BAYESIAN STATISTICS
Lecture Notes 2015

B.J.K. Kleijn

University of Amsterdam
Korteweg-de Vries Institute for Mathematics
# Contents

## Preface

1 Introduction

1.1 Frequentist statistics

1.1.1 Data and model

1.1.2 Frequentist estimation

1.2 Bayesian statistics

1.3 The frequentist analysis of Bayesian methods

1.4 Exercises

2 Bayesian basics

2.1 Bayes’s rule, prior and posterior distributions

2.1.1 Bayes’s rule

2.1.2 Bayes’s billiard

2.1.3 The Bayesian view of the model

2.1.4 A frequentist’s view of the posterior

2.2 Bayesian point estimators

2.2.1 Posterior predictive and posterior mean

2.2.2 Small-ball and formal Bayes estimators

2.2.3 The Maximum-A-Posteriori estimator

2.3 Confidence sets and credible sets

2.3.1 Frequentist confidence intervals

2.3.2 Bayesian credible sets

2.4 Tests and Bayes factors

2.4.1 Neyman-Pearson tests

2.4.2 Randomized tests and the Neyman-Pearson lemma

2.4.3 Test sequences and minimax optimality

2.4.4 Bayes factors

2.5 Decision theory and classification

2.5.1 Decision theory

2.5.2 Bayesian decision theory
Preface

These lecture notes were written for the course ‘Bayesian Statistics’, taught at University of Amsterdam in the spring of 2007. The course was aimed at first-year MSc.-students in statistics, mathematics and related fields. The aim was for students to understand the basic properties of Bayesian statistical methods; to be able to apply this knowledge to statistical questions and to know the extent (and limitations) of conclusions based thereon. Considered were the basic properties of the procedure, choice of the prior by objective and subjective criteria and Bayesian methods of inference. In addition, some non-parametric Bayesian modelling and basic posterior asymptotic behaviour will received due attention.

An attempt has been made to make these lecture notes as self-contained as possible. Nevertheless the reader is expected to have been exposed to some statistics, preferably from a mathematical perspective. It is not assumed that the reader is familiar with asymptotic statistics. Where possible, definitions, lemmas and theorems have been formulated such that they cover parametric and nonparametric models alike. An index, references and an extensive bibliography are included.

Since Bayesian statistics is formulated in terms of probability theory, some background in measure theory is prerequisite to understanding these notes in detail. However the reader is not supposed to have all measure-theorical knowledge handy: appendix A provides an overview of relevant measure-theoretic material. In the description and handling of nonparametric statistical models, functional analysis and topology play a role. Of the latter two, however, only the most basic notions are used and all necessary detail in this respect will be provided during the course.

For corrections to early versions of these notes the author thanks C. Muris.

Bas Kleijn, Amsterdam, January 2007
Chapter 1

Introduction

The goal of statistical inference is to understand, describe and estimate (aspects of) the randomness of measured data. Quite naturally, this invites the assumption that the data represents a sample from an unknown but fixed probability distribution.

1.1 Frequentist statistics

Any frequentist inferential procedure relies on three basic ingredients: the data, a model and an estimation procedure. The central assumption in frequentism is that the data has a definite but unknown, underlying distribution to which all inference pertains.

1.1.1 Data and model

The \textit{data} is a measurement or observation which we denote by $Y$, taking values in a corresponding samplespace.

\textbf{Definition 1.1.} The samplespace for an observation $Y$ is a measurable space $(\mathcal{Y}, \mathcal{B})$ (see definition A.2) containing all values that $Y$ can take upon measurement.

Measurements and data can take any form, ranging from \textit{categorical data} (sometimes referred to as \textit{nominal data} where the samplespace is simply a (usually finite) set of points or labels with no further mathematical structure), \textit{ordinal data} (also known as \textit{ranked data}, where the samplespace is endowed with a total ordering), to \textit{interval data} (where in addition to having an ordering, the samplespace allows one to compare differences or distances between points), to \textit{ratio data} (where we have all the structure of the real line). Moreover $Y$ can collect the results of a number of measurements, so that it takes its values in the form of a vector (think of an experiment involving repeated, stochastically independent measurements of the same quantity, leading to a so-called independent and identically distributed (or \textit{i.i.d.}) sample). The data $Y$ may even be \textit{functional data}, taking its values in a space of functions or...
in other infinite-dimensional spaces, for example, in the statistical study of continuous-time time-series.

The samplespace $\mathcal{Y}$ is assumed to be a measurable space to enable the consideration of probability measures on $\mathcal{Y}$, formalizing the uncertainty in measurement of $Y$. As was said in the opening words of this chapter, frequentist statistics hinges on the assumption that there exists a probability measure $P_0 : \mathcal{B} \to [0,1]$ on the samplespace $\mathcal{Y}$ representing the “true distribution of the data”:

$$Y \sim P_0$$

(1.1)

Hence from the frequentist perspective, statistics revolves around the central question: “What does the data make clear about $P_0$?”, which may be considered in parts by questions like, “From the data, what can we say about the mean of $P_0$?”, “Based on the data that we have, how sharp can we formulate hypotheses concerning the value of the variance of $P_0$?”, etcetera.

The second ingredient of a statistical procedure is a model, which contains all explanations under consideration of the randomness in $Y$.

**Definition 1.2.** A (frequentist) statistical model $\mathcal{P}$ is a collection of probability measures $P : \mathcal{B} \to [0,1]$ on the samplespace $(\mathcal{Y}, \mathcal{B})$. The distributions $P$ are called model distributions. For every sample space $(\mathcal{Y}, \mathcal{B})$, the collection of all probability distributions is called the full model (sometimes referred to as the full non-parametric model), denoted $\mathcal{M}(\mathcal{Y}, \mathcal{B})$.

The model $\mathcal{P}$ contains the candidate distributions for $Y$ that the statistician finds “reasonable” explanations of the uncertainty he observes (or expects to observe) in $Y$. As such, it constitutes a choice of the statistician analyzing the data rather than a given. From a more mathematical perspective we observe that a model $\mathcal{P}$ on $(\mathcal{Y}, \mathcal{B})$ is a subset of the space $ca(\mathcal{Y}, \mathcal{B})$ of all signed measures $\mu : \mathcal{B} \to \mathbb{R}$ (that is, all countably additive, real-valued set functions) that are of finite total variation. Equipped with the total-variational norm (see appendix A, definition A.7), $\mu \mapsto \|\mu\|$, $ca(\mathcal{Y}, \mathcal{B})$ is a Banach space, in which the full model can be characterized as follows:

$$\mathcal{M}(\mathcal{Y}, \mathcal{B}) = \{ P \in ca(\mathcal{Y}, \mathcal{B}) : P \geq 0, P(\mathcal{Y}) = 1 \}.$$ 

Often, we describe models as families of probability densities rather than distributions.

**Definition 1.3.** If there exists a $\sigma$-finite measure $\mu : \mathcal{B} \to [0,\infty]$ such that for all $P \in \mathcal{P}$, $P \ll \mu$, we say that the model is dominated.

The Radon-Nikodym theorem (see theorem A.26) guarantees that we may represent a dominated probability measure $P$ in terms of a probability density function $p = dP/d\mu : \mathcal{Y} \to [0,\infty)$ that satisfies $\int_A p(y) \, d\mu(y) = P(A)$ for all $A \in \mathcal{B}$. For dominated models, it makes sense to adopt a slightly different mathematical perspective: if $\mu$ dominates $\mathcal{P}$, we map $\mathcal{P}$ to the space of all $\mu$-integrable functions $L_1(\mu)$ by means of the Radon-Nikodym mapping.
Example 1.4. Suppose that \( \mathcal{Y} \) is countable (and let \( \mathcal{B} \) be the powerset of \( \mathcal{Y} \)): then the measure \( \mu \) that puts mass one at every point in \( \mathcal{Y} \), also known as the counting measure on \( \mathcal{Y} \), is \( \sigma \)-finite and dominates every other (finite) measure on \( \mathcal{Y} \). Consequently, any model on \((\mathcal{Y}, \mathcal{B})\) can be represented in terms of elements \( p \) in the Banach space \( L_1(\mu) \), more commonly denoted as \( \ell_1 \),

\[
\ell_1 = \{(f_1, f_2, \ldots) \in [0, 1]^\infty : \sum_{i \geq 1} |f_i| < \infty\}.
\]

where it is noted that \( p_i \geq 0 \) and \( \|p\| = \sum_i |p_i| = \sum_i p_i = 1 \) for all \( P \in \mathcal{M}(\mathcal{Y}, \mathcal{B}) \).

In case the sample space is not discrete, the full model is not dominated by a \( \sigma \)-finite measure. However, suppose that a \( \sigma \)-finite measure on the sample space is given. The Radon-Nikodym mapping maps every \( \mu \)-dominated model \( P \) to a subset of,

\[
\mathcal{M}(\mu) = \{p \in L_1(\mu) : p \geq 0, \int_\mathcal{Y} p(y) \, d\mu(y) = 1\}.
\]

Lemma 1.5. The mapping between a model \( \mathcal{P} \) dominated by a \( \sigma \)-finite measure \( \mu \) and its \( L_1(\mu) \)-representation is an isometry: for all \( P_1, P_2 \in \mathcal{P} \), with \( \mu \)-densities \( p_1, p_2 \),

\[
\|P_1 - P_2\| = \frac{1}{2} \int_\mathcal{Y} |p_1(y) - p_2(y)| \, d\mu(y).
\]

Note that for any dominated model, the dominating measure is not unique and hence, that the \( L_1 \)-representation of \( \mathcal{P} \) depends on the particular choice of dominating measure \( \mu \).

The common way of representing a model is a description in terms of a parameterization.

Definition 1.6. A model \( \mathcal{P} \) is parameterized with parameter space \( \Theta \), if there exists a surjective map \( \Theta \to \mathcal{P} : \theta \mapsto P_\theta \), called the parameterization of \( \mathcal{P} \).

Parametrizations are motivated by the context of the statistical question and the parameter \( \theta \) usually has a clear interpretation when viewed in this context. The formulation of parametric model descriptions is the modelling step of statistics: to the statistician, it transforms the data from a mere list of numbers to an informative (but noisy) representation of an underlying truth.

Definition 1.7. A parameterization of a statistical model \( \mathcal{P} \) is said to be identifiable, if the map \( \Theta \to \mathcal{P} : \theta \mapsto P_\theta \) is injective.

Injectivity of the parameterization means that for all \( \theta_1, \theta_2 \in \Theta, \theta_1 \neq \theta_2 \) implies that \( P_{\theta_1} \neq P_{\theta_2} \). In other words, no two different parameter values \( \theta_1 \) and \( \theta_2 \) give rise to the same distribution. Clearly, in order for \( \theta \in \Theta \) to serve as a useful representation for the candidate distributions \( P_\theta \), identifiability is a first requirement. Other common conditions on the map \( \theta \mapsto P_\theta \) are continuity (for example, with respect to the total-variational norm), differentiability (the definition of which may involve technical subtleties in case \( \Theta \) is infinite-dimensional) and other smoothness conditions.
Remark 1.8. Although strictly speaking ambivalent, it is commonplace to refer to both $\mathcal{P}$
and the parameterizing space $\Theta$ as “the model”. This practice is not unreasonable in view of
the fact that, in practice, almost all models are parameterized in an identifiable way, so that
there exists a bijective correspondence between $\Theta$ and $\mathcal{P}$. In these lecture notes, reference to
the model concerns $\mathcal{P}$ while $\Theta$ is always called the parameter space.

A customary assumption in frequentist statistics is that the model is well-specified.

Definition 1.9. A model $\mathcal{P}$ is said to be well-specified if it contains the true distribution of
the data $P_0$, i.e.

$$P_0 \in \mathcal{P}. \quad (1.2)$$

If (1.2) does not hold, the model is said to be mis-specified.

Clearly if $\mathcal{P}$ is parameterized by $\Theta$, (1.2) implies the existence of a point $\theta_0 \in \Theta$ such that
$P_{\theta_0} = P_0$; if, in addition, the model is identifiable, the parameter value $\theta_0$ is unique.

If the full non-parametric model is used, (1.2) holds trivially. However, for smaller models,
(1.2) has the status of an assumption on the unknown quantity of interest $P_0$ and may, as
such, be hard to justify. The reason for (the somewhat odd and certainly very contentious)
assumption (1.2) lies in the interpretation of statistical conclusions: an estimate of a parameter
is of value if that parameter can be ascribed to the “true” distribution of the data. If, on
the other hand, one assumes that the model is mis-specified, parameter estimates may reflect
aspects of the true distribution but cannot be associated with the true distribution of the
data directly any more.

The model we use in a statistical procedure constitutes a choice rather than a given: pre-
sented with a particular statistical problem, different statisticians may choose to use different
models. The only condition is that (1.2) is satisfied, which is why we have to choose the model
in a “reasonable way” given the nature of $Y$. When choosing the model, two considerations
compete: on the one hand, small models are easy to handle mathematically and statistically
and parameters usually have clear interpretations, on the other hand, for large models,
assumption (1.2) is more realistic since they have a better chance of containing $P_0$ (or at least
approximate it more closely). The amount of data available plays a crucial role: if we have
a limited sample, simple models have a better chance of leading to sensible results, while an
abundance of data enables more sophisticated forms of statistical analysis. In this respect the
most important distinction is made in terms of the dimension of the model.

Definition 1.10. A model $\mathcal{P}$ is said to be parametric of dimension $d$, if there exists an
identifiable parameterization $\Theta \to \mathcal{P}: \theta \mapsto P_\theta$, where $\Theta \subset \mathbb{R}^d$ with non-empty interior
$\Theta \neq \emptyset$.

The requirement regarding the interior of $\Theta$ in definition 1.10 ensures that the dimension
$d$ really concerns $\Theta$ and not just the dimension of the space $\mathbb{R}^d$ (in which $\Theta$ could otherwise
be a lower-dimensional subset).
Example 1.11. The normal model for a single, real measurement $Y$, is the collection of all normal distributions on $\mathbb{R}$, i.e.

$$\mathcal{P} = \{ N(\mu, \sigma^2) : (\mu, \sigma) \in \Theta \}$$

where the parameterizing space $\Theta$ equals $\mathbb{R} \times (0, \infty)$. The map $(\mu, \sigma) \mapsto N(\mu, \sigma^2)$ is surjective and injective, i.e. the normal model is a two-dimensional, identifiable parametric model. Moreover, the normal model is dominated by the Lebesgue measure on the samplespace $\mathbb{R}$ and can hence be described in terms of Lebesgue-densities:

$$p_{\mu, \sigma}(y) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}.$$

Note that for any fixed $y \in \mathcal{Y}$, the dependence $\Theta \rightarrow \mathbb{R} : (\mu, \sigma) \mapsto p_{\mu, \sigma}(y)$ is continuous on all of $\Theta$. So if $(\mu_n, \sigma_n)$ converges to $(\mu, \sigma)$ in $\Theta$, then $p_n(y) := p_{\mu_n, \sigma_n}(y)$ converges to $p(y) := p_{\mu, \sigma}(y)$. Then total-variational distance between the distributions $P_n$ and $P$ (associated with the densities $p_n$ and $p$ respectively) satisfies,

$$\| P_n - P \| = \frac{1}{2} \int_{\mathcal{Y}} |p_n(y) - p(y)| \, d\mu(y) \rightarrow 0.$$

by lemma 1.5 and Scheffé’s lemma. Conclude that the parametrization $\Theta \rightarrow \mathcal{P} : (\mu, \sigma) \mapsto P_{\mu, \sigma}$ is continuous with respect to the total-variational metric on $\mathcal{P}$.

