimising* $\sum_{i=1}^{n}(x_i - \mu)^2$. Hence, finding the mean by **maximum likelihood assuming a normal model** is **equivalent to least-squares estimation**!

Note that least-squares estimation has been in use at least since the early 1800s [1] and thus predates maximum likelihood (1922). Due to its simplicity it is still very popular in particular in regression and the link

[1] Stigler, S. M. 1981. *Gauss and the invention of least squares*. Ann. Statist. **9**:465–474. https://doi.org/10.1214/aos/1176345451

with maximum likelihood and normality allows to understand why it usually works well.

See also Example 3.3 and Example 4.5 for further links of the normal distribution with squared error.

Bias of maximum likelihood estimates

Example 8.5 is interesting because it shows that maximum likelihood can result in both biased and as well as unbiased estimators.

Recall that $x \sim N(\mu, \sigma^2)$. As a result

$$\hat{\mu}_{ML} = \bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

with $\mathrm{E}(\hat{\mu}_{ML}) = \mu$ and

$$\widehat{\sigma^2}_{\mathrm{ML}} \sim W_1\left(s^2 = \frac{\sigma^2}{n}, k = n-1\right)$$

(see Section A.8) with mean $\mathrm{E}(\widehat{\sigma^2}_{ML}) = \frac{n-1}{n}\sigma^2$.

Therefore, the MLE of μ is unbiased as

$$\mathrm{Bias}(\hat{\mu}_{ML}) = \mathrm{E}(\hat{\mu}_{ML}) - \mu = 0$$

In contrast, however, the MLE of σ^2 is negatively biased because

$$\mathrm{Bias}(\widehat{\sigma^2}_{ML}) = \mathrm{E}(\widehat{\sigma^2}_{ML}) - \sigma^2 = -\frac{1}{n}\sigma^2$$

Thus, in the case of the variance parameter of the normal distribution the MLE is *not* recovering the well-known unbiased estimator of the variance

$$\widehat{\sigma^2}_{UB} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{n}{n-1}\widehat{\sigma^2}_{ML}$$

In other words, the unbiased variance estimate is not a maximum likelihood estimate!

Therefore it is worth keeping in mind that maximum likelihood can result in biased estimates for finite n. For large n, however, the bias disappears as MLEs are consistent.

▶ Minimal sufficient statistics and maximal data reduction

In all the examples discussed above the sufficient statistic was typically either $\bar{x}$ and $\overline{x^2}$ (or both). This is not a coincidence since all of the examples are exponential families with canonical statistics x and x^2, and in exponential families a sufficient statistic can be obtained as the average of the canonical statistics.

Crucially, in the above examples the identified sufficient statistics are also **minimal sufficient statistics** where the dimension of sufficient statistic is equal to the dimension of the parameter vector, and as such as low as possible. Minimal sufficient statistics provide maximal data reduction as will be discussed later.

9 Observed Fisher information

9.1 Definition of the observed Fisher information

Visual inspection of the log-likelihood function (e.g. Figure 9.1) suggests that it contains more information about the parameter $\boldsymbol{\theta}$ than just the location of the maximum point at $\hat{\boldsymbol{\theta}}_{ML}$.

In particular, in a regular model the **curvature** of the log-likelihood function at the MLE seems to be related to the accuracy of $\hat{\boldsymbol{\theta}}_{ML}$: if the likelihood surface is flat near the maximum (low curvature) then if is more difficult to find the optimal parameter (also numerically). Conversely, if the likelihood surface is sharply peaked (strong curvature) then the maximum point is well defined.

The curvature can be quantified by the second-order derivatives (Hessian matrix) of the log-likelihood function.

Accordingly, the **observed Fisher information** (matrix) is defined as the negative Hessian of the log-likelihood function $\ell_n(\boldsymbol{\theta})$ at the MLE $\hat{\boldsymbol{\theta}}_{ML}$:

$$\boldsymbol{J}_n(\hat{\boldsymbol{\theta}}_{ML}) = -\boldsymbol{H}(\hat{\boldsymbol{\theta}}_{ML}) = -\nabla\nabla^T \ell_n(\hat{\boldsymbol{\theta}}_{ML})$$

Sometimes this is simply called the "observed information". To avoid confusion with the **expected Fisher information**

$$\boldsymbol{I}^{\text{Fisher}}(\boldsymbol{\theta}) = -\mathrm{E}_{F_\theta}\left(\nabla\nabla^T \log f(x|\boldsymbol{\theta})\right)$$

introduced earlier it is necessary to always use the qualifier "observed" when referring to $\boldsymbol{J}_n(\hat{\boldsymbol{\theta}}_{ML})$.

We will see in more detail later that the observed Fisher information plays an important role at quantifying the uncertainty of a maximum likelihood estimate.

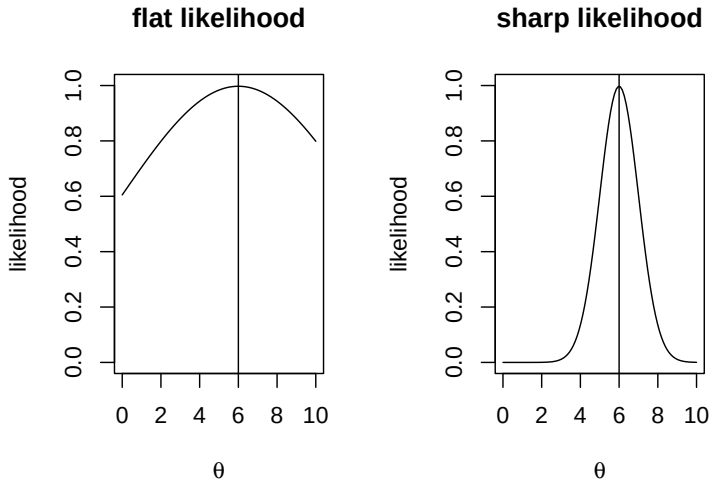


Figure 9.1: Flat and sharp log-likelihood function.

Transformation properties

As a consequence of the invariance of the score function and curvature function the **observed Fisher information is invariant against transformations of the sample space**. This is the same invariance also shown by the expected Fisher information and by the KL divergence.

▶ Like the expected Fisher information the observed Fisher information (as a Hessian matrix) transforms covariantly under change of model parameters — see Section 5.1.

Relationship between observed and expected Fisher information

The observed Fisher information $\boldsymbol{J}_n(\hat{\boldsymbol{\theta}}_{ML})$ and the expected Fisher information $\boldsymbol{I}^{\text{Fisher}}(\boldsymbol{\theta})$ are related but also two clearly different entities.

Curvature based:

- Both types of Fisher information are based on computing second order derivatives (Hessian matrix), thus both are based on the curvature of a function.

Transformation properties:

- Both quantities are invariant against changes of the parametrisation of the sample space.
- ▶ Both transform covariantly when changing the parameter of the distribution.

Data-based vs. model only:

- The observed Fisher information is computed from the log-likelihood function. Therefore it takes both the model and the observed data D into account and explicitly depends on the sample size n. It contains estimates of the parameters but not the parameters themselves. While the curvature of the log-likelihood function may be computed for any point of the log-likelihood function the observed Fisher information specifically refers to the curvature at the MLE $\hat{\boldsymbol{\theta}}_{ML}$. It is linked to the (asymptotic) variance of the MLE (see the examples and as will be discussed in more detail later).
- In contrast, the expected Fisher information is derived directly from the log-density of the model family. It does not depend on the observed data, and thus does not depend on sample size. It makes sense and can be computed for any value of the parameters. It describes the geometry of the space of the model family, and is the local approximation of KL information.

Large sample equivalence:

- Assume that for large sample size n the MLE converges to $\hat{\boldsymbol{\theta}}_{ML} \to \boldsymbol{\theta}_0$. It follows from the construction of the observed Fisher information and the law of large numbers that asymptotically for large sample size $\boldsymbol{J}_n(\hat{\boldsymbol{\theta}}_{ML}) \to n\boldsymbol{I}^{\text{Fisher}}(\boldsymbol{\theta}_0)$ (i.e. the expected Fisher information for a set of iid random variables, see Section 5.1.

▶ Finite sample equivalence for exponential families:

- In a very important class of models, namely for **exponential families**, we find that $\boldsymbol{J}_n(\hat{\boldsymbol{\theta}}_{ML}) = n\boldsymbol{I}^{\text{Fisher}}(\hat{\boldsymbol{\theta}}_{ML})$ is valid also for finite sample size n. This can be directly seen from the special instances of exponential families such as the Bernoulli distribution

(Example 5.1 and Example 9.1), the normal distribution with one parameter (Example 5.3 and Example 9.2), the normal distribution with two parameters (Example 5.4 and Example 9.4) and the categorical distribution (Example 5.5 and Example 9.5).

- However, exponential families are an exception. In a general model $\boldsymbol{J}_n(\hat{\boldsymbol{\theta}}_{ML}) \neq n\boldsymbol{I}^{\text{Fisher}}(\hat{\boldsymbol{\theta}}_{ML})$ for finite sample size n. As an example consider the location-scale t-distribution $t_\nu(\mu, \tau^2)$ with unknown median parameter μ and known scale parameter τ^2 and given degree of freedom ν. This is not an exponential family model (unless $\nu \to \infty$ when it becomes the normal distribution). It can be shown that the expected Fisher information is $I^{\text{Fisher}}(\mu) = \frac{\nu+1}{\nu+3}\frac{1}{\tau^2}$ but the observed Fisher information $J_n(\hat{\mu}_{ML}) \neq n\frac{\nu+1}{\nu+3}\frac{1}{\tau^2}$ with a maximum likelihood estimate $\hat{\mu}_{ML}$ that can only be computed numerically with no closed form available.

9.2 Observed Fisher information examples

Models with a single parameter

Example 9.1. Observed Fisher information for the Bernoulli model $\text{Ber}(\theta)$:

We continue Example 8.1. The second derivative of the log-likelihood function is

$$H_n(\theta) = \frac{dS_n(\theta)}{d\theta} = -n\left(\frac{\bar{x}}{\theta^2} + \frac{1-\bar{x}}{(1-\theta)^2}\right)$$

The observed Fisher information is therefore

$$\begin{aligned} J_n(\hat{\theta}_{ML}) &= -H(\hat{\theta}_{ML}) \\ &= n\left(\frac{\bar{x}}{\hat{\theta}_{ML}^2} + \frac{1-\bar{x}}{(1-\hat{\theta}_{ML})^2}\right) \\ &= n\left(\frac{1}{\hat{\theta}_{ML}} + \frac{1}{1-\hat{\theta}_{ML}}\right) \\ &= \frac{n}{\hat{\theta}_{ML}(1-\hat{\theta}_{ML})} \end{aligned}$$

The inverse of the observed Fisher information is:

$$J_n(\hat{\theta}_{ML})^{-1} = \frac{\hat{\theta}_{ML}(1-\hat{\theta}_{ML})}{n}$$

Compare this with $\mathrm{Var}\left(\frac{x}{n}\right) = \frac{\theta(1-\theta)}{n}$ for $x \sim \mathrm{Bin}(n, \theta)$.

Example 9.2. Observed Fisher information for the normal distribution with unknown mean and known variance:

This is the continuation of Example 8.2. The second derivative of the log-likelihood function is

$$H_n(\mu) = \frac{dS_n(\mu)}{d\mu} = -\frac{n}{\sigma^2}$$

The observed Fisher information at the MLE is therefore

$$J_n(\hat{\mu}_{ML}) = -H_n(\hat{\mu}_{ML}) = \frac{n}{\sigma^2}$$

and the inverse of the observed Fisher information is

$$J_n(\hat{\mu}_{ML})^{-1} = \frac{\sigma^2}{n}$$

For $x_i \sim N(\mu, \sigma^2)$ we have $\mathrm{Var}(x_i) = \sigma^2$ and hence $\mathrm{Var}(\bar{x}) = \sigma^2/n$, which is equal to the inverse observed Fisher information.

Example 9.3. Observed Fisher information for the normal distribution with known mean and unknown variance:

This is the continuation of Example 8.3. The second derivative of the log-likelihood function is

$$H_n(\sigma^2) = \frac{dS_n(\sigma^2)}{d\sigma^2} = -\frac{n}{2\sigma^4}\left(\frac{2}{\sigma^2}\overline{(x-\mu)^2} - 1\right)$$

Correspondingly, the observed Fisher information is

$$J_n(\widehat{\sigma^2}_{ML}) = -H_n(\widehat{\sigma^2}_{ML}) = \frac{n}{2}\left(\widehat{\sigma^2}_{ML}\right)^{-2}$$

and its inverse is

$$J_n(\widehat{\sigma^2}_{ML})^{-1} = \frac{2}{n}\left(\widehat{\sigma^2}_{ML}\right)^{2}$$

With $x_i \sim N(\mu, \sigma^2)$ the empirical variance $\widehat{\sigma^2}_{ML}$ follows a one-dimensional Wishart distribution

$$\widehat{\sigma^2}_{\mathrm{ML}} \sim \mathrm{Wis}\left(s^2 = \frac{\sigma^2}{n}, k = n-1\right)$$

(see Section A.8) and hence has variance $\mathrm{Var}(\widehat{\sigma^2}_{ML}) = \frac{n-1}{n}\frac{2\sigma^4}{n}$. For large n this becomes $\mathrm{Var}\left(\widehat{\sigma^2}_{ML}\right) \overset{a}{=} \frac{2}{n}\left(\sigma^2\right)^2$ which is (apart from the "hat") the inverse of the observed Fisher information.

Models with multiple parameters

Example 9.4. Observed Fisher information for the normal distribution with mean and variance parameter:

This is the continuation of Example 8.5.

The Hessian matrix of the log-likelihood function is

$$\begin{aligned}\boldsymbol{H}_n(\mu, \sigma^2) &= \nabla\nabla^T \ell_n(\mu, \sigma^2) \\ &= \begin{pmatrix} -\frac{n}{\sigma^2} & -\frac{n}{\sigma^4}(\bar{x} - \mu) \\ -\frac{n}{\sigma^4}(\bar{x} - \mu) & \frac{n}{2\sigma^4} - \frac{n}{\sigma^6}\left(\overline{x^2} - 2\mu\bar{x} + \mu^2\right) \end{pmatrix}\end{aligned}$$

The negative Hessian at the MLE, i.e. at $\hat{\mu}_{ML} = \bar{x}$ and $\widehat{\sigma^2}_{ML} = \overline{x^2} - \bar{x}^2$, yields the **observed Fisher information matrix**:

$$\begin{aligned}\boldsymbol{J}_n(\hat{\mu}_{ML}, \widehat{\sigma^2}_{ML}) &= -\boldsymbol{H}(\hat{\mu}_{ML}, \widehat{\sigma^2}_{ML}) \\ &= \begin{pmatrix} \frac{n}{\widehat{\sigma^2}_{ML}} & 0 \\ 0 & \frac{n}{2(\widehat{\sigma^2}_{ML})^2} \end{pmatrix}\end{aligned}$$

The observed Fisher information matrix is diagonal with positive entries. Therefore its eigenvalues are all positive as required for a maximum, because for a diagonal matrix the eigenvalues are simply the the entries on the diagonal.

The inverse of the observed Fisher information matrix is

$$\boldsymbol{J}_n(\hat{\mu}_{ML}, \widehat{\sigma^2}_{ML})^{-1} = \begin{pmatrix} \frac{\widehat{\sigma^2}_{ML}}{n} & 0 \\ 0 & \frac{2(\widehat{\sigma^2}_{ML})^2}{n} \end{pmatrix}$$

Example 9.5. ▶ Observed Fisher information of the categorical distribution:

We continue Example 8.7. We first need to compute the negative Hessian matrix of the log likelihood function $-\nabla\nabla^T \ell_n(\pi_1, \ldots, \pi_{K-1})$ and then evaluate it at the MLEs $\hat{\pi}_1^{ML}, \ldots, \hat{\pi}_{K-1}^{ML}$.

The diagonal entries of the Hessian matrix (with $i = 1, \ldots, K-1$) are

$$\frac{\partial^2}{\partial \pi_i^2}\ell_n(\pi_1, \ldots, \pi_{K-1}) = -n\left(\frac{\bar{x}_i}{\pi_i^2} + \frac{\bar{x}_K}{\pi_K^2}\right)$$

and its off-diagonal entries are (with $j = 1, \dots, K-1$)

$$\frac{\partial^2}{\partial \pi_i \partial \pi_j} \ell_n(\pi_1, \dots, \pi_{K-1}) = -\frac{n \bar{x}_K}{\pi_K^2}$$

Thus, the observed Fisher information matrix at the MLE for a categorical distribution is the $K-1 \times K-1$ dimensional matrix

$$\begin{aligned} \boldsymbol{J}_n(\hat{\pi}_1^{ML}, \dots, \hat{\pi}_{K-1}^{ML}) &= n \begin{pmatrix} \frac{1}{\hat{\pi}_1^{ML}} + \frac{1}{\hat{\pi}_K^{ML}} & \cdots & \frac{1}{\hat{\pi}_K^{ML}} \\ \vdots & \ddots & \vdots \\ \frac{1}{\hat{\pi}_K^{ML}} & \cdots & \frac{1}{\hat{\pi}_{K-1}^{ML}} + \frac{1}{\hat{\pi}_K^{ML}} \end{pmatrix} \\ &= n \text{Diag}\left(\frac{1}{\hat{\pi}_1^{ML}}, \dots, \frac{1}{\hat{\pi}_{K-1}^{ML}} \right) + \frac{n}{\hat{\pi}_K^{ML}} \mathbf{1} \end{aligned}$$

Note that the observed Fisher information matrix is the sum of a positive definite matrix (the first diagonal matrix) and a positive semi-definite matrix (the second matrix which is a multiple of $\mathbf{1}$), hence it is positive definite. Hence, the Hessian at the MLE is negative definite as required for a maximum.

For $K = 2$ (cf. Example 9.1) this reduces to the observed Fisher information of a Bernoulli variable

$$\begin{aligned} J_n(\hat{\theta}_{ML}) &= n \left(\frac{1}{\hat{\theta}_{ML}} + \frac{1}{1 - \hat{\theta}_{ML}} \right) \\ &= \frac{n}{\hat{\theta}_{ML}(1 - \hat{\theta}_{ML})} \end{aligned}$$

The inverse of the observed Fisher information is:

$$\boldsymbol{J}_n(\hat{\pi}_1^{ML}, \dots, \hat{\pi}_{K-1}^{ML})^{-1} = \frac{1}{n} \begin{pmatrix} \hat{\pi}_1^{ML}(1 - \hat{\pi}_1^{ML}) & \cdots & -\hat{\pi}_1^{ML} \hat{\pi}_{K-1}^{ML} \\ \vdots & \ddots & \vdots \\ -\hat{\pi}_{K-1}^{ML} \hat{\pi}_1^{ML} & \cdots & \hat{\pi}_{K-1}^{ML}(1 - \hat{\pi}_{K-1}^{ML}) \end{pmatrix}$$

To show that this is indeed the inverse we use the Woodbury matrix identity

$$(\boldsymbol{A} + \boldsymbol{U}\boldsymbol{B}\boldsymbol{V})^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{U}(\boldsymbol{B}^{-1} + \boldsymbol{V}\boldsymbol{A}^{-1}\boldsymbol{U})^{-1}\boldsymbol{V}\boldsymbol{A}^{-1}$$

with

- $B = 1$,
- $\boldsymbol{u} = (\pi_1, \dots, \pi_{K-1})^T$,
- $\boldsymbol{v} = -\boldsymbol{u}^T$,
- $\boldsymbol{A} = \text{Diag}(\boldsymbol{u})$ and its inverse $\boldsymbol{A}^{-1} = \text{Diag}(\pi_1^{-1}, \dots, \pi_{K-1}^{-1})$.

Then $\boldsymbol{A}^{-1}\boldsymbol{u} = \mathbf{1}_{K-1}$ and $1 - \boldsymbol{u}^T\boldsymbol{A}^{-1}\boldsymbol{u} = \pi_K$. With this

$$\boldsymbol{J}_n(\hat{\pi}_1^{ML}, \dots, \hat{\pi}_{K-1}^{ML})^{-1} = \frac{1}{n} \left(\boldsymbol{A} - \boldsymbol{u}\boldsymbol{u}^T \right)$$

and

$$\boldsymbol{J}_n(\hat{\pi}_1^{ML}, \dots, \hat{\pi}_{K-1}^{ML}) = n \left(\boldsymbol{A}^{-1} + \frac{1}{\pi_K} \mathbf{1}_{K-1 \times K-1} \right)$$

10 Quadratic approximation and normal asymptotics

10.1 Approximate distribution of maximum likelihood estimates

Quadratic log-likelihood of the multivariate normal model

Assume we observe a single sample $\boldsymbol{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with known covariance. Noting that the multivariate normal density is

$$f(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{d}{2}} \det(\boldsymbol{\Sigma})^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

the corresponding log-likelihood for $\boldsymbol{\mu}$ is

$$\ell_1(\boldsymbol{\mu}) = C - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$$

where C is a constant that does not depend on $\boldsymbol{\mu}$. Note that the log-likelihood is a quadratic function (both for $\boldsymbol{x}$ and $\boldsymbol{\mu}$) and the maximum of the function lies at $\boldsymbol{\mu} = \boldsymbol{x}$ with value C.

Quadratic approximation of a log-likelihood function

Now consider the quadratic approximation of a general log-likelihood function $\ell_n(\boldsymbol{\theta})$ for $\boldsymbol{\theta}$ around the MLE $\hat{\boldsymbol{\theta}}_{ML}$ (Figure 10.1).

We assume the underlying model is regular and that $\nabla \ell_n(\hat{\boldsymbol{\theta}}_{ML}) = 0$, i.e. the gradient at the maximum vanishes. The Taylor series approximation of scalar-valued function $f(\boldsymbol{x})$ around $\boldsymbol{x}_0$ is

$$f(\boldsymbol{x}) = f(\boldsymbol{x}_0) + \nabla^T f(\boldsymbol{x}_0)\,(\boldsymbol{x}-\boldsymbol{x}_0) + \frac{1}{2}(\boldsymbol{x}-\boldsymbol{x}_0)^T \nabla\nabla^T f(\boldsymbol{x}_0)(\boldsymbol{x}-\boldsymbol{x}_0) + \ldots$$

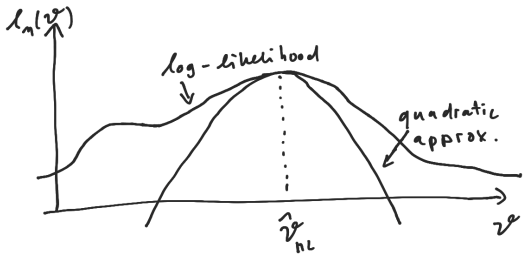


Figure 10.1: Quadratic approximation of the log-likelihood function.

Applied to the log-likelihood function this yields

$$\ell_n(\boldsymbol{\theta}) \approx \ell_n(\hat{\boldsymbol{\theta}}_{ML}) - \frac{1}{2}(\hat{\boldsymbol{\theta}}_{ML}-\boldsymbol{\theta})^T J_n(\hat{\boldsymbol{\theta}}_{ML})(\hat{\boldsymbol{\theta}}_{ML}-\boldsymbol{\theta})$$

This is a quadratic function with maximum at $(\hat{\boldsymbol{\theta}}_{ML}, \ell_n(\hat{\boldsymbol{\theta}}_{ML}))$. Note the appearance of the observed Fisher information $J_n(\hat{\boldsymbol{\theta}}_{ML})$ in the quadratic term. There is no linear term because of the vanishing gradient at the MLE.

Crucially, this approximated log-likelihood takes the same form as if $\hat{\boldsymbol{\theta}}_{ML}$ was sampled from a multivariate normal distribution with mean $\boldsymbol{\theta}$ and with covariance given by the *inverse* observed Fisher information.

Note that this requires a positive definite observed Fisher information matrix so that $J_n(\hat{\boldsymbol{\theta}}_{ML})$ is actually invertible!

Example 10.1. Quadratic approximation of the log-likelihood for a proportion:

From Example 8.1 we have the log-likelihood

$$\ell_n(p) = n\left(\bar{x}\log p + (1-\bar{x})\log(1-p)\right)$$

and the MLE

$$\hat{p}_{ML} = \bar{x}$$

and from Example 9.1 the observed Fisher information

$$J_n(\hat{p}_{ML}) = \frac{n}{\bar{x}(1-\bar{x})}$$

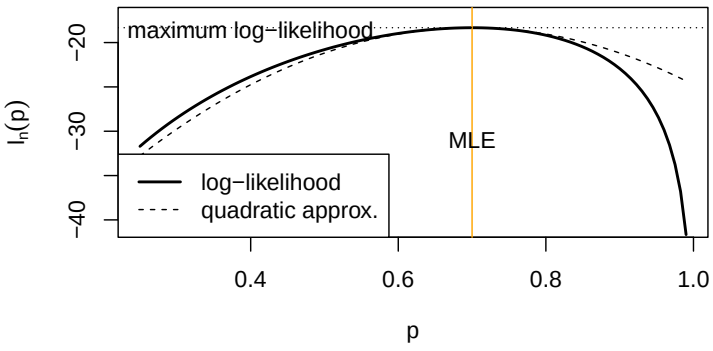


Figure 10.2: Quadratic approximation of the log-likelihood for a Bernoulli model.

The log-likelihood at the MLE is

$$\ell_n(\hat{p}_{ML}) = n\left(\bar{x}\log\bar{x} + (1-\bar{x})\log(1-\bar{x})\right)$$

This allows us to construct the quadratic approximation of the log-likelihood around the MLE as

$$\begin{aligned}\ell_n(p) &\approx \ell_n(\hat{p}_{ML}) - \frac{1}{2}J_n(\hat{p}_{ML})(p-\hat{p}_{ML})^2 \\ &= n\left(\bar{x}\log\bar{x} + (1-\bar{x})\log(1-\bar{x}) - \frac{(p-\bar{x})^2}{2\bar{x}(1-\bar{x})}\right) \\ &= C + \frac{\bar{x}p - \frac{1}{2}p^2}{\bar{x}(1-\bar{x})/n}\end{aligned}$$

The constant C does not depend on p, its function is to match the approximate log-likelihood at the MLE with that of the corresponding original log-likelihood. The approximate log-likelihood takes on the form of a normal log-likelihood (Example 8.2) for one observation of $\hat{p}_{ML} = \bar{x}$ from $N\left(p, \frac{\bar{x}(1-\bar{x})}{n}\right)$.

Figure 10.2 shows the Bernoulli log-likelihood function and its quadratic approximation illustrated for data with $n = 30$ and $\bar{x} = 0.7$:

Asymptotic normality of maximum likelihood estimates

Intuitively, it makes sense to associate large amount of curvature of the log-likelihood at the MLE with low variance of the MLE (and conversely, low amount of curvature with high variance).

From the above we see that for regular models:

- normality implies a quadratic log-likelihood,
- conversely, taking an quadratic approximation of the log-likelihood implies approximate normality, and
- in the quadratic approximation **the inverse observed Fisher information plays the role of the covariance** of the MLE.

