Statistics 2: Likelihood and Bayes

Korbinian Strimmer

10 December 2023

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–. This version is from 10 December 2023.

The notes will be updated from time to time. To view the current version visit the online MATH27720 Statistics 2 lecture notes in a web browser.

You may also wish to download the MATH27720 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

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.

The notes are for the version of MATH27720 Statistics 2 taught in spring 2024 at the University of Manchester.

I hope that you enjoy the course and that you will find the notes useful! If you have any questions, comments, or corrections please email me at korbinian.strimmer@manchester.ac.uk.

About the module

Topics covered

The MATH27720 Statistics 2 module is designed to run over the course of 11 weeks. It has the following structure:

1. Overview and refresher (W1)
2. Likelihood estimation and inference (W2–W6)
3. Bayesian learning and inference (W7–W11)

This module focuses on conceptual understanding and methods, not on theory. Specifically, you will learn about the foundations of statistical learning using likelihood and Bayesian approaches and also how these are underpinned by entropy.

As such, the presentation in this course is non-technical. The aim is to offer insights how diverse statistical approaches are linked and to demonstrate that statistics offers a concise and coherent theory of information rather than being an adhoc collection of "recipes" for data analysis (a common but wrong perception of statistics).

Prerequisites

For this module it is important that you refresh your knowledge in:

- Introduction to statistics
- Probability
- R data analysis and programming

In addition you will need to some elements of matrix algebra and how to compute the gradient and the curvature of a function of several variables.

Check the Appendix of these notes for a brief refresher of the essential material.

Additional support material

If you are a University of Manchester student and enrolled in this module you will find on Blackboard:

- a weekly learning plan for an 11 week study period (plus one additional week for revision),
- weekly worksheets with examples and solutions and R code, and
- exam papers of previous years.

Furthermore, there is also a MATH27720 Statistics 2 online reading list hosted by the University of Manchester library.

Acknowledgements

These notes are in part based 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.

Part I

Likelihood estimation and inference

Chapter 1

Overview of statistical learning

1.1 How to learn from data?

A fundamental question is how to extract information from data in an optimal way, and to make predictions based on this information.

For this purpose, a number of competing **theories of information** have been developed. **Statistics** is the oldest science of information and is concerned with offering principled ways to learn from data and to extract and process information using probabilistic models. However, there are other theories of information (e.g. Vapnik-Chernov theory of learning, computational learning) that are more algorithmic than analytic and sometimes not even based on probability theory.

Furthermore, there are other disciplines, such computer science and machine learning that are closely linked with and also have substantial overlap with statistics. The field of "data science" today comprises of both statistics and machine learning and brings together mathematics, statistics and computer science. Also the growing field of so-called "artificial intelligence" makes substantial use of statistical and machine learning techniques.

The recent popular science book "The Master Algorithm" by Domingos (2015) provides an accessible informal overview over the various schools of science of information. It discusses the main algorithms used in machine learning and statistics:

- Starting as early as 1763, the **Bayesian school** of learning was started which later turned out to be closely linked with *likelihood inference* established in 1922 by Ronald A. Fisher (1890–1962) and generalised in 1951 to **entropy learning** by Kullback and Leibler.

- It was also in the 1950s that the concept of artificial **neural network** arises, essentially a nonlinear input-output map with no underlying probabilistic modelling. This field saw another leap in the 1980s and further progressed from 2010 onwards with the development of *deep dearning*. It is now one of the most popular (and most effective) methods for analysing imaging data. Even your mobile phone most likely now has a dedicated computer chip with special neural network hardware. Despite their non-probabilistic origins, modern interpretations of neural networks view them as high-dimensional nonlinear statistical models.
- Further advanced theories of information were developed in the 1960 under the term of **computational learning**, most notably 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".
- With the advent of large-scale genomic and other high-dimensional data there has been a surge of new and exciting developments in the field of high-dimensional (large dimension) and also big data (large dimension and large sample size), both in statistics and in machine learning.

The connections between various fields of information is still not perfectly understood, but it is clear that an overarching theory will need to be based on probabilistic learning.

1.2 Probability theory versus statistical learning

When you study statistics (or any other information theory) you need to be aware that there is a fundamental difference between probability theory and statistics, and that relates to the **distinction between "randomness" and "uncertainty"**.

Probability theory studies **randomness**, by developing mathematical models for randomness (such as probability distributions), and studying corresponding mathematical properties (including asymptotics etc). Probability theory may in fact be viewed as a branch of measure theory, and as such it belongs to the domain of pure mathematics.

Probability theory provides probabilistic generative models for data, both for simulation of new data and as well as for explaining observed data. Methods and theory how to best identify probabilistic models from observations and to use them to predict future observations belongst to the domain of applied mathematics, specifically statistics and the related areas of machine learning and data science.

Note that statistics, in contrast to probability, is therefore not at all concerned with randomness. Instead, the focus is about measuring and elucidating the **uncertainty** of events, predictions, outcomes, model parameters and this

uncertainty measures the **state of knowledge** about these quantities. As soon as new data or information becomes available, the state of knowledge and thus the uncertainty changes! Thus, **uncertainty is an epistemological property**.

The uncertainty most often is due to our ignorance of the true underlying processes (on purpose or not), but not because the underlying process is actually random. The success of statistics is based on the fact that we can mathematically model the uncertainty without knowing any detailed specifics of the underlying processes, and yet we still can furnish procedures for optimal inference despite the uncertainty.

In short, statistics is about describing the state of knowledge of the world, which may be uncertain and incomplete, and to make decisions and predictions in the face of uncertainty, and this uncertaintly sometimes derives from randomness but most often from our ignorance (and sometimes this ignorance even helps to create a simple yet effective model).

1.3 Cartoon 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 overfitting (not enough data for the complexity) but also underfitting (model is too simplistic).

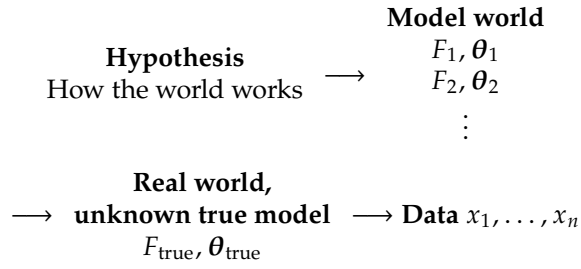
We observe data $D = \{x_1, \ldots, x_n\}$ assumed to result from an underlying true data-generating model F_{true}, the distribution for x.

To explain the observed data, and also to predict future data, we will make hypotheses in the form of candidate models $F_1, F_2, \ldots$. Often these candidate models form a model family $F_{\boldsymbol{\theta}}$ indexed by a parameter vector $\boldsymbol{\theta}$, with specific values for each model so that we can also write $F_{\boldsymbol{\theta}_1}, F_{\boldsymbol{\theta}_2}, \ldots$ for the various models. These parameters may, or may not, have some direct interpretation. Ideally, the models are also identifiable within a family, i.e. each distinct model is identified by a unique parameter so that $F_{\boldsymbol{\theta}_1} = F_{\boldsymbol{\theta}_2}$ implies $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$, hence if models are the same then the corresponding parameters must also be the same.

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

The various candidate models $F_1, F_2, \ldots$ in the **model world** will at best be good approximations to the true underlying data generating model F_{true}. 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 $F_{\text{true}} = F_{\boldsymbol{\theta}_{\text{true}}}$. 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.



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 *overfitting* (models being too complex and don't generalise well) or *underfitting* (models being too simplistic and hence also don't predict well).

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

1.4 Common distributions used in statistical models

Models employed in statistical analysis are typically multivariate comprising many random variables. As such these models can be very complex, with hierarchical or network-like structures linking observed and latent variables, and possibly exhibiting nonlinear functional relationships.

However, nonetheless even the most complex models will normally be composed of more elementary building blocks. For example, the following distributions frequently occur as simple parametric models:

- **Bernoulli distribution** $\text{Ber}(\theta)$ and **categorical distribution** $\text{Cat}(\boldsymbol{\pi})$: used to model frequencies (on the domain $[0, 1]$). Repeated application yields the **binomial distribution** $\text{Bin}(n, \theta)$ and **multinomial distribution** $\text{Mult}(n, \boldsymbol{\pi})$.
- **Normal distribution** in both the univariate $N(\mu, \sigma^2)$ and multivariate $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ version: commonly used to model mean values (on the domain $[-\infty, \infty]$).
- **Gamma distribution** $\text{Gam}(\alpha, \theta)$: used to model scale factors (on the domain $[0, \infty]$). It is also known (with different parameterisation) as

univariate **Wishart distribution** $W_1\left(s^2, k\right)$ or as **scaled chi-squared distribution** $s^2\chi^2_k$. Special cases include the **chi-squared distribution** χ^2_k and the **exponential distribution** $\text{Exp}(\theta)$.

All the above parametric distributions are examples of **exponential families**. A distribution family $P_{\boldsymbol{\eta}}$ is an exponential family with canonical parameters $\boldsymbol{\eta}$ if the family results from **exponential tilting** of a **base distribution** B. In this case its density or probability mass function can be written as

$$p(x|\boldsymbol{\eta}) = b(x)\, e^{\boldsymbol{\eta}^T \boldsymbol{u}(x)} / e^{\psi(\boldsymbol{\eta})}$$

where $\boldsymbol{u}(x)$ are the canonical statistics and

$$\psi(\boldsymbol{\eta}) = \log \int_x b(x)\, e^{\boldsymbol{\eta}^T \boldsymbol{u}(x)}\, dx$$

is the **log-normaliser** or **log-partition function**. Exponential families have many favourable properties that facilitate estimation and inference, such as existence of all moments, easy to optimise, analytic solutions for parameter estimates, among many others.

Another commonly used parametric distributuion is a generalisation of the normal model:

- **Location-scale t-distribution** $\text{lst}(\mu, \tau^2, \nu)$: similar to the normal distibution $N(\mu, \sigma^2)$ but with heavier tails. It emerges as the sampling distribution for the t-statistic and as compound distribution in Bayesian learning. Special cases include the **Student's t_ν distribution** and **Cauchy distribution** $\text{Cau}(\mu, \tau)$. Note that this is not an exponential family and due to the heavy tails, depending on the choice of the degrees of freedom ν, not all moments may exist.

Finally, **nonparametric approaches** are also often used to describe and analyse the observed data. Rather than specifying a parametric model for F one focuses on using the whole distribution to define meaningful **statistical functionals** $\theta = g(F)$, such as the mean and the variance.

The above models will be used throughout the module — for a refresher of the technical details please refer to the Appendix. In the second part of the module (Bayesian statistics) we will encounter further distributions such as the **beta distribution** or **inverse gamma distribution**.

Subsequent modules in later study years (year 3 and 4) introduce more complex models, related to temporal and spatial modelling, regression analysis and generalised linear models, and multivariate statistics and machine learning.

Exponential families are studied in more detail the module about generalised linear models.

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 **nonparametric 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

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

and of the variance

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

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 $P_{\boldsymbol{\theta}}$ with density or mass function $p(x|\boldsymbol{\theta})$ where $\boldsymbol{\theta}$ 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 then defined as

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

The parameter that maximises the likelihood is the **maximum likelihood estimate**.

The first part of this module is devoted to exploring the method of maximum likelihood both practically and more theoretically. We start by considering the justification of the method of maximum likelihood. Historically, the likelihood function was introduced (and still often is interpreted) as the probability to observe the data given the model with specified parameters $\boldsymbol{\theta}$. However, this view is incorrect as it not only breaks down for continuous random variables due to the use of densities and 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, the true foundation of maximum likelihood lies in information theory, specifically in its close link with the relative entropy between the unknown true distribution F and the model $P_{\boldsymbol{\theta}}$. As a

result we will see that **maximum likelihood extends empirical estimation to parametric models**. This insight allows to shed light both on the optimality properties as well as the limitations of maximum likelihood inference.

In the second part we then introduce 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.

The aim of this module is therefore

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 (a still widespread but wrong perception of statistics).

Chapter 2

From entropy to maximum likelihood

2.1 Overview

The purpose of this chapter is to introduce various important information criteria used in statistics and machine learning. These are based on entropy and provide the foundation for the method of maximum likelihood. Thus, this will also enable a better understanding the properties of maximum likelihood.

The modern definition of (relative) entropy was first discovered in the 1870s by physicist Ludwig Boltzmann (1844–1906) in the context of thermodynamics. The probabilistic interpretation of statistical mechanics and entropy was further developed by Josiah W. Gibbs (1839–1903).

In essence, in physics entropy describes the spread or concentration of energy.

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 are affiliated with the University of Manchester.

Entropy ↗ Shannon Entropy (Shannon 1948)
↘ Relative Entropy (Kullback-Leibler 1951)

Fisher information → Likelihood theory (Fisher 1922)

Mutual Information → Information theory (Shannon 1948, Lindley 1953)

2.2 Shannon entropy and differential entropy

2.2.1 Surprise, surprisal or Shannon information

The **surprise** to observe an event of probability p is **defined** as $-\log(p)$. This is also called **surprisal** or **Shannon information**.

Thus, the surprise to observe a certain event (with $p = 1$) is zero, and conversely the surprise to observe an event that is certain not to happen (with $p = 0$) is infinite.

The **log-odds ratio** can be viewed as the difference of the surprise of an event and the surprise of the complementary event:

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

In this module we always use the *natural logarithm* by default, and will explicitly write $\log_2$ and $\log_{10}$ for logarithms with respect to base 2 and 10, respectively.

Surprise and entropy computed with the natural logarithm (log) is given in "nats" (=natural information units). Using $\log_2$ leads to "bits" and using $\log_{10}$ to "ban" or "Hartley".

2.2.2 Shannon entropy

Assume we have a categorical distribution P with K classes/categories. The corresponding class probabilities are $p_1, \ldots, p_K$ with Pr("class k") $= p_k$ and $\sum_{k=1}^{K} p_k = 1$. The probability mass function (PMF) is $p(x = \text{"class k"}) = p_k$.

As the random variable x is discrete the categorical distribution P is a discrete distribution. However, P is generally also known as *the* discrete distribution.

The **Shannon entropy** (1948) [1] of the distribution P is defined as the **expected**

[1] 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

surprise or **expected Shannon information**, i.e. the negative expected log-probability

$$\begin{aligned} H(P) &= -\mathrm{E}_P\left(\log p(x)\right) \\ &= -\sum_{k=1}^{K} p_k \log(p_k) \end{aligned}$$

As all $p_k \in [0, 1]$ by construction Shannon entropy must be larger or equal to 0.

Furthermore, it is bounded above by $\log K$. This can be seen by maximising Shannon entropy as a function with regard to the p_k under the constraint $\sum_{k=1}^{K} p_k = 1$, e.g., by constrained optimisation using Lagrange multipliers. The maximum is achieved for P being the discrete uniform - see Example 2.3.

Hence for any categorical distribution P with K categories we have

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

In statistical physics, the Shannon entropy is known as Gibbs entropy (1878).

2.2.3 Combinatorial derivation of Shannon entropy

There is a close link of Shannon entropy with multinomial coefficients counting the permutations of n items (samples) of K distinct types (classes):

Example 2.1. Large sample asymptotics of the log-multinomial coefficient and link to Shannon entropy:

The number of possible permutation of n items of K distinct types, with n_1 of type 1, n_2 of type 2 and so on, is given by the multinomial coefficient

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

with $\sum_{k=1}^{K} n_k = n$ and $K \leq n$.

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

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

With this

$$\begin{aligned}
\log W &= \log \binom{n}{n_1, \ldots, n_K} \\
&= \log n! - \sum_{k=1}^{K} \log n_k! \\
&\approx n \log n - n - \sum_{k=1}^{K} (n_k \log n_k - n_k) \\
&= n \log n - \sum_{k=1}^{K} n_k \log n_k \\
&= \sum_{k=1}^{K} n_k \log n - \sum_{k=1}^{K} n_k \log n_k \\
&= -n \sum_{k=1}^{K} \frac{n_k}{n} \log \left(\frac{n_k}{n} \right)
\end{aligned}$$

and thus

$$\begin{aligned}
\frac{1}{n} \log \binom{n}{n_1, \ldots, n_K} &\approx - \sum_{k=1}^{K} \hat{p}_k \log \hat{p}_k \\
&= H(\hat{P})
\end{aligned}$$

where $\hat{P}$ is the empirical categorical distribution with $\hat{p}_k = \frac{n_k}{n}$.

The combinatorial derivation of Shannon entropy is credited in statistics to Wallis (1962) but has already been used a century earlier by Boltzmann (1877) who discovered it in his work in statistical mechanics (recall $S = k_b \log W$ is the Boltzmann entropy).

2.2.4 Differential entropy

Shannon entropy is only defined for discrete random variables.

Differential Entropy results from applying the definition of Shannon entropy to a *continuous* random variable x with density $p(x)$:

$$H(P) = -\mathrm{E}_P(\log p(x)) = -\int_x p(x) \log p(x)\, dx$$

Despite having essentially the same formula the different name is justified because differential entropy exhibits different properties compared to Shannon entropy, because the logarithm is taken of a density which in contrast to a probability can assume values larger than one. As a consequence, differential entropy is *not* bounded below by zero and can be negative.

2.2.5 Entropy examples

A) Models with single parameter

Example 2.2. Consider the uniform distribution $U(0, a)$ with $a > 0$, support from 0 to a and density $p(x) = 1/a$. As $-\int_0^a p(x) \log p(x) dx = -\int_0^a \frac{1}{a} \log(\frac{1}{a}) dx = \log a$ the differential entropy is

$$H(U(0, a)) = \log a \, .$$

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

B) Models with multiple parameters

Example 2.3. Discrete uniform distribution U_K: let $p_1 = p_2 = \ldots = p_K = \frac{1}{K}$. Then

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

Note that $\log K$ is the largest value the Shannon entropy can assume with K classes and indicates maximum spread.

Example 2.4. 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 entropy

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

Note that 0 is the smallest value that Shannon entropy can assume and that it corresponds to maximum concentration.

Example 2.5. 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 $\mathrm{E}((x-\mu)^2) = \sigma^2$

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

Interestingly, $H(P)$ only depends on the variance and not on the mean, and the entropy grows with the variance. Note that for $\sigma^2 < 1/(2\pi e) \approx 0.0585$ the differential entropy is negative.

2.2.6 Maximum entropy principle to characterise distributions

Both maximum Shannon entropy and differential entropy are useful to characterise distributions:

As seen in the examples above, **large entropy** implies that the **distribution is spread out** whereas **small entropy** means the **distribution is concentrated**.

Correspondingly, maximum entropy distributions can be considered minimally informative about a random variable. The higher the entropy the more spread out (and hence more uninformative) the distribution.

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

Using maximum entropy to characterise maximally uninformative distributions was advocated by E.T. Jaynes (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.

2.3 Cross-entropy and Kullback-Leibler divergence

2.3.1 Definition of cross-entropy

If we modify the definition of Shannon entropy (and differential entropy) such that the expectation over the log-density (say $g(x)$ of distribution G) is taken with regard to a different distribution F we arrive at the **cross-entropy**

$$H(F, G) = -\mathrm{E}_F\left(\log g(x)\right)$$

Thus, cross-entropy is a functional of two distributions F and G.

For discrete distributions F and G with class probabilities $f_1, \ldots, f_K$ and $g_1, \ldots, g_K$ the cross-entropy is computed as the weighted sum

$$H(F, G) = -\sum_{k=1}^{K} f_k \log g_k$$

For continuous distributions F and G with densities $f(x)$ and $g(x)$ we compute the integral

$$H(F, G) = -\int_x f(x) \log g(x)\, dx$$

Note that

- Cross-entropy is not symmetric with regard to F and G, because the expectation is taken with reference to F.
- By construction if both distributions are identical cross-entropy reduces to Shannon and differential entropy, respectively: $H(F, F) = H(F)$.

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

$$H(F, G) \geq H(F)$$

with equality for $F = G$. This is known as **Gibbs' inequality**.

Equivalently we can write

$$\underbrace{H(F, G) - H(F)}_{\text{relative entropy}} \geq 0$$

In fact, this recalibrated cross-entropy (known as KL divergence or relative entropy) turns out to be more fundamental than both cross-entropy and entropy. It will be studied in detail in the next section.

2.3.2 Definition of KL divergence

Also known as **relative entropy** and **discrimination information**.

The **relative entropy** measures the **divergence** of a distribution G from the distribution F and is defined as

$$\begin{aligned} D_{\mathrm{KL}}(F, G) &= \mathrm{E}_F \log\left(\frac{dF}{dG}\right) \\ &= \mathrm{E}_F \log\left(\frac{f(x)}{g(x)}\right) \\ &= \underbrace{-\mathrm{E}_F(\log g(x))}_{\text{cross-entropy}} - \Big(\underbrace{-\mathrm{E}_F(\log f(x))}_{\text{(differential) entropy}} \Big) \\ &= H(F, G) - H(F) \end{aligned}$$

- $D_{\mathrm{KL}}(F, G)$ measures the amount of information lost if G is used to approximate F.
- If F and G are identical (and no information is lost) then $D_{\mathrm{KL}}(F, G) = 0$.