Definition 1.12. If there is no finite-dimensional subset $\Theta$ that parameterizes $\mathcal{P}$, then $\mathcal{P}$ is called a non-parametric model.

For instance, the full model $\mathcal{M}(\mathcal{Y}, \mathcal{B})$ is non-parametric unless the samplespace contains a finite number of points.

Example 1.13. Let $\mathcal{Y}$ be a finite set containing $n \geq 1$ points $y_1, y_2, \ldots, y_n$ and let $\mathcal{B}$ be the power-set $2^\mathcal{Y}$ of $\mathcal{Y}$. Any probability measure $P : \mathcal{B} \rightarrow [0, 1]$ on $(\mathcal{Y}, \mathcal{B})$ is absolutely continuous with respect to the counting measure on $\mathcal{Y}$ (see example A.11). The density of $P$ with respect to the counting measure is a map $p : \mathcal{Y} \rightarrow \mathbb{R}$ such that $p \geq 0$ and

$$\sum_{i=1}^{n} p(y_i) = 1.$$

As such, $P$ can be identified with an element of the so-called simplex $S_n$ in $\mathbb{R}^n$, defined as follows

$$S_n = \left\{ p = (p_1, \ldots, p_n) \in \mathbb{R}^n : p_i \geq 0, \sum_{i=1}^{n} p_i = 1 \right\}.$$

This leads to an identifiable parameterization $S_n \rightarrow \mathcal{P} : p \mapsto P$ of the full non-parametric model on $(\mathcal{Y}, \mathcal{B})$, of dimension $n-1$. Note that $S_n$ has empty interior in $\mathbb{R}^n$, but can be brought in one-to-one correspondence with a compact set in $\mathbb{R}^{n-1}$ with non-empty interior by
the embedding:
\[
\left\{(p_1, \ldots, p_{n-1}) \in \mathbb{R}^{n-1} : p_i \geq 0, \sum_{i=1}^{n-1} p_i \leq 1\right\} \rightarrow S_n : \\
(p_1, \ldots, p_{n-1}) \mapsto \left(p_1, \ldots, p_{n-1}, 1 - \sum_{i=1}^{n-1} p_i\right).
\]

1.1.2 Frequentist estimation

The third ingredient of a frequentist inferential procedure is an estimation method. Clearly not all statistical problems involve an explicit estimation step and of those that do, not all estimate the distribution $P_0$ directly. Nevertheless, one may regard the problem of point-estimation in the model $\mathcal{P}$ as prototypical.

**Definition 1.14.** A point-estimator (or estimator) for $P_0$ is a map $\hat{P} : \mathcal{Y} \rightarrow \mathcal{P}$, representing our “best guess” $\hat{P} \in \mathcal{P}$ for $P_0$ based on the data $Y$ (and other known quantities).

Note that a point-estimator is a statistic: since a point-estimator must be calculable in practice, it may depend only on information that is known to the statistician after he has performed the measurement realized as $Y = y$. Also note that a point-estimator is a stochastic quantity: $\hat{P} = \hat{P}(Y)$ depends on $Y$ and is hence random. Upon measurement of $Y$ resulting in a realisation $Y = y$, the realisation of the estimator is an estimate $\hat{P}(y)$, a definite point in $\mathcal{P}$. If the model is parameterized, one may define a point-estimator $\hat{\theta} : \mathcal{Y} \rightarrow \Theta$ for $\theta_0$, from which we obtain $\hat{P} = P_{\hat{\theta}}$ as an estimator for $P_0$. If the model is identifiable, estimation of $\theta_0$ in $\Theta$ is equivalent to estimation of $P_0$ in $\mathcal{P}$.

Aside from estimates for the distribution $P_0$, one is often interested in estimating only certain aspects of $P_0$.

**Example 1.15.** Suppose that a bank tries to assess market risk for an asset: they have the asset on the books for price $x$ but tomorrow’s market will say that it is worth a price $X$, distributed according to an unknown $P_0$. To assess the risk of holding the position until tomorrow, the absolute return $X - x$ is of importance. Of course, the bank would prefer to have a reliable estimate for $P_0$ (and thus for the distribution of $X - x$) but that question is often too hard to answer and reliability cannot be guaranteed. Instead, the bank will resort to a simplification by focussing on the aspect of the distribution $P_0$ that they find most important for their risk assessment. A popular notion in this context is a quantity called value-at-risk: given a time-horizon (in this case, tomorrow) and a significance level $\alpha \in (0, 1)$ (often chosen equal to 0.05), value-at-risk $q$ is defined as the minimal $q$ at which,

\[
P_0(X - x < q) \geq \alpha.
\]

To interpret $q$, note that losses exceeding value-at-risk occur on only an expected fraction $\alpha$ of all trading days. In mathematical terms, $q$ is a quantile of $P_0$. So to come to grips with this
particular form of risk assessment, estimation of the quantile \( q \) suffices: we estimate only an aspect of the distribution \( P_0 \).

Another example occurs in parametric models: if the dimension \( d \) of a parametric model is greater than one, we may choose to estimate only one component of \( \theta \) (called the parameter of interest) and disregard other components (called nuisance parameters). More generally, we may choose to estimate certain properties of \( P_0 \) (e.g. its expectation, variance) rather than \( P_0 \) itself and in many cases, direct estimation of the property of interest of \( P_0 \) is more efficient than estimation through \( \hat{P} \).

**Example 1.16.** Consider a model \( \mathcal{P} \) consisting of distributions on \( \mathbb{R} \) with finite expectation and define the functional \( e : \mathcal{P} \to \mathbb{R} \) by the expectation \( e(P) = PX \). Suppose that we are interested in the expectation \( e_0 = e(P_0) \) of the true distribution. Obviously, based on an estimator \( \hat{P} \) for \( P_0 \) we may define an estimator,

\[
\hat{e} = \int_{\mathbb{R}} x \, d\hat{P}(x) \tag{1.3}
\]

to estimate \( e_0 \). For instance, assume that \( X \) is integrable under \( P_0 \) and \( Y = (X_1, \ldots, X_n) \) collects the results of an i.i.d. experiment with \( X_i \sim P_0 \) marginally (for all \( 1 \leq i \leq n \)), then the empirical expectation of \( X \), defined simply as the sample-average of \( X 
[
\]
provides an estimator for \( e_0 \). (Note that the sample-average is also of the form (1.3) if we choose as our point-estimator for \( P_0 \) the empirical distribution \( \hat{P} = \mathbb{P}_n \).) The law of large numbers guarantees that \( \mathbb{P}_n X \) converges to \( e_0 \) almost-surely as \( n \to \infty \), and (if \( X \) is quadratically integrable) the central limit theorem asserts that this convergence proceeds at rate \( n^{-1/2} \) (and that the limit distribution is zero-mean normal with \( P_0(X - P_0 X)^2 \) as its variance). Many parameterizations \( \theta \mapsto P_\theta \) are such that (components of) \( \theta \) coincide with expectations. Often, other properties of \( P_0 \) can also be related to expectations: for example, if \( X \in \mathbb{R} \), the probabilities \( F_0(s) = P_0(X \leq s) = P_01\{X \leq s\} \) can be estimated by

\[
\hat{F}(s) = \frac{1}{n} \sum_{i=1}^{n} 1\{X_i \leq s\}
\]

i.e. as the empirical expectation of the function \( x \mapsto 1\{x \leq s\} \). This leads to a step-function with \( n \) jumps of size \( 1/n \) at samplepoints, which estimates the distribution function \( F_0 \). Generalizing, any property of \( P_0 \) that can be expressed in terms of an expectation of a \( P_0 \)-integrable function of \( X \), \( P_0(g(X)) \), is estimable by the corresponding empirical expectation, \( \mathbb{P}_n g(X) \).

To estimate a probability distribution (or any of its properties or parameters), many different estimators may exist. Therefore, the use of any particular estimator constitutes (another) choice made by the statistician analyzing the problem. Whether such a choice is a
good or a bad one depends on optimality criteria, which are either dictated by the particular nature of the problem (see section 2.5 which extends the purely inferential point of view), or based on more generically desirable properties of the estimator (note the use of the rather ambiguous qualification “best guess” in definition 1.14).

**Example 1.17.** To illustrate what we mean by “desirable properties”, note the following. When estimating \( P_0 \) one may decide to use an estimator \( \hat{P} \) because it has the property that it is close to the true distribution of \( Y \) in total variation. To make this statement specific, the property that make such an estimator \( \hat{P} \) attractive is that there exist small constants \( \epsilon > 0 \) and \( \alpha > 0 \), such that for all \( P \in \mathcal{P} \),

\[
P(\|\hat{P}(Y) - P\| < \epsilon) > 1 - \alpha,
\]

i.e. if \( Y \sim P \), then \( \hat{P} \) lies close to \( P \) with high \( P \)-probability. Note that we formulate this property “for all \( P \) in the model”: since \( P_0 \in \mathcal{P} \) is unknown, the only way to guarantee that this property holds under \( P_0 \), is to prove that it holds for all \( P \in \mathcal{P} \), provided that (1.2) holds.

A popular method of estimation that satisfies common optimality criteria in many (but certainly not all, see [67]) problems is maximum-likelihood estimation.

**Definition 1.18.** Suppose that the model \( \mathcal{P} \) is dominated by a \( \sigma \)-finite measure \( \mu \). The likelihood principle says that one should pick \( \hat{P} \in \mathcal{P} \) as an estimator for the distribution \( P_0 \) of \( Y \) such that the density \( \hat{p} \) satisfies,

\[
\hat{p}(Y) = \sup_{P \in \mathcal{P}} p(Y).
\]

So \( \hat{P} \) is the point in the model for which the likelihood map evaluated in \( Y \), \( \mathcal{P} \rightarrow [0, \infty) : P \mapsto p(Y) \) attains its maximum. This defines the maximum-likelihood estimator (or MLE) \( \hat{P} \) for \( P_0 \).

**Remark 1.19.** The MLE \( \hat{P} \) does not depend on the particular dominating measure \( \mu \) (see exercise 1.1).

A word of caution is in order: mathematically, the above “definition” of the MLE begs questions of existence and uniqueness: regarding \( P \mapsto p(Y) \) as a (stochastic) map on the model (called the likelihood), there may not be any point in \( \mathcal{P} \) where the likelihood takes on its supremal value nor is there any guarantee that such a maximal point is unique with \( P_0 \)-probability equal to one.

**Remark 1.20.** If \( \Theta \rightarrow \mathcal{P} : \theta \mapsto P_\theta \) parameterizes \( \mathcal{P} \), the above is extended to the maximum-likelihood estimator \( \hat{\theta}(Y) \) for \( \theta_0 \), when we note that \( \sup_{\theta \in \Theta} p_\theta(Y) = \sup_{P \in \mathcal{P}} p(Y) \).
for $Y$ exists. It makes clear, however, that frequentist inference concerns itself primarily with the stochastics of the random variable $Y$ and not with the context in which $Y$ resides. Other than the fact that the model has to be chosen “reasonably” based on the nature of $Y$, frequentist inference does not involve any information regarding the background of the statistical problem in its procedures unless one chooses to use such information explicitly (see, for example, remark 2.32 on penalized maximum-likelihood estimation). In Bayesian statistics the use of background information is an integral part of the procedure unless one chooses to disregard it: by the definition of a prior measure, the statistician may express that he believes in certain points of the model more strongly than others. This thought is elaborated on further in section 1.2 (*e.g.* example 1.22).

Similarly, results of estimation procedures are sensitive to the context in which they are used: two statistical experiments may give rise to the same model formally, but the estimator used in one experiment may be totally unfit for use in the other experiment.

**Example 1.21.** For example, if we interested in a statistic that predicts the rise or fall of a certain share-price on the stockmarket based on its value over the past week, the estimator we use does not have to be a very conservative one: we are interested primarily in its long-term performance and not in the occasional mistaken prediction. However, if we wish to predict the rise or fall of white-bloodcell counts in an HIV-patient based on last week’s counts, overly optimistic predictions can have disastrous consequences.

Although in the above example, data and model are very similar in these statistical problems, the estimator used in the medical application should be much more conservative than the estimator used in the stock-market problem. The inferential aspects of both questions are the same, but the context in which such inference is made calls for adaptation. Such considerations form the motivation for statistical decision theory, as explained further in section 2.5.

### 1.2 Bayesian statistics

The subject of these lecture notes is an alternative approach to statistical questions known as Bayesian statistics, after Rev. Thomas Bayes, the author of “An essay towards solving a problem in the doctrine of chances” published posthumously in 1763 [4]. Bayes considered a number of probabilistic questions in which data and parameters are treated on equal footing. The Bayesian procedure itself is explained in detail in chapter 2 and further chapters explore its properties. In this section we have the more modest goal of illustrating the conceptual differences with frequentist statistical analysis.

In Bayesian statistics, data and model form two factors of the same space, *i.e.* no formal distinction is made between measured quantities $Y$ and parameters $\theta$. One may envisage the process of generating a measurement outcome $Y = y$ as two draws, one draw from $\Theta$ (or $\mathcal{P}$) to select a value of $\theta$ (or distribution $P_{\theta}$) and a subsequent draw from $P_{\theta}$ to arrive at $Y = y$. 

This perspective may seem rather absurd in view of the definitions made in section 1.1, but in [4], Bayes gives examples in which this perspective is perfectly reasonable (see example 2.11 in these lecture notes). An element \( P_\theta \) of the model is interpreted simply as the distribution of \( Y \) given the parameter value \( \theta \), i.e. as the conditional distribution of \( Y|\theta \). The joint distribution of \((Y,\theta)\) then follows upon specification of the marginal distribution of \( \theta \) on \( \Theta \), which is called the prior. Based on the joint distribution for the data \( Y \) and the parameters \( \theta \), straightforward conditioning on \( Y \) gives rise to a distribution for the parameters \( \theta | Y \) called the posterior distribution on the model \( \Theta \). Hence, given the model, the data and a prior distribution, the Bayesian procedure leads to a posterior distribution that incorporates the information provided by the data.

Often in applications, the nature of the data and the background of the problem suggest that certain values of \( \theta \) are more “likely” than others, even before any measurements are done. The model \( \mathcal{P} \) describes possible probabilistic explanations of the data and, in a sense, the statistician believes more strongly in certain explanations than in others. This is illustrated by the following example, which is due to L. Savage (1961) [78].

**Example 1.22.** Consider the following three statistical experiments:

1. A lady who drinks milk in her tea claims to be able to tell which was poured first, the tea or the milk. In ten trials, she determines correctly whether it was tea or milk that entered the cups first.

2. A music expert claims to be able to tell whether a page of music was written by Haydn or by Mozart. In ten trials conducted, he correctly determines the composer every time.