This suggests the following theorem:

Asymptotically, the MLE of the parameters of a regular model is normally distributed around the true parameter and with covariance equal to the inverse of the observed Fisher information:

$$\hat{\boldsymbol{\theta}}_{ML} \overset{a}{\sim} \underbrace{N_d}_{\text{multivariate normal}}\left(\underbrace{\boldsymbol{\theta}_{\text{true}}}_{\text{mean vector}}, \underbrace{\boldsymbol{J}_n(\hat{\boldsymbol{\theta}}_{ML})^{-1}}_{\text{covariance matrix}}\right)$$

This theorem about the distributional properties of MLEs greatly enhances the usefulness of the method of maximum likelihood. It implies that in regular settings maximum likelihood is not just a method for obtaining point estimates but also also provides estimates of their uncertainty.

Remarks on the asympotics

However, we need to clarify what "asymptotic" actually means in the context of the above theorem:

1) Primarily, it means to have sufficient sample size so that the log-likelihood $\ell_n(\boldsymbol{\theta})$ is sufficiently well approximated by a quadratic function around $\hat{\boldsymbol{\theta}}_{ML}$. The better the local quadratic approximation the better the normal approximation!

2) In a regular model with positive definite observed Fisher information matrix this is guaranteed for large sample size $n \to \infty$ thanks to the central limit theorem).

3) However, n going to infinity is in fact not always required for the normal approximation to hold! Depending on the particular model a good local fit to a quadratic log-likelihood may be available also for finite n. As a trivial example, for the normal log-likelihood it is valid for any n.

4) In the other hand, in non-regular models (with nondifferentiable log-likelihood at the MLE and/or a singular Fisher information matrix) no amount of data, not even $n \to \infty$, will make the quadratic approximation work.

Remarks:

- The asymptotic normality of MLEs was first discussed in Fisher (1925) [1]
- The technical details of the above considerations are worked out in the theory of locally asymptotically normal (LAN) models pioneered in 1960 by Lucien LeCam (1924–2000).
- There are also methods to obtain higher-order (higher than quadratic and thus non-normal) asymptotic approximations. These relate to so-called saddle point approximations.

Information inequality and asymptotic optimal efficiency

Assume now that $\hat{\boldsymbol{\theta}}$ is an arbitrary and unbiased estimator for $\boldsymbol{\theta}$ and the underlying data-generating model is regular with density $f(\boldsymbol{x}|\boldsymbol{\theta})$.

H. Cramér (1893–1985), C. R. Rao (1920–) and others demonstrated in 1945 the so-called **information inequality**,

$$\operatorname{Var}(\hat{\boldsymbol{\theta}}) \geq \frac{1}{n}\boldsymbol{I}^{\text{Fisher}}(\boldsymbol{\theta}_{\text{true}})^{-1}$$

which puts a lower bound on the variance of an estimator for $\boldsymbol{\theta}$. (Note for $d > 1$ this is a matrix inequality, meaning that the difference matrix is positive semidefinite).

For large sample size with $n \to \infty$ and $\hat{\boldsymbol{\theta}}_{ML} \to \boldsymbol{\theta}$ the observed Fisher information becomes $J_n(\hat{\boldsymbol{\theta}}) \to n\boldsymbol{I}^{\text{Fisher}}(\boldsymbol{\theta})$ and therefore we can write the asymptotic distribution of $\hat{\boldsymbol{\theta}}_{ML}$ as

$$\hat{\boldsymbol{\theta}}_{ML} \overset{a}{\sim} N_d\left(\boldsymbol{\theta}_{\text{true}}, \frac{1}{n}\boldsymbol{I}^{\text{Fisher}}(\boldsymbol{\theta}_{\text{true}})^{-1}\right)$$

[1] Fisher R. A. 1925. *Theory of statistical estimation.* Math. Proc. Cambridge Philos. Soc. **22**:700–725. https://doi.org/10.1017/S0305004100009580

This means that for large n in regular models $\hat{\boldsymbol{\theta}}_{ML}$ achieves the lowest variance possible according to the Cramér-Rao information inequality. In other words, for large sample size maximum likelihood is optimally efficient and thus the best available estimator will in fact be the MLE!

However, as we will see later this does not hold for small sample size where it is indeed possible (and necessary) to improve over the MLE (e.g. via Bayesian estimation or regularisation).

Non-regular models

For non-regular models the asymptotic normality does not hold. Instead, the (asympotic) distribution of the MLE must be obtained in by other means.

Example 10.2. Distribution of MLE for upper bound θ of the uniform distribution:

This continues Example 8.4 to find the distribution of $\hat{\theta}_{ML}$. Due to a discontinuity in the density at the MLE the observed Fisher information cannot be computed and the normal approximation for the distribution of $\hat{\theta}_{ML}$ is not valid.

Nonetheless, one can still obtain the sampling distribution of $\hat{\theta}_{ML} = x_{[n]}$. However, *not* via asymptotic arguments but instead by understanding that $x_{[n]}$ is an order statistic (see https://en.wikipedia.org/wiki/Order_statistic) with the following properties:

$$x_{[n]} \sim \theta\,\text{Beta}(n, 1) \qquad \text{"n-th order statistic"}$$

$$\text{E}\left(x_{[n]}\right) = \tfrac{n}{n+1}\theta$$

$$\text{Var}\left(x_{[n]}\right) = \tfrac{n}{(n+1)^2(n+2)}\theta^2 \quad \approx \tfrac{\theta^2}{n^2}$$

Note that the variance decreases with $\frac{1}{n^2}$ which is much faster than the usual $\frac{1}{n}$ of an "efficient" estimator. Correspondingly, $\hat{\theta}_{ML}$ is a so-called "super efficient" estimator.

10.2 Quantifying the uncertainty of maximum likelihood estimates

Estimating the variance of MLEs

In the previous section we saw that MLEs are asymptotically normally distributed, with the inverse Fisher information (both expected and observed) linked to the asymptotic variance.

This leads to the question whether to use the observed Fisher information $J_n(\hat{\boldsymbol{\theta}}_{ML})$ or the expected Fisher information at the MLE $n\boldsymbol{I}^{\text{Fisher}}(\hat{\boldsymbol{\theta}}_{ML})$ to estimate the variance of the MLE?

- Clearly, for $n \to \infty$ both can be used interchangeably.
- However, they can be very different for finite n in particular for models that are not exponential families.
- Also normality may occur well before n goes to ∞.

Therefore one needs to choose between the two, considering also that

- the expected Fisher information at the MLE is the average curvature at the MLE, whereas the observed Fisher information is the actual observed curvature, and
- the observed Fisher information naturally occurs in the quadratic approximation of the log-likelihood.

All in all, the observed Fisher information as estimator of the variance is more appropriate as it is based on the actual observed data and also works for large n (in which case it yields the same result as using expected Fisher information):

$$\widehat{\text{Var}}(\hat{\boldsymbol{\theta}}_{ML}) = \boldsymbol{J}_n(\hat{\boldsymbol{\theta}}_{ML})^{-1}$$

and its square-root as the estimate of the standard deviation

$$\widehat{\text{SD}}(\hat{\boldsymbol{\theta}}_{ML}) = \boldsymbol{J}_n(\hat{\boldsymbol{\theta}}_{ML})^{-1/2}$$

Note that in the above we use *matrix inversion* and the (inverse) *matrix square root*.

The reasons for preferring observed Fisher information are made mathematically precise in a classic paper by Efron and Hinkley (1978) [2] .

[2] Efron, B., and D. V. Hinkley. 1978. *Assessing the accuracy of the maximum likelihood estimator: observed versus expected Fisher information.* Biometrika **65**:457–482. https://doi.org/10.1093/biomet/65.3.457

Examples for the estimated variance and asymptotic normal distribution

Example 10.3. Estimated variance and distribution of the MLE of a proportion:

From Example 8.1 and Example 9.1 we know the MLE

$$\hat{p}_{ML} = \bar{x} = \frac{k}{n}$$

and the corresponding observed Fisher information

$$J_n(\hat{p}_{ML}) = \frac{n}{\hat{p}_{ML}(1-\hat{p}_{ML})}$$

The estimated variance of the MLE is therefore

$$\widehat{\text{Var}}(\hat{p}_{ML}) = \frac{\hat{p}_{ML}(1-\hat{p}_{ML})}{n}$$

and the corresponding asymptotic normal distribution is

$$\hat{p}_{ML} \overset{a}{\sim} N\left(p, \frac{\hat{p}_{ML}(1-\hat{p}_{ML})}{n}\right)$$

Example 10.4. Estimated variance and distribution of the MLE of the mean parameter for the normal distribution with known variance:

From Example 8.2 and Example 9.2 we know that

$$\hat{\mu}_{ML} = \bar{x}$$

and that the corresponding observed Fisher information at $\hat{\mu}_{ML}$ is

$$J_n(\hat{\mu}_{ML}) = \frac{n}{\sigma^2}$$

The estimated variance of the MLE is therefore

$$\widehat{\text{Var}}(\hat{\mu}_{ML}) = \frac{\sigma^2}{n}$$

and the corresponding asymptotic normal distribution is

$$\hat{\mu}_{ML} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Note that in this case the distribution is not asymptotic but is **exact**, i.e. valid also for small n (as long as the data x_i are actually from $N(\mu, \sigma^2)$!).

Wald statistic

Centering the MLE $\hat{\boldsymbol{\theta}}_{ML}$ with $\boldsymbol{\theta}_0$ followed by standardising with $\widehat{\mathrm{SD}}(\hat{\boldsymbol{\theta}}_{ML})$ yields the **Wald statistic** (named after Abraham Wald, 1902–1950):

$$\begin{aligned}\boldsymbol{t}(\boldsymbol{\theta}_0) &= \widehat{\mathrm{SD}}(\hat{\boldsymbol{\theta}}_{ML})^{-1}(\hat{\boldsymbol{\theta}}_{ML} - \boldsymbol{\theta}_0) \\ &= \boldsymbol{J}_n(\hat{\boldsymbol{\theta}}_{ML})^{1/2}(\hat{\boldsymbol{\theta}}_{ML} - \boldsymbol{\theta}_0)\end{aligned}$$

The **squared Wald statistic** is a scalar defined as

$$\begin{aligned}t(\boldsymbol{\theta}_0)^2 &= \boldsymbol{t}(\boldsymbol{\theta}_0)^T \boldsymbol{t}(\boldsymbol{\theta}_0) \\ &= (\hat{\boldsymbol{\theta}}_{ML} - \boldsymbol{\theta}_0)^T \boldsymbol{J}_n(\hat{\boldsymbol{\theta}}_{ML})(\hat{\boldsymbol{\theta}}_{ML} - \boldsymbol{\theta}_0)\end{aligned}$$

Note that in the literature both $\boldsymbol{t}(\boldsymbol{\theta}_0)$ and $t(\boldsymbol{\theta}_0)^2$ are commonly referred to as Wald statistics. In this text we use the qualifier "squared" if we refer to the latter.

We now assume that the true underlying parameter is $\boldsymbol{\theta}_0$. Since the MLE is asymptotically normal the Wald statistic is asymptotically **standard normal** distributed:

$$\begin{aligned}\boldsymbol{t}(\boldsymbol{\theta}_0) &\overset{a}{\sim} N_d(\mathbf{0}_d, \boldsymbol{I}_d) && \text{for vector } \boldsymbol{\theta} \\ t(\theta_0) &\overset{a}{\sim} N(0,1) && \text{for scalar } \theta\end{aligned}$$

Correspondingly, the **squared** Wald statistic is chi-squared distributed:

$$\begin{aligned}t(\boldsymbol{\theta}_0)^2 &\overset{a}{\sim} \chi^2_d && \text{for vector } \boldsymbol{\theta} \\ t(\theta_0)^2 &\overset{a}{\sim} \chi^2_1 && \text{for scalar } \theta\end{aligned}$$

The degree of freedom of the chi-squared distribution is the dimension d of the parameter vector $\boldsymbol{\theta}$.

Examples of the (squared) Wald statistic

Example 10.5. Wald statistic for a proportion:

We continue from Example 10.3. With $\hat{p}_{ML} = \bar{x}$ and $\widehat{\mathrm{Var}}(\hat{p}_{ML}) = \frac{\hat{p}_{ML}(1-\hat{p}_{ML})}{n}$ and thus $\widehat{\mathrm{SD}}(\hat{p}_{ML}) = \sqrt{\frac{\hat{p}_{ML}(1-\hat{p}_{ML})}{n}}$ we get as **Wald statistic**:

$$t(p_0) = \frac{\bar{x} - p_0}{\sqrt{\bar{x}(1-\bar{x})/n}} \overset{a}{\sim} N(0,1)$$

The **squared Wald statistic** is:

$$t(p_0)^2 = n\frac{(\bar{x} - p_0)^2}{\bar{x}(1-\bar{x})} \overset{a}{\sim} \chi^2_1$$

Example 10.6. Wald statistic for the mean parameter of a normal distribution with known variance:

We continue from Example 10.4. With $\hat{\mu}_{ML} = \bar{x}$ and $\widehat{\mathrm{Var}}(\hat{\mu}_{ML}) = \frac{\sigma^2}{n}$ and thus $\widehat{\mathrm{SD}}(\hat{\mu}_{ML}) = \frac{\sigma}{\sqrt{n}}$ we get as **Wald statistic**:

$$t(\mu_0) = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$$

Note this is the one sample t-statistic with given σ. The **squared Wald statistic** is:

$$t(\mu_0)^2 = \frac{(\bar{x} - \mu_0)^2}{\sigma^2/n} \sim \chi^2_1$$

Again, in this instance this is the exact distribution, not just the asymptotic one.

Using the Wald statistic or the squared Wald statistic we can test whether a particular μ_0 can be rejected as underlying true parameter, and we can also construct corresponding confidence intervals.

Example 10.7. Wald statistic for the categorical distribution:

The squared Wald statistic is

$$\begin{aligned}t(\boldsymbol{p}_0)^2 &= (\hat{\pi}_1^{ML} - p_1^0, \ldots, \hat{\pi}_{K-1}^{ML} - p_{K-1}^0)\boldsymbol{J}_n(\hat{\pi}_1^{ML}, \ldots, \hat{\pi}_{K-1}^{ML}) \begin{pmatrix} \hat{\pi}_1^{ML} - p_1^0 \\ \vdots \\ \hat{\pi}_{K-1}^{ML} - p_{K-1}^0 \end{pmatrix} \\ &= n\left(\sum_{k=1}^{K-1} \frac{(\hat{\pi}_k^{ML} - p_k^0)^2}{\hat{\pi}_k^{ML}} + \frac{\left(\sum_{k=1}^{K-1}(\hat{\pi}_k^{ML} - p_k^0)\right)^2}{\hat{\pi}_K^{ML}}\right) \\ &= n\left(\sum_{k=1}^{K} \frac{(\hat{\pi}_k^{ML} - p_k^0)^2}{\hat{\pi}_k^{ML}}\right) \\ &= nD_{\text{Neyman}}(\mathrm{Cat}(\hat{\boldsymbol{\pi}}_{ML}), \mathrm{Cat}(\boldsymbol{p}_0))\end{aligned}$$

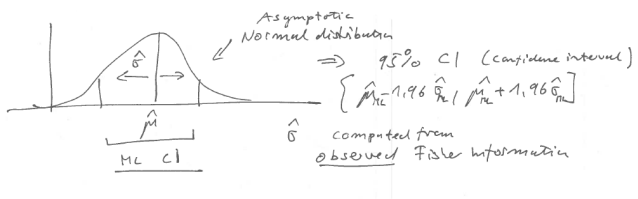


Figure 10.3: Construction of a 95% symmetric normal confidence interval for a maximum likelihood estimate.

With $n_1, \ldots, n_K$ the observed counts with $n = \sum_{k=1}^{K} n_k$ and $\hat{\pi}_k^{ML} = \frac{n_k}{n} = \bar{x}_k$, and $n_1^{\text{expect}}, \ldots, n_K^{\text{expect}}$ the expected counts $n_k^{\text{expect}} = np_k^0$ under $\boldsymbol{p}_0$ we can write the squared Wald statistic as follows:

$$t(\boldsymbol{p}_0)^2 = \sum_{k=1}^{K} \frac{(n_k - n_k^{\text{expect}})^2}{n_k} = \chi^2_{\text{Neyman}}$$

This is known as the Neyman chi-squared statistic (note the *observed* counts in its denominator) and it is asymptotically distributed as χ^2_{K-1} because there are $K-1$ free parameters in $\boldsymbol{p}_0$.

Normal confidence intervals using the Wald statistic

See Section A.10 to review relevant background from year 1.

The asymptotic normality of MLEs derived from regular models enables us to construct a corresponding normal confidence interval (Figure 10.3). For example, to construct the asymptotic normal CI for the MLE of a scalar parameter θ we use the MLE $\hat{\theta}_{ML}$ as estimate of the mean and its standard deviation $\widehat{\text{SD}}(\hat{\theta}_{ML})$ computed from the observed Fisher information:

$$\text{CI} = [\hat{\theta}_{ML} \pm c_{\text{normal}} \widehat{\text{SD}}(\hat{\theta}_{ML})]$$

Here c_{normal} is a critical value for the standard-normal symmetric confidence interval chosen to achieve the desired nominal coverage. The critical values are computed using the inverse standard normal distribution function via $c_{\text{normal}} = \Phi^{-1}\left(\frac{1+\kappa}{2}\right)$. A list of critical values for the standard normal distribution is found in Table A.1. For example, for a CI with 95% coverage one uses the factor 1.96 so that

$$\text{CI} = [\hat{\theta}_{ML} \pm 1.96 \widehat{\text{SD}}(\hat{\theta}_{ML})]$$

The normal CI can be expressed using Wald statistic as follows:

$$\text{CI} = \{\theta_0 : |t(\theta_0)| < c_{\text{normal}}\}$$

Similary, it can also be expressed using the squared Wald statistic:

$$\text{CI} = \{\theta_0 : t(\boldsymbol{\theta}_0)^2 < c_{\text{chisq}}\}$$

Note that this form facilitates the construction of normal confidence intervals for a parameter vector $\boldsymbol{\theta}_0$.

A list of critical values for the chi-squared distribution with one degree of freedom is found in Table A.2.

The following lists contains the critical values resulting from the chi-squared distribution with degree of freedom $m = 1$ for the three most common choices of coverage κ for a normal CI for a univariate parameter: For example, for a 95% interval the critical value equals 3.84 (which is the square of the critical value 1.96 for the standard normal).

Normal tests using the Wald statistic

Finally, recall the **duality between confidence intervals and statistical tests**. Specifically, a confidence interval with coverage κ can be also used for testing as follows:

- for every θ_0 inside the CI the data do not allow to reject the hypothesis that θ_0 is the true parameter with significance level $\alpha = 1 - \kappa$.
- Conversely, all values θ_0 outside the CI can be rejected to be the true parameter with significance level $\alpha = 1 - \kappa$.

Hence, in order to test whether $\boldsymbol{\theta}_0$ is the true underlying parameter value we can compute the corresponding (squared) Wald statistic, find the desired critical value and then decide on rejection.

Examples for normal confidence intervals and corresponding tests

Example 10.8. Asymptotic normal confidence interval for a proportion:

We continue from Example 10.3 and Example 10.5. Assume we observe $n = 30$ measurements with average $\bar{x} = 0.7$. Then $\hat{p}_{ML} = \bar{x} = 0.7$ and $\widehat{\text{SD}}(\hat{p}_{ML}) = \sqrt{\frac{\bar{x}(1-\bar{x})}{n}} \approx 0.084$.

The symmetric asymptotic normal CI for p with 95% coverage is given by $\hat{p}_{ML} \pm 1.96\,\widehat{\mathrm{SD}}(\hat{p}_{ML})$ which for the present data results in the interval $[0.536, 0.864]$.

Example 10.9. Asymptotic normal test for a proportion:

We continue from Example 10.8.

We now consider two possible values ($p_0 = 0.5$ and $p_0 = 0.8$) as potentially true underlying proportion.

The value $p_0 = 0.8$ lies inside the 95% confidence interval $[0.536, 0.864]$. This implies we cannot reject the hypthesis that this is the true underlying parameter on 5% significance level. In contrast, $p_0 = 0.5$ is outside the confidence interval so we can indeed reject this value. In other words, data plus model exclude this value as statistically implausible.

This can be verified more directly by computing the corresponding (squared) Wald statistics (see Example 10.5) and comparing them with the relevant critical value (3.84 from chi-squared distribution for 5% significance level):

- $t(0.5)^2 = \frac{(0.7-0.5)^2}{0.084^2} = 5.71 > 3.84$ hence $p_0 = 0.5$ can be rejected.
- $t(0.8)^2 = \frac{(0.7-0.8)^2}{0.084^2} = 1.43 < 3.84$ hence $p_0 = 0.8$ cannot be rejected.

Note that the squared Wald statistic at the boundaries of the normal confidence interval is equal to the critical value.

Example 10.10. Normal confidence interval for the mean:

We continue from Example 10.4 and Example 10.6. Assume that we observe $n = 25$ measurements with average $\bar{x} = 10$, from a normal with unknown mean and variance $\sigma^2 = 4$.

Then $\hat{\mu}_{ML} = \bar{x} = 10$ and $\widehat{\mathrm{SD}}(\hat{\mu}_{ML}) = \sqrt{\frac{\sigma^2}{n}} = \frac{2}{5}$.

The symmetric asymptotic normal CI for p with 95% coverage is given by $\hat{\mu}_{ML} \pm 1.96\,\widehat{\mathrm{SD}}(\hat{\mu}_{ML})$ which for the present data results in the interval $[9.216, 10.784]$.

Example 10.11. Normal test for the mean:

We continue from Example 10.10.

We now consider two possible values ($\mu_0 = 9.5$ and $\mu_0 = 11$) as potentially true underlying mean parameter.

The value $\mu_0 = 9.5$ lies inside the 95% confidence interval $[9.216, 10.784]$. This implies we cannot reject the hypthesis that this is the true underlying parameter on 5% significance level. In contrast, $\mu_0 = 11$ is outside the confidence interval so we can indeed reject this value. In other words, data plus model exclude this value as a statistically implausible.

This can be verified more directly by computing the corresponding (squared) Wald statistics (see Example 10.6) and comparing them with the relevant critical values:

- $t(9.5)^2 = \frac{(10-9.5)^2}{4/25} = 1.56 < 3.84$ hence $\mu_0 = 9.5$ cannot be rejected.
- $t(11)^2 = \frac{(10-11)^2}{4/25} = 6.25 > 3.84$ hence $\mu_0 = 11$ can be rejected.

The squared Wald statistic at the boundaries of the confidence interval equals the critical value.

Note that this is the standard one-sample test of the mean, and that it is exact, not an approximation.

11 Likelihood-based confidence interval and likelihood ratio

11.1 Likelihood-based confidence intervals and Wilks statistic

General idea and definition of Wilks log-likelihood ratio statistic

Instead of relying on the normal resp. quadratic approximation, we can also use the log-likelihood directly to find **likelihood confidence intervals** (Figure 11.1).

Idea: find all $\boldsymbol{\theta}_0$ that have a log-likelihood that is almost as good as $ell_n(\hat{\boldsymbol{\theta}}_{ML})$.

$$\text{CI} = \{\boldsymbol{\theta}_0 : \ell_n(\hat{\boldsymbol{\theta}}_{ML}) - \ell_n(\boldsymbol{\theta}_0) \leq \Delta\}$$

Here Δ is our tolerated deviation from the maximum log-likelihood. We will see below how to determine a suitable Δ.

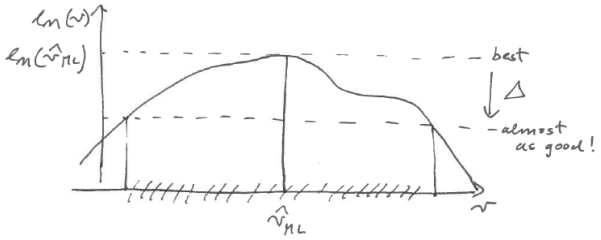


Figure 11.1: Construction of a likelihood-based confidence intervals.

The above leads naturally to the **Wilks log-likelihood ratio statistic** $W(\boldsymbol{\theta}_0)$ defined as:

$$\begin{aligned} W(\boldsymbol{\theta}_0) &= 2\log\left(\frac{L(\hat{\boldsymbol{\theta}}_{ML}|D)}{L(\boldsymbol{\theta}_0|D)}\right) \\ &= 2(\ell_n(\hat{\boldsymbol{\theta}}_{ML}) - \ell_n(\boldsymbol{\theta}_0)) \end{aligned}$$

With its help we can write the likelihood CI as follows:

$$\text{CI} = \{\boldsymbol{\theta}_0 : W(\boldsymbol{\theta}_0) \leq 2\Delta\}$$

The Wilks statistic is named after Samuel S. Wilks (1906–1964).

Advantages of using a likelihood-based CI:

- not restricted to be symmetric
- enables to construct multivariate CIs for parameter vector easily even in non-normal cases
- contains normal CI as special case

Example 11.1. The likelihood ratio statistic:

As alternative to the Wilks log-likelhood ratio statistic $W(\boldsymbol{\theta}_0)$ one may also use the **likelihood ratio** statistic

$$\Lambda(\boldsymbol{\theta}_0) = \frac{L(\boldsymbol{\theta}_0|D)}{L(\hat{\boldsymbol{\theta}}_{ML}|D)}$$

The two statistics can be transformed into each other using

$$W(\boldsymbol{\theta}_0) = -2\log\Lambda(\boldsymbol{\theta}_0)$$

and

$$\Lambda(\boldsymbol{\theta}_0) = e^{-W(\boldsymbol{\theta}_0)/2}$$

Hence large values of $W(\boldsymbol{\theta}_0$ correspond to small values of $\Lambda(\boldsymbol{\theta}_0)$ and the other way round.

In this course we will only use $W(\boldsymbol{\theta}_0)$ as it is both easier to compute and its sampling distribution is easier to obtain.

Examples of the Wilks log-likelihood ratio statistic

Example 11.2. Wilks statistic for the proportion:

The log-likelihood function for the parameter θ is (cf. Example 8.1)

$$\ell_n(\theta) = n(\bar{x}\log\theta + (1-\bar{x})\log(1-\theta))$$

Hence the Wilks statistic is with $\hat{\theta}_{ML} = \bar{x}$

$$\begin{aligned} W(\theta_0) &= 2(\ell_n(\hat{\theta}_{ML}) - \ell_n(\theta_0)) \\ &= 2n\left(\bar{x}\log\left(\frac{\bar{x}}{\theta_0}\right) + (1-\bar{x})\log\left(\frac{1-\bar{x}}{1-\theta_0}\right)\right) \end{aligned}$$

Comparing with Example 4.4 we see that in this case the Wilks statistic is essentially (apart from a scale factor $2n$) the KL divergence between two Bernoulli distributions:

$$W(\theta_0) = 2nD_{\mathrm{KL}}(\mathrm{Ber}(\hat{\theta}_{ML}), \mathrm{Ber}(\theta_0))$$

Example 11.3. Wilks statistic for the mean parameter of a normal model:

The Wilks statistic is

$$W(\mu_0) = \frac{(\bar{x}-\mu_0)^2}{\sigma^2/n}$$

See Worksheet L2 for a derivation of the Wilks statistic from the normal log-likelihood function.