(Note: here "divergence" measures the dissimilarity between probability distributions. This type of divergence is not related and should not be confused with divergence (div) as used in vector analysis.)

The use of the term "divergence" rather than "distance" serves to emphasise that the distributions F and G are not interchangeable in $D_{\mathrm{KL}}(F, G)$.

There exist various notations for KL divergence in the literature. Here we use $D_{\mathrm{KL}}(F, G)$ but you often find as well $\mathrm{KL}(F||G)$ and $I^{KL}(F; G)$.

Some authors (e.g. Efron) call twice the KL divergence $2D_{\mathrm{KL}}(F, G) = D(F, G)$ the **deviance** of G from F.

2.3.3 Properties of KL divergence

1. $D_{\mathrm{KL}}(F, G) \neq D_{\mathrm{KL}}(G, F)$, i.e. the KL divergence is not symmetric, F and G cannot be interchanged.
2. $D_{\mathrm{KL}}(F, G) = 0$ if and only if $F = G$, i.e., the KL divergence is zero if and only if F and G are identical.
3. $D_{\mathrm{KL}}(F, G) \geq 0$, proof via **Jensen's inequality**.
4. $D_{\mathrm{KL}}(F, G)$ remains invariant under coordinate transformations, i.e. it is an invariant geometric quantity.

Note that in the KL divergence the expectation is taken over a ratio of densities (or ratio of probabilities for discrete random variables). This is what creates the transformation invariance.

For more details and proofs of properties 3 and 4 see Worksheet E1.

2.3.4 Origin of KL divergence and application in statistics

Historically, in physics (negative) relative entropy was discovered by Boltzmann (1878). [2] In statistics and information theory it was introduced by Kullback and Leibler (1951). [3]

In statistics the typical roles of the distribution F and G in $D_{\mathrm{KL}}(F, G)$ are:

- F is the (unknown) underlying true model for the data generating process
- G is the approximating model (typically a distribution family indexed by parameters)

Optimising (i.e. minimising) the KL divergence with regard to G amounts to *approximation* and optimising with regard to F to *imputation*. Later we will see how this leads to the method of maximum likelihood and to Bayesian learning, respectively.

2.3.5 Cross-entropy and KL divergence examples

A) Models with a single parameter

Example 2.6. KL divergence between two Bernoulli distributions $\mathrm{Ber}(\theta_1)$ and $\mathrm{Ber}(\theta_2)$:

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

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

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

Assume $F_{\mathrm{ref}} = N(\mu_{\mathrm{ref}}, \sigma^2)$ and $F = N(\mu, \sigma^2)$.

Then we get

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

B) Models with multiple parameters

Example 2.8. KL divergence between two categorical distributions with K classes:

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

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

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

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

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

Example 2.9. Cross-entropy between two normals:

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

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

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

Example 2.10. If $\mu_{\text{ref}} = \mu$ and $\sigma^2_{\text{ref}} = \sigma^2$ then the cross-entropy $H(F_{\text{ref}}, F)$ in Example 2.9 degenerates to the differential entropy $H(F_{\text{ref}}) = \frac{1}{2}\left(\log(2\pi\sigma^2_{\text{ref}}) + 1\right)$.

Example 2.11. KL divergence between two univariate normals with different means and variances:

Assume $F_{\text{ref}} = N(\mu_{\text{ref}}, \sigma^2_{\text{ref}})$ and $F = N(\mu, \sigma^2)$. Then

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

If variances are equal then we recover the previous Example 2.7 as special case.

2.4 Local quadratic approximation and expected Fisher information

2.4.1 Definition of expected Fisher information

KL information measures the divergence of two distributions. We may thus use relative entropy to measure the divergence between two distributions in the same family, separated in parameter space only by some small $\boldsymbol{\varepsilon}$.

First, we consider $D_{\text{KL}}(F_{\boldsymbol{\theta}}, F_{\boldsymbol{\theta}+\boldsymbol{\varepsilon}}) = \text{E}_{F_{\boldsymbol{\theta}}}\left(\log f(\boldsymbol{x}|\boldsymbol{\theta}) - \log f(\boldsymbol{x}|\boldsymbol{\theta}+\boldsymbol{\varepsilon})\right) = h(\boldsymbol{\varepsilon})$ where $\boldsymbol{\theta}$ is kept constant and $\boldsymbol{\varepsilon}$ is varying. From the properties of the KL divergence we know that $D_{\text{KL}}(F_{\boldsymbol{\theta}}, F_{\boldsymbol{\theta}+\boldsymbol{\varepsilon}}) \geq 0$ and that it becomes zero only if $\boldsymbol{\varepsilon} = 0$. Thus, by construction the function $h(\boldsymbol{\varepsilon})$ achieves a true minimum of $h(0) = 0$ at $\boldsymbol{\varepsilon} = 0$, with a vanishing gradient $\nabla h(0) = 0$ and a positive definite Hessian matrix $\nabla\nabla^T h(0)$. Therfore we can approximate it by a quadratic function around $\boldsymbol{\varepsilon} = 0$:

$$h(\boldsymbol{\varepsilon}) \approx \frac{1}{2}\boldsymbol{\varepsilon}^T \nabla\nabla^T h(0)\, \boldsymbol{\varepsilon}$$

The Hessian matrix $\nabla\nabla^T h(0)$ is computed as $\boldsymbol{I}^{\text{Fisher}}(\boldsymbol{\theta}) = -\text{E}_{F_{\boldsymbol{\theta}}} \nabla\nabla^T \log f(\boldsymbol{x}|\boldsymbol{\theta})$ and is the negative expected Hessian matrix of the log-density at $\boldsymbol{\theta}$. It is called the **expected Fisher information** at $\boldsymbol{\theta}$. The KL divergence can thus be locally approximated by

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

As second possibility 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 $\boldsymbol{\varepsilon}$:

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

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

2.4.2 Expected Fisher information for a set of independent and identically distributed random variables

Furthermore, we can also compute the expected Fisher information for a set of 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 $\boldsymbol{I}^{\text{Fisher}}(\boldsymbol{\theta})$. The expected Fisher information $\boldsymbol{I}^{\text{Fisher}}_{x_1,\ldots,x_n}(\boldsymbol{\theta})$ 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 yields

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

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

2.5 Expected Fisher information examples

2.5.1 Models with a single parameter

Example 2.12. Expected Fisher information for the Bernoulli distribution:

The log-probability mass function of the Bernoulli $\text{Ber}(\theta)$ 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 $\text{E}(x) = \theta$ we get as Fisher information

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

Example 2.13. Quadratic approximations of the KL divergence between two Bernoulli distributions:

From Example 2.6 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 2.12 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.

2.5.2 Models with multiple parameters

Example 2.14. 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 2.15. 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

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

For $K = 2$ and $\pi_1 = \theta$ this reduces to the expected Fisher information of a Bernoulli variable, see Example 2.12.

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

Example 2.16. 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 $D_{\mathrm{KL}}(P, Q)$ between the categorical distribution $P = \mathrm{Cat}(\boldsymbol{p})$ with probabilities $\boldsymbol{p} = (p_1, \ldots, p_K)^T$ with the categorical distribution $Q = \mathrm{Cat}(\boldsymbol{q})$ with probabilities $\boldsymbol{q} = (q_1, \ldots, q_K)^T$.

From Example 2.8 we already know the KL divergence and from Example 2.15 the corresponding expected Fisher information.

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

$$
\begin{aligned}
D_{\mathrm{KL}}(P, Q = P + \varepsilon) &= D_{\mathrm{KL}}(\mathrm{Cat}(\boldsymbol{p}), \mathrm{Cat}(\boldsymbol{p} + \boldsymbol{\varepsilon})) \\
&\approx \frac{1}{2}(\varepsilon_1, \ldots, \varepsilon_{K-1})\, \boldsymbol{I}^{\mathrm{Fisher}}(p_1, \ldots, p_{K-1}) \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_{K-1} \end{pmatrix} \\
&= \frac{1}{2}\left(\sum_{k=1}^{K-1} \frac{\varepsilon_k^2}{p_k} + \frac{\left(\sum_{k=1}^{K-1} \varepsilon_k\right)^2}{p_K} \right) \\
&= \frac{1}{2} \sum_{k=1}^{K} \frac{\varepsilon_k^2}{p_k} \\
&= \frac{1}{2} \sum_{k=1}^{K} \frac{(p_k - q_k)^2}{p_k} \\
&= \frac{1}{2} D_{\mathrm{Neyman}}(P, Q)
\end{aligned}
$$

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

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

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 $D_{\mathrm{Neyman}}(P, Q) = D_{\mathrm{Pearson}}(Q, P)$.

2.6 From entropy learning to maximum likelihood

2.6.1 The relative entropy between true model and approximating model

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

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

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

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 $D_{\mathrm{KL}}(F, G_{\boldsymbol{\theta}})$ 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 $n \to \infty$:
$$\hat{F}_n \overset{a.s.}{\to} F$$
- 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 $H(F, G_{\boldsymbol{\theta}})$ is approximated by the **empirical cross-entropy** where the expectation is taken with regard to $\hat{F}_n$ rather than F:

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

This turns out to be equal to the negative log-likelihood standardised by the sample size n! Or in other words, the **log-likelihood** is the **negative empirical cross-entropy multiplied by sample size** n.

From the link of the multinomial coefficient with Shannon entropy (Example 2.1) we already know that for large sample size

$$H(\hat{F}) \approx \frac{1}{n} \log \binom{n}{n_1, \ldots, n_K}$$

The KL divergence $D_{\mathrm{KL}}(F, G_{\boldsymbol{\theta}})$ can therefore be approximated by

$$D_{\mathrm{KL}}(F, G_{\boldsymbol{\theta}}) \approx -\frac{1}{n} \left(\log \binom{n}{n_1, \ldots, n_K} + l_n(\boldsymbol{\theta}|D) \right)$$

Thus, with the KL divergence we obtain not just the log-likelihood (the cross-entropy part) but also the multiplicity factor taking account of the possible orderings of the data (the entropy part).

2.6.2 Minimum KL divergence and maximum likelihood

If we knew F we would simply minimise $D_{\mathrm{KL}}(F, G_{\boldsymbol{\theta}})$ to find the particular model $G_{\boldsymbol{\theta}}$ that is closest to the true model. Equivalently, we would minimise the cross-entropy $H(F, G_{\boldsymbol{\theta}})$. 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 $D_{\mathrm{KL}}(F, G_{\boldsymbol{\theta}})$ we simply minimise $H(\hat{F}_n, G_{\boldsymbol{\theta}})$ which is the same as maximising the log-likelihood $l_n(\boldsymbol{\theta}|D)$.

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!

$$\begin{aligned} \hat{\boldsymbol{\theta}}^{ML} &= \arg\max_{\boldsymbol{\theta}} \; l_n(\boldsymbol{\theta}|D) \\ &= \arg\min_{\boldsymbol{\theta}} \; H(\hat{F}_n, G_{\boldsymbol{\theta}}) \\ &\approx \arg\min_{\boldsymbol{\theta}} \; D_{\mathrm{KL}}(F, G_{\boldsymbol{\theta}}) \end{aligned}$$

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 $G_{\boldsymbol{\theta}}$ from the unkown true model F. In other words, **maximum likelihood estimators are minimum empirical relative entropy estimators**.

As the relative entropy 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 relative entropy maximum likelihood inherits for large n (and only then!) all the optimality properties from KL divergence. These will be discussed in more detail later in the course.

2.6.3 Further connections

Since minimising KL divergence contains ML estimation as special case you may wonder whether there is a broader justification of relative entropy in the context of statistical data analysis?

Indeed, KL divergence has strong geometrical interpretation that forms the basis of *information geometry*. In this field the manifold of distributions is studied using tools from differential geometry. The expected Fisher information plays an important role as metric tensor in the space of distributions.

Furthermore, it is also linked to probabilistic forecasting. In the framework of so-called **scoring rules**. the only local proper scoring rule is the negative log-probability ("surprise"). The expected "surprise" is the cross-entropy and relative entropy is the corresponding natural divergence connected with the log scoring rule.

Furthermore, another intriguing property of KL divergence is that the relative entropy $D_{\mathrm{KL}}(F, G)$ is the *only divergence measure* that is both a Bregman and an f-divergence. Note that f-divergences and Bregman-divergences (in turn related to proper scoring rules) are two large classes of measures of similarity and divergence between two probability distributions.

Finally, not only the likelihood estimation but also the Bayesian update rule (as discussed later in this module) is another special case of entropy learning.

Chapter 3

Maximum likelihood estimation

3.1 Outline of maximum likelihood estimation

3.1.1 General procedure

The starting points in an ML analysis are

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

From this we construct the likelihood function:

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

Historically, the likelihood is also often interpreted as the probability of the data given the model. However, this is not strictly correct. First, this interpretation only applies to discrete random variables. Second, since the samples are iid even in this case one would still need to add a factor accounting for the multiplicity of 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 $f(x|\boldsymbol{\theta})$ is a density, not a probability.

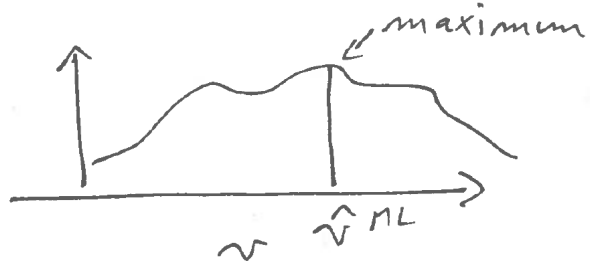
As we have seen in the previous chapter the origin of the likelihood function lies in its connection to relative entropy. Specifically, the log-likelihood function

- $l_n(\boldsymbol{\theta}|D) = \sum_{i=1}^{n} \log f(x_i|\boldsymbol{\theta})$

divided by sample size n is a large sample approximation of the cross-entropy between the unknown true data generating model and the approximating model $F_{\boldsymbol{\theta}}$. Note that the log-likelihood is additive over the samples x_i.

The maximum likelihood point estimate $\hat{\boldsymbol{\theta}}^{ML}$ is then given by maximising the (log)-likelihood

$$\hat{\boldsymbol{\theta}}^{ML} = \arg\max l_n(\boldsymbol{\theta}|D)$$



Thus, finding the MLE is an optimisation problem that in practise is most often solved numerically on the computer, using approaches such as *gradient ascent* (or for negative log-likelihood *gradient descent*) and related algorithms. The numerical optimisation is usually done on the log-likelihood rather than on the likelihood function to avoid problems with the computer representation of small floating point numbers.

Depending on the complexity of the likelihood function finding the maximum can indeed be very difficult. On the other hand, for likelihood functions constructed from exponential families maximum likelihood estimation is very straightforward.

3.1.2 Obtaining MLEs for a regular model

In regular situations, i.e. when

- the log-likelihood function is twice differentiable with regard to the parameters,
- the maximum (peak) of the likelihood function lies inside the parameter space and not at a boundary,
- the parameters of the model are all identifiable (in particular the model is not overparameterised), and
- the second derivative of the log-likelihood at the maximum is negative and not zero (for more than one parameter: the Hessian matrix at the maximum is negative definite and not singular)

then in order to maximise $l_n(\boldsymbol{\theta}|D)$ one may use the **score function** $S(\boldsymbol{\theta})$ which is the first derivative of the log-likelihood function with regard to the parameter:

$S_n(\theta) = \frac{dl_n(\theta|D)}{d\theta}$ scalar parameter θ: first derivative of log-likelihood function

$\boldsymbol{S}_n(\boldsymbol{\theta}) = \nabla l_n(\boldsymbol{\theta}|D)$ gradient if $\boldsymbol{\theta}$ is a vector (i.e. if there's more than one parameter)

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:

$$\frac{d^2 l_n(\hat{\theta}_{ML}|D)}{d\theta^2} < 0$$

In the case of a parameter vector (multivariate $\boldsymbol{\theta}$) you need to compute the Hessian matrix (matrix of second order derivatives) at the MLE:

$$\nabla\nabla^T l_n(\hat{\boldsymbol{\theta}}_{ML}|D)$$

and then verify that this matrix is negative definite (i.e. all its eigenvalues must be negative).

As we will see later the second order derivatives of the log-likelihood function also play an important role for assessing the uncertainty of the MLE.

3.2 Maximum likelihood estimation in practise

3.2.1 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 3.1. Estimation of a proportion – maximum likelihood 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 Ber(θ): Pr("success") = θ 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 Ber(θ) 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:

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

  Note how the log-likelihood depends on the data only through $\bar{x}$! This is an example of a *sufficient statistic* for the parameter θ (in fact it is also a *minimally* sufficient statistic). This will be discussed in more detail later.

- Score function:

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

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

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

  With $\frac{dS_n(\theta)}{d\theta} = -n\left(\frac{\bar{x}}{\theta^2} + \frac{1-\bar{x}}{(1-\theta)^2}\right) < 0$ 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 3.2. Normal distribution with unknown mean and known variance:

- $x \sim N(\mu, \sigma^2)$ with $\text{E}(x) = \mu$ and $\text{Var}(x) = \sigma^2$
- 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:

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

- Log-Density:

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

- Log-likelihood function:

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

  Note how the non-constant terms of the log-likelihood depend on the data only through $\bar{x}$!

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

The constant term C in the log-likelihood function collects all terms that do not depend on the parameter. After taking the first derivative with regard to the parameter this term disappears thus C **is not relevant for finding the MLE** of the parameter. **In the future we will often omit such constant terms from the log-likelihood function without further mention.**

Example 3.3. 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:

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

- Log-Density:

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

- Log-likelihood function:

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

  Note how the log-likelihood function depends on the data only through $\overline{(x-\mu)^2}$!

- 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 $\frac{dS_n(\sigma^2)}{d\sigma^2} = -\frac{n}{2\sigma^4}\left(\frac{2}{\sigma^2}\overline{(x-\mu)^2} - 1\right)$ and hence $\frac{dS_n(\widehat{\sigma^2}_{ML})}{d\sigma^2} = -\frac{n}{2}\left(\widehat{\sigma^2}_{ML}\right)^{-2} < 0$ the optimum is indeed the maximum.

3.2.2 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 3.4. Normal distribution with mean and variance both unknown:

- $x \sim N(\mu, \sigma^2)$ with $\mathrm{E}(x) = \mu$ and $\mathrm{Var}(x) = \sigma^2$
- 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:

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

  Note how the log-likelihood function depends on the data only through $\bar{x}$ and $\overline{x^2}$!

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

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

- 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 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 3.9.

Example 3.5. Maximum likelihood estimates of the parameters of the multivariate normal distribution:

The results from Example 3.4 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 3.6. Maximum likelihood estimation of the parameters of the categorical distribution:

Maximum likelihood estimation seems trivial at first sight but it is in fact a bit more complicated since there are only $K-1$ free parameters, and not K. So 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 $\sum_{k=1}^{K} \pi_k = 1$ (for example by using a Lagrange multiplier).

- 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 $\bar{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i = (\bar{x}_1, \ldots, \bar{x}_K)^T$ 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

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

- Score function (gradient)

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

  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 of the occurance of a class among the n samples.

3.3 Properties of maximum likelihood estimation

3.3.1 Invariance property of the maximum likelihood

The invariance principle states that the **maximum likelihood is invariant against reparameterisation**.

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 reason why this works is that maximisation is a procedure that is invariant against transformations of the argument of the function that is maximised. Consider a function $h(x)$ with a maximum at $x_{\max} = \arg\max h(x)$. Now we relabel the argument using $y = g(x)$ where g is an invertible function. Then the function in terms of y is $h(g^{-1}(y))$. and clearly this function has a maximum at $y_{\max} = g(x_{\max})$ since $h(g^{-1}(y_{\max})) = h(x_{\max})$.

The invariance property can be very useful in practise 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.

3.3.2 Consistency of maximum likelihood estimates

One important property of the method of maximum likelihood is that in general it produces **consistent estimates**.

Specifically, if the true underlying model F_{true} with parameter $\boldsymbol{\theta}_{\text{true}}$ is contained in the set of specified candidates models $F_{\boldsymbol{\theta}}$

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

then

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

This is a consequence of $D_{\text{KL}}(F_{\text{true}}, F_{\boldsymbol{\theta}}) \to 0$ for $F_{\boldsymbol{\theta}} \to F_{\text{true}}$, and that maximisation of the likelihood function is for large n equivalent to minimising the relative entropy.

Thus given sufficient data the MLE will converge to the true value. As a consequence, **MLEs 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 F_θ is misspecified (i.e. it does not contain the actual true model) the MLE is still optimal in the sense in that it will find the closest possible model.