3. A drunken friend says that he can predict the outcome of a fair coin-flip. In ten trials, he is right every time.

Let us analyze these three experiments in a frequentist fashion, e.g. we assume that the trials are independent and possess a definite Bernoulli distribution, c.f. (1.1). In all three experiments, \( \theta_0 \in \Theta = [0,1] \) is the per-trial probability that the person gives the right answer. We test their respective claims posing the hypotheses:

\[
H_0 : \quad \theta_0 = \frac{1}{2}, \quad H_1 : \quad \theta_0 > \frac{1}{2}.
\]

The total number of successes out of ten trials is a sufficient statistic for \( \theta \) and we use it as our test-statistics, noting that its distribution is binomial with \( n = 10 \), \( \theta = \theta_0 \) under \( H_0 \). Given the data \( Y \) with realization \( y \) of ten correct answers, applicable in all three examples, we reject \( H_0 \) at p-value \( 2^{-10} \approx 0.1\% \). So there is strong evidence to support the claims made in all three cases. Note that there is no difference in the frequentist analyses: formally, all three cases are treated exactly the same.

Yet intuitively (and also in every-day practice), one would be inclined to treat the three claims on different footing: in the second experiment, we have no reason to doubt the expert’s
claim, whereas in the third case, the friend’s condition makes his claim less than plausible. In the first experiment, the validity of the lady’s claim is hard to guess beforehand. The outcome of the experiments would be as expected in the second case and remarkable in the first. In the third case, one would either consider the friend extremely lucky, or begin to doubt the fairness of the coin being flipped.

The above example convincingly makes the point that in our intuitive approach to statistical issues, we include all knowledge we have, even resorting to strongly biased estimators if the model does not permit a non-biased way to incorporate it. The Bayesian approach to statistics allows us to choose priors that reflect this subjectivity: from the outset, we attach more prior mass to parameter-values that we deem more likely, or that we believe in more strongly. In the above example, we would choose a prior that concentrates more mass at high values of \( \theta \) in the second case and at low values in the third case. In the first case, the absence of prior knowledge would lead us to remain objective, attaching equal prior weights to high and low values of \( \theta \). Although the frequentist’s testing procedure can be adapted to reflect subjectivity, the Bayesian procedure incorporates it rather more naturally through the choice of a prior.

Subjectivist Bayesians view the above as an advantage; objectivist Bayesians and frequentists view it as a disadvantage. Subjectivist Bayesians argue that personal beliefs are an essential part of statistical reasoning, deserving of an explicit role in the formalism and interpretation of results. Objectivist Bayesians and frequentists reject this thought because scientific reasoning should be devoid of any personal beliefs or interpretation. So the above freedom in the choice of the prior is also the Achilles heel of Bayesian statistics: fervent frequentists and objectivist Bayesians take the point of view that the choice of prior is an undesirable source of ambiguity, rather than a welcome way to incorporate “expert knowledge” as in example 1.22. After all, if the subjectivist Bayesian does not like the outcome of his analysis, he can just go back and change the prior to obtain a different outcome. Similarly, if two subjectivist Bayesians analyze the same data they may reach completely different conclusions, depending on the extent to which their respective priors differ.

To a certain extent such ambiguity is also present in frequentist statistics, since frequentists have the freedom to choose biased point-estimators. For example, the use of either a maximum-likelihood or penalized maximum-likelihood estimator leads to differences, the size of which depends on the relative sizes of likelihood and penalty. (Indeed, through the maximum-a-posteriori Bayesian point-estimator (see definition 2.31), one can demonstrate that the log-prior-density can be viewed as a penalty term in a penalized maximum-likelihood procedure, c.f. remark 2.32.) Yet the natural way in which subjectivity is expressed in the Bayesian setting is more explicit. Hence the frequentist or objectivist Bayesian sees in this a clear sign that subjective Bayesian statistics lacks universal value unless one imposes that the prior should not express any bias (see section 3.2).

A second difference in philosophy between frequentist and Bayesian statisticians arises as a
result of the fact that the Bayesian procedure does not require that we presume the existence of a “true, underlying distribution” $P_0$ of $Y$ (compare with (1.1)). The subjectivist Bayesian views the model with (prior or posterior) distribution as his own, subjective explanation of the uncertainty in the data. For that reason, subjectivists prefer to talk about their (prior or posterior) “belief” concerning parameter values rather than implying objective validity of their assertions. On the one hand, such a point of view makes intrinsic ambiguities surrounding statistical procedures explicit; on the other hand, one may wonder about the relevance of strictly personal belief in a scientific tradition that emphasizes universality of reported results.

The philosophical debate between Bayesians and frequentist has raged with varying intensity for decades, but remains undecided to this date. In practice, the choice for a Bayesian or frequentist estimation procedure is usually not motivated by philosophical considerations, but by far more practical issues, such as ease of computation and implementation, common custom in the relevant field of application, specific expertise of the researcher or other forms of simple convenience. Recent developments [3] suggest that the philosophical debate will be put to rest in favour of more practical considerations as well.

1.3 The frequentist analysis of Bayesian methods

Since this point has the potential to cause great confusion, we emphasize the following: this course presents Bayesian statistics from a hybrid perspective, i.e. we consider Bayesian techniques but analyze them with frequentist methods.

We take the frequentist point of view with regard to the data, e.g. assumption (1.1); we distinguish between samplespace and model and we do not adhere to subjectivist interpretations of results (although their perspective is discussed in the main text). On the other hand, we endow the model with a prior probability measure and calculate the posterior distribution, i.e. we use concepts and definitions from Bayesian statistics. This enables us to assess Bayesian methods on equal footing with frequentist statistical methods and extends the range of interesting questions. Moreover, it resolves the inherent ambiguity haunting the subjectivist interpretation of statistical results. Note, however, that the derivation of expression (2.12) (for example), is the result of subjectivist Bayesian assumptions on data and model. Since these assumptions are at odds with the frequentist perspective, we shall take (2.12) as a definition rather than a derived form (see subsection 2.1.4).

Much of the material covered in these lecture notes does not depend on any particular philosophical point of view, especially when the subject matter is purely mathematical. Nevertheless, it is important to realize when philosophical issues may come into play and there will be points where this is the case. In particular when discussing asymptotic properties of Bayesian procedures, adoption of assumption (1.1) is instrumental, basically because discussing convergence requires a limit-point.
Notation and conventions

Throughout these notes, we make use of notation that is common in the mathematical-statistics literature. In addition, the following notational conventions are used. The expectation of a random variable $Z$ distributed according to a probability distribution $P$ is denoted $P_Z$. Samples are denoted $Y$ with realization $y$, or in the case of $n$ i.i.d.-$P_0$ observations, $X_1, \ldots, X_n$. The sample-average (or empirical expectation) for a sample $X_1, \ldots, X_n$, denoted $P_nX$, is defined as $P_nX = n^{-1}\sum_{i=1}^n X_i$ (where it is assumed that $X$ is $P_0$-integrable); the empirical process $G_n$ is defined as $G_nX = n^{1/2}(P_n - P_0)X$ (where it is assumed that $P_0(X - P_0X)^2 < \infty$). The distribution function of the standard normal distribution is denoted $\Phi: \mathbb{R} \rightarrow [0, 1]$. The transpose of a vector $\ell \in \mathbb{R}^d$ is denoted $\ell^T$; the transpose of a matrix $I$ is denoted $I^T$. The formulation "$A(n)$ holds for large enough $n$" should be read as "there exists an $N \geq 1$ such that for all $n \geq N$, $A(n)$ holds".

1.4 Exercises

**Exercise 1.1.** Let $Y \in \mathcal{Y}$ be a random variable with unknown distribution $P_0$. Let $\mathcal{P}$ be a model for $Y$, dominated by a $\sigma$-finite measure $\mu$. Assume that the maximum-likelihood estimator $\hat{P}$ (see definition 1.18) is well-defined, $P_0$-almost-surely. Show that if $\nu$ is a $\sigma$-finite measure dominating $\mu$ and we calculate the likelihood using $\nu$-densities, then the associated MLE is equal to $\hat{P}$. Conclude that the MLE does not depend on the dominating measure used, c.f. remark 1.19.

**Exercise 1.2.** Prove lemma 1.5.

**Exercise 1.3.** In case the sample space is not discrete and the $\sigma$-algebra is rich enough (for example, if $\mathcal{Y} = \mathbb{R}$ with the Borel $\sigma$-algebra), the full model is not dominated by a $\sigma$-finite measure. Demonstrate this fact by example.

**Exercise 1.4.** Although customarily the model is defined first and estimators follow, it is possible to reverse the order: suppose that we have a certain fixed estimator in mind, how should we choose the model in order for the fixed estimator to perform.

More explicitly, consider a data vector $Y = (X_1, \ldots, X_n)$ that forms an i.i.d. sample from a unknown distribution $P_0$ on $\mathbb{R}$. We are interested in estimation of the quantity $\psi = P_0g(X)$, assumed to be finite, where $g: \mathbb{R} \rightarrow \mathbb{R}$ is a given measurable function defined on the sample space for $X$. Examples: if $g$ is the identity map, then $\psi$ is the expectation of $X$; if $g = (X - P_0X)^2$, then $\psi$ is the variance of $X$; if $g = 1\{X \leq x\}$ for some $x \in \mathbb{R}$, then $\psi = F_0(x)$, the value of the distribution function associated with $P_0$ at $x$. In such cases, estimation of $\psi$ by the sample mean $\hat{\psi}_n := \mathbb{P}_ng$ appears sensible for large $n$.

a. Based on the limit $n \rightarrow \infty$, characterize the largest model $\mathcal{P}$ in which estimation of $\psi$ by $\hat{\psi}_n$ makes sense. What can be said of $\mathcal{P}$ if $g$ is a bounded function?
b. Show that $\hat{\psi}_n$ is consistent: $\lim_n \hat{\psi}_n = \psi$, $P_0$-almost-surely.

To analyze the behaviour of $\hat{\psi}_n$ in some more detail, consider the following question.

c. Restrict the model $\mathcal{P}$ further to demonstrate that $\hat{\psi}_n$ converges to $\psi$ at rate $n^{-1/2}$, i.e. for every sequence $M_n \to \infty$, that is,

$$P^n_0 \left( n^{1/2} |\hat{\psi}_n - \psi| > M_n \right) \to 0,$$

as $n \to \infty$.

d. Compare the property under c. above with example 1.17 and state in words how the quality of $\hat{\psi}_n$ as an estimator for $\psi$ improves as $n \to \infty$.

**Exercise 1.5.** In the three experiments of example 1.22, describe the Neyman-Person test for hypotheses $H_0$ and $H_1$ at level $\alpha \in (0,1)$ (but do not try to solve the equation for the critical level). Calculate the $p$-value of the realization of 10 successes and 0 failures (in 10 Bernoulli trials according to $H_0$).
In this chapter, we consider the basic definitions and properties of Bayesian statistical and decision-theoretic methods. We derive the posterior distribution from data, model and prior and we discuss how the posterior should be viewed if one assumes the frequentist point of view of section 2.1. In section 2.2 we consider point estimators derived from the posterior and in section 2.3 we discuss confidence sets and credible sets. Section 2.4 discusses the Neyman-Pearson theory of hypothesis testing, as well as a brief introduction to the Le Cam-Birgé theory of asymptotically optimal test sequences and, of course, posterior odds and Bayes factors. Section 2.5 concludes the chapter with a discussion of minimax risk and Bayes risk, and the differences that occur in decision theory. Throughout the chapter the explicit goal is to consider frequentist methods side-by-side with the Bayesian procedures, for comparison and reference.

It should be stressed that the material presented here covers only the most basic Bayesian concepts; further reading is recommended. Various books providing overviews of Bayesian statistics are recommended, depending on the background and interest of the reader: a highly theoretical treatment can be found in Le Cam (1986) [66], which develops a general, mathematical framework for statistics and decision theory, dealing with Bayesian methods as an important area of its application. For a more down-to-earth version of this work, applied only to smooth parametric models, the interested reader is referred to Le Cam and Yang (1990) [68]. The book by Van der Vaart (1998) [87] contains a chapter on Bayesian statistics focusing on the Bernstein-Von Mises theorem. A general reference of a more decision-theoretic inclination with a clear focus on Bayesian statistics, is the book by Berger (1985) [8]; a more recent reference of a similar nature is Bernardo and Smith (1993) [13]. Both Berger and Bernardo and Smith devote a great deal of attention to philosophical arguments in favour of the Bayesian approach to statistics, staying rather terse with regard to mathematical detail and focusing almost exclusively on parametric models. Recommendable is also Robert’s “The Bayesian choice” (2001) [76], which offers a very useful explanation on computational
aspects of Bayesian statistics. Finally, Ripley (1996) [77] discusses Bayesian methods with a very pragmatic focus on pattern classification. The latter reference relates all material with applications in mind but does so based on a firm statistical and decision-theoretic background.

2.1 Bayes’s rule, prior and posterior distributions

In this section, we introduce the basic definitions and procedures in Bayesian statistics. Formalizing the Bayesian procedure can be done in several ways. We start this section with considerations that are traditionally qualified as being of a “subjectivist” nature: in subsection 2.1.1 we derive the relation between data, model and prior on the one hand, and the posterior on the other based on Bayes’s Rule without reference to the frequentist’s “true distribution of the data”. We take due care with regard to the support of the prior (see subsection 2.1.3) and consider a prototypical example usually referred to as Bayes’s Billiard in subsection 2.1.2. In subsection 2.1.4 we revert to the “frequentist” point of view through an assumption relating the “true distribution of the data” to the prior predictive distribution (see definition 2.4).

2.1.1 Bayes’s rule

The Bayesian framework does not just view the data \( Y \) as a random variable but casts the parameter in that form as well. The parameter space \( \Theta \) is assumed to be a measurable space, with \( \sigma \)-algebra \( \mathcal{G} \) and, rather than just taking on fixed values \( \theta \) as in the frequentist case, the parameter is represented by a random variable \( \vartheta \) taking values in \( \Theta \). We assume that on the product-space \( \mathcal{Y} \times \Theta \) (with product \( \sigma \)-algebra \( \mathcal{F} = \sigma(\mathcal{B} \times \mathcal{G}) \)) we have a probability measure

\[
\Pi : \mathcal{F} \to [0,1],
\]

which is not a product measure. The probability measure \( \Pi \) provides a joint probability distribution for \( (Y, \vartheta) \), where \( Y \) is the observation and \( \vartheta \) (the random variable associated with) the parameter of the model.

Implicitly the choice for the measure \( \Pi \) defines the model in Bayesian context, by the possibility to condition the distribution of \( Y \) on fixed values \( \vartheta = \theta \) in \( \Theta \). The conditional distribution for \( Y|\vartheta \) describes the distribution of the observation \( Y \) given the parameter \( \vartheta \).

(For a discussion of conditional probabilities, see appendix A.5). As such, it can be identified with the elements \( P\theta \) of the model \( \mathcal{P} = \{P\theta : \theta \in \Theta \} \).

