Statistical Methods: Likelihood, Bayes and Regression

Korbinian Strimmer

28 February 2023

Contents

Welcome

These are the lecture notes for MATH20802, a course in **Statistical Methods** for second year mathematics students at the Department of Mathematics of the University of Manchester.

The course text was written by Korbinian Strimmer from 2019–2023. This version is from 28 February 2023.

The notes will be updated from time to time. To view the current version visit the online MATH20802 lecture notes.

You may also download the MATH20802 lecture notes as PDF. For a paper copy it is recommended to print two pages per sheet.

License

These notes are licensed to you under Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International 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 MATH20802 taught in spring 2023 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 MATH20802 module is designed to run over the course of 11 weeks. It has three main parts:

1. Likelihood estimation and inference (W1–W4)
2. Bayesian learning and inference (W5–W8)
3. Linear regression (W9–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 MATH20802 online reading list hosted by the University of Manchester library.

Acknowledgements

Many thanks to Beatriz Costa Gomes for her help in creating the 2019 version of the lecture notes when I was teaching this 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 R.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 that works in a non-probabilistic way. 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 has a dedicated computer chip with special neural network hardware, for example.
- 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).
- 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 thus it belongs to the domain of pure mathematics.

Probability theory provides probabilistic generative models for data, for simulation of data or for use in learning from data, i.e. inference about the model from observations. Methods and theory how to best learn from data is 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 in fact not at all concerned with randomness. Instead, the focus is about measuring and elucidating the **uncertainty** of events, predictions, outcomes, parameters and this uncertainty measures the **state of knowledge**. Note that if 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 specifics of the underlying processes, and we still have procedures for optimal inference despite the uncertainty.

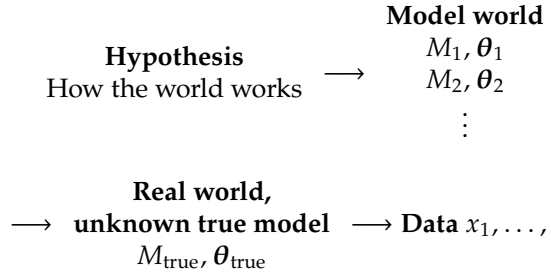
In short, statistics is about describing the state of knowledge of the world, which may be uncertain and incomplete, and to make decisions and predictions in the face of uncertainty, and this uncertaintly sometimes derives from randomness but most often from our ignorance (and sometimes this ignorance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

1.3 Cartoon of statistical learning

We observe data $D = \{x_1, \ldots, x_n\}$ assumed to have been generated by an underlying true model M_{true} with true parameters $\boldsymbol{\theta}_{\text{true}}$

To explain the data, and make predictions, we make hypotheses in the form of candidate models $M_1, M_2, \ldots$ and corresponding parameters $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots$. The true model itself is unknown and cannot be observed. However, what we can observe is data D from the true model by measuring properties of objects interest (our observations from experiments). Sometimes we can also perturb the model and see what the effect is (interventional study).

The various candidate models $M_1, M_2, \ldots$ in the **model world** will never be perfect or correct as the true model M_{true} will only be among the candidate models in an idealised situation. However, 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 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** matches the complexity of the unknown true model and also the complexity of the data observed from the unknown true model.

1.4 Likelihood

In statistics and machine learning most models that are being used are probabilistic to take account of both randomness and uncertainty. A core task in statistical learning is to identify those models that explain the existing data well and that also generalise well to unseen data.

For this we need, among other things, a measure of how well a candidate model approximates the (typically unknown) true data generating model and an approach to choose the best model(s). One such approach is provided by the method of maximum likelihood that enables us to estimate parameters of models and to find the particular model that is the best fit to the data.

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 defined as

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

Typically, instead of the likelihood one uses the log-likelihood function:

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

Reasons for preferring the log-likelihood (rather than likelihood) include that

- the log-density is in fact the more "natural" and relevant quantity (this will become clear in the upcoming chapters) and that
- addition is numerically more stable than multiplication on a computer.

For discrete random variables for which $p(x|\boldsymbol{\theta})$ is a probability mass function the likelihood is often interpreted as the probability to observe the data given the model with specified parameters $\boldsymbol{\theta}$. In fact, this was indeed the way how the likelihood was historically introduced. However, this view is not strictly correct. First, given that the samples are iid and thus the ordering of the x_i is not important, an additional factor accounting for the possible permutations is needed in the likelihood to obtain the actual probability of the data. Moreover, for continuous random variables this interpretation breaks down due to the use

of densities rather than probability mass functions in the likelihood. Thus, the view that the likelihood is the probability of the data is in fact too simplistic.

In the next chapter we will see that the justification for using likelihood rather stems from its close link to the Kullback-Leibler information and cross-entropy. This also helps to understand why using likelihood for estimation is only optimal in the limit of large sample size.

In the first part of the MATH28082 "Statistical Methods" module we will study likelihood estimation and inference in much detail. We will provide links to related methods of inference and discuss its information-theoretic foundations. We will also discuss the optimality properties as well as the limitations of likelihood inference. Extensions of likelihood analysis, in particular Bayesian learning, which will be discussed in the second part of the module. In the third part of the module we will apply statistical learning to linear regression.

Chapter 2

From entropy to maximum likelihood

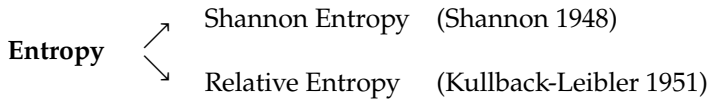
2.1 Entropy

2.1.1 Overview

In this chapter we discuss various information criteria and their connection to maximum likelihood.

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

In the 1940–1950's the notion of entropy turned out to be central in information theory, a field pioneered by mathematicians such as R. Hartley (1888–1970), S. Kullback (1907–1994), A. Turing (1912–1954), R. Leibler (1914–2003), I. J. Good (1916–2009), C. Shannon (1916–2001), and E. T. Jaynes (1922–1998), and later further explored by S. Amari (1936–), I. Ciszár (1938–), B. Efron (1938–), A. P. Dawid (1946–) and many others.



Fisher information	→	Likelihood theory	(Fisher 1922)
Mutual Information	→	Information theory	(Shannon 1948, Lindley 1953)

2.1.2 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".

2.1.3 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 $p(x =$ "class k") $= p_k$.

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

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

[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

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 Langrange multipliers. The maximum is achieved for P being the discrete uniform - see Example 2.1.

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).

Example 2.1. 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 this is the largest value the Shannon entropy can assume with K classes.

Example 2.2. 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 corresponds to maximum concentration.

Thus, **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.

This interpretation is also supported by the close link of Shannon entropy with multinomial coefficients counting the permutations of n items (samples) of K distinct types (classes).

Example 2.3. 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 combinatorical derivation of Shannon entropy is now credited 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).

2.1.4 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.

Example 2.4. 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.

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.

2.1.5 Maximum entropy principle to characterise distributions

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

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.

The higher the entropy the more spread out (and more uninformative) is a distribution.

Using maximum entropy to characterise maximally uniformative 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.

2.1.6 Cross-entropy

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

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

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

Therefore, cross-entropy is a measure linking two distributions F and G.

Note that

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

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.

Example 2.6. Cross-entropy between two normals:

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

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

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

Example 2.7. If $\mu_{\text{ref}} = \mu$ and $\sigma^2_{\text{ref}} = \sigma^2$ then the cross-entropy $H(F_{\text{ref}}, F)$ in Example 2.6 degenerates to the differential entropy $H(F_{\text{ref}}) = \frac{1}{2}\left(\log(2\pi\sigma^2_{\text{ref}}) + 1\right)$.

2.2 Kullback-Leibler divergence

2.2.1 Definition

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

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

- $D_{\text{KL}}(F, G)$ 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 $D_{\text{KL}}(F, G) = 0$.

(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 $D_{\text{KL}}(F, G)$.

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

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

2.2.2 Properties of KL divergence

1. $D_{\text{KL}}(F, G) \neq D_{\text{KL}}(G, F)$, i.e. the KL divergence is not symmetric, F and G cannot be interchanged.
2. $D_{\text{KL}}(F, G) = 0$ if and only if $F = G$, i.e., the KL divergence is zero if and only if F and G are identical.
3. $D_{\text{KL}}(F, G) \geq 0$, proof via **Jensen's inequality**.
4. $D_{\text{KL}}(F, G)$ 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.

2.2.3 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.

2.2.4 KL divergence examples

Example 2.8. KL divergence between two Bernoulli distributions $\mathrm{Ber}(p)$ and $\mathrm{Ber}(q)$:

The "success" probabilities for the two distributions are p and q, respectively, and the complementary "failure" probabilities are $1-p$ and $1-q$. With this we get for the KL divergence

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

Example 2.9. KL divergence between two univariate normals with different means and variances:

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

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

[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

Example 2.10. KL divergence between two univariate normals with different means and common variance:

An important special case of the previous Example 2.9 occurs if the variances are equal. Then we get

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

2.3 Local quadratic approximation and expected Fisher information

2.3.1 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 $\boldsymbol{\varepsilon}$.

Let $h(\boldsymbol{\theta}) = D_{\mathrm{KL}}(F_{\boldsymbol{\theta}_0}, F_{\boldsymbol{\theta}}) = \mathrm{E}_{F_{\boldsymbol{\theta}_0}}\left(\log f(\boldsymbol{x}|\boldsymbol{\theta}_0) - \log f(\boldsymbol{x}|\boldsymbol{\theta})\right)$. Note that the first distribution in the KL divergence is fixed at $F_{\boldsymbol{\theta}_0}$ and the second distribution is varied. Then $h(\boldsymbol{\theta}_0 + \boldsymbol{\varepsilon}) = D_{\mathrm{KL}}(F_{\boldsymbol{\theta}_0}, F_{\boldsymbol{\theta}_0+\boldsymbol{\varepsilon}})$. Since the KL divergence vanishes only when the two arguments are identical $h(\boldsymbol{\theta})$ reaches a minimum at $\boldsymbol{\theta}_0$ with $h(\boldsymbol{\theta}_0) = 0$ and flat gradient $\nabla h(\boldsymbol{\theta}_0) = 0$.

We can therefore approximate $h(\boldsymbol{\theta}_0 + \boldsymbol{\varepsilon})$ by a quadratic function around $\boldsymbol{\theta}_0$

$$\begin{aligned} D_{\mathrm{KL}}(F_{\boldsymbol{\theta}_0}, F_{\boldsymbol{\theta}_0+\boldsymbol{\varepsilon}}) = h(\boldsymbol{\theta}_0 + \boldsymbol{\varepsilon}) &\approx \frac{1}{2}\boldsymbol{\varepsilon}^T \nabla\nabla^T h(\boldsymbol{\theta}_0)\boldsymbol{\varepsilon} \\ &= \frac{1}{2}\boldsymbol{\varepsilon}^T \left(-\mathrm{E}_{F_{\boldsymbol{\theta}_0}} \nabla\nabla^T \log f(\boldsymbol{x}|\boldsymbol{\theta}_0)\right) \boldsymbol{\varepsilon} \\ &= \frac{1}{2}\boldsymbol{\varepsilon}^T \underbrace{\boldsymbol{I}^{\mathrm{Fisher}}(\boldsymbol{\theta}_0)}_{\text{expected Fisher information}} \boldsymbol{\varepsilon} \end{aligned}$$

This yields the **expected Fisher information** at $\boldsymbol{\theta}_0$ as the negative expected Hessian matrix of the log-density at $\boldsymbol{\theta}_0$. Since $\boldsymbol{\theta}_0$ is a minimum the expected Fisher information matrix must be positive definite!

We can use the above approximation also to compute the divergence $D_{\mathrm{KL}}(F_{\boldsymbol{\theta}_0+\boldsymbol{\varepsilon}}, F_{\boldsymbol{\theta}_0})$ where the first argument varies and the second is kept fixed:

$$D_{\mathrm{KL}}(F_{\boldsymbol{\theta}_0+\boldsymbol{\varepsilon}}, F_{\boldsymbol{\theta}_0}) \approx \frac{1}{2}\boldsymbol{\varepsilon}^T \boldsymbol{I}^{\mathrm{Fisher}}(\boldsymbol{\theta}_0 + \boldsymbol{\varepsilon})\, \boldsymbol{\varepsilon}$$

In a linear approximation $\boldsymbol{I}^{\mathrm{Fisher}}(\boldsymbol{\theta}_0 + \boldsymbol{\varepsilon}) \approx \boldsymbol{I}^{\mathrm{Fisher}}(\boldsymbol{\theta}_0) + \boldsymbol{\Delta}_{\boldsymbol{\varepsilon}}$ each element of the matrix $\boldsymbol{\Delta}_{\boldsymbol{\varepsilon}}$ is the scalar product of $\boldsymbol{\varepsilon}$ and the gradient of the corresponding element

in $\boldsymbol{I}^{\text{Fisher}}(\boldsymbol{\theta}_0)$ evaluated at $\boldsymbol{\theta}_0$. Therefore $\boldsymbol{\varepsilon}^T \boldsymbol{\Delta}_{\boldsymbol{\varepsilon}} \boldsymbol{\varepsilon}$ is of *cubic order* in $\boldsymbol{\varepsilon}$ and hence

$$\begin{aligned} D_{\text{KL}}(F_{\boldsymbol{\theta}_0+\boldsymbol{\varepsilon}}, F_{\boldsymbol{\theta}_0}) &\approx \frac{1}{2} \boldsymbol{\varepsilon}^T \boldsymbol{I}^{\text{Fisher}}(\boldsymbol{\theta}_0 + \boldsymbol{\varepsilon}) \boldsymbol{\varepsilon} \\ &\approx \frac{1}{2} \boldsymbol{\varepsilon}^T \boldsymbol{I}^{\text{Fisher}}(\boldsymbol{\theta}_0) \boldsymbol{\varepsilon} + \boldsymbol{\varepsilon}^T \boldsymbol{\Delta}_{\boldsymbol{\varepsilon}} \boldsymbol{\varepsilon} \\ &\approx \frac{1}{2} \boldsymbol{\varepsilon}^T \boldsymbol{I}^{\text{Fisher}}(\boldsymbol{\theta}_0) \boldsymbol{\varepsilon} \end{aligned}$$

keeping only terms quadratic in $\boldsymbol{\varepsilon}$.

Note that there is no data involved in computing the expected Fisher information, hence it is purely a property of the model, or more precisely of the space of the models 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.

2.3.2 Examples

Example 2.11. Expected Fisher information for the Bernoulli distribution:

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

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

where p is the proportion of "success". The second derivative with regard to the parameter p is

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

Since $\text{E}(x) = p$ we get as Fisher information

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

Example 2.12. Quadratic approximations of the KL divergence between two Bernoulli distributions:

From Example 2.8 we have as KL divergence

$$D_{\text{KL}}\left(\text{Ber}(p_1), \text{Ber}(p_2)\right) = p_1 \log\left(\frac{p_1}{p_2}\right) + (1 - p_1) \log\left(\frac{1 - p_1}{1 - p_2}\right)$$

and from Example 2.11 the corresponding expected Fisher information.

The quadratic approximation implies that

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

and also that

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

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

Example 2.13. 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 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

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

2.4 Entropy learning and maximum likelihood

2.4.1 The relative entropy between true model and approximating model

Assume we have observations $x_1, \ldots, x_n$. 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" :

- By the strong law of large numbers the empirical distribution $\hat{F}_n$ based on observed data $D = \{x_1, \ldots, x_n\}$ converges to the true underlying distribution F as $n \to \infty$ almost surely:

$$\hat{F}_n \overset{a.s.}{\to} F$$

- 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 2.3) 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).

2.4.2 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 $l_n(\boldsymbol{\theta}|D)$. Note that the entropy of the true distribution F (and the corresponding empirical distribution $\hat{F}$) that does not depend on the parameters $\boldsymbol{\theta}$ and hence it does not matter when minimising the divergence.

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.

As a consequence of the close link of maximum likelhood 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.

2.4.3 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 Principle of maximum likelihood estimation

3.1.1 Outline

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

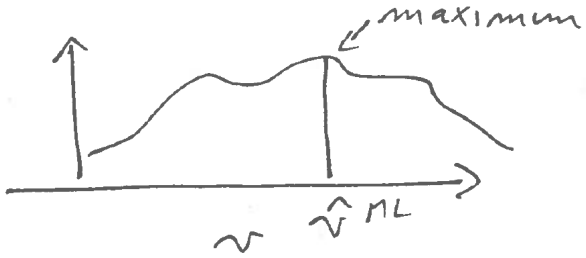
- $l_n(\boldsymbol{\theta}|D) = \sum_{i=1}^{n} \log f(x_i|\boldsymbol{\theta})$

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. Depending on the complexity of the likelihood function finding the maximum can be very difficult.

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** $\boldsymbol{S}(\boldsymbol{\theta})$ which is the first derivative of the log-likelihood function:

$S_n(\theta) = \frac{dl_n(\theta|D)}{d\theta}$ 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 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.

3.1.3 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 λ using some invertible function $g()$ so that $\lambda = g(\theta)$. Then the maximum likelihood estimate $\hat{\lambda}_{ML}$ of the new parameter λ is simply the transformation of the maximum likelihood estimate $\hat{\theta}_{ML}$ of the original parameter θ with $\hat{\lambda}_{ML} = g(\hat{\theta}_{ML})$. 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.

3.1.4 Consistency of maximum likelihood estimates

One important property of maximum likelihood is that 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 $F_{\boldsymbol{\theta}}$ 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.

3.2 Maximum likelihood estimation in practise

3.2.1 Estimation of a proportion

Example 3.1. Maximum likelihood estimation for the Bernoulli model:

We aim to estimate the true proportion p 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(p): Pr("success") = p and Pr("failure") = $1 - p$.
- 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: p: probability of "success".

What is the MLE of p?

- the observations $D = \{x_1, \ldots, x_n\}$ take on values 0 or 1.

- the average of the data points is $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{n_1}{n}$.
- the probability mass function (PMF) of the Bernoulli distribution $\text{Ber}(p)$ is:

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

- log-PMF:

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

- log-likelihood function:

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

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

- Score function:

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

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

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

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

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

Note that to analyse the coin tossing experiment and to estimate p we may equally well use the binomial distribution $\text{Bin}(n, p)$ as model for the number of successes. In this case we then have only a single observation, namely the observed k . This results in the same MLE for p but the likelihood function based on the binomial PMF includes the binomial coefficient $\binom{n}{k}$. However, as this factor does not depend on p it disappears in the score function and has no influence in the derivation of the MLE.