It is possible to find inconsistent MLEs, but this occurs only in situations where the dimension of the model / number of parameters increases with sample size, or when the MLE is at a boundary or when there are singularities in the likelihood function.

3.3.3 Relationship of maximum likelihood with least squares estimation

In Example 3.2 the form of the log-likelihood function is a function of the sum of squared differences. Maximising $l_n(\mu|D) = -\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu)^2$ 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 with maximum likelihood and normality allows to understand why it usually works well!

3.3.4 Bias of maximum likelihood estimates

Example 3.4 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 Appendix) 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

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

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

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

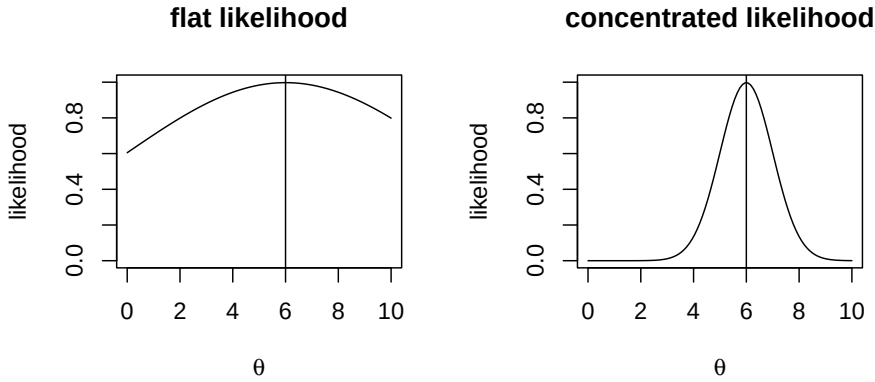
$$\widehat{\sigma^2}_{UB} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{n}{n-1}\widehat{\sigma^2}_{ML}$$

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.

3.4 Observed Fisher information

3.4.1 Motivation and definition of observed Fisher information



By inspection of some log-likelihood curves it is apparent that the log-likelihood function contains more information about the parameter $\boldsymbol{\theta}$ than just the maximum point $\hat{\boldsymbol{\theta}}_{ML}$.

In particular the **curvature** of the log-likelihood function at the MLE must be somehow related 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 peaked (strong curvature) then the maximum point is clearly defined.

The curvature is described by the second-order derivatives (Hessian matrix) of the log-likelihood function.

For univariate θ the Hessian is a scalar:

$$\frac{d^2 l_n(\theta|D)}{d\theta^2}$$

For multivariate parameter vector $\boldsymbol{\theta}$ of dimension d the Hessian is a matrix of size $d \times d$:

$$\nabla\nabla^T l_n(\boldsymbol{\theta}|D)$$

By construction the Hessian is negative definite at the MLE (i.e. its eigenvalues are all negative) to ensure the the function is concave at the MLE (i.e. peak shaped).

The **observed Fisher information** (matrix) is defined as the negative curvature at the MLE $\hat{\boldsymbol{\theta}}_{ML}$:

$$\boldsymbol{J}_n(\hat{\boldsymbol{\theta}}_{ML}) = -\nabla\nabla^T l_n(\hat{\boldsymbol{\theta}}_{ML}|D)$$

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

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

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

3.4.2 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:

- Both types of Fisher information are based on computing second order derivatives (Hessian matrix), thus both are based on the curvature of a function.
- The observed Fisher information is computed from the log-likelihood function. Therefore it takes 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 discuss in more detail later).
- In contrast, the expected Fisher information is derived directly from the log-density. It does not depend on the observed data, and thus does not depend on sample size. It can be computed for any value of the parameters. It describes the geometry of the space of the models, and is the local approximation of relative entropy.

- Assume that for large sample size n the MLE converges to $\hat{\boldsymbol{\theta}}_{ML} \rightarrow \boldsymbol{\theta}_0$. It follows from the construction of the observed Fisher information and the law of large numbers that asymptotically for large sample size $\boldsymbol{J}_n(\hat{\boldsymbol{\theta}}_{ML}) \rightarrow n\boldsymbol{I}^{\text{Fisher}}(\boldsymbol{\theta}_0)$ (i.e. the expected Fisher information for a set of iid random variables, see Chapter 2).
- In a very important class of models, namely in an **exponential family model**, 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 is in fact the case for all the examples discussed above (e.g. see Examples 3.7 and 2.12 for the Bernoulli distribution and Examples 3.9 and 2.14 for the normal distribution).
- However, this is 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. An example is provided by the Cauchy distribution with median parameter θ. It is not an exponential family model and has expected Fisher information $I^{\text{Fisher}}(\theta) = \frac{1}{2}$ regardless of the choice the median parameter whereas the observed Fisher information $J_n(\hat{\theta}_{ML})$ depends on the MLE $\hat{\theta}_{ML}$ of the median parameter and is not simply $\frac{n}{2}$.

3.5 Observed Fisher information examples

3.5.1 Models with a single parameter

Example 3.7. Bernoulli model Ber(θ):

We continue Example 3.1. Recall that $\hat{\theta}_{ML} = \bar{x} = \frac{n_1}{n}$ and the score function $S_n(\theta) = n\left(\frac{\bar{x}}{\theta} - \frac{1-\bar{x}}{1-\theta}\right)$. The negative second derivative of the log-likelihood function is

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

The observed Fisher information is therefore

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

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

Example 3.8. Normal distribution with unknown mean and known variance:

This is the continuation of Example 3.2. Recall the MLE for the mean $\hat{\mu}_{ML} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$ and the score function $S_n(\mu) = \frac{n}{\sigma^2}(\bar{x} - \mu)$. The negative second derivative of the score function is

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

The observed Fisher information at the MLE is therefore

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

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 $\mathrm{Var}(\bar{x}) = \frac{\sigma^2}{n}$, which is equal to the inverse observed Fisher information.

3.5.2 Models with multiple parameters

Example 3.9. Normal distribution with mean and variance parameter:

This is the continuation of Example 3.4. Recall the MLE for the mean and variance:

$$\hat{\mu}_{ML} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$$

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

with score function

$$\boldsymbol{S}_n(\mu, \sigma^2) = \nabla l_n(\mu, \sigma^2|D) = \begin{pmatrix} \frac{n}{\sigma^2}(\bar{x} - \mu) \\ -\frac{n}{2\sigma^2} + \frac{n}{2\sigma^4}\left(\overline{x^2} - 2\mu\bar{x} + \mu^2\right) \end{pmatrix}$$

The Hessian matrix of the log-likelihood function is

$$\nabla\nabla^T l_n(\mu, \sigma^2|D) = \begin{pmatrix} -\frac{n}{\sigma^2} & -\frac{n}{\sigma^4}(\bar{x} - \mu) \\ -\frac{n}{\sigma^4}(\bar{x} - \mu) & \frac{n}{2\sigma^4} - \frac{n}{\sigma^6}\left(\overline{x^2} - 2\mu\bar{x} + \mu^2\right) \end{pmatrix}$$

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**:

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

Note that 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

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

Recall that $x \sim N(\mu, \sigma^2)$ and therefore

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

Hence $\text{Var}(\hat{\mu}_{ML}) = \frac{\sigma^2}{n}$. If you compare this with the first diagonal entry of the inverse observed Fisher information matrix you see that this is essentially the same expression (apart from the "hat").

The empirical variance $\widehat{\sigma^2}_{ML}$ follows a one-dimensional Wishart distribution

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

(see Appendix) with variance $\text{Var}(\widehat{\sigma^2}_{ML}) = \frac{n-1}{n}\frac{2\sigma^4}{n}$. For large n this becomes $\text{Var}(\widehat{\sigma^2}_{ML}) \overset{a}{=} \frac{2\sigma^4}{n}$ which is essentially (apart from the "hat") the second diagonal entry of the inverse observed Fisher information matrix.

Example 3.10. Observed Fisher information of the categorical distribution:

We continue Example 3.6. We first need to compute the negative Hessian matrix of the log likelihood function $-\nabla\nabla^T l_n(\pi_1, \ldots, \pi_{K-1})$ 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

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

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

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

Thus, the observed Fisher information matrix at the MLE for a categorical distribution is

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

For $K = 2$ (cf. Example 3.7) 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:

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

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

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

with

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

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

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

and

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

Chapter 4

Quadratic approximation and normal asymptotics

4.1 Approximate distribution of maximum likelihood estimates

4.1.1 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

$$l_1(\boldsymbol{\mu}|\boldsymbol{x}) = C - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$$

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 the point $(\boldsymbol{x}, C)$.

4.1.2 Quadratic approximation of a log-likelihood function

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

We assume the underlying model is regular and that $\nabla l_n(\hat{\boldsymbol{\theta}}_{ML}|D) = 0$.

The Taylor series approximation of scalar-valued function $f(\boldsymbol{x})$ around $\boldsymbol{x}_0$ is

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

Applied to the log-likelihood function this yields

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

This is a quadratic function with maximum at $(\hat{\boldsymbol{\theta}}_{ML}, l_n(\hat{\boldsymbol{\theta}}_{ML}|D))$. 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 4.1. Quadratic approximation of the log-likelihood for a proportion:

From Example 3.1 we have the log-likelihood

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

and the MLE

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

and from Example 3.7 the observed Fisher information

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

The log-likelihood at the MLE is

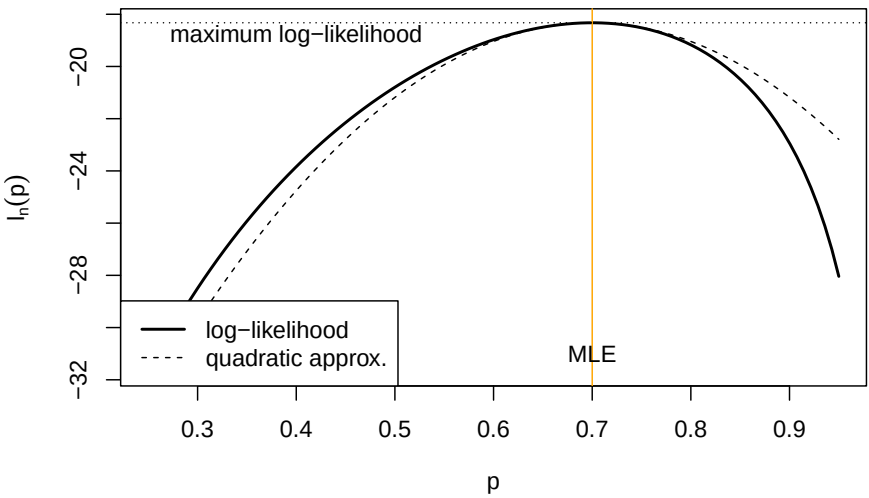
$$l_n(\hat{p}_{ML}|D) = n\left(\bar{x}\log\bar{x} + (1 - \bar{x})\log(1 - \bar{x})\right)$$

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

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

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 3.2) for one observation of $\hat{p}_{ML} = \bar{x}$ from $N\left(p, \frac{\bar{x}(1-\bar{x})}{n}\right)$.

The following figure shows the above log-likelihood function and its quadratic approximation for example data with $n = 30$ and $\bar{x} = 0.7$:



4.1.3 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

- 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 is normally distributed around the true parameter and with covariance equal to the inverse of the observed Fisher information**:

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

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.

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 $l_n(\boldsymbol{\theta})$ 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]

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

- 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.

4.1.4 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**,

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

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

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

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).

4.2 Quantifying the uncertainty of maximum likelihood estimates

4.2.1 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] .

4.2.2 Examples for the estimated variance and asymptotic normal distribution

Example 4.2. Estimated variance and distribution of the MLE of a proportion:

From Examples 3.1 and 3.7 we know the MLE

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

and the corresponding observed Fisher information

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

[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

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

From Examples 3.2 and 3.8 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 $N(\mu, \sigma^2)$!).

4.2.3 Wald statistic

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

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

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:

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

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

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

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

4.2.4 Examples of the (squared) Wald statistic

Example 4.4. Wald statistic for a proportion:

We continue from Example 4.2. With $\hat{p}_{ML} = \bar{x}$ and $\widehat{\text{Var}}(\hat{p}_{ML}) = \frac{\hat{p}_{ML}(1-\hat{p}_{ML})}{n}$ and thus $\widehat{\text{SD}}(\hat{p}_{ML}) = \sqrt{\frac{\hat{p}_{ML}(1-\hat{p}_{ML})}{n}}$ 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 4.5. Wald statistic for the mean parameter of a normal distribution with known variance:

We continue from Example 4.3. With $\hat{\mu}_{ML} = \bar{x}$ and $\widehat{\text{Var}}(\hat{\mu}_{ML}) = \frac{\sigma^2}{n}$ and thus $\widehat{\text{SD}}(\hat{\mu}_{ML}) = \frac{\sigma}{\sqrt{n}}$ 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 4.6. Wald statistic for the categorical distribution:

The squared Wald statistic is

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

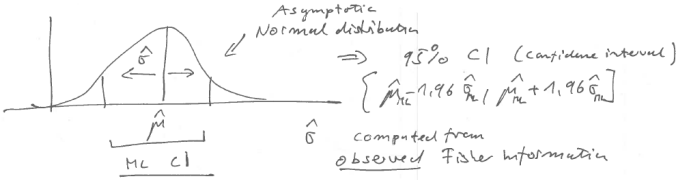
With $n_1, \ldots, n_K$ the observed counts with $n = \sum_{k=1}^K n_k$ and $\hat{\pi}_k^{ML} = \frac{n_k}{n} = \bar{x}_k$, and $n_1^{\text{expect}}, \ldots, n_K^{\text{expect}}$ the expected counts $n_k^{\text{expect}} = n p_k^0$ under $\boldsymbol{p}_0$ we can write the squared Wald statistic as follows:

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

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 $\boldsymbol{p}_0$.

4.2.5 Normal confidence intervals using the Wald statistic

The asymptotic normality of MLEs derived from regular models enables us to construct a corresponding normal confidence interval (CI):



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:

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

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 $c_{\text{normal}} = \Phi^{-1}\left(\frac{1+\kappa}{2}\right)$ (cf. refresher section in the Appendix).

coverage κ	Critical value c_{normal}
0.9	1.64
0.95	1.96
0.99	2.58

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 $\boldsymbol{\theta}_0$.

The following lists containst 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:

coverage κ	Critical value c_{chisq} $(m = 1)$
0.9	2.71
0.95	3.84
0.99	6.63

4.2.6 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 $1 - \kappa$.
- Conversely, all values θ_0 outside the CI can be rejected to be the true parameter with significance level $1 - \kappa$.

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.

4.2.7 Examples for normal CI and tests

Example 4.7. Asymptotic normal confidence interval for a proportion:

We continue from Examples 4.2 and 4.4. Assume we observe $n = 30$ measurements with average $\bar{x} = 0.7$. Then $\hat{p}_{ML} = \bar{x} = 0.7$ and $\widehat{\mathrm{SD}}(\hat{p}_{ML}) = \sqrt{\frac{\bar{x}(1-\bar{x})}{n}} \approx 0.084$.

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 4.8. Asymptotic normal test for a proportion:

We continue from Example 4.7.

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 4.4) 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 4.9. Normal confidence interval for the mean:

We continue from Examples 4.3 and 4.5. 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 4.10. Normal test for the mean:

We continue from Example 4.9.

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 4.5) 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.

4.3 Example of a non-regular model

Not all models allow a quadratic approximation of the log-likelihood function around the MLE. This is the case when the log-likelihood function is not differentiable at the MLE. These models are called non-regular and for those models the normal approximation is not available.

Example 4.11. Uniform distribution with upper bound θ:

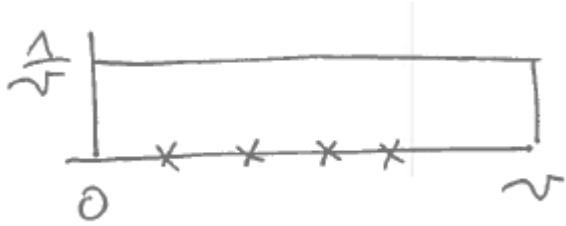
$$x_1, \ldots, x_n \sim U(0, \theta)$$

With $x_{[i]}$ we denote the *ordered* observations with $0 \leq x_{[1]} < x_{[2]} < \ldots < x_{[n]} \leq \theta$ and $x_{[n]} = \max(x_1, \ldots, x_n)$.

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

The probability density function of $U(0, \theta)$ is

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



and on the log-scale

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

Since all observed data $D = \{x_1, \ldots, x_n\}$ lie in the interval $[0, \theta]$ we get as log-likelihood function

$$l_n(\theta|D) = \begin{cases} -n \log \theta & \text{for } x_{[n]} \leq \theta \\ -\infty & \text{otherwise} \end{cases}$$

Obtaining the MLE of θ is straightforward: $-n \log \theta$ is monotonically decreasing with θ and $\theta \geq x_{[n]}$ hence the log-likelihood function has a maximum at $\hat{\theta}_{ML} = x_{[n]}$.

However, there is a discontinuity in $l_n(\theta|D)$ at $x_{[n]}$ and therefore $l_n(\theta|D)$ **is not differentiable** at $\hat{\theta}_{ML}$. Thus, **there is no quadratic approximation around** $\hat{\theta}_{ML}$ and the **observed Fisher information cannot be computed**. Hence, the normal approximation for the distribution of $\hat{\theta}_{ML}$ is not valid regardless of sample size, i.e. not even asymptotically for $n \to \infty$.

Nonetheless, we can in fact 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:

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

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

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

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.

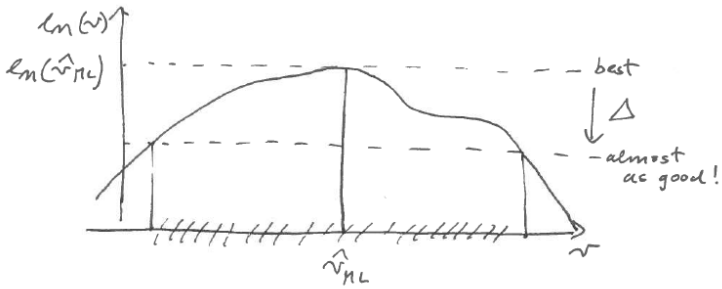
Chapter 5

Likelihood-based confidence interval and likelihood ratio

5.1 Likelihood-based confidence intervals and Wilks statistic

5.1.1 General idea and definition of Wilks log-likelihood ratio statistic

Instead of relying on normal / quadratic approximation, we can also use the log-likelihood directly to find the so called **likelihood confidence intervals**:



Idea: find all θ_0 that have a log-likelihood that is almost as good as $l_n(\hat{\theta}_{ML}|D)$.

$$\text{CI} = \{\theta_0 : l_n(\hat{\theta}_{ML}|D) - l_n(\theta_0|D) \leq \Delta\}$$

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

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

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

With its help we can write the likelihood CI 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

5.1.2 Examples of the Wilks statistic

Example 5.1. Wilks statistic for the proportion:

The log-likelihood for the parameter θ is (cf. Example 3.1)

$$l_n(\theta|D) = n(\bar{x}\log\theta + (1-\bar{x})\log(1-\theta))$$

Hence the Wilks statistic is

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

Comparing with Example 2.6 we see that in this case the Wilks statistic is essentially (apart from a scale factor $2n$) the KL divergence between two Bernoulli distributions:

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

Example 5.2. Wilks statistic for the mean parameter of a normal model:

The Wilks statistic is

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

See Worksheet L3 for a derivation of the Wilks statistic directly from the log-likelihood function.

Note this is the same as the squared Wald statistic discussed in Example 4.5.

Comparing with Example 2.7 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 5.3. Wilks log-likelihood ratio statistic for the categorical distribution:

The Wilks log-likelihood ratio is

$$W(\boldsymbol{p}_0) = 2(l_n(\hat{\pi}_1^{ML}, \ldots, \hat{\pi}_{K-1}^{ML}) - l_n(p_1^0, \ldots, p_{K-1}^0))$$

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

$$l_n(\hat{\pi}_1^{ML}, \ldots, \hat{\pi}_{K-1}^{ML}) = n \sum_{k=1}^{K} \bar{x}_k \log \hat{\pi}_k^{ML} = n \sum_{k=1}^{K} \bar{x}_k \log \bar{x}_k$$

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

$$l_n(p_1^0, \ldots, p_{K-1}^0) = n \sum_{k=1}^{K} \bar{x}_k \log p_k^0$$

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 $n_k^{\text{expect}} = np_k^0$ we can write the Wilks statistic respectively the G-statistic as follows:

$$W(\boldsymbol{p}_0) = 2 \sum_{k=1}^{K} n_k \log \left(\frac{n_k}{n_k^{\text{expect}}} \right)$$

5.1.3 Quadratic approximation of the Wilks statistic

Recall the *quadratic approximation* of the log-likelihood function $l_n(\boldsymbol{\theta}_0|D)$ (= second order Taylor series around the MLE $\hat{\boldsymbol{\theta}}_{ML}$):

$$l_n(\boldsymbol{\theta}_0|D) \approx l_n(\hat{\boldsymbol{\theta}}_{ML}|D) - \frac{1}{2}(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}_{ML})^T \boldsymbol{J}_n(\hat{\boldsymbol{\theta}}_{ML})(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}_{ML})$$

With this we can then approximate the Wilks statistic:

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

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.