**Definition 2.1.** The distribution of the data \( Y \) conditional on the parameter \( \vartheta \) (c.f. definition A.34) is a map,

\[
\Pi_{Y|\vartheta} : \mathcal{B} \times \Theta \to [0,1],
\]

called (a version of) the model distributions.
Bayes’s rule, prior and posterior

(Note that potentially, this definition leaves room for problems regarding the regularity of the conditional distribution (see definition A.35). Whenever this subtlety is relevant, it is assumed that $\Theta$ is a Polish space (see definition A.36) which guarantees the existence of regular versions of conditional probabilities (see theorem A.37).) Since conditional probabilities are defined almost-surely with respect to the marginal (see definition A.34), the Bayesian notion of a model is represented only up to null-sets of the marginal distribution of $\vartheta$, referred to in Bayesian context as the prior for the parameter $\vartheta$.

**Definition 2.2.** The marginal distribution $\Pi : \mathcal{G} \to [0,1]$ for $\vartheta$ is called the prior.

The prior is interpreted in the subjectivist’s philosophy as the “degree of belief” attached to subsets of the model a priori, that is, before any observation has been made or incorporated in the calculation. Central to the Bayesian framework is the conditional distribution for $\vartheta$ given $Y$, called the posterior. The transition from prior to posterior represents the way in which “prior belief” is turned into “posterior belief” (concerning the parameter) based on the data. The posterior is interpreted as a data-amended version of the prior, that is to say, the subjectivist’s original “degree of belief” corrected by observation of $Y$ through conditioning. Below we define the posterior in conjunction with the marginal distribution for the data, the so-called prior predictive distribution.

**Definition 2.3.** The distribution $\Pi_{\vartheta|Y} : \mathcal{G} \times \mathcal{Y} \to [0,1]$ for $\vartheta|Y$ is called the posterior distribution.

The definition of the posterior is almost-sure with respect to the marginal data distribution $P^\Pi$ (see definition A.33 and the concluding remarks of subsection A.5).

**Definition 2.4.** The marginal distribution $P^\Pi : \mathcal{B} \to [0,1]$ for $Y$ is called the prior predictive distribution. If, in the above, one replaces the prior by the posterior, the resulting distribution for $Y$ is referred to as the posterior predictive distribution.

In the subjectivist philosophy, the prior predictive distribution describes a subjectivist’s expectations concerning the observation $Y$ based only on model and prior, i.e. before involving the data or realizations thereof. Given model and prior, the prior predictive distribution is of mixture form.

**Lemma 2.5.** The prior predictive $P^\Pi$ can be expressed in terms of model distributions $P_\theta$ and the prior, taking the form,

$$P^\Pi(Y \in A) = \int_{\Theta} P_\theta(A) d\Pi(\theta), \quad (2.3)$$

for all $A \in \mathcal{B}$.

In most statistical settings model, model distributions and prior are choices: the statistician chooses $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ and $\Pi$, after which the measure $\Pi$ on $\mathcal{Y} \times \Theta$ of (2.1) arises as:

$$\Pi(B \times G) = \int_G P_\theta(B) \Pi(\theta),$$
Bayesian basics

(with slight abuse of notation). In that case one does not condition to obtain model distributions and regularity does not play a role.

The Bayesian symmetry between observation and parameter invites an identity expressing its essence. Bayes’s Rule relates model distributions, prior, posterior and prior predictive distribution through \( \Pi(\theta \in G|Y \in B) \Pi(Y \in B) = \Pi(Y \in B|\theta \in G) \Pi(\theta \in G) \) (see lemma A.32). The following theorem restates this fact in terms of the concepts we have introduced above, in a property which is sometimes referred to as a disintegration of the joint measure on model times sample space.

**Theorem 2.6.** Posterior, prior predictive, model distributions and prior are related through Bayes’s Rule,

\[
\int_B \Pi(G|Y = y) \, dP(y) = \int_G P_\theta(B) \, d\Pi(\theta),
\]  

(2.4)

for all \( B \in \mathcal{B} \) and \( G \in \mathcal{G} \).

**Proof** Equality (2.4) follows since both sides are equal to \( \Pi(B \times G) \), c.f. definition A.34. \( \square \)

Note that, given model and prior, property (2.4) characterizes the posterior, up to re-definition on null sets of the prior predictive distribution \( P_{\Pi} \). Consequently, we may turn this theorem around and use property (2.4) as the defining property of the posterior

**Definition 2.7.** Given model and prior, any map \( \pi : \mathcal{G} \times \mathcal{Y} \to [0,1] \) such that \( y \mapsto \pi(G,y) \) is measurable for all \( G \in \mathcal{G} \) and such that \( \pi \) satisfies,

\[
\int_B \pi(G,y) \, dP_{\Pi}(y) = \int_G P_\theta(B) \, d\Pi(\theta),
\]  

(2.5)

is called a version of the posterior.

Unfortunately property (2.5) does not imply that \( \pi \) is a regular conditional probability, so we are left with an equivalence in which property 2 of definition A.35 remains as a condition.

**Theorem 2.8.** A map \( \pi : \mathcal{G} \times \mathcal{Y} \to [0,1] \) is a regular version of the posterior iff \( \pi \) satisfies (2.5), \( y \mapsto \pi(G,y) \) is \( \mathcal{B} \)-measurable for all \( G \in \mathcal{G} \) and \( G \mapsto \pi(G,y) \) is a measure on \( \mathcal{G} \) for \( P_{\Pi} \)-almost-all \( y \in \mathcal{Y} \).

**Proof** The map \( \pi \) satisfies the three properties listed in definition A.35. \( \square \)

By contrast, the following construction does imply regularity. Assuming that the model \( \mathcal{P} \) is dominated, there exists an expression for the posterior in terms of model densities. Since most statistical models are defined as families of densities (e.g. Lebesgue-densities on \( \mathbb{R} \) or \( \mathbb{R}^n \)) this accessible form of the posterior is used often throughout this text (and in practice).

**Theorem 2.9.** Assume that the model \( \mathcal{P} = \{ P_\theta : \theta \in \Theta \} \) is well-specified and dominated by a \( \sigma \)-finite measure \( \mu \) on \( (\mathcal{Y}, \mathcal{B}) \) with densities \( p_\theta = dP_\theta / d\mu \). Then the posterior can be expressed as,

\[
\Pi(\theta \in G|Y) = \int_G p_\theta(Y) \, d\Pi(\theta) / \int_\Theta p_\theta(Y) \, d\Pi(\theta),
\]  

(2.6)

for all \( G \in \mathcal{G} \). This version of the posterior is regular.
Proof Since the model is dominated, the prior predictive distribution has a density with respect to $\mu$, because for every $B \in \mathcal{B}$,

$$P^\Pi(B) = \int_\Theta P_\theta(B) d\Pi(\theta) = \int_\Theta \int_B p_\theta(y) d\mu(y) d\Pi(\theta) = \int_B \left( \int_\Theta P_\theta(y) d\Pi(\theta) \right) d\mu(y).$$

in accordance with the Fubini and Radon-Nikodym theorems. The prior predictive density $p^\Pi : \mathcal{Y} \to \mathbb{R}$ is therefore equal to the denominator on the r.h.s. of (2.6). Let $B \in \mathcal{B}$ and $G \in \mathcal{G}$ be given. Substituting (2.6) into the l.h.s. of (2.4), we obtain,

$$\int_B \Pi(G|Y = y) dP^\Pi(y) = \int_B \left( \int_G p_\theta(Y) d\Pi(\theta) \right) dP^\Pi(y)$$

$$= \int_B \int_G p_\theta(y) d\Pi(\theta) d\mu(y) = \int_G P_\theta(B) d\Pi(\theta).$$

According to theorem 2.6, (2.6) is a version of the posterior and property 3 of definition A.35 is satisfied. Property 1 of definition A.35 follows from Fubini’s theorem (which guarantees measurability of the r.h.s. of (2.6)). Since $P^\Pi(p^\Pi > 0) = 1$, the denominator in (2.6) is non-zero $P^\Pi$-almost-surely and the posterior is well-defined (as a map $G \to [0,1]$), $P^\Pi$-almost-surely. In addition, for all $y$ such that $p^\Pi(y) > 0$ and any sequence $(G_n)$ of disjoint, $\mathcal{G}$-measurable sets,

$$\Pi(\theta \in \bigcup_{n \geq 1} G_n \mid Y = y) = C(y) \int_{\bigcup_n G_n} p_\theta(y) d\Pi(\theta) = C(y) \int \sum_{n \geq 1} 1_{\theta \in G_n} p_\theta(y) d\Pi(\theta)$$

$$= \sum_{n \geq 1} C(y) \int_{G_n} p_\theta(y) d\Pi(\theta) = \sum_{n \geq 1} \Pi(\theta \in G_n \mid Y = y),$$

by dominated convergence. We have established that on an event of $P^\Pi$-measure one, this version of the posterior is well-defined and $\sigma$-additive, so that also property 2 of definition A.35 holds. Conclude that (2.6) is a regular version of the posterior. \hfill \Box

In the rest of this text, we shall not concern ourselves with regularity of posteriors any more and not make explicit mention of it as a condition (although strictly speaking we should): in most cases, theorem 2.9 applies and regularity is automatic. However that theorem does not cover all cases: the following example should serve as a word of caution that not all models are dominated and regularity is not always automatically guaranteed.

Example 2.10. Suppose that the samplespace is $\mathbb{R}$ and the model $\mathcal{P}$ consists of all measures of the form (see example A.12):

$$P = \sum_{j=1}^m \alpha_j \delta_{x_j},$$

(2.7)

for some $m \geq 1$, with $\alpha_1, \ldots, \alpha_m$ satisfying $\alpha_j \geq 0$, $\sum_{j=1}^m \alpha_j = 1$ and $x_1, \ldots, x_m \in \mathbb{R}$. A suitable prior for this model exists: distributions drawn from the so-called Dirichlet process prior (see section 3.6) are of the form (2.7) with (prior) probability one. There is no $\sigma$-finite dominating measure for this model and hence the model can not be represented by a family of densities, c.f. definition 1.3. From the frequentist’s point of view, if the true distribution $P_0$
Bayesian basics

for the observation is also a convex combination of Dirac measures, distributions in the model are singular with respect to $P_0$ unless they happen to have support-points in common with $P_0$. Consequently definition (2.6) does not give sensible results in this case. We have to resort to the subjectivist definition 2.3 in order to make sense of the posterior distribution.

2.1.2 Bayes’s billiard

To many who have been introduced to statistics from the frequentist point of view, treating the parameter $\theta$ for the model as a random variable $\vartheta$ seems somewhat unnatural because the frequentist role for the parameter is entirely different from that of the data. The following example demonstrates that in certain situations the Bayesian point of view is not unnatural at all.

**Example 2.11.** In the posthumous publication of “An essay towards solving a problem in the doctrine of chances” in 1763 [4], Thomas Bayes included an example of a situation in which the above, subjectivist perspective arises quite naturally. It involves a number of red balls and one white ball placed on a table and has become known in the literature as Bayes’s billiard.

We consider the following experiment: unseen by the statistician, someone places $n$ red balls and one white ball on a billiard table of length 1. The statistician will be reported the number $K$ of red balls that is closer to the cushion than the white ball ($K$ plays the role of the data in this example) and is asked to give a distribution reflecting his beliefs concerning the position of the white ball $X$ ($X$ plays the role of the parameter) based on $K$. Calling the distance between the white ball and the bottom cushion of the table $X$ and the distances between the red balls and the bottom cushion $Y_i$, ($i = 1, \ldots, n$), it is known to the statistician that their joint distribution is:

$$ (X; Y_1, \ldots, Y_n) \sim U[0, 1]^{n+1}, \quad (2.8) $$

i.e. all balls are placed independently with uniform distribution. (this distribution gives rise both to the model (for $K$) and to the prior (for $X$)). His prior knowledge concerning $X$ (i.e. without knowing the observed value $K = k$) offers little information: the best that can be said is that $X \sim U[0,1]$, the marginal distribution of $X$, i.e. the prior. The question is how this distribution for $X$ changes when we incorporate the observation $K = k$, that is, when we use the observation to arrive at posterior beliefs.

Since for every $i$, $Y_i$ and $X$ are independent c.f. (2.8), we have,

$$ P(Y_i \leq X \mid X = x) = P(Y_i \leq x) = x, $$

for any $x \in [0, 1]$. So for each of the red balls, determining whether it lies closer to the cushion than the white ball amounts to a Bernoulli experiment with parameter $x$. Since in addition the positions $Y_1, \ldots, Y_n$ are independent, counting the number $K$ of red balls closer to the cushion than the white ball amounts to counting “successes” in a sequence of independent Bernoulli
experiments. We conclude that $K$ has a binomial distribution $\text{Bin}(n; x)$, i.e.

$$P(K = k|X = x) = \frac{n^k}{k!(n-k)!} x^k (1 - x)^{n-k}.$$  

It is possible to obtain the density for the distribution of $X$ conditional on $K = k$ from the above display using Bayes’s Rule:

$$p(x|K = k) = P(K = k|X = x) \frac{p(x)}{P(K = k)}, \quad (2.9)$$

but in order to use it, we need the two marginal densities $p(x)$ and $P(K = k)$ in the fraction. From (2.8) it is known that $p(x) = 1$ and $P(K = k)$ can be obtained by integrating,

$$P(K = k) = \int_0^1 P(K = k|X = x) p(x) \, dx.$$  

Substituting in (2.9), we find:

$$p(x|K = k) = \frac{P(K = k|X = x) p(x)}{\int_0^1 P(K = k|X = x) p(x) \, dx} = B(n, k) x^k (1 - x)^{n-k},$$

where $B(n, k)$ is a normalization factor. The $x$-dependence of the density in the above display reveals that $X|K = k$ is distributed according to a Beta-distribution, $\text{B}(k+1, n-k+1)$, so that the normalization factor $B(n, k)$ must equal $B(n, k) = \Gamma(n + 2) / \Gamma(k + 1) \Gamma(n - k + 1)$.

This provides the statistician with distributions reflecting his beliefs concerning the position of the white ball for all possible values $k$ for the observation $X$. Through conditioning on $K = k$, the prior distribution of $X$ is changed into the posterior for $X$: if a relatively small number of red balls is closer to the cushion than the white ball (i.e. in case $k$ is small compared to $n$), then the white ball is probably close to the cushion; if $k$ is relatively large, the white ball is probably far from the cushion (see figure 2.1). The illustration on the cover of these lecture notes is Bayes’s own reflection of his Billiard, complete with Beta-density drawn along the bottom.

### 2.1.3 The Bayesian view of the model

Based on the definitions of subsection 2.1.1 a remark is in order with regard to the notion of the *model* in Bayesian statistics: if, for a subset $P_1 \subset P$, the prior assigns mass zero, then for all practical purposes $P_1$ does not play a role since omission of $P_1$ from $P$ does not influence the posterior. As long as the model is parametric, i.e. $\Theta \subset \mathbb{R}^d$, we can always use priors that dominate the Lebesgue measure, ensuring that any $P_1$ of prior measure zero has Lebesgue measure zero in $\Theta$ and can therefore be thought of as negligibly small. However, in non-parametric models null-sets of the prior and posterior may be much larger than expected intuitively.