3.2.2 Estimation of the mean and variance of a normal distribution

Example 3.2. Normal distribution with unknown mean and known variance:

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

What's the MLE of 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:

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

  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 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:

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

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

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

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

Note that to obtain the second component of the score function the partial derivative needs to be taken with regard to the variance parameter σ^2 — not with regard to σ! Hint: replace $\sigma^2 = v$ in the log-likelihood function, then take the partial derivative with regard to v, then backsubstitute $v = \sigma^2$ in the result.

- Maximum likelihood estimate:

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

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

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

- To confirm that we actually have maximum we need to verify that the eigenvalues of the Hessian matrix are all negative. This is indeed the case, for details see Example 3.6.

3.2.3 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!

3.2.4 Bias and maximum likelihood estimates

Example 3.3 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

$$\widehat{\sigma^2}_{ML} \sim \frac{\sigma^2}{n}\chi^2_{n-1}$$

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

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

Therefore, the MLE of μ is unbiased as

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

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

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

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

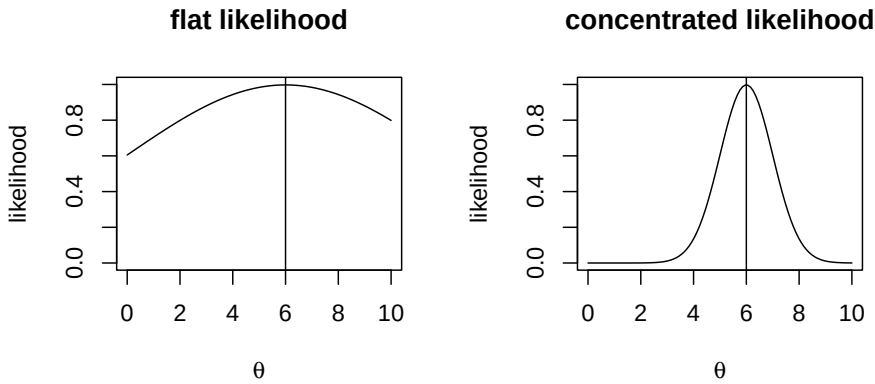
$$\widehat{\sigma^2}_{UB} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{n}{n-1}\widehat{\sigma^2}_{ML}$$

Conversely, the unbiased estimator 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.

3.3 Observed Fisher information

3.3.1 Motivation and definition



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

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

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

3.3.2 Examples of observed Fisher information

Example 3.4. Bernoulli model Ber(p):

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

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

The observed Fisher information is therefore

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

The inverse of the observed Fisher information is:

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

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

Example 3.5. 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.

Example 3.6. Normal distribution with mean and variance parameter:

This is the continuation of Example 3.3. 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

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

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 scaled chi-squared distribution

$$\widehat{\sigma^2}_{ML} \sim \frac{\sigma^2}{n} \chi^2_{n-1}$$

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.

3.3.3 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 curvature at the MLE $\hat{\boldsymbol{\theta}}_{ML}$. It is linked to the (asymptotic) variance of the MLE as we have seen in the examples and 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 $\hat{\boldsymbol{\theta}}_{ML} \to \boldsymbol{\theta}_0$. It follows from the construction of the observed Fisher information and the law of large numbers that asymptotically for large sample size $\boldsymbol{J}_n(\hat{\boldsymbol{\theta}}_{ML}) \to n\boldsymbol{I}^{\text{Fisher}}(\boldsymbol{\theta}_0)$.

- 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 3.4 and 2.11 for the Bernoulli distribution and Examples 3.6 and 2.13 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 $\frac{n}{2}$.

Chapter 4

Quadratic approximation and normal asymptotics

4.1 Multivariate statistics for random vectors

4.1.1 Covariance and correlation

Assume a scalar random variable x with mean $\mathrm{E}(x) = \mu$. The corresponding variance is given by

$$\begin{aligned}\mathrm{Var}(x) &= \mathrm{E}\left((x-\mu)^2\right) \\ &= \mathrm{E}\left((x-\mu)(x-\mu)\right) \\ &= \mathrm{E}(x^2) - \mu^2\end{aligned}$$

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 $\sigma_{ij} = \text{Cov}(x_i, x_j)$ 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 $\sigma_{ii} = \text{Cov}(x_i, x_i) = \text{Var}(x_i) = \sigma_i^2$ correspond to the variances of the components of $\boldsymbol{x}$. The covariance matrix is **positive semi-definite**, i.e. the eigenvalues of $\boldsymbol{\Sigma}$ are all positive or equal to zero. However, in practise one aims to use non-singular covariance matrices, with all eigenvalues positive, so that they are invertible.

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

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

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

4.1.2 Multivariate normal distribution

The density of a normally distributed scalar variable $x \sim N(\mu, \sigma^2)$ with mean $\text{E}(x) = \mu$ and variance $\text{Var}(x) = \sigma^2$ is

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

The univariate normal distribution for a scalar x generalises to the **multivariate normal distribution** for a 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

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

For $d = 1$ we have $\boldsymbol{x} = x$, $\boldsymbol{\mu} = \mu$ and $\boldsymbol{\Sigma} = \sigma^2$ so that the multivariate normal density reduces to the univariate normal density.

Example 4.1. Maximum likelihood estimates of the parameters of the multivariate normal distribution:

Maximising the log-likelihood based on the multivariate normal density yields the MLEs for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. These are generalisations of the MLEs for the mean μ and variance σ^2 of the univariate normal as encountered in Example 3.3.

The estimates can be written in three different ways:

a) data vector notation

with $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ the n vector-valued observations from the multivariate normal:

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

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

b) data component notation

with x_{ki} the i-th component of the k-th sample $\boldsymbol{x}_k$:

$$\hat{\mu}_i = \frac{1}{n} \sum_{k=1}^{n} x_{ki} \text{ with } \hat{\boldsymbol{\mu}} = \begin{pmatrix} \hat{\mu}_1 \\ \vdots \\ \hat{\mu}_d \end{pmatrix}$$

$$\hat{\sigma}_{ij} = \frac{1}{n}\sum_{k=1}^{n}(x_{ki} - \hat{\mu}_i)(x_{kj} - \hat{\mu}_j) \text{ with } \widehat{\Sigma} = (\hat{\sigma}_{ij})$$

c) data matrix notation

with $\boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1^T \\ \dots \\ \boldsymbol{x}_n^T \end{pmatrix}$ as a data matrix containing the samples in its rows. Note that this is the *statistics convention* — in much of the engineering and computer science literature the data matrix is often transposed and samples are stored in the columns. Thus, the formulas below are only correct assuming the statistics convention.

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\boldsymbol{X}^T\mathbf{1}_n$$

Here $\mathbf{1}_n$ is a vector of length n containing 1 at each component.

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\boldsymbol{X}^T\boldsymbol{X} - \hat{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}^T$$

To simplify the expression for the estimate of the covariance matrix one often assumes that the data matrix is centered, i.e. that $\hat{\boldsymbol{\mu}} = 0$.

Because of the ambiguity in convention (machine learning versus statistics convention) and the often implicit use of centered data matrices the matrix notation is often a source of confusion. Hence, using the other two notations is generally preferable.

4.2 Approximate distribution of maximum likelihood estimates

4.2.1 Quadratic log-likelihood resulting from normal model

Assume we observe a single sample $\boldsymbol{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}^2)$ with known covariance. 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 exactly quadratic and the maximum lies at $(\boldsymbol{x}, C)$.

4.2.2 Quadratic approximation of a log-likelihood function

Now consider the quadratic approximation of the log-likelihood function $l_n(\boldsymbol{\theta}|D)$ for 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 natural 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, we realise that the approximation has the same form as if $\hat{\boldsymbol{\theta}}_{ML}$ was a sample 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 4.2. 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 3.4 the observed Fisher information

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

The log-likelihood at the MLE is

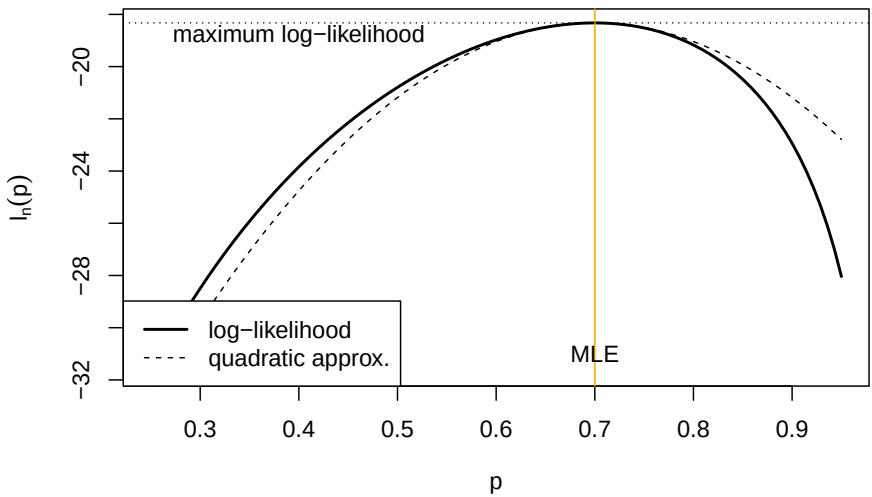
$$l_n(\hat{p}_{ML}|D) = n\left(\bar{x}\log\bar{x} + (1 - \bar{x})\log(1 - \bar{x})\right)$$

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 $\bar{x} = 0.7$:



4.2.3 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 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.

4.2.4 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).

4.3 Quantifying the uncertainty of maximum likelihood estimates

4.3.1 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):

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

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

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

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)[1] .

Example 4.3. Estimated variance and distribution of the MLE of a proportion:

From Examples 3.1 and 3.4 we know the MLE

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

and the corresponding observed Fisher information

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

The estimated variance of the MLE is therefore

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

and the corresponding asymptotic normal distribution is

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

[1] 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

Example 4.4. Estimated variance and distribution of the MLE of the mean parameter for the normal distribution with known variance:

From Examples 3.2 and 3.5 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 $N(\mu, \sigma^2)$!).

4.3.2 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:

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

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

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

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

Example 4.5. Wald statistic for a proportion:

We continue from Example 4.3. 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 4.6. Wald statistic for the mean parameter of a normal distribution with known variance:

We continue from Example 4.4. 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:

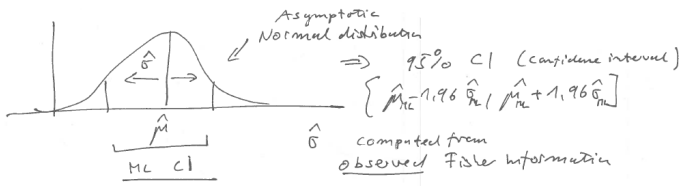
$$t(\mu_0)^2 = \frac{(\bar{x} - \mu_0)^2}{\sigma^2/n} \sim \chi^2_1$$

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.

4.3.3 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

Example 4.7. Asymptotic normal confidence interval for a proportion:

We continue from Examples 4.3 and 4.5. Assume we observe $n = 30$ measurements with average $\bar{x} = 0.7$. Then $\hat{p}_{ML} = \bar{x} = 0.7$ and $\widehat{\text{SD}}(\hat{p}_{ML}) = \sqrt{\frac{\bar{x}(1-\bar{x})}{n}} \approx 0.084$.

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

Example 4.8. Normal confidence interval for the mean:

We continue from Examples 4.4 and 4.6. Assume that we observe $n = 25$ measurements with average $\bar{x} = 10$, from a normal with unknown mean and variance $\sigma^2 = 4$.

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

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

4.3.4 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.

Example 4.9. 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 4.5) and comparing them with the relevant critical value (3.84 from chi-squared distribution for 5% significance level):

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

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

Example 4.10. Normal confidence interval and test for the mean:

We continue from Example 4.8.

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 4.6) and comparing them with the relevant critical values:

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

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

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

4.4 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:

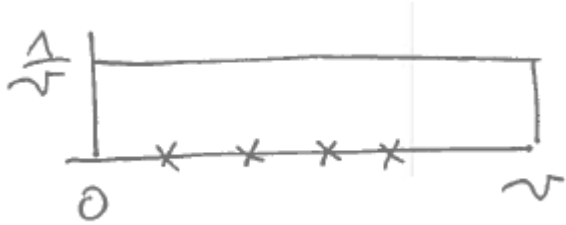
$$x_1, \ldots, x_n \sim U(0, \theta)$$

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.

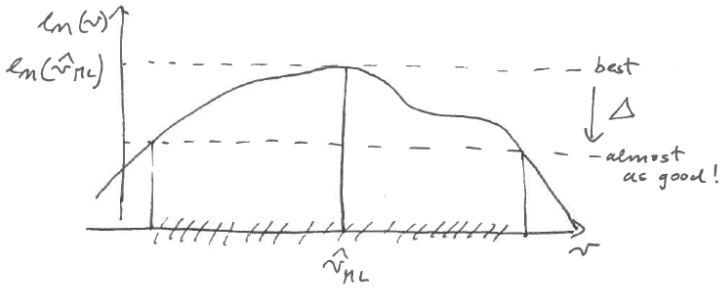
Chapter 5

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 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 $\boldsymbol{\theta}_0$ that have a log-likelihood that is almost as good as $l_n(\hat{\boldsymbol{\theta}}_{ML}|D)$.

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

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

Question: how to choose Δ, i.e how to calibrate the likelihood interval? Essentially, by comparing with a normal CI!

Example 5.1. Wilks statistic for the proportion:

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

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

Hence the Wilks statistic is

$$\begin{aligned} W(p_0) &= 2(l_n(\hat{p}_{ML}|D) - l_n(p_0|D)) \\ &= 2n\left(\bar{x}\log\left(\frac{\bar{x}}{p_0}\right) + (1-\bar{x})\log\left(\frac{1-\bar{x}}{1-p_0}\right)\right) \end{aligned}$$

Comparing with Example 2.8 we see that in this case the Wilks statistic is essentially (apart from a scale factor $2n$) the KL divergence between two Bernoulli distributions:

$$W(p_0) = 2nD_{\text{KL}}(\text{Ber}(\hat{p}_{ML}), \text{Ber}(p_0))$$

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 4.6.

Comparing with Example 2.10 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:

$$W(p_0) = 2nD_{\mathrm{KL}}(N(\hat{\mu}_{ML}, \sigma^2), N(\mu_0, \sigma^2))$$

5.1.2 Quadratic approximation of Wilks statistic and squared Wald 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.

Example 5.3. 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 4.5.

Example 5.4. 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.

5.1.3 Distribution of the Wilks statistic

The connection with the squared Wald 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**.

5.1.4 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

Example 5.5. Likelihood confidence interval for a proportion:

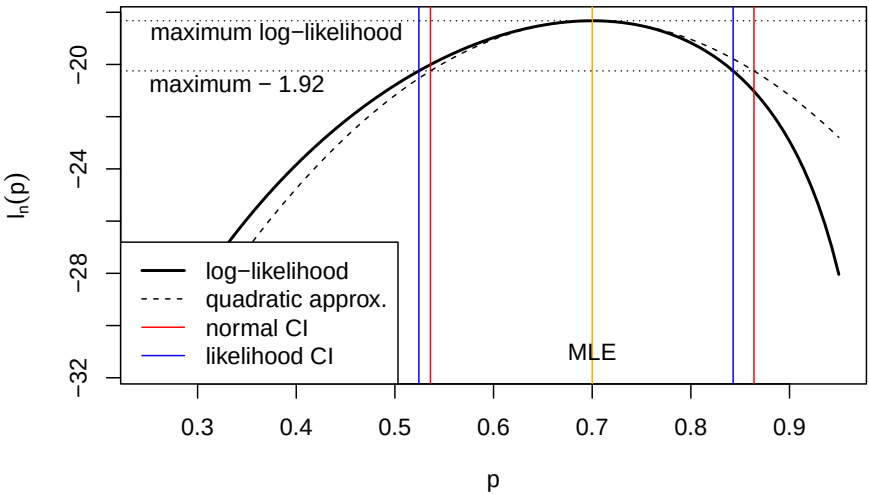
We continue from Example 5.1, and as in Example 4.7 we asssume we have data with $n = 30$ and $\bar{x} = 0.7$.

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 4.2). 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!



5.1.5 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 5.6. Likelihood test for a proportion:

We continue from Example 5.5 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 4.9.

5.1.6 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

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

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 5.7. 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 $n = n_1 + n_2$):

$$x_1, \ldots, x_{n_1} \sim N(\mu_1, \sigma^2)$$

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:

[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

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:

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

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 estimated means are related by

$$\hat{\mu}_0 = \frac{n_1}{n}\hat{\mu}_1 + \frac{n_2}{n}\hat{\mu}_2$$

so the overall mean is the weighted average of the two individual group means.

Moreover, the two estimated variances are related by

$$\begin{aligned}\widehat{\sigma_0^2} &= \widehat{\sigma^2} + \frac{n_1 n_2}{n^2}(\hat{\mu}_1 - \hat{\mu}_2)^2 \\ &= \widehat{\sigma^2}\left(1 + \frac{1}{n}\frac{(\hat{\mu}_1 - \hat{\mu}_2)^2}{\frac{n}{n_1 n_2}\widehat{\sigma^2}}\right) \\ &= \widehat{\sigma^2}\left(1 + \frac{t_{ML}^2}{n}\right)\end{aligned}$$

with

$$t_{ML} = \frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\widehat{\sigma^2}}}$$

This is an example of a variance decomposition, with

- $\widehat{\sigma_0^2}$ being the estimated total variance,
- $\widehat{\sigma^2}$ the estimated within-group variance and
- $\widehat{\sigma^2}\frac{t_{ML}^2}{n} = \frac{n_1 n_2}{n^2}(\hat{\mu}_1 - \hat{\mu}_2)^2$ the estimated between-group variance.

and

- $\frac{\widehat{\sigma^2}}{\widehat{\sigma_0^2}} = 1 + \frac{t_{ML}^2}{n}$

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:

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

The last step uses the decomposition for the total variance $\widehat{\sigma_0^2}$. If an unbiased estimate of the variance is used ($\widehat{\sigma^2}_{\text{UB}} = \frac{n}{n-2}\widehat{\sigma^2}$) rather than the MLE then

$$W = n\log\left(1 + \frac{1}{n-2}t^2\right)$$

with

$$t = \frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \widehat{\sigma^2}_{\text{UB}}}}$$

$\longrightarrow$ the GRLT is a monotone function 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^2_1. Hence, we reject the null model on 5% significance level for all $W > 3.84$.

More generally, it turns out that not just the two-sample t-test but many other commonly used familiar statistical tests can be interpreted as GLRTs!

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 in 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.