5.1.4 Examples of quadratic approximations

Example 5.4. Quadratic approximation of the Wilks statistic for a proportion (continued from Example 5.1):

A Taylor series of second order (for p_0 around $\bar{x}$) yields

$$\log\left(\frac{\bar{x}}{p_0}\right) \approx -\frac{p_0 - \bar{x}}{\bar{x}} + \frac{(p_0 - \bar{x})^2}{2\bar{x}^2}$$

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 this 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 4.4.

Example 5.5. Quadratic approximation of the Wilks statistic for the mean parameter of a normal model (continued from Example 5.2):

The normal log-likelihood is already quadratic in the mean parameter (cf. Example 3.2). 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 5.6. 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).

5.1.5 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**.

5.1.6 Cutoff values for the likelihood CI

coverage κ	$\Delta = \frac{c_{\text{chisq}}}{2}$ $(m = 1)$
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 — note that $2\Delta = c_{\text{chisq}}$ where c_{chisq} is the critical value for a specified coverage κ. This yields the above table for scalar parameter

Essentially, this means that to calibrate the interval we simply compare it with a normal CI.

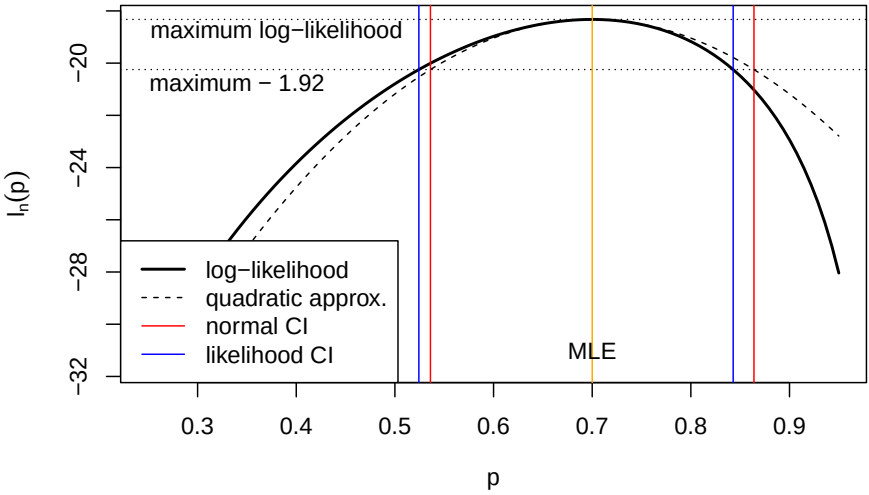
Example 5.7. Likelihood confidence interval for a proportion:

We continue from Example 5.1, and as in Example 4.7 we asssume we have data with $n = 30$ and $\bar{x} = 0.7$.

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 4.7.

The following figure illustrate the relationship between the normal CI, the likelihood CI and also shows the role of the quadratic approximation (see also Example 4.1). 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!



5.1.7 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.

A statistic equivalent to $W(\boldsymbol{\theta}_0)$ is the **likelihood ratio**

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

The two statistics can be transformed into each other by $W(\boldsymbol{\theta}_0) = -2\log\Lambda(\boldsymbol{\theta}_0)$ and $\Lambda(\boldsymbol{\theta}_0) = e^{-W(\boldsymbol{\theta}_0)/2}$. We **reject** H_0 for **small values of** Λ.

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 sigificance level) it will maximise the power (=1- type II error, probability of correctly accepting H_1). This is known as the **Neyman-Pearson theorem**.

Example 5.8. Likelihood test for a proportion:

We continue from Example 5.7 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 4.8.

5.1.8 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

$$2n(D_{\mathrm{KL}}(F, G_B) - D_{\mathrm{KL}}(F, G_A)) \approx 2(l_n(\boldsymbol{\theta}_A|D) - l_n(\boldsymbol{\theta}_B|D))$$

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

$$W(\boldsymbol{\theta}_0) = 2(l_n(\hat{\boldsymbol{\theta}}_{ML}|D) - l_n(\boldsymbol{\theta}_0|D))$$

thus compares the best-fit distribution with $\hat{\boldsymbol{\theta}}_{ML}$ as the parameter to the distribution with parameter $\boldsymbol{\theta}_0$.

For some specific models 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 is the case for the examples 5.1 and 5.2 and also more generally for exponential family models, but it is not true in general.

5.2 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 restricted model space
$H_1 : \boldsymbol{\theta} \in \omega_1 = \Omega \setminus \omega_0$	True model is not the restricted model 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

$$W = 2\log\left(\frac{L(\hat{\theta}_{ML}|D)}{L(\hat{\theta}^0_{ML}|D)}\right) \text{ and } \Lambda = \frac{\max\limits_{\theta\in\omega_0} L(\theta|D)}{\max\limits_{\theta\in\Omega} L(\theta|D)}$$

where $L(\hat{\theta}_{ML}|D)$ is the maximised likelihood assuming the full model (with parameter space Ω) and $L(\hat{\theta}^0_{ML}|D)$ is the maximised likelihood for the restricted model (with 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 5.9. 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_0^2 (so that $x_1, \ldots, x_n \sim N(\mu_0, \sigma_0^2)$).

Ω: full model with three parameters μ_1, μ_2, σ^2.

Corresponding log-likelihood functions:

Restricted model ω_0:

$$\log L(\mu_0, \sigma_0^2 | D) = -\frac{n}{2} \log(\sigma_0^2) - \frac{1}{2\sigma_0^2} \sum_{i=1}^{n} (x_i - \mu_0)^2$$

[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

Full model Ω:

$$\begin{aligned}\log L(\mu_1, \mu_2, \sigma^2|D) &= \left(-\frac{n_1}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n_1}(x_i-\mu_1)^2\right) + \\ &\quad \left(-\frac{n_2}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=n_1+1}^{n}(x_i-\mu_2)^2\right) \\ &= -\frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\left(\sum_{i=1}^{n_1}(x_i-\mu_1)^2 + \sum_{i=n_1+1}^{n}(x_i-\mu_2)^2\right)\end{aligned}$$

Corresponding MLEs:

$$\begin{array}{lll} \omega_0: & \hat{\mu}_0 = \frac{1}{n}\sum_{i=1}^{n} x_i & \widehat{\sigma_0^2} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\mu}_0)^2 \\ \Omega: & \hat{\mu}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} x_i & \widehat{\sigma^2} = \frac{1}{n}\left\{\sum_{i=1}^{n_1}(x_i-\hat{\mu}_1)^2 + \sum_{i=n_1+1}^{n}(x_i-\hat{\mu}_2)^2\right\} \\ & \hat{\mu}_2 = \frac{1}{n_2}\sum_{i=n_1+1}^{n} x_i & \end{array}$$

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

$$t_{\mathrm{ML}} = \sqrt{n\hat{\pi}_1\hat{\pi}_2}\frac{\hat{\mu}_1 - \hat{\mu}_2}{\widehat{\sigma^2}}$$

is the two sample t-statistic based on the ML variance estimate $\widehat{\sigma^2}$ and $t_{\mathrm{UB}} = t_{ML}\sqrt{\frac{n-2}{n}}$ is the conventional two sample t-statistic based on the unbiased variance estimate $\widehat{\sigma^2}_{\mathrm{UB}} = \frac{n}{n-2}\widehat{\sigma^2}$ (see Appendix).

Corresponding maximised log-likelihood:

Restricted model:

$$\log L(\hat{\mu}_0, \widehat{\sigma_0^2}|D) = -\frac{n}{2}\log(\widehat{\sigma_0^2}) - \frac{n}{2}$$

Full model:

$$\log L(\hat{\mu}_1, \hat{\mu}_2, \widehat{\sigma^2}|D) = -\frac{n}{2}\log(\widehat{\sigma^2}) - \frac{n}{2}$$

Likelihood ratio statistic:

$$\begin{aligned} W &= 2\log\left(\frac{L(\hat{\mu}_1, \hat{\mu}_2, \widehat{\sigma^2}|D)}{L(\hat{\mu}_0, \widehat{\sigma_0^2}|D)}\right) \\ &= 2\log L\left(\hat{\mu}_1, \hat{\mu}_2, \widehat{\sigma^2}|D\right) - 2\log L\left(\hat{\mu}_0, \widehat{\sigma_0^2}|D\right) \\ &= n\log\left(\frac{\widehat{\sigma_0^2}}{\widehat{\sigma^2}}\right) \\ &= n\log\left(\frac{t_{\mathrm{ML}}^2}{n} + 1\right) \\ &= n\log\left(\frac{t_{\mathrm{UB}}^2}{n-2} + 1\right) \end{aligned}$$

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 many other commonly used familiar statistical tests and test statistics can be interpreted as GLRTs. This shows the wide applicability of this procedure.

Chapter 6

Optimality properties and conclusion

6.1 Properties of maximum likelihood encountered so far

1. MLE is a special case of relative entropy 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 $D_{\mathrm{KL}}(F_{\text{true}}, F_{\boldsymbol{\theta}})$
↓
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_{true} with parameter $\boldsymbol{\theta}_{\text{true}}$ is contained in the set of specified candidates models $F_{\boldsymbol{\theta}}$ then the MLE will converge to the true model.

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).

6.2 Summarising data and the concept of (minimal) sufficiency

6.2.1 Sufficient statistic

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

Generally, a **statistic** $T(D)$ is function of the observed data $D = \{x_1, \ldots, x_n\}$. 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 statistic $T(D)$ is said to be **sufficient** for a parameter $\boldsymbol{\theta}$ in a model if the corresponding likelihood function can be written using only $T(D)$ in the terms that involve $\boldsymbol{\theta}$ such that

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

where $h()$ and $k()$ are positive-valued functions, and or equivalently on log-scale

$$l_n(\boldsymbol{\theta}) = \log h(T(D), \boldsymbol{\theta}) + \log k(D)\,.$$

This is known as the **Fisher-Pearson factorisation**.

By construction, estimation and inference about $\boldsymbol{\theta}$ based on the factorised likelihood $L(\boldsymbol{\theta})$ is mediated through the sufficient statistic $T(D)$ and does not

require the original data D. Instead, the sufficient statistic $T(D)$ contains all the information in D required to learn about the parameter $\boldsymbol{\theta}$.

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** $T(D)$. If the MLE is not unique then it can be chosen to be function of $T(D)$. Note that **a sufficient statistic always exists** since the data D are themselves sufficient statistics, with $T(D) = D$. Furthermore, sufficient statistics are **not unique** since applying a one-to-one transformation to $T(D)$ yields another sufficient statistic.

6.2.2 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 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 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 $\mathcal{X}_t$.

6.2.3 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.

6.2.4 Example: normal distribution

Example 6.1. 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

$$l_n(\boldsymbol{\theta}) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

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

$$l_n(\boldsymbol{\theta}) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{n}{2\sigma^2}(\overline{x^2} - 2\bar{x}\mu + \mu^2)$$

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

$$l_n(\boldsymbol{\theta}) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{n}{2\sigma^2}(s^2 + (\bar{x} - \mu)^2)$$

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!

6.2.5 MLEs of parameters of an exponential family are minimal sufficient statistics

The conclusion from Example 6.1 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.

6.3 Concluding remarks on maximum likelihood

6.3.1 Remark on KL divergence

Finding the model $F_{\boldsymbol{\theta}}$ that best approximates the underlying true model F_0 is done by minimising the relative entropy $D_{\mathrm{KL}}(F_0, F_{\boldsymbol{\theta}})$. For large sample size n we may approximate F_0 by the empirical distribution $\hat{F}_0$, and minimising $D_{\mathrm{KL}}(\hat{F}_0, F_{\boldsymbol{\theta}})$ then yields the method of maximum likelihood, as discussed earlier.

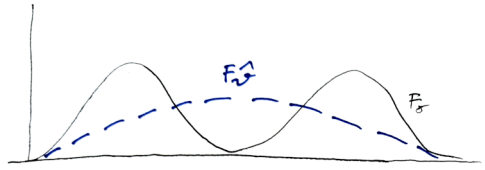
However, since the KL divergence is not symmetric there are in fact two ways to minimise the divergence between a fixed F_0 and the family $F_{\boldsymbol{\theta}}$, each with different properties:

a) **forward KL**, **approximation KL**: $\min_{\boldsymbol{\theta}} D_{\mathrm{KL}}(F_0, F_{\boldsymbol{\theta}})$

   Note that 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: $f_{\boldsymbol{\theta}}(x) > 0$ 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 $F_{\hat{\boldsymbol{\theta}}}$ will seek to cover all modes.

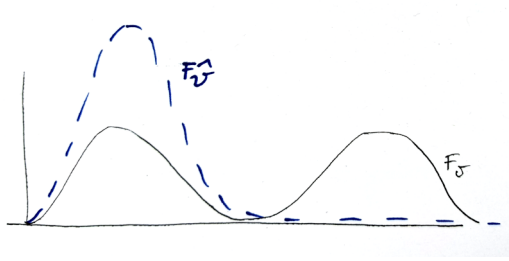


b) **reverse KL, inference KL**: $\min_{\theta} D_{\mathrm{KL}}(F_{\theta}, F_0)$

Note that 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.



Maximum likelihood is based on "forward KL", whereas Bayesian updating and Variational Bayes approximations use "reverse KL".

6.3.2 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

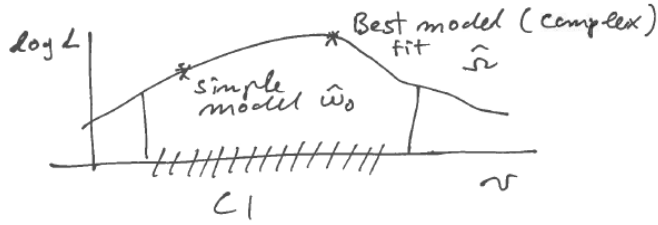
which are essentially two sides of the same coin.

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).

6.3.3 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

Chapter 7

Conditioning and Bayes rule

In this chapter we review conditional probabilities. Conditional probability is essential for Bayesian statistical modelling.

7.1 Conditional probability

Assume we have two random variables x and y with 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 x and y are given by $p(x) = \int_y p(x, y)dy$ and $p(y) = \int_x f(x, y)dx$. Thus, when computing the marginal densities a variable is removed from the joint density by integrating over all possible states of that variable. It follows also that $\int_x p(x)dx = 1$ and $\int_y p(y)dy = 1$, i.e. the marginal densities also integrate to 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, say 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

$$p(x|y = y_0) = \frac{p(x, y = y_0)}{p(y = y_0)}$$

The denominator $p(y = y_0) = \int_x p(x, y = y_0)dx$ is needed to ensure that $\int_x p(x|y = y_0)dx = 1$, thus it renormalises $p(x, y = y_0)$ so that it is a proper density.

To simplify notation, the specific value on which a variable is conditioned is often left out so we just write $p(x|y)$.

7.2 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 two random variables x and y. It thus allows to reverse the ordering 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.

7.3 Conditional mean and variance

The mean $\mathrm{E}(x|y)$ and variance $\mathrm{Var}(x|y)$ of the conditional distribution with density $p(x|y)$ are called **conditional mean** and **conditional variance**.

The **law of total expectation** states that

$$\mathrm{E}(x) = \mathrm{E}(\mathrm{E}(x|y))$$

The **law of total variance** states that

$$\mathrm{Var}(x) = \mathrm{Var}(\mathrm{E}(x|y)) + \mathrm{E}(\mathrm{Var}(x|y))$$

The first term is the "explained" or "between-group" variance, and the second the "unexplained" or "mean within group" variance.

Example 7.1. Mean and variance of a mixture model:

Assume K groups indicated by a discrete variable $y = 1, 2, \ldots, K$ 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 $\sum_{y=1}^{K} \pi_y p(x|y)$. This is called a mixture model.

The total mean $\mathrm{E}(x) = \mu_0$ is equal to $\sum_{y=1}^{K} \pi_y \mu_y$.

The total variance $\mathrm{Var}(x) = \sigma_0^2$ is equal to

$$\sum_{y=1}^{K} \pi_y (\mu_y - \mu_0)^2 + \sum_{y=1}^{K} \pi_y \sigma_y^2$$

7.4 Conditional entropy and entropy chain rules

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.

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.

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** or **conditional relative entropy** 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 relative entropy can also be computed from the conditional (cross-)entropies by

$$D_{\mathrm{KL}}(Q_{y|x}, P_{y|x}) = H(Q_{y|x}, P_{y|x}) - H(Q_{y|x})$$

The above decompositions for the entropy, the cross-entropy and relative entropy are known as **entropy chain rules**.

7.5 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.

Chapter 8

Models with latent variables and missing data

8.1 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**

$$l_n(\boldsymbol{\theta}|D_x, D_y) = \sum_{i=1}^{n} \log p(x_i, y_i|\boldsymbol{\theta})$$

to estimate $\boldsymbol{\theta}$. Recall that

$$l_n(\boldsymbol{\theta}|D_x, D_y) = -nH(\hat{Q}_{x,y}, P_{x,y|\boldsymbol{\theta}})$$

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**

$$l_n(\boldsymbol{\theta}|D_x) = \sum_{i=1}^{n} \log p(x_i|\boldsymbol{\theta}) = \sum_{i=1}^{n} \log \int_y p(x_i, y|\boldsymbol{\theta})dy$$

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

$$l_n(\boldsymbol{\theta}|D_x) = -nH(\hat{Q}_x, P_{x|\boldsymbol{\theta}})$$

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 8.1. 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

$$l_n(\boldsymbol{\theta}|D_x, D_y) = \sum_{i=1}^{n} \log \pi_{y_i} + \sum_{i=1}^{n} \log N(x_i|\mu_{y_i}, \sigma_{y_i})$$

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

$$l_n(\boldsymbol{\theta}|D_x) = \sum_{i=1}^{n} \log \sum_{y=1}^{2} \pi_y N(x|\mu_y, \sigma_y^2)$$

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 8.2. Alternative computation of the observed data likelihood:

An alternative way to arrive at the observed data likelihood is to marginalise the complete data likelihood.

$$L_n(\boldsymbol{\theta}|D_x, D_y) = \prod_{i=1}^{n} p(x_i, y_i|\boldsymbol{\theta})$$

and

$$L_n(\boldsymbol{\theta}|D_x) = \int_{y_1,\ldots,y_n} \prod_{i=1}^{n} p(x_i, y_i|\boldsymbol{\theta}) dy_1 \ldots dy_n$$

The integration (sum) and the multiplication can be interchanged as per Generalised Distributive Law leading to

$$L_n(\boldsymbol{\theta}|D_x) = \prod_{i=1}^{n} \int_{y} p(x_i, y|\boldsymbol{\theta}) dy$$

which is the same as constructing the likelihood from the marginal density.

8.2 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 8.3. Latent states of two group normal mixture model:

Continuing from Example 8.1 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

$$p(y_i|x_i, \hat{\boldsymbol{\theta}}) = \frac{\hat{\pi}_{y_i} N(x_i|\hat{\mu}_{y_i}, \widehat{\sigma_{y_i}^2})}{\hat{p}N(x_i|\hat{\mu}_1, \widehat{\sigma_1^2}) + (1-\hat{p})N(x_i|\hat{\mu}_2, \widehat{\sigma_2^2})}$$

8.3 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 $D_{\mathrm{KL}}(\hat{Q}_x Q_{y|x}, P_{x,y}^{\boldsymbol{\theta}})$ iteratively

1) with regard to $Q_{y|x}$ ("E" step) and
2) with regard to the parameters $\boldsymbol{\theta}$ of $P_{x,y}^{\boldsymbol{\theta}}$ ("M" step")

one minimises $D_{\mathrm{KL}}(\hat{Q}_x, P_x^{\boldsymbol{\theta}})$ with regard to the parameters of $P_x^{\boldsymbol{\theta}}$.

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, in the above we see that we can learn about unobservable states by means of Bayes theorem. By extending this same principle to learning about parameters and models we arrive at Bayesian learning.

Chapter 9

Essentials of Bayesian statistics

9.1 Principle of Bayesian learning

9.1.1 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 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**.

$$\underbrace{p(\boldsymbol{\theta}|x)}_{\text{posterior}} = \underbrace{p(\boldsymbol{\theta})}_{\text{prior}} \frac{p(x|\boldsymbol{\theta})}{p(x)}$$

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**.

9.1.2 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.

9.1.3 Bayesian update and likelihood

After independent and identically distributed 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 $p(D) = \int_{\boldsymbol{\theta}} p(\boldsymbol{\theta})L(\boldsymbol{\theta}|D)d\boldsymbol{\theta}$ with $\boldsymbol{\theta}$ integrated out.