**Example 2.12.** Taking the above argument to the extreme, consider a normal location model $P = \{ \mathcal{N}(\theta, 1) : \theta \in \mathbb{R} \}$ with a prior $\Pi = \delta_{\theta_1}$ (see example A.12), for some $\theta_1 \in \Theta$, defined...
Figure 2.1 Posterior densities for the position $X$ of the white ball, given the number $k$ of red balls closer to the cushion of the billiard (out of a total of $n = 6$ red balls). For the lower values of $k$, the white ball is close to the cushion with high probability, since otherwise more red balls would probably lie closer to the cushion. This is reflected by the posterior density for $X|K = 1$, for example, by the fact that it concentrates much of its mass close to $x = 0$.

**on the Borel $\sigma$-algebra $\mathcal{B}$.** Then the model is dominated by the Lebesgue measure and the posterior takes the form:

$$
\Pi(\theta \in A | Y) = \int_A p_\theta(Y) \, d\Pi(\theta) / \int_\Theta p_\theta(Y) \, d\Pi(\theta) = \frac{p_{\theta_1}(Y)}{p_{\theta_1}(Y)} \Pi(A) = \Pi(A).
$$

for any $A \in \mathcal{B}$. In other words, the posterior equals the prior, concentrating all its mass in the point $\theta_1$. Even though we started out with a model that suggests estimation of location, effectively the model consists of only one point due to the degeneracy of the prior. In subjectivist terms, the prior belief is fully biased towards $\theta_1$, leaving no room for amendment by the data when we condition to obtain the posterior.

This example raises the question which part of the model proper $\mathcal{P}$ plays a role in the Bayesian approach. In that respect, it is helpful to make the following definition.

**Definition 2.13.** In addition to $(\Theta, \mathcal{G}, \Pi)$ being a probability space, let $(\Theta, \mathcal{T})$ be a topological space and assume that $\mathcal{G}$ contains the Borel $\sigma$-algebra $\mathcal{B}$ corresponding to the topology $\mathcal{T}$. The **support** $\text{supp}(\Pi)$ of the prior $\Pi$ is defined as the smallest closed set $S$ such that $\Pi(\Theta \setminus S) = 0$.

It is tempting to equate the support of a prior to the set described by the following intersection.

$$
S = \bigcap \{ G \in \mathcal{G} : G \text{ closed, } \Pi(\Theta \setminus G) = 0 \}. \tag{2.10}
$$
Perhaps somewhat surprisingly, the validity of this identification is hard to establish: for any \((\Theta, \mathcal{G})\) as in definition 2.13, \(S\) is measurable: \(S\) is an intersection of closed sets so \(S\) is closed. Since the Borel \(\sigma\)-algebra is generated by the open sets (see definition A.4 and the example that follows) \(S \in \mathcal{B} \subset \mathcal{G}\). To show that \(\Pi(\Theta \setminus S) = 0\) requires extra conditions on the space \(\Theta\); the following lemma covers a large class of models (for another example, see exercise 2.7).

**Lemma 2.14.** In addition to \((\Theta, \mathcal{G}, \Pi)\) being a probability space, let \((\Theta, \mathcal{T})\) be a topological space and assume that \(\mathcal{G}\) contains the Borel \(\sigma\)-algebra \(\mathcal{B}\) corresponding to the topology \(\mathcal{T}\). If \(\mathcal{T}\) has a countable basis, \(S = \text{supp}(\Pi)\).

**Proof** Since \(S\) is closed, \(\Theta \setminus S\) is open so it can be written as a union of basis elements for the topology \(\mathcal{T}\). By assumption, the basis is countable so this union is a countable union: there exists a sequence \((U_k)_{k \geq 1}\) of (open) basis sets such that \(\Theta \setminus S = \bigcup_{k \geq 1} U_k\). For each \(k \geq 1\) there exists a closed \(G \subset \Theta\) such that \(\Pi(\Theta \setminus G) = 0\) and \(U \subset \Theta \setminus G\). Hence \(\Pi(U_k) = 0\) and, by the \(\sigma\)-additivity of \(\Pi\),\(\Pi(\Theta \setminus S) \leq \sum_k \Pi(U_k) = 0\). \(\square\)

This implies, for example, that the support of \(\Pi\) is of the form (2.10) if \((\Pi, \mathcal{T})\) is a separable metrizable space (the interested reader should describe the countable basis for \(\mathcal{T}\) in that case).

**Example 2.15.** In example 2.12, the model \(\mathcal{P}\) consists of all normal distributions of the form \(N(\theta, 1), \theta \in \mathbb{R}\), but the support of the prior \(\text{supp}(\Pi)\) equals the singleton \(\{N(\theta_1, 1)\} \subset \mathcal{P}\).

Note that the support of the prior is defined based on a topology, the Borel \(\sigma\)-algebra of which must belong to the domain of the prior measure. In parametric models this assumption is rarely problematic but in non-parametric models, finding such a prior may be difficult and the support may be an ill-defined concept. Therefore we may choose to take a less precise but more generally applicable perspective: the model is viewed as the support of the prior \(\Pi\), but only up to \(\Pi\)-null-sets (c.f. the \(\Pi\)-almost-sure nature of the identification (2.2)). That means that we may add to or remove from the model at will, as long as we make sure that the changes have prior measure equal to zero: the model itself is a \(\Pi\)-almost-sure concept.

### 2.1.4 A frequentist’s view of the posterior

So far, we have not discussed the details of the data \(Y\), we have treated \(Y\) completely abstractly. In this section we consider, firstly, the relation between the frequentist distribution of \(Y\) (the “true” \(P_0\)) and the Bayesian distribution of \(Y\) (the marginal \(P^{\Pi}\)). Secondly, we consider samples of independent, repeated measurements of a random variable \(X\). We shall see that the Bayesian way to describe data and statistical experiments is in contradiction with the frequentist assumption. The section concludes with a sort of recipe listing the steps involved in the Bayesian analysis of a data set. These steps are illustrated with a very simple parametric example.
The derivation of the posterior in subsection 2.1.1 does not refer to the any “true, underlying distribution of the data” but it does involve a marginal distribution for \( Y \), the prior predictive distribution of definition 2.4. If one adopts the frequentist framework to analyze Bayesian tools like the posterior, a discrepancy arises since \( P_0 \) and \( P^\Pi \) are two distributions for the data \( Y \) that are not equal (for a striking instance of the discrepancy, see remark 2.16). To the frequentist, \( P^\Pi \) is a side-product of the Bayesian construction that has no realistic interpretation. There is, however, a clear technical issue: all definitions and derivations in subsection 2.1.1 are almost-sure with respect to the prior predictive distribution. To ensure that all of this continues to make sense after we adopt assumption (1.1) we require that \( P^\Pi \) dominates \( P_0 \):

\[
P_0 \ll P^\Pi.
\]  

In that case, null-sets of \( P^\Pi \) are also null-sets of \( P_0 \), so that all \( P^\Pi \)-almost-sure statements and definitions are also \( P_0 \)-almost-sure. In particular, expression (2.6) for the posterior in a dominated model is satisfies the regularity condition not only \( P^\Pi \)-but also \( P_0 \)-almost-surely, if we assume (2.11). Throughout the rest of these lecture notes we shall adopt the frequentist philosophy to analyse Bayesian tools, i.e. we assume (1.1) and (2.11).

In many experiments or observations, the data consists of a sample of \( n \) repeated, stochastically independent measurements of the same quantity (an \( i.i.d. \) sample). To accommodate this situation formally, we choose \( \mathcal{Y} \) equal to the \( n \)-fold product of a sample space \( (\mathcal{X}, \mathcal{A}) \) and the observation takes the form \( Y = (X_1, X_2, \ldots, X_n) \). In frequentist statistics we can incorporate any assumption concerning the data directly on the true distribution of \( Y \). In the case of \( i.i.d. \) data we change the notation slightly and use \( P_0 \) to denote the frequentist, true marginal distribution for any of the \( X_i \), so that \( Y \sim P_0^n \). The frequentist composes his model accordingly, that is, he describes a family of marginal distributions and models the sample through \( n \)-fold products.

By contrast, in Bayesian statistics the \( i.i.d. \) assumption can only be reflected through the model distributions: to the Bayesian, the \( i.i.d. \) assumption takes the form of conditional independence of the observations, given \( \vartheta = \theta \):

\[
\Pi_{Y|\vartheta}(X_1 \in A_1, \ldots, X_n \in A_n \mid \vartheta = \theta) = \prod_{i=1}^{n} \Pi_{Y_i|\vartheta}(X_i \in A_i \mid \vartheta = \theta) = \prod_{i=1}^{n} P_\theta(A_i),
\]

for all \((A_1, \ldots, A_n) \in \mathcal{A}^n\). Similarly we see that the prior predictive distribution for \( i.i.d. \) data takes the form:

\[
P^\Pi_n(X_1 \in A_1, \ldots, X_n \in A_n) = \int_{\Theta} \prod_{i=1}^{n} P_\theta(A_i) \, d\Pi(\theta).
\]

The posterior is now a solution for Bayes’s Rule in the following form,

\[
\int_{A} \Pi(G|X_1 = x_1, \ldots, X_n = x_n) \, dP^\Pi_n(x_1, \ldots, x_n) = \int_{G} \prod_{i=1}^{n} P_\theta(A_i) \, d\Pi(\theta),
\]
Bayes’s rule, prior and posterior

where $A = A_1 \times \ldots \times A_n$, $G$ is any measurable model subset. Assuming that the model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ for the marginal distributions is dominated by a $\sigma$-finite measure $\mu$ on $\mathcal{X}$, the above can also be expressed in terms of $\mu$-densities $p_\theta = dP_\theta/d\mu$. Using theorem 2.9 we obtain the following expression for the posterior distribution:

$$
\Pi_n(\theta \in G | X_1, X_2, \ldots, X_n) = \frac{\int_G \prod_{i=1}^n p_\theta(X_i) \, d\Pi(\theta)}{\int_{\Theta} \prod_{i=1}^n p_\theta(X_i) \, d\Pi(\theta)}, \quad (2.12)
$$

where $G \in \mathcal{G}$ is any measurable subset of the model $\mathcal{P}$. We shall use this formula in most calculations of posteriors throughout the rest of this text. In a dominated model, the Radon-Nikodym derivative (see theorem A.26) of the posterior with respect to the prior is the likelihood function, normalized to be a probability density function:

$$
\frac{d\Pi(\cdot | X_1, \ldots, X_n)}{d\Pi}(\theta) = \frac{\prod_{i=1}^n p_\theta(X_i)}{\int_{\Theta} \prod_{i=1}^n p_\theta(X_i) \, d\Pi(\theta)}, \quad (2.13)
$$

$P_\Pi^n$-almost-surely, and under (2.11), also $P_0^n$-almost-surely. The latter fact explains why such strong relations exist between Bayesian and maximum-likelihood methods. Indeed, the proportionality of the posterior density and the likelihood provides a useful qualitative picture of the posterior as a measure that concentrates on regions in the model where the likelihood is relatively high. This may serve as a direct, Fisherian motivation for the use of Bayesian methods in a frequentist context, c.f. section 1.3.

Remark 2.16. Note that the prior predictive distribution for i.i.d. data is itself not a product distribution but a mixture of product distributions! This illustrates the discrepancy between $P_0$ and $P_\Pi$ quite clearly: while the true distribution of the data describes an i.i.d. random vector, the prior predictive distribution describes a random vector that is just exchangeable (in accordance with De Finetti’s theorem (see theorem A.2.2)).

Remark 2.17. For the frequentist to use Bayesian tools, e.g. a posterior calculated using (2.12), he has to assume condition (2.11). In the context of i.i.d. samples, that requirement takes the form,

$$
P_0^n \ll P_n, \quad (for \ all \ n \geq 1).
$$

To summarize, the Bayesian procedure consists of the following steps

(i) Based on the background of the data $Y$, the statistician chooses a model $\mathcal{P}$, usually with some measurable parameterization $\Theta \to \mathcal{P} : \theta \mapsto P_\theta$.

(ii) A prior measure $\Pi$ on $\mathcal{P}$ is chosen (reflecting “belief” c.f. subjectivism). Usually a measure on $\Theta$ is defined, inducing a measure on $\mathcal{P}$.

(iii) Based on definition 2.3, on expression (2.6) or in the case of an i.i.d. sample, on (2.12), we calculate the posterior as a function of the data $Y$. 

Bayesian basics

(iv) We observe a realization of the data \( Y = y \) and use it to calculate a realisation of the posterior.

The statistician may then infer properties of the parameter \( \theta \) from the posterior \( \Pi(\cdot | Y = y) \), giving them a subjectivist or objectivist interpretation. One important point: when reporting the results of any statistical procedure, one is obliged to also reveal all relevant details concerning the procedure followed and the data. So in addition to inference on \( \theta \), the statistician should report on the nature and size of the sample used and, in the Bayesian case, should always report choice of model and prior as well, with a clear motivation.

Example 2.18. To illustrate the above "recipe" with a concrete example, consider the one-dimensional parametric model \( \mathcal{P} \) consisting of exponential distributions:

\[
\mathcal{P} = \{ \text{Exp}(\theta) : \theta \in \Theta = (0, \infty) \}.
\]

Lebesgue measure dominates the model and densities take the form \( p_\theta(x) = \theta \exp(-\theta x) \), for \( x \geq 0 \). Assume that the data consists of \( n \) observations, (conditionally) independent and identically distributed. As a prior on the model, we take another exponential distribution with density \( \pi(\theta) = \exp(-\theta) \) (for \( \theta \in \Theta \)). The posterior density relative to Lebesgue measure on \( \Theta \)
takes the form,

\[
d\Pi(\theta | X_1, \ldots, X_n) = C(X_1, \ldots, X_n) \left( \prod_{i=1}^{n} \theta e^{-\theta X_i} 1_{\{X_i \geq 0\}} \right) e^{-\theta} d\theta
\]

where \( C(X_1, \ldots, X_n) \) denotes the (data-dependent) normalization factor that makes the posterior a probability measure. We calculate,

\[
d\Pi(\theta | X_1, \ldots, X_n) = C(X_1, \ldots, X_n) \theta^{n} e^{-\theta(1+\sum_i X_i)} 1_{\{X_{(1)} \geq 0\}} d\theta
\]

(where \( X_{(1)} = \min_i X_i \)). Since,

\[
\int_{0}^{\infty} \theta^n e^{-\alpha \theta} d\theta = \frac{n!}{\alpha^{n+1}},
\]

we see that \( C(X_1, \ldots, X_n) \) must be equal to \( (1 + \sum_i X_i)^{n+1}/n! \). So for any measurable \( A \subset \Theta \), the posterior probability is given by:

\[
\Pi(\theta \in A | X_1, \ldots, X_n) = \frac{1}{n!} \left( 1 + \sum_{i=1}^{n} X_i \right)^{n+1} 1_{\{X_{(1)} \geq 0\}} \int_{A} \theta^n e^{-\theta(1+\sum_{i=1}^{n} X_i)} d\theta.
\]

Note that the posterior density collapses to zero (and no longer describes a probability distribution!) if \( X_i < 0 \) for some \( 1 \leq i \leq n \). As Bayesians, we insist that the data must be compatible with the model in this sense. As frequentists, we involve an underlying distribution \( P_0 \) for a single observation \( X \) and require that \( P_0(X \geq 0) = 1 \), so that the posterior is well-defined \( P_0^n \)-almost-surely.
2.2 Bayesian point estimators

When considering questions of statistical estimation, the outcome of a frequentist procedure is of a different nature than the outcome of a Bayesian procedure: a point-estimator (the frequentist outcome) is a point in the model, whereas the posterior is a distribution on the model. A first question, then, concerns the manner in which to compare the two. The connection between Bayesian procedures and frequentist (point-)estimation methods is provided by point-estimators derived from the posterior. Needless to say, comparison of frequentist and Bayesian point-estimators requires that we assume the frequentist philosophy to analyze Bayesian methodology, c.f. subsection 2.1.4.