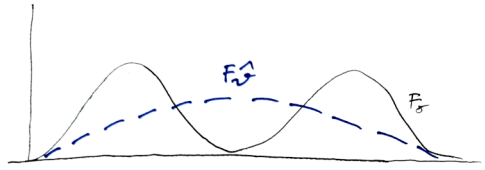
However, since the KL divergence is not symmetric there are in fact two ways to minimise the divergence between a fixed F_0 and the family $F_{\boldsymbol{\theta}}$, each with different properties:

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.

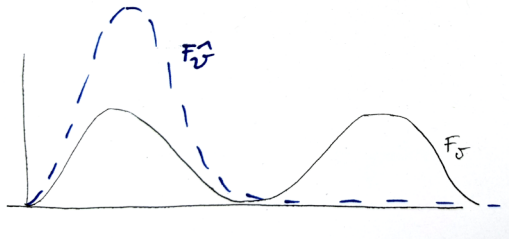


b) **reverse KL, inference KL**: $\min_{\theta} D_{\mathrm{KL}}(F_{\theta}, F_0)$

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

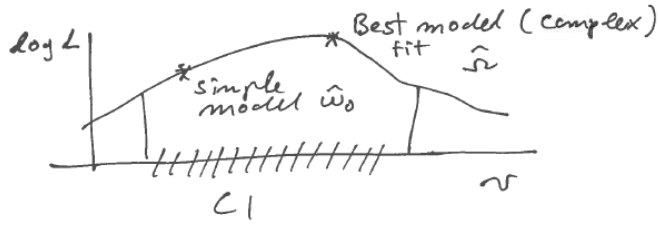
which are essentially two sides of the same coin.

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

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

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 (also known as "pooled" 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

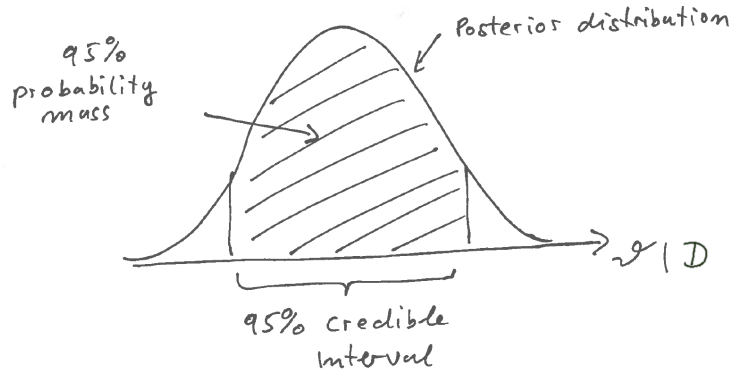
The Bayesian estimate is the full complete posterior distribution!

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 paper on the Stein estimator — an estimator of the mean that dominates ML estimator. His estimator is always better in terms of MSE than the ML estimator, and this was very puzzling at that time!
- Only from the 1950s the use of the term "Bayesian statistics" became prevalent — see Fienberg (2006)[3]

Due to advances in personal computing from 1970 onwards Bayesian learning has become more pervasive!

[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

- 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.
- work on reference priors (Bernado 1979).
- EM algorithm published in 1977 which 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 that came arose as generalisation of the EM algorithm and are also linked to models and methods from 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/ binomial model as in the analogous example in Part I:

Repeated Bernoulli experiment (binomial model):

$x \in \{0, 1\}$ (e.g. "tails" vs. "heads")

probability mass function (pmf): $\Pr(x = 1) = p$, $\Pr(x = 0) = 1 - p$
Mean: $\mathrm{E}(x) = p$
Variance $\mathrm{Var}(x) = p(1 - p)$

$\mathrm{Bin}(n, p)$ (sum of n Bernoulli experiments)
$x \in \{0, 1, \ldots, n\}$
Mean: $\mathrm{E}(x) = np$
Variance: $\mathrm{Var}(x) = np(1 - p)$

Standardised binomial (average of n Bernoulli experiments):
$\frac{x}{n} \in \{0, \frac{1}{n}, \ldots, 1\}$
Mean: $\mathrm{E}(\frac{x}{n}) = p$
Variance: $\mathrm{Var}(\frac{x}{n}) = \frac{p(1-p)}{n}$

From part I (likelihood theory) we know that the *maximum likelihood estimate* of the proportion is the frequency $\hat{p}_{ML} = \frac{x}{n}$ given x (number of "heads") is

observed in n repeats.

10.1.2 Excursion: Properties of the Beta distribution

The density of the Beta distribution Beta(α, β) for $x \in [0, 1]$ and $\alpha > 0$ and $\beta > 0$ is

$$f(x|\alpha, \beta) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

The mean is $\mathrm{E}(x) = \mu = \frac{\alpha}{\alpha+\beta}$ and the variance $\mathrm{Var}(x) = \frac{\mu(1-\mu)}{\alpha+\beta+1}$.

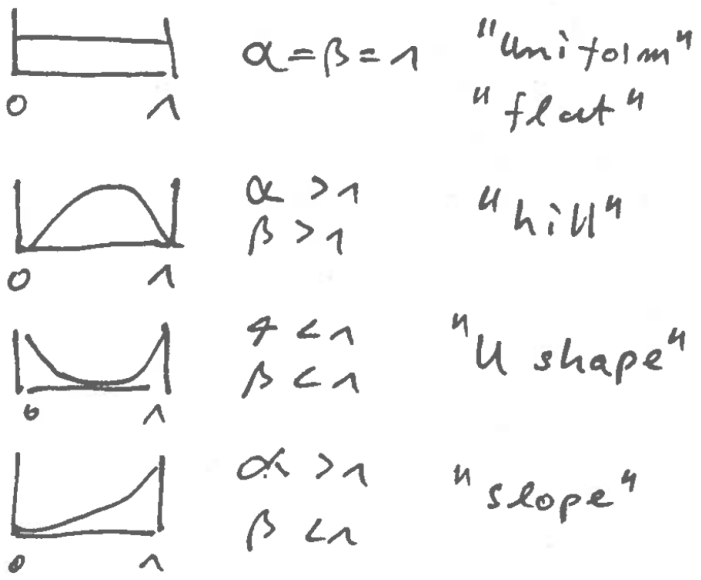
The density depends on the Beta function $B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ which in turn is defined via Euler's Gamma function

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

Note that $\Gamma(x) = (x-1)!$ for any positive integer x

A useful reparameterisation of the Beta distribution is in terms of the parameters $\mu \in [0, 1]$ and $m > 0$, yielding the original parameters via $\alpha = \mu m$ and $\beta = (1-\mu)m$. Conversely, $m = \alpha + \beta$ and $\mu = \frac{\alpha}{\alpha+\beta}$.

The Beta distribution is very flexible and can assume a number of different shapes, depending on the value of α and β:



10.1.3 Beta prior distribution

In Bayesian learning we need to make explicit our uncertainty about p.

p has support $[0, 1] \rightarrow$ we use the **Beta distribution Beta**(α, β) **as prior for** p with parameters $\alpha \geq 0$ and $\beta \geq 0$:

$$p \sim \text{Beta}(\alpha, \beta)$$

Note this does not actually mean that p is random! It only means that we model the uncertainty about p using a Beta random variable!

The flexibility of the Beta distribution allows to accomodate a large variety of possible scenarios for our prior knowledge.

The prior mean is

$$\text{E}(p) = \frac{\alpha}{m} = \mu_{\text{prior}}$$

and the prior variance

$$\text{Var}(p) = \frac{\mu_{\text{prior}}(1 - \mu_{\text{prior}})}{m + 1}$$

where $m = \alpha + \beta$.

Note the similarity to the moments of the standardised binomial above!

10.1.4 Computing the posterior distribution

Bayes' theorem for continuous random variables to compute posterior density:

$$f(p|x) = \frac{f(x|p)f(p)}{\int_{p'} f(x|p')f(p')dp'}$$

We use in our analysis the Beta-Binomial model:

a) **Beta prior:**

$$p \sim \text{Beta}(\alpha, \beta)$$

$$f(p) = \frac{1}{B(\alpha, \beta)} p^{\alpha-1}(1 - p)^{\beta-1}$$

b) **Binomial likelihood:**

$$x|p \sim \text{Bin}(n, p)$$

$$f(x|p) = \binom{n}{x} p^x (1 - p)^{(n-x)}$$

Applying Bayes' theorem results in

c) **Beta posterior distribution**

$$p|x \sim \text{Beta}(\alpha + x, \beta + n - x)$$

$$f(p|x) = \frac{1}{B(\alpha + x, \beta + n - x)} p^{\alpha+x-1}(1-p)^{\beta+n-x-1}$$

(for a proof see Worksheet B1!)

The posterior can be summarised by its first two moments (mean and variance):

Posterior mean:

$$\mu_{\text{posterior}} = \text{E}(p|x) = \frac{x + \alpha}{n + m}$$

Posterior variance:

$$\sigma^2_{\text{posterior}} = \text{Var}(p|x) = \frac{\mu_{\text{posterior}}(1 - \mu_{\text{posterior}})}{n + m + 1}$$

10.2 Properties of Bayesian learning

The Beta-Binomial models allows to observe a number of intriguing features and properties of Bayesian learning. Many of these extend also to other models as we will see later.

10.2.1 Prior acting as pseudodata

In the expression for the posterior mean and variance you can see that $m = \alpha + \beta$ behaves like an implicit sample size connected with prior information!

Specifically, α and β act as **pseudocounts** that influence both the posterior mean and the posterior variance, exactly in the same way as conventional data.

For example, the larger m (and thus larger α and β) the smaller is the posterior variance, with variance decreasing proportional to the inverse of m. If the prior is highly concentrated, i.e. if it has low variance and large precision (=inverse variance) then the implicit data size m is large. Conversely, if the prior has a large variance, then the prior is vague and the implicit data size m 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 effective sample size.

Another interpretation is that any prior summarises data that may have been available previously as observations.

10.2.2 Linear shrinkage of mean

The posterior mean $\mu_{\text{posterior}}$ is a linearly adjusted $\hat{\mu}_{ML}$. This becomes evident by writing $\mu_{\text{posterior}}$ as

$$\mu_{\text{posterior}} = \lambda \mu_{\text{prior}} + (1 - \lambda)\hat{\mu}_{ML}$$

with weight $\lambda \in [0, 1]$

$$\lambda = \frac{m}{m + n}.$$

The **posterior mean is a convex combination (i.e. the weighted average) of the ML estimate and the prior mean**. The factor λ is called the **shrinkage intensity** — note that it is the ratio of the "prior sample size" (m) and the "effective overall sample size" ($m + n$).

1. This is called *shrinkage*, because the ML estimator is "shrunk" towards the prior mean (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 implies $\alpha = 0$ and $\beta = 0$, or $n \to \infty$.

   Note that using maximum likelihood to estimate the proportion p (for moderate or small n) is the same as Bayesian posterior mean estimation using the Beta-Binomial model with prior $\alpha = 0$ and $\beta = 0$. 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 m is very large (implying that the prior is sharply concentrated around the prior mean).

4. Since the ML estimate $\hat{\mu}_{ML}$ 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, Bayesian learning typically produces biased estimators (but asymptotically they will be unbiased like in ML).

5. That the posterior mean is a linear combination of the ML estimate 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].

6. Furthermore, it is possible (and indeed quite useful for computational reasons!) to formulate Bayes learning in terms of linear shrinkage and using only second moments, see e.g. Hartigan (1969)[2]. The resulting theory

[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

is called "Bayes linear statistics" (Goldstein and Wooff, 2007)[3].

10.2.3 Conjugacy of prior and posterior distribution

In the Beta-Binomial model for estimating the proportion p 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.

In the Beta-Binomial model the likelihood is based on the binomial distribution and has the following form (only terms depending on the parameter p are shown):

$$p^x(1-p)^{n-x}$$

The form of the Beta prior is (again, only showing terms depending on p):

$$p^{\alpha-1}(1-p)^{\beta-1}$$

Since the posterior is proportional to the product of prior and likelihood the posterior will have exactly the same form as the prior:

$$p^{\alpha+x-1}(1-p)^{\beta+n-x-1}$$

Choosing the prior distribution from a family conjugate to 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:

$$\alpha \longrightarrow \alpha + x$$

$$\beta \longrightarrow \beta + n - x$$

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 $n >> \alpha, \beta$ the posterior mean and variance become asympotically

$$\mu_{\text{posterior}} \stackrel{a}{=} \frac{x}{n} = \hat{\mu}_{ML}$$

[3]Goldstein, M., and D. Wooff. 2007. *Bayes Linear Statistics: Theory and Methods*. Wiley. https://doi.org/10.1002/9780470065662

and

$$\sigma^2_{\text{posterior}} \overset{a}{=} \frac{\hat{\mu}_{ML}(1-\hat{\mu}_{ML})}{n}$$

Thus, if sample size is large 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 values of α and β 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 centered around the MLE and with the variance of the MLE:

$$\text{for large } n\text{: } p(\boldsymbol{\theta}|\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n) \rightarrow N(\hat{\boldsymbol{\theta}}_{ML}, \text{Var}(\hat{\boldsymbol{\theta}}_{ML}))$$

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 our example of the Beta-Bernoulli model.

The Bayesian CLT is generally known as the **Bernstein-van 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 grapically.

10.2.6 Posterior variance for finite n

Previously we have the derived a Bayesian point estimate for the proportion p as the posterior mean

$$\text{E}(p|x) = \frac{x+\alpha}{n+m} = \hat{p}_{\text{Bayes}}$$

with posterior variance

$$\text{Var}(p|x) = \frac{\hat{p}_{\text{Bayes}}(1-\hat{p}_{\text{Bayes}})}{n+m+1}$$

Asymptotically, we have seen that 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 estimate will be indistinguishable from the MLE and shares its favourable properties.

In addition, for finite sample size the posterior variance will tyically be *smaller* than both the asymptotic posterior variance (for large n) and the prior variance, showing that combining the information 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

For the **likelihood** we assume as data-generating model the normal distribution with known fixed variance σ^2

$$x|\mu \sim N(\mu, \sigma^2)$$

This yields as the MLE $\hat{\mu}_{ML} = \bar{x}$.

10.3.2 Normal prior distribution

To model the uncertainty about μ we use the normal distribution $N(\mu, \sigma^2/k)$ parameterised by the two parameters μ and k (remember σ^2 is fixed).

With $\mu = \mu_0$ and $k = m$ we get the **normal prior**

$$\mu \sim N(\mu_0, \sigma^2/m)$$

with prior mean $\mathrm{E}(\mu) = \mu_0$ and prior variance $\mathrm{Var}(\mu) = \frac{\sigma^2}{m}$ where m is the implied sample size from the prior. Note that m does not need to be an integer value!

10.3.3 Normal posterior distribution

The **posterior distribution** after observing n samples $x_1, \ldots x_n$ is normal with $\mu = \mu_1$ and $k = m + n$

$$\mu|x_1, \ldots x_n \sim N(\mu_1, \sigma^2/(m+n))$$

with posterior mean

$$\mathrm{E}(\mu|x_1, \ldots x_n) = \mu_1 = \frac{m\mu_0 + n\bar{x}}{n+m} = \lambda\mu_0 + (1-\lambda)\hat{\mu}_{ML}$$

with $\lambda = \frac{m}{n+m}$. Note the linear shrinkage of $\hat{\mu}_{ML}$ towards μ_0!

The corresponding posterior variance is

$$\text{Var}(\mu|x_1, \dots x_n) = \frac{\sigma^2}{n+m}$$

Thus, the **normal distribution is the conjugate distribution to the mean parameter in the normal likelihood**.

10.3.4 Large sample asymptotics and Stein paradox

For n large and $n >> m$ we get

$$\text{E}(\mu|x_1, \dots x_n) \stackrel{a}{=} \hat{\mu}_{ML}$$

$$\text{Var}(\mu|x_1, \dots x_n) \stackrel{a}{=} \frac{\sigma^2}{n}$$

i.e. the MLE and its asymptotic variance!

Note that the posterior variance $\frac{\sigma^2}{n+m}$ is smaller than the asymptotic variance $\frac{\sigma^2}{n}$ and the prior variance $\frac{\sigma^2}{m}$.

10.4 Estimating the variance using the inverse-Gamma-Normal model

10.4.1 Inverse Gamma distribution

Next, we study a common Bayesian model for estimating the variance parameter of the normal distribution. For this we use the inverse Gamma distribution:

$$x \sim \text{Inv-Gam}(\alpha, \beta)$$

This distribution is closely linked with the Gamma distribution — the inverse of x is Gamma-distributed with inverted scale parameter:

$$\frac{1}{x} \sim \text{Gam}(\alpha, \beta^{-1})$$

For use as prior and posterior we employ a different parameterisation with $\mu = \beta/(\alpha - 1)$ and $k = 2(\alpha - 1)$:

$$x \sim \text{Inv-Gam}\left(\alpha = \frac{k+2}{2}, \beta = \frac{k\mu}{2}\right) = \text{Inv-Gam}(\mu, k)$$

The reason for choosing the mean parameterisation using μ and k instead of α and β is that this parameterisation simplifies the Bayesian update rule for the mean.

The first two moments of the inverse Gamma distribution are

$$\mathrm{E}(x) = \frac{\beta}{\alpha - 1} = \mu$$

and

$$\mathrm{Var}(x) = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)} = \frac{2\mu^2}{k-2}$$

The inverse Gamma distribution is also known under two further alternative names: 1) inverse scaled chi-squared distribution and 2) one-dimensional inverse Wishart distribution.

10.4.2 Normal likelihoood

As data likelihood / generating model we use normal distribution $N(\mu, \sigma^2)$ with given fixed mean μ.

This yields as MLE $\widehat{\sigma}^2_{ML} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$

10.4.3 Inverse Gamma prior distribution

For the prior distribution we use the inverse Gamma distribution with $k = m$ and $\mu = \sigma_0^2$

$$\sigma^2 \sim \text{Inv-Gam}(\mu = \sigma_0^2, k = m)$$

The corresponding prior mean is

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

and the prior variance is

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

(note that $m > 2$)

10.4.4 Inverse Gamma posterior distribution

As the inverse Gammma distribution is conjugate to the normal likelihood the posterior distribution is inverse Gamma as well:

$$\sigma^2 | x_1 \ldots, x_n \sim \text{Inv-Gam}(\mu = \sigma_1^2, k = m + n)$$

with $\sigma_1^2 = \frac{\sigma_0^2 m + n\widehat{\sigma}^2_{ML}}{m+n}$.