Note this is the same as the squared Wald statistic discussed in Example 10.6.

Comparing with Example 4.5 we see that in this case the Wilks statistic is essentially (apart from a scale factor $2n$) the KL divergence between two normal distributions with different means and variance equal to σ^2:

$$W(p_0) = 2nD_{\mathrm{KL}}(N(\hat{\mu}_{ML}, \sigma^2), N(\mu_0, \sigma^2))$$

Example 11.4. Wilks log-likelihood ratio statistic for the categorical distribution:

The Wilks log-likelihood ratio is

$$W(\boldsymbol{p}_0) = 2\left(\ell_n(\hat{\pi}_1^{ML}, \ldots, \hat{\pi}_{K-1}^{ML}) - \ell_n(p_1^0, \ldots, p_{K-1}^0)\right)$$

with $\boldsymbol{p}_0 = c(p_1^0, \ldots, p_K^0)^T$. As the probabilities sum up to 1 there are only $K-1$ free parameters.

The log-likelihood at the MLE is

$$\ell_n(\hat{\pi}_1^{ML}, \ldots, \hat{\pi}_{K-1}^{ML}) = n\sum_{k=1}^{K}\bar{x}_k\log\hat{\pi}_k^{ML} = n\sum_{k=1}^{K}\bar{x}_k\log\bar{x}_k$$

with $\hat{\pi}_k^{ML} = \frac{n_k}{n} = \bar{x}_k$. Note that here and in the following the sums run from 1 to K where the K-th component is always computed from the components 1 to $K-1$, as in the previous section. The log-likelihood at $\boldsymbol{p}_0$ is

$$\ell_n(p_1^0, \ldots, p_{K-1}^0) = n\sum_{k=1}^{K}\bar{x}_k\log p_k^0$$

so that the Wilks statistic becomes

$$W(\boldsymbol{p}_0) = 2n\sum_{k=1}^{K}\bar{x}_k\log\left(\frac{\bar{x}_k}{p_k^0}\right)$$

It is asymptotically chi-squared distributed with $K-1$ degrees of freedom.

Note that for this model the Wilks statistic is equal to the KL divergence

$$W(\boldsymbol{p}_0) = 2nD_{\mathrm{KL}}(\mathrm{Cat}(\hat{\boldsymbol{\pi}}_{ML}), \mathrm{Cat}(\boldsymbol{p}_0))$$

The Wilks log-likelihood ratio statistic for the categorical distribution is also known as the G test statistic where $\hat{\boldsymbol{\pi}}_{ML}$ corresponds to the observed frequencies (as observed in data) and $\boldsymbol{p}_0$ are the expected frequencies (i.e. hypothesised to be the true frequencies).

Using observed counts n_k and expected counts $n_k^{\mathrm{expect}} = np_k^0$ we can write the Wilks statistic respectively the G-statistic as follows:

$$W(\boldsymbol{p}_0) = 2\sum_{k=1}^{K}n_k\log\left(\frac{n_k}{n_k^{\mathrm{expect}}}\right)$$

Quadratic approximation of the Wilks statistic

Recall the *quadratic approximation* of the log-likelihood function $\ell_n(\boldsymbol{\theta}_0)$ (= second order Taylor series around the MLE $\hat{\boldsymbol{\theta}}_{ML}$):

$$\ell_n(\boldsymbol{\theta}_0) \approx \ell_n(\hat{\boldsymbol{\theta}}_{ML}) - \frac{1}{2}(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}_{ML})^T \boldsymbol{J}_n(\hat{\boldsymbol{\theta}}_{ML})(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}_{ML})$$

With this we can then approximate the Wilks statistic:

$$\begin{aligned} W(\boldsymbol{\theta}_0) &= 2(\ell_n(\hat{\boldsymbol{\theta}}_{ML}) - \ell_n(\boldsymbol{\theta}_0)) \\ &\approx (\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}_{ML})^T \boldsymbol{J}_n(\hat{\boldsymbol{\theta}}_{ML})(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}_{ML}) \\ &= t(\boldsymbol{\theta}_0)^2 \end{aligned}$$

Thus the **quadratic approximation of the Wilks statistic yields the squared Wald statistic**.

Conversely, the Wilks statistic can be understood a generalisation of the squared Wald statistic.

Examples of quadratic approximations

Example 11.5. Quadratic approximation of the Wilks statistic for a proportion (continued from Example 11.2):

The Wilks statistic is

$$W(\theta_0) = 2n\left(\bar{x}\log\left(\frac{\bar{x}}{\theta_0}\right) + (1-\bar{x})\log\left(\frac{1-\bar{x}}{1-\theta_0}\right)\right)$$

Computing Taylor series of second order (for p_0 around $\bar{x}$) yields the following approximations:

$$\log\left(\frac{\bar{x}}{p_0}\right) \approx -\frac{p_0 - \bar{x}}{\bar{x}} + \frac{(p_0-\bar{x})^2}{2\bar{x}^2}$$

and

$$\log\left(\frac{1-\bar{x}}{1-p_0}\right) \approx \frac{p_0 - \bar{x}}{1-\bar{x}} + \frac{(p_0-\bar{x})^2}{2(1-\bar{x})^2}$$

With the above we can approximate the Wilks statistic of the proportion as

$$\begin{aligned} W(p_0) &\approx 2n\left(-(p_0-\bar{x}) + \frac{(p_0-\bar{x})^2}{2\bar{x}} + (p_0-\bar{x}) + \frac{(p_0-\bar{x})^2}{2(1-\bar{x})}\right) \\ &= n\left(\frac{(p_0-\bar{x})^2}{\bar{x}} + \frac{(p_0-\bar{x})^2}{(1-\bar{x})}\right) \\ &= n\left(\frac{(p_0-\bar{x})^2}{\bar{x}(1-\bar{x})}\right) \\ &= t(p_0)^2\,. \end{aligned}$$

This verifies that the quadratic approximation of the Wilks statistic leads back to the squared Wald statistic of Example 10.5.

Example 11.6. Quadratic approximation of the Wilks statistic for the mean parameter of a normal model (continued from Example 11.3):

The normal log-likelihood is already quadratic in the mean parameter (cf. Example 8.2). Correspondingly, the Wilks statistic is quadratic in the mean parameter as well. Hence in this particular case the quadratic "approximation" is in fact exact and the Wilks statistic and the squared Wald statistic are identical!

Correspondingly, confidence intervals and tests based on the Wilks statistic are identical to those obtained using the Wald statistic.

Example 11.7. Quadratic approximation of the Wilks log-likelihood ratio statistic for the categorical distribution:

Developing the Wilks statistic $W(\boldsymbol{p}_0)$ around the MLE $\hat{\boldsymbol{\pi}}_{ML}$ yields the squared Wald statistic which for the categorical distribution is the Neyman chi-squared statistic:

$$\begin{aligned} W(\boldsymbol{p}_0) &= 2nD_{\mathrm{KL}}(\mathrm{Cat}(\hat{\boldsymbol{\pi}}_{ML}), \mathrm{Cat}(\boldsymbol{p}_0)) \\ &\approx nD_{\mathrm{Neyman}}(\mathrm{Cat}(\hat{\boldsymbol{\pi}}_{ML}), \mathrm{Cat}(\boldsymbol{p}_0)) \\ &= \sum_{k=1}^{K} \frac{(n_k - n_k^{\mathrm{expect}})^2}{n_k} \\ &= \chi^2_{\mathrm{Neyman}} \end{aligned}$$

If instead we approximate the KL divergence assuming $\boldsymbol{p}_0$ as fixed we arrive at

$$\begin{aligned} 2nD_{\mathrm{KL}}(\mathrm{Cat}(\hat{\boldsymbol{\pi}}_{ML}), \mathrm{Cat}(\boldsymbol{p}_0)) &\approx nD_{\mathrm{Pearson}}(\mathrm{Cat}(\hat{\boldsymbol{\pi}}_{ML}), \mathrm{Cat}(\boldsymbol{p}_0)) \\ &= \sum_{k=1}^{K} \frac{(n_k - n_k^{\mathrm{expect}})^2}{n_k^{\mathrm{expect}}} \\ &= \chi^2_{\mathrm{Pearson}} \end{aligned}$$

which is the well-known Pearson chi-squared statistic (note the *expected* counts in its denominator).

Distribution of the Wilks statistic

The connection with the squared Wald statistic as quadratic approximation of the Wilks log-likelihood ratio statistic implies that both have asympotically the same distribution.

Hence, under $\boldsymbol{\theta}_0$ the Wilks statistic is distributed asymptotically as

$$W(\boldsymbol{\theta}_0) \overset{a}{\sim} \chi^2_d$$

where d is the number of parameters in $\boldsymbol{\theta}$, i.e. the dimension of the model.

For scalar θ (i.e. single parameter and $d = 1$) this becomes

$$W(\theta_0) \overset{a}{\sim} \chi^2_1$$

This fact is known as **Wilks' theorem**.

Cutoff values for the likelihood CI

Table 11.1: Cutoff values for construction of likelihood confidence intervals for a single parameter.

coverage κ	$\Delta = \frac{c_{\text{chisq}}}{2}$
0.9	1.35
0.95	1.92
0.99	3.32

The asymptotic distribution for W is useful to choose a suitable Δ for the likelihood CI noting that $2\Delta = c_{\text{chisq}}$ where c_{chisq} is the critical value from Table A.2 for a specified coverage κ. This yields Table 11.1 valid for a scalar parameter.

Hence, in order to calibrate the likelihood interval we in effect compare it with a normal confidence interval.

Example 11.8. Likelihood confidence interval for a proportion:

We continue from Example 11.2, and as in Example 10.8 we asssume we have data with $n = 30$ and $\bar{x} = 0.7$.

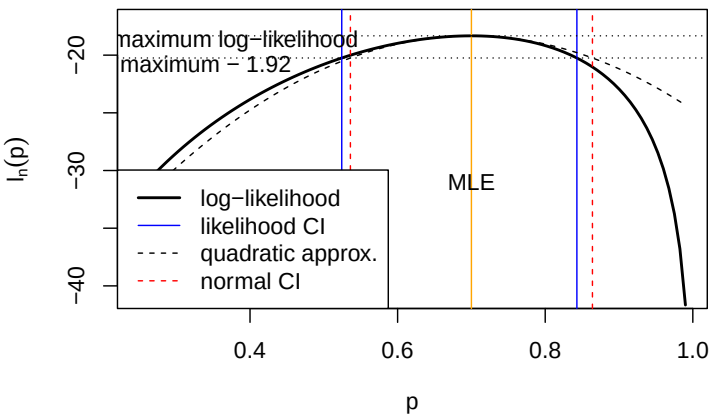


Figure 11.2: Likelihood-based CI and normal interval for the Bernoulli model.

This yields (via numerical root finding) as the 95% likelihood confidence interval the interval $[0.524, 0.843]$. It is similar but not identical to the corresponding asymptotic normal interval $[0.536, 0.864]$ obtained in Example 10.8.

Figure 11.2 illustrates the relationship between the normal CI and the likelihood CI and also shows the role of the quadratic approximation (see also Example 10.1). Note that:

- the normal CI is symmetric around the MLE whereas the likelihood CI is not symmetric
- the normal CI is identical to the likelihood CI when using the quadratic approximation!

Likelihood ratio test (LRT) using Wilks statistic

As in the normal case (with Wald statistic and normal CIs) one can also construct a test using the Wilks statistic:

$H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$	True model is $\boldsymbol{\theta}_0$	Null hypothesis
$H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$	True model is **not** $\boldsymbol{\theta}_0$	Alternative hypothesis

As test statistic we use the Wilks log likelihood ratio $W(\boldsymbol{\theta}_0)$. Extreme values of this test statistic imply evidence against H_0.

Note that the null model is "simple" (= a single parameter value) whereas the alternative model is "composite" (= a set of parameter values).

Remarks:

- The composite alternative H_1 is represented by a single point (the MLE).
- **Reject** H_0 for **large values of** $W(\boldsymbol{\theta}_0)$
- under H_0 and for large n the statistic $W(\boldsymbol{\theta}_0)$ is chi-squared distributed, i.e. $W(\boldsymbol{\theta}_0) \overset{a}{\sim} \chi^2_d$. This allows to compute critical values (i.e tresholds to declared rejection under a given significance level) and also p-values corresponding to the observed test statistics.
- Models **outside** the CI are **rejected**
- Models **inside** the CI **cannot be rejected**, i.e. they can't be statistically distinguished from the best alternative model.

It can be shown that the likelihood ratio test to compare two simple models is optimal in the sense that for any given specified type I error (=probability of wrongly rejecting H_0, i.e. the significance level) it will maximise the power (=1- type II error, probability of correctly accepting H_1). This is known as the **Neyman-Pearson theorem**.

As we have seen previousy, the likelihood-based confidence interval differs from the the confidence interval based on the quadratic / normal approximation. Correspondingly, tests based on the log-likelihood ratio $W(\boldsymbol{\theta}_0)$ and on the squared Wald statistic $t(\theta_0)^2$ will also yield different outcomes (e.g. rejection due to lying outside the confidence interval) even though both test statistics share the same asymptotic distribution and critical values.

Example 11.9. Likelihood test for a proportion:

We continue from Example 11.8 with 95% likelihood confidence interval $[0.524, 0.843]$.

The value $p_0 = 0.5$ is outside the CI and hence can be rejected whereas $p_0 = 0.8$ is insided the CI and hence cannot be rejected on 5% significance level.

The Wilks statistic for $p_0 = 0.5$ and $p_0 = 0.8$ takes on the following values:

- $W(0.5) = 4.94 > 3.84$ hence $p_0 = 0.5$ can be rejected.
- $W(0.8) = 1.69 < 3.84$ hence $p_0 = 0.8$ cannot be rejected.

Note that the Wilks statistic at the boundaries of the likelihood confidence interval is equal to the critical value (3.84 corresponding to 5% significance level for a chi-squared distribution with 1 degree of freedom).

Compare also with the normal test for a proportion in Example 10.9.

Origin of likelihood ratio statistic

The likelihood ratio statistic is asymptotically linked to differences in the KL divergences of the two compared models with the underlying true model.

Assume that F is the true (and unknown) data-generating model and that $G_{\boldsymbol{\theta}}$ is a family of models and we would like to compare two candidate models G_A and G_B corresponding to parameters $\boldsymbol{\theta}_A$ and $\boldsymbol{\theta}_B$ on the basis of observed data $D = \{x_1, \ldots, x_n\}$. The KL divergences $D_{\mathrm{KL}}(F, G_A)$ and $D_{\mathrm{KL}}(F, G_B)$ indicate how close each of the models G_A and G_B fit the true F. The difference of the two divergences is a way to measure the relative fit of the two models, and can be computed as

$$D_{\mathrm{KL}}(F, G_B) - D_{\mathrm{KL}}(F, G_A) = \mathrm{E}_F \log \frac{g(x|\boldsymbol{\theta}_A)}{g(x|\boldsymbol{\theta}_B)}$$

Replacing F by the empirical distribution $\hat{F}_n$ leads to the large sample approximation

$$2n(D_{\mathrm{KL}}(F, G_B) - D_{\mathrm{KL}}(F, G_A)) \approx 2(\ell_n(\boldsymbol{\theta}_A) - \ell_n(\boldsymbol{\theta}_B))$$

Hence, the difference in the log-likelihoods provides an estimate of the difference in the KL divergence of the two models involved.

The Wilks log likelihood ratio statistic

$$W(\boldsymbol{\theta}_0) = 2(\ell_n(\hat{\boldsymbol{\theta}}_{ML}) - \ell_n(\boldsymbol{\theta}_0))$$

thus compares the best-fit distribution with $\hat{\boldsymbol{\theta}}_{ML}$ as the parameter to the distribution with parameter $\boldsymbol{\theta}_0$.

For exponential families the Wilks statistic can also be written in the form of the KL divergence:

$$W(\boldsymbol{\theta}_0) = 2nD_{\mathrm{KL}}(F_{\hat{\boldsymbol{\theta}}_{ML}}, F_{\boldsymbol{\theta}_0})$$

This has been seen in Example 11.2 and Example 11.3. However, this is not true in general.

11.2 Generalised likelihood ratio test (GLRT)

Also known as **maximum likelihood ratio test (MLRT)**. The Generalised Likelihood Ratio Test (GLRT) works just like the standard likelihood ratio test with the difference that now the null model H_0 is also a composite model.

$H_0 : \boldsymbol{\theta} \in \omega_0 \subset \Omega$	True model lies in the restricted space
$H_1 : \boldsymbol{\theta} \in \Omega$	True model lies in the unstricted space

Both H_0 and H_1 are now composite hypotheses. Ω represents the unrestricted model space with dimension (=number of free parameters) $d = |\Omega|$. The constrained space ω_0 has degree of freedom $d_0 = |\omega_0|$ with $d_0 < d$. Note that in the standard LRT the set ω_0 is a simple point with $d_0 = 0$ as the null model is a simple distribution. Thus, LRT is contained in GLRT as special case!

The corresponding generalised (log) likelihood ratio statistic is given by

$$\begin{aligned} W &= 2\log\left(\frac{L(\hat{\theta}_{ML}|D)}{L(\hat{\theta}^0_{ML}|D)}\right) \\ &= 2(\ell_n(\hat{\theta}_{ML}) - \ell_n(\hat{\theta}^0_{ML})) \end{aligned}$$

and

$$\Lambda = \frac{\max\limits_{\theta\in\omega_0} L(\theta|D)}{\max\limits_{\theta\in\Omega} L(\theta|D)}$$

where $L(\hat{\theta}_{ML}|D)$ is the maximised likelihood assuming the full model (with full parameter space Ω) and $L(\hat{\theta}^0_{ML}|D)$ is the maximised likelihood for the restricted model (with restricted parameter space ω_0). Hence, to compute the GRLT test statistic we need to perform two optimisations, one for the full and another for the restricted model.

Remarks:

- MLE in the restricted model space ω_0 is taken as a representative of H_0.
- The likelihood is **maximised** in **both numerator** and **denominator**.
- The restricted model is a special case of the full model (i.e. the two models are nested).
- The asymptotic distribution of W is chi-squared with degree of freedom depending on both d and d_0:

$$W \overset{a}{\sim} \chi^2_{d-d_0}$$

- This result is due to Wilks (1938) [1]. Note that it assumes that the true model is contained among the investigated models.
- If H_0 is a simple hypothesis (i.e. $d_0 = 0$) then the standard LRT (and corresponding CI) is recovered as special case of the GLRT.

Example 11.10. GLRT example:

Case-control study: (e.g. "healthy" vs. "disease")
we observe normal data $D = \{x_1, \ldots, x_n\}$ from two groups with sample size n_1 and n_2 (and $n = n_1 + n_2$), with two different means μ_1 and μ_2 and common variance σ^2:

$$x_1, \ldots, x_{n_1} \sim N(\mu_1, \sigma^2)$$

and

$$x_{n_1+1}, \ldots, x_n \sim N(\mu_2, \sigma^2)$$

Question: are the two means μ_1 and μ_2 the same in the two groups?

$H_0 : \mu_1 = \mu_2$ (with variance unknown, i.e. treated as nuisance parameter)
$H_1 : \mu_1 \neq \mu_2$

Restricted and full models:

ω_0: restricted model with two parameters μ_0 and σ^2_0 (so that $x_1, \ldots, x_n \sim N(\mu_0, \sigma^2_0)$).

Ω: full model with three parameters μ_1, μ_2, σ^2.

[1] Wilks, S. S. 1938. *The large-sample distribution of the likelihood ratio for testing composite hypotheses.* Ann. Math. Statist. **9**:60–62. https://doi.org/10.1214/aoms/1177732360

Corresponding log-likelihood functions:

Restricted model ω_0:

$$\ell_n(\mu_0, \sigma_0^2) = -\frac{n}{2}\log(\sigma_0^2) - \frac{1}{2\sigma_0^2}\sum_{i=1}^{n}(x_i - \mu_0)^2$$

Full model Ω:

$$\begin{aligned}\ell_n(\mu_1, \mu_2, \sigma^2) &= \left(-\frac{n_1}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n_1}(x_i - \mu_1)^2\right) + \\ &\quad \left(-\frac{n_2}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=n_1+1}^{n}(x_i - \mu_2)^2\right) \\ &= -\frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\left(\sum_{i=1}^{n_1}(x_i - \mu_1)^2 + \sum_{i=n_1+1}^{n}(x_i - \mu_2)^2\right)\end{aligned}$$

Corresponding MLEs:

$$\begin{aligned}\omega_0: &\quad \hat{\mu}_0 = \tfrac{1}{n}\textstyle\sum_{i=1}^{n} x_i \qquad \widehat{\sigma_0^2} = \tfrac{1}{n}\textstyle\sum_{i=1}^{n}(x_i - \hat{\mu}_0)^2 \\ \Omega: &\quad \begin{matrix}\hat{\mu}_1 = \tfrac{1}{n_1}\textstyle\sum_{i=1}^{n_1} x_i \\ \hat{\mu}_2 = \tfrac{1}{n_2}\textstyle\sum_{i=n_1+1}^{n} x_i\end{matrix} \qquad \widehat{\sigma^2} = \tfrac{1}{n}\left\{\textstyle\sum_{i=1}^{n_1}(x_i - \hat{\mu}_1)^2 + \sum_{i=n_1+1}^{n}(x_i - \hat{\mu}_2)^2\right\}\end{aligned}$$

Note that the three estimated means are related by

$$\begin{aligned}\hat{\mu}_0 &= \frac{n_1}{n}\hat{\mu}_1 + \frac{n_2}{n}\hat{\mu}_2 \\ &= \hat{\pi_1}\hat{\mu}_1 + \hat{\pi_2}\hat{\mu}_2\end{aligned}$$

so the overall mean is the weighted average of the two individual group means.

Moreover, the two estimated variances are related by

$$\widehat{\sigma_0^2} = \hat{\pi_1}\hat{\pi_2}(\hat{\mu}_1 - \hat{\mu}_2)^2 + \widehat{\sigma^2}$$

Note that this is an example of variance decomposition, where

- $\widehat{\sigma_0^2}$ is the estimated total variance,

- $\hat{\pi_1}\hat{\pi_2}(\hat{\mu}_1 - \hat{\mu}_2)^2$ the estimated between-group (explained) variance, and
- $\widehat{\sigma^2}$ is the estimated average within-group (unexplained) variance.

For the following we also note that

$$\begin{aligned}\frac{\widehat{\sigma_0^2}}{\widehat{\sigma^2}} &= \hat{\pi_1}\hat{\pi_2}\frac{(\hat{\mu}_1 - \hat{\mu}_2)^2}{\widehat{\sigma^2}} + 1 \\ &= \frac{t_{\text{ML}}^2}{n} + 1 \\ &= \frac{t_{\text{UB}}^2}{n-2} + 1\end{aligned}$$

where

$$t_{\text{ML}} = \sqrt{n\hat{\pi_1}\hat{\pi_2}}\frac{\hat{\mu}_1 - \hat{\mu}_2}{\widehat{\sigma^2}}$$

is the two sample t-statistic based on the ML variance estimate $\widehat{\sigma^2}$ and $t_{\text{UB}} = t_{ML}\sqrt{\frac{n-2}{n}}$ is the conventional two sample t-statistic based on the unbiased variance estimate $\widehat{\sigma^2}_{\text{UB}} = \frac{n}{n-2}\widehat{\sigma^2}$ (see Section A.9).

Corresponding maximised log-likelihood:

Restricted model:

$$\ell_n(\hat{\mu}_0, \widehat{\sigma_0^2}) = -\frac{n}{2}\log(\widehat{\sigma_0^2}) - \frac{n}{2}$$

Full model:

$$\ell_n(\hat{\mu}_1, \hat{\mu}_2, \widehat{\sigma^2} = -\frac{n}{2}\log(\widehat{\sigma^2}) - \frac{n}{2}$$

Likelihood ratio statistic:

$$
\begin{aligned}
W &= 2\left(\ell_n\left(\hat{\mu}_1, \hat{\mu}_2, \widehat{\sigma^2}\right) - \ell_n\left(\hat{\mu}_0, \widehat{\sigma_0^2}\right)\right) \\
&= n \log\left(\frac{\widehat{\sigma_0^2}}{\widehat{\sigma^2}}\right) \\
&= n \log\left(\frac{t_{\mathrm{ML}}^2}{n} + 1\right) \\
&= n \log\left(\frac{t_{\mathrm{UB}}^2}{n-2} + 1\right)
\end{aligned}
$$

Thus, the log-likelihood ratio statistic W is a monotonic function (a one-to-one transformation!) of the (squared) two sample t-statistic!

Asymptotic distribution:

The degree of freedom of the full model is $d = 3$ and that of the constrained model $d_0 = 2$ so the generalised log likelihood ratio statistic W is distributed asymptotically as χ_1^2. Hence, we reject the null model on 5% significance level for all $W > 3.84$.

Other application of GLRTs

As shown above, the two sample t statistic can be derived as a likelihood ratio statistic.

More generally, it turns out that many other commonly used familiar statistical tests and test statistics can be interpreted as GLRTs. This shows the wide applicability of this procedure.

12 Optimality properties and conclusion

12.1 Properties of maximum likelihood encountered so far

1. MLE is a special case of KL divergence minimisation *valid for large samples*.
2. MLE can be seen as generalisation of least squares (and conversely, least squares is a special case of ML).

Kullback-Leibler 1951
Entropy learning: minimise $D_{\mathrm{KL}}(F, P_{\boldsymbol{\theta}})$
↓
large n
↓
Fisher 1922
Maximise Likelihood $L(\boldsymbol{\theta}|D)$
↓
normal model
↓
Gauss 1805
Minimise squared error $\sum_i (x_i - \theta)^2$

3. Given a model, derivation of the MLE is basically automatic (only optimisation required)!
4. MLEs are **consistent**, i.e. if the true underlying model F is contained in the set of specified candidates models $F_{\boldsymbol{\theta}}$ then the MLE will converge to the true model corresponding to parameter $\boldsymbol{\theta}_{\mathrm{true}}$.
5. Correspondingly, **MLEs are asympotically unbiased**.
6. However, MLEs are *not* necessarily unbiased in finite samples (e.g. the MLE of the variance parameter in the normal distribution).

7. The maximum likelihood is invariant against parameter transformations.

8. In regular situations (when local quadratic approximation is possible) MLEs are **asympotically normally distributed**, with the asymptotic variance determined by the observed Fisher information.

9. In regular situations and for large sample size MLEs are **asympotically optimally efficient** (Cramer-Rao theorem): For large samples the MLE achieves the lowest possible variance possible in an estimator — this is the so-called Cramer-Rao lower bound. The variance decreases to zero with $n \to \infty$ typically with rate $1/n$.