2.2.1 Posterior predictive and posterior mean

We think of a Bayesian point-estimator as a point in the model around which posterior mass is accumulated most, a point around which the posterior distribution is concentrated in some way. As such, any reasonable Bayesian point-estimator should represent the “location” of the posterior distribution. But as is well known from probability theory, there is no unique definition for the “location” of a distribution. Accordingly, there are many different ways to define Bayesian point-estimators.

Remark 2.19. Arguably, there are distributions for which even the existence of a “location” is questionable. For instance, consider the convex combination of point-masses $P = \frac{1}{2} \delta_{-1} + \frac{1}{2} \delta_{1}$ on $(\mathbb{R}, \mathcal{B})$. Reasonable definitions of location, like the mean and the median of $P$, all assign as the location of $P$ the point $0 \in \mathbb{R}$. Yet small neighbourhoods of $0$ do not receive any $P$-mass, so $0$ can hardly be viewed as a point around which $P$ concentrates its mass. The problem is not of a mathematical nature, it is conceptual: when we think of the “location” of a distribution we normally think of unimodal distributions which have unambiguous “locations”. However, it is common practice to formulate the notion for all distributions by the same definitions.

One quantity that is often used to represent a distribution’s location is its expectation. This motivates the first definition of a Bayesian point-estimator: the posterior mean.

Definition 2.20. Consider a statistical problem involving data $Y$ taking values in a sample-space $(\mathcal{Y}, \mathcal{B})$ and a model $(\mathcal{P}, \mathcal{G})$ with prior $\Pi$. Assume that the maps $P \mapsto P(B)$, $(B \in \mathcal{B})$ are measurable with respect to $\mathcal{G}$ and that the posterior $\Pi(\cdot | Y)$ is a regular conditional distribution. The posterior predictive distribution is a data-dependent map $\hat{P}$, defined by,

$$\hat{P}(B) = \int_{\mathcal{P}} P(B) \, d\Pi( P | Y ),\quad (2.14)$$

for every event $B \in \mathcal{B}$.

Lemma 2.21. The posterior predictive distribution $\hat{P} : \mathcal{B} \to [0,1]$ is a probability measure, almost surely.
Proof (Assuming that the posterior is a regular conditional distribution) \( \hat{P} \) is defined \( P^\Pi \)-almost-surely as a mapping \( \mathcal{B} \to [0, 1] \). Let \( F \in \mathcal{B} \) denote the event that \( \hat{P} \) is well-defined and let \( y \in F \) be given. Clearly, for all \( B \in \mathcal{B} \), \( 0 \leq \hat{P}(B) \leq 1 \). Let \( (B_i)_{i \geq 1} \subset \mathcal{B} \) be any sequence of disjoint events. Since \( (P, i) \mapsto P(B_i) \) is non-negative and measurable, Fubini’s theorem applies in the third equality below:

\[
\hat{P}\left(\bigcup_{i \geq 1} B_i\right) = \int_{\mathcal{B}} \hat{P}\left(\bigcup_{i \geq 1} B_i\right) d\Pi(P | Y = y) = \int_{\mathcal{B}} \sum_{i \geq 1} P(B_i) d\Pi(P | Y = y) = \sum_{i \geq 1} \hat{P}(B_i),
\]

which proves \( \sigma \)-additivity of \( \hat{P} \) for all \( y \in F \), that is, \( P^\Pi \)-almost-surely.

Remark 2.22. The qualification “almost surely” in the formulation of lemma 2.21 has distinct explanations for Bayesians and for frequentists: for the Bayesian, the data \( Y \) is marginally distributed according to the prior predictive distribution, so it is with respect to null sets of \( P^\Pi \) that “almost surely” is to be interpreted in that case. By contrast, the frequentist assumes that \( Y \sim P_0 \), so he is forced to adopt assumption (2.11) and the interpretation of “almost surely” refers to null sets of \( P_0 \) in that case. This dual purpose of the phrase “almost surely” re-occurs in many places below.

Although we refer to \( \hat{P} \) as a point-estimator (see definition 1.14), strictly speaking, we have not shown that \( \hat{P} \in \mathcal{P} \) and indeed, generically, \( \hat{P} \notin \mathcal{P} \) as the following lemma shows.

Lemma 2.23. Let \((\mathcal{Y}, \mathcal{B})\) be a sample space with model \( \mathcal{P} \). Endow \( \mathcal{P} \) with the total-variational norm as a metric, assume that the model is relatively compact in the corresponding metric topology and let \( \Pi \) be a Borel prior on \( \mathcal{P} \). Then \( \hat{P} \) lies in the closed convex hull of \( \mathcal{P} \), almost-surely.

Proof Let \( \epsilon > 0 \) be given. Since \( \mathcal{P} \) is relatively compact by assumption, there exists an \( N \geq 1 \) and \( \{P_1, \ldots, P_N\} \subset \mathcal{P} \) such that the total-variation balls \( B_1 = \{P' \in \mathcal{P} : \|P' - P_i\| < \epsilon\} \) form a finite cover of \( \mathcal{P} \), i.e. \( \mathcal{P} = \bigcup_{i=1}^{N} B_i \). Through definition of \( C_{i+1} = B_{i+1} \setminus C_i \) (for all \( i \geq 1 \), with \( C_1 = B_1 \)) this cover is turned into a finite partition \( \{C_1, \ldots, C_N\} \) of \( \mathcal{P} \). Define \( \lambda_i = \Pi(C_i | Y = y) \) and note that,

\[
\|\hat{P} - \sum_{i=1}^{N} \lambda_i P_i\| = \sup_{B \in \mathcal{B}} \left| \sum_{i=1}^{N} \int_{C_i} (P(B) - P_i(B)) d\Pi(P | Y = y) \right| \\
\leq \sup_{B \in \mathcal{B}} \sum_{i=1}^{N} \int_{C_i} |P(B) - P_i(B)| d\Pi(P | Y = y) \\
\leq \sum_{i=1}^{N} \int_{C_i} \sup_{B \in \mathcal{B}} |P(B) - P_i(B)| d\Pi(P | Y = y) \leq \epsilon.
\]

We have shown that there exist elements in the convex hull \( \text{co}(\mathcal{P}) \) of \( \mathcal{P} \) that are arbitrarily close to \( \hat{P} \) in total variation and conclude that \( \hat{P} \) lies in its closure \( \overline{\text{co}}(\mathcal{P}) \). \( \square \)
In many practical situations, the model $\mathcal{P}$ is parametric with parameterization $\Theta \to \mathcal{P}: \theta \mapsto P_\theta$. In case $\Theta$ is convex, a different form of “averaging over the model” applies.

**Definition 2.24.** Let $\mathcal{P}$ be a model parameterized by a closed, convex subset $\Theta$ of $\mathbb{R}^d$. Let $\Pi$ be a Borel prior defined on $\Theta$ with a regular posterior $\Pi(\cdot | Y)$. If $\vartheta$ is integrable with respect to the posterior, the posterior mean is defined

$$\hat{\theta}_1(Y) = \int_\Theta \theta \, d\Pi(\theta | Y) \in \Theta,$$

almost-surely.

**Example 2.25.** In example 2.18 the posterior takes the form:

$$\Pi(\vartheta \in A | X_1, \ldots, X_n) = \frac{1}{n!} \left( 1 + \sum_{i=1}^n X_i \right)^{n+1} 1_{\{X_1(1) \geq 0\}} \int_A \theta^n e^{-\theta(1 + \sum_{i=1}^n X_i)} d\theta.$$ 

Assuming that $P_0(X \geq 0) = 1$, we omit the indicator for $X_1(1) \geq 0$ and write the posterior mean of definition 2.24 as follows:

$$\hat{\theta}_1(Y) = \int_\Theta \theta \, d\Pi(\theta | Y) = \frac{1}{n!} \left( 1 + \sum_{i=1}^n X_i \right)^{n+1} \int_0^\infty \theta^{n+1} e^{-\theta(1 + \sum_{i=1}^n X_i)} d\theta$$

$$= \frac{1}{n!} \left( 1 + \sum_{i=1}^n X_i \right)^{-1} \int_0^\infty \zeta^{n+1} e^{-\zeta} d\zeta = (n+1) \left( 1 + \sum_{i=1}^n X_i \right)^{-1},$$

where we have used that $\int_0^\infty \zeta^{n+1} e^{-\zeta} d\zeta = \Gamma(n+2) = (n+1)!$.

In definition 2.24 convexity of $\Theta$ is a condition (rather than an afterthought, as with definition 2.20): if $\Theta$ is not convex there is no guarantee that $\hat{\theta}_1 \in \Theta$, in which case $P_{\hat{\theta}_1}$ is not defined since $\hat{\theta}_1$ does not lie in the domain of the parameterization. It is tempting to think that there is no difference between the posterior predictive and the posterior mean if the model is parametric. This is not the case, however, as demonstrated by the following (counter)example.

**Example 2.26.** Consider a normal location model in two dimensions for an observation $Y$, where the location $\mu \in \mathbb{R}^2$ lies on the unit circle and the covariance $\Sigma$ is fixed and known:

$$\mathcal{P} = \{ P_\theta = N(\mu(\theta), \Sigma) : \mu(\theta) = (\cos \theta, \sin \theta), \theta \in [0, 2\pi) \}.$$ 

This is an identifiable, one-dimensional parametric model with closed convex parameterizing space $\Theta = [0, 2\pi)$. Assume that $\Xi$ is the uniform distribution on $\Theta$ ($\Xi$ plays the role of the posterior; it does not matter what shape the posterior really has, all we need is a counterexample). We define the corresponding measure $\Xi'$ on $\mathcal{P}$ by applying $\Xi$ to the pre-image of the parameterization. By rotational symmetry of $\Xi$ and Fubini’s theorem, the expectation of $Y$ under $\hat{P}$ is

$$\int_Y d\hat{P} = \int_{\mathcal{P}} PY \, d\Xi'(P) = \int_\Theta P_\theta Y \, d\Xi(\theta) = \frac{1}{2\pi} \int_0^{2\pi} \mu(\theta) \, d\theta = (0, 0).$$
Note that none of the distributions in $\mathcal{P}$ has the origin as its expectation. We can also calculate the expectation of $Y$ under $P_{\hat{\theta}}$ in this situation:

$$\hat{\theta}_1(Y) = \int_{\Theta} \theta \, d\Xi(\theta) = \frac{1}{2\pi} \int_0^{2\pi} \theta \, d\theta = \pi,$$

which leads to $P_{\hat{\theta}}Y = P_\pi Y = (-1, 0)$. Clearly, the posterior predictive distribution does not equal the model distribution corresponding to the posterior mean. In fact, the above illustrates that the posterior predictive distribution does not belong to the model necessarily: $\hat{P} \not\in \mathcal{P}$.

The fact that the expectations of $\hat{P}$ and $P_{\hat{\theta}}$ in example 2.26 differ makes it clear that $\hat{P} \neq P_{\hat{\theta}}$ unless special circumstances apply: if we consider a parameterization $\theta \mapsto P_\theta$ from a (closed, convex) parameterizing space $\Theta$ with posterior measure $\Xi(d\theta)$ onto a space of probability measures $\mathcal{P}$ (with induced posterior $\Pi(dP)$), it makes a difference whether we consider definition (2.14), or calculate $P_{\hat{\theta}}$.

2.2.2 Small-ball and formal Bayes estimators

Since there are multiple ways of defining the location of a distribution, there are multiple ways of obtaining point-estimators from the posterior distribution. A straightforward alternative to the posterior mean is given in the following definition which requires that the model is one-dimensional.

**Definition 2.27.** Let $\Theta$ be a subset of $\mathbb{R}$ with non-empty interior and let $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ be a parametric model with Borel prior $\Pi$ on $\Theta$ and posterior $\Pi(\cdot | Y)$. The posterior median is defined by,

$$\tilde{\theta}(Y) = \inf\{s \in \mathbb{R} : \Pi(\theta \leq s | Y) \geq 1/2\},$$

almost-surely.

Thus the posterior median represents the smallest value for $\theta$ such that the posterior mass to its left is greater than or equal to 1/2. This definition simplifies drastically in case the posterior has a continuous, (strictly) monotone distribution function: in that case the above definition reduces to the perhaps more familiar definition as the (unique) point $\tilde{\theta} \in \Theta$ where $\Pi(\theta \leq \tilde{\theta} | Y) = 1/2$. In some situations, the posterior median offers an advantage over the posterior mean since its definition does not depend on integrability requirements.

Another alternative is decision-theoretic in essence (see section 2.5), that is, one takes the perspective in which an assessment of loss is inherent. Suppose that we consider estimation in a metric model $(\mathcal{P}, d)$ and we quantify errors in estimation as follows: if the true distribution of the data is $P_0$ and we estimate that it is $P$, then we incur a loss (to be specified further by the context of the problem) that is a monotone increasing function $\ell : [0, \infty) \mapsto [0, \infty)$ of the distance $d(P_0, P)$. If we assume that the posterior concentrates its mass around $P_0$ (as well as possible) then the expected loss relative to the posterior should provide guidance and the following point-estimators make sense [66].
Definition 2.28. Let $\mathcal{P}$ be a model with metric $d : \mathcal{P} \times \mathcal{P} \to \mathbb{R}$ and a Borel prior $\Pi$ on $\mathcal{P}$. Let $\ell : [0, \infty) \to [0, \infty)$ be a convex loss-function such that $\ell(0) = 0$. The formal Bayes estimator is the minimizer of the function,

$$
\mathcal{P} \to \mathbb{R} : P \mapsto \int_{\mathcal{P}} \ell(d(P, Q)) \, d\Pi(Q \mid Y),
$$

over the model $\mathcal{P}$. Provided that such a minimizer exists and is unique, the formal Bayes estimator is defined almost-surely.

Note that the posterior expectation of $\ell(d(P, Q))$ is non-negative and may be infinite. Another useful point estimator that does not presuppose integrability conditions relative to the posterior is defined as follows.

Definition 2.29. Let the data $Y$ with model $\mathcal{P}$, metric $d$ and prior $\Pi$ be given. Suppose that the $\sigma$-algebra on which $\Pi$ is defined contains the Borel $\sigma$-algebra generated by the metric topology. For given $\epsilon > 0$, the small-ball estimator is defined to be the maximizer of the function

$$
P \mapsto \Pi(B_d(P, \epsilon) \mid Y), \quad (2.17)
$$

over the model, where $B_d(P, \epsilon)$ is the $d$-ball in $\mathcal{P}$ of radius $\epsilon$ centred on $P$. Provided that such a maximizer exists and is unique, it is defined almost-surely.

Remark 2.30. Similarly to definition 2.29, for a fixed value $p$ such that $1/2 < p < 1$, we may define a Bayesian point estimator as the centre point of the smallest $d$-ball with posterior mass greater than or equal to $p$ (if it exists and is unique).