The posterior mean is

$$\mathrm{E}(\sigma^2 | x_1 \ldots, x_n) = \sigma_1^2$$

and the posterior variance

$$\text{Var}(\sigma^2|x_1 \ldots, x_n) = \frac{2\sigma_1^4}{m+n-2}$$

The update formula for the posterior mean of the variance follows the usual linear shrinkage rule:

$$\sigma_1^2 = \lambda\sigma_0^2 + (1-\lambda)\widehat{\sigma}_{ML}^2$$

with $\lambda = \frac{m}{m+n}$.

10.4.5 Large sample asymptotics

For n large and $n >> m$ we get

$$\text{E}(\sigma^2|x_1, \ldots x_n) \stackrel{a}{=} \widehat{\sigma}_{ML}^2$$

$$\text{Var}(\sigma^2|x_1, \ldots x_n) \stackrel{a}{=} \frac{2\sigma^4}{n}$$

which is indeed the MLE of σ^2 and its asymptotic variance!

10.4.6 Gamma-Normal model for the precision

Instead of estimating the variance we may wish to estimate the precision, i.e. the inverse variance. In the above we have used an inverse Gamma distribution for the prior and posterior of the variance. Thus, to model the precision we may therefore use a Gamma prior distribution and a normal likelihood, resulting in a Gamma posterior distribution.

10.4.7 Joint estimation of mean and variance

It is possible to combine the Normal-Normal for the mean and the Inverse-Gamma-Normal model into a joint model for the mean and variance.

This implies having a joint prior and a joint posterior for μ and σ^2.

The resulting joint point estimates are identical to the above individual estimates.

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 with fixed mean and variance, or a Binomial distribution with a set 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 with a given mean but unspecified variance, or a Binomial model with unspecified parameter p, 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 p. If p follows a Beta distribution then the marginal likelihood with p integrated out is the Beta-Binomial distribution (see also Worksheet B2). 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 Inverse-Gamma-Norma 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

$$\hat{\theta}_{\text{shrink}} = \text{E}(\theta|x_1, \ldots, x_n) = \lambda\theta_0 + (1-\lambda)\hat{\theta}_{\text{ML}}$$

of the MLE $\hat{\theta}_{\text{ML}}$ towards the prior mean θ_0, with shrinkage intensity $\lambda = \frac{m}{m+n}$ determined by the parameter m (the implicit sample size) and the sample size n.

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 hyper-parameter in this setting is m (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 interests by probability
- In the light of new evidence (observed data) the uncertainty is updated, i.e. the prior distribution is combined with the likelihood to form the posterior distribution

Example: Beta-Binomial model

- Binomial likelihood
- n observations: x "heads", $n - x$ "tails"
- Frequency $\hat{\theta}_{ML} = \frac{x}{n}$
- Beta prior $\theta \sim \text{Beta}(\alpha_0, \beta_0)$ with mean $\theta_0 = \frac{\alpha_0}{m}$ and $m = \alpha_0 + \beta_0$
- Beta posterior $\theta|x, n \sim \text{Beta}(\alpha_1, \beta_1)$ with mean $\theta_1 = \frac{\alpha_1}{\alpha_1+\beta_1}$ and $\alpha_1 = \alpha_0 + x$ and $\beta_1 = \beta_0 + n - x$
- Update of prior mean to posterior mean by shrinkage of MLE:

$$\theta_1 = \lambda\theta_0 + (1 - \lambda)\hat{\theta}_{ML}$$

  with shrinkage intensity $\lambda = \frac{m}{n+m}$
- m can be interpreted as prior sample size

13.1.1 Remarks

- If posterior in same family as prior → conjugate prior.
- In an exponential family the Bayesian update of the mean is always expressible as linear shrinkage of the MLE.

- For sample size $n \to \infty$ then $\lambda \to 0$ and $\theta_1 \to \hat{\theta}_{ML}$ (for large samples posterior mean = maximum likelihood estimator).
- For $n \to 0$ then $\lambda \to 1$ and $\theta_1 \to \hat{\theta}_0$ (if no data is available fall back to prior).
- Note that the Bayesian estimator is biased for finite n by construction (but asymptotically unbiased like the MLE).

13.1.2 Advantages

- Adding prior information has regularisation properties. This is very important in more complex models with many parameters, e.g., in estimation of a covariance matrix (to avoid singularity).
- Improves small-sample accuracy (e.g. MSE)
- Bayesian estimators tend to perform better than MLEs is not surprising - they use the obseved data plus extra information!
- Bayesian credible intervals are conceptually much more simple than frequentist confidence intervals.

13.1.3 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 hyper-parameter λ.

13.1.4 Specifying the prior — problem or advantage?

In Bayesian statistics the analysist 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 $m = 2$ or a transformation-invariant prior Beta$(1/2, 1/2)$ (=Jeffreys prior) with implicit sample size $m = 1$ 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.

Part III

Regression

Chapter 14

Overview over regression modelling

14.1 General setup

- y: **response variable**, also known as **outcome** or **label**
- $x_1, x_2, x_3, \ldots, x_d$: **predictor variables**, also known as **covariates** or **covariables**
- The relationship between the outcomes and the predictor variables is assumed to follow

$$y = f(x_1, x_2, \ldots, x_d) + \varepsilon$$

where f is the **regression function** (not a density) and ε represents **noise**.

14.2 Objectives

1. **Understand the relationship** between the response y and the predictor variables x_i by **learning the regression function** f from observed data (training data). The estimated regression function is $\hat{f}$.
2. **Prediction of outcomes**

$$\underbrace{\hat{y}}_{\substack{\text{predicted response} \\ \text{using fitted } \hat{f}}} = \hat{f}(x_1, x_2, \ldots, x_d)$$

   If instead of the fitted function $\hat{f}$ the known regression function f is used we denote this by

$$\underbrace{y^\star}_{\substack{\text{predicted response} \\ \text{using known } f}} = f(x_1, x_2, \ldots, x_d)$$

3. **Variable importance**
   - which covariates are most relevant in predicting the outcome?
   - allows to better understand the data and model
     → variable selection (to build simpler model with same predictive capability)

14.3 Regression as a form of supervised learning

Regression modelling is a special case of **supervised learning**.

In supervised learning we make use of labelled data, i.e. $\boldsymbol{x}_i$ has an associated *label* y_i. Thus, the data is consists of pairs $(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_n, y_n)$.

The *supervision* part of in supervised learning refers to the fact that the labels are given.

In **regression** typically the label y_i is continuous and called the *response*.

On the other hand, if the label y_i is discrete/categorical then supervised learning is called **classification**.

$$\text{Supervised Learning} \begin{cases} \longrightarrow \text{Discrete } y & \longrightarrow \text{Classification Methods} \\ \longrightarrow \text{Continuous } y & \longrightarrow \text{Regression Methods} \end{cases}$$

Another important type of statistical learning is **unsupervised learning** where labels y are inferred from the data $\boldsymbol{x}$ (this is also known as **clustering**). Furthermore, there is also *semi-supervised learning* with labels only partly known.

Note that there are regression models (e.g. logistic regression) with discrete response that are performing classification, so one may argue that "supervised learning"="generalised regression".

14.4 Various regression models used in statistics

In this course we only study linear multiple regression. However, you should be aware that the linear model is in fact just a special cases of some much more general regression approaches.

General regression model:

$$y = f(x_1, \ldots, x_d) + \text{"noise"}$$

- The function f is estimated nonparametrically - splines - Gaussian processes
- Generalised Additive Models (GAM): - the function f is assumed to be the sum of individual functions $f_i(x_i)$
- Generalised Linear Models (GLM): - f is a transformed linear predictor $h(\sum b_i x_i)$, noise is assumed from an exponential family
- Linear Model (LM): - linear predictor $\sum b_i x_i$, normal noise

In R the linear model is implemented in the function lm(), and generalised linear models in the function glm(). Generalised additive models are available in the package "mgcv".

In the following we focus on the linear regression model with continuous response.

Chapter 15

Linear Regression

15.1 The linear regression model

In this module we assume that f is a linear function:

$$f(x_1, \ldots, x_d) = \beta_0 + \sum_{j=1}^{d} \beta_j x_j = y^\star$$

In vector notation:

$$f(\boldsymbol{x}) = \beta_0 + \boldsymbol{\beta}^T \boldsymbol{x} = y^\star$$

with $\boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_d \end{pmatrix}$ and $\boldsymbol{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix}$

Therefore, the linear regression model is

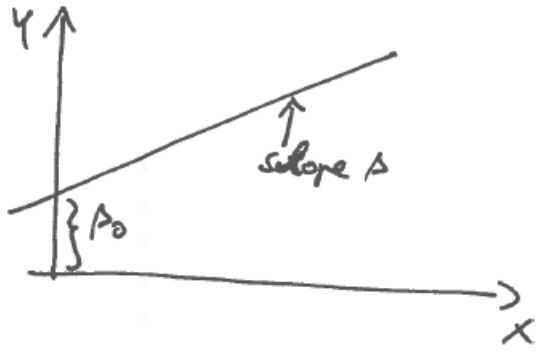
$$\begin{aligned} y &= \beta_0 + \sum_{j=1}^{d} \beta_j x_j + \varepsilon \\ &= \beta_0 + \boldsymbol{\beta}^T \boldsymbol{x} + \varepsilon \\ &= y^\star + \varepsilon \end{aligned}$$

where:

- β_0 is the **intercept**
- $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)^T$ are the **regression coefficients**
- $\boldsymbol{x} = (x_1, \ldots, x_d)^T$ is the predictor vector containing the **predictor variables**

15.2 Interpretation of regression coefficients and intercept

- The regression coefficient β_i corresponds to the slope (first partial derivative) of the regression function in the direction of x_i. In other words, the gradient of $f(\boldsymbol{x})$ are the regression coefficients: $\nabla f(\boldsymbol{x}) = \boldsymbol{\beta}$
- The intercept β_0 is the offset at the origin ($x_1 = x_2 = \ldots = x_d = 0$):



15.3 Different types of linear regression:

- **Simple linear regression**: $y = \beta_0 + \beta x + \varepsilon$ (=single predictor)
- **Multiple linear regression**: $y = \beta_0 + \sum_{j=1}^{d} \beta_j x_j + \varepsilon$ (= multiple predictor variables)
- **Multivariate regression**: multivariate response $\boldsymbol{y}$

15.4 Distributional assumptions and properties

General assumptions:

- We treat y and $x_1, \ldots, x_d$ as the primary observables that can be described by random variables.
- $\beta_0, \boldsymbol{\beta}$ are parameters to be inferred from the observations on y and $x_1, \ldots, x_d$.
- Specifically, will we assume that response and predictors have a mean and a (cov)variance:

  i. Response:
  $\mathrm{E}(y) = \mu_y$
  $\mathrm{Var}(y) = \sigma_y^2$
  The **variance of the response** $\mathrm{Var}(y)$ is also called the **total variation** .

ii. Predictors:
$\mathrm{E}(x_i) = \mu_{x_i}$ (or $\mathrm{E}(\boldsymbol{x}) = \boldsymbol{\mu_x}$)
$\mathrm{Var}(x_i) = \sigma^2_{x_i}$ and $\mathrm{Cor}(x_i, x_j) = \rho_{ij}$ (or $\mathrm{Var}(\boldsymbol{x}) = \boldsymbol{\Sigma_x}$)
The **signal variance** $\mathrm{Var}(y^\star) = \mathrm{Var}(\beta_0 + \boldsymbol{\beta}^T\boldsymbol{x}) = \boldsymbol{\beta}^T\boldsymbol{\Sigma_x}\boldsymbol{\beta}$ is also called the **explained variation**.

- We assume that y and $\boldsymbol{x}$ are jointly distributed with correlation $\mathrm{Cor}(y, x_j) = \rho_{y,x_j}$ between each predictor variable x_j and the response y.
- In contrast to y and $\boldsymbol{x}$ the noise variable ε is only indirectly observed via the difference $\varepsilon = y - y^\star$. We denote the mean and variance of the noise by $\mathrm{E}(\varepsilon)$ and $\mathrm{Var}(\varepsilon)$.
The **noise variance** $\mathrm{Var}(\varepsilon)$ is also called the **unexplained variation** or the **residual variance**. The **residual standard error** is $\mathrm{SD}(\varepsilon)$.

Identifiability assumptions:

In a statistical analysis we would like to be able to separate signal ($y^\star$) from noise (ε). To achieve this we require some **distributional assumptions to ensure identifiability** and avoid confounding:

1) **Assumption 1:** ε and $y^\star$ are are independent. This implies $\mathrm{Var}(y) = \mathrm{Var}(y^\star) + \mathrm{Var}(\varepsilon)$, or equivalently $\mathrm{Var}(\varepsilon) = \mathrm{Var}(y) - \mathrm{Var}(y^\star)$.

   Thus, this assumption implies the **decomposition of variance**, i.e. that the **total variation** $\mathrm{Var}(y)$ equals the sum of the **explained variation**$\mathrm{Var}(y^\star)$ and the **unexplained variation**$\mathrm{Var}(\varepsilon)$.

2) **Assumption 2:** $\mathrm{E}(\varepsilon) = 0$. This allows to identify the intercept β_0 and implies $\mathrm{E}(y) = \mathrm{E}(y^\star)$.

Optional assumptions (often but not always):

- The noise ε is normally distributed
- The response y and and the predictor variables x_i are continuous variables
- The response and predictor variables are jointly normally distributed

Further properties:

- As a result of the independence assumption 1) we can only choose two out of the three variances freely:
  i. in a generative perspective we will choose signal variance $\mathrm{Var}(y^\star)$ (or equivalently the variances $\mathrm{Var}(x_j)$) and the noise variance $\mathrm{Var}(\varepsilon)$, then the variance of the response $\mathrm{Var}(y)$ follows.
  ii. in an observational perspective we will observe the variance of the reponse $\mathrm{Var}(y)$ and the variances $\mathrm{Var}(x_j)$, and then the error variance $\mathrm{Var}(\varepsilon)$ follows.
- As we will see later the regression coefficients β_j depend on the correlations between the response y and and the predictor variables x_j. Thus, the choice of regression coefficients implies a specific correlation pattern, and vice

versa (in fact, we will use this correlation pattern to infer the regression coefficients from data!).

15.5 Regression in data matrix notation

We can also write the regression in terms of actual observed data (rather than in terms of random variables):

Data matrix for the predictors:

$$\boldsymbol{X} = \begin{pmatrix} x_{11} & \dots & x_{1d} \\ \vdots & \ddots & \vdots \\ x_{n1} & \dots & x_{nd} \end{pmatrix}$$

Note the statistics convention: the n rows of $\boldsymbol{X}$ contain the samples, and the d columns contain variables.

Response data vector: $(y_1, \dots, y_n)^T = \boldsymbol{y}$

Then the regression equation is written in data matrix notation:

$$\underbrace{\boldsymbol{y}}_{n\times 1} = \underbrace{\mathbf{1}_n \beta_0}_{n\times 1} + \underbrace{\boldsymbol{X}}_{n\times d} \underbrace{\boldsymbol{\beta}}_{d\times 1} + \underbrace{\underbrace{\boldsymbol{\varepsilon}}_{n\times 1}}_{\text{residuals}}$$

where $\mathbf{1}_n = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$ is a column vector of length n (size $n \times 1$).

Note that here the regression coefficients are now multiplied *after* the data matrix (compare with the original vector notation where the *transpose* of regression coefficients come *before* the vector of the predictors).

The **observed noise** values (i.e. realisations of the random variable ε) are called the **residuals**.

15.6 Centering and vanishing of the intercept β_0

If $\boldsymbol{x}$ and y are centered, i.e. if $\mathrm{E}(\boldsymbol{x}) = \boldsymbol{\mu_x} = 0$ and $\mathrm{E}(y) = \mu_y = 0$, then the intercept β_0 disappears:

The regression equation is

$$y = \beta_0 + \boldsymbol{\beta}^T \boldsymbol{x} + \varepsilon$$

with $E(\varepsilon)$. Taking the expectation on both sides we get $\mu_y = \beta_0 + \boldsymbol{\beta}^T \boldsymbol{\mu}_{\boldsymbol{x}}$ and therefore

$$\beta_0 = \mu_y - \boldsymbol{\beta}^T \boldsymbol{\mu}_{\boldsymbol{x}}$$

This is zero if the mean of the response μ_y and the mean of predictors $\boldsymbol{\mu}_{\boldsymbol{x}}$ vanish. Conversely, if we assume that the intercept vanishes ($\beta_0 = 0$) this is only possible for general $\boldsymbol{\beta}$ if both $\boldsymbol{\mu}_{\boldsymbol{x}} = 0$ and $\mu_y = 0$.

Thus, in the linear model is always possible to transform y and $\boldsymbol{x}$ (or data $\boldsymbol{y}$ and $\boldsymbol{X}$) so that the intercept vanishes. To simplify equations we will therefore often set $\beta_0 = 0$.

15.7 Objectives in data analysis using linear regression

1. Understand functional relationship: find estimates of the intercept ($\hat{\beta}_0$) and the regression coefficients ($\hat{\beta}_j$), as well as the associated errors.
2. Prediction:
   - Known coefficients β_0 and $\boldsymbol{\beta}$: $y^\star = \beta_0 + \boldsymbol{\beta}^T \boldsymbol{x}$
   - Estimated coefficients $\hat{\beta}_0$ and $\hat{\boldsymbol{\beta}}$ (note the "hat"!): $\hat{y} = \hat{\beta}_0 + \sum_{j=1}^{d} \hat{\beta}_j x_j = \hat{\beta}_0 + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}$

   For each point prediction find the **corresponding prediction error!**
3. Variable importance: Which predictors x_j are most relevant?
   $\rightarrow$ test whether $\beta_j = 0$
   $\rightarrow$ find measures of variable importance

   Remark: as we will see β_j or $\hat{\beta}_j$ itself is **not** a measure of variable importance!

Chapter 16

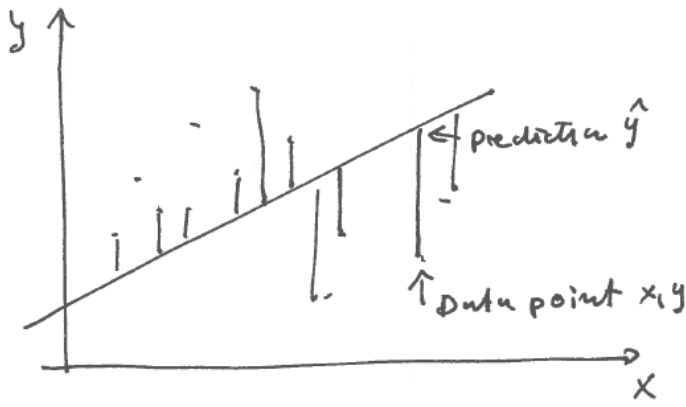
Estimating regression coefficients

In this chapter we discuss various ways to estimate the regression coefficients. First, we discuss estimation by Ordinary Least Squares (OLS) by minimising the residual sum of squares. This yields the famous Gauss estimator. Second, we derive estimates of the regression coefficients using the methods of maximum likelihood assuming normal errors. This also leads to the Gauss estimator. Third, we show that the coefficients in linear regression can written and interpreted in terms of two covariance matrices and that the Gauss estimator of the regression coefficients is a plug-in estimator using the MLEs of these covariance matrices. Furthermore, we show that the (population version) of the Gauss estimator can also be derived by finding the best linear predictor and by conditioning. Finally, we discuss special cases of regression coefficients and their relationship to marginal correlation.