The marginal likelihood serves as a standardising factor so that the posterior density for $\boldsymbol{\theta}$ integrates to 1:

$$\int_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|D)d\boldsymbol{\theta} = \frac{1}{p(D)} \int_{\boldsymbol{\theta}} p(\boldsymbol{\theta})L(\boldsymbol{\theta}|D)d\boldsymbol{\theta} = 1$$

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).

9.1.4 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 $p(x_2|x_1) = \int_{\boldsymbol{\theta}} p(x_2|\boldsymbol{\theta})p(\boldsymbol{\theta}|x_1)d\boldsymbol{\theta}$. 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). It is straightforward to understand why the probability of the new x_i depends on the previously observed data points — because the uncertainty about the model parameter $\boldsymbol{\theta}$ depends on how much data we have already observed. 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 $p(D|\boldsymbol{\theta}) = \prod_{i=1}^{n} p(x_i|\boldsymbol{\theta})$.

9.1.5 Summaries of posterior distributions and credible intervals

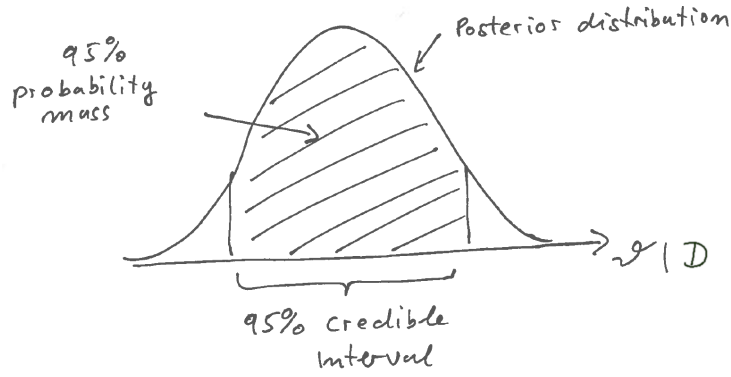
The Bayesian estimate is the full complete posterior distribution!

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.



Bayesian credible intervals (unlike their frequentist confidence counterparts) are thus very easy to interpret - they simply correspond to the area in the parameter space in which the 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.

9.1.6 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 practise, 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 Aesara/JAX as backend, NumPyro using JAX as backend, TensorFlow Probability on JAX using JAX as backend, PyMC3 using Theano 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.

9.2 Some background on Bayesian statistics

9.2.1 Bayesian interpretation of probability

9.2.1.1 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).

9.2.1.2 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.

9.2.1.3 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 = frequency (of an event in a long-running series of identically repeated experiments)

This is the *ontological view* of probability (i.e. probability "exists" and is identical to something that can be observed.).

It is also a very restrictive view of probability. 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!

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 rather than something that is an inherent property of an object.

Note that this 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.

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).

9.2.2 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.
- 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]

[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.

[3] Fienberg, S. E. 2006. *When did Bayesian inference become "Bayesian"?* Bayesian Analysis **1**:1–40. https://doi.org/10.1214/06-BA101

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)
- 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.

Chapter 10

Bayesian learning in practise

In this chapter we discuss how three basic problems, namely how to estimate a proportion, the mean and the variance in a Bayesian framework.

10.1 Estimating a proportion using the beta-binomial model

10.1.1 Binomial likelihood

In order to apply Bayes' theorem we first need to find a suitable likelihood. We use the Bernoulli model as in Example 3.1:

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 3.1 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.

10.1.2 Beta prior distribution

In Bayesian statistics we need not only to specify the data generating process but also a prior distribution over the parameters of the likelihood function.

Therefore, we need to **explicitly specify our prior uncertainty about θ.**

The parameter θ has support $[0, 1]$. Therefore we may use a **beta distribution Beta**(α_1, α_2) **as prior for** θ (see the Appendix for properties of this distribution). 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

$$p(\theta) = \frac{1}{B(\alpha_1, \alpha_2)} \theta^{\alpha_1 - 1}(1-\theta)^{\alpha_2 - 1}$$

In terms of mean parameterisation Beta(μ_0, k_0) this corresponds to:

- The prior concentration parameter is set to $k_0 = \alpha_1 + \alpha_2$
- The prior mean parameter is set to $\mu_0 = \alpha_1/k_0$.

The prior mean is therefore

$$\mathrm{E}(\theta) = \mu_0$$

and the prior variance

$$\mathrm{Var}(\theta) = \frac{\mu_0(1-\mu_0)}{k_0+1}$$

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".

10.1.3 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:

$$\theta|D \sim \mathrm{Beta}(\alpha_1 + n_2, \alpha_2 + n_2)$$

with density

$$p(\theta|D) = \frac{1}{B(\alpha_1+n_1, \alpha_2+n_2)}\theta^{\alpha_1+n_1-1}(1-\theta)^{\alpha_2+n_2-1}$$

(For a proof see Worksheet B1.)

In the corresponding mean parameterisation $\mathrm{Beta}(\mu_1, k_1)$ this results in the following updates:

- The concentration parameter is updated to $k_1 = k_0 + n$
- The mean parameter 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

$$\mathrm{Var}(\theta|D) = \frac{\mu_1(1-\mu_1)}{k_1+1}$$

10.2 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.

10.2.1 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.

10.2.2 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 parameterised 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

[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

10.2.3 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:

$$\alpha_1 \rightarrow \alpha_1 + n_1 = \alpha_1 + n\hat{\theta}_{ML}$$

$$\alpha_2 \rightarrow \alpha_2 + n_2 = \alpha_2 + n(1-\hat{\theta}_{ML})$$

Thus, conjugate prior distributions are very convenient choices. However, in their application it must be ensured that the prior distribution is flexible 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).

10.2.4 Large sample limits of mean and variance

If n is large and $n >> \alpha, \beta$ then $\lambda \rightarrow 0$ and hence the posterior mean and variance become asympotically

$$\mathrm{E}(\theta|D) \stackrel{a}{=} \frac{n_1}{n} = \hat{\theta}_{ML}$$

and

$$\text{Var}(\theta|D) \overset{a}{=} \frac{\hat{\theta}_{ML}(1-\hat{\theta}_{ML})}{n}$$

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!

10.2.5 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:

$$\text{for large } n\text{: } p(\boldsymbol{\theta}|D) \rightarrow N(\hat{\boldsymbol{\theta}}_{ML}, \text{Var}(\hat{\boldsymbol{\theta}}_{ML}))$$

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 as by Laplace.

In the Worksheet B1 the asymptotic convergence of the posterior distribution to a normal distribution is demonstrated graphically.

10.2.6 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

$$\text{E}(\theta|D) = \frac{\alpha_1 + n_1}{k_1} = \hat{\theta}_{\text{Bayes}}$$

along with the posterior variance

$$\text{Var}(\theta|D) = \frac{\hat{\theta}_{\text{Bayes}}(1-\hat{\theta}_{\text{Bayes}})}{k_1 + 1}$$

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.

10.3 Estimating the mean using the normal-normal model

10.3.1 Normal likelihood

As in Example 3.2 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 $D = \{x_1, \ldots x_n\}$. 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.

10.3.2 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

$$\text{Var}(\mu) = \frac{\sigma^2}{k_0}$$

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.

10.3.3 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

$$\text{E}(\mu|D) = \mu_1$$

and the posterior variance is

$$\text{Var}(\mu|D) = \frac{\sigma^2}{k_1}$$

10.3.4 Large sample asymptotics

For n large and $n >> k_0$ the shrinkage intensity $\lambda \to 0$ and and $k_1 \to n$. As a result

$$\text{E}(\mu|D) \overset{a}{=} \hat{\mu}_{ML}$$

$$\text{Var}(\mu|D) \overset{a}{=} \frac{\sigma^2}{n}$$

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 $\frac{\sigma^2}{k_0}$.

10.4 Estimating the variance using the inverse-gamma-normal model

10.4.1 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.

10.4.2 IG prior distribution

To model the uncertainty about the variance we use the inverse-gamma (IG) distribution, also known as inverse Wishart (IW) distribution (see Appendix for details of this distribution). The IG distribution is conjugate for the variance parameter in the normal likelihood, hence both the prior and the posterior distribution are IG.
As we use the Wishart parameterisation we may equally well call this an inverse Wishart (IW) prior, and the whole model IW-normal model.

Specifically, as prior distribution for σ^2 we assume using the mean parameter μ and concentration parameter κ:

$$\sigma^2 \sim W_1^{-1}(\psi = \kappa_0\sigma_0^2, \nu = \kappa_0 + 2)$$

- The prior concentration parameter is set to κ_0
- The prior mean parameter is set to σ_0^2

The corresponding prior mean is

$$\mathrm{E}(\sigma^2) = \sigma_0^2$$

and the prior variance is

$$\mathrm{Var}(\sigma^2) = \frac{2\sigma_0^4}{\kappa_0 - 2}$$

(note that $\kappa_0 > 2$ for the variance to exist)

10.4.3 IG posterior distribution

After observing $D = \{x_1 \ldots, x_n\}$ the posterior distribution is also IG with updated parameters:

$$\sigma^2|D \sim W_1^{-1}(\psi = \kappa_1\sigma_1^2, \nu = \kappa_1 + 2)$$

- 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

$$\mathrm{E}(\sigma^2|D) = \sigma_1^2$$

and the posterior variance

$$\mathrm{Var}(\sigma^2|D) = \frac{2\sigma_1^4}{\kappa_1 - 2}$$

10.4.4 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

$$\mathrm{E}(\sigma^2|D) \stackrel{a}{=} \widehat{\sigma^2}_{ML}$$

and the asymptotic posterior variance

$$\mathrm{Var}(\sigma^2|D) \stackrel{a}{=} \frac{2(\widehat{\sigma^2}_{ML})^2}{n}$$

Thus we recover the MLE of σ^2 and its asymptotic variance.

10.4.5 Other equivalent update rules

Above the update rule from prior to posterior inverse gamma distribution is stated for the mean parameterisation:

- $\kappa_0 \to \kappa_1 = \kappa_0 + n$
- $\sigma_0^2 \to \sigma_1^2 = \lambda\sigma_0^2 + (1 - \lambda)\widehat{\sigma^2}_{ML}$ with $\lambda = \frac{\kappa_0}{\kappa_1}$

This has the advantage that the mean of the inverse gamma 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 parameterisations. In terms of the conventional parameters α and β of the inverse gamma 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 ν of the univariate inverse Wishart distribution the update rule is

- $\nu_0 \rightarrow \nu_1 = \nu_0 + n$
- $\psi_0 \rightarrow \psi_1 = \psi_0 + n\widehat{\sigma^2}_{ML} = \psi_0 + \sum_{i=1}^{n}(x_i - \mu)^2$

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 B3 for proof of equivalence of all above update rules.)

10.5 Estimating the precision using the gamma-normal model

10.5.1 MLE of the precision

Instead of estimating the variance σ^2 we may wish to estimate the precision $w1/\sigma^2$, i.e. the inverse of 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). As a consequence, as seen below, the posterior mean of w cannot be written as linear adjustment of the MLE.

10.5.2 Gamma (Wishart) prior

For modelling the variance we have used an inverse gamma (inverse Wishart) distribution for the prior and posterior distributions. Thus, in order to model the precision we therefore now use a gamma (Wishart) distribution.

Specifically, we use the Wishart distribution in the mean parameterisation (see Appendix):

$$w \sim W_1(s^2 = w_0/k_0, k = k_0)$$

- The prior concentration parameter is set to k_0
- The prior mean parameter is set to w_0

The corresponding prior mean is

$$\mathrm{E}(w) = w_0$$

and the prior variance is

$$\mathrm{Var}(\sigma^2) = 2w_0^2/k_0$$

10.5.3 Gamma / Wishart posterior

After observing $D = \{x_1 \ldots, x_n\}$ the posterior distribution is also gamma resp. Wishart with updated parameters:

$$w|D \sim W_1(s^2 = w_1/k_1, k = k_1)$$

- The posterior concentration parameter is updated to $k_1 = k_0 + n$
- The posterior mean parameter update follows the update:

$$\frac{1}{w_1} = \lambda \frac{1}{w_0} + (1 - \lambda)\frac{1}{\hat{w}_{ML}}$$

  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.
- Equivalent update rules are for the inverse scale parameter s^2

$$\frac{1}{s_1^2} = \frac{1}{s_0^2} + n\widehat{\sigma^2}_{ML}$$

  and for the rate parameter $\beta = 1/(2s^2)$ of the gamma distribution

$$\beta_1 = \beta_0 + \frac{n}{2}\widehat{\sigma^2}_{ML}$$

  This is the form you will find most often in textbooks.

The posterior mean is

$$\mathrm{E}(w|D) = w_1$$

and the posterior variance

$$\mathrm{Var}(w|D) = 2w_1^2/k_1$$

Chapter 11

Bayesian model comparison

11.1 Marginal likelihood as model likelihood

11.1.1 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 11.1. 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 B3). This is an example of a compound probability distribution.

11.1.2 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

11.1.3 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.

11.2 The Bayes factor for comparing two models

11.2.1 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.

11.2.2 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

$$\text{log-posterior odds} = \text{weight of evidence} + \text{log-prior odds}$$

11.2.3 Scale for the Bayes factor

Following Harold Jeffreys (1961) [1] one may interpret the strength of the Bayes factor as follows:

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

More recently, Kass and Raftery (1995) [2] proposed to use the following slightly modified scale:

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

11.2.4 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

[1] Jeffreys, H. *Theory of Probability*. 3rd ed. Oxford University Press.

[2] Kass, R.E., and A.E. Raftery. 1995. *Bayes factors*. JASA **90**:773–795. https://doi.org/10.1080/01621459.1995.10476572

11.3 Approximate computations

The marginal likelihood and the Bayes factor can be difficult to compute in practise. 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.

11.3.1 Schwarz (1978) approximation of log-marginal likelihood

The logarithm of the marginal likelihood of a model can be approximated following Schwarz (1978) [3] as follow:

$$\log p(D|M) \approx l_n^M(\hat{\boldsymbol{\theta}}_{ML}^M) - \frac{1}{2} d_M \log n$$

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.

11.3.2 Bayesian information criterion (BIC)

The BIC (Bayesian information criterion) of the model M is the approximated log-marginal likelihood times the factor -2:

$$BIC(M) = -2 l_n^M(\hat{\boldsymbol{\theta}}_{ML}^M) + d_M \log n$$

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

11.3.3 Approximating the weight of evidence (log-Bayes factor) with BIC

Using BIC (twice) the log-Bayes factor can be approximated as

$$\begin{aligned} 2 \log B_{12} &\approx -BIC(M_1) + BIC(M_2) \\ &= 2\left(l_n^{M_1}(\hat{\boldsymbol{\theta}}_{ML}^{M_1}) - l_n^{M_2}(\hat{\boldsymbol{\theta}}_{ML}^{M_2})\right) - \log(n)(d_{M_1} - d_{M_2}) \end{aligned}$$

i.e. it is the penalised log-likelihood ratio of model M_1 vs. M_2.

11.4 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.

11.4.1 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!

11.4.2 Test errors

11.4.2.1 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"

11.4.2.2 Specificity and Sensitivity

From counts of TP, TN, FN, FP we can derive further quantities:

- True Negative Rate TNR, **specificity**: $TNR = \frac{TN}{TN+FP} = 1 - FPR$ with FPR=False Positive Rate = $1 - \alpha_I$
- True Positive Rate TPR, **sensitivity**, **power**, recall: $TPR = \frac{TP}{TP+FN} = 1-FNR$ with FNR=False negative rate = $1 - \alpha_{II}$
- 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.

11.4.2.3 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).

11.4.3 Bayesian perspective

11.4.3.1 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).

11.4.3.2 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.

11.4.3.3 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).

11.4.4 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.

Chapter 12

Choosing priors in Bayesian analysis

12.1 Choosing a prior

12.1.1 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).

12.1.2 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!

12.2 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.

12.2.1 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:

$$p(\boldsymbol{\theta}) \propto \sqrt{\det \boldsymbol{I}^{\text{Fisher}}(\boldsymbol{\theta})}$$

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 $p(\mu) = 1$.

[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.

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.

12.2.2 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!).

12.3 Empirical Bayes

In empirical Bayes the data analysist specifies a family of prior distribution (say a Beta distribution with free 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.

12.3.1 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:

$$p(D|\lambda) = \int_{\boldsymbol{\theta}} p(D|\boldsymbol{\theta})p(\boldsymbol{\theta}|\lambda)d\boldsymbol{\theta}$$

[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

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".

12.3.2 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 $\lambda = \frac{k_0}{k_0}$ determined by the ration 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 12.1. 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.

Chapter 13

Optimality properties and summary

13.1 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..

13.1.1 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.

13.1.2 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.

13.1.3 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!

$\rightarrow$ 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.

13.2 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.

13.3 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). Intringuingly, 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 relative entropy 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.

13.4 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.

Appendix

Appendix A

Refresher

Statistics is a mathematical science that requires practical use of tools from probability, vector and matrices, analysis etc.

Here we briefly list some essentials that are needed for MATH27720 Statistics 2. Please familiarise yourself (again) with these topics.

A.1 Basic mathematical notation

Summation:

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \ldots + x_n$$

Multiplication:

$$\prod_{i=1}^{n} x_i = x_1 \times x_2 \times \ldots \times x_n$$

Indicator function:

$$1_A = \begin{cases} 1 & \text{if } A \text{ is true} \\ 0 & \text{if } A \text{ is not true} \end{cases}$$

A.2 Vectors and matrices

Vector and matrix notation.

Vector algebra.

Eigenvectors and eigenvalues for a real symmetric matrix.

Eigenvalue (spectral) decomposition of a real symmetric matrix.

Positive and negative definiteness of a real symmetric matrix (containing only positive or only negative eigenvalues).

Singularity of a real symmetric matrix (containing one or more eigenvalues identical to zero).

Singular value decomposition of a real matrix.

Inverse of a matrix.

Trace and determinant of a square matrix.

Connection with eigenvalues (trace = sum of eigenvalues, determinant = product of eigenvalues).

A.3 Functions

A.3.1 Gradient

The **gradient** of a scalar-valued function $h(\boldsymbol{x})$ with vector argument $\boldsymbol{x} = (x_1, \ldots, x_d)^T$ is the vector containing the first order partial derivatives of $h(\boldsymbol{x})$ with regard to each $x_1, \ldots, x_d$:

$$\begin{aligned}\nabla h(\boldsymbol{x}) &= \begin{pmatrix} \frac{\partial h(\boldsymbol{x})}{\partial x_1} \\ \vdots \\ \frac{\partial h(\boldsymbol{x})}{\partial x_d} \end{pmatrix} \\ &= \frac{\partial h(\boldsymbol{x})}{\partial \boldsymbol{x}} \\ &= \text{grad } h(\boldsymbol{x})\end{aligned}$$

The symbol ∇ is called the **nabla operator** (also known as **del operator**).

Note that we write the gradient as a **column vector**. This is called the **denominator layout** convention, see https://en.wikipedia.org/wiki/Matrix_calculus for details. In contrast, many textbooks (and also earlier versions of these lecture notes) assume that gradients are row vectors, following the so-called numerator layout convention.

Example A.1. Examples for the gradient:

- $h(\boldsymbol{x}) = \boldsymbol{a}^T\boldsymbol{x} + b$. Then $\nabla h(\boldsymbol{x}) = \frac{\partial h(\boldsymbol{x})}{\partial \boldsymbol{x}} = \boldsymbol{a}$.
- $h(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{x}$. Then $\nabla h(\boldsymbol{x}) = \frac{\partial h(\boldsymbol{x})}{\partial \boldsymbol{x}} = 2\boldsymbol{x}$.
- $h(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}$. Then $\nabla h(\boldsymbol{x}) = \frac{\partial h(\boldsymbol{x})}{\partial \boldsymbol{x}} = (\boldsymbol{A} + \boldsymbol{A}^T)\boldsymbol{x}$.

A.3.2 Hessian matrix

The matrix of all second order partial derivates of scalar-valued function with vector-valued argument is called the **Hessian matrix**:

$$\begin{aligned}\nabla\nabla^T h(\boldsymbol{x}) &= \begin{pmatrix} \frac{\partial^2 h(\boldsymbol{x})}{\partial x_1^2} & \frac{\partial^2 h(\boldsymbol{x})}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 h(\boldsymbol{x})}{\partial x_1 \partial x_d} \\ \frac{\partial^2 h(\boldsymbol{x})}{\partial x_2 \partial x_1} & \frac{\partial^2 h(\boldsymbol{x})}{\partial x_2^2} & \cdots & \frac{\partial^2 h(\boldsymbol{x})}{\partial x_2 \partial x_d} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 h(\boldsymbol{x})}{\partial x_d \partial x_1} & \frac{\partial^2 h(\boldsymbol{x})}{\partial x_d \partial x_2} & \cdots & \frac{\partial^2 h(\boldsymbol{x})}{\partial x_d^2} \end{pmatrix} \\ &= \left(\frac{\partial h(\boldsymbol{x})}{\partial x_i \partial x_j} \right) \\ &= \frac{\partial^2 h(\boldsymbol{x})}{\partial \boldsymbol{x} \partial \boldsymbol{x}^T}\end{aligned}$$

By construction the Hessian matrix is square and symmetric.