Suitable conditions for the existence of small-ball estimators form the subject of exercise 2.12.

2.2.3 The Maximum-A-Posteriori estimator

If the posterior is dominated by a $\sigma$-finite measure $\mu$, the posterior density with respect to $\mu$ can be used as a basis for defining Bayesian point estimators.

Definition 2.31. Let $\mathcal{P}$ be a model with prior $\Pi$. Assume that the posterior is absolutely continuous with respect to a $\sigma$-finite measure $\mu$ on $\mathcal{P}$, with $\mu$-density $\theta \mapsto \pi(\theta \mid Y)$. The maximum-a-posteriori estimator (or MAP-estimator, or posterior mode) $\hat{\theta}_2$ for $\theta$ is defined as the point in the model where the posterior density takes on its maximal value:

$$
\pi(\hat{\theta}_2 \mid Y) = \sup_{\theta \in \Theta} \pi(\theta \mid Y). \quad (2.18)
$$

Provided that such a point exists and is unique, the MAP-estimator is defined almost-surely.

The MAP-estimator has a serious weak point: a different choice of dominating measure $\mu$ leads to a different MAP estimator! A MAP-estimator is therefore not fully specified
unless we indicate which dominating measure was used to define the posterior density. It is
with respect to this dominating measure that we define our estimator, so a motivation for
the dominating measure used is inherently necessary. Often the Lebesgue measure is used
without further comment, or objective measures (see section 3.2) are used. Another option is
to use the prior measure as the dominating measure, in which case the MAP estimator equals
the maximum-likelihood estimator.

Remark 2.32. There is an interesting connection between (Bayesian) MAP-estimation and
(frequentist) maximum-likelihood estimation. Referring to formula (2.12) we see that in an
i.i.d. experiment with parametric model, the MAP-estimator maximizes:

\[
\Theta \to \mathbb{R} : \theta \mapsto \prod_{i=1}^{n} p_{\theta}(X_i) \pi(\theta),
\]

where it is assumed that the model is dominated and that the prior has a density \(\pi\) with respect
to the Lebesgue measure \(\mu\). If the prior had been uniform, the last factor would have dropped
out and maximization of the posterior density is maximization of the likelihood. Therefore,
differences between ML and MAP estimators are entirely due to non-uniformity of the prior.
Subjectivist interpretation aside, prior non-uniformity has an interpretation in the frequentist
setting as well, through what is called penalized maximum likelihood estimation (see, for exam-
ple, van de Geer (2000) [39]): Bayes’s rule applied to the posterior density \(\pi_n(\theta|X_1,\ldots,X_n)\)
gives:

\[
\log \pi_n(\theta|X_1,\ldots,X_n) = \log \pi_n(X_1,\ldots,X_n|\theta) + \log \pi(\theta) + D(X_1,\ldots,X_n),
\]

where \(D\) is a (\(\theta\)-independent, but stochastic) normalization constant. The first term equals the
log-likelihood and the logarithm of the prior plays the role of a penalty term when maximizing
over \(\theta\). Hence, maximizing the posterior density over the model \(\Theta\) can be identified with
maximization of a penalized likelihood over \(\Theta\). So defining a penalized MLE \(\hat{\theta}_n\) with the
logarithm of the prior density \(\theta \mapsto \log \pi(\theta)\) in the role of the penalty, the MAP-estimator
coincides with \(\hat{\theta}_n\). The above offers a direct connection between Bayesian and frequentist
methods of point-estimation. As such, it provides an frequentist interpretation of the prior as
a penalty in the ML procedure.

All Bayesian point estimators defined above as maximizers or minimizers over the model
suffer from the usual existence and uniqueness issues associated with extrema. However, there
are straightforward methods to overcome such issues. We illustrate using the MAP-estimator.
Questions concerning the existence and uniqueness of MAP-estimators should be compared
to those of the existence and uniqueness of \(M\)-estimators in frequentist statistics. Although
it is hard to formulate conditions of a general nature to guarantee that the MAP-estimator
exists, often one can use the following lemma to guarantee existence.

Lemma 2.33. Consider a parameterized model \(\Theta \to \mathcal{P} : \theta \mapsto P_{\theta}\); If the parameter space \(\Theta\)
is compact and the posterior density \(\theta \mapsto \pi(\theta|Y)\) is upper-semi-continuous, then the MAP-
estimator exists almost surely.
To prove uniqueness one has to be aware of various possible problems among which, for instance, identifiability of the model (see section 1.1, in particular definition 1.7).

**Example 2.34.** Assuming that $P_0(X \geq 0) = 1$, the posterior density in example 2.18 has the form:

$$
\pi(\theta|X_1, \ldots, X_n) = \frac{1}{n!} \left(1 + \sum_{i=1}^{n} X_i\right)^{n+1} \theta^n e^{-\theta(1+\sum_{i=1}^{n} X_i)},
$$

$P_0^n$-almost-surely, where $\theta > 0$. Setting the $\theta$-derivative to zero, we find that the MAP estimator is given by:

$$\hat{\theta}_2(Y) = n \left(1 + \sum_{i=1}^{n} X_i\right)^{-1}.$$

Note the similarity with the posterior mean (2.16) in this problem: although theoretically, it is possible that the posterior mean and the MAP estimate differ substantially, in many (e.g. unimodal) cases the maximum of the posterior density lies in the bulk that determines the posterior mean as well, and MAP and posterior mean are close. Asymptotically the difference disappears entirely, in many cases.

### 2.3 Confidence sets and credible sets

Besides point-estimation, frequentist statistics has several other inferential techniques at its disposal. The two most prominent are the analysis of confidence intervals and the testing of statistical hypotheses. In the next section, we consider frequentist testing of hypotheses, in this section, we discuss frequentist confidence sets and their Bayesian counterparts, called credible sets.

#### 2.3.1 Frequentist confidence intervals

Assume that we have a model $\mathcal{P}$ parametrized by an identifiable parameter $\theta$ in a parameter set $\Theta$, assuming that the true distribution of the data $Y \sim P_0$ belongs to the model, that is, $P_0 = P_{\theta_0}$ for some $\theta_0 \in \Theta$. The inferential goal the confidence set is to use the data $Y$ to define an model subset $C$ that contains $\theta_0$ with a high probability. Here, “high” probability requires quantification in terms of a level $\alpha$, called the confidence level.

**Definition 2.35.** Let $\Theta \rightarrow \mathcal{P} : \theta \mapsto P_\theta$ be an identifiable model and assume that $Y \sim P_{\theta_0}$ for some $\theta_0 \in \Theta$. Choose a confidence level $\alpha \in (0, 1)$. Let $C_\alpha$ be subset of $\Theta$ dependent only on the data $Y$. Then $C_\alpha$ is a confidence set for $\theta$ of confidence level $\alpha$, if it solves the equation,

$$P_\theta(\theta \in C_\alpha) \geq 1 - \alpha, \quad (2.19)$$

for all $\theta \in \Theta$.

The condition that $C_\alpha$ be dependent on the data $Y$ is meant to express the requirement that $C_\alpha$ can be calculated once the data has been observed. The confidence set may also
Bayesian basics

depend on other quantities that are known to the statistician, so \( C_\alpha \) is a statistic (see definition 1.14). The dependence of \( C_\alpha \) on the data \( Y \) makes \( C_\alpha \) a random subset of the model. Compare this to point estimation, in which the (data-dependent) estimator is a random point in the model.

Clearly confidence sets are not unique and small confidence sets are more informative than large ones: if, for some confidence level \( \alpha \), we have two different procedures of finding confidence sets, leading to sets \( C_\alpha \) and \( D_\alpha \) of confidence level \( \alpha \) respectively, and \( C_\alpha \subset D_\alpha \), \( P_\theta \)-almost-surely for all \( \theta \), then \( C_\alpha \) is preferred over \( D_\alpha \).

Example 2.36. Let \( Y_n = (X_1, \ldots, X_n) \) be an i.i.d. sample from a normal distribution \( N(\mu_0, \sigma^2) \), with known variance \( \sigma^2 > 0 \) and unknown \( \mu_0 \in \mathbb{R} \). As is well-known, the sample average is normally distributed,

\[
\hat{\mu}_n = \frac{1}{n} \sum X_i \sim N(\mu_0, \sigma^2/n),
\]

with a variance \( \sigma_n^2 = \sigma^2/n \). We rephrase as follows:

\[
P_{\theta_0}(\frac{\frac{1}{n} \sum X_i - \mu_0}{\sigma_n} \leq x) = \Phi(x),
\]

for all \( x \in \mathbb{R} \), where \( \Phi \) denotes the distribution function of the standard normal distribution. Consequently,

\[
P_{\theta_0}(\frac{1}{n} \sum X_i - \frac{\sigma x}{\sqrt{n}} < \mu_0 \leq \frac{1}{n} \sum X_i + \frac{\sigma x}{\sqrt{n}}) = \Phi(x) - \Phi(-x).
\]

Fixing some confidence level \( \alpha > 0 \), we solve for \( x_{\alpha/2} \) in the equation \( \Phi(x_{\alpha/2}) - \Phi(-x_{\alpha/2}) = 1 - \alpha \) to arrive at the conclusion that the interval,

\[
C_\alpha = \left[ \frac{1}{n} \sum X_i - \frac{\sigma x_{\alpha/2}}{\sqrt{n}}, \frac{1}{n} \sum X_i + \frac{\sigma x_{\alpha/2}}{\sqrt{n}} \right]
\]

is a level-\( \alpha \) confidence set for the parameter \( \mu \).

Like in the example 2.36, confidence intervals for a parameter \( \theta \) are often derived from estimators for \( \theta \): in the case of example 2.36 the sample average estimates \( \mu \) and it is the distribution of the sample average around \( \mu \) that determines the confidence interval.

Definition 2.37. Let \( \Theta \rightarrow \mathcal{P} : \theta \mapsto P_\theta \) be an identifiable model and assume that \( Y \sim P_{\theta_0} \) for some \( \theta_0 \in \Theta \). Let \( \hat{\theta} : \mathcal{Y} \rightarrow \Theta \) be an estimator for \( \theta \). The distribution of \( \hat{\theta} \) over \( \Theta \) is called the sampling distribution of the estimator.

Note that the randomness in \( \hat{\theta} \) occurs because it is a function of the data, which is distributed according to \( P_{\theta_0} \) itself. Therefore the sampling distribution of \( \hat{\theta} \) depends on the parameter \( \theta_0 \); in fact, it is exactly this dependence that allows us to draw statistical conclusions. In example 2.36, the location of the sampling distribution is \( \mu_0 \), so it may be assumed that the sample average \( \hat{\mu}_n \) lies close to its expectation, the true mean \( \mu_0 \). Exactly how close
Confidence sets and credible sets

is quantified by the variance of the sampling distribution of $\hat{\mu}_n$. A confidence set is an inferential instrument to express how much uncertainty remains concerning the true value of a parameter after estimation.

In case we have a very large sample, the analysis is sometimes more conveniently done based on limits, in particular the weak limit in the central limit theorem. To that end, we also define sequences of confidence sets that reach the required confidence level only in the limit $n \to \infty$.

**Definition 2.38.** Assume that data and model are as in definition 2.35. Let a confidence level $\alpha \in (0, 1)$ be given. If, for all $n \geq 1$, $Y_n = (X_1, \ldots, X_n)$ is i.i.d.-$P_0$, a sequence $(C_{\alpha,n})$ of sets that depend on $Y_n$ and solves,

$$\lim \inf_{n \to \infty} P_{\theta_0}^n(\theta \in C_{\alpha,n}) \geq 1 - \alpha, \quad (2.21)$$

for all $\theta \in \Theta$ is called an asymptotic confidence set of level $\alpha$.

As it turns out, example 2.36 applies in an approximate form in all asymptotic cases where the central limit theorem applies, as the following example demonstrates.

**Example 2.39.** Let $\Theta \to \mathcal{P} : \theta \mapsto P_\theta$ be an identifiable model for measurements $X$ and assume that $Y = (X_1, \ldots, X_n)$ is an i.i.d. sample from a distribution $P_{\theta_0}$ for some $\theta_0 \in \Theta$. Suppose that the data $Y$ forms an i.i.d. sample from a distribution $P_\theta = P_{\theta_0}$ on $\mathbb{R}$ and that $P_\theta X = \theta$ for all $\theta \in \Theta$. Moreover, we assume that for some known constant $S > 0$, $\sigma^2(\theta) = \text{Var}_\theta X \leq S^2$, for all $\theta \in \Theta$. Consider the sample-average $\hat{\theta}_n = \mathbb{P}_n X$. According to the central limit theorem, estimators that are sample averages have sampling distributions that converge weakly to normal distributions. Choose a confidence level $\alpha \in (0, 1)$ and note that,

$$P_{\theta_0}^n \left( -\frac{\sigma(\theta_0) x_{\alpha/2}}{\sqrt{n}} < \hat{\theta}_n - \theta_0 \leq \frac{\sigma(\theta_0) x_{\alpha/2}}{\sqrt{n}} \right) \to 1 - \alpha, \quad (n \to \infty). \quad (2.22)$$

Define $C_n$ by

$$C_n = \left[ \hat{\theta}_n - \frac{S x_{\alpha/2}}{\sqrt{n}}, \hat{\theta}_n + \frac{S x_{\alpha/2}}{\sqrt{n}} \right].$$

Then $P_{\theta_0}^n(\theta_0 \in C_n) \geq 1 - \alpha$. Note that if we had not used $S$ but $\sigma(\theta_0)$ instead, the $\theta_0$-dependence of $\sigma(\theta_0)$ would violate the requirement that $C_n$ be a statistic: since the true value $\theta_0$ of $\theta$ is unknown, so is $\sigma(\theta_0)$. Substituting the (known) upper-bound $S$ for $\sigma(\theta_0)$ enlarges the $\sigma(\theta_0)$-interval that follows from (2.22) to its maximal extent, eliminating the $\theta_0$-dependence. In a practical situation one would not assume that there is some upper bound $S > 0$, but substitute $\sigma(\theta_0)$ by an estimator $\hat{\sigma}_n$ (a practice known as studentization, after the Student $t$-distribution one obtains upon plugging in $\hat{\sigma}$ for finite $n \geq 1$ with $X_i$ that are marginally normal; refer to the case of example 2.36 if $\sigma^2$ had not been known.) Since the asymptotics of the studentized version are equal to those of the version based on $\sigma(\theta_0)$, studentization does not change the conclusions we based on (2.22).
2.3.2 Bayesian credible sets

The Bayesian analogs of confidence sets are called credible sets and are derived from the posterior distribution. The rationale behind the definition of credible sets is exactly the same one that motivated confidence sets: we look for a subset $D$ of the model that is as small as possible while receiving a certain minimal probability. Presently, however, the word “probability” is not based on the sampling distribution of an estimator, but in line with the Bayesian notion, that is, “probability” according to the posterior distribution.