10. The likelihood ratio can be used to construct optimal tests (in the sense of the Neyman-Pearson theorem).

12.2 Summarising data and the concept of (minimal) sufficiency

Induced partioning of data space and likelihood equivalence

Every sufficient statistic $t(D)$ induces a partitioning of the space of data sets by clustering all hypothetical outcomes for which the statistic $t(D)$ assumes the same value t:

$$\mathcal{X}_t = \{D : t(D) = t\}$$

The **data sets in $\mathcal{X}_t$ are equivalent in terms of the sufficient statistic** $t(D)$. Note that this implies that $t(D)$ is not a 1:1 transformation of D. Instead of n data points $x_1, \ldots, x_n$ as few as one or two summaries (such as empirical mean and variance) may be sufficient to fully convey all the information in the data about the model parameters. Thus, transforming data D using a sufficient statistic $t(D)$ may result in substantial **data reduction**.

Two data sets D_1 and D_2 for which the ratio of the corresponding likelihoods $L(\boldsymbol{\theta}|D_1)/L(\boldsymbol{\theta}|D_2)$ does not depend on $\boldsymbol{\theta}$ (so the two likelihoods are proportional to each other by a constant) are called **likelihood equivalent** because a likelihood-based procedure to learn about $\boldsymbol{\theta}$ will draw identical conclusions from D_1 and D_2. For data sets $D_1, D_2 \in \mathcal{X}_t$ which

are equivalent with respect to a sufficient statistic T it follows directly from the Fisher-Pearson factorisation

$$L(\boldsymbol{\theta}|D) = h(\boldsymbol{t}(D), \boldsymbol{\theta})\, k(D)$$

that the ratio

$$L(\boldsymbol{\theta}|D_1)/L(\boldsymbol{\theta}|D_2) = k(D_1)/k(D_2)$$

and thus is constant with regard to $\boldsymbol{\theta}$. As a result, all **data sets in $\mathcal{X}_t$ are likelihood equivalent**. However, the converse is not true: depending on the sufficient statistics there usually will be many likelihood equivalent data sets that are not part of the same set $\mathcal{X}_t$.

Minimal sufficient statistics

Of particular interest is therefore to find those sufficient statistics that achieve the coarsest partitioning of the sample space and thus may allow the highest data reduction. Specifically, a **minimal sufficient statistic** is a sufficient statistic for which all likelihood equivalent data sets also are equivalent under this statistic.

Therefore, to check whether a sufficient statistic $t(D)$ is minimally sufficient we need to verify whether for any two likelihood equivalent data sets D_1 and D_2 it also follows that $t(D_1) = t(D_2)$. If this holds true then T is a minimally sufficient statistic.

An equivalent non-operational definition is that a minimal sufficient statistic $t(D)$ is a sufficient statistic that can be computed from any other sufficient statistic $S(D)$. This follows from the above directly: assume any sufficient statistic $S(D)$, this defines a corresponding set $\mathcal{X}_s$ of likelihood equivalent data sets. By implication any $D_1, D_2 \in \mathcal{X}_s$ will necessarily also be in $\mathcal{X}_t$, thus whenever $S(D_1) = S(D_2)$ we also have $t(D_1) = t(D_2)$, and therefore $t(D_1)$ is a function of $S(D_1)$.

A trivial but **important example of a minimal sufficient statistic is the likelihood function itself** since by definition it can be computed from any set of sufficient statistics. Thus the likelihood function $L(\boldsymbol{\theta})$ captures all information about $\boldsymbol{\theta}$ that is available in the data. In other words, it provides an *optimal summary* of the observed data with regard to a model. Note that in Bayesian statistics (to be discussed in Part 2 of the module) the likelihood function is used as proxy/summary of the data.

Example: normal distribution

Example 12.1. Sufficient statistics for the parameters of the normal distribution:

The normal model $N(\mu, \sigma^2)$ with parameter vector $\boldsymbol{\theta} = (\mu, \sigma^2)^T$ and log-likelihood

$$\ell_n(\boldsymbol{\theta}) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

One possible set of minimal sufficient statistics for $\boldsymbol{\theta}$ are $\bar{x}$ and $\overline{x^2}$, and with these we can rewrite the log-likelihood function without any reference to the original data $x_1, \ldots, x_n$ as follows

$$\ell_n(\boldsymbol{\theta}) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{n}{2\sigma^2}(\overline{x^2} - 2\bar{x}\mu + \mu^2)$$

An alternative set of minimal sufficient statistics for $\boldsymbol{\theta}$ consists of $s^2 = \overline{x^2} - \bar{x}^2 = \widehat{\sigma^2}_{ML}$ as and $\bar{x} = \hat{\mu}_{ML}$. The log-likelihood written in terms of s^2 and $\bar{x}$ is

$$\ell_n(\boldsymbol{\theta}) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{n}{2\sigma^2}(s^2 + (\bar{x} - \mu)^2)$$

Note that in this example the dimension of the parameter vector $\boldsymbol{\theta}$ equals the dimension of the minimal sufficient statistic, and furthermore, that the MLEs of the parameters are in fact minimal sufficient!

MLEs of parameters of an exponential family are minimal sufficient statistics

The conclusion from Example 12.1 holds true more generally: **in an exponential family model** (such as the normal distribution as particular important case) **the MLEs of the parameters are minimal sufficient statistics**. Thus, there will typically be substantial dimension reduction from the raw data to the sufficient statistics.

However, outside exponential families the MLE is not necessarily a minimal sufficient statistic, and may not even be a sufficient statistic. This is because **a (minimal) sufficient statistic of the same dimension as the parameters does not always exist**. A classic example is the Cauchy distribution for which the minimal sufficient statistics are the ordered observations, thus the MLE of the parameters do not constitute sufficient

statistics, let alone minimal sufficient statistics. However, the MLE is of course still a function of the minimal sufficient statistic.

In summary, the likelihood function acts as perfect data summariser (i.e. as minimally sufficient statistic), and in exponential families (e.g. normal distribution) the MLEs of the parameters $\hat{\boldsymbol{\theta}}_{ML}$ are minimal sufficient.

Finally, while sufficiency is clearly a useful concept for data reduction one needs to keep in mind that this is always in reference to a specific model. Therefore, unless one strongly believes in a certain model it is generally a good idea to keep (and not discard!) the original data.

12.3 Concluding remarks on maximum likelihood

Application of KL divergence in statistics

In statistics the typical roles of the distribution Q and P in the KL divergence $D_{\mathrm{KL}}(Q, P)$ are:

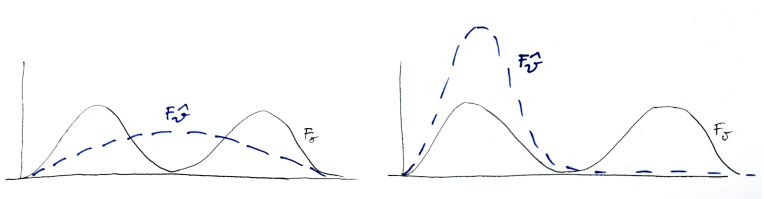
- Q is the (unknown) underlying true model for the data-generating process
- P is the approximating model (typically a parametric distribution family)

Optimising (i.e. minimising) the KL divergence with regard to P amounts to *approximation* and optimising with regard to Q to *imputation*.

In previous chapters we have seen how the KL divergence leads to maximum likelihood (via minimum empirical cross-entropy) and also allows to choose distribution families (via maximum entropy). Later we will also see how KL divergence is linked to Bayesian learning.

Since the KL divergence is not symmetric there two distinct ways to minimise the divergence between a fixed F_0 and the family $F_{\boldsymbol{\theta}}$ (see Figure 12.1). minimising the parameter $\boldsymbol{\theta}$ in $D_{\mathrm{KL}}(\hat{F}_0, F_{\boldsymbol{\theta}})$ ("forward KL") and in $D_{\mathrm{KL}}(F_{\boldsymbol{\theta}}, \hat{F}_0)$ ("backward KL").

Each way has different properties:



(a) Forward KL (b) Reverse KL

Figure 12.1: Illustration of (a) forward KL and b) reverse KL optimisation.

a) **forward KL**, **approximation KL**: $\min_{\theta} D_{\mathrm{KL}}(F_0, F_\theta)$

   Here we keep the first argument fixed and minimise KL by changing the second argument. This is also called an "M (Moment) projection". It has a **zero avoiding** property: $f_\theta(x) > 0$ whenever $f_0(x) > 0$.

   This procedure is mean-seeking and inclusive, i.e. when there are multiple modes in the density of F_0 a fitted unimodal density $F_{\hat{\theta}}$ will seek to cover all modes (**mass covering** property).

   Maximum likelihood is based on "forward KL".

b) **reverse KL**, **inference KL**: $\min_{\theta} D_{\mathrm{KL}}(F_\theta, F_0)$

   Here we keep the second argument fixed and minimise KL by changing the first argument. This is also called an "I (Information) projection". It has a **zero forcing** property: $f_\theta(x) = 0$ whenever $f_0(x) = 0$.

   This procedure is mode-seeking and exclusive, i.e. when there are multiple modes in the density of F_0 a fitted unimodal density $F_{\hat{\theta}}$ will seek out one mode to the exclusion of the others (**mode attracting** property).

   Bayesian updating and variational Bayes approximations use "reverse KL".

What happens if n is small?

From the long list of optimality properties of ML it is clear that for large sample size n the best estimator will typically be the MLE.

However, for **small sample size it is indeed possible (and necessary) to improve over the MLE** (e.g. via Bayesian estimation or regularisation). Some of these ideas will be discussed in Part II.

- Likelihood will *overfit*!

Alternative methods need to be used:

- regularised/penalised likelihood
- Bayesian methods

which are essentially two sides of the same coin.

Classic example of a simple non-ML estimator that is better than the MLE: **Stein's example / Stein paradox** (C. Stein, 1955):

- Problem setting: estimation of the mean in multivariate case
- Maximum likelihood estimation breaks down! → average (=MLE) is worse in terms of MSE than Stein estimator.
- For small n the asymptotic distributions for the MLE and for the LRT are not accurate, so for inference in these situations the distributions may need to be obtained by simulation (e.g. parametric or nonparametric bootstrap).

Model selection

- CI are sets of models that are not statistically distinguishable from the best ML model
- in doubt, choose the simplest model compatible with data
- better prediction, avoids overfitting
- Useful for model exploration and model building.
- Note that, by construction, the model with more parameters always has a higher likelihood, implying likelihood favours complex models
- Complex model may overfit!

- For comparison of models penalised likelihood or Bayesian approaches may be necessary
- Model selection in small samples and high dimension is challenging

- Recall that the aim in statistics is **not** about rejecting models (this is easy as for large sample size any model will be rejected!)

- Instead, the aim is model building, i.e. to find a model that **explains the data well** and that **predicts well**!

- Typically, this will not be the best-fit ML model, but rather a simpler model that is close enough to the best / most complex model.

Part II

Bayesian statistics

13 Conditioning and Bayes rule

In this chapter we review conditional probabilities. Conditional probability is essential for Bayesian statistical modelling.

13.1 Conditional probability

We consider two random variables x and y and assume a **joint density** (or joint PMF) $p(x, y)$. By definition $\int_{x,y} p(x, y)dxdy = 1$.

The **marginal densities** for the individual random variables x and y are given by $p(x) = \int_y p(x, y)dy$ and $p(y) = \int_x f(x, y)dx$, respectively. Therefore, the marginal densities are derived from the joint density by integrating over all possible states of the variable that is being excluded. As necessary for any density, the marginal densities also integrate to one, i.e. $\int_x p(x)dx = 1$ and $\int_y p(y)dy = 1$.

As alternative to integrating out a random variable in the joint density $p(x, y)$ we may wish to keep it fixed at some value. For instance we may want to keep y fixed at y_0. In this case $p(x, y = y_0)$ is proportional to the **conditional density** (or PMF) given by the ratio

$$p(x|y = y_0) = \frac{p(x, y = y_0)}{p(y = y_0)}$$

In this formula the denominator $p(y = y_0) = \int_x p(x, y = y_0)dx$ ensures that $\int_x p(x|y = y_0)dx = 1$, thus it renormalises $p(x, y = y_0)$ so that it becomes a density integrating to one. To simplify notation, the particular value on which a variable is conditioned is often left out, and we just write $p(x|y)$.

13.2 Bayes' theorem

Thomas Bayes (1701-1761) was the first to state Bayes' theorem on conditional probabilities.

Using the definition of conditional probabilities we see that the joint density can be written as the product of marginal and conditional density in two different ways:

$$p(x, y) = p(x|y)p(y) = p(y|x)p(x)$$

This directly leads to **Bayes' theorem**:

$$p(x|y) = p(y|x)\frac{p(x)}{p(y)}$$

This rule relates the two possible conditional densities (or conditional probability mass functions) for the random variables x and y, enabling the change of the order of conditioning.

Bayes's theorem was published in 1763 only after his death by Richard Price (1723-1791):

Pierre-Simon Laplace independently published Bayes' theorem in 1774 and he was in fact the first to routinely apply it to statistical calculations.

13.3 Conditional mean and variance

The conditional distribution $P_{x|y}$ with density $p(x|y)$ has mean $\mathrm{E}(x|y)$ and variance $\mathrm{Var}(x|y)$. These are called the **conditional mean** and **conditional variance**, respectively.

The conditional mean $\mathrm{E}(x|y)$ is als denoted by $\mathrm{E}\left(P_{x|y}\right)$ and $\mathrm{E}_{P_{x|y}}(x)$. It is obtained by calculating

$$\mathrm{E}(x|y) = \mathrm{E}\left(P_{x|y}\right) = \mathrm{E}_{P_{x|y}}(x) = \begin{cases} \sum_x p(x|y)\, x & \text{discrete case} \\ \int_x p(x|y)\, x\, dx & \text{continuous case} \end{cases}$$

The conditional variance $\mathrm{Var}(x|y)$ is als denoted by $\mathrm{Var}\left(P_{x|y}\right)$ and $\mathrm{Var}_{P_{x|y}}(x)$.

$$\mathrm{Var}(x|y) = \mathrm{Var}\left(P_{x|y}\right) = \mathrm{Var}_{P_{x|y}}(x) = \mathrm{E}_{P_{x|y}}((x - \mathrm{E}_{P_{x|y}}(x))^2)) = \mathrm{E}((x - \mathrm{E}(x|y))^2|y)$$

The **law of total expectation** links the means of the marginal distribution and of the conditional distributions, stating that

$$\begin{aligned}\mathrm{E}(x) &= \mathrm{E}\left(\mathrm{E}(x|y)\right) \\ &= \mathrm{E}_{P_y}\mathrm{E}_{P_{x|y}}(x) \\ &= \mathrm{E}_{P_{x,y}}(x)\end{aligned}$$

Hence, the total mean (left side) is the weighted average (outer expectation) of the various conditional means (inner expectation).

Similarly, the **law of total variance** states that

$$\begin{aligned}\mathrm{Var}(x) &= \mathrm{Var}(\mathrm{E}(x|y)) + \mathrm{E}(\mathrm{Var}(x|y)) \\ &= \mathrm{Var}_{P_y}\mathrm{E}_{P_{x|y}}(x) + \mathrm{E}_{P_y}\mathrm{Var}_{P_{x|y}}(x)\end{aligned}$$

The total variance (left side) decomposes into the "explained" variance or "between-group" variance (first term on right side) and the "unexplained" variance or "mean within group" variance (second term on the right side). Again, the outer expectations are with regard to P_y and the inner expecations with regard to $P_{x|y}$.

Example 13.1. Mean and variance of a mixture model:

Assume K groups indicated by a discrete variable $y \in \{1, 2, \ldots, K\}$ with probability $p(y) = \pi_y$. In each group the observations x follow a density $p(x|y)$ with conditional mean $E(x|y) = \mu_y$ and conditional variance $\mathrm{Var}(x|y) = \sigma_y^2$. The joint density for x and y is $p(x, y) = \pi_y p(x|y)$. The marginal density for x is $p(x) = \sum_{y=1}^{K} \pi_y p(x|y)$. This is called a mixture model.

The total mean $\mathrm{E}(x) = \mu_0$ is equal to

$$\mu_0 = \sum_{y=1}^{K} \pi_y \mu_y$$

The total variance $\mathrm{Var}(x) = \sigma_0^2$ is equal to

$$\sigma_0^2 = \sum_{y=1}^{K} \pi_y (\mu_y - \mu_0)^2 + \sum_{y=1}^{K} \pi_y \sigma_y^2$$

13.4 Conditional entropy and entropy chain rules

Similar to mean and variance one can also define conditional versions of entropies. These lead to decomposition of the joint entropy that are also known as **entropy chain rules**.

Conditional entropy

For the entropy of the joint distribution we find that

$$\begin{aligned}H(P_{x,y}) &= -\mathrm{E}_{P_{x,y}} \log p(x, y) \\ &= -\mathrm{E}_{P_x}\mathrm{E}_{P_{y|x}}(\log p(x) + \log p(y|x)) \\ &= -\mathrm{E}_{P_x} \log p(x) - \mathrm{E}_{P_x}\mathrm{E}_{P_{y|x}} \log p(y|x) \\ &= H(P_x) + H(P_{y|x})\end{aligned}$$

thus it decomposes into the entropy of the marginal distribution and the **conditional entropy** defined as

$$H(P_{y|x}) = -\mathrm{E}_{P_x}\mathrm{E}_{P_{y|x}} \log p(y|x)$$

Note that to simplify notation by convention the expectation E_{P_x} over the variable x that we condition on (x) is implicitly assumed.

Conditional cross-entropy

Similarly, for the cross-entropy we get

$$\begin{aligned}H(Q_{x,y}, P_{x,y}) &= -\mathrm{E}_{Q_{x,y}} \log p(x, y) \\ &= -\mathrm{E}_{Q_x}\mathrm{E}_{Q_{y|x}} \log\left(p(x)\, p(y|x)\right) \\ &= -\mathrm{E}_{Q_x} \log p(x) - \mathrm{E}_{Q_x}\mathrm{E}_{Q_{y|x}} \log p(y|x) \\ &= H(Q_x, P_x) + H(Q_{y|x}, P_{y|x})\end{aligned}$$

where the **conditional cross-entropy** is defined as

$$H(Q_{y|x}, P_{y|x}) = -\mathrm{E}_{Q_x}\mathrm{E}_{Q_{y|x}} \log p(y|x)$$

Note again the implicit expectation E_{Q_x} over x implied in this notation.

Conditional KL divergence

The KL divergence between the joint distributions can be decomposed as follows:

$$\begin{aligned} D_{\mathrm{KL}}(Q_{x,y}, P_{x,y}) &= \mathrm{E}_{Q_{x,y}} \log\left(\frac{q(x,y)}{p(x,y)}\right) \\ &= \mathrm{E}_{Q_x}\mathrm{E}_{Q_{y|x}} \log\left(\frac{q(x)q(y|x)}{p(x)p(y|x)}\right) \\ &= \mathrm{E}_{Q_x} \log\left(\frac{q(x)}{p(x)}\right) + \mathrm{E}_{Q_x}\mathrm{E}_{Q_{y|x}} \log\left(\frac{q(y|x)}{p(y|x)}\right) \\ &= D_{\mathrm{KL}}(Q_x, P_x) + D_{\mathrm{KL}}(Q_{y|x}, P_{y|x}) \end{aligned}$$

with the **conditional KL divergence** defined as

$$D_{\mathrm{KL}}(Q_{y|x}, P_{y|x}) = \mathrm{E}_{Q_x}\mathrm{E}_{Q_{y|x}} \log\left(\frac{q(y|x)}{p(y|x)}\right)$$

(again the expectation E_{Q_x} is usually dropped for convenience). The conditional KL divergence can also be computed from the conditional (cross-)entropies by the familiar relationship

$$D_{\mathrm{KL}}(Q_{y|x}, P_{y|x}) = H(Q_{y|x}, P_{y|x}) - H(Q_{y|x})$$

Conditional Boltzmann relative entropy

The Boltzmann entropy is the negative of the KL divergence.

Hence:

$$\begin{aligned} B(Q_{x,y}, P_{x,y}) &= B(Q_x, P_x) + B(Q_{y|x}, P_{y|x}) \\ &= B(Q_y, P_y) + B(Q_{x|y}, P_{x|y}) \end{aligned}$$

13.5 Entropy bounds for the marginal variables

The chain rule for KL divergence directly shows that

$$\begin{aligned} \underbrace{D_{\mathrm{KL}}(Q_{x,y}, P_{x,y})}_{\text{upper bound}} &= D_{\mathrm{KL}}(Q_x, P_x) + \underbrace{D_{\mathrm{KL}}(Q_{y|x}, P_{y|x})}_{\geq 0} \\ &\geq D_{\mathrm{KL}}(Q_x, P_x) \end{aligned}$$

This means that the KL divergence between the joint distributions forms an **upper bound for the KL divergence between the marginal distributions**, with the difference given by the conditional KL divergence $D_{\mathrm{KL}}(Q_{y|x}, P_{y|x})$.

Equivalently, we can state an **upper bound for the marginal cross-entropy**:

$$\begin{aligned} \underbrace{H(Q_{x,y}, P_{x,y}) - H(Q_{y|x})}_{\text{upper bound}} &= H(Q_x, P_x) + \underbrace{D_{\mathrm{KL}}(Q_{y|x}, P_{y|x})}_{\geq 0} \\ &\geq H(Q_x, P_x) \end{aligned}$$

Instead of an upper bound we may as well express this as **lower bound for the negative marginal cross-entropy**

$$\begin{aligned} -H(Q_x, P_x) &= \underbrace{-H(Q_x Q_{y|x}, P_{x,y}) + H(Q_{y|x})}_{\text{lower bound}} + \underbrace{D_{\mathrm{KL}}(Q_{y|x}, P_{y|x})}_{\geq 0} \\ &\geq F\left(Q_x, Q_{y|x}, P_{x,y}\right) \end{aligned}$$

.

Since entropy and KL divergence is closedly linked with maximum likelihood the above bounds play a major role in statistical learning of models with unobserved latent variables (here y). They form the basis of important methods such as the EM algorithm as well as of variational Bayes.

14 Models with latent variables and missing data

14.1 Complete data log-likelihood versus observed data log-likelihood

It is frequently the case that we need to employ models where not all variables are observable and the corresponding data are missing.

For example consider two random variables x and y with a joint density

$$p(x, y|\boldsymbol{\theta})$$

and parameters $\boldsymbol{\theta}$. If we observe data $D_x = \{x_1, \ldots, x_n\}$ and $D_y = \{y_1, \ldots, y_n\}$ for n samples we can use the **complete data log-likelihood**

$$\ell(\boldsymbol{\theta}|D_x, D_y) = \sum_{i=1}^{n} \log p(x_i, y_i|\boldsymbol{\theta})$$

to estimate $\boldsymbol{\theta}$. Recall that

$$\ell(\boldsymbol{\theta}|D_x, D_y) = -nH(\hat{Q}_{x,y}, P_{x,y|\boldsymbol{\theta}})$$

where $\hat{Q}_{x,y}$ is the empirical joint distribution based on both D_x and D_y and $P_{x,y|\boldsymbol{\theta}}$ the joint model, so maximising the complete data log-likelihood minimises the cross-entropy $H(\hat{Q}_{x,y}, P_{x,y|\boldsymbol{\theta}})$.

Now assume that y is not observable and hence is a so-called **latent variable**. Then we don't have observations D_y and therefore cannot use the complete data likelihood. Instead, for maximum likelihood estimation with missing data we need to use the **observed data log-likelihood**.

From the joint density we obtain the marginal density for x by integrating out the unobserved variable y:

$$p(x|\boldsymbol{\theta}) = \int_y p(x, y|\boldsymbol{\theta}) dy$$

Using the marginal model we then compute the **observed data log-likelihood**

$$\ell(\boldsymbol{\theta}|D_x) = \sum_{i=1}^{n} \log p(x_i|\boldsymbol{\theta}) = \sum_{i=1}^{n} \log \int_y p(x_i, y|\boldsymbol{\theta}) dy$$

Note that only the data D_x are used.

Maximum likelihood estimation based on the marginal model proceeds as usual by maximising the corresponding observed data likelihood function which is

$$\ell(\boldsymbol{\theta}|D_x) = -nH(\hat{Q}_x, P_{x|\boldsymbol{\theta}})$$

where $\hat{Q}_x$ is the empirical distribution based only on D_x and $P_{x|\boldsymbol{\theta}}$ is the model family. Hence, maximising the observed data log-likelihood minimises the cross-entropy $H(\hat{Q}_x, P_{x|\boldsymbol{\theta}})$.

Example 14.1. Two group normal mixture model:

Assume we have two groups labelled by $y = 1$ and $y = 2$ (thus the variable y is discrete). The data x observed in each group are normal with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2, respectively. The probability of group 1 is $\pi_1 = p$ and the probability of group 2 is $\pi_2 = 1 - p$. The density of the joint model for x and y is

$$p(x, y|\boldsymbol{\theta}) = \pi_y N(x|\mu_y, \sigma_y)$$

The model parameters are $\boldsymbol{\theta} = (p, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2)^T$ and they can be inferred from the complete data comprised of $D_x = \{x_1, \ldots, x_n\}$ and the group allocations $D_y = \{y_1, \ldots, y_n\}$ of each sample using the complete data log-likelihood

$$\ell(\boldsymbol{\theta}|D_x, D_y) = \sum_{i=1}^{n} \log \pi_{y_i} + \sum_{i=1}^{n} \log N(x_i|\mu_{y_i}, \sigma_{y_i})$$

However, typically we do not know the class allocation y and thus we need to use the marginal model for x alone which has density

$$\begin{aligned} p(x|\boldsymbol{\theta}) &= \sum_{y=1}^{2} \pi_y N(\mu_y, \sigma_y^2) \\ &= pN(x|\mu_1, \sigma_1^2) + (1-p)N(x|\mu_2, \sigma_2^2) \end{aligned}$$

This is an example of a **two-component mixture model**. The corresponding observed data log-likelihood is

$$\ell(\boldsymbol{\theta}|D_x) = \sum_{i=1}^{n} \log \sum_{y=1}^{2} \pi_y N(x|\mu_y, \sigma_y^2)$$

Note that the form of the observed data log-likelihood is more complex than that of the complete data log-likelihood because it contains the logarithm of a sum that cannot be simplified. It is used to estimate the model parameters $\boldsymbol{\theta}$ from D_x without requiring knowledge of the class allocations D_y.

Example 14.2. Alternative computation of the observed data likelihood:

An alternative way to arrive at the observed data likelihood is to marginalise the complete data likelihood.

$$L(\boldsymbol{\theta}|D_x, D_y) = \prod_{i=1}^{n} p(x_i, y_i|\boldsymbol{\theta})$$

and

$$L(\boldsymbol{\theta}|D_x) = \int_{y_1,\ldots,y_n} \prod_{i=1}^{n} p(x_i, y_i|\boldsymbol{\theta}) dy_1 \ldots dy_n$$

The integration (sum) and the multiplication can be interchanged as per Generalised Distributive Law leading to

$$L(\boldsymbol{\theta}|D_x) = \prod_{i=1}^{n} \int_{y} p(x_i, y|\boldsymbol{\theta}) dy$$

which is the same as constructing the likelihood from the marginal density.