16.1 Ordinary Least Squares (OLS) estimator of regression coefficients

Now we show the classic way (Gauss 1809; Legendre 1805) to **estimate regression coefficients** by the method of **ordinary least squares (OLS)**.

Idea: choose regression coefficients such as to *minimise* the *squared error* between observations and the prediction.



In data matrix notation (note we assume $\beta_0 = 0$ and thus *centered data* $\boldsymbol{X}$ and $\boldsymbol{y}$):

$$\text{RSS}(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$$

RSS is an abbreviation for "Residual Sum of Squares" which is is a function of $\boldsymbol{\beta}$. Minimising RSS yields the OLS estimate:

$$\widehat{\boldsymbol{\beta}}_{\text{OLS}} = \arg\min_{\boldsymbol{\beta}} \text{RSS}(\boldsymbol{\beta})$$

$$\text{RSS}(\boldsymbol{\beta}) = \boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}$$

Gradient:

$$\nabla\text{RSS}(\boldsymbol{\beta}) = -2\boldsymbol{X}^T\boldsymbol{y} + 2\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}$$

$$\nabla\text{RSS}(\widehat{\boldsymbol{\beta}}) = 0 \longrightarrow \boldsymbol{X}^T\boldsymbol{y} = \boldsymbol{X}^T\boldsymbol{X}\widehat{\boldsymbol{\beta}}$$

$$\Longrightarrow \widehat{\boldsymbol{\beta}}_{\text{OLS}} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

Note the similarities in the procedure to maximum likelihood (ML) estimation (with minimisation instead of maximisation)! In fact, as we see next this is not by chance as OLS *is* indeed a special case of ML! This also implies that OLS is generally a good method — but only if sample size n is large!

The above Gauss' estimator is fundamental in statistics so it is worthwile to memorise it!

16.2 Maximum likelihood estimation of regression coefficients

16.2.1 Normal log-likelihood function for regression coefficients and noise variance

We now show how to estimate regression coefficients using the method of maximum likelihood. This is a second method to derive $\hat{\boldsymbol{\beta}}$.

We recall the basic regression equation

$$y = \beta_0 + \boldsymbol{\beta}^T \boldsymbol{x} + \varepsilon$$

with independent noise ε and observed data $y_1, \ldots, y_n$ and $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$.

Assuming $\mathrm{E}(\varepsilon) = 0$ the intercept is identified as

$$\beta_0 = \mu_y - \boldsymbol{\beta}^T \boldsymbol{\mu}_{\boldsymbol{x}}$$

Combining the two above equations we see that noise variable equals

$$\varepsilon = (y - \mu_y) - \boldsymbol{\beta}^T (\boldsymbol{x} - \boldsymbol{\mu}_{\boldsymbol{x}})$$

Assuming joint (multivariate) normality for the observed data, the response y and predictors $\boldsymbol{x}$, we get as the MLEs for the respective means and (co)variances:

- $\hat{\mu}_y = \hat{\mathrm{E}}(y) = \frac{1}{n} \sum_{i=1}^n y_i$
- $\hat{\sigma}_y^2 = \widehat{\mathrm{Var}}(y) = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{\mu}_y)^2$
- $\hat{\boldsymbol{\mu}}_{\boldsymbol{x}} = \hat{\mathrm{E}}(\boldsymbol{x}) = \frac{1}{n} \sum_{i=1}^n \boldsymbol{x}_i$
- $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{xx}} = \widehat{\mathrm{Var}}(\boldsymbol{x}) = \frac{1}{n} \sum_{i=1}^n (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_{\boldsymbol{x}})(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_{\boldsymbol{x}})^T$
- $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}y} = \widehat{\mathrm{Cov}}(\boldsymbol{x}, y) = \frac{1}{n} \sum_{i=1}^n (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_{\boldsymbol{x}})(y_i - \hat{\mu}_y)$

Note that these are are sufficient statistics and hence summarize perfectly the observed data for $\boldsymbol{x}$ and y under the normal assumption

Consequently, the residuals (indirect observations of the noise variable) for a given choice of regression coefficients $\boldsymbol{\beta}$ and the observed data for $\boldsymbol{x}$ and y are

$$\varepsilon_i = (y_i - \hat{\mu}_y) - \boldsymbol{\beta}^T (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_{\boldsymbol{x}})$$

Assuming that the noise $\varepsilon \sim N(0, \sigma_\varepsilon^2)$ is normally distributed with mean 0 and variance $\mathrm{Var}(\varepsilon) = \sigma_\varepsilon^2$. we can write down the normal log-likelihood function for σ_ε^2 and $\boldsymbol{\beta}$:

$$\log L(\boldsymbol{\beta}, \sigma_\varepsilon^2) = -\frac{n}{2} \log \sigma_\varepsilon^2 - \frac{1}{2\sigma_\varepsilon^2} \sum_{i=1}^n \left((y_i - \hat{\mu}_y) - \boldsymbol{\beta}^T (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_{\boldsymbol{x}}) \right)^2$$

Maximising this function leads to the MLEs of σ_ε^2 and $\boldsymbol{\beta}$!

Note that the residual sum of squares appears in the log-likelihood function (with a minus sign), which implies that ML assuming normal distribution will recover the OLS estimator for the regression coefficients! So OLS is a special case of ML !

16.2.2 Detailed derivation of the MLEs

The gradient with regard to $\boldsymbol{\beta}$ is

$$\begin{aligned}\nabla_\beta \log L(\boldsymbol{\beta}, \sigma_\varepsilon^2) &= \frac{1}{\sigma_\varepsilon^2} \sum_{i=1}^{n} \left((\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_{\boldsymbol{x}})(y_i - \hat{\mu}_y) - (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_{\boldsymbol{x}})(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_{\boldsymbol{x}})^T \boldsymbol{\beta} \right) \\ &= \frac{n}{\sigma_\varepsilon^2} \left(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}y} - \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}\boldsymbol{x}} \boldsymbol{\beta} \right)\end{aligned}$$

Setting this equal to zero yields the Gauss estimator

$$\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}\boldsymbol{x}}^{-1} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}y}$$

By plugin we the get the MLE of β_0 as

$$\hat{\beta}_0 = \hat{\mu}_y - \hat{\boldsymbol{\beta}}^T \hat{\boldsymbol{\mu}}_{\boldsymbol{x}}$$

Taking the derivative of $\log L(\hat{\boldsymbol{\beta}}, \sigma_\varepsilon^2)$ with regard to σ_ε^2 yields

$$\frac{\partial}{\partial \sigma_\varepsilon^2} \log L(\hat{\boldsymbol{\beta}}, \sigma_\varepsilon^2) = -\frac{n}{2\sigma_\varepsilon^2} + \frac{1}{2\sigma_\varepsilon^4} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

with $\hat{y}_i = \hat{\beta}_0 + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i$ and the residuals $y_i - \hat{y}_i$ resulting from the fitted linear model. This leads to the MLE of the noise variance

$$\widehat{\sigma_\varepsilon^2} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Note that the MLE $\widehat{\sigma_\varepsilon^2}$ is a biased estimate of σ_ε^2. The unbiased estimate is $\frac{1}{n-d-1} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$, where d is the dimension of $\boldsymbol{\beta}$ (i.e. the number of predictors).

16.2.3 Asymptotics

The advantage of using maximum likelihood is that we also get the (asympotic) variance associated with each estimator and typically can also assume asymptotic normality.

Specifically, for $\hat{\boldsymbol{\beta}}$ we get via the observed Fisher information at the MLE an asymptotic estimator of its variance

$$\widehat{\mathrm{Var}}(\widehat{\boldsymbol{\beta}}) = \frac{1}{n}\widehat{\sigma_\varepsilon^2}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{xx}}^{-1}$$

Similarly, for $\hat{\beta}_0$ we have

$$\widehat{\mathrm{Var}}(\widehat{\beta}_0) = \frac{1}{n}\widehat{\sigma_\varepsilon^2}(1 + \hat{\boldsymbol{\mu}}^T\hat{\boldsymbol{\Sigma}}_{\boldsymbol{xx}}^{-1}\hat{\boldsymbol{\mu}})$$

For finite sample size n with known $\mathrm{Var}(\varepsilon)$ one can show that the variances are

$$\mathrm{Var}(\widehat{\boldsymbol{\beta}}) = \frac{1}{n}\sigma_\varepsilon^2\hat{\boldsymbol{\Sigma}}_{\boldsymbol{xx}}^{-1}$$

and

$$\mathrm{Var}(\widehat{\beta}_0) = \frac{1}{n}\sigma_\varepsilon^2(1 + \hat{\boldsymbol{\mu}}_{\boldsymbol{x}}^T\hat{\boldsymbol{\Sigma}}_{\boldsymbol{xx}}^{-1}\hat{\boldsymbol{\mu}}_{\boldsymbol{x}})$$

and that the regression coefficients and the intercept are normally distributed according to

$$\widehat{\boldsymbol{\beta}} \sim N_d(\boldsymbol{\beta}, \mathrm{Var}(\widehat{\boldsymbol{\beta}}))$$

and

$$\widehat{\beta}_0 \sim N(\beta_0, \mathrm{Var}(\widehat{\beta}_0))$$

We may use this to test whether whether $\beta_j = 0$ and $\beta_0 = 0$.

16.3 Covariance plug-in estimator of regression coefficients

16.3.1 Regression coeffients as product of variances

We now try to understand regression coefficients in terms of covariances (thus obtaining a third way to compute and estimate them).

We recall that the Gauss regression coefficients are given by

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

where $\boldsymbol{X}$ is the $n \times d$ data matrix (in statistics convention)

$$\boldsymbol{X} = \begin{pmatrix} x_{11} & \dots & x_{1d} \\ \vdots & \ddots & \vdots \\ x_{n1} & \dots & x_{nd} \end{pmatrix}$$

Note that we assume that the data matrix $\boldsymbol{X}$ is centered (i.e. column sums $\boldsymbol{X}^T\mathbf{1}_n = \mathbf{0}$ are zero).

Likewise $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ is the response data vector (also centered with $\boldsymbol{y}^T\mathbf{1}_n = 0$).

Noting that

$$\hat{\boldsymbol{\Sigma}}_{xx} = \frac{1}{n}(\boldsymbol{X}^T\boldsymbol{X})$$

is the MLE of covariance matrix among $\boldsymbol{x}$ and

$$\hat{\boldsymbol{\Sigma}}_{xy} = \frac{1}{n}(\boldsymbol{X}^T\boldsymbol{y})$$

is the MLE of the covariance between $\boldsymbol{x}$ and y we see that the OLS estimate of the regression coefficients can be expressed as

$$\widehat{\boldsymbol{\beta}} = \left(\hat{\boldsymbol{\Sigma}}_{xx}\right)^{-1}\hat{\boldsymbol{\Sigma}}_{xy}$$

We can write down a population version (with no hats!):

$$\boldsymbol{\beta} = \boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}$$

Thus, OLS regression coefficients can be interpreted as plugin estimator using MLEs of covariances! In fact, we may also use the unbiased estimates since the scale factor ($1/n$ or $1/(n-1)$) cancels out so it does not matter which one you use!

16.3.2 Importance of positive definiteness of estimated covariance matrix

Note that $\hat{\boldsymbol{\Sigma}}_{xx}$ is inverted in $\widehat{\boldsymbol{\beta}} = \left(\hat{\boldsymbol{\Sigma}}_{xx}\right)^{-1}\hat{\boldsymbol{\Sigma}}_{xy}$.

- Hence, the estimate $\hat{\boldsymbol{\Sigma}}_{xx}$ needs to be positive definite!
- But $\hat{\boldsymbol{\Sigma}}_{xx}^{\text{MLE}}$ is only positive definite if $n > d$!

Therefore we can use the ML estimate (empirical estimator) only for large $n > d$, otherwise we need to employ a different (regularised) estimation approach (e.g. Bayes or a penalised ML)!

Remark: writing $\hat{\boldsymbol{\beta}}$ explicitly based on covariance estimates has the advantage that we can construct plug-in estimators of regression coefficients based on regularised covariance estimators that improve over ML for small sample size. This leads to the so-called SCOUT method (=covariance-regularized regression by Witten and Tibshirani, 2008).

16.4 Standardised regression coefficients and their relationship to correlation

We recall the relationship between regression coefficients $\boldsymbol{\beta}$ and the marginal covariance $\boldsymbol{\Sigma}_{xy}$ and the covariances among the predictors $\boldsymbol{\Sigma}_{xx}$:

$$\boldsymbol{\beta} = \boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}$$

We can rewrite the regression coefficients in terms of marginal correlations $\boldsymbol{P}_{xy}$ and correlations $\boldsymbol{P}_{xx}$ among the predictors using the variance-correlation decompositions $\boldsymbol{\Sigma}_{xx} = \boldsymbol{V}_x^{1/2}\boldsymbol{P}_{xx}\boldsymbol{V}_x^{1/2}$ and $\boldsymbol{\Sigma}_{xy} = \boldsymbol{V}_x^{1/2}\boldsymbol{P}_{xy}\sigma_y$:

$$\begin{aligned}\boldsymbol{\beta} &= \underbrace{\boldsymbol{V}_x^{-1/2}}_{\text{(inverse) scale of } x_i}\ \boldsymbol{P}_{xx}^{-1}\boldsymbol{P}_{xy}\ \underbrace{\sigma_y}_{\text{scale of } y} \\ &= \boldsymbol{V}_x^{-1/2}\,\boldsymbol{\beta}_{\text{std}}\,\sigma_y\end{aligned}$$

Thus the regression coefficients $\boldsymbol{\beta}$ contain the scale of the variables, and take into account the correlations among the predictors ($\boldsymbol{P}_{xx}$) in addition to the marginal correlations between the response y and the predictors x_i ($\boldsymbol{P}_{xy}$).

This decomposition allows to understand a number special cases for which the regression coefficients simplify further:

a) If the response and the predictors are standardised to have variance one, i.e. $\text{Var}(y) = 1$ and $\text{Var}(x_i) = 1$, then $\boldsymbol{\beta}$ becomes equal to the **standardised regression coefficients**

$$\boldsymbol{\beta}_{\text{std}} = \boldsymbol{P}_{xx}^{-1}\boldsymbol{P}_{xy}$$

Note that standardised regression coefficients do not make use of variances and and thus are scale-independent.

b) If there is no correlation among the predictors , i.e. $\boldsymbol{P}_{xx} = \boldsymbol{I}$ the the regression coefficients reduce to

$$\boldsymbol{\beta} = \boldsymbol{V}_x^{-1}\boldsymbol{\Sigma}_{xy}$$

where $\boldsymbol{V}_x$ is a diagonal matrix containing the variances of the predictors. This is also called **marginal regression**. Note that the inversion of $\boldsymbol{V}_x$ is trival since you only need to invert each diagonal element individually.

c) If both a) and b) apply simultaneously (i.e. there is no correlation among predictors and response and predictors and predictors are standardised) then the regression coefficients simplify even further to

$$\boldsymbol{\beta} = \boldsymbol{P}_{xy}$$

Thus, in this very special case the regression coefficients are identical to the correlations between the response and the predictors!

16.5 Further ways to obtain regression coefficients

16.5.1 Best linear predictor

The **best linear predictor** is a fourth way to arrive at the linear model. This is closely related to OLS and minimising squared residual error.

Without assuming normality the above multiple regression model can be shown to be optimal linear predictor under the minimum mean squared prediction error:

Assumptions:

- y and $\boldsymbol{x}$ are random variables
- we construct a new variable (the linear predictor) $y^{\star\star} = b_0 + \boldsymbol{b}^T\boldsymbol{x}$ to optimally approximate y

Aim:

- choose b_0 and $\boldsymbol{b}$ such to minimize the mean squared prediction error $\mathrm{E}((y - y^{\star\star})^2)$

16.5.1.1 Result:

The mean squared prediction error $MSPE$ in dependence of $(b_0, \boldsymbol{b})$ is

$$\begin{aligned}
\mathrm{E}((y - y^{\star\star})^2) &= \mathrm{Var}(y - y^{\star\star}) + \mathrm{E}(y - y^{\star\star})^2 \\
&= \mathrm{Var}(y - b_0 - \boldsymbol{b}^T\boldsymbol{x}) + (\mathrm{E}(y) - b_0 - \boldsymbol{b}^T\mathrm{E}(\boldsymbol{x}))^2 \\
&= \sigma_y^2 + \mathrm{Var}(\boldsymbol{b}^T\boldsymbol{x}) + 2\,\mathrm{Cov}(y, -\boldsymbol{b}^T\boldsymbol{x}) + (\mu_y - b_0 - \boldsymbol{b}^T\boldsymbol{\mu}_{\boldsymbol{x}})^2 \\
&= \sigma_y^2 + \boldsymbol{b}^T\boldsymbol{\Sigma}_{\boldsymbol{xx}}\boldsymbol{b} - 2\,\boldsymbol{b}^T\boldsymbol{\Sigma}_{\boldsymbol{xy}} + (\mu_y - b_0 - \boldsymbol{b}^T\boldsymbol{\mu}_{\boldsymbol{x}})^2 \\
&= MSPE(b_0, \boldsymbol{b})
\end{aligned}$$

We look for

$$(\beta_0, \boldsymbol{\beta}) = \arg\min_{b_0, \boldsymbol{b}} MSPE(b_0, \boldsymbol{b})$$

In order to find the minimum we compute the gradient with regard to $(b_0, \boldsymbol{b})$

$$\nabla MSPE = \begin{pmatrix} -2(\mu_y - b_0 - \boldsymbol{b}^T\boldsymbol{\mu}_{\boldsymbol{x}}) \\ 2\,\boldsymbol{\Sigma}_{\boldsymbol{xx}}\boldsymbol{b} - 2\,\boldsymbol{\Sigma}_{\boldsymbol{xy}} - 2\boldsymbol{\mu}_{\boldsymbol{x}}(\mu_y - b_0 - \boldsymbol{b}^T\boldsymbol{\mu}_{\boldsymbol{x}}) \end{pmatrix}$$

and setting this equal to zero yields

$$\begin{pmatrix} \beta_0 \\ \boldsymbol{\beta} \end{pmatrix} = \begin{pmatrix} \mu_y - \boldsymbol{\beta}^T\boldsymbol{\mu}_{\boldsymbol{x}} \\ \boldsymbol{\Sigma}_{\boldsymbol{xx}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{xy}} \end{pmatrix}$$

Thus, the optimal values for b_0 and $\boldsymbol{b}$ in the best linear predictor correspond to the previously derived coefficients β_0 and $\boldsymbol{\beta}$!