Example A.2. $h(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x}$. Then $\nabla\nabla^T h(\boldsymbol{x}) = \frac{\partial^2 h(\boldsymbol{x})}{\partial \boldsymbol{x} \partial \boldsymbol{x}^T} = (\boldsymbol{A} + \boldsymbol{A}^T)$.

A.3.3 Conditions for a local extremum of a function

To check if x_0 or $\boldsymbol{x}_0$ is a local extremum, i.e. a local maximum or a local minimum, of a differentiable function $h(x)$ or $h(\boldsymbol{x})$ we can use the following conditions:

For a function of a single variable:

i) First derivative is zero at the extremum: $h'(x_0) = 0$.
ii) If the second derivative $h''(x_0) < 0$ at the extremum is negative then it is a maximum.
iii) If the second derivative $h''(x_0) > 0$ at the extremum is positive it is a minimum.

Note that conditions ii) and iii) are sufficient but not necessary. For a minimum, it is necessary that the second derivative is non-negative, and for a maximum that the second derivative is non-positive. As example, x^4 has a minimum at $x_0 = 0$ with vanishing first derivative and also a vanishing second derivative.

For a function of several variables:

i) Gradient vanishes at extremum: $\nabla h(\boldsymbol{x}_0) = 0$.
ii) If the Hessian $\nabla\nabla^T h(\boldsymbol{x}_0)$ is negative definite (= all eigenvalues of Hessian matrix are negative) then the extremum is a maximum.
iii) If the Hessian is positive definite (= all eigenvalues of Hessian matrix are positive) then the extremum is a minimum.

Again, conditions ii) and iii) are sufficient but not necessary. For a minimum it is necessary that the Hessian is positive semi-definite, and for a maximum that the Hessian is negative semi-definite.

A.3.4 Linear and quadratic approximation

A linear and quadratic approximation of a differentiable function is given by a Taylor series of first and second order, respectively.

Applied to a scalar-valued function of a scalar:

$$h(x) \approx h(x_0) + h'(x_0)(x - x_0) + \frac{1}{2}h''(x_0)(x - x_0)^2$$

Note that $h'(x_0) = h'(x) \,|\, x_0$ is first derivative of $h(x)$ evaluated at x_0 and $h''(x_0) = h''(x) \,|\, x_0$ is the second derivative of $h(x)$ evaluated x_0.

With $x = x_0 + \varepsilon$ the approximation can also be written as

$$h(x_0 + \varepsilon) \approx h(x_0) + h'(x_0)\,\varepsilon + \frac{1}{2}h''(x_0)\,\varepsilon^2$$

Applied to a scalar-valued function of a vector:

$$h(\boldsymbol{x}) \approx h(\boldsymbol{x}_0) + \nabla h(\boldsymbol{x}_0)^T(\boldsymbol{x} - \boldsymbol{x}_0) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}_0)^T\,\nabla\nabla^T h(\boldsymbol{x}_0)\,(\boldsymbol{x} - \boldsymbol{x}_0)$$

Note that $\nabla h(\boldsymbol{x}_0)$ is the gradient of $h(\boldsymbol{x})$ evaluated at $\boldsymbol{x}_0$ and $\nabla\nabla^T h(\boldsymbol{x}_0)$ the Hessian matrix of $h(\boldsymbol{x})$ evaluated at $\boldsymbol{x}_0$.

With $\boldsymbol{x} = \boldsymbol{x}_0 + \boldsymbol{\varepsilon}$ this approximation can also be written as

$$h(\boldsymbol{x}_0 + \boldsymbol{\varepsilon}) \approx h(\boldsymbol{x}_0) + \nabla h(\boldsymbol{x}_0)^T\boldsymbol{\varepsilon} + \frac{1}{2}\boldsymbol{\varepsilon}^T\,\nabla\nabla^T h(\boldsymbol{x}_0)\,\boldsymbol{\varepsilon}$$

Example A.3. Commonly occurring Taylor series approximations of second order are for example

$$\log(x_0 + \varepsilon) \approx \log(x_0) + \frac{\varepsilon}{x_0} - \frac{\varepsilon^2}{2x_0^2}$$

and

$$\frac{x_0}{x_0 + \varepsilon} \approx 1 - \frac{\varepsilon}{x_0} + \frac{\varepsilon^2}{x_0^2}$$

Example A.4. Around a local extremum $\boldsymbol{x}_0$ (maximum or minimum) we can approximate the function using

$$h(\boldsymbol{x}_0 + \boldsymbol{\varepsilon}) \approx h(\boldsymbol{x}_0) + \frac{1}{2}\boldsymbol{\varepsilon}^T\nabla\nabla^T h(\boldsymbol{x}_0)\boldsymbol{\varepsilon}$$

Note this is a quadratic approximation with the linear term missing due to the gradient being zero at $\boldsymbol{x}_0$.

A.3.5 Convex and concave functions

A function $h(\boldsymbol{x})$ is **convex** if for all $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ the line segment from point $(\boldsymbol{x}_1, h(\boldsymbol{x}_1))$ to point $(\boldsymbol{x}_2, h(\boldsymbol{x}_2))$ never lies below the function. Moreover, the function is strictly convex if the line segment always lies above the curve, apart from the two end points:

$$\lambda h(\boldsymbol{x}_1) + (1-\lambda)h(\boldsymbol{x}_2) \geq h(\lambda \boldsymbol{x}_1 + (1-\lambda)\boldsymbol{x}_2)$$

for all $\lambda \in [0, 1]$.

Equivalently, a differentiable function $h(\boldsymbol{x})$ is convex (strictly convex) if for all $\boldsymbol{x}_0$ the function $h(\boldsymbol{x})$ never lies below (always lies above, except at $\boldsymbol{x}_0$) the linear approximation through the point $(\boldsymbol{x}_0, h(\boldsymbol{x}_0))$:

$$h(\boldsymbol{x}) \geq h(\boldsymbol{x}_0) + \nabla h(\boldsymbol{x}_0)^T(\boldsymbol{x} - \boldsymbol{x}_0)$$

For a convex function a vanishing gradient at $\boldsymbol{x}_0$ indicates a minimum at $\boldsymbol{x}_0$. Furthermore, any local minimum must also be a global minimum (for a differentiable function this follows directly from the last inequality). For a strictly convex function the minimum is unique so there is at most one local/global minimum in that case.

If $h(\boldsymbol{x})$ is convex, then $-h(\boldsymbol{x})$ is **concave**, and the criteria above can be adapted accordingly to check for concavity and strict concavity, as well as to identify local/global maxima.

(Strictly) convex and concave functions are convenient objective functions in optimisation as it is straightforward to find their local/global extrema, both analytically and numerically.

As the shape of a convex function resembles that of a valley, one way to memorise that fact is that a **va**lley is con**ve**x.

Example A.5. Convex functions.

This is a convex function but not a strictly convex function:

- $\max(x^2, |x|)$

The following are strictly convex functions:

- x^2,
- x^4,
- e^x,
- $x \log(x)$ for $x > 0$.

On the other hand, this is not a convex function:

- $\frac{1}{x^2}$ for all $x \neq 0$.

However, the function in last example is strictly convex if the domain is restricted to either $x > 0$ or $x < 0$.

Example A.6. Concave functions.

The following are strictly concave function:

- $-x^2$,
- $\log(x)$ for $x > 0$,
- $\sqrt{x}$ for $x > 0$.

A.4 Combinatorics

A.4.1 Number of permutations

The number of possible orderings, or permutations, of n distinct items is the number of ways to put n items in n bins with exactly one item in each bin. It is given by the factorial

$$n! = \prod_{i=1}^{n} i = 1 \times 2 \times \ldots \times n$$

where n is a positive integer. For $n = 0$ the factorial is defined as

$$0! = 1$$

as there is exactly one permutation of zero objects.

The factorial can also be obtained using the gamma function

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

which can be viewed as continuous version of the factorial with $\Gamma(x) = (x-1)!$ for any positive integer x.

A.4.2 Multinomial and binomial coefficient

The number of possible permutation of n items of K distinct types, with n_1 of type 1, n_2 of type 2 and so on, equals the number of ways to put n items into K bins with n_1 items in the first bin, n_2 in the second and so on. It is given by the **multinomial** coefficient

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

with $\sum_{k=1}^{K} n_k = n$ and $K \leq n$. Note that it equals the number of permutation of all items divided by the number of permutations of the items in each bin (or of each type).

If all $n_k = 1$ and hence $K = n$ the multinomial coefficient reduces to the factorial.

If there are only two bins / types ($K = 2$) the multinomial coefficients becomes the **binomial coefficient**

$$\binom{n}{n_1} = \binom{n}{n_1, n - n_1} = \frac{n!}{n_1!(n - n_1)!}$$

which counts the number of ways to choose n_1 elements from a set of n elements.

A.4.3 De Moivre-Sterling approximation of the factorial

The factorial is frequently approximated by the following formula derived by Abraham de Moivre (1667–1754) and James Stirling (1692-1770)

$$n! \approx \sqrt{2\pi} n^{n+\frac{1}{2}} e^{-n}$$

or equivalently on logarithmic scale

$$\log n! \approx \left(n + \frac{1}{2}\right) \log n - n + \frac{1}{2} \log (2\pi)$$

The approximation is good for small n (but fails for $n = 0$) and becomes more and more accurate with increasing n. For large n the approximation can be simplified to

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

A.5 Probability

A.5.1 Random variables

A **random variable** describes a random experiment. The set of all possible outcomes is the **sample space** or **state space** of the random variable and is denoted by $\Omega = \{\omega_1, \omega_2, \ldots\}$. The outcomes ω_i are the **elementary events**. The sample space Ω can be finite or infinite. Depending on type of outcomes the random variable is **discrete** or **continuous**.

An event $A \subseteq \Omega$ is a subset of Ω and thus itself a set composed of elementary events: $A = \{a_1, a_2, \ldots\}$. This includes as special cases the full set $A = \Omega$, the empty set $A = \emptyset$, and the elementary events $A = \omega_i$. The complementary event A^C is the complement of the set A in the set Ω so that $A^C = \Omega \backslash A = \{\omega_i \in \Omega : \omega_i \notin A\}$.

The probability of an event A is denoted by $\Pr(A)$. Essentially, to obtain this probability we need to count the elementary elements corresponding to A. To do this we assume as axioms of probability that

- $\Pr(A) \geq 0$, probabilities are positive,
- $\Pr(\Omega) = 1$, the certain event has probability 1, and
- $\Pr(A) = \sum_{a_i \in A} \Pr(a_i)$, the probability of an event equals the sum of its constituting elementary events a_i. This sum is taken over a finite or countable infinite number of elements.

This implies

- $\Pr(A) \leq 1$, i.e. probabilities all lie in the interval $[0, 1]$
- $\Pr(A^C) = 1 - \Pr(A)$, and
- $\Pr(\emptyset) = 0$

Assume now that we have two events A and B. The probability of the event "A and B" is then given by the probability of the set intersection $\Pr(A \cap B)$. Likewise the probability of the event "A or B" is given by the probability of the set union $\Pr(A \cup B)$.

From the above it is clear that the definition and theory of probability is closely linked to set theory, and in particular to measure theory. Indeed, viewing probability as a special type of measure allows for an elegant treatment of both discrete and continuous random variables (but one which is out of scope of this module).

A.5.2 Probability mass and density function, distribution function and quantile function

To describe a random variable x with state space Ω we need a way to effectively store the probabilities of the corresponding elementary outcomes $x \in \Omega$. Note that for convenience we use the same symbol to denote the random variable and its elementary outcomes.

For a discrete random variable we define the event $A = \{x : x = a\} = \{a\}$ and get the probability

$$\Pr(A) = \Pr(x = a) = f(a)$$

directly from the **probability mass function** (PMF), here denoted by lower case f (but we frequently also use p or q). The PMF has the property that $\sum_{x \in \Omega} f(x) = 1$ and that $f(x) \in [0, 1]$.

For continuous random variables we need to use a **probability density function** (PDF) instead. We define the event $A = \{x : a < x \leq a + da\}$ as an infinitesimal interval and then assign the probability

$$\Pr(A) = \Pr(a < x \leq a + da) = f(a)da\,.$$

The PDF has the property that $\int_{x \in \Omega} f(x)dx = 1$ but in contrast to a PMF the density $f(x) \geq 0$ may take on values larger than 1.

As alternative to using PMF/PDFs we may also use a **distribution function** to describe the random variable. This assumes an ordering exist among the elementary events so that we can define the event $A = \{x : x \leq a\}$ and compute its probability as

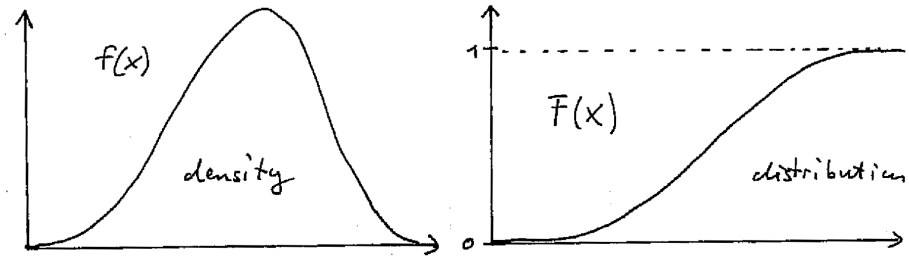
$$F(a) = \Pr(A) = \Pr(x \leq a) = \begin{cases} \sum_{x \in A} f(x) & \text{discrete case} \\ \int_{x \in A} f(x)dx & \text{continuous case} \end{cases}$$

This is also known **cumulative distribution function** (CDF) and is denoted by upper case F (or P and Q). By construction the distribution function is monotonically non-decreasing and its value ranges from 0 to 1. With its help we can compute the probability of an interval set such as

$$\Pr(a < x \leq b) = F(b) - F(a)\,.$$

The inverse of the distribution function $y = F(x)$ is the **quantile function** $x = F^{-1}(y)$. The 50% quantile $F^{-1}\left(\frac{1}{2}\right)$ is the **median**.

If the random variable x has distribution function F we write $x \sim F$.



A.5.3 Expectation of a random variable

The expected value $\mathrm{E}(x)$ of a random variable is defined as the weighted average over all possible outcomes, with the weight given by the PMF / PDF $f(x)$:

$$\mathrm{E}_F(x) = \begin{cases} \sum_{x\in\Omega} f(x)x & \text{discrete case} \\ \int_{x\in\Omega} f(x)x\,dx & \text{continuous case} \end{cases}$$

Note the notation to emphasise that the expectation is taken with regard to the distribution F. The subscript F is usually left out if there are no ambiguities. Furthermore, because the sum or integral may diverge the expectation is not necessarily always defined (in contrast to quantiles).

The expected value of a function of a random variable $h(x)$ is obtained similarly:

$$\mathrm{E}_F(h(x)) = \begin{cases} \sum_{x\in\Omega} f(x)h(x) & \text{discrete case} \\ \int_{x\in\Omega} f(x)h(x)dx & \text{continuous case} \end{cases}$$

This is called the "law of the unconscious statistician", or short LOTUS. Again, to highlight that the random variable x has distribution F we write $\mathrm{E}_F(h(x))$.

A.5.4 Jensen's inequality for the expectation

If $h(\boldsymbol{x})$ is a *convex* function then the following inequality holds:

$$\mathrm{E}(h(\boldsymbol{x})) \geq h(\mathrm{E}(\boldsymbol{x}))$$

Recall: a con**ve**x function (such as x^2) has the shape of a "**va**lley".

A.5.5 Probability as expectation

Probability itself can also be understood as an expectation. For an event A we can define a corresponding indicator function $1_{x \in A}$ for an elementary element x to be part of A. From the above it then follows

$$\mathrm{E}(1_{x \in A}) = \Pr(A)\,,$$

Interestingly, one can develop the whole theory of probability from this perspective. [1]

A.5.6 Moments and variance of a random variable

The moments of a random variable are defined as follows:

- Zeroth moment: $\mathrm{E}(x^0) = 1$ by construction of PDF and PMF,
- First moment: $\mathrm{E}(x^1) = \mathrm{E}(x) = \mu$, the mean,
- Second moment: $\mathrm{E}(x^2)$
- The variance is the second moment centred about the mean μ:

$$\mathrm{Var}(x) = \mathrm{E}((x - \mu)^2) = \sigma^2$$

- The variance can also be computed by $\mathrm{Var}(x) = \mathrm{E}(x^2) - \mathrm{E}(x)^2$. Note this is an example of Jensen's inequality, with $\mathrm{E}(x^2) = \mathrm{E}(x)^2 + \mathrm{Var}(x) \geq \mathrm{E}(x)^2$.

A distribution does not necessarily need to have any finite first or higher moments. An example is the Cauchy distribution that does not have a mean or variance (or any other higher moment).

A.5.7 Transformation of random variables

Linear transformation of random variables: if a and b are constants and x is a random variable, then the random variable $y = a + bx$ has mean $\mathrm{E}(y) = a + b\mathrm{E}(x)$ and variance $\mathrm{Var}(y) = b^2\mathrm{Var}(x)$.

For a general invertible coordinate transformation $y = h(x) = y(x)$ the backtransformation is $x = h^{-1}(y) = x(y)$.

The transformation of the infinitesimal volume element is $dy = \left|\frac{dy}{dx}\right| dx$.

[1] Whittle, P. 2000. Probability via Expectation (3rd ed.). Springer. https://doi.org/10.1007/978-1-4612-0509-8

The transformation of the density is $f_y(y) = \left|\frac{dx}{dy}\right| f_x(x(y))$.

Note that $\left|\frac{dx}{dy}\right| = \left|\frac{dy}{dx}\right|^{-1}$.

A.5.8 Random vectors and covariance matrix

Instead of scalar random variables one often also considers random vectors and also random matrices.

For a random vector $\boldsymbol{x} = (x_1, x_2, ..., x_d)^T$ the mean $\mathrm{E}(\boldsymbol{x}) = \boldsymbol{\mu}$ is simply comprised of the means of its components, i.e. $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)^T$. Thus, the mean of a random vector of dimension is a vector of of the same length.

The variance of a random vector of length d, however, is not a vector but a matrix of size $d \times d$. This matrix is called the **covariance matrix**:

$$\begin{aligned}\mathrm{Var}(\boldsymbol{x}) &= \underbrace{\boldsymbol{\Sigma}}_{d\times d} = (\sigma_{ij}) = \begin{pmatrix} \sigma_{11} & \ldots & \sigma_{1d} \\ \vdots & \ddots & \vdots \\ \sigma_{d1} & \ldots & \sigma_{dd} \end{pmatrix} \\ &= \mathrm{E}\left(\underbrace{(\boldsymbol{x}-\boldsymbol{\mu})}_{d\times 1} \underbrace{(\boldsymbol{x}-\boldsymbol{\mu})^T}_{1\times d} \right) \\ &= \mathrm{E}(\boldsymbol{x}\boldsymbol{x}^T) - \boldsymbol{\mu}\boldsymbol{\mu}^T\end{aligned}$$

The entries of the covariance matrix $\sigma_{ij} = \mathrm{Cov}(x_i, x_j)$ describe the covariance between the random variables x_i and x_j. The covariance matrix is symmetric, hence $\sigma_{ij} = \sigma_{ji}$. The diagonal entries $\sigma_{ii} = \mathrm{Cov}(x_i, x_i) = \mathrm{Var}(x_i) = \sigma_i^2$ correspond to the variances of the components of $\boldsymbol{x}$. The covariance matrix is by construction **positive semi-definite**, i.e. the eigenvalues of $\boldsymbol{\Sigma}$ are all positive or equal to zero.

However, wherever possible one will aim to use models with non-singular covariance matrices, with all eigenvalues positive, so that the covariance matrix is invertible.

For univariate x and scalar constant a the variance of ax equals $\mathrm{Var}(ax) = a^2\mathrm{Var}(x)$. For a random vector $\boldsymbol{x}$ of dimension d and constant matrix $\boldsymbol{A}$ of dimension $m \times d$ this generalises to $\mathrm{Var}(\boldsymbol{A}x) = \boldsymbol{A}\mathrm{Var}(\boldsymbol{x})\boldsymbol{A}^T$.