14.2 Estimation of the unobservable latent states using Bayes theorem

After estimating the marginal model it is straightforward to obtain a probabilistic prediction about the state of the latent variables $y_1, \ldots, y_n$. Since

$$p(x, y|\boldsymbol{\theta}) = p(x|\boldsymbol{\theta})\, p(y|x, \boldsymbol{\theta}) = p(y|\boldsymbol{\theta})\, p(x|y, \boldsymbol{\theta})$$

given an estimate $\hat{\boldsymbol{\theta}}$ we are able to compute for each observation x_i

$$p(y_i|x_i, \hat{\boldsymbol{\theta}}) = \frac{p(x_i, y_i|\hat{\boldsymbol{\theta}})}{p(x_i|\hat{\boldsymbol{\theta}})} = \frac{p(y_i|\hat{\boldsymbol{\theta}})\, p(x_i|y_i, \hat{\boldsymbol{\theta}})}{p(x_i|\hat{\boldsymbol{\theta}})}$$

the probabilities / densities of all states of y_i (note this an application of Bayes' theorem).

Example 14.3. Latent states of two group normal mixture model:

Continuing from Example 14.1 above we assume the marginal model has been fitted with parameter values $\hat{\boldsymbol{\theta}} = (\hat{p}, \hat{\mu}_1, \hat{\mu}_2, \widehat{\sigma_1^2}, \widehat{\sigma_2^2})^T$. Then for each sample x_i we can get probabilistic prediction about group assocation of each sample by

$$p(y_i|x_i, \hat{\boldsymbol{\theta}}) = \frac{\hat{\pi}_{y_i} N(x_i|\hat{\mu}_{y_i}, \widehat{\sigma_{y_i}^2})}{\hat{p} N(x_i|\hat{\mu}_1, \widehat{\sigma_1^2}) + (1 - \hat{p}) N(x_i|\hat{\mu}_2, \widehat{\sigma_2^2})}$$

14.3 EM Algorithm

Computing and maximising the observed data log-likelihood can be difficult because of the integration over the unobserved variable (or summation in case of a discrete latent variable). In contrast, the complete data log-likelihood function may be easier to compute.

The widely used **EM algorithm**, formally described by Dempster and others (1977) but also used before, addresses this problem and maximises the observed data log-likelihood indirectly in an iterative procedure comprising two steps:

1) First ("E" step), the missing data D_y is imputed using Bayes' theorem. This provides probabilities ("soft allocations") for each possible state of the latent variable.
2) Subsequently ("M" step), the expected complete data log-likelihood function is computed, where the expectation is taken with regard to the distribution over the latent states, and it is maximised with regard to $\boldsymbol{\theta}$ to estimate the model parameters.

The EM algorithm leads to the exact same estimates as if the observed data log-likelihood would be optimised directly. Therefore the EM algorithm is in fact *not* an approximation, it is just a different way to find the MLEs.

The EM algorithm and application to clustering is discussed in more detail in the module MATH38161 Multivariate Statistics and Machine Learning.

In a nutshell, the justication for the EM algorithm follows from the entropy chain rules and the corresponding bounds, such as $D_{\mathrm{KL}}(Q_{x,y}, P_{x,y}) \geq D_{\mathrm{KL}}(Q_x, P_x)$ (see previous chapter). Given observed data for x we know the empirical distribution $\hat{Q}_x$. Hence, by minimising $D_{\mathrm{KL}}(\hat{Q}_x Q_{y|x}, P^{\boldsymbol{\theta}}_{x,y})$ iteratively

1) with regard to $Q_{y|x}$ ("E" step) and
2) with regard to the parameters $\boldsymbol{\theta}$ of $P^{\boldsymbol{\theta}}_{x,y}$ ("M" step")

one minimises $D_{\mathrm{KL}}(\hat{Q}_x, P^{\boldsymbol{\theta}}_x)$ with regard to the parameters of $P^{\boldsymbol{\theta}}_x$.

Interestingly, in the "E" step the first argument of the KL divergence is optimised ("I" projection) and in the "M" step the second argument ("M" projection).

Alternatively, instead of bounding the marginal KL divergence one can also either minimise the upper bound of the cross-entropy or maximise the lower bound of the negative cross-entropy. All of these three procedures yield the same EM algorithm.

Note that the optimisation of the entropy bound in the "E" step requires variational calculus since the argument is a distribution! The EM algorithm is therefore in fact a special case of a **variational Bayes algorithm** since it not only provides estimates of $\boldsymbol{\theta}$ but also yields the distribution of the latent states by means of the calculus of variations.

Finally, the previous discussion illustrates that we can gain insights into unobservable states using Bayes' theorem. By applying the same principle to parameters and models, we arrive at the Bayesian approach to statistical learning.

15 Essentials of Bayesian statistics

15.1 Principle of Bayesian learning

From prior to posterior distribution

Bayesian statistical learning applies Bayes' theorem to update our state of knowledge about a parameter in the light of data.

Ingredients:

- $\boldsymbol{\theta}$ parameter(s) of interest, unknown and fixed.
- prior distribution with density $p(\boldsymbol{\theta})$ describing the *uncertainty* (not randomness!) about $\boldsymbol{\theta}$
- data-generating process $p(x|\boldsymbol{\theta})$

Note the **model underlying the Bayesian approach is the joint distribution**

$$p(\boldsymbol{\theta}, x) = p(\boldsymbol{\theta})p(x|\boldsymbol{\theta})$$

as both a prior distribution over the parameters as well as a data-generating process have to be specified.

Question: new information in the form of a new observation x arrives - how does the uncertainty about $\boldsymbol{\theta}$ change?

Answer: use Bayes' theorem to **update the prior density to the posterior density** (see Figure 15.1).

$$\underbrace{p(\boldsymbol{\theta}|x)}_{\text{posterior}} = \underbrace{p(\boldsymbol{\theta})}_{\text{prior}} \frac{p(x|\boldsymbol{\theta})}{p(x)}$$

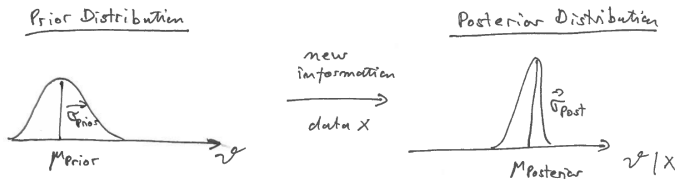


Figure 15.1: Bayesian learning by updating the prior distribution to the posterior distribution.

For the denominator in Bayes formula we need to compute $p(x)$. This is obtained by

$$\begin{aligned} p(x) &= \int_\theta p(x, \boldsymbol{\theta}) d\boldsymbol{\theta} \\ &= \int_\theta p(x|\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta} \end{aligned}$$

i.e. by marginalisation of the parameter $\boldsymbol{\theta}$ from the joint distribution of $\boldsymbol{\theta}$ and x. (For discrete $\boldsymbol{\theta}$ replace the integral by a sum). Depending on the context this quantity is either called the

- **normalisation constant** as it ensures that the posterior density $p(\boldsymbol{\theta}|x)$ integrates to one.
- **prior predictive density** of the data x given the model M before seeing any data. To emphasise the implicit conditioning on a model we may write $p(x|M)$. Since all parameters have been integrated out M in fact refers to a model *class*.
- **marginal likelihood** of the underlying **model** (class) M given data x. To emphasise this may write $L(M|x)$. Sometimes it is also called **model likelihood**.

Zero forcing property

It is easy to see that if in Bayes rule the prior density/probability is zero for some parameter value $\boldsymbol{\theta}$ then the posterior density/probability will remain at zero for that $\boldsymbol{\theta}$, regardless of any data collected. This **zero-forcing property** of the Bayes update rule has been called **Cromwell's rule** by Dennis Lindley (1923–2013). Therefore, assigning prior density/probability 0 to an event should be avoided.

Note that this implies that assigning prior probability 1 should be avoided, too.

Bayesian update and likelihood

After *independent and identically distributed* (iid) data $D = \{x_1, \ldots, x_n\}$ have been observed the Bayesian posterior is computed by

$$\underbrace{p(\boldsymbol{\theta}|D)}_{\text{posterior}} = \underbrace{p(\boldsymbol{\theta})}_{\text{prior}} \frac{L(\boldsymbol{\theta}|D)}{p(D)}$$

involving the likelihood $L(\boldsymbol{\theta}|D) = \prod_{i=1}^n p(x_i|\boldsymbol{\theta})$ and the marginal likelihoood $p(D) = \int_\theta p(\boldsymbol{\theta}) L(\boldsymbol{\theta}|D) d\boldsymbol{\theta}$ with $\boldsymbol{\theta}$ integrated out.

The marginal likelihood serves as a standardising factor so that the posterior density for $\boldsymbol{\theta}$ integrates to 1:

$$\int_\theta p(\boldsymbol{\theta}|D) d\boldsymbol{\theta} = \frac{1}{p(D)} \int_\theta p(\boldsymbol{\theta}) L(\boldsymbol{\theta}|D) d\boldsymbol{\theta} = 1$$

Unfortunately, the integral to compute the marginal likelihood is typically analytically intractable and requires numerical integration and/or approximation.

Comparing likelihood and Bayes procedures note that

- conducting a Bayesian statistical analysis requires integration respectively averaging (to compute the marginal likelihood)
- in contrast to a likelihood analysis that requires optimisation (to find the maximum likelihood).

Sequential updates

Note that the Bayesian update procedure can be repeated again and again: we can use the posterior as our new prior and then update it with further data. Thus, we may also update the posterior density sequentially, with the data points $x_1, \ldots, x_n$ arriving one after the other, by computing first $p(\boldsymbol{\theta}|x_1)$, then $p(\boldsymbol{\theta}|x_1, x_2)$ and so on until we reach $p(\boldsymbol{\theta}|x_1, \ldots, x_n) = p(\boldsymbol{\theta}|D)$.

For example, for the first update we have

$$p(\boldsymbol{\theta}|x_1) = p(\boldsymbol{\theta}) \frac{p(x_1|\boldsymbol{\theta})}{p(x_1)}$$

with $p(x_1) = \int_{\boldsymbol{\theta}} p(x_1|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}$.

The second update yields

$$\begin{aligned} p(\boldsymbol{\theta}|x_1, x_2) &= p(\boldsymbol{\theta}|x_1)\frac{p(x_2|\boldsymbol{\theta}, x_1)}{p(x_2|x_1)} \\ &= p(\boldsymbol{\theta}|x_1)\frac{p(x_2|\boldsymbol{\theta})}{p(x_2|x_1)} \\ &= p(\boldsymbol{\theta})\frac{p(x_1|\boldsymbol{\theta})p(x_2|\boldsymbol{\theta})}{p(x_1)p(x_2|x_1)} \end{aligned}$$

with $p(x_2|x_1) = \int_{\boldsymbol{\theta}} p(x_2|\boldsymbol{\theta})p(\boldsymbol{\theta}|x_1)d\boldsymbol{\theta}$.

Note that *given* $\boldsymbol{\theta}$ the x_i are independent, hence $p(x_2|\boldsymbol{\theta}, x_1) = p(x_2|\boldsymbol{\theta})$ and the joint distribution given $\boldsymbol{\theta}$ is $p(x_1, x_2|\boldsymbol{\theta}) = p(x_1|\boldsymbol{\theta})p(x_2|\boldsymbol{\theta})$. In contrast, the joint distribution with $\boldsymbol{\theta}$ integrated out is $p(x_1, x_2) = p(x_1)p(x_2|x_1) \neq p(x_1)p(x_2)$ so without specification of $\boldsymbol{\theta}$ out the x_i are dependent.

The final step is

$$p(\boldsymbol{\theta}|D) = p(\boldsymbol{\theta}|x_1, \ldots, x_n) = p(\boldsymbol{\theta})\frac{\prod_{i=1}^n p(x_i|\boldsymbol{\theta})}{p(D)}$$

with the marginal likelihood factorising into

$$p(D) = \prod_{i=1}^{n} p(x_i|x_{<i})$$

with

$$p(x_i|x_{<i}) = \int_{\boldsymbol{\theta}} p(x_i|\boldsymbol{\theta})p(\boldsymbol{\theta}|x_{<i})d\boldsymbol{\theta}$$

The last factor is the **posterior predictive density** of the new data x_i after seeing data $x_1, \ldots, x_{i-1}$ (given the model class M).

Intuitively, we can understand why the probability of the new x_i depends on the previously observed data points. This is because the uncertainty about the model parameter $\boldsymbol{\theta}$ depends on how many data points we have already observed, and typically with more data the uncertainty about $\boldsymbol{\theta}$ decreases. Therefore the marginal likelihood $p(D)$ is *not* simply the product of the marginal densities $p(x_i)$ at each x_i but instead the product of the conditional densities $p(x_i|x_{<i})$.

Only when the parameter is fully known, and there is no uncertainty about $\boldsymbol{\theta}$, the observations x_i are independent. This leads back to the standard likelihood where we condition on a particular $\boldsymbol{\theta}$ and the likelihood is the product $p(D|\boldsymbol{\theta}) = \prod_{i=1}^n p(x_i|\boldsymbol{\theta})$.

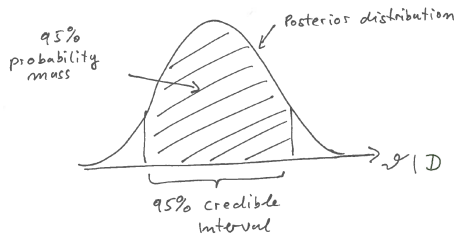


Figure 15.2: Bayesian credible interval based on the posterior distribution.

Summaries of posterior distributions and credible intervals

The Bayesian estimate is the full complete posterior distribution!

However, it is useful to summarise aspects of the posterior distribution:

- Posterior mean $\mathrm{E}(\boldsymbol{\theta}|D)$
- Posterior variance $\mathrm{Var}(\boldsymbol{\theta}|D)$
- Posterior mode etc.

In particular the mean of the posterior distribution is often taken as a *Bayesian point estimate*.

The posterior distribution also allows to define **credible regions** or **credible intervals**. These are the **Bayesian equivalent to confidence intervals** and are constructed by finding the areas of highest probability mass (say 95%) in the posterior distribution (Figure 15.2).

Bayesian credible intervals, unlike frequentist confidence intervals, are thus very easy to interpret as they simply correspond to the part of the parameter space in which we can find the parameter with a given specified probability. In contrast, in frequentist statistics it does not make sense to assign a probability to a parameter value.

Note that there are typically many credible intervals with the given specified coverage α (say 95%). Therefore, we may need further criteria to construct these intervals.

For univariate parameter θ a **two-sided equal-tail credible interval** is obtained by finding the corresponding lower $1 - \alpha/2$ and upper $\alpha/2$ quantiles. Typically this type of credible interval is easy to compute. However, note that the density values at the left and right boundary

points of such an interval are typically different. Also this does not generalise well to a multivariate parameter $\boldsymbol{\theta}$.

As alternative, a **highest posterior density (HPD)** credible interval of coverage α is found by identifying the shortest interval (i.e. with smallest support) for the given α probability mass. Any point within an HDP credible interval has higher density than a point outside the HDP credible interval. Correspondingly, the density at the boundary of an HPD credible interval is constant taking on the same value everywhere along the boundary.

A Bayesian HPD credible interval is constructed in a similar fashion as a likelihood-based confidence interval, starting from the mode of the posterior density and then looking for a common threshold value for the density to define the boundary of the credible interval. When the posterior density has multiple modes the HPD interval may be disjoint. HPD intervals are also well defined for multivariate $\boldsymbol{\theta}$ with the boundaries given by the contour lines of the posterior density resulting from the threshold value.

In the Worksheet B1 examples for both types of credible intervals are given and compared visually.

Practical application of Bayes statistics on the computer

As we have seen Bayesian learning is *conceptually straightforward*:

1) Specify prior uncertainty $p(\boldsymbol{\theta})$ about the parameters of interest $\boldsymbol{\theta}$.
2) Specify the data-generating process for a specified parameter: $p(x|\boldsymbol{\theta})$.
3) Apply Bayes' theorem to update prior uncertainty in the light of the new data.

In practice, however, computing the posterior distribution can be *computationally very demanding*, especially for complex models.

For this reason specialised software packages have been developed for computational Bayesian modelling, for example:

- Bayesian statistics in R: https://cran.r-project.org/web/views/Bayesian.html
- Stan probabilistic programming language (interfaces with R, Python, Julia and other languages) — https://mc-stan.org

- Bayesian statistics in Python: PyMC using PyTensor as backend, NumPyro using JAX as backend, TensorFlow Probability on JAX using JAX as backend, Pyro using PyTorch as backend, TensorFlow Probability using Tensorflow as backend.
- Bayesian statistics in Julia: Turing.jl
- Bayesian hierarchical modelling with BUGS, JAGS and NIMBLE.

In addition to numerical procedures to sample from the posterior distribution there are also many procedures aiming to approximate the Bayesian posterior, employing the Laplace approximation, integrated nested Laplace approximation (INLA), variational Bayes etc.

15.2 Some background on Bayesian statistics

Bayesian interpretation of probability

What makes you "Bayesian"?

If you use Bayes' theorem are you therefore automatically a Bayesian? No!!

Bayes' theorem is a mathematical fact from probability theory. Hence, Bayes' theorem is valid for everyone, whichever form for statistical learning your are subscribing (such as frequentist ideas, likelihood methods, entropy learning, Bayesian learning).

As we discuss now the key difference between Bayesian and frequentist statistical learning lies in the differences in *interpretation of probability*, not in the mathematical formalism for probability (which includes Bayes' theorem).

Mathematics of probability

The mathematics of probability in its modern foundation was developed by Andrey Kolmogorov (1903–1987). In this book Foundations of the Theory of Probability (1933) he establishes probability in terms of set theory/ measure theory. This theory provides a coherent mathematical framework to work with probabilities.

However, Kolmogorov's theory does *not* provide an interpretation of probability!

→ The Kolmogorov framework is the basis for both the frequentist and the Bayesian interpretation of probability.

Interpretations of probability

Essentially, there are two major commonly used interpretation of probability in statistics - the **frequentist interpretation** and the **Bayesian interpretation**.

A: Frequentist interpretation

Probability is interpreted as a relative frequency (of an event in a long-running series of identically repeated experiments, ideally based on infinitely many trials)

Note that the law of large numbers provides justification to interpret long-running relative frequencies as probabilities. However, the converse assumption — and the key frequentist assumption! — that all probabilities have a frequentist interpretation does *not* follow from the law of large numbers.

By definition, the frequentist view of probability is very restrictive. For example, frequentist probability cannot be used to describe events that occur only a single time. Frequentist probability thus can only be applied asymptotically, for large samples.

The frequentist view of probability is also called the *ontological view* of probability as it assumes that probability "exists" independently of an observer and our knowledge.

In turn, ontological probability implies actual inherent randomness which, as we have discussed already in the introduction, amounts only to a small fraction of settings in which probability is used.

As a result, in classical frequentist statistics there are constructs other than probability to assess the uncertainty about parameters (e.g. confidence intervals, sampling distributions).

B: Bayesian probability

"Probability does not exist" — famous quote by Bruno de Finetti (1906–1985), a Bayesian statistician.

What does this mean?

Probability is a **description of the state of knowledge** and of **uncertainty**.

Probability is thus an *epistemological quantity* that is assigned and that changes with more information rather than something that is an inherent property.

Note that this notion of probability does not require any repeated experiments. The Bayesian interpretation of probability is valid regardless of sample size or the number or repetitions of an experiment.

Epistemiological probability can be applied both to single repetitions and to large number of trials, in the latter case (by means of the law of large numbers) it agrees numerically with frequentist probability.

Hence, the key difference between frequentist and Bayesian approaches is not the use of Bayes' theorem. Rather it is whether you consider probability as ontological (frequentist) or epistemological entity (Bayesian).

As a result, in Bayesian statistics to assess the uncertainty about parameters one can use probability directy (in the form of prior and posterior distributions).

Historical developments

- Bayesian statistics is named after Thomas Bayes (1701-1761). His paper [1] introducing the famous theorem was published only after his death (1763).
- Pierre-Simon Laplace (1749-1827) was the first to practically use Bayes' theorem for statistical calculations, and he also independently discovered Bayes' theorem in 1774 [2]
- This activity was then called "inverse probability" and not "Bayesian statistics".
- Between 1900 and 1940 classical mathematical statistics was developed and the field was heavily influenced and dominated by R.A. Fisher (who invented likelihood theory and ANOVA, among other things - he was also working in biology and was professor of genetics). Fisher was very much opposed to Bayesian statistics.

[1] Bayes, T. 1763. *An essay towards solving a problem in the doctrine of chances*. The Philosophical Transactions **53**:370–418. https://doi.org/10.1098/rstl.1763.0053

[2] Laplace, P.-S. 1774. *Mémoire sur la probabilité de causes par les évenements*. Mémoires de mathématique et de physique, présentés à l'Académie Royale des sciences par divers savants et lus dans ses assemblées. Paris, Imprimerie Royale, pp. 621–657.

- 1931 Bruno de Finetti publishes his "representation theorem". This shows that the joint distribution of a sequence of exchangeable events (i.e. where the ordering can be permuted) can be represented by a mixture distribution that can be constructed via Bayes' theorem. (Note that exchangeability is a weaker condition than i.i.d.) This theorem is often used as a justification of Bayesian statistics (along with the so-called Dutch book argument, also by de Finetti).

- 1933 publication of Andrey Kolmogorov's book on probability theory.

- 1946 Cox theorem by Richard T. Cox (1898–1991): the aim to generalise classical logic from TRUE/FALSE statements to continuous measures of uncertainty inevitably leads to probability theory and Bayesian learning! This justification of Bayesian statistics was later popularised by Edwin T. Jaynes (1922–1998) in various books (1959, 2003).

- 1955 Stein Paradox - Charles M. Stein (1920–2016) publishes a paper on the Stein estimator — an estimator of the mean that dominates the ML estimator (i.e. the sample average). The Stein estimator is better in terms of MSE than the ML estimator, which was very puzzling at that time but it is easy to understand from a Bayesian perspective.

- Only from the 1950s the use of the term "Bayesian statistics" became prevalent — see Fienberg (2006) [3]

Due to advances in personal computing from 1970 onwards Bayesian learning has become more pervasive!

- Computers allow to do the complex (numerical) calculations needed in Bayesian statistics .
- Metropolis-Hastings algorithm published in 1970 (which allows to sample from a posterior distribution without explicitly computing the marginal likelihood).
- Development of regularised estimation techniques such as penalised likelihood in regression (e.g. ridge regression 1970).
- penalised likelihood via KL divergence for model selection (Akaike 1973).
- A lot of work on interpreting Stein estimators as empirical Bayes estimators (Efron and Morris 1975)

[3] Fienberg, S. E. 2006. *When did Bayesian inference become "Bayesian"?* Bayesian Analysis **1**:1–40. https://doi.org/10.1214/06-BA101

- regularisation originally was only meant to make singular systems/matrices invertible, but then it turned out regularisation has also a Bayesian interpretation.
- Reference priors (Bernardo 1979) proposed as default priors for models with multiple parameters.
- The EM algorithm (published in 1977) uses Bayes theorem for imputing the distribution of the latent variables.

Another boost was in the 1990/2000s when in science (e.g. genomics) many complex and high-dimensional data set were becoming the norm, not the exception.

- Classical statistical methods cannot be used in this setting (overfitting!) so new methods were developed for high-dimensional data analysis, many with a direct link to Bayesian statistics
- 1996 lasso (L1 regularised) regression invented by Robert Tibshirani.
- Machine learning methods for non-parametric and extremely highly parametric models (neural network) require either explicit or implicit regularisation.
- Many Bayesians in this field, many using variational Bayes techniques which may be viewed as generalisation of the EM algorithm and are also linked to methods used in statistical physics.

16 Bayesian learning in practice

In this chapter we discuss how three basic problems, namely how to estimate a proportion, the mean and the variance in a Bayesian framework.

16.1 Estimating a proportion using the beta-binomial model

Binomial likelihood

In order to apply Bayes' theorem we first need to find a suitable likelihood. We use the Bernoulli model as in Example 8.1:

Repeated Bernoulli experiment (binomial model):

Bernoulli data-generating process:

$$x \sim \text{Ber}(\theta)$$

- $x \in \{0, 1\}$ (e.g. "success" vs. "failure")
- The "success" is indicated by outcome $x = 1$ and the "failure" by $x = 0$
- Parameter: θ is the probability of "success"
- probability mass function (PMF): $\Pr(x = 1) = \theta$, $\Pr(x = 0) = 1 - \theta$
- Mean: $\text{E}(x) = \theta$
- Variance $\text{Var}(x) = \theta(1 - \theta)$

Binomial model $\text{Bin}(n, \theta)$ (sum of n Bernoulli experiments):

- $y \in \{0, 1, \ldots, n\} = \sum_{i=1}^{n} x_i$
- Mean: $\text{E}(y) = n\theta$
- Variance: $\text{Var}(y) = n\theta(1 - \theta)$
- Mean of standardised y: $\text{E}(y/n) = \theta$
- Variance of standardised y: $\text{Var}(y/n) = \frac{\theta(1-\theta)}{n}$

Maximum likelihood estimate of θ:

- We conduct n Bernoulli trials and observe data $D = \{x_1, \ldots, x_n\}$ with average $\bar{x}$ and n_1 successes and $n_2 = n - n_1$ failures.
- Binomial likelihood:

$$L(\theta|D) = \binom{n}{n_1} \theta^{n_1}(1 - \theta)^{n_2}$$

  Note that the binomial coefficient arises as the ordering of the x_i is irrelevant but it may be discarded as is does not contain the parameter θ.
- From Example 8.1 we know that the maximum likelihood estimate of the proportion θ is the frequency

$$\hat{\theta}_{ML} = \frac{n_1}{n} = \bar{x}$$

  Thus, the MLE $\hat{\theta}_{ML}$ can be expressed as an average (of the individual data points). This seemingly trivial fact is important for Bayesian estimation of θ using linear shrinkage, as will become evident below.

Beta distribution

A random variable with support $x \in [0, 1]$ is often described by a **beta distribution**

$$x \sim \text{Beta}(\alpha_1, \alpha_2)$$

with parameters $\alpha_1 \geq 0$ and $\alpha_2 \geq 0$.

For $\alpha_1 = \alpha_2 = 1$ the beta distribution reduces to the uniform distribution $\text{Unif}(0, 1)$.

The density of the beta distribution is

$$p(x|\alpha_1, \alpha_2) = \frac{1}{B(\alpha_1, \alpha_2)} x^{\alpha_1 - 1}(1 - x)^{\alpha_2 - 1}$$

where $B(\alpha_1, \alpha_2) = \frac{\Gamma(\alpha_1)\Gamma(\alpha_2)}{\Gamma(\alpha_1 + \alpha_2)}$ is the beta function.

The mean is

$$\text{E}(x) = \frac{\alpha_1}{\alpha_1 + \alpha_2}$$

and the variance is

$$\text{Var}(x) = \frac{\mu(1 - \mu)}{\alpha_1 + \alpha_2 + 1}$$

Instead of the parameters α_1 and α_2 it is often convenient to consider the mean parametrisation

$$x \sim \text{Beta}(\alpha_1 = m\mu, \alpha_2 = m(1-\mu))$$

with $\mu = \alpha_1/m$ as mean parameter and $m = \alpha_1 + \alpha_2 \geq 0$ as concentration parameter.