**Definition 2.40.** Let $(\Theta, \mathcal{G})$ be a measurable space parameterizing an identifiable model $\Theta \rightarrow \mathcal{P} : \theta \mapsto P_\theta$ for data $Y$, with prior $\Pi : \mathcal{G} \rightarrow [0, 1]$. Choose a level $\alpha \in (0, 1)$. Let $D_\alpha \in \mathcal{G}$ be a subset of $\Theta$. Then $D_\alpha$ is a level-$\alpha$ credible set for $\vartheta$, if it solves the equation,

$$\Pi(\vartheta \in D_\alpha \mid Y) \geq 1 - \alpha,$$

almost-surely.

To find credible sets in examples one starts by calculating the posterior distribution from the prior and the data and, based on that, proceeds to derive a subset $D_\alpha$ such that (2.23) is satisfied. From a frequentist perspective, credible sets are statistics since they are defined based entirely on the posterior (which is a statistic itself). A credible set is sometimes referred to as a credible region, or, if $D$ is an interval in a one-dimensional parameterspace, a credible interval. In the following, frequentist definition we assume that the data forms an i.i.d. sample from a marginal distribution $P_0$.

**Definition 2.41.** If, for all $n \geq 1$, $Y_n = (X_1, \ldots, X_n)$ is i.i.d.-$P_0$, a sequence of model subsets $(D_{\alpha,n})$ that solves,

$$\liminf_{n \to \infty} \Pi(\vartheta \in D_{\alpha,n} \mid Y_n) \geq 1 - \alpha,$$

$P_0$-almost-surely, is called an asymptotic credible set of level $\alpha$.

**Remark 2.42.** In smooth, parametric models for i.i.d. data there is an close, asymptotic relation between Bayesian credible sets and frequentist confidence intervals centred on the maximum-likelihood estimator: the Bernstein-von Mises theorem implies that level-$\alpha$ credible sets coincide with abovementioned level-$\alpha$ confidence sets asymptotically! In situations where it is hard to calculate the ML estimator or to construct the corresponding confidence interval explicitly, it is sometimes relatively easy to obtain credible sets (based on a simulated sample from the posterior, as obtained from the MCMC procedure, for example). In such cases, one can calculate credible sets and conveniently interpret them as confidence intervals centred on the MLE, due to the Bernstein-von Mises theorem.

Definition 2.40 suffices to capture the concept of a credible set, but offers too much freedom in the choice of $D$: given a level $\alpha > 0$, many sets will satisfy (2.23), just like confidence sets can be chosen in many different ways. Note that, also here, we prefer smaller sets over
large ones: if, for some level $\alpha$, two different level-$\alpha$ credible sets $F_\alpha$ and $G_\alpha$ are given, both satisfying (2.23) and $F_\alpha \subset G_\alpha$ then $F_\alpha$ is preferred over $G_\alpha$. If the posterior is dominated with density $\theta \mapsto \pi(\theta|Y)$, we can be more specific. We define, for every $k \geq 0$, the level-set

$$D(k) = \{ \theta \in \Theta : \pi(\theta|Y) \geq k \},$$

and consider the following.

**Definition 2.43.** Let $(\Theta, \mathcal{F})$ a measurable space parameterizing a model $\Theta \to \mathcal{P} : \theta \mapsto P_\theta$ for data $Y \in \mathcal{Y}$, with prior $\Pi : \mathcal{F} \to [0,1]$. Assume that the posterior is dominated by a $\sigma$-finite measure $\mu$ on $(\Theta, \mathcal{F})$, with density $\pi(\cdot|Y) : \Theta \to \mathbb{R}$. Choose $\alpha \in (0,1)$. A level-$\alpha$ HPD-credible set (from highest posterior density) for $\vartheta$ is the subset $D_\alpha = D(k_\alpha)$, where $k_\alpha$ equals:

$$k_\alpha = \sup \{ k \geq 0 : \Pi(\vartheta \in D(k)|Y) \geq 1 - \alpha \}.$$

In other words, $D_\alpha$ is the smallest level-set of the posterior density that receives posterior mass greater than or equal to $1 - \alpha$. Note that HPD-credible sets depend on the choice of dominating measure: if we had chosen to use a different measure $\mu$, HPD-credible sets would have changed as well!

A default choice for the dominating measure for the posterior density would be the prior. In that case, the posterior density is proportional to the likelihood, c.f. (2.13), with a data-dependent normalization constant. Such a perspective is attractive to the frequentist who does not want the prior to play a role in the procedures. However, it is exactly opposite to what the subjectivist wants: a prior chosen according to subjectivist criteria places high mass in subsets of the model that the statistician attaches “high belief” to. Therefore the density of the posterior with respect to the prior can be expected to be relatively small in those subsets! As a result, sets of high prior density tend to fall outside HPD credible sets based on the posterior density relative to the prior. For that reason, the Lebesgue measure on $\Theta \subset \mathbb{R}^d$ is the default choice for the dominating measure.

### 2.4 Tests and Bayes factors

Having discussed confidence sets and credible sets in the previous section, we now turn to the related subject of hypothesis testing. We start with a discussion of the Neyman-Pearson framework and the famous lemma concerning the optimal test for testing one distribution versus another.

#### 2.4.1 Neyman-Pearson tests

Assume that we have data $Y$ and a parameterized model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ such that $Y \sim P_{\theta_0}$ for some $\theta_0 \in \Theta$. For simplicity, we assume that $\Theta \subset \mathbb{R}$ whenever the dimension of $\Theta$ is of importance. We start with the central ideas and definitions that play a role in the
Bayesian basics

Neyman-Pearson approach to statistical hypothesis testing. In this context, the hypotheses are presumptions one can make concerning the distribution of the data. Since the model contains all distributions the frequentist is willing to consider as candidates for $P_0$, the hypotheses are formulated in terms of a partition of the model (or its parameterization) into two disjoint subsets.

**Definition 2.44.** Testing of hypotheses proceeds through choice of a model subset $\Theta_0$ corresponding to the so-called null hypothesis $H_0$ and it’s complement $\Theta_1 = \Theta \setminus \Theta_0$, called the alternative hypothesis $H_1$. We distinguish between so-called simple hypotheses which consist of a single point in $\Theta$ and composite hypotheses which consist of bigger subsets in $\Theta$.

**Example 2.45.** As an example consider an i.i.d. real-valued sample $X_1, \ldots, X_n$ from $P_0$ and let the model be a location family: we take some distribution $Q$ and translate over $\theta \in \Theta \subset \mathbb{R}$, $P_{\theta}(A) = Q(A - \theta)$. for all $A$. Let a $\tilde{\theta} \in \Theta$ be a chosen point in the model and consider (simple) null-hypothesis $H_0$ and alternative $H_1$ concerning the true value $\theta_0$ of the parameter:

$$H_0 : \theta_0 = \tilde{\theta}, \quad H_1 : \theta_0 \in \Theta \setminus \{\tilde{\theta}\}, \quad (2.26)$$

The procedure now requires that we use the data to investigate the hypotheses further.

An immediate warning is in order because there is a potential for the reader to trip over the following conceptual confusion: the null and alternative hypotheses do not play symmetric roles in the Neyman-Pearson procedure: the goal of Neyman-Pearson hypothesis testing is to find out whether or not the data contains “enough” evidence to reject the null hypothesis as a likely explanation when compared to explanations offered by the alternative. By no means does Neyman-Pearson testing equate to the corresponding classification problem, in which one would treat $H_0$ and $H_1$ symmetrically and make a choice for one or the other based on the data. (In subsection 2.4.3 we consider a symmetric form of testing, and for more on frequentist and Bayesian classification, see section 2.5 on decision theory.)

For the hypotheses (2.26), the most-used version of the Neyman-Pearson method seeks to find a so-called test statistic $T(Y) \in \mathbb{R}^d$, displaying different behaviour depending on whether the data $Y$ is distributed according to distributions in $H_0$ or $H_1$.

**Example 2.46.** If, in the situation of example 2.45, we specify that $Q$ has a mean that equals 0, the hypotheses concern the mean of $P_0$, so it makes sense to use the sample average as our test statistic: if the mean of $P_0$ equals our chosen $\tilde{\theta}$, the sample average may be expected to lie close to $\tilde{\theta}$ with high probability, while a situation in which the mean of $P_0$ lies far away from $\tilde{\theta}$ produces sample averages at some distance from $\tilde{\theta}$ with high probability. Therefore the sample mean is a statistic that contains statistical information pertaining to the hypotheses.

The amount of uncertainty left in this information, depends on the sampling distribution of the sample mean.

To make this precise, one also chooses a so-called critical set $K \subset \mathbb{R}^d$ such that $P_0(T \in K)$ is “small” and $P_0(T \notin K)$ is “small” for all $\theta \in \Theta_1$. What we mean by “small” probabilities
in this context is a choice, the so-called significance level $\alpha \in (0, 1)$ associated with the test. The significance level sets the standard for which probabilities are deemed “small”. That way, upon realization $Y = y$, a distribution in the hypothesis $H_0$ makes an outcome $T(y) \in K$ improbable (and hence, makes $H_0$ implausible). This set-up is formalised in the following definition.

**Definition 2.47.** Let $\Theta \to \mathcal{P} : \theta \to P_\theta$ be an identifiable, parametrized model for a sample $Y$. Formulate two hypotheses $H_0$ and $H_1$ by introducing a two-set partition $\{\Theta_0, \Theta_1\}$ of the model $\Theta$:

$$
H_0 : \quad \theta_0 \in \Theta_0, \quad H_1 : \quad \theta_0 \in \Theta_1.
$$

We say that a test for these hypotheses based on a test-statistic $T$ with critical set $K$ is of level $\alpha \in (0, 1)$ if the power function $\pi : \Theta \to [0, 1]$, defined by

$$
\pi(\theta) = P_\theta(T(Y) \in K),
$$

is uniformly small over $\Theta_0$:

$$
\sup_{\theta \in \Theta_0} \pi(\theta) \leq \alpha. \tag{2.27}
$$

From the above definition we arrive at the conclusion that if $Y = y$ and $T(y) \in K$, hypothesis $H_0$ is improbable enough to be rejected, since $H_0$ forms an “unlikely” explanation of observed data (at said significance level). The degree of “unlikeness” can be quantified in terms of the so-called $p$-value, which is the lowest significance level at which the realised value of the test statistic $T(y)$ would have led us to reject $H_0$.

Of course there is the possibility that our decision is wrong and $H_0$ is actually true but $T(y) \in K$ nevertheless, so that our rejection of the null hypothesis is unwarranted. This is called a type-I error; a type-II error is made when we do not reject $H_0$ while $H_0$ is not true. The significance level $\alpha$ represents a fixed upper bound for the probability of a type-I error, c.f. (2.27). If we have several possible tests, all of which satisfy the Type-I error bound (2.27), the Neyman-Pearson approach calls for minimization of the Type-II error probability: of all the pairs $(T, K)$ satisfying (2.27), one prefers a pair that minimizes $P_\theta(T(Y) \notin K)$, (ideally) for all $\theta \in \Theta_1$. However, generically such uniformly most-powerful tests do not exist due to the possibility that different $(T, K)$ pairs satisfying (2.27) are most powerful over distinct subsets of the alternative.

We consider the Neyman-Pearson approach to testing in some more detail in the following example in the context of normally distributed data.

**Example 2.48.** Denote by $P_\mu$ the normal distribution $N(\mu, \sigma^2)$ with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$. Let $Y_n = (X_1, \ldots, X_n)$ be an i.i.d. sample from a normal distribution $P_0 = P_{\mu_0}$, assuming that the variance $\sigma^2 > 0$ is known, and unknown $\mu_0 \in \mathbb{R}$. As we have seen in (2.20), the sample average is normally distributed, $\hat{\mu}_n = \frac{1}{n}X \sim N(\mu_0, \sigma^2/n)$. For any $\mu$, we have,

$$
P_\mu^n\left(\frac{\sqrt{n}(\hat{\mu}_n - \mu)}{\sigma} \leq x\right) = \Phi(x),
$$
Bayesian basics

for all \( x \in \mathbb{R} \), where \( \Phi \) denotes the distribution function of the standard normal distribution. Re-write to obtain,

\[
P^\mu_n \left( \frac{x \sigma}{\sqrt{n}} < \hat{\mu}_n - \mu \leq \frac{x \sigma}{\sqrt{n}} \right) = \Phi(x) - \Phi(-x),
\]
for \( x > 0 \). By choosing some \( \tilde{\mu} \) and a significance level \( \alpha \), we formulate null- and alternative hypotheses as in (2.26),

\[
H_0 : \mu_0 = \tilde{\mu}, \quad H_1 : \mu_0 \neq \tilde{\mu},
\]
and note that under the null-hypothesis,

\[
P^\mu_n \left( \tilde{\mu} - \frac{x_{\alpha/2} \sigma}{\sqrt{n}} < \hat{\mu}_n \leq \tilde{\mu} + \frac{x_{\alpha/2} \sigma}{\sqrt{n}} \right) = 1 - \alpha,
\]
if we choose the quantiles \( x_{\alpha/2} \) like in example 2.36. Hence the null-hypothesis makes it improbable to observe \( |\hat{\mu}_n - \tilde{\mu}| > \frac{n^{-1/2}}{\alpha} x_{\alpha/2} \), which gives rise to the following definition of the critical set \( K_\alpha \)

\[
K_\alpha = \mathbb{R} \setminus \left[ \tilde{\mu} - \frac{x_{\alpha/2} \sigma}{\sqrt{n}}, \tilde{\mu} + \frac{x_{\alpha/2} \sigma}{\sqrt{n}} \right],
\]
which enables us to formulate our decision on the null hypothesis, that is, to reject \( H_0 \) or not:

(i) if the realised \( \hat{\mu}_n \in K_\alpha \), we reject \( H_0 \) at significance level \( \alpha \), and,

(ii) if the realised \( \hat{\mu}_n \notin K_\alpha \), we do not see enough evidence in the data to reject \( H_0 \) at significance level \( \alpha \).

Beware of a common pitfall regarding the interpretation: under case (ii), we do not draw the conclusion that \( H_0 \) is accepted. The data does not provide enough evidence to reject the null hypothesis but that does not imply that we accept it!

Note the behaviour of the procedure with varying sample-size: keeping the significance level fixed, the width of the critical sets \( K_\alpha \) is of order \( O(n^{-1/2}) \), so smaller and smaller critical sets can be used as more information concerning the distribution \( P_0 \) (read, data) comes available. Conversely, if we keep the critical set fixed, the probability for a Type-I error decreases (exponentially) with growing sample-size. Analogous to example 2.39, it is common practice to use asymptotic tests because sampling distributions are rarely available in closed form: in cases where the sample-size is deemed to be large enough, asymptotic approximations to sampling distributions an be used (e.g. the central limit theorem when estimators are based on sample averages).

2.4.2 Randomized tests and the Neyman-Pearson lemma

According to the Neyman-Pearson approach, tests are only considered if they satisfy (2.27) and are optimal if, in addition, they maximize testing power over the alternative uniformly. Optimality in this sense is sometimes not achievable, so one wonders if, mathematically, anything can be said at all. The Neyman-Pearson lemma answers this question in the affirmative
for problems where the null- and alternative hypotheses each contain a single distribution. In many models, (optimal) tests for composite hypotheses can be found.

To formulate the Neyman-Pearson lemma, however, we have to generalize the testing procedure slightly: as it turns out the existence of an optimal test can only be guaranteed if we allow for a randomization of