A.5.9 Correlation matrix

A covariance matrix can be factorised into the product

$$\boldsymbol{\Sigma} = \boldsymbol{V}^{\frac{1}{2}}\boldsymbol{P}\boldsymbol{V}^{\frac{1}{2}}$$

where $\boldsymbol{V}$ is a diagonal matrix containing the variances

$$\boldsymbol{V} = \begin{pmatrix} \sigma_{11} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_{dd} \end{pmatrix}$$

and the matrix $\boldsymbol{P}$ ("upper case rho") is the symmetric **correlation matrix**

$$\boldsymbol{P} = (\rho_{ij}) = \begin{pmatrix} 1 & \dots & \rho_{1d} \\ \vdots & \ddots & \vdots \\ \rho_{d1} & \dots & 1 \end{pmatrix} = \boldsymbol{V}^{-\frac{1}{2}} \boldsymbol{\Sigma} \boldsymbol{V}^{-\frac{1}{2}}$$

Thus, the correlation between x_i and x_j is defined as

$$\rho_{ij} = \text{Cor}(x_i, x_j) = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}$$

A.6 Statistics

A.6.1 Statistical learning

The aim in statistics — data science — machine learning is to use data (from experiments, observations, measurements) to learn about and understand the world using models. In statistics we employ probabilistic models.

Let denote data by $D = \{x_1, \dots, x_n\}$ and models by $p(x|\theta)$ where θ represents the parameters of the model. Often (but not always) θ can be interpreted as and/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.

Note that it is generally easy to find one or several models that explain the data but these then often do not predict well.

Therefore, one would like to avoid **overfitting** the data and identify models that are appropriate for the data at hand (i.e. not too simple but also not too complex).

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. Specfically a confidence interval contains all parameter values that are not significantly different from the best parameter.

A.6.2 Sampling properties of a point estimator $\hat{\theta}$

A point estimator $\hat{\theta}$ depends on the data, hence it exibits **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:

$$\begin{aligned} \text{Bias:} \quad \text{Bias}(\hat{\theta}) &= \mathrm{E}(\hat{\theta}) - \theta \\ \text{Variance:} \quad \text{Var}(\hat{\theta}) &= \mathrm{E}\left((\hat{\theta} - \mathrm{E}(\hat{\theta}))^2\right) \\ \text{Mean squared error:} \quad \text{MSE}(\hat{\theta}) &= \mathrm{E}((\hat{\theta} - \theta)^2) \\ &= \text{Var}(\hat{\theta}) + \text{Bias}(\hat{\theta})^2 \end{aligned}$$

The last identity about MSE follows from $\mathrm{E}(x^2) = \text{Var}(x) + \mathrm{E}(x)^2$.

At first sight it seems desirable to focus on unbiased (for finite n) estimators. However, requiring strict unbiasedness is not always a good idea. In many situations it is better to allow for some small bias and in order to achieve a smaller variance and an overall total 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)

A.6.3 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 but rather weak requirement for any reasonable estimator. Of all consistent estimators we typically select the estimators that are most **efficient** (i.e. with fasted decrease in MSE) and that therefore have smallest variance and/or MSE for given finite n.

Consistency implies we recover the true model in the limit of infinite data if the model class contains the true data generating model. If the model class does not contain the true model then strict consistency cannot be achieved but we still wish to get as close as possible to the true model when choosing model parameters.

A.6.4 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

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} 1_{x_i \leq x}$$

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.

Note this is effectively a variant of the **law of large numbers** applied to the whole distribution, rather than just the mean (see below).

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.6.5 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.6.6 Law of large numbers

The **law of large numbers** was discovered by Jacob Bernoulli (1655-1705) and states that the average converges to the mean.

Since $\hat{F}_n$ convergences strongly to F as $n \to \infty$ there is corresponding convergence of the average $\mathrm{E}_{\hat{F}_n}(h(x)) = \frac{1}{n}\sum_{i=1}^{n} h(x_i)$ to the expectation $\mathrm{E}_F(h(x))$.

In other words, if the mean exists then for sufficiently large n it can be substituted by the empirical mean.

Likewise, the law of large numbers can be applied to empirical estimators to show that they will converge to the corresponding true quantities for sufficiently large n.

Furthermore, one may use the law of large numbers as a **justification to interpret large-sample limits of frequencies as probabilities**. However, **the converse** , namely requesting that all probabilities must have a frequentist interpretation, **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 says nothing about the finite sample properties of estimators.

A.6.7 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 $N(\mu, \sigma^2)$ 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

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

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

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

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 $\mathrm{Var}(\widehat{\sigma^2}_{\mathrm{UB}}) > \mathrm{Var}(\widehat{\sigma^2}_{\mathrm{ML}})$ and $\mathrm{MSE}(\widehat{\sigma^2}_{\mathrm{UB}}) > \mathrm{MSE}(\widehat{\sigma^2}_{\mathrm{ML}})$ so that the biased empirical estimator has both lower variance and lower mean squared error than the unbiased estimator.

A.6.8 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 $s^2_{\mathrm{UB}} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ then one arrives at the one sample t-statistic

$$t_{\mathrm{UB}} = \frac{\bar{x} - \mu}{\sqrt{s^2_{\mathrm{UB}}/n}} \sim t_{n-1}\,.$$

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 (ML) estimate of the variance $s^2_{\mathrm{ML}} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{n-1}{n}s^2_{\mathrm{UB}}$ is used then this leads to a slightly different statistic

$$t_{\mathrm{ML}} = \frac{\bar{x} - \mu}{\sqrt{s^2_{\mathrm{ML}}/n}} = \sqrt{\frac{n}{n-1}}\,t_{\mathrm{UB}}$$

with

$$t_{\mathrm{ML}} \sim \mathrm{lst}\left(0, \tau^2 = \frac{n}{n-1}, n-1\right)$$

Thus, t_{ML} follows a location-scale t-distribution, with mean 0 for $n > 2$ and variance $n/(n-3)$ for $n > 3$.

A.6.9 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 $s^2_{\mathrm{UB}} = \frac{1}{n-2}\left(\sum_{i=1}^{n_1}(x_i - \hat{\mu}_1)^2 + \sum_{i=n_1+1}^{n}(x_i - \hat{\mu}_2)^2\right)$ (note the factor $n-2$) or by the empirical (ML) estimate $s^2_{\mathrm{ML}} = \frac{1}{n}\left(\sum_{i=1}^{n_1}(x_i - \hat{\mu}_1)^2 + \sum_{i=n_1+1}^{n}(x_i - \hat{\mu}_2)^2\right) = \frac{n-2}{n}s^2_{\mathrm{UB}}$. 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

$$t_{\mathrm{UB}} = \frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)s^2_{\mathrm{UB}}}} = \frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{\left(\frac{n}{n_1 n_2}\right)s^2_{\mathrm{UB}}}}$$

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

$$t_{\mathrm{UB}} = \sqrt{n}\frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{\left(\frac{1}{\hat{\pi}_1} + \frac{1}{\hat{\pi}_2}\right)s^2_{\mathrm{UB}}}} = \sqrt{n\hat{\pi}_1\hat{\pi}_2}\frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{s^2_{\mathrm{UB}}}}$$

The two sample t-statistic is distributed as

$$t_{\mathrm{UB}} \sim t_{n-2}$$

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

$$\begin{aligned} t_{\mathrm{ML}} &= \frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) s^2_{\mathrm{ML}}}} = \frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{\left(\frac{n}{n_1 n_2}\right) s^2_{\mathrm{ML}}}} \\ &= \sqrt{n} \frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{\left(\frac{1}{\hat{\pi}_1} + \frac{1}{\hat{\pi}_2}\right) s^2_{\mathrm{ML}}}} = \sqrt{n \hat{\pi}_1 \hat{\pi}_2} \frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{s^2_{\mathrm{ML}}}} \\ &= \sqrt{\frac{n}{n-2}} t_{\mathrm{UB}} \end{aligned}$$

with

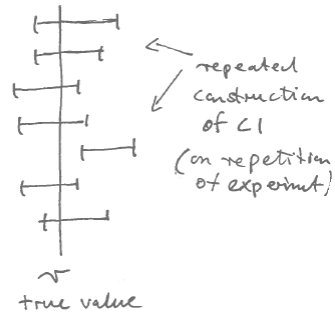
$$t_{\mathrm{ML}} \sim \mathrm{lst}\left(0, \tau^2 = \frac{n}{n-2}, n-2\right)$$

Thus, t_{ML} follows a location-scale t-distribution, with mean 0 for $n > 3$ and variance $n/(n-4)$ for $n > 4$.

A.6.10 Confidence intervals

- A **confidence** interval (CI) is an **interval estimate** with a **frequentist** interpretation.
- Definition of **coverage** κ of a CI: how often (in repeated identical experiment) does the estimated CI overlap the true parameter value θ
  - Eg.: Coverage $\kappa = 0.95$ (95%) means that in 95 out of 100 case the estimated CI will contain the (unknown) true value (i.e. it will "cover" θ).

Illustration of the repeated construction of a CI for θ:



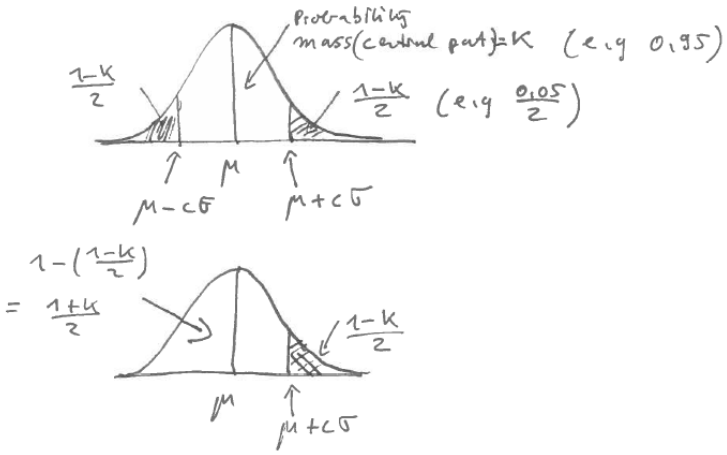
- Note that a CI is actually an **estimate**: $\widehat{\mathrm{CI}}(x_1, \ldots, x_n)$, i.e. it depends on data and has a random (sampling) variation.

- A good CI has high coverage and is compact.

Note: the coverage probability is **not** the probability that the true value is contained in a given estimated interval (that would be the Bayesian *credible* interval).

A.6.11 Symmetric normal confidence interval

For a normally distributed univariate random variable it is straightforward to construct a symmetric two-sided CI with a given desired coverage κ.



For a normal random variable $X \sim N(\mu, \sigma^2)$ with mean μ and variance σ^2 and density function $f(x)$ we can compute the probability

$$\Pr(x \leq \mu + c\sigma) = \int_{-\infty}^{\mu+c\sigma} f(x)dx = \Phi(c) = \frac{1+\kappa}{2}$$

Note $\Phi(c)$ is the cumulative distribution function (CDF) of the standard normal $N(0, 1)$:

From the above we obtain the critical point c from the quantile function, i.e. by inversion of Φ:

$$c = \Phi^{-1}\left(\frac{1+\kappa}{2}\right)$$

The following table lists c for the three most commonly used values of κ - it is useful to memorise these values!

Coverage κ	Critical value c
0.9	1.64
0.95	1.96

Coverage κ	Critical value c
0.99	2.58

A **symmetric standard normal CI** with nominal coverage κ for

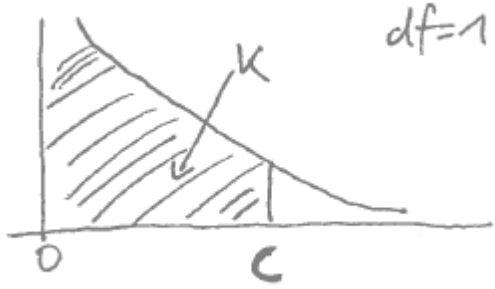
- a scalar parameter θ
- with normally distributed estimate $\hat{\theta}$ and
- with estimated standard deviation $\hat{\mathrm{SD}}(\hat{\theta}) = \hat{\sigma}$

is then given by

$$\widehat{\mathrm{CI}} = [\hat{\theta} \pm c\hat{\sigma}]$$

where c is chosen for desired coverage level κ.

A.6.12 Confidence interval based on the chi-squared distribution



As for the normal CI we can compute critical values but for the chi-squared distribution we use a one-sided interval:

$$\Pr(x \leq c) = \kappa$$

As before we get c by the quantile function, i.e. by inverting the CDF of the chi-squared distribution.

The following list the critical values for the three most common choice of κ for $m = 1$ (one degree of freedom):

Coverage κ	Critical value c ($m = 1$)
0.9	2.71
0.95	3.84
0.99	6.63

A one-sided CI with nominal coverage κ is then given by $[0, c]$.

Appendix B

Essential distributions

This appendix revisits a number of essential distributions that are frequently employed statistical analysis and probabilistic modelling. You should already be familiar with most of these distributions from earlier modules (Probability 1 and 2, Statistics 1).

B.1 Bernoulli distribution

The **Bernoulli distribution** Ber(θ) is simplest distribution possible. It is named after Jacob Bernoulli (1655-1705) who also discovered the law of large numbers.

It describes a discrete binary random variable with two states $x = 0$ ("failure") and $x = 1$ ("success"), where the parameter $\theta \in [0, 1]$ is the probability of "success". Often the Bernoulli distribution is also referred to as "coin tossing" model with the two outcomes "heads" and "tails".

Correspondingly, the probability mass function of Ber(θ) is

$$p(x = 0) = \Pr(\text{"failure"}) = 1 - \theta$$

and

$$p(x = 1) = \Pr(\text{"success"}) = \theta$$

A compact way to write the PMF of the Bernoulli distribution is

$$p(x|\theta) = \theta^x (1 - \theta)^{1-x}$$

The log PMF is

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

If a random variable x follows the Bernoulli distribution we write

$$x \sim \text{Ber}(\theta)\,.$$

The expected value is $\mathrm{E}(x) = \theta$ and the variance is $\mathrm{Var}(x) = \theta(1 - \theta)$.

B.2 Binomial distribution

Closely related to the Bernoulli distribution is the **binomial distribution** $\text{Bin}(n, \theta)$ which results from repeating a Bernoulli experiment n times and counting the number of successes among the n trials (without keeping track of the ordering of the experiments). Thus, if $x_1, \ldots, x_n$ are n independent $\text{Ber}(\theta)$ random variables then $y = \sum_{i=1}^{n} x_i$ is distributed as $\text{Bin}(n, \theta)$.

Its probability mass function is:

$$p(y|n, \theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y}$$

for $y \in \{0, 1, 2, \ldots, n\}$. The binomial coefficient $\binom{n}{x}$ is needed to account for the multiplicity of ways (orderings of samples) in which we can observe y sucesses.

The expected value is $\text{E}(y) = n\theta$ and the variance is $\text{Var}(y) = n\theta(1-\theta)$.

If a random variable y follows the binomial distribution we write

$$y \sim \text{Bin}(n, \theta)$$

For $n = 1$ it reduces to the Bernoulli distribution $\text{Ber}(\theta)$.

In R the PMF of the binomial distribution is called `dbinom()`. The binomial coefficient itself is computed by `choose()`.

B.3 Categorical distribution

The **categorical distribution** is a generalisation of the Bernoulli distribution from two classes to K classes.

The categorical distribution $\text{Cat}(\boldsymbol{\pi})$ describes a discrete random variable with K states ("categories", "classes", "bins") where the parameter vector $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)^T$ specifies the probability of each of class so that Pr("class k") $= \pi_k$. The parameters satisfy $\pi_k \in [0, 1]$ and $\sum_{k=1}^{K} \pi_k = 1$, hence there are $K - 1$ independent parameters in a categorical distribution (and not K).

There are two main ways to numerically represent "class k":

i) by "integer encoding", i.e. by the corresponding integer k.
ii) by "one hot encoding", i.e. by an indicator vector $\boldsymbol{x} = (x_1, \ldots, x_K)^T = (0, 0, \ldots, 1, \ldots, 0)^T$ containing zeros everywhere except for the element $x_k = 1$ at position k. Thus all $x_k \in \{0, 1\}$ and $\sum_{k=1}^{K} x_k = 1$.

In the following we use "one hot encoding". Therefore sampling from a categorical distribution with parameters $\boldsymbol{\pi}$

$$\boldsymbol{x} \sim \text{Cat}(\boldsymbol{\pi})$$

yields a random index vector $\boldsymbol{x}$.

The corresponding probability mass function (PMF) can be written conveniently in terms of x_k as

$$p(\boldsymbol{x}|\boldsymbol{\pi}) = \prod_{k=1}^{K} \pi_k^{x_k} = \begin{cases} \pi_k & \text{if } x_k = 1 \end{cases}$$

and the log PMF as

$$\log p(\boldsymbol{x}|\boldsymbol{\pi}) = \sum_{k=1}^{K} x_k \log \pi_k = \begin{cases} \log \pi_k & \text{if } x_k = 1 \end{cases}$$

In order to be more explicit that the categorical distribution has $K-1$ and not K parameters we rewrite the log-density with $\pi_K = 1 - \sum_{k=1}^{K-1} \pi_k$ and $x_K = 1 - \sum_{k=1}^{K-1} x_k$ as

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

Note that there is no particular reason to choose π_K as dependent of the probabilities of the other classes, in its place any other of the π_k may be selected.

For $K = 2$ the categorical distribution reduces to the Bernoulli $\text{Ber}(\theta)$ distribution, with $\pi_1 = \theta$ and $\pi_2 = 1 - \theta$.

The expected value is $\text{E}(\boldsymbol{x}) = \boldsymbol{\pi}$, in component notation $\text{E}(x_k) = \pi_k$. The covariance matrix is $\text{Var}(\boldsymbol{x}) = \text{Diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^T$, which in component notation is $\text{Var}(x_i) = \pi_i(1 - \pi_i)$ and $\text{Cov}(x_i, x_j) = -\pi_i \pi_j$.

The form of the categorical covariance matrix follows directly from the definition of the variance $\text{Var}(\boldsymbol{x}) = \text{E}(\boldsymbol{x}\boldsymbol{x}^T) - \text{E}(\boldsymbol{x})\text{E}(\boldsymbol{x})^T$ and noting that $x_i^2 = x_i$ and $x_i x_j = 0$ if $i \neq j$. Furthermore, the categorical covariance matrix is singular by construction, as the K random variables $x_1, \ldots, x_K$ are dependent through the constraint $\sum_{k=1}^{K} x_k = 1$.

B.4 Multinomial distribution

The **multinomial distribution** $\text{Mult}(n, \boldsymbol{\pi})$ arises from repeated categorical sampling, in the same fashion as the binomial distribution arises from repeated Bernoulli sampling. Thus, if $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are n independent $\text{Cat}(\boldsymbol{\pi})$ random categorical variables then $\boldsymbol{y} = \sum_{i=1}^{n} \boldsymbol{x}_i$ is distributed as $\text{Mult}(n, \boldsymbol{\pi})$.

The corresponding PMF describes the probability of a pattern $y_1, \ldots, y_K$ of samples distributed across K classes (with $n = \sum_{k=1}^{K} y_k$):

$$p(\boldsymbol{y}|n, \theta) = \binom{n}{y_1, \ldots, y_n} \prod_{k=1}^{K} \pi_k^{y_k}$$

where $\binom{n}{y_1,\ldots,y_n}$ is the multinomial coefficient.

The expected value is $\mathrm{E}(\boldsymbol{y}) = n\boldsymbol{\pi}$, in component notation $\mathrm{E}(y_k) = n\pi_k$. The covariance matrix is $\mathrm{Var}(\boldsymbol{y}) = n\mathrm{Diag}(\boldsymbol{\pi}) - n\boldsymbol{\pi}\boldsymbol{\pi}^T$, which in component notation is $\mathrm{Var}(x_i) = n\pi_i(1 - \pi_i)$ and $\mathrm{Cov}(x_i, x_j) = -n\pi_i\pi_j$.

The multinomial distribution is further described in in the module MATH38161 Multivariate Statistics and Machine Learning.

B.5 Normal distribution

The **normal distribution** is the most important continuous probability distribution. It is also called **Gaussian distribution** named after Carl Friedrich Gauss (1777–1855).

The univariate normal distribution $N(\mu, \sigma^2)$ has two parameters μ (location) and σ^2 (scale):

$$x \sim N(\mu, \sigma^2)$$

with mean

$$\mathrm{E}(x) = \mu$$

and variance

$$\mathrm{Var}(x) = \sigma^2$$

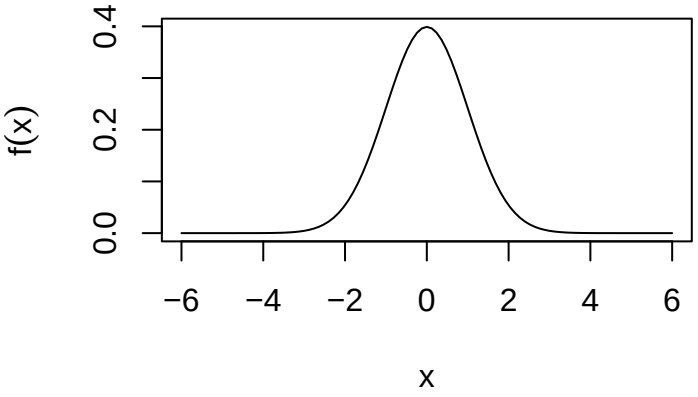
Probability density function (PDF):

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

In R the density function is called `dnorm()`.