In mean parametrisation the mean is

$$\text{E}(x) = \mu$$

and the variance

$$\text{Var}(x) = \frac{\mu(1-\mu)}{m+1}$$

See the Probability and Distribution Refresher notes for further properties of the beta distribution and related distributions (such as the Dirichlet distribution as its multivariate version).

Beta prior distribution

In Bayesian statistics the model is comprised the data-generating process (as for maximum likelihood) but we also require the specification of a prior distribution over the parameters. Therefore, we need to **explicitly modely our prior uncertainty about θ.**

The parameter θ has support $[0, 1]$. Therefore it is natural to use a **beta distribution Beta(α_1, α_2) as prior for** θ We will see below that the beta distribution is a natural choice as a prior in conjunction with a binomial likelihood.

The parameters of a prior (here $\alpha_1 \geq 0$ and $\alpha_2 \geq 0$) are also known as the **hyperparameters** of the model to distinguish them from the parameters of the likelihood function (here θ).

We write for the prior distribution

$$\theta \sim \text{Beta}(\alpha_1, \alpha_2)$$

with density

$$p(\theta) = \frac{1}{B(\alpha_1, \alpha_2)} \theta^{\alpha_1 - 1}(1-\theta)^{\alpha_2 - 1}$$

and prior mean

$$\text{E}(\theta) = \frac{\alpha_1}{\alpha_1 + \alpha_2}$$

It is important that this does not actually mean that θ is random. It only means that we model the uncertainty about θ using a beta-distributed random variable. The flexibility of the beta distribution allows to accommodate a large variety of possible scenarios for our prior knowledge using just two parameters.

Note the mean and variance of the beta prior and the mean and variance of the standardised binomial variable y/n have the same form. This is further indication that the binomial likelihood and the beta prior are well matched — see the discussion below about "conjugate priors".

Computing the posterior distribution

After observing data $D = \{x_1, \ldots, x_n\}$ with n_1 "successes" and $n_2 = n - n_1$ "failures" we can compute the posterior density over θ using Bayes' theorem:

$$p(\theta|D) = \frac{p(\theta)L(\theta|D)}{p(D)}$$

Applying Bayes' theorem results in the posterior distribution:

$$\theta|D \sim \text{Beta}(\alpha_1 + n_1, \alpha_2 + n_2)$$

with density

$$p(\theta|D) = \frac{1}{B(\alpha_1 + n_1, \alpha_2 + n_2)} \theta^{\alpha_1 + n_1 - 1}(1-\theta)^{\alpha_2 + n_2 - 1}$$

and posterior mean

$$\text{E}(\theta|D) = \frac{\alpha_1 + n_1}{\alpha_1 + \alpha_2 + n_1 + n_2}$$

For a proof see Worksheet B1.

Thus, when updating from the prior to the posterior the parameters of the beta distribution are updated in the following simple fashion:

- $\alpha_1 \longrightarrow \alpha_1 + n_1$
- $\alpha_2 \longrightarrow \alpha_2 + n_2$

Update from prior to posterior in terms of mean parametrisation

In is instructive to consider the update in terms of mean and concentration parameters.

- The prior concentration parameter m is set to $k_0 = \alpha_1 + \alpha_2$
- The prior mean parameter μ is set to $\mu_0 = \alpha_1 / k_0$.

The prior mean is therefore

$$\mathrm{E}(\theta) = \mu_0$$

and the prior variance

$$\mathrm{Var}(\theta) = \frac{\mu_0(1-\mu_0)}{k_0 + 1}$$

The posterior mean and concentration parameters are then as follows:

- The concentration parameter m is updated to $k_1 = k_0 + n$
- The mean parameter μ is updated to

$$\mu_1 = \frac{\alpha_1 + n_1}{k_1}$$

This can be written as

$$\begin{aligned}\mu_1 &= \frac{\alpha_1}{k_1} + \frac{n_1}{k_1} \\ &= \frac{k_0}{k_1}\frac{\alpha_1}{k_0} + \frac{n}{k_1}\frac{n_1}{n} \\ &= \lambda\mu_0 + (1-\lambda)\hat{\theta}_{ML}\end{aligned}$$

with $\lambda = \frac{k_0}{k_1}$. Hence, μ_1 is a convex combination of the prior mean and the MLE.

Therefore, the posterior mean is

$$\mathrm{E}(\theta|D) = \mu_1$$

and the posterior variance is

$$\mathrm{Var}(\theta|D) = \frac{\mu_1(1-\mu_1)}{k_1 + 1}$$

Thus the prior to posterior update in mean parametrisation is:

- $k_0 \longrightarrow k_1 = k_0 + n$
- $\mu_0 \longrightarrow \mu_1 = \lambda\mu_0 + (1-\lambda)\hat{\theta}_{ML}$ with $\lambda = k_0/k_1$

16.2 Properties of Bayesian learning

The beta-binomial model, even though it is one of the simplest possible models, already allows to observe a number of important features and properties of Bayesian learning. Many of these apply also to other models as we will see later.

Prior acting as pseudodata

In the expression for the mean and variance you can see that the concentration parameter $k_0 = \alpha_1 + \alpha_2$ behaves like an implicit sample size connected with the prior information about θ.

Specifically, α_1 and α_2 act as **pseudocounts** that influence both the posterior mean and the posterior variance, exactly in the same way as conventional observations.

For example, the larger k_0 (and thus the larger α_1 and α_2) the smaller is the posterior variance, with variance decreasing proportional to the inverse of k_0. If the prior is highly concentrated, i.e. if it has low variance and large precision (=inverse variance) then the implicit data size k_0 is large. Conversely, if the prior has large variance, then the prior is vague and the implicit data size k_0 is small.

Hence, a prior has the same effect as if one would add data — but without actually adding data! This is precisely this why a prior acts as a regulariser and prevents overfitting, because it increases the effective sample size.

Another interpretation is that a prior summarises data that may have been available previously as observations.

Linear shrinkage of mean

In the beta-binomial model the **posterior mean is a convex combination (i.e. the weighted average) of the ML estimate and the prior mean** as can be seen from the update formula

$$\mu_1 = \lambda\mu_0 + (1-\lambda)\hat{\theta}_{ML}$$

with weight $\lambda \in [0, 1]$

$$\lambda = \frac{k_0}{k_1}\,.$$

Thus, the posterior mean μ_1 is a linearly adjusted $\hat{\theta}_{ML}$. The factor λ is called the **shrinkage intensity** — note that this is the ratio of the "prior sample size" (k_0) and the "effective total sample size" (k_1).

1. This adjustment of the MLE is called *shrinkage*, because the $\hat{\theta}_{ML}$ is "shrunk" towards the prior mean μ_0 (which is often called the "target", and sometimes the target is zero, and then the terminology "shrinking" makes most sense).

2. If the shrinkage intensity is zero ($\lambda = 0$) then the ML point estimator is recovered. This happens when $\alpha_1 = 0$ and $\alpha_2 = 0$ or for $n \to \infty$.

   Remark: using maximum likelihood to estimate θ (for moderate or small n) is the same as Bayesian posterior mean estimation using the beta-binomial model with prior $\alpha_1 = 0$ and $\alpha_2 = 0$. This prior is extremely "u-shaped" and the implicit prior for the ML estimation. Would you use such a prior intentionally?

3. If the shrinkage intensity is large ($\lambda \to 1$) then the posterior mean corresponds to the prior. This happens if $n = 0$ or if k_0 is very large (implying that the prior is sharply concentrated around the prior mean).

4. Since the ML estimate $\hat{\theta}_{ML}$ is unbiased the Bayesian point estimate is biased (for finite n!). And the bias is induced by the prior mean deviating from the true mean. This is also true more generally as Bayesian learning typically produces biased estimators (but asymptotically they will be unbiased like in ML).

5. The fact that the posterior mean is a linear combination of the MLE and the prior mean is not a coincidence. In fact, this is true for all distributions that are exponential families, see e.g. Diaconis and Ylvisaker (1979)[1]. Crucially, exponential families can always be parametrised such that the corresponding MLEs are expressed as averages of functions of the data (more technically: the MLE of the mean parameter in an EF is the average of the canonical statistic). In conjunction with a particular type of prior (conjugate priors, always existing for exponential families, see below) this allows to write the update from the prior to posterior mean as a linear adjustment of the MLE.

6. Furthermore, it is possible (and indeed quite useful for computational reasons!) to formulate Bayes learning assuming only first and second moments (i.e. without full distributions) and in terms of linear shrinkage, see e.g. Hartigan (1969)[2]. The resulting theory is called "Bayes linear statistics" (Goldstein and Wooff, 2007)[3].

[1] Diaconis, P., and D Ylvisaker. 1979. *Conjugate Priors for Exponential Families.* Ann. Statist. **7**:269–281. https://doi.org/10.1214/aos/1176344611

Conjugacy of prior and posterior distribution

In the beta-binomial model for estimating the proportion θ the choice of the **beta distribution as prior distribution** along with the binomial likelihood resulted in having the **beta distribution as posterior distribution** as well.

If the prior and posterior belong to the same distributional family the prior is called a **conjugate prior**. This will be the case if the prior has the same functional form as the likelihood. Therefore one also says that the prior is conjugate for the likelihood.

It can be shown that conjugate priors exist for all likelihood functions that are based on data-generating models that are exponential families.

In the beta-binomial model the likelihood is based on the binomial distribution and has the following form (only terms depending on the parameter θ are shown):

$$\theta^{n_1}(1-\theta)^{n_2}$$

The form of the beta prior is (again, only showing terms depending on θ):

$$\theta^{\alpha_1-1}(1-\theta)^{\alpha_2-1}$$

Since the posterior is proportional to the product of prior and likelihood the posterior will have exactly the same form as the prior:

$$\theta^{\alpha_1+n_1-1}(1-\theta)^{\alpha_2+n_2-1}$$

Choosing the prior distribution from a family conjugate for the likelihood greatly simplifies Bayesian analysis since the Bayes formula can then be written in form of an update formula for the parameters of the beta distribution:

$$\alpha_1 \to \alpha_1 + n_1 = \alpha_1 + n\hat{\theta}_{ML}$$
$$\alpha_2 \to \alpha_2 + n_2 = \alpha_2 + n(1-\hat{\theta}_{ML})$$

Thus, conjugate prior distributions are very convenient choices. However, in their application it must be ensured that the prior distribution is flexible

[2] Hartigan, J. A. 1969. *Linear Bayesian methods.* J. Roy. Statist. Soc. B **31**:446-454 https://doi.org/10.1111/j.2517-6161.1969.tb00804.x

[3] Goldstein, M., and D. Wooff. 2007. *Bayes Linear Statistics: Theory and Methods.* Wiley. https://doi.org/10.1002/9780470065662

enough to encapsulate all prior information that may be available. In cases where this is not the case alternative priors should be used (and most likely this will then require to compute the posterior distribution numerically rather than analytically).

Large sample limits of mean and variance

If n is large and $n >> \alpha, \beta$ then $\lambda \to 0$ and hence the posterior mean and variance become asympotically

$$\mathrm{E}(\theta|D) \stackrel{a}{=} \frac{n_1}{n} = \hat{\theta}_{ML}$$

and

$$\mathrm{Var}(\theta|D) \stackrel{a}{=} \frac{\hat{\theta}_{ML}(1-\hat{\theta}_{ML})}{n}$$

Thus, if the sample size is large then the Bayes' estimator turns into the ML estimator! Specifically, the posterior mean becomes the ML point estimate, and the posterior variance is equal to the asymptotic variance computed via the observed Fisher information.

Thus, for large n the data dominate and any details about the prior (such as the settings of the hyperparameters α_1 and α_2) become irrelevant!

Asymptotic normality of the posterior distribution

Also known as **Bayesian Central Limit Theorem (CLT)**.

Under some regularity conditions (such as regular likelihood and positive prior probability for all parameter values, finite number of parameters, etc.) for large sample size the Bayesian posterior distribution converges to a normal distribution centred around the MLE and with the variance of the MLE:

$$\text{for large } n\text{: } p(\boldsymbol{\theta}|D) \to N(\hat{\boldsymbol{\theta}}_{ML}, \mathrm{Var}(\hat{\boldsymbol{\theta}}_{ML}))$$

So not only are the posterior mean and variance converging to the MLE and the variance of the MLE for large sample size, but also the posterior distribution itself converges to the sampling distribution!

This holds generally in many regular cases, not just in the simple case above.

The Bayesian CLT is generally known as the **Bernstein-von Mises theorem** (who discovered it at around 1920–30), but special cases were already known by Laplace.

In the Worksheet B1 the asymptotic convergence of the posterior distribution to a normal distribution is demonstrated graphically.

Posterior variance for finite n

From the Bayesian posterior we can obtain a Bayesian point estimate for the proportion θ by computing the posterior mean

$$\mathrm{E}(\theta|D) = \frac{\alpha_1 + n_1}{k_1} = \hat{\theta}_{\text{Bayes}}$$

along with the posterior variance

$$\mathrm{Var}(\theta|D) = \frac{\hat{\theta}_{\text{Bayes}}(1-\hat{\theta}_{\text{Bayes}})}{k_1 + 1}$$

Asymptotically for large n the posterior mean becomes the maximum likelihood estimate (MLE), and the posterior variance becomes the asymptotic variance of the MLE. Thus, for large n the Bayesian point estimate will be indistinguishable from the MLE and shares its favourable properties.

In addition, for finite sample size the posterior variance will typically be *smaller* than both the asymptotic posterior variance (for large n) and the prior variance, showing that combining the information available in the prior and in the data leads to a more efficient estimate.

16.3 Estimating the mean using the normal-normal model

Normal likelihood

As in Example 8.2 where we estimated the mean parameter by maximum likelihood we assume as data-generating model the normal distribution with unknown mean μ and known variance σ^2:

$$x \sim N(\mu, \sigma^2)$$

We observe n samples $D = \{x_1, \dots x_n\}$. This yields using maximum likelihood the estimate $\hat{\mu}_{ML} = \bar{x}$.

We note that the MLE $\hat{\mu}_{ML}$ is expressed as an average of the data points, which is what enables the linear shrinkage seen below.

Normal prior distribution

The **normal distribution is the conjugate distribution for the mean parameter of a normal likelihood**, so if we use a normal prior then posterior for μ is normal as well.

To model the uncertainty about μ we use the normal distribution in the form $N(\mu, \sigma^2/k)$ with a mean parameter μ and a concentration parameter $k > 0$ (remember that σ^2 is given and is also used in the likelihood).

Specifically, we use as normal **prior distribution** for the mean

$$\mu \sim N\left(\mu_0, \frac{\sigma^2}{k_0}\right)$$

- The prior concentration parameter is set to k_0
- The prior mean parameter is set to μ_0

Hence the prior mean is

$$\mathrm{E}(\mu) = \mu_0$$

and the prior variance

$$\mathrm{Var}(\mu) = \frac{\sigma^2}{k_0}$$

where the concentration parameter k_0 corresponds the implied sample size of the prior. Note that k_0 does not need to be an integer value.

Normal posterior distribution

After observing data D the **posterior distribution** is also normal with updated parameters $\mu = \mu_1$ and k_1

$$\mu|D \sim N\left(\mu_1, \frac{\sigma^2}{k_1}\right)$$

- The posterior concentration parameter is updated to $k_1 = k_0 + n$

- The posterior mean parameter is updated to

$$\mu_1 = \lambda\mu_0 + (1 - \lambda)\hat{\mu}_{ML}$$

  with $\lambda = \frac{k_0}{k_1}$. This can be seen as linear shrinkage of $\hat{\mu}_{ML}$ towards the prior mean μ_0.

(For a proof see Worksheet B2.)

The posterior mean is

$$\mathrm{E}(\mu|D) = \mu_1$$

and the posterior variance is

$$\mathrm{Var}(\mu|D) = \frac{\sigma^2}{k_1}$$

Large sample asymptotics

For n large and $n >> k_0$ the shrinkage intensity $\lambda \to 0$ and $k_1 \to n$. As a result

$$\mathrm{E}(\mu|D) \stackrel{a}{=} \hat{\mu}_{ML}$$

$$\mathrm{Var}(\mu|D) \stackrel{a}{=} \frac{\sigma^2}{n}$$

i.e. we recover the MLE and its asymptotic variance!

Note that for finite n the posterior variance $\frac{\sigma^2}{n+k_0}$ is smaller than both the asymptotic variance $\frac{\sigma^2}{n}$ of the MLE and the prior variance $\frac{\sigma^2}{k_0}$.

16.4 Estimating the variance using the IW-normal model

Normal likelihood

As data-generating model we use normal distribution

$$x \sim N(\mu, \sigma^2)$$

with unknown variance σ^2 and known mean μ. This yields as maximum likelihood estimate for the variance

$$\widehat{\sigma^2}_{ML} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

Note that, again, the MLE is an average (of a quadratic function of the individual data points). This enables linear shrinkage of the MLE as seen below.

IW prior distribution

To model the uncertainty about the variance we use the inverse-gamma (IG) distribution, also known as the univariate inverse Wishart (IW) distribution (refer to the Probability and Distribution Refresher notes for details of this distribution). The IG resp. univariate IW distribution are identical apart from parametrisation. They are conjugate for the variance parameter in the normal likelihood, hence both the prior and the posterior distribution are also IG resp. IW.

In the following we use the Wishart parametrisation, hence we call this an inverse Wishart (IW) prior, and the whole model IW-normal model. However, this model is also often called IG-normal model if the IG distribution an parametrisation is employed as prior.

Specifically, as prior distribution for σ^2 we use the IW distribution in mean parametrisation:

$$\sigma^2 \sim \text{IW}\left(\psi = \kappa_0\sigma_0^2, k = \kappa_0 + 2\right)$$

- The prior concentration parameter is κ_0
- The prior mean parameter is σ_0^2

The corresponding prior mean is

$$\text{E}(\sigma^2) = \sigma_0^2$$

and the prior variance is

$$\text{Var}(\sigma^2) = \frac{2\sigma_0^4}{\kappa_0 - 2}$$

(note that $\kappa_0 > 2$ is required for the variance to exist)

IW posterior distribution

After observing $D = \{x_1 \ldots, x_n\}$ the posterior distribution is also IW with updated parameters:

$$\sigma^2|D \sim \text{IW}\left(\psi = \kappa_1\sigma_1^2, k = \kappa_1 + 2\right)$$

- The posterior concentration parameter is updated to $\kappa_1 = \kappa_0 + n$
- The posterior mean parameter update follows the standard linear shrinkage rule:

  $$\sigma_1^2 = \lambda\sigma_0^2 + (1-\lambda)\widehat{\sigma^2}_{ML}$$

  with $\lambda = \frac{\kappa_0}{\kappa_1}$.

The posterior mean is

$$\text{E}(\sigma^2|D) = \sigma_1^2$$

and the posterior variance

$$\text{Var}(\sigma^2|D) = \frac{2\sigma_1^4}{\kappa_1 - 2}$$

Large sample asymptotics

For large sample size n with $n >> \kappa_0$ the shrinkage intensity vanishes ($\lambda \to 0$) and therefore $\sigma_1^2 \to \widehat{\sigma^2}_{ML}$. We also find that $\kappa_1 - 2 \to n$.

This results in the asymptotic posterior mean

$$\text{E}(\sigma^2|D) \overset{a}{=} \widehat{\sigma^2}_{ML}$$

and the asymptotic posterior variance

$$\text{Var}(\sigma^2|D) \overset{a}{=} \frac{2(\widehat{\sigma^2}_{ML})^2}{n}$$

Thus we recover the MLE of σ^2 and its asymptotic variance.

Other equivalent update rules

Above the update rule from prior to posterior IW distribution is stated for the mean parametrisation:

- $\kappa_0 \to \kappa_1 = \kappa_0 + n$

- $\sigma_0^2 \rightarrow \sigma_1^2 = \lambda\sigma_0^2 + (1-\lambda)\widehat{\sigma^2}_{ML}$ with $\lambda = \frac{\kappa_0}{\kappa_1}$

This has the advantage that the mean of the IW distribution is updated directly, and that the prior and posterior variance is also straightforward to compute.

The same update rule can also be expressed in terms of the other parametrisations. In terms of the conventional parameters α and β of the IG distribution the update rule is

- $\alpha_0 \rightarrow \alpha_1 = \alpha_0 + \frac{n}{2}$
- $\beta_0 \rightarrow \beta_1 = \beta_0 + \frac{n}{2}\widehat{\sigma^2}_{ML} = \beta_0 + \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2$

For the parameters ψ and k of the univariate inverse Wishart distribution the update rule is

- $k_0 \rightarrow k_1 = k_0 + n$
- $\psi_0 \rightarrow \psi_1 = \psi_0 + n\widehat{\sigma^2}_{ML} = \psi_0 + \sum_{i=1}^{n}(x_i - \mu)^2$

For the parameters τ^2 and ν of the scaled inverse chi-squared distribution the update rule is

- $\nu_0 \rightarrow \nu_1 = \nu_0 + n$
- $\tau_0^2 \rightarrow \tau_1^2 = \frac{\nu_0}{\nu_1}\tau_0^2 + \frac{n}{\nu_1}\widehat{\sigma^2}_{ML}$

(See Worksheet B2 for proof of the equivalence of all the above update rules.)

16.5 Estimating the precision using the Wishart-normal model

MLE of the precision

Instead of estimating the variance σ^2 we may wish to estimate the precision $w = 1/\sigma^2$, i.e. the inverse of the variance.

As above the data-generating model is a normal distribution

$$x \sim N(\mu, 1/w)$$

with unknown precision w and known mean μ. This yields as maximum likelihood estimate (easily derived thanks to the invariance principle!)

$$\hat{w}_{ML} = \frac{1}{\widehat{\sigma^2}_{ML}} = \frac{1}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2}$$

Crucially, the MLE of the precision w is not an average itself (instead, it is a function of an average). Consequently, as shown below, the posterior mean of w cannot be expressed as a linear adjustment of the MLE.

Wishart prior

For modelling the variance we have used an IW resp. IG distribution for the prior and posterior distributions. Thus, in order to model the precision we therefore now use a univariate Wishart resp. gamma distribution. Recall that hese two distributions are identical apart from their parametrisation.

Specifically, we use the Wishart distribution in the mean parametrisation (see Probability and Distribution Refresher notes for details):

$$w \sim \text{Wis}\left(s^2 = \frac{w_0}{k_0}, k = k_0\right)$$

- The prior concentration parameter is set to k_0
- The prior mean parameter is set to w_0

The corresponding prior mean of the precision is

$$\text{E}(w) = w_0$$

and the prior variance is

$$\text{Var}(w) = \frac{2w_0^2}{k_0}$$

Wishart posterior

After observing $D = \{x_1 \ldots, x_n\}$ the posterior distribution is also Wishart with updated parameters:

$$w|D \sim \text{Wis}\left(s^2 = \frac{w_1}{k_1}, k = k_1\right)$$

- The posterior concentration parameter is updated to $k_1 = k_0 + n$

- The posterior mean parameter update follows the rule:

$$\frac{1}{w_1} = \lambda \frac{1}{w_0} + (1 - \lambda) \frac{1}{\hat{w}_{ML}}$$

  with $\lambda = \frac{k_0}{k_1}$. Crucially, the linear update is applied to the inverse of the precision but **not** to the precision itself. This is because the MLE of the precision parameter cannot be expressed as an average.

The posterior mean is

$$\mathrm{E}(w|D) = w_1$$

and the posterior variance

$$\mathrm{Var}(w|D) = \frac{2w_1^2}{k_1}$$

Equivalent update rules directly for the parameters of a corresponding gamma distribution $\mathrm{Gam}(\alpha = k/2, \beta = 1/(2s^2))$ with mean $\alpha/\beta = ks^2$ are given as follows. The shape parameter α is updated according to

$$\alpha_1 = \alpha_0 + \frac{n}{2}$$

and the rate parameter β according to

$$\beta_1 = \beta_0 + \frac{n}{2} \widehat{\sigma^2}_{ML}$$

This is the form you will find most often in textbooks. While elegant in terms of parameter updates for the gamma prior it obscures the fact the mean of the precision is not linearly updated.

17 Bayesian model comparison

17.1 Marginal likelihood as model likelihood

Simple and composite models

In the introduction of the Bayesian learning we already encountered the marginal likelihood $p(D|M)$ of a model class M in the denominator of Bayes' rule:

$$p(\boldsymbol{\theta}|D, M) = \frac{p(\boldsymbol{\theta}|M)p(D|\boldsymbol{\theta}, M)}{p(D|M)}$$

Computing this marginal likelihood is different for simple and composite models.

A model is called "simple" if it directly corresponds to a specific distribution, say, a normal distribution with fixed mean and variance, or a binomial distribution with a given probability for the two classes. Thus, a simple model is a point in the model space described by the parameters of a distribution family (e.g. μ and σ^2 for the normal family $N(\mu, \sigma^2)$. For a simple model M the density $p(D|M)$ corresponds to standard likelihood of M and there are no free parameters.

On the other hand, a model is "composite" if it is composed of simple models. This can be a finite set, or it can be comprised of infinite number of simpple models. Thus a composite model represent a model class. For example, a normal distribution with a given mean but unspecified variance, or a binomial model with unspecified class probability, is a composite model.

If M is a composite model, with the underlying simple models indexed by a parameter $\boldsymbol{\theta}$, the likelihood of the model is obtained by marginalisation over $\boldsymbol{\theta}$:

$$\begin{aligned} p(D|M) &= \int_{\boldsymbol{\theta}} p(D|\boldsymbol{\theta}, M)p(\boldsymbol{\theta}|M)d\boldsymbol{\theta} \\ &= \int_{\boldsymbol{\theta}} p(D, \boldsymbol{\theta}|M)d\boldsymbol{\theta} \end{aligned}$$

i.e. we *integrate* over all parameter values $\boldsymbol{\theta}$.

If the distribution over the parameter $\boldsymbol{\theta}$ of a model is strongly concentrated around a specific value $\boldsymbol{\theta}_0$ then the composite model degenerates to a simple point model, and the marginal likelihood becomes the likelihood of the parameter $\boldsymbol{\theta}_0$ under that model.

Example 17.1. Beta-binomial distribution:

Assume that likelihood is binomial with mean parameter θ. If θ follows a Beta distribution then the marginal likelihood with θ integrated out is the beta-binomial distribution (see also Worksheet B2). This is an example of a compound probability distribution.

Log-marginal likelihood as penalised maximum log-likelihood

By rearranging Bayes' rule we see that

$$\log p(D|M) = \log p(D|\boldsymbol{\theta}, M) - \log \frac{p(\boldsymbol{\theta}|D, M)}{p(\boldsymbol{\theta}|M)}$$

The above is valid for all $\boldsymbol{\theta}$.

Assuming concentration of the posterior around the MLE $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$ we will have $p(\hat{\boldsymbol{\theta}}_{\mathrm{ML}}|D, M) > p(\hat{\boldsymbol{\theta}}_{\mathrm{ML}}|M)$ and thus

$$\log p(D|M) = \underbrace{\log p(D|\hat{\boldsymbol{\theta}}_{\mathrm{ML}}, M)}_{\text{maximum log-likelihood}} \underbrace{- \log \frac{p(\hat{\boldsymbol{\theta}}_{\mathrm{ML}}|D, M)}{p(\hat{\boldsymbol{\theta}}_{\mathrm{ML}}|M)}}_{\text{penalty} > 0}$$

Therefore, the log-marginal likelihood is essentially a penalised version of the maximum log-likelihood, and the penalty depends on the concentration of the posterior around the MLE.