16.5.1.2 Irreducible Error

The minimum achieved MSPE (=**irreducible error**) is

$$MSPE(\beta_0, \boldsymbol{\beta}) = \sigma_y^2 - \boldsymbol{\beta}^T \boldsymbol{\Sigma}_{xx} \boldsymbol{\beta} = \sigma_y^2 - \boldsymbol{\Sigma}_{yx} \boldsymbol{\Sigma}_{xx}^{-1} \boldsymbol{\Sigma}_{xy}$$

With the abbreviation $\Omega^2 = \boldsymbol{P}_{yx} \boldsymbol{P}_{xx}^{-1} \boldsymbol{P}_{xy} = \sigma_y^{-2} \boldsymbol{\Sigma}_{yx} \boldsymbol{\Sigma}_{xx}^{-1} \boldsymbol{\Sigma}_{xy}$ we can simplify this to

$$MSPE(\beta_0, \boldsymbol{\beta}) = \sigma_y^2 (1 - \Omega^2) = \mathrm{Var}(\varepsilon)$$

Writing $b_0 = \beta_0 + \Delta_0$ and $\boldsymbol{b} = \boldsymbol{\beta} + \boldsymbol{\Delta}$ it is easy to see that the mean squared predictive error is a quadratic function around the minimum:

$$MSPE(\beta_0 + \Delta_0, \boldsymbol{\beta} + \boldsymbol{\Delta}) = \mathrm{Var}(\varepsilon) + \Delta_0^2 + \boldsymbol{\Delta}^T \boldsymbol{\Sigma}_{xx} \boldsymbol{\Delta}$$

Note that usually $y^\star = \beta_0 + \boldsymbol{\beta}^T \boldsymbol{x}$ does not perfectly approximate y so there *will* be an irreducible error (= noise variance)

$$\mathrm{Var}(\varepsilon) = \sigma_y^2 (1 - \Omega^2) > 0$$

which implies $\Omega^2 < 1$.

The quantity Ω^2 has a further interpretation of the population version of as the squared multiple correlation coefficient between the response and the predictors and plays a vital role in decomposition of variance, as discussed later.

16.5.2 Regression by conditioning

Conditioning is a fifth way to arrive at the linear model. This is also the most general way and can be used to derive many other regression models (not just the simple linear model).

16.5.2.1 General idea:

- two random variables y (response, scalar) and $\boldsymbol{x}$ (predictor variables, vector)
- we assume that y and $\boldsymbol{x}$ have a joint distribution $F_{y,\boldsymbol{x}}$
- compute *conditional* random variable $y|\boldsymbol{x}$ and the corresponding distribution $F_{y|\boldsymbol{x}}$

16.5.2.2 Multivariate normal assumption

Now we assume that y and $\boldsymbol{x}$ are (jointly) multivariate normal. Then the conditional distribution $F_{y|\boldsymbol{x}}$ is a univariate normal with the following moments (you can verify this by looking up the general conditional multivariate normal distribution):

a) Conditional expectation:

$$\mathrm{E}(y|\boldsymbol{x}) = y^\star = \beta_0 + \boldsymbol{\beta}^T\boldsymbol{x}$$

with coefficients $\boldsymbol{\beta} = \boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}$ and intercept $\beta_0 = \mu_y - \boldsymbol{\beta}^T\boldsymbol{\mu}_{\boldsymbol{x}}$.
Note that as $y^\star$ depends on $\boldsymbol{x}$ it is a random variable itself with mean

$$\mathrm{E}(y^\star) = \beta_0 + \boldsymbol{\beta}^T\boldsymbol{\mu}_{\boldsymbol{x}} = \mu_y$$

and variance

$$\begin{aligned}\mathrm{Var}(y^\star) &= \mathrm{Var}(\mathrm{E}(y|\boldsymbol{x})) \\ &= \boldsymbol{\beta}^T\boldsymbol{\Sigma}_{xx}\boldsymbol{\beta} = \boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy} \\ &= \sigma_y^2\boldsymbol{P}_{yx}\boldsymbol{P}_{xx}^{-1}\boldsymbol{P}_{xy} \\ &= \sigma_y^2\Omega^2\end{aligned}$$

b) Conditional variance:

$$\begin{aligned}\mathrm{Var}(y|\boldsymbol{x}) &= \sigma_y^2 - \boldsymbol{\beta}^T\boldsymbol{\Sigma}_{xx}\boldsymbol{\beta} \\ &= \sigma_y^2 - \boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy} \\ &= \sigma_y^2(1-\Omega^2)\end{aligned}$$

Note this is a constant so $\mathrm{E}(\mathrm{Var}(y|\boldsymbol{x})) = \sigma_y^2(1-\Omega^2)$ as well.

Chapter 17

Squared multiple correlation and variance decomposition in linear regression

In this chapter we first introduce the (squared) multiple correlation and the multiple and adjusted R^2 coefficients as estimators. Subsequently we discuss variance decomposition.

17.1 Squared multiple correlation Ω^2 and the R^2 coefficient

In the previous chapter we encountered the following quantity:

$$\Omega^2 = \boldsymbol{P}_{yx}\boldsymbol{P}_{xx}^{-1}\boldsymbol{P}_{xy} = \sigma_y^{-2}\boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}$$

With $\boldsymbol{\beta} = \boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}$ and $\beta_0 = \mu_y - \boldsymbol{\beta}^T\boldsymbol{\mu}_x$ it is straightforward to verify the following:

- the cross-covariance between y and $y^\star$ is

$$\begin{aligned}\mathrm{Cov}(y, y^\star) &= \boldsymbol{\Sigma}_{yx}\boldsymbol{\beta} = \boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy} \\ &= \sigma_y^2\boldsymbol{P}_{yx}\boldsymbol{P}_{xx}^{-1}\boldsymbol{P}_{xy} = \sigma_y^2\Omega^2\end{aligned}$$

- the (signal) variance of $y^\star$ is

$$\begin{aligned}\mathrm{Var}(y^\star) &= \boldsymbol{\beta}^T\boldsymbol{\Sigma}_{xx}\boldsymbol{\beta} = \boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy} \\ &= \sigma_y^2\boldsymbol{P}_{yx}\boldsymbol{P}_{xx}^{-1}\boldsymbol{P}_{xy} = \sigma_y^2\Omega^2\end{aligned}$$

hence the correlation $\text{Cor}(y, y^\star) = \frac{\text{Cov}(y,y^\star)}{\text{SD}(y)\text{SD}(y^\star)} = \Omega$ with $\Omega \geq 0$.

This helps to understand the Ω and Ω^2 coefficients:

- Ω is the linear correlation between the response (y) and prediction $y^\star$.
- Ω^2 is called the **squared multiple correlation** between the scalar y and the vector $\boldsymbol{x}$.
- Note that if we only have one predictor (if x is a scalar) then $\boldsymbol{P}_{xx} = 1$ and $\boldsymbol{P}_{yx} = \rho_{yx}$ so the multiple squared correlation coefficient reduces to squared correlation $\Omega^2 = \rho^2_{yx}$ between two scalar random variables y and x.

17.1.1 Estimation of Ω^2 and the multiple R^2 coefficient

The multiple squared correlation coefficient Ω^2 can be estimated by plug-in of empirical estimates for the corresponding correlation matrices:

$$R^2 = \hat{\boldsymbol{P}}_{yx}\hat{\boldsymbol{P}}^{-1}_{xx}\hat{\boldsymbol{P}}_{xy} = \hat{\sigma}_y^{-2}\hat{\boldsymbol{\Sigma}}_{yx}\hat{\boldsymbol{\Sigma}}^{-1}_{xx}\hat{\boldsymbol{\Sigma}}_{xy}$$

This estimator of Ω^2 is called the **multiple R^2 coefficient**.

If the same scale factor $1/n$ or $1/(n-1)$ is used in estimating the variance σ^2_y and the covariances $\boldsymbol{\Sigma}_{xx}$ and $\boldsymbol{\Sigma}_{yx}$ then this factor will cancel out.

Above we have seen that Ω^2 is directly linked with the noise variance via

$$\text{Var}(\varepsilon) = \sigma^2_y(1 - \Omega^2)\,.$$

so we can express the squared multiple correlation as

$$\Omega^2 = 1 - \text{Var}(\varepsilon)/\sigma^2_y$$

The **maximum likelihood estimate** of the noise variance $\text{Var}(\varepsilon)$ (also called **residual variance**) can be computed from the residual sum of squares $RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2$ as follows:

$$\widehat{\text{Var}}(\varepsilon)_{ML} = \frac{RSS}{n}$$

whereas the **unbiased estimate** is obtained by

$$\widehat{\text{Var}}(\varepsilon)_{UB} = \frac{RSS}{n-d-1} = \frac{RSS}{df}$$

where the **degree of freedom** is $df = n - d - 1$ and d is the number of predictors.

Similarly, we can find the maximum likelihood estimate v_y^{ML} for σ^2_y (with factor $1/n$) as well as an unbiased estimate v_y^{UB} (with scale factor $1/(n-1)$)

The **multiple R^2 coefficient** can then be written as

$$R^2 = 1 - \widehat{\text{Var}}(\varepsilon)_{ML}/v_y^{ML}$$

Note we use MLEs.

In contrast, the so-called **adjusted multiple R^2 coefficient** is given by

$$R^2_{\text{adj}} = 1 - \widehat{\text{Var}}(\varepsilon)_{UB}/v_y^{UB}$$

where the unbiased variances are used.

Both R^2 and R^2_{adj} are estimates of Ω^2 and are related by

$$1 - R^2 = (1 - R^2_{\text{adj}})\,\frac{df}{n-1}$$

17.1.2 R commands

In R the command `lm()` fits the linear regression model.

In addition to the regression cofficients (and derived quantities) the R function `lm()` also lists

- the multiple R-squared R^2,
- the adjusted R-squared R^2_{adj},
- the degrees of freedom df and
- the residual standard error $\sqrt{\widehat{\text{Var}}(\varepsilon)_{UB}}$ (computed from the unbiased variance estimate).

See also Worksheet R3 which provides R code to reproduce the exact output of the native `lm()` R function.

17.2 Variance decomposition in regression

The squared multiple correlation coefficient is useful also because it plays an important role in the decomposition of the total variance:

- total variance: $\text{Var}(y) = \sigma_y^2$
- unexplained variance (irreducible error): $\sigma_y^2(1-\Omega^2) = \text{Var}(\varepsilon)$
- the explained variance is the complement: $\sigma_y^2\Omega^2 = \text{Var}(y^\star)$

In summary:

$$\text{Var}(y) = \text{Var}(y^\star) + \text{Var}(\varepsilon)$$

becomes

$$\underbrace{\sigma_y^2}_{\text{total variance}} = \underbrace{\sigma_y^2\Omega^2}_{\text{explained variance}} + \underbrace{\sigma_y^2(1-\Omega^2)}_{\text{unexplained variance}}$$

The unexplained variance measures the fit after introducing predictors into the model (smaller means better fit). The total variance measures the fit of the model without any predictors. The explained variance is the difference between total and unexplained variance, it indicates the increase in model fit due to the predictors.

17.2.1 Law of total variance and variance decomposition

The **law of total variance** is

$$\underbrace{\mathrm{Var}(y)}_{\text{total variance}} = \underbrace{\mathrm{Var}(\mathrm{E}(y|\boldsymbol{x}))}_{\text{explained variance}} + \underbrace{\mathrm{E}(\mathrm{Var}(y|\boldsymbol{x}))}_{\text{unexplained variance}}$$

provides a very general decomposition in explained and unexplained parts of the variance that is valid regardless of the form of the distributions $F_{y,\boldsymbol{x}}$ and $F_{y|\boldsymbol{x}}$.

In regression it conncects variance decomposition and conditioning. If you plug-in the conditional expections for the multivariate normal model (cf. previous chapter) we recover

$$\underbrace{\sigma_y^2}_{\text{total variance}} = \underbrace{\sigma_y^2\Omega^2}_{\text{explained variance}} + \underbrace{\sigma_y^2(1-\Omega^2)}_{\text{unexplained variance}}$$

17.2.2 Related quantities

Using the above three quantities (total variance, explained variance, and unexplained variance) we can construct a number of scores:

1) **coefficient of determination**, **squared multiple correlation**:

$$\frac{\text{explained var}}{\text{total var}} = \frac{\sigma_y^2\Omega^2}{\sigma_y^2} = \Omega^2$$

(range 0 to 1, with 1 indicating perfect fit)

2) **coefficient of non-determination**, **coefficient of alienation**:

$$\frac{\text{unexplained var}}{\text{total var}} = \frac{\sigma_y^2(1-\Omega^2)}{\sigma_y^2} = 1-\Omega^2$$

(range 0 to 1, with 0 indicating perfect fit)

3) F **score**, t^2 **score**:

$$\frac{\text{explained var}}{\text{unexplained var}} = \frac{\sigma_y^2\Omega^2}{\sigma_y^2(1-\Omega^2)} = \frac{\Omega^2}{1-\Omega^2} = \mathcal{F} = \frac{\tau^2}{n}$$

(range 0 to ∞, with ∞ indicating perfect fit)

Note that the $\mathcal{F}$ and τ^2 scores are population versions of the F and t^2 statistics.

Also note that $\Omega^2 = \frac{\tau^2}{\tau^2+n} = \frac{\mathcal{F}}{\mathcal{F}+1}$ links squared correlation with squared t-scores and F-scores.

17.3 Sample version of variance decomposition

If Ω^2 and σ_y^2 are replaced by their MLEs this can be written in a sample version as follows using data points y_i, predictions $\hat{y}_i$ and $\bar{y} = \frac{1}{n}\sum_{i=1}^n y_i$

$$\underbrace{\sum_{i=1}^n (y_i - \bar{y})^2}_{\text{total sum of squares (TSS)}} = \underbrace{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}_{\text{explained sum of squares (ESS)}} + \underbrace{\sum_{i=1}^n (y_i - \hat{y}_i)^2}_{\text{residual sum of squares (RSS)}}$$

Note that TSS, ESS and RSS all scale with n. Using data vector notation the sample-based variance decomposition can be written in form of the Pythagorean theorem:

$$\underbrace{||\boldsymbol{y} - \bar{y}\mathbf{1}\,||^2}_{\text{total sum of squares (TSS)}} = \underbrace{||\hat{\boldsymbol{y}} - \bar{y}\mathbf{1}||^2}_{\text{explained sum of squares (ESS)}} + \underbrace{||\boldsymbol{y} - \hat{\boldsymbol{y}}||^2}_{\text{residual sum of squares (RSS)}}$$

17.3.1 Geometric interpretation of regression as orthogonal projection:

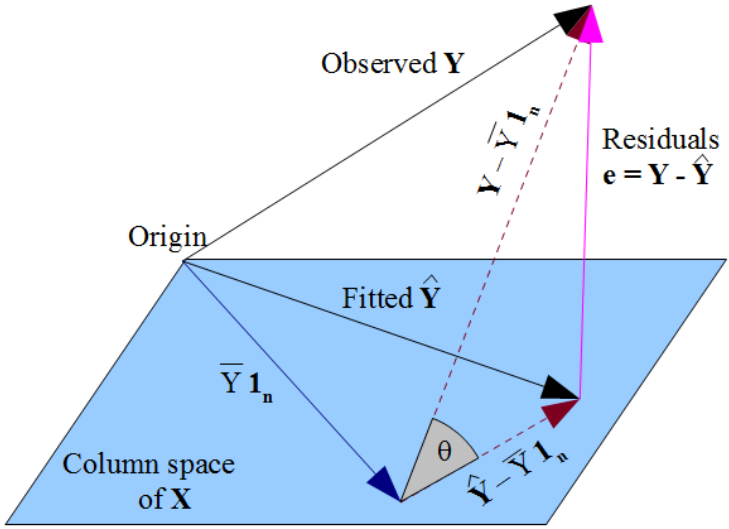
The above equation can be further simplified to

$$||\boldsymbol{y}||^2 = ||\hat{\boldsymbol{y}}||^2 + \underbrace{||\boldsymbol{y} - \hat{\boldsymbol{y}}||^2}_{\text{RSS}}$$

Geometrically speaking, this implies $\hat{\boldsymbol{y}}$ is an orthogonal projection of $\boldsymbol{y}$, since the residuals $\boldsymbol{y} - \hat{\boldsymbol{y}}$ and the predictions $\hat{\boldsymbol{y}}$ are orthogonal (by construction!).

This also valid for the centered versions of the vectors, i.e. $\hat{\boldsymbol{y}} - \bar{y}\mathbf{1}_n$ is an orthogonal projection of $\boldsymbol{y} - \bar{y}\mathbf{1}_n$ (see Figure).

Also note that the angle θ between the two centered vectors is directly related to the (estimated) multiple correlation, with $R = \cos(\theta) = \frac{||\hat{\boldsymbol{y}} - \bar{y}\mathbf{1}_n||}{||\boldsymbol{y} - \bar{y}\mathbf{1}_n||}$, or $R^2 = \cos(\theta)^2 = \frac{||\hat{\boldsymbol{y}} - \bar{y}\mathbf{1}_n||^2}{||\boldsymbol{y} - \bar{y}\mathbf{1}_n||^2} = \frac{\text{ESS}}{\text{TSS}}$.



Source of Figure: Stack Exchange

Chapter 18

Prediction and variable selection

In this chapter we discuss how to compute (lower bounds) of the prediction error and how to select variables relevant for prediction

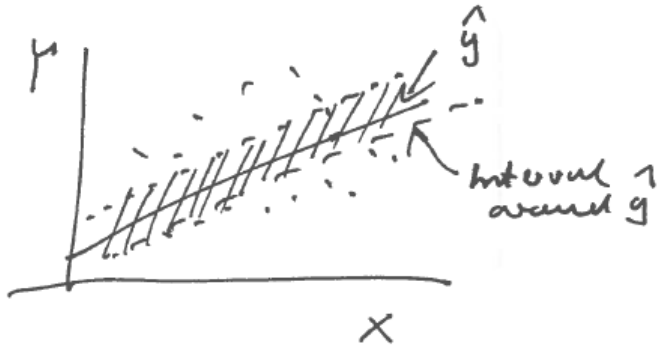
18.1 Prediction and prediction intervals

Learning the regression function from (training) data is only the first step in application of regression models.