The standard normal distribution is $N(0, 1)$ with mean 0 and variance 1.

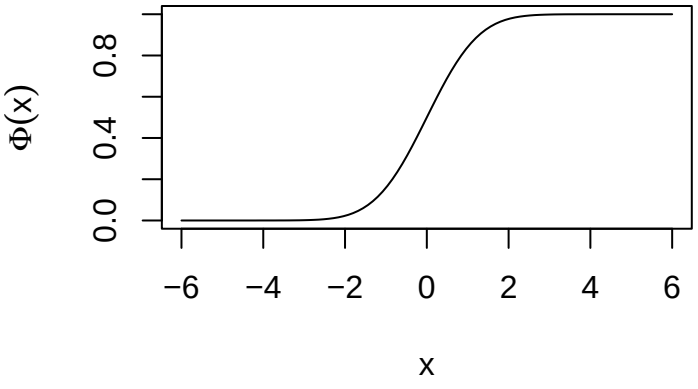
Plot of the PDF of the standard normal:



The cumulative distribution function (CDF) of the standard normal $N(0, 1)$ is

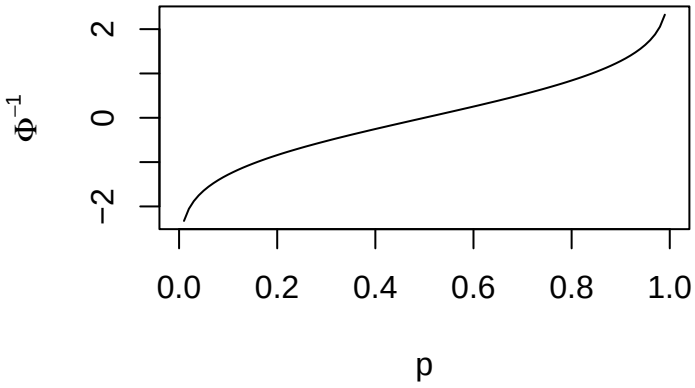
$$\Phi(x) = \int_{-\infty}^{x} p(x'|\mu = 0, \sigma^2 = 1)dx'$$

There is no analytic expression for $\Phi(x)$. In R the function is called `pnorm()`. Plot of the CDF of the standard normal:



The inverse $\Phi^{-1}(p)$ is called the quantile function of the standard normal. In R

the function is called `qnorm()`.



The sum of two normal random variables is also normal (with the appropriate mean and variance).

The **central limit theorem** (first postulated by Abraham de Moivre (1667–1754)) asserts that in many cases the distribution of the mean of identically distributed independent random variables converges to a normal distribution, even if the individual random variables are not normal.

B.6 Multivariate normal distribution

The univariate normal distribution for a random scalar x generalises to the **multivariate normal distribution** for a random vector $\boldsymbol{x} = (x_1, x_2, ..., x_d)^T \sim N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with with mean $\mathrm{E}(\boldsymbol{x}) = \boldsymbol{\mu}$ and covariance matrix $\mathrm{Var}(\boldsymbol{x}) = \boldsymbol{\Sigma}$. The corresponding density is

$$p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{d}{2}} \det(\boldsymbol{\Sigma})^{-\frac{1}{2}} \exp\left(\underbrace{-\frac{1}{2} \underbrace{(\boldsymbol{x}-\boldsymbol{\mu})^T}_{1\times d} \underbrace{\boldsymbol{\Sigma}^{-1}}_{d\times d} \underbrace{(\boldsymbol{x}-\boldsymbol{\mu})}_{d\times 1}}_{1\times 1 = \text{scalar!}} \right)$$

The expectation is $\mathrm{E}(\boldsymbol{x}) = \boldsymbol{\mu}$ and the variance $\mathrm{Var}(\boldsymbol{x}) = \boldsymbol{\Sigma}$.

For $d = 1$ we get $\boldsymbol{x} = x$, $\boldsymbol{\mu} = \mu$ and $\boldsymbol{\Sigma} = \sigma^2$ so that the multivariate normal density reduces to the univariate normal density.

B.7 Gamma distribution (aka Wishart and scaled chi-squared distribution) and special cases (chi-squared and exponential distribution)

The gamma distribution is widely used in statistics, and appears in various parameterisations and under different names, which may be confusing at times.

B.7.1 Standard parameterisation

Another important continous distribution is the gamma distribution $\text{Gam}(\alpha, \theta)$. It has two parameters $\alpha > 0$ (shape) and $\theta > 0$ (scale):

$$x \sim \text{Gam}(\alpha, \theta)$$

with mean

$$\text{E}(x) = \alpha\theta$$

and variance

$$\text{Var}(x) = \alpha\theta^2$$

The gamma distribution is also often used with a rate parameter $\beta = 1/\theta$ (so one needs to pay attention which parameterisation is used).

Probability density function (PDF):

$$p(x|\alpha, \theta) = \frac{1}{\Gamma(\alpha)\theta^\alpha} x^{\alpha-1} e^{-x/\theta}$$

The density of the gamma distribution is available in the R function `dgamma()`. The cumulative density function is `pgamma()` and the quantile function is `qgamma()`.

B.7.2 Wishart parameterisation and scaled chi-squared distribution

The gamma distribution is often used with a different set of parameters $k = 2\alpha$ and $s^2 = \theta/2$ (hence conversely $\alpha = k/2$ and $\theta = 2s^2$). In this form it is known as **one-dimensional Wishart distribution**

$$W_1\left(s^2, k\right)$$

named after John Wishart (1898–1954). In the Wishart parameterisation the mean is

$$\text{E}(x) = ks^2$$

and the variance

$$\mathrm{Var}(x) = 2ks^4$$

Another name for the one-dimensional Wishart distribution with exactly the same parameterisation is **scaled chi-squared distribution** denoted as

$$s^2\chi^2_k$$

Finally, note we often employ the Wishart distribution in **mean parameterisation** $W_1\left(s^2 = \mu/k, k\right)$ with $\mu = ks^2$ and k (and thus $\theta = 2\mu/k$). It has mean

$$\mathrm{E}(x) = \mu$$

and variance

$$\mathrm{Var}(x) = \frac{2\mu^2}{k}$$

B.7.3 Construction as sum of squared normals

A gamma distributed variable can be constructed as follows. Assume k independent normal random variables with mean 0 and variance s^2:

$$z_1, z_2, \ldots, z_k \sim N(0, s^2)$$

Then the sum of the squares

$$x = \sum_{i=1}^{k} z_i^2$$

follows

$$x \sim \sigma^2\chi^2_k = W_1\left(s^2, k\right)$$

or equivalently

$$x \sim \mathrm{Gam}\left(\alpha = \frac{k}{2}, \theta = 2s^2\right)$$

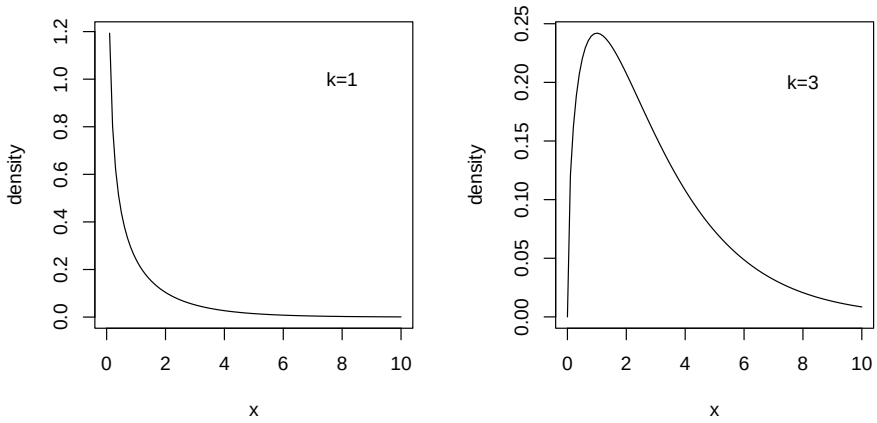
B.7.4 Special cases of the gamma distribution (chi-squared and exponential distribution)

B.7.4.1 Chi-squared distribution

The **chi-squared distribution** χ^2_k is a special one-parameter restriction of the gamma resp. Wishart distribution obtained when setting $s^2 = 1$ or, equivalently, $\theta = 2$ or $\mu = k$.

It has mean $\mathrm{E}(x) = k$ and variance $\mathrm{Var}(x) = 2k$. The chi-squared distribution χ^2_k equals $\mathrm{Gam}(\alpha = k/2, \theta = 2) = W_1(1, k)$.

Here is a plot of the density of the chi-squared distribution for degrees of freedom $k = 1$ and $k = 3$:



In R the density of the chi-squared distribution is given by `dchisq()`. The cumulative density function is `pchisq()` and the quantile function is `qchisq()`.

B.7.4.2 Exponential distribution

The **exponential distribution** $\text{Exp}(\theta)$ with scale parameter θ is another special one-parameter restriction of the gamma distribution with shape parameter set to $\alpha = 1$ (or equivalently $k = 2$).

It thus equals $\text{Gam}(\alpha = 1, \theta) = W_1(s^2 = \theta/2, k = 2)$. It has mean θ and variance θ^2.

Just like the gamma distribution the exponential distribution is also often specified using a rate parameter $\beta = 1/\theta$ instead of a scale parameter θ.

In R the command `dexp()` returns the density of the exponential distribution, `pexp()` is the corresponding cumulative density function and `qexp()` is the quantile function.

B.8 Location-scale t-distribution and special cases (Student's t and Cauchy distribution)

B.8.1 Location-scale t-distribution

The location-scale t-distribution $\text{lst}(\mu, \tau^2, \nu)$ is a generalisation of the normal distribution. It has an additional parameter $\nu > 0$ (degrees of freedom) that controls the probability mass in the tails. For small values of ν the distribution is

heavy-tailed — indeed so heavy that for $\nu \leq 1$ even the mean is not defined and for $\nu \leq 2$ the variance is undefined.

The probability density of $\text{lst}(\mu, \tau^2, \nu)$ is

$$p(x|\mu, \tau^2, \nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\pi\nu\tau^2}\,\Gamma(\frac{\nu}{2})} \left(1 + \frac{(x-\mu)^2}{\nu\tau^2}\right)^{-(\nu+1)/2}$$

The mean is (for $\nu > 1$)

$$\text{E}(x) = \mu$$

and the variance (for $\nu > 2$)

$$\text{Var}(x) = \tau^2 \frac{\nu}{\nu - 2}$$

For $\nu \to \infty$ the location-scale t-distribution $\text{lst}(\mu, \tau^2, \nu)$ becomes the normal distribution $N(\mu, \tau^2)$.

In the R `extraDistr` package the command `dlst()` returns the density of the location-scale t-distribution, `plst()` is the corresponding cumulative density function and `qlst()` is the quantile function.

B.8.2 Student's t-distribution

For $\mu = 0$ and $\tau^2 = 1$ the location-scale t-distribution becomes the Student's t-distribution t_ν with mean 0 (for $\nu > 1$) and variance $\frac{\nu}{\nu-2}$ (for $\nu > 2$).

It can thus be viewed as a generalisation of the standard normal distribution $N(0, 1)$.

If $y \sim t_\nu$ then $x = \mu + \tau y$ is distributed as $x \sim \text{lst}(\mu, \tau^2, \nu)$.

For $\nu \to \infty$ the t-distribution becomes equal to $N(0, 1)$.

In R the command `dt()` returns the density of the t-distribution, `pt()` is the corresponding cumulative density function and `qt()` is the quantile function.

B.8.3 Cauchy and standard Cauchy distribution

For $\nu = 1$ the location-scale t-distribution becomes the Cauchy distribution $\text{Cau}(\mu, \tau)$ with density $p(x|\mu, \tau) = \frac{\tau}{\pi(\tau^2 + (x-\mu)^2)}$.

For $\nu = 1$ the t-distribution becomes the standard Cauchy distribution $\text{Cau}(0, 1)$ with density $p(x) = \frac{1}{\pi(1+x^2)}$.

Appendix C

Further distributions used in Bayesian analysis

This appendix introduces a number of distributions essential for Bayesian analysis.

See in particular the Chapter "Bayesian learning in practise".

C.1 Beta distribution

C.1.1 Standard parameterisation

The density of the beta distribution Beta(α, β) is

$$p(x|\alpha, \beta) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

with $x \in [0, 1]$ and $\alpha > 0$ and $\beta > 0$. The density depends on the beta function $B(z_1, z_1) = \frac{\Gamma(z_1)\Gamma(z_2)}{\Gamma(z_1+z_2)}$ which in turn is defined via Euler's gamma function $\Gamma(x)$. Note that $\Gamma(x) = (x-1)!$ for any positive integer x.

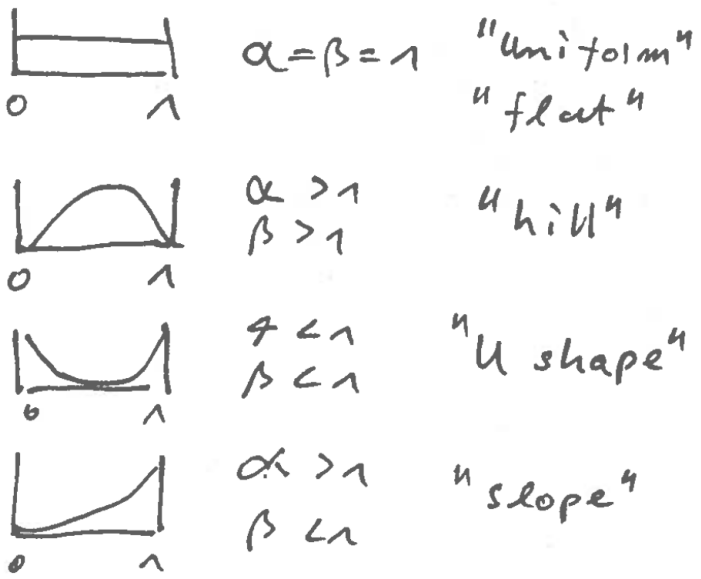
The mean of the beta distribution is

$$\mathrm{E}(x) = \frac{\alpha}{\alpha + \beta}$$

and its variance is

$$\mathrm{Var}(x) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

The beta distribution is very flexible and can assume a number of different shapes, depending on the value of α and β:



C.1.2 Mean parameterisation

A useful reparameterisation Beta(μ, k) of the beta distribution is in terms of a mean parameter $\mu \in [0, 1]$ and a concentration parameter $k > 0$. These are given by

$$k = \alpha + \beta$$

and

$$\mu = \frac{\alpha}{\alpha + \beta}$$

The original parameters can be recovered by

$$\alpha = \mu k$$

and

$$\beta = (1 - \mu)k$$

The mean and variance of the beta distribution expressed in terms of μ and k are

$$\mathrm{E}(x) = \mu$$

and

$$\mathrm{Var}(x) = \frac{\mu(1 - \mu)}{k + 1}$$

With increasing concentration parameter k the variance decreases and thus the probability mass becomes more concentrated around the mean.

C.2 Inverse gamma (inverse Wishart) distribution

C.2.1 Standard parameterisation

The inverse gamma (IG) distribution Inv-Gam(α, β) has density

$$\frac{\beta^\alpha}{\Gamma(\alpha)}(1/x)^{\alpha+1}e^{-\beta/x}$$

with two parameters $\alpha > 0$ (shape parameter) and $\beta > 0$ (scale parameter) and support $x > 0$.

The mean of the inverse gamma distribution is

$$\mathrm{E}(x) = \frac{\beta}{\alpha - 1}$$

and the variance

$$\mathrm{Var}(x) = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)}$$

Thus, for the mean to exist we have the restriction $\alpha > 1$ and for the variance to exist $\alpha > 2$.

The IG distribution is closely linked with the gamma distribution. If $x \sim$ Inv-Gam(α, β) is IG-distributed then the inverse of x is gamma distributed:

$$\frac{1}{x} \sim \mathrm{Gam}(\alpha, \theta = \beta^{-1})$$

where α is the shared shape parameter and θ the scale parameter of the gamma distribution.

C.2.2 Wishart parameterisation

The inverse gamma distribution is frequently used with a different set of parameters $\psi = 2\beta$ (scale parameter) and $\nu = 2\alpha$ (shape parameter), or conversely $\alpha = \nu/2$ and $\beta = \psi/2$. In this form it is called **one-dimensional inverse Wishart distribution** $W_1^{-1}(\psi, \nu)$ with mean and variance given by

$$\mathrm{E}(x) = \frac{\psi}{\nu - 2} = \mu$$

for $\nu > 2$ and

$$\mathrm{Var}(x) = \frac{2\psi^2}{(\nu-4)(\nu-2)^2} = \frac{2\mu^2}{\nu-4}$$

for $\nu > 4$.

Instead of ψ and ν we may also equivalently use μ and $\kappa = \nu - 2$ as parameters for the inverse Wishart distribution, so that $W_1^{-1}(\psi = \kappa\mu, \nu = \kappa + 2)$ has mean

$$\mathrm{E}(x) = \mu$$

with $\kappa > 0$ and the variance is

$$\mathrm{Var}(x) = \frac{2\mu^2}{\kappa - 2}$$

with $\kappa > 2$. This **mean parameterisation** is useful when employing the IG distribution as prior and posterior.

Finally, with $W_1^{-1}(\psi = \nu\tau^2, \nu)$, where $\tau^2 = \mu\frac{\kappa}{\kappa+2} = \frac{\psi}{\nu}$ is a biased mean parameter, we get the **scaled inverse chi-squared distribution** $\tau^2\text{Inv-}\chi^2_\nu$ with

$$\mathrm{E}(x) = \tau^2\frac{\nu}{\nu - 2}$$

for $\nu > 2$ and

$$\mathrm{Var}(x) = \frac{2\tau^4}{\nu - 4}\frac{\nu^2}{(\nu - 2)^2}$$

for $\nu > 4$.

The inverse Wishart and Wishart distributions are linked. If $x \sim W_1^{-1}(\psi, \nu)$ is inverse-Wishart distributed then the inverse of x is Wishart distributed with inverted scale parameter:

$$\frac{1}{x} \sim W_1(s^2 = \psi^{-1}, k = \nu)$$

where k is the shape parameter and s^2 the scale parameter of the Wishart distribution.

C.3 Location-scale t-distribution as compound distribution

Suppose that

$$x|s^2 \sim N(\mu, s^2)$$

with corresponding density $p(x|s^2)$ and mean $\mathrm{E}(x|s^2) = \mu$ and variance $\mathrm{Var}(x|s^2) = s^2$.

Now let the variance s^2 be distributed as inverse gamma / inverse Wishart

$$s^2 \sim W^{-1}(\psi = \kappa\sigma^2, \nu = \kappa + 2) = W^{-1}(\psi = \tau^2\nu, \nu)$$

with corresponding density $p(s^2)$ and mean $\mathrm{E}(s^2) = \sigma^2 = \tau^2\nu/(\nu - 2)$. Note we use here both the mean parameterisation (σ^2, κ) and the inverse chi-squared parameterisation (τ^2, ν).

The joint density for x and s^2 is $p(x, s^2) = p(x|s^2)p(s^2)$. We are interested in the marginal density for x:

$$p(x) = \int p(x, s^2)ds^2 = \int p(s^2)p(x|s^2)ds^2$$

This is a compound distribution of a normal with fixed mean μ and variance s^2 varying according the inverse gamma distribution. Calculating the integral results in the location-scale t-distribution with parameters

$$x \sim \text{lst}\left(\mu, \sigma^2\frac{\kappa}{\kappa+2}, \kappa+2\right) = \text{lst}\left(\mu, \tau^2, \nu\right)$$

with mean

$$\text{E}(x) = \mu$$

and variance

$$\text{Var}(x) = \sigma^2 = \tau^2\frac{\nu}{\nu-2}$$

From the law of total expectation and variance we can also directly verify that

$$\text{E}(x) = \text{E}(\text{E}(x|s^2)) = \mu$$

and

$$\text{Var}(x) = \text{E}(\text{Var}(x|s^2)) + \text{Var}(\text{E}(x|s^2)) = \text{E}(s^2) = \sigma^2 = \tau^2\frac{\nu}{\nu-2}$$

Appendix D

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.

D.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.

D.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.

Bibliography

Agresti, A., and M. Kateri. 2022. *Foundations of Statistics for Data Scientists*. Chapman; Hall/CRC.

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.