Model complexity and Occams razor

Intriguingly, the penality implicit in the log-marginal likelihood is linked to the complexity of the model, in particular to the number of parameters of M. We will see this directly in the Schwarz approximation of the log-marginal likelihood discussed below.

Thus, the averaging over $\boldsymbol{\theta}$ in the marginal likelihood has the effect of automatically penalising complex models. Therefore, when comparing models using the marginal likelihood a complex model may be ranked below simpler models. In contrast, when selecting a model by comparing maximum likelihood directly the model with the highest number of parameters always wins over simpler models. Hence, the penalisation implicit in the marginal likelihood prevents overfitting that occurs with maximum likelihood.

The principle of preferring a less complex model is called **Occam's razor** or the **law of parsimony**.

When choosing models a simpler model is often preferable over a more complex model, because the simpler model is typically better suited to both explaining the currently observed data as well as future data, whereas a complex model will typically only excel in fitting the current data but will perform poorly in prediction.

17.2 The Bayes factor for comparing two models

Definition of the Bayes factor

The **Bayes factor** is the ratio of the likelihoods of the two models:

$$B_{12} = \frac{p(D|M_1)}{p(D|M_2)}$$

The **log-Bayes factor** $\log B_{12}$ is also called the **weight of evidence** for M_1 over M_2.

Bayes theorem in terms of the Bayes factor

We would like to compare two models M_1 and M_2. Before seeing data D we can check their **Prior odds** (= ratio of prior probabilities of the models M_1 and M_2):

$$\frac{\Pr(M_1)}{\Pr(M_2)}$$

After seeing data $D = \{x_1, \ldots, x_n\}$ we arrive at the **Posterior odds** (= ratio of posterior probabilities):

$$\frac{\Pr(M_1|D)}{\Pr(M_2|D)}$$

Using Bayes Theorem $\Pr(M_i|D) = \Pr(M_i)\frac{p(D|M_i)}{p(D)}$ we can rewrite the posterior odds as

$$\underbrace{\frac{\Pr(M_1|D)}{\Pr(M_2|D)}}_{\text{posterior odds}} = \underbrace{\frac{p(D|M_1)}{p(D|M_2)}}_{\text{Bayes factor } B_{12}} \underbrace{\frac{\Pr(M_1)}{\Pr(M_2)}}_{\text{prior odds}}$$

The **Bayes factor** is the multiplicative factor that updates the prior odds to the posterior odds.

On the log scale we see that

$$\text{log-posterior odds} = \text{weight of evidence} + \text{log-prior odds}$$

Interpretive scales for the Bayes factor

Table 17.1: Scale for the Bayes factor according to Jeffreys (1961).

B_{12}	$\log B_{12}$	evidence in favour of M_1 versus M_2
> 100	> 4.6	decisive
10 to 100	2.3 to 4.6	strong
3.2 to 10	1.16 to 2.3	substantial
1 to 3.2	0 to 1.16	not worth more than a bare mention

Following Harold Jeffreys (1961) [1] one may interpret the strength of the Bayes factor as listed in Table 17.1.

[1] Jeffreys, H. *Theory of Probability*. 3rd ed. Oxford University Press.

Table 17.2: Scale for the Bayes factor according to Kass and Raftery (1995).

B_{12}	$\log B_{12}$	evidence in favour of M_1 versus M_2
> 150	> 5	very strong
20 to 150	3 to 5	strong
3 to 20	1 to 3	positive
1 to 3	0 to 1	not worth more than a bare mention

More recently, Kass and Raftery (1995) [2] proposed to use the a slightly modified scale (Table 17.2).

Bayes factor versus likelihood ratio

If both M_1 and M_2 are simple models then the **Bayes factor is identical to the likelihood ratio** of the two models.

However, if one of the two models is composite then the Bayes factor and the generalised likelihood ratio differ: In the Bayes factor the representative of a composite model is the **model average** of the simple models indexed by $\boldsymbol{\theta}$, with weights taken from the prior distribution over the simple models contained in M. In contrast, in the generalised likelihood ratio statistic the representative of a composite model is chosen by *maximisation*.

Thus, **for composite models**, the **Bayes factor does *not* equal the corresponding generalised likelihood ratio statistic.** In fact, the key difference is that the Bayes factor is a penalised version of the likelihood ratio, with the penality depending on the difference in complexity (number of parameters) of the two models

17.3 Approximate computations

The marginal likelihood and the Bayes factor can be difficult to compute in practice. Therefore, a number of approximations have been developed. The most important is the so-called Schwarz (1978) approximation of the log-marginal likelihood. It is used to approximate the log-Bayes factor and also yields the BIC (Bayesian information criterion) which can be interpreted as penalised maximum likelihood.

[2] Kass, R.E., and A.E. Raftery. 1995. *Bayes factors*. JASA **90**:773–795. https://doi.org/10.1080/01621459.1995.10476572

Schwarz (1978) approximation of log-marginal likelihood

The logarithm of the marginal likelihood of a model can be approximated following Schwarz (1978) [3] as follows:

$$\log p(D|M) \approx \ell_n^M(\hat{\boldsymbol{\theta}}_{ML}^M) - \frac{1}{2} d_M \log n$$

where d_M is the dimension of the model M (number of parameters in $\boldsymbol{\theta}$ belonging to M) and n is the sample size and $\hat{\boldsymbol{\theta}}_{ML}^M$ is the MLE. For a simple model $d_M = 0$ so then there is no approximation as in this case the marginal likelihood equals the likelihood.

The above formula can be obtained by quadratic approximation of the likelihood **assuming large** n and assuming that the prior is locally uniform around the MLE. The Schwarz (1978) approximation is therefore a special case of a Laplace approximation.

Note that the approximation is the maximum log-likelihood minus a penalty that depends on the model complexity (as measured by dimension d), hence this is an example of penalised ML! Also note that the distribution over the parameter $\boldsymbol{\theta}$ is not required in the approximation.

Bayesian information criterion (BIC)

The BIC (Bayesian information criterion) of the model M is the approximated log-marginal likelihood times the factor -2:

$$BIC(M) = -2\ell_n^M(\hat{\boldsymbol{\theta}}_{ML}^M) + d_M \log n$$

Thus, when comparing models one aimes to maximise the marginal likelihood or, as approximation, minimise the BIC.

The reason for the factor "-2" is simply to have a quantity that is on the same scale as the Wilks log likelihood ratio. Some people / software packages also use the factor "2".

[3] Schwarz, G. 1978. *Estimating the dimension of a model*. Ann. Statist. **6**:461–464. https://doi.org/10.1214/aos/1176344136

Approximating the weight of evidence (log-Bayes factor) with BIC

Using BIC (twice) the log-Bayes factor can be approximated as

$$\begin{aligned} 2\log B_{12} &\approx -BIC(M_1) + BIC(M_2) \\ &= 2\left(\ell_n^{M_1}(\hat{\boldsymbol{\theta}}_{ML}^{M_1}) - \ell_n^{M_2}(\hat{\boldsymbol{\theta}}_{ML}^{M_2})\right) - \log(n)(d_{M_1} - d_{M_2}) \end{aligned}$$

i.e. it is the penalised log-likelihood ratio of model M_1 vs. M_2.

17.4 Bayesian testing using false discovery rates

We introduce False Discovery Rates (FDR) as a Bayesian method to distinguish a null model from an alternative model. This is closely linked with classical frequentist multiple testing procedures.

Setup for testing a null model H_0 versus an alternative model H_A

We consider two models:

H_0 : null model, with density $f_0(x)$ and distribution $F_0(x)$

H_A : alternative model, with density $f_A(x)$ and distribution $F_A(x)$

Aim: given observations $x_1, \ldots, x_n$ we would like to decide for each x_i whether it belongs to H_0 or H_A.

This is done by a critical decision threshold x_c: if $x_i > x_c$ then x_i is called "significant" and otherwise called "not significant".

In classical statistics one of the the most widely used approach to find the decision threshold is by computing p-values from the x_i (this uses only the null model but not the alternative model), and then thresholding the p-values a a certain level (say 5%). If n is large then often the test is modified by adjusting the p-values or the threshold (e.g. if Bonferroni correction).

Note that this procedure ignores any information we may have about the alternative model!

Test errors

True and false positives and negatives

For any decision threshold x_c we can distinguish the following errors:

- False positives (FP), "false alarm", type I error: x_i belongs to null but is called "significant"
- False negative (FN), "miss", type II error: x_i belongs to alternative, but is called "not significant"

In addition we have:

- True positives (TP), "hits": belongs to alternative and is called "significant"
- True negatives (TN), "correct rejections": belongs to null and is called "not significant"

Specificity and Sensitivity

From counts of TP, TN, FN, FP we can derive further quantities:

- False Positive Rate FPR, false alarm rate, type I error probability: $FPR = \alpha_I = \frac{FP}{TN+FP} = 1 - TNR$
- False Negative Rate FNR, miss rate, type II error probability: $FNR = \alpha_{II} = \frac{FN}{TP+FN} = 1 - TPR$
- True Negative Rate TNR, **specificity**: $TNR = \frac{TN}{TN+FP} = 1 - FPR = 1 - \alpha_I$
- True Positive Rate TPR, **sensitivity**, **power**, recall: $TPR = \frac{TP}{TP+FN} = 1 - FNR = 1 - \alpha_{II}$
- Accuracy: $ACC = \frac{TP+TN}{TP+TN+FP+FN}$

Another common way to choose the decision threshold x_d in classical statistics is to balance sensitivity/power vs. specificity (maximising both power and specificity, or equivalently, minimising both false positive and false negative rates). ROC curves plot TPR/sensitivity vs. FPR = 1-specificity.

FDR and FNDR

It is possible to link the above with the observed counts of TP, FP, TN, FN:

- False Discovery Rate (FDR): $FDR = \frac{FP}{FP+TP}$
- False Nondiscovery Rate (FNDR): $FNDR = \frac{FN}{TN+FN}$
- Positive predictive value (PPV), True Discovery Rate (TDR), precision: $PPV = \frac{TP}{FP+TP} = 1 - FDR$
- Negative predictive value (NPV): $NPV = \frac{TN}{TN+FN} = 1 - FNDR$

In order to choose the decision threshold it is natural to balance FDR and FDNR (or PPV and NPV), by minimising both FDR and FNDR or maximising both PPV and NPV.

In machine learning it is common to use "precision-recall plots" that plot precision (=PPV, TDR) vs. recall (=power, sensitivity).

Bayesian perspective

Two component mixture model

In the Bayesian perspective the problem of choosing the decision threshold is related to computing the posterior probability

$$\Pr(H_0|x_i),$$

i.e. probability of the null model given the observation x_i, or equivalently computing

$$\Pr(H_A|x_i) = 1 - \Pr(H_0|x_i)$$

the probability of the alternative model given the observation x_i.

This is done by assuming a mixture model

$$f(x) = \pi_0 f_0(x) + (1 - \pi_0) f_A(x)$$

where $\pi_0 = \Pr(H_0)$ is the prior probability of H_0 and. $\pi_A = 1 - \pi_0 = \Pr(H_A)$ the prior probabiltiy of H_A.

Note that the weights π_0 can in fact be estimated from the observations by fitting the mixture distribution to the observations $x_1, \ldots, x_n$ (so it is effectively an empirical Bayes method where the prior is informed by the data).

Local FDR

The posterior probability of the null model given a data point is then given by

$$\Pr(H_0|x_i) = \frac{\pi_0 f_0(x_i)}{f(x_i)} = LFDR(x_i)$$

This quantity is also known as the **local FDR** or **local False Discovery Rate**.

In the given one-sided setup the local FDR is large (close to 1) for small x, and will become close to 0 for large x. A common decision rule is given by thresholding local false discovery rates: if $LFDR(x_i) < 0.1$ the x_i is called significant.

q-values

In correspondence to p-values one can also define tail-area based false discovery rates:

$$Fdr(x_i) = \Pr(H_0|X > x_i) = \frac{\pi_0 F_0(x_i)}{F(x_i)}$$

These are called **q-values**, or simply **False Discovery Rates (FDR)**. Intriguingly, these also have a frequentist interpretation as adjusted p-values (using a Benjamini-Hochberg adjustment procedure).

Software

There are a number of R packages to compute (local) FDR values:

For example:

- locfdr
- qvalue
- fdrtool

and many more.

Using FDR values for screening is especially useful in high-dimensional settings (e.g. when analysing genomic and other high-throughput data).

FDR values have both a Bayesian as well as frequentist interpretation, providing further evidence that good classical statistical methods do have a Bayesian interpretation.

18 Choosing priors in Bayesian analysis

18.1 Choosing a prior

Prior as part of the model

It is **essential in a Bayesian analysis to specify your prior uncertainty about the model parameters**. Note that this is simply **part of the modelling process**! Thus in a Bayesian approach the data analyst needs to be more explicit about all modelling assumptions.

Typically, when choosing a suitable prior distribution we consider the overall form (shape and domain) of the distribution as well as its key characteristics such as the mean and variance. As we have learned the precision (inverse variance) of the prior may often be viewed as implied sample size.

For large sample size n the posterior mean converges to the maximum likelihood estimate (and the posterior distribution to normal distribution centered around the MLE), so for large n we may ignore specifying a prior.

However, for small n it is essential that a prior is specified. In non-Bayesian approaches this prior is still there but it is either implicit (maximum likelihood estimation) or specified via a penality (penalised maximum likelihood estimation).

Some guidelines

So the question remains what are good ways to choose a prior? Two useful ways are:

1. Use a weakly informative prior. This means that you do have an idea (even if only vague) about the suitable values of the parameter of interest, and you use a corresponding prior (for example with moderate variance) to model the uncertainty. This acknowledges that there are no uninformative priors and but also aims that the prior does not dominate the likelihood (i.e. the data). The result is a weakly regularised estimator. Note that it is often desirable that the prior adds information (if only a little) so that it can act as a regulariser.
2. Empirical Bayes methods can often be used to determine one or all of the hyperparameters (i.e. the parameters in the prior) from the observed data. There are several ways to do this, one of them is to tune the shrinkage parameter λ to achieve minimum MSE. We discuss this further below.

Furthermore, there also exist many proposals advocating so-called "uninformative priors" or "objective priors". However, there are no actually unformative priors, since a prior distribution that looks uninformative (i.e. "flat") in one coordinate system can be informative in another — this is a simple consequence of the rule for transformation of probability densities. As a result, often the suggested objective priors are in fact improper, i.e. are not actually probability distributions!

18.2 Default priors or uninformative priors

Objective or for default priors are attempts 1) to automatise specification of a prior and 2) to find uniformative priors.

Jeffreys prior

The most well-known non-informative prior is given by a proposal by Harold Jeffreys (1891–1989) in 1946 [1].

Specifically, this prior is constructed from the expected Fisher information and thus promises automatic construction of objective uninformative priors using the likelihood:

$$p(\boldsymbol{\theta}) \propto \sqrt{\det \boldsymbol{I}^{\text{Fisher}}(\boldsymbol{\theta})}$$

[1] Jeffreys, H. 1946. *An invariant form for the prior probability in estimation problems*. Proc. Roy. Soc. A **186**:453–461. https://doi.org/10.1098/rspa.1946.0056

The reasoning underlying this prior is **invariance against transformation of the coordinate system of the parameters**.

For the Beta-Binomial model the Jeffreys prior corresponds to Beta($\frac{1}{2}, \frac{1}{2}$). Note this is not the uniform distribution but a U-shaped prior.

For the normal-normal model it corresponds to the flat improper prior $p(\mu) = 1$.

For the IG-normal model the Jeffreys prior is the improper prior $p(\sigma^2) = \frac{1}{\sigma^2}$.

This already illustrates the main problem with this type of prior – namely that it often is improper, i.e. the prior distribution is not actually a probability distribution (i.e. the density does not integrate to 1).

Another issue is that Jeffreys priors are usually not conjugate which complicates the update from the prior to the posterior.

Furthermore, if there are multiple parameters ($\boldsymbol{\theta}$ is a vector) then Jeffreys priors do not usually lead to sensible priors.

Reference priors

An alternative to Jeffreys priors are the so-called **reference priors** developed by Bernardo (1979) [2]. This type of priors aims to choose the prior such that there is maximal "correlation" between the data and the parameter. More precisely, the mutual information between θ and x is maximised (i.e. the the expected KL divergence between the posterior and prior distribution). The underlying motivation is that the data and parameters should be maximally linked (thereby minimising the influence of the prior).

For univariate settings the reference priors are identical to Jeffreys priors. However, reference prior also provide reasonable priors in multivariate settings.

In both Jeffreys' and the reference prior approach the choice of prior is by expectation over the data, i.e. not for the specific data set at hand (this can be seen both as a positive and negative!).

[2] Bernardo, J. M. 1979. *Reference posterior distributions for Bayesian inference (with discussion).* JRSS B **41**:113–147. https://doi.org/10.1111/j.2517-6161.1979.tb01066.x

18.3 Empirical Bayes

In empirical Bayes the data analysist specifies a family of prior distribution (say a beta distribution with two parameters), and then the data at hand are used to find an optimal choise for the hyper-parameters (hence the name "empirical"). Thus the hyper-parameters are not specified but themselves estimated.

Type II maximum likelihood

In particular, assuming data D, a likelihood $p(D|\boldsymbol{\theta})$ for some model with parameters $\boldsymbol{\theta}$ as well as a prior $p(\boldsymbol{\theta}|\lambda)$ for $\boldsymbol{\theta}$ with hyper-parameter λ the marginal likelihood now depends on λ:

$$p(D|\lambda) = \int_{\boldsymbol{\theta}} p(D|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\lambda) d\boldsymbol{\theta}$$

We can therefore use maximum (marginal) likelihood find optimal values of λ given the data.

Since maximum-likelihood is used in a second level step (the hyper-parameters) this type of empirical Bayes is also often called "type II maximum likelihood".

Shrinkage estimation using empirical risk minimisation

An alternative (but related) way to estimate hyper-parameters is by minimising the empirical risk.

In the examples for Bayesian estimation that we have considered so far the posterior mean of the parameter of interest was obtained by linear shrinkage

$$\hat{\theta}_{\text{shrink}} = \text{E}(\theta|D) = \lambda\theta_0 + (1 - \lambda)\hat{\theta}_{\text{ML}}$$

of the MLE $\hat{\theta}_{\text{ML}}$ towards the prior mean θ_0, with shrinkage intensity $\lambda = \frac{k_0}{k_1}$ determined by the ratio of the prior and posterior concentration parameters k_0 and k_1.

The resulting point estimate $\hat{\theta}_{\text{shrink}}$ is called *shrinkage estimate* and is a convex combination of θ_0 and $\hat{\theta}_{\text{ML}}$. The prior mean θ_0 is also called the "target".

The hyperparameter in this setting is k_0 (linked to the precision of the prior) and or equivalently the shrinkage intensity λ.

An optimal value for λ can be obtained by minimising the mean squared error of the estimator $\hat{\theta}_{\text{shrink}}$.

In particular, by construction, the target θ_0 has low or even zero variance but non-vanishing and potentially large bias, whereas the MLE $\hat{\theta}_{\text{ML}}$ will have low or zero bias but a substantial variance. By combinining these two estimators with opposite properties the aim is to achieve a *bias-variance tradeoff* so that the resulting estimator $\hat{\theta}_{\text{shrink}}$ has lower MSE than either θ_0 and $\hat{\theta}_{\text{ML}}$.

Specifically, the aim is to find

$$\lambda^\star = \arg\min_\lambda \text{E}\left((\theta - \hat{\theta}_{\text{shrink}})^2\right)$$

It turns out that this can be minimised without knowing the actual true value of θ and the result for an unbiased $\hat{\theta}_{\text{ML}}$ is

$$\lambda^\star = \frac{\text{Var}(\hat{\theta}_{\text{ML}})}{\text{E}((\hat{\theta}_{\text{ML}} - \theta_0)^2)}$$

Hence, the shrinkage intensity will be small if the variance of the MLE is small and/or if the target and the MLE differ substantially. On the other hand, if the variance of the MLE is large and/or the target is close to the MLE the shrinkage intensity will be large.

Choosing the shrinkage parameter by optimising expected risk (here mean squared error) is also a form empirical Bayes.

Example 18.1. James-Stein estimator:

Empirical risk minimisation to estimate the shrinkage parameter of the normal-normal model for a single observation yields the James-Stein estimator (1955).

Specifically, James and Stein propose the following estimate for the multivariate mean $\boldsymbol{\mu}$ of using a single sample $\boldsymbol{x}$ drawn from the multivariate normal $N_d(\boldsymbol{\mu}, \boldsymbol{I})$:

$$\hat{\boldsymbol{\mu}}_{JS} = \left(1 - \frac{d-2}{||\boldsymbol{x}||^2}\right)\boldsymbol{x}$$

Here, we recognise $\hat{\boldsymbol{\mu}}_{ML} = \boldsymbol{x}$, $\boldsymbol{\mu}_0 = 0$ and shrinkage intensity $\lambda^\star = \frac{d-2}{||\boldsymbol{x}||^2}$.

Efron and Morris (1972) and Lindley and Smith (1972) later generalised the James-Stein estimator to the case of multiple observations $\boldsymbol{x}_1, \ldots \boldsymbol{x}_n$ and target $\boldsymbol{\mu}_0$, yielding an empirical Bayes estimate of μ based on the normal-normal model.

19 Optimality properties and summary

19.1 Bayesian statistics in a nutshell

- Bayesian statistics explicitly models the uncertainty about the parameters of interest by probability
- In the light of new evidence (observed data) the uncertainty is updated, i.e. the prior distribution is combined via Bayes rule with the likelihood to form the posterior distribution
- If the posterior distribution is in same family as the prior → conjugate prior.
- In an exponential family the Bayesian update of the mean is always expressible as linear shrinkage of the MLE.
- For large sample size the posterior mean becomes maximum likelihood estimator and the prior playes no role.
- Conversely, for small sample size if no data is available the posterior stays close the prior.

Advantages

- Adding prior information has regularisation properties. This is very important in more complex models with many parameters, e.g., in the estimation of a covariance matrix (to avoid singularity).
- Improves small-sample accuracy (e.g. MSE)
- Bayesian estimators tend to perform better than MLEs - this is not surprising as they use the observed data plus the extra information available in the prior.
- Bayesian credible intervals are conceptually much more simple than frequentist confidence intervals.

Frequentist properties of Bayesian estimators

A Bayesian point estimator (e.g. the posterior mean) can also be assessed by its frequentist properties.

- First, by construction due to introducing a prior the Bayesian estimator will be biased for finite n even if the MLE is unbiased.
- Second, intriguingly it turns out that the sampling variance of the Bayes point estimator (not to be confused with the posterior variance!) can be smaller than the variance of the MLE. This depends on the choice of the shrinkage parameter λ that also determines the posterior variance.

As a result, Bayesian estimators may have smaller MSE (=squared bias + variance) than the ML estimator for finite n.

In statistical decision theory this is called the theorem of **admissibility of Bayes rules**. It states that under mild conditions every admissible estimation rule (i.e. one that dominates all other estimators with regard to some expected loss, such as the MSE) is in fact a Bayes estimator with some prior.

Unfortunately, this theorem does not tell which prior is needed to achive optimality, however an optimal estimator can often be found by tuning the hyperparameters.

Specifying the prior — problem or advantage?

In Bayesian statistics the data analyst needs to be very explicit about the modelling assumptions:

Model = data-generating process (likelihood) + prior uncertainty (prior distribution)

Note that alternative statistical methods can often be interpreted as Bayesian methods assuming a specific *implicit* prior!

For example, likelihood estimation for the binomial model is equivalent to Bayes estimation using the Beta-Binomial model with a Beta$(0, 0)$ prior (=Haldane prior).
However, when choosing a prior explicitly for this model, interestingly most analysts would rather use a flat prior Beta$(1, 1)$ (=Laplace prior) with implicit sample size $k_0 = 2$ or a transformation-invariant prior Beta$(1/2, 1/2)$ (=Jeffreys prior) with implicit sample size $k_0 = 1$ rather than the Haldane prior!

→ be aware about the implicit priors!!

Better to acknowledge that a prior is being used (even if implicit!)
Being specific about all your assumptions is enforced by the Bayesian approach.

Specifying a prior is thus best understood as an intrinsic part of model specification. It helps to improve inference and it may only be ignored if there is lots of data.

19.2 Optimality of Bayesian inference

The optimality of Bayesian model making use of full model specification (likelihood plus prior) can be shown from a number of different perspectives. Correspondingly, there are many theorems that prove (or at least indicate) this optimality:

1) Richard Cox's theorem: generalising classical logic invariably leads to Bayesian inference.

2) de Finetti's representation theorem: joint distribution of exchangeable observations can always be expressed as weighted mixture over a prior distribution for the parameter of the model. This implies the existence of the prior distribution and the requirement of a Bayesian approach.

3) Frequentist decision theory: all admissible decision rules are Bayes rules!

4) Entropy perspective: The posterior density (a function!) is obtained as a result of optimising an entropy criterion. Bayesian updating may thus be viewed as a *variational optimisation problem*. Specifically, Bayes theorem is the minimal update when new information arrives in form of observations (see below).

Remark: there exist a number of further (often somewhat esoteric) suggestions for propagating uncertainty such as "fuzzy logic", imprecise probabilities, etc. These contradict Bayesian learning and are thus in direct violation of the above theorems.

19.3 Connection with entropy learning

The *Bayesian update rule* is a very general form of learning when the *new information arrives in the form of data*. But actually there is an even more general principle of which the Bayesian update rule is just a special case: the **principle of minimal information update** (e.g. Jaynes 1959, 2003) or **principle of minimum information discrimination (MDI) (Kullback 1959)**.

It can be summarised as follows: **Change your beliefs only as much as necessary to be coherent with new evidence!**

Under this principle of "inertia of beliefs" when new information arrives the uncertainty about a parameter is only minimally adjusted, only as much as needed to account for the new information. To implement this principle KL divergence is a natural measure to quantify the change of the underlying beliefs. This is known as **entropy learning**.

The Bayes rule emerges a special case of entropy learning:

- The KL divergence between the joint posterior $Q_{x,\boldsymbol{\theta}}$ and joint prior distribution $P_{x,\boldsymbol{\theta}}$ is computed, with the posterior distribution $Q_{\boldsymbol{\theta}|x}$ as free parameter.
- The conditional distribution $Q_{\boldsymbol{\theta}|x}$ is found by minimising the KL divergence $D_{\mathrm{KL}}(Q_{x,\boldsymbol{\theta}}, P_{x,\boldsymbol{\theta}})$.
- The optimal solution to this **variational optimisation problem** is given by Bayes' rule!

This application of the KL divergence is an example of **reverse KL optimisation** (aka I-projection, see Part I of the notes). Intriguingly, this explains the zero forcing property of Bayes' rule (because that this is a general property of an I-projection).

Applying entropy learning therefore includes Bayesian learning as special case:

1) If information arrives in form of data → update prior by Bayes' theorem (Bayesian learning).