The next step is to actually make **prediction** of future outcomes y^{test} given test data $\boldsymbol{x}^{\text{test}}$:

$$y^{\text{test}} = \hat{y}(\boldsymbol{x}^{\text{test}}) = \hat{f}_{\hat{\beta}_0,\hat{\boldsymbol{\beta}}}(\boldsymbol{x}^{\text{test}})$$

Note that y^{test} is a point estimator. Is it possible also to construct a corresponding interval estimate?



The answer is yes, and leads back to the conditioning approach:

$$y^\star = \mathrm{E}(y|\boldsymbol{x}) = \beta_0 + \boldsymbol{\beta}^T\boldsymbol{x}$$

$$\mathrm{Var}(\varepsilon) = \mathrm{Var}(y|\boldsymbol{x}) = \sigma_y^2(1-\Omega^2)$$

We know that the mean squared prediction error for $y^\star$ is $\mathrm{E}((y-y^\star)^2) = \mathrm{Var}(\varepsilon)$ and that this is the minimal irreducible error. Hence, we may use $\mathrm{Var}(\varepsilon)$ as the *minimum* variability for the prediction.

The corresponding prediction interval is

$$\left[y^\star(\boldsymbol{x}^{\text{test}}) \pm c \times \mathrm{SD}(\varepsilon)\right]$$

where c is some suitable constant (e.g. 1.96 for symmetric 95% normal intervals).

However, please note that the prediction interval constructed in this fashion will be an *underestimate*. The reason is that this assumes that we employ $y^\star = \beta_0 + \boldsymbol{\beta}^T\boldsymbol{x}$ but in reality we actually use $\hat{y} = \hat{\beta}_0 + \hat{\boldsymbol{\beta}}^T\boldsymbol{x}$ for prediction — note the estimated coefficients! We recall from an earlier chapter (best linear predictor) that this leads to increase of MSPE compared with using the optimal β_0 and $\boldsymbol{\beta}$.

Thus, for better prediction intervals we would need to consider the mean squared prediction error of $\hat{y}$ that can be written as $\mathrm{E}((y-\hat{y})^2) = \mathrm{Var}(\varepsilon) + \delta$ where δ is an **additional error term due to using an estimated rather than the true regression function**. δ typically declines with $1/n$ but can be substantial for small n (in particular as it usually depends on the number of predictors d).

For more details on this we refer to later modules on regression.

18.2 Variable importance and prediction

Another key question in regression modelling is to find out predictor variables $x_1, x_2, \ldots, x_d$ are actually important for predicting the outcome y.

→ We need to study variable importance measures (VIM).

18.2.1 How to quantify variable importance?

A variable x_i is **important** if it **improves prediction** of the response y.

Recall variance decomposition:

$$\mathrm{Var}(y) = \sigma_y^2 = \underbrace{\sigma_y^2\Omega^2}_{\text{explained variance}} + \underbrace{\sigma_y^2(1-\Omega^2)}_{\text{unexplained/residual variance} = \mathrm{Var}(\varepsilon)}$$

- Ω^2 squared multiple correlation $\in [0, 1]$
- Ω^2 large $\rightarrow 1$ predictor variables explain most of σ_y^2
- Ω^2 small $\rightarrow 0$ linear model fails and predictors do not explain variability
- $\Rightarrow$ If a predictor helps to $\begin{matrix}\text{increase explained variance}\\ \text{decrease unexplained variance}\end{matrix}$ then it is important!
- $\Omega^2 = \boldsymbol{P}_{yx}\boldsymbol{P}_{xx}^{-1}\boldsymbol{P}_{xy} \hat{=}$ a function of the X!

VIM: which predictors contribute most to Ω^2

18.2.2 Some candidates for VIMs

1. The regression coefficients $\boldsymbol{\beta}$
   - $\boldsymbol{\beta} = \boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy} = \boldsymbol{V}_x^{-1/2}\boldsymbol{P}_{xx}^{-1}\boldsymbol{P}_{xy}\sigma_y$
   - Not a good VIM since $\boldsymbol{\beta}$ contains the scale!
   - Large $\hat{\beta}_i$ does not indicate that x_i is important.
   - Small $\hat{\beta}_i$ does not indicate that x_i is not important.
2. Standardised regression coefficients $\boldsymbol{\beta}_{\text{std}}$
   - $\boldsymbol{\beta}_{\text{std}} = \boldsymbol{P}_{xx}^{-1}\boldsymbol{P}_{xy}$
   - implies $\text{Var}(y) = 1$, $\text{Var}(x_i) = 1$
   - These do not contain the scale (so better than $\hat{\beta}$)
   - But still unclear how this relates to decomposition of variance
3. Squared marginal correlations ρ_{y,x_i}^2

   Consider case of uncorrelated predictors: $\boldsymbol{P}_{xx} = \boldsymbol{I}$ (no correlation among x_i)

   $$\Rightarrow \Omega^2 = \boldsymbol{P}_{yx}\boldsymbol{P}_{xy} = \sum_{i=1}^{d} \rho_{y,x_i}^2$$

   $\rho_{y,x_i}^2 = \text{Cor}(y, x_i)$ is the marginal correlation between y and x_i, and Ω^2 is (for uncorrelated predictors) the sum of squared marginal correlations.
   - If $\boldsymbol{P}_{xx} = \boldsymbol{I}$, then *ranking* predictors by ρ_{y,x_i}^2 is optimal!
   - The predictor with largest marginal correlation reduces the unexplained variance most!
   - good news: even if there is weak correlation among predictors the marginal correlations are still good as VIM (but then they will not perfectly add up to Ω^2)
   - Advantage: very simple but often also very effective.
   - Caution! If there is strong correlation in $\boldsymbol{P}_{xx}$, then there is **colinearity** (in this case it is oftern best to remove one of the strongly correlated variables, or to merge the correlated variables).

Often, ranking predictors by their squared marginal correlations is done as a prefiltering step (independence screening).

18.3 Regression t-scores.

18.3.1 Wald statistic for regression coefficients

So far, we discussed three obvious candidates for for variable importance measures (regression coefficients, standardised regression coefficients, marginal correlations).

In this section we consider a further quantity, the **regression $t-$score**:

Recall that ML estimation of the regression coefficients yields

- a point estimate $\hat{\boldsymbol{\beta}}$
- the (asymptotic) variance $\widehat{\text{Var}}(\hat{\boldsymbol{\beta}})$
- the asymptotic normal distribution $\hat{\boldsymbol{\beta}} \overset{a}{\sim} N_d(\boldsymbol{\beta}, \widehat{\text{Var}}(\hat{\boldsymbol{\beta}}))$

Corresponding to each predictor x_i we can construct from the above a t-score

$$t_i = \frac{\hat{\beta}_i}{\widehat{\text{SD}}(\hat{\beta}_i)}$$

where the standard deviations are computed by $\widehat{\text{SD}}(\hat{\beta}_i) = \text{Diag}(\widehat{\text{Var}}(\hat{\boldsymbol{\beta}}))_i$. This corresponds to the **Wald statistic** to test that the underlying true regression coefficient is zero ($\beta_i = 0$).

Correspondingly, under the null hypthesis that $\beta_i = 0$ asymptotically for large n the regression t-score is standard normal distributed:

$$t_i \overset{a}{\sim} N(0, 1)$$

This allows to compute (symmetric) p-values $p = 2\Phi(-|t_i|)$ where Φ is the standard normal distribution function.

For finite n, assuming normality of the observation and using the unbiased estimate for variance when computing t_i, the exact distribution of i_i is given by the Student-t distribution:

$$t_i \sim t_{n-d-1}$$

Regression t-scores can thus be used to test whether a regression coefficient is zero. A large magnitude of the t_i score indicates that the hypothesis $\beta_i = 0$ can be rejected. Thus, a small p-value (say smaller than 0.05) signals that the regression coefficient is non-zero and hence that the corresponding predictor variable should be included in the model.

This allows rank predictor variables by $|t_i|$ or the corresponding p-values with regard to their relevance in the linear model. Typically, in order to simplify a

model, predictors with the largest p-values (and thus smallest absolute t-scores) may be removed from a model. However, note that having a p-value say larger than 0.05 by itself is not sufficient to declare a regression coefficient to be zero (because in classical statistical testing you can only reject the null hypothesis, but not accept it!).

Note that by construction the regression t-scores do not depend on the scale, so when the original data are rescaled this will not affect the corresponding regression t-scores. Furthermore, if $\widehat{\mathrm{SD}}(\hat{\beta}_i)$ is small, then the regression t-score t_i can still be large even if $\hat{\beta}_i$ is small!

18.3.2 Computing

When you perform regression analysis in R (or another statistical software package) the computer will return the following:

$\hat{\beta}_i$	$\widehat{\mathrm{SD}}(\hat{\beta}_i)$	$t_i = \frac{\hat{\beta}_i}{\widehat{\mathrm{SD}}(\hat{\beta}_i)}$	p-values	Indicator of
Estimated repression coefficient	Error of $\hat{\beta}_i$	t-score computed from first two columns	for t_i based on t-distribution	Significance * 0.9 ** 0.95 *** 0.99

In the `lm()` function in R the standard deviation is the square root of the unbiased estimate of the variance (but note that it itself is not unbiased!).

18.3.3 Connection with partial correlation

The deeper reason why ranking predictors by regression t-scores and associated p-values is useful is their link with **partial correlation**.

In particular, the (squared) regression t-score can be 1:1 transformed into the (estimated) (squared) partial correlation

$$\hat{\rho}^2_{y,x_i|x_{j\neq i}} = \frac{t_i^2}{t_i^2 + df}$$

with $df = n - d - 1$, and it can be shown that the p-values for testing that $\beta_i = 0$ are exactly the same as the p-values for testing that the partial correlation $\rho_{y,x_i|x_{j\neq i}}$ vanishes!

Therefore, ranking the predictors x_i by regression t-scores leads to exactly the same ranking and p-values as partial correlation!

18.3.4 Squared Wald statistic and the F statistic

In the above we looked at individual regression coefficients. However, we can also construct a Wald test using the complete vector $\hat{\boldsymbol{\beta}}$. The squared Wald statistic to test that $\boldsymbol{\beta} = 0$ is given by

$$\begin{aligned} t^2 &= \hat{\boldsymbol{\beta}}^T \widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}^{-1})\hat{\boldsymbol{\beta}} \\ &= \left(\hat{\boldsymbol{\Sigma}}_{yx}\hat{\boldsymbol{\Sigma}}_{xx}^{-1}\right)\left(\frac{n}{\widehat{\sigma_\varepsilon^2}}\hat{\boldsymbol{\Sigma}}_{xx}\right)\left(\hat{\boldsymbol{\Sigma}}_{xx}^{-1}\hat{\boldsymbol{\Sigma}}_{xy}\right) \\ &= \frac{n}{\widehat{\sigma_\varepsilon^2}}\hat{\boldsymbol{\Sigma}}_{yx}\hat{\boldsymbol{\Sigma}}_{xx}^{-1}\hat{\boldsymbol{\Sigma}}_{xy} \\ &= \frac{n}{\widehat{\sigma_\varepsilon^2}}\hat{\sigma}_y^2 R^2 \end{aligned}$$

With $\widehat{\sigma_\varepsilon^2}/\hat{\sigma}_y^2 = 1 - R^2$ we finally get the related F statistic

$$\frac{t^2}{n} = \frac{R^2}{1 - R^2} = F$$

which is a function of R^2. If $R^2 = 0$ then $F = 0$. If R^2 is large (< 1) then F is large as well ($< \infty$) and the null hypothesis $\boldsymbol{\beta} = 0$ can be rejected, which implies that at least one regression coefficient is non-zero. Note that the squared Wald statistic t^2 is asymptotically χ_d^2 distributed which is useful to find critical values and to compute p-values.

18.4 Further approaches for variable selection

In addition to ranking by marginal and partial correlation, there are many other approaches for variable selection in regression!

a) Search-based methods:

- search through subsets of linear models for d variables, ranging from full model (including all predictors) to the empty model (includes no predictor) and everything inbetween.
- Problem: exhaustive search not possible even for relatively small d as space of models is very large!
- Therefore heuristic approaches such as forward selection (adding predictors), backward selection (removing predictors), or monte-carlo random search are employed.
- Problem: maximum likelihood cannot be used for choosing among the models - since ML will always pick the best model. Therefore, penalised ML criteria such as AIC or Bayesian criteria are often employed instead.

b) Integrative estimation and variable selection:

- there are methods that fit the regression model and perform variable selection *simultaneously*.
- the most well-known approach of this type is "lasso" regression (Tibshirani 1996)
- This applies a (generalised) linear model with ML plus L1 penalty.
- Alternative: Bayesian variable selection and estimation procedures

c) Entropy-based variable selection:

As seen above, two of the most popular approaches in linear models are based on correlation, either marginal correlation or partial correlation (via regression t-scores).

Correlation measures can be generalised to non-linear settings. One very popular measure is the **mutual information** which is computed using the KL divergence. In case of two variables x and y with joint normal distribution and correlation ρ the mutual information is a function of the correlation:

$$\mathrm{MI}(x, y) = \frac{1}{2} \log(1 - \rho^2)$$

In regression he mutual information between the response y and predictor x_i is $\mathrm{MI}(y, x_i)$, and this widely used for feature selection, in particular in machine learning.

d) FDR based variable selection in regression:

Feature selection controling the false discovery rate (FDR) among the selected features are becoming more popular, in particular a number of procedures using so-called "knockoffs", see https://web.stanford.edu/group/candes/knockoffs/ .

e) Variable importance using Shapley values:

Borrowing a concept from game theory Shapley values have recently become popular in machine learning to evaluate the variable importance of predictors in nonlinear models. Their relationship to other statistical methods for measuring variable importance is the focus of current research.

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 "Statistical Methods". 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:

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

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 Convex and concave functions

A function $h(x)$ is convex if the second derivative $h''(x) \geq 0$ for all x. More generally, a function $h(\boldsymbol{x})$, where $\boldsymbol{x}$ is a vector, is convex if the Hessian matrix $\nabla\nabla^T h(\boldsymbol{x})$ is positive definite, i.e. if the eigenvalues of the Hessian matrix are all positive.

If $h(\boldsymbol{x})$ is convex, then $-h(\boldsymbol{x})$ is *concave*. A function is concave if the Hessian matrix is negative definite, i.e. if the eigenvalues of the Hessian matrix are all negative.

Example A.3. The logarithm $\log(x)$ is an example of a concave function whereas x^2 is a convex function.

To memorise, a **v**alley is con**v**ex.

A.3.4 Linear and quadratic approximation

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

Applied to 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 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.4. 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

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

A.3.5 Conditions for local optimum of a function

To check if x_0 or $\boldsymbol{x}_0$ is a local maximum or minimum we can use the following conditions:

For a function of a single variable:

i) First derivative is zero at optimum $h'(x_0) = 0$.
ii) If the second derivative $h''(x_0) < 0$ at the optimum is negative the function is locally concave and the optimum is a maximum.
iii) If the second derivative $h''(x_0) > 0$ is positive at the optimum the function is locally convex and the optimum is a minimum.

For a function of several variables:

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

Around the local optimum $\boldsymbol{x}_0$ we can approximate the function quadratically 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 the linear term is missing due to the gradient being zero at $\boldsymbol{x}_0$.

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

$$n! = \Gamma(n + 1)$$

which can be viewed as continuous version of the factorial.

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 possible outcomes is the **sample space** or **state space** 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 subset of Ω and thus itself a set 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 $A^C = \Omega \setminus A = \{\omega_i \in \Omega : \omega_i \notin A\}$.

The probability of an event is denoted by $\Pr(A)$. We assume 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 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 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 probability theory is closely linked to set theory, and in particular to measure theory. This allows for an unified treatment of discrete and continuous random variables (an elegant framework but not needed for this module).

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

To describe a random variable x we need to assign probabilities to the corresponding elementary outcomes $x \in \Omega$. For convenience we use the same name to denote the random variable and the elementary outcomes.

For a discrete random variable we employ a probability mass function (PMF). We denote the it by a lower case f but occasionally we also use p or q. In the discrete case we can define the event $A = \{x : x = a\} = \{a\}$ and obtain the probability directly from the PMF:

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

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.

Assuming an ordering we can define the event $A = \{x : x \leq a\}$ and compute its probability

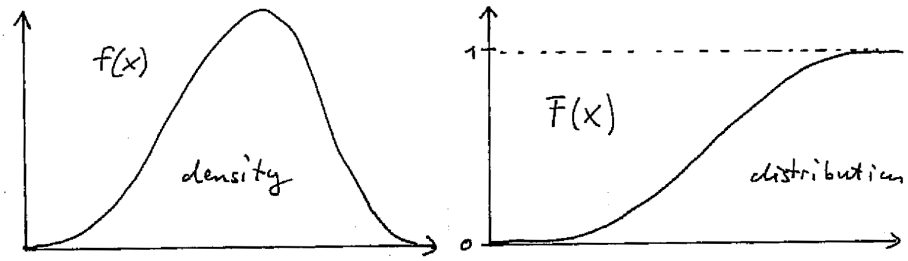
$$F(a) = \Pr(A) = \Pr(x \leq a) = \begin{cases} \sum_{x\in A} f(x) & \text{discrete case} \\ \int_{x\in A} f(x)dx & \text{continuous case} \end{cases}$$

This is known as the **distribution function**, or **cumulative distribution function** (CDF) and is denoted by upper case F if the corresponding PDF/PMF is f (or P and Q if the corresponding PDF/PMF are p and q). By construction the distribution function is monotonically increasing and its value ranges from 0 to 1. With its help we can compute the probability of general interval sets 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 Expection and variance 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 $f(x)$:

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

To emphasise that the expecation is taken with regard to the distribution F we write $\mathrm{E}_F(x)$ with the distribution F as subscript. The expectation is not necessarily always defined for a continuous random variable as the integral may diverge.

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

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

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 $\mathrm{E}_F(h(x))$.

For an event A we can define a corresponding **indicator function**

$$1_A(x) = \begin{cases} 1 & x \in A \\ 0 & x \notin A \end{cases}$$

Intriguingly,

$$\mathrm{E}(1_A(x)) = \Pr(A)$$

i.e. the expectation of the indicator variable for A is the probability of A.

The moments of random variables are also defined by expectation:

- Zeroth moment: $\mathrm{E}(x^0) = 1$ by definition 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 momented centered about the mean μ:

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

- The variance can also be computed by $\mathrm{Var}(x) = \mathrm{E}(x^2) - \mathrm{E}(x)^2$.

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.5.4 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 $dy = |\frac{dy}{dx}|dx$.

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.5.5 Law of large numbers:

Suppose we observe data $D = \{x_1, \ldots, x_n\}$ with each $x_i \sim F$.

- By the strong law of large numbers the empirical distribution $\hat{F}_n$ based on data $D = \{x_1, \ldots, x_n\}$ converges to the true underlying distribution F as $n \to \infty$ almost surely:

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

  The Glivenko–Cantelli theorem asserts that the convergence is uniform. Since the strong law implies the weak law we also have convergence in probability:

  $$\hat{F}_n \xrightarrow{P} F$$

- 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 $\mathrm{E}_F(h(x))$.

A.5.6 Jensen's inequality

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

for a *convex* function $h(\boldsymbol{x})$.

Recall: a con**v**ex function (such as x^2) has the shape of a "**v**alley".

A.6 Distributions

A.6.1 Bernoulli distribution and binomial distribution

The Bernoulli distribution $\text{Ber}(p)$ is simplest distribution possible. It is named after Jacob Bernoulli (1655-1705) who also invented 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 $p \in [0, 1]$ 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 $\text{Ber}(p)$ is

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

and

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

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

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

If a random variable x follows the Bernoulli distribution we write

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

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

Closely related to the Bernoulli distribution is the binomial distribution $\text{Bin}(m, p)$ which results from repeating a Bernoulli experiment m times and counting the number of successes among the m trials (without keeping track of the ordering of the experiments).

Its probability mass function is:

$$f(x|p) = \binom{m}{x} p^x(1 - p)^{m-x}$$

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

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

If a random variable x follows the binomial distribution we write

$$x \sim \text{Bin}(m, p)$$

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

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

A.6.2 Normal distribution

Univariate normal distribution:

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

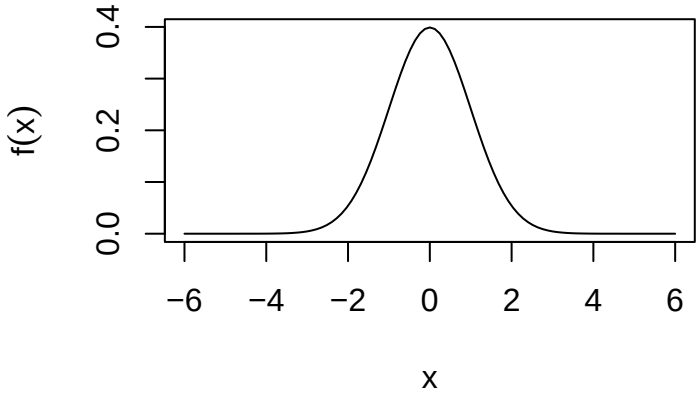
Probability density function (PDF):

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

In R the density function is called `dnorm()`.

The standard normal distribution is $N(0, 1)$ with mean 1 and variance 1.

Plot of the PDF of the standard normal:

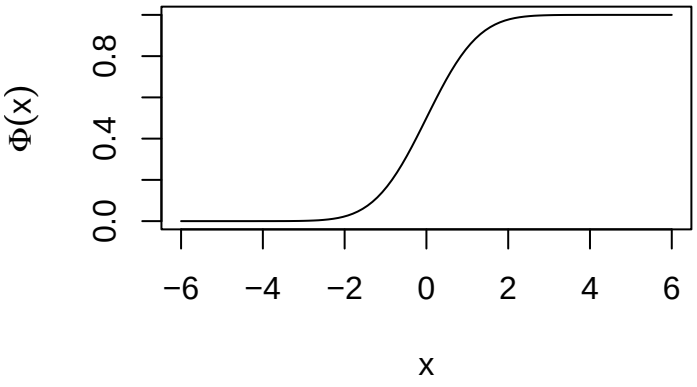


The cumulative distribution function (CDF) of the standard normal $N(0, 1)$ is

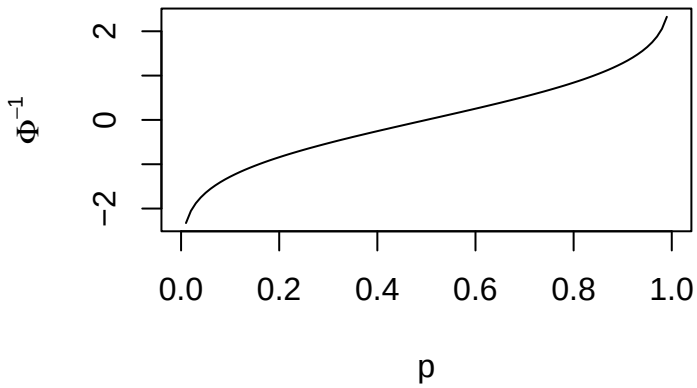
$$\Phi(x) = \int_{-\infty}^{x} f(x'|\mu = 0, \sigma^2 = 1)dx'$$

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).

A.6.3 Gamma distribution aka scaled chi-squared and one-dimensional Wishart distribution

Assume m independent normal random variables

$$z_1, z_2, \ldots, z_m \sim N(0, \sigma_z^2)$$

Then the sum of the squares

$$x = \sum_{i=1}^{m} z_i^2$$

with

$$\mu_x = m\sigma_z^2$$

follows a **scaled chi-squared distribution**

$$x \sim \frac{\mu_x}{m}\chi_m^2 = W_1\left(\frac{\mu_x}{m}, m\right) = \text{Gam}\left(\frac{1}{2}m, 2\frac{\mu_x}{m}\right)$$

with degree of freedom m and $x \geq 0$. The mean and variance of a scaled chi-squared distributed variable is $\text{E}(x) = \mu_x$ and $\text{Var}(x) = \frac{2\mu_x^2}{m}$.

Another name for the scaled chi-squared distribution is **one-dimensional Wishart distribution** $W_1(\frac{\mu_x}{m}, m)$.

The **gamma distribution** $\text{Gam}(\alpha, \beta)$ is another name for the scaled chi-squared distribution with a different parameterisation in terms of a shape parameter α and a scale parameter β. The scaled chi-squared distribution $\frac{\mu_x}{m}\chi_m^2$ equals $\text{Gam}(\alpha = \frac{1}{2}m, \beta = 2\frac{\mu_x}{m})$. The mean of $\text{Gam}(\alpha, \beta)$ is $\alpha\beta = \mu_x$ and its variance is $\alpha\beta^2 = \frac{2\mu_x^2}{m}$.

The density of the gamma distribution (aka scaled chi-squared distribution) is available in the R function `dgamma()`. The cumulative density function is `pgamma()` and the quantile function is `qgamma()`.

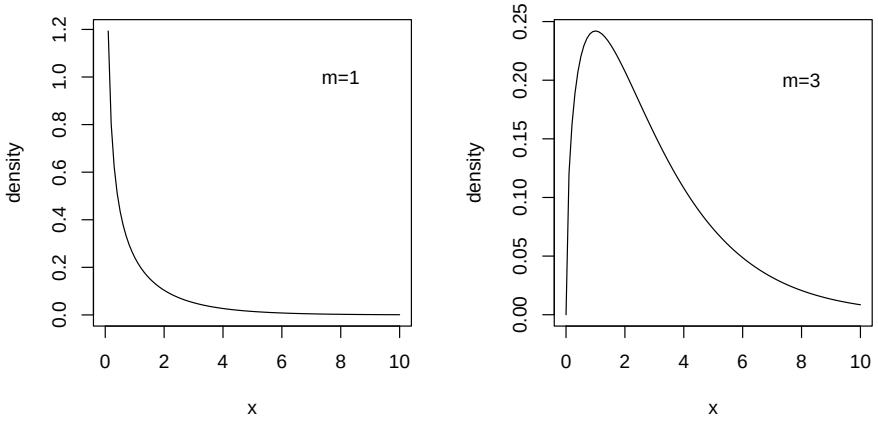
A.6.4 Special cases of the gamma distribution: exponential distribution and chi-squared distibution

The **chi-squared distribution** is a special case with $\sigma_z^2 = 1$ and hence $\frac{\mu_x}{m} = 1$. It has mean $\text{E}(x) = m$ and variance $\text{Var}(x) = 2m$. The chi-squared distribution χ_m^2 equals $\text{Gam}(\frac{m}{2}, 2)$.

The **exponential distribution** $\text{Exp}(\beta)$ with scale parameter β is another special case of the gamma distribution with shape parameter $\alpha = 1$ (or $m = 2$) and thus equals $\text{Gam}(1, \beta)$. It has mean β and variance β^2.

Instead of the scale parameter the exponential distribution is also often specified using a rate parameter $\lambda = \frac{1}{\beta}$.

Here is a plot of the density of the chi-squared distribution for degrees of freedom $m = 1$ and $m = 3$:



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()`.

Likewise, in R `dexp()` gives the density of the exponential distribution, and `pexp()` and `qexp()` are the corresponding cumulutative density and quantile functions.

A.7 Statistics

A.7.1 Statistical learning

The aim in statistics - data science - machine learning is to learn from data (from experiments, observations, measurements) to learn about and understand the world.

Specifically, to identify the best model(s) for the data in order to

- to explain the current data, and
- to enable good prediction of future data

Note that it is easy to get models that only explain the data but do not predict well!

This is called *overfitting* the data and happens in particular if the model is overparameterized for the amount of data available.

Specifically, we have data $x_1, \ldots, x_n$ and models $f(x|\theta)$ that are indexed the parameter θ.

Often (but not always) θ can be interpreted and/or is associated with some property of the model.

If there is only a single parameter we write θ (scalar parameter). For a parameter vector we write $\boldsymbol{\theta}$ (in bold type).

A.7.2 Point and interval estimation

- There is a parameter θ of interest in a model
- we are uncertain about this parameter (i.e. we don't know the exact value)
- we would like to learn about this parameter by observing data $x_1, \ldots, x_n$ from the model

Often the parameter(s) of interest are related to moments (such as mean and variance) or to quantiles of the distribution representing the model.

Estimation:

- An **estimator for** θ is a function $\hat{\theta}(x_1, \ldots, x_n)$ that maps the data (input) to a "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.

Simple estimators of mean and variance:

Suppose we have data $x_1, \ldots, x_n$ all sampled independently from a distribution F.

- The average (also known as empirical mean) $\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i$ is an estimate of the mean of F.
- The empirical variance $\widehat{\sigma^2}_{\text{ML}} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})^2$ is an estimate of the variance of F. Note the factor $1/n$. It is the maximum likelihood estimate assuming a normal model.
- The unbiased sample variance $\widehat{\sigma^2}_{\text{UB}} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \hat{\mu})^2$ is another estimate of the variance of F. Note the factor $1/(n-1)$ therefore $n \geq 2$ is required for this estimator.

A.7.3 Sampling properties of a point estimator $\hat{\theta}$

A point estimator $\hat{\theta}$ depends on the data, hence it has 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.7.4 Sampling distribution of mean and variance estimators for normal data

Suppose we have data $x_1, \ldots, x_n$ all sampled from a normal distribution $N(\mu, \sigma^2)$.

- 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 $\text{Var}(\hat{\mu}) = \frac{\sigma^2}{n}$. The estimate $\hat{\mu}$ is unbiased since $\mathrm{E}(\hat{\mu}) - \mu = 0$ The mean squared error of $\hat{\mu}$ is $\text{MSE}(\hat{\mu}) = \frac{\sigma^2}{n}$.

- The empirical variance $\widehat{\sigma^2}_{\text{ML}} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2$ for normal data follows a scaled chi-squared distribution or equivalently a Gamma distribution

$$\widehat{\sigma^2}_{\text{ML}} \sim \frac{\sigma^2}{n}\chi^2_{n-1} = \text{Gam}\left(\underbrace{\frac{n-1}{2}}_{\text{shape}}, \underbrace{\frac{2\sigma^2}{n}}_{\text{scale}}\right)$$

  Thus, $\mathrm{E}(\widehat{\sigma^2}_{\text{ML}}) = \frac{n-1}{n}\sigma^2$ and $\text{Var}(\widehat{\sigma^2}_{\text{ML}}) = \frac{2(n-1)}{n^2}\sigma^4$. The estimate $\widehat{\sigma^2}_{\text{ML}}$ is biased since $\mathrm{E}(\widehat{\sigma^2}_{\text{ML}}) - \sigma^2 = -\frac{1}{n}\sigma^2$. The mean squared error is $\text{MSE}(\widehat{\sigma^2}_{\text{ML}}) = \frac{2(n-1)}{n^2}\sigma^4 + \frac{1}{n^2}\sigma^4 = \frac{2n-1}{n^2}\sigma^4$.

- The unbiased variance $\widehat{\sigma^2}_{\text{UB}} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{\mu})^2$ for normal data follows a scaled chi-squared distribution or equivalently a Gamma distribution

$$\widehat{\sigma^2}_{\text{UB}} \sim \frac{\sigma^2}{n-1}\chi^2_{n-1} = \text{Gam}\left(\underbrace{\frac{n-1}{2}}_{\text{shape}}, \underbrace{\frac{2\sigma^2}{n-1}}_{\text{scale}}\right)$$

  Thus, $\text{E}(\widehat{\sigma^2}_{\text{UB}}) = \sigma^2$ and $\text{Var}(\widehat{\sigma^2}_{\text{UB}}) = \frac{2}{n-1}\sigma^4$. The estimate $\widehat{\sigma^2}_{\text{ML}}$ is unbiased since $\text{E}(\widehat{\sigma^2}_{\text{UB}}) - \sigma^2 = 0$. The mean squared error is $\text{MSE}(\widehat{\sigma^2}_{\text{UB}}) = \frac{2}{n-1}\sigma^4$.

  Note that for any $n > 1$ we find that $\text{Var}(\widehat{\sigma^2}_{\text{UB}}) > \text{Var}(\widehat{\sigma^2}_{\text{ML}})$ and $\text{MSE}(\widehat{\sigma^2}_{\text{UB}}) > \text{MSE}(\widehat{\sigma^2}_{\text{ML}})$ so that the biased empirical estimator has both lower variance and lower mean squared error than the unbiased estimator.

A.7.5 Asymptotics

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.

The typical rate of decrease of variance of a good estimator is $\frac{1}{n}$. Thus, when sample size is doubled the variance is divided by 2 (and the standard deviation is divided by $\sqrt{2}$).

Consistency: $\hat{\theta}$ is called consistent if

$$\text{MSE}(\hat{\theta}) \longrightarrow 0 \text{ with } n \to \infty$$

The three estimators discussed above (empirical mean, empirical variance, unbiased variance) are all consistent as their MSE goes to zero with large sample size n.

Consistency is a *minimum* essential requirement for any reasonable estimator! Of all consistent estimators we typically prefer the estimator that is most efficient (i.e. with fasted decrease in MSE) and that therefore has 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.

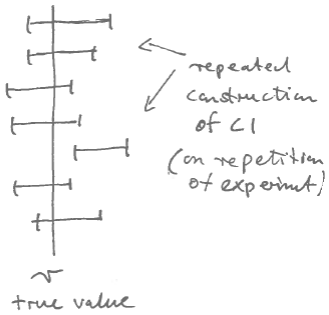
If the model class does not contain the true model then strict consistency cannot be achived but we still wish to get as close as possible to the true model when choosing model parameters.

A.7.6 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:

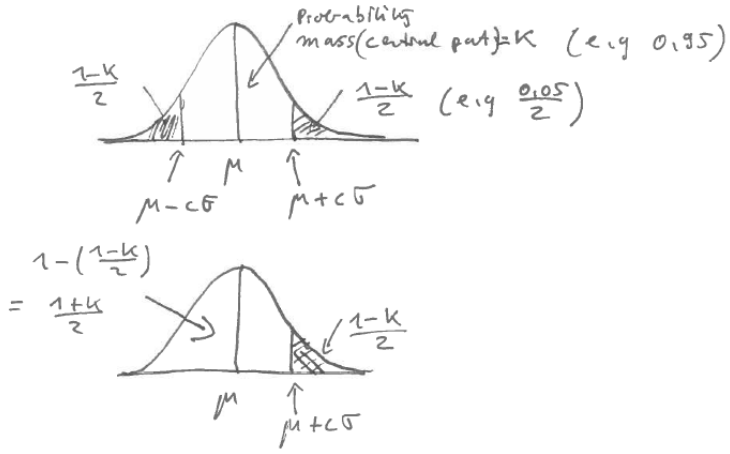


- Note that a CI is actually an **estimate**: $\widehat{\text{CI}}(x_1, \ldots, x_n)$, i.e. it depends on data and has a random (sampling) variation.
- A good CI has high coverage and is compact.

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.7.7 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
0.99	2.58

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

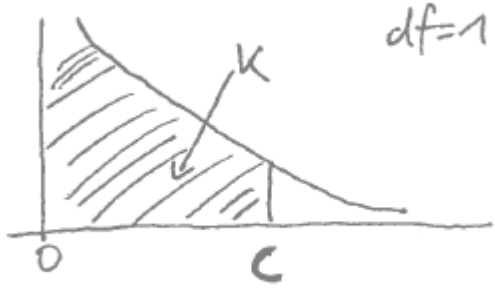
- a scalar parameter θ
- with normally distributed estimate $\hat{\theta}$ and
- with estimated standard deviation $\hat{\text{SD}}(\hat{\theta}) = \hat{\sigma}$

is then given by

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

where c is chosen for desired coverage level κ.

A.7.8 Confidence interval for 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

Further study

In this module we can only touch the surface of likelihood and Bayes inference. As a starting point for further reading the following text books are recommended.

B.1 Recommended reading

- Faraway (2015) *Linear Models with R (second edition)*. Chapman and Hall/CRC.
- Held and Bové (2020) *Applied Statistical Inference: Likelihood and Bayes (2nd edition)*. Springer.
- Agresti and Kateri (2022) *Foundations of Statistics for Data Scientists*. Chapman and Hall/CRC.

B.2 Additional references

- Heard (2021) *An Introduction to Bayesian Inference, Methods and Computation*. Springer.
- Gelman et al. (2014) *Bayesian data analysis (3rd edition)*. CRC Press.
- Wood (2015) *Core Statistics*. Cambridge University Press. PDF available from https://www.maths.ed.ac.uk/~swood34/core-statistics-nup.pdf

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.

Faraway, J. J. 2015. *Linear Models with R*. 2nd ed. Chapman; Hall/CRC.

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.

Wood, S. 2015. *Core Statistics*. Cambridge University Press.