Interestingly, entropy learning will lead to other update rules for other types of information:

2) If information arrives in the form of another distribution → update using R. Jeffrey's rule of conditioning (1965).

3) If the information is presented in the form of constraints → Kullback's principle of minimum MDI (1959), E. T. Jaynes maximum entropy (MaxEnt) principle (1957).

This shows (again) how fundamentally important KL divergence is in statistics. It not only leads to likelihood inference (via forward KL) but also to Bayesian learning, as well as to other forms of information updating (via reverse KL).

Furthermore, in Bayesian statistics the KL divergence is useful to choose priors (e.g. reference priors) and it also helps in (Bayesian) experimental design to quantify the information provided by an experiment.

19.4 Conclusion

Bayesian statistics offers a coherent framework for statistical learning from data, with methods for

- estimation
- testing
- model building

There are a number of theorems that show that "optimal" estimators (defined in various ways) are all Bayesian.

It is conceptually very simple — but can be computationally very involved!

It provides a coherent generalisation of classical TRUE/FALSE logic (and therefore does not suffer from some of the inconsistencies prevalent in frequentist statistics).

Bayesian statistics is a non-asymptotic theory, it works for any sample size. Asympotically (large n) it is consistent and converges to the true model (like ML!). But Bayesian reasoning can also be applied to events that take place only once — no assumption of hypothetical infinitely many repetitions as in frequentist statistics is needed.

Moreover, many classical (frequentist) procedures may be viewed as *approximations* to Bayesian methods and estimators, so using classical approaches in the correct application domain is perfectly in line with the Bayesian framework.

Bayesian estimation and inference also automatically regularises (via the prior) which is important for complex models and when there is the problem of overfitting.

Bibliography

Agresti, A., and M. Kateri. 2022. *Foundations of Statistics for Data Scientists*. Chapman; Hall/CRC.

Diaconis, P., and B. Skyrms. 2018. *Ten Great Ideas about Chance*. Princeton University Press.

Domingos, P. 2015. *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*. Basic Books.

Gelman, A., J. B. Carlin, H. A. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. 2014. *Bayesian Data Analysis*. 3rd ed. CRC Press.

Heard, N. 2021. *An Introduction to Bayesian Inference, Methods and Computation*. Springer.

Held, L., and D. S. Bové. 2020. *Applied Statistical Inference: Likelihood and Bayes*. Second. Springer.

Jaynes, E. T. 2003. *Probability Theory: The Logic of Science*. Cambridge University Press.

McGrayne, S. B. 2011. *The Theory That Would Not Die*. Yale University Press.

A Statistics refresher

Below you find a brief overview over some relevant concepts in statistics that you should be familiar with from earlier modules.

A.1 Data and statistics as functions of data

Broadly, by "data" we refer to quantitative observations and measurements collected in experiments.

We denote the observed data by $D = \{x_1, \ldots, x_n\}$ where n denotes the number of data points (the **sample size**). Each data point can be scalar or a multivariate quantity.

Generally, a **statistic** $t(D)$ is function of the observed data D. The statistic $t(D)$ can be of any type and value (scalar, vector, matrix etc. — even a function). $t(D)$ is called a *summary statistic* if it describes important aspects of the data such as location (e.g. the average $\text{avg}(D) = \bar{x}$, the median) or scale (e.g. standard deviation, interquartile range).

A.2 Statistical learning

The aim in statistics, and by extension in data science and machine learning, is to use data to learn about and better understand the world. It is a key feature of statistics to employ probabilistic models for that purpose.

Let denote data models by $p(x|\theta)$ where θ represents the parameters of the model. Often (but not always) θ can be interpreted as or is associated with some manifest property of the model. If there is only a single parameter we write θ (scalar parameter). If we wish to highlight that there are multiple parameters we write $\boldsymbol{\theta}$ (in bold type).

Specifically, our aim is to identify the best model(s) for the data in order to both

- explain the current data, and
- to enable good prediction of future data.

By choosing a sufficiently complex model the first aim (to explain the observed data) is often easily achieved. However, if the model is too complex it can fail to address the second aim (to predict unseen data well). Thus, when choosing a model we would like to avoid both the problem of **underfitting** (i.e. choosing an overly simplistic model) as well **overfitting** (i.e. choosing an overly complex model). Finding this balances will also help with interpreting the fitted model.

Typically, we focus the analysis to a specific model family with a some parameter θ.
An **estimator for** θ is a function $\hat{\theta}(D)$ of the data that maps the data (input) to an informed guess (output) about θ.

- A **point estimator** provides a single number for each parameter
- An **interval estimator** provides a set of possible values for each parameter.

Interval estimators can be linked to the concept of testing specified values for a parameter. Specifically a confidence interval contains all parameter values that are not significantly different from the best parameter.

A.3 Sampling properties of a point estimator

A point estimator $\hat{\theta}$ depends on the data, hence it exhibits **sampling variation**, i.e. estimate will be different for a new set of observations.

Thus $\hat{\theta}$ can be seen as a random variable, and its distribution is called **sampling distribution** (across different experiments).

Properties of this distribution can be used to evaluate how far the estimator deviates (on average across different experiments) from the true value:

$$\begin{aligned}
\text{Bias:} \quad & \text{Bias}(\hat{\theta}) && = \text{E}(\hat{\theta}) - \theta \\
\text{Variance:} \quad & \text{Var}(\hat{\theta}) && = \text{E}\left((\hat{\theta} - \text{E}(\hat{\theta}))^2\right) \\
\text{Mean squared error:} \quad & \text{MSE}(\hat{\theta}) && = \text{E}((\hat{\theta} - \theta)^2) \\
& && = \text{Var}(\hat{\theta}) + \text{Bias}(\hat{\theta})^2
\end{aligned}$$

The last identity about MSE follows from $\text{E}(x^2) = \text{Var}(x) + \text{E}(x)^2$.

At first sight it seems desirable to focus on unbiased (for finite sample size n) estimators. However, requiring strict unbiasedness is not always a good idea. In many situations it is better to accept some bias in an estimator in order to achieve a smaller variance and an overall smaller MSE. This is called **bias-variance tradeoff** — as more bias is traded for smaller variance (or, conversely, less bias is traded for higher variance). This is also related to the above mentioned problems of underfitting (large bias) and overfitting (large variance).

A.4 Efficiency and consistency of an estimator

Typically, Bias, Var and MSE all decrease with increasing sample size so that with more data $n \to \infty$ the errors become smaller and smaller.

Efficiency: An estimator $\hat{\theta}_A$ is said to more efficient than estimator $\hat{\theta}_B$ if for same sample size n it has smaller error (e.g. MSE) than the competing estimator.

The typical rate of decrease in variance of a good estimator is $\frac{1}{n}$ and the rate of decrease in the standard deviation is $\frac{1}{\sqrt{n}}$. Note that this implies that to get one digit more accuracy in an estimate (standard deviation decreasing by factor of 10) we need 100 times more data!

Consistency: $\hat{\theta}$ is called consistent if

$$\text{MSE}(\hat{\theta}) \longrightarrow 0 \text{ with } n \to \infty$$

Consistency is an essential yet relatively weak requirement for any reasonable estimator. Among all consistent estimators we generally choose those that are most **efficient**, meaning that they exhibit the smallest variance and/or MSE for a given finite n.

Consistency implies that, given an infinite amount of data, the true model can be accurately identified, provided that the model class includes the actual data-generating model. If the model class does not encompass the true model, strict consistency cannot be attained. Nevertheless, our goal remains to choose a model that is closest to the true model and approximates it as best as possible.

A.5 Law of large numbers

The **law of large numbers**, discovered by Jacob Bernoulli (1655-1705), asserts that, if the mean exists, the sample average will converge to the mean as the sample size n becomes large. Therefore, when the mean is defined, it can be approximated by the empirical mean for sufficiently large values of n.

A variant of the law of large numbers is that the empirical distribution $\hat{F}_n$ convergences strongly to F (Section A.6).

As a result, with $n \to \infty$ there's also convergence of the average of a function of the observed samples to the corresponding expectation of the function of the random variable:

$$\mathrm{E}_{\hat{F}_n}(h(x)) = \frac{1}{n}\sum_{i=1}^{n} h(x_i) \to \mathrm{E}_F(h(x))$$

Hence, the law of large numbers also ensures that empirical estimators (Section A.7) will converge to the corresponding true values for sufficiently large n.

Moreover, the law of large numbers provides a **justification for interpreting large-sample limits of frequencies as probabilities**. However, **the converse** assumption that all probabilities can be interpreted in frequentist manner **does not follow** from the law of large numbers or from the axioms of probability.

Finally, it is worth pointing out that the law of large number doesn't say anything about the finite sample properties of an estimator, it is only concerned with the asymptotic domain (large n).

A.6 Empirical distribution function

Suppose we observe data $D = \{x_1, \ldots, x_n\}$ with each $x_i \sim F$ sampled independently and identically. The empirical cumulative distribution function $\hat{F}_n(x)$ based on data D is then given by

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n}[x_i \leq x]$$

where $[A]$ is the indicator function in Iverson bracket notation which equals 1 if A is true and 0 otherwise. Thus $\hat{F}_n(x)$ counts how many

observations $x_i \in D$ are smaller or equal than x and then standardises by the total number of samples n.

The empirical distribution function is monotonically non-decreasing from 0 to 1 in discrete steps.

In R the empirical distribution function is computed by `ecdf()`.

Crucially, the empirical distribution $\hat{F}_n$ converges strongly (almost surely) to the underlying distribution F as $n \to \infty$:

$$\hat{F}_n \overset{a.s.}{\to} F$$

The Glivenko-Cantelli theorem additionally asserts that the convergence is uniform.

This theorem is a variant of the **law of large numbers** (Section A.5) applied to the whole distribution, rather than just to the mean.

As a result, we may use the empirical distribution $\hat{F}_n$ based on data D as an estimate of the underlying unknown true distribution F. From the convergence theorems we know that $\hat{F}_n$ is consistent.

However, for $\hat{F}_n$ to work well as an estimate of F the number of observations n must be sufficiently large so that the approximation provided by $\hat{F}_n$ is adequate.

A.7 Empirical estimators

The fact that for large sample size n the empirical distribution $\hat{F}_n$ may be used as a substitute for the unknown F allows us to easily construct empirical estimators.

Specifically, parameters of a model can typically be expressed as a functional of the distribution $\theta = g(F)$. An **empirical estimator** $\hat{\theta}$ is constructed by substituting the true distribution by the empirical distribution $\hat{\theta} = g(\hat{F}_n)$.

An example is the mean $\mathrm{E}_F(x)$ with regard to F. The **empirical mean** is the expectation with regard to the empirical distribution which equals the **average of the samples**:

$$\hat{\mathrm{E}}(x) = \hat{\mu} = \mathrm{E}_{\hat{F}_n}(x) = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$$

Similarly, other empirical estimators can be constructed simply by replacing the expectation in the definition of the quantity of interest by the sample average. For example, the **empirical variance** with unknown mean is given by

$$\widehat{\mathrm{Var}}(x) = \widehat{\sigma^2} = \mathrm{E}_{\hat{F}_n}((x - \hat{\mu})^2) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Note the factor $1/n$ before the summation sign. We can also write the empirical variance in terms of $\overline{x^2} = \frac{1}{n}\sum_{k=1}^{n} x^2$ as

$$\widehat{\mathrm{Var}}(x) = \overline{x^2} - \bar{x}^2$$

By construction, as a result of the strong convergence of $\hat{F}_n$ to F empirical estimators are consistent, with their MSE, variance and bias all decreasing to zero with large sample size n. However, for finite sample size they do have a finite variance and may also be biased.

For example, the empirical variance given above is biased with $\mathrm{Bias}(\widehat{\sigma^2}) = -\sigma^2/n$. Note this bias decreases with n. An unbiased estimator can be obtained by rescaling the empirical estimator by the factor $n/(n-1)$:

$$\widehat{\sigma^2}_{\mathrm{UB}} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

The empirical estimators for the mean and variance can also be obtained for random vectors $\boldsymbol{x}$. In this case the data $D = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ is comprised of n vector-valued observations.

For the mean get

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{k=1}^{n}\boldsymbol{x}_k = \bar{\boldsymbol{x}}$$

and for the covariance

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{k=1}^{n}(\boldsymbol{x}_k - \bar{\boldsymbol{x}})\ (\boldsymbol{x}_k - \bar{\boldsymbol{x}})^T$$

Note the factor $\frac{1}{n}$ in the estimator of the covariance matrix.

With $\overline{\boldsymbol{x}\boldsymbol{x}^T} = \frac{1}{n}\sum_{k=1}^{n}\boldsymbol{x}_k\boldsymbol{x}_k^T$ we can also write

$$\widehat{\boldsymbol{\Sigma}} = \overline{\boldsymbol{x}\boldsymbol{x}^T} - \bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^T$$

A.8 Sampling distribution of mean and variance estimators for normal data

If the underlying distribution family of $D = \{x_1, \ldots, x_n\}$ is known we can often obtain the exact distribution of an estimator.

For example, assuming normal distribution $x_i \sim N(\mu, \sigma^2)$ we can derive the sampling distribution for the empirical mean and variance:

- The empirical estimator of the mean parameter μ is given by $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Under the normal assumption the distribution of $\hat{\mu}$ is

  $$\hat{\mu} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

  Thus $\mathrm{E}(\hat{\mu}) = \mu$ and $\mathrm{Var}(\hat{\mu}) = \frac{\sigma^2}{n}$. The estimate $\hat{\mu}$ is unbiased as $\mathrm{E}(\hat{\mu}) - \mu = 0$. The mean squared error of $\hat{\mu}$ is $\mathrm{MSE}(\hat{\mu}) = \frac{\sigma^2}{n}$.

- The empirical variance $\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ for normal data follows a one-dimensional Wishart distribution

  $$\widehat{\sigma^2} \sim \mathrm{Wis}\left(s^2 = \frac{\sigma^2}{n}, k = n - 1\right)$$

  Thus, $\mathrm{E}(\widehat{\sigma^2}) = \frac{n-1}{n}\sigma^2$ and $\mathrm{Var}(\widehat{\sigma^2}_{\mathrm{ML}}) = \frac{2(n-1)}{n^2}\sigma^4$. The estimate $\widehat{\sigma^2}$ is biased since $\mathrm{E}(\widehat{\sigma^2}_{\mathrm{ML}}) - \sigma^2 = -\frac{1}{n}\sigma^2$. The mean squared error is $\mathrm{MSE}(\widehat{\sigma^2}) = \frac{2(n-1)}{n^2}\sigma^4 + \frac{1}{n^2}\sigma^4 = \frac{2n-1}{n^2}\sigma^4$.

- The unbiased variance estimate $\widehat{\sigma^2}_{\mathrm{UB}} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ for normal data follows a one-dimensional Wishart distribution

  $$\widehat{\sigma^2}_{\mathrm{UB}} \sim \mathrm{Wis}\left(s^2 = \frac{\sigma^2}{n-1}, k = n - 1\right)$$

  Thus, $\mathrm{E}(\widehat{\sigma^2}_{\mathrm{UB}}) = \sigma^2$ and $\mathrm{Var}(\widehat{\sigma^2}_{\mathrm{UB}}) = \frac{2}{n-1}\sigma^4$. The estimate $\widehat{\sigma^2}_{\mathrm{ML}}$ is unbiased since $\mathrm{E}(\widehat{\sigma^2}_{\mathrm{UB}}) - \sigma^2 = 0$. The mean squared error is $\mathrm{MSE}(\widehat{\sigma^2}_{\mathrm{UB}}) = \frac{2}{n-1}\sigma^4$.

  Interestingly, for any $n > 1$ we find that $\mathrm{Var}\left(\widehat{\sigma^2}_{\mathrm{UB}}\right) > \mathrm{Var}\left(\widehat{\sigma^2}_{\mathrm{ML}}\right)$ and $\mathrm{MSE}\left(\widehat{\sigma^2}_{\mathrm{UB}}\right) > \mathrm{MSE}\left(\widehat{\sigma^2}_{\mathrm{ML}}\right)$ so that the biased empirical estimator has both lower variance and lower mean squared error than the unbiased estimator.

A.9 t-statistics

One sample t-statistic

Suppose we observe n independent data points $x_1, \ldots, x_n \sim N(\mu, \sigma^2)$. Then the average $\bar{x} = \sum_{i=1}^{n} x_i$ is distributed as $\bar{x} \sim N(\mu, \sigma^2/n)$ and correspondingly

$$z = \frac{\bar{x} - \mu}{\sqrt{\sigma^2/n}} \sim N(0, 1)$$

Note that z uses the *known variance* σ^2.

If the variance is unknown and is estimated by the *unbiased variance*

$$s^2_{\text{UB}} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

then one arrives at the one sample t-statistic

$$t_{\text{UB}} = \frac{\bar{x} - \mu}{\sqrt{s^2_{\text{UB}}/n}} \sim t_{n-1}\,.$$

It is distributed according to a Student's t-distribution with $n-1$ degrees of freedom, with mean 0 for $n > 2$ and variance $(n-1)/(n-3)$ for $n > 3$.

If instead of the unbiased estimate the *empirical variance* (i.e. the maximum likelihood estimate, ML)

$$s^2_{\text{ML}} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{n-1}{n} s^2_{\text{UB}}$$

is used then this leads to a slightly different statistic

$$t_{\text{ML}} = \frac{\bar{x} - \mu}{\sqrt{s^2_{\text{ML}}/n}} = \sqrt{\frac{n}{n-1}}\, t_{\text{UB}}$$

with

$$t_{\text{ML}} \sim t_{n-1}\left(0, \tau^2 = \frac{n}{n-1}\right)$$

Thus, t_{ML} follows a location-scale t-distribution, with mean 0 for $n > 2$ and variance $n/(n-3)$ for $n > 3$.

Two sample t-statistic with common variance

Now suppose we observe normal data $D = \{x_1, \ldots, x_n\}$ from two groups with sample size n_1 and n_2 (and $n = n_1 + n_2$) with two different means μ_1 and μ_2 and common variance σ^2:

$$x_1, \ldots, x_{n_1} \sim N(\mu_1, \sigma^2)$$

and

$$x_{n_1+1}, \ldots, x_n \sim N(\mu_2, \sigma^2)$$

Then $\hat{\mu}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} x_i$ and $\hat{\mu}_2 = \frac{1}{n_2} \sum_{i=n_1+1}^{n} x_i$ are the sample averages within each group.

The common variance σ^2 may be estimated either by the *unbiased estimate*

$$s^2_{\text{UB}} = \frac{1}{n-2} \left(\sum_{i=1}^{n_1} (x_i - \hat{\mu}_1)^2 + \sum_{i=n_1+1}^{n} (x_i - \hat{\mu}_2)^2 \right)$$

(note the factor $n-2$) or by the *empirical estimate* (ML)

$$s^2_{\text{ML}} = \frac{1}{n} \left(\sum_{i=1}^{n_1} (x_i - \hat{\mu}_1)^2 + \sum_{i=n_1+1}^{n} (x_i - \hat{\mu}_2)^2 \right) = \frac{n-2}{n} s^2_{\text{UB}}$$

The estimator for the common variance is a often referred to as *pooled variance estimate* as information is pooled from two groups to obtain the estimate.

Using the unbiased pooled variance estimate the two sample t-statistic is given by

$$t_{\text{UB}} = \frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) s^2_{\text{UB}}}} = \frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{\left(\frac{n}{n_1 n_2}\right) s^2_{\text{UB}}}}$$

In terms of empirical frequencies $\hat{\pi}_1 = \frac{n_1}{n}$ and $\hat{\pi}_2 = \frac{n_2}{n}$ it can also be written as

$$t_{\text{UB}} = \sqrt{n} \frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{\left(\frac{1}{\hat{\pi}_1} + \frac{1}{\hat{\pi}_2}\right) s^2_{\text{UB}}}} = \sqrt{n \hat{\pi}_1 \hat{\pi}_2} \frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{s^2_{\text{UB}}}}$$

The two sample t-statistic is distributed as

$$t_{\text{UB}} \sim t_{n-2}$$

i.e. according to a Student's t-distribution with $n-2$ degrees of freedom, with mean 0 for $n > 3$ and variance $(n-2)/(n-4)$ for $n > 4$. Large values of the two sample t-statistic indicates that there are indeed two groups rather than just one.

The two sample t-statistic using the empirical (ML) pooled estimate of the variance is

$$\begin{aligned} t_{\text{ML}} &= \frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) s^2_{\text{ML}}}} = \frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{\left(\frac{n}{n_1 n_2}\right) s^2_{\text{ML}}}} \\ &= \sqrt{n} \frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{\left(\frac{1}{\hat{\pi}_1} + \frac{1}{\hat{\pi}_2}\right) s^2_{\text{ML}}}} = \sqrt{n\hat{\pi}_1\hat{\pi}_2} \frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{s^2_{\text{ML}}}} \\ &= \sqrt{\frac{n}{n-2}} t_{\text{UB}} \end{aligned}$$

with

$$t_{\text{ML}} \sim t_{n-2}\left(0, \tau^2 = \frac{n}{n-2}\right)$$

Thus, t_{ML} follows a location-scale t-distribution, with mean 0 for $n > 3$ and variance $n/(n-4)$ for $n > 4$.

A.10 Confidence intervals

General concept

A **confidence** interval (CI) is an **interval estimate** $\widehat{\text{CI}}(x_1, \ldots, x_n)$ that depends on data and has a random (sampling) variation as well as a **frequentist** interpretation.

A key property of a confidence interval is its **coverage probability**

$$\kappa = \Pr(\theta \in \widehat{\text{CI}})$$

which describes how often — in repeated identical experiments — the estimated confidence interval overlaps the true parameter value θ, i.e. how often it will "cover" θ (see Figure A.1). In the above θ is fixed and $\widehat{\text{CI}}$ is random. Note that κ is explicitly **not** the probability that the true value is contained in a specific instance of an estimated confidence interval. Specifically, any particular confidence interval either covers θ or it doesn't.

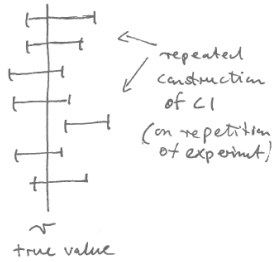


Figure A.1: Coverage property of confidence intervals.

For example, a coverage probability $\kappa = 0.95$ (95%) implies that in 95 out of 100 repeated experiments the corresponding estimated confidence interval will contain the (unknown) true value.

It is trivial to create a confidence with high coverage, simply by assuming a wide interval. Therefore, a useful confidence interval must be both **compact** and have **high coverage**.

Finally, there is also a direct relationship between confidence intervals and statistical testing procedures. Specifically, a confidence interval can be interpreted as the set of parameter values that cannot be rejected. The complement $\alpha = 1 - \kappa$ is called the **rejection probability** or **significance level**.

Symmetric normal confidence interval

For a normally distributed univariate random variable it is straightforward to construct a symmetric two-sided confidence interval with a given desired coverage κ (Figure A.2). The confidence interval corresponds to the central part of the density and contains probability mass $\kappa = 1 - \alpha$ whereas both tails each contain mass $(1-\kappa)/2 = \alpha/2$ and correspond together to the rejection region.

A **symmetric normal confidence interval** with nominal coverage κ for

- a scalar parameter θ
- with normally distributed estimate $\hat{\theta} \sim N(\theta, \sigma^2)$

is given by

$$\widehat{\text{CI}} = [\hat{\theta} \pm c\sigma]$$

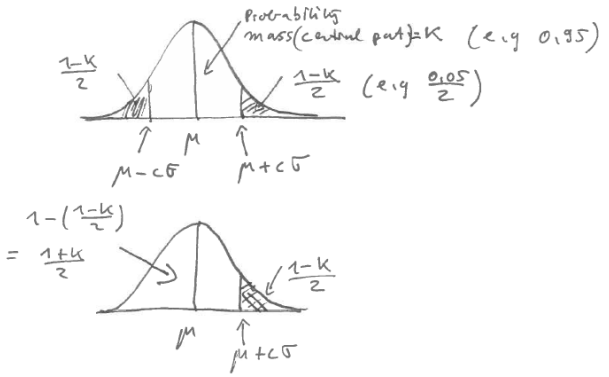


Figure A.2: Construction of a symmetric two-sided normal confidence interval.

where the critical value c is chosen to achieve the desired coverage probability

$$\kappa = \Pr(\hat{\theta} - c\sigma \leq \theta \leq \hat{\theta} + c\sigma)$$

The critical value c is obtained as the $(1 + \kappa)/2 = 1 - \alpha/2$ quantile $z_{(1+\kappa)/2} = z_{1-\alpha/2}$ of the standard normal distribution $N(0, 1)$ so that

$$c = \Phi^{-1}\left(\frac{1+\kappa}{2}\right) = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$$

where where $\Phi(c)$ is the cumulative distribution function (CDF) of the standard normal $N(0, 1)$ distribution. Its inverse Φ^{-1} is the standard normal quantile function.

Table A.1: Critical values for the standard normal distribution.

Coverage κ	Critical value c	Quantile $z_{(1+\kappa)/2}$
0.90	1.6449	$z_{0.95}$
0.95	1.9600	$z_{0.975}$
0.99	2.5758	$z_{0.995}$

Table A.1 lists the critical values c for the three most commonly used values of κ. It is useful to memorise these values as they are used frequently.

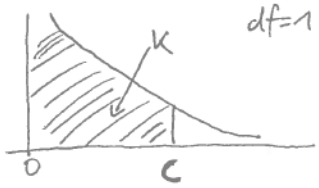


Figure A.3: Construction of a one-sided confidence interval based on a chi-squared distribution with one degree of freedom.

One-sided confidence interval based on the chi-squared distribution

For a chi-squared distributed statistic commonly a one-sided confidence interval of the form $[0, c]$ is used with nominal coverage probability $\kappa = \Pr(x \leq c)$ (see Figure A.3). The right tail contains $1 - \kappa = \alpha$ probability mass.

We obtain the critical value c as the $\kappa = 1 - \alpha$ quantile of the chi-squared distribution by using the quantile function, i.e. by inverting the CDF of the chi-squared distribution.

Table A.2: Critical values for the chi-squared distribution with one degree of freedom.

Coverage κ	Critical value c	Quantile x_κ
0.90	2.7055	$x_{.90}$
0.95	3.8415	$x_{.95}$
0.99	6.6349	$x_{.99}$

Table A.2 lists the critical values for the three most common choices of the coverage probability κ for a chi-squared distribution with one degree of freedom. Note that these critical values are the squared values of the corresponding thresholds in Table A.1 (with small discrepancies due to rounding).

B Further study

In this module we can only touch the surface of likelihood and Bayes inference. As a starting point for further reading the following text books are recommended.

B.1 Recommended reading

- Held and Bové (2020) *Applied Statistical Inference: Likelihood and Bayes (2nd edition)*. Springer.
- Agresti and Kateri (2022) *Foundations of Statistics for Data Scientists*. Chapman and Hall/CRC.

B.2 Additional references

- Heard (2021) *An Introduction to Bayesian Inference, Methods and Computation*. Springer.
- Gelman et al. (2014) *Bayesian data analysis (3rd edition)*. CRC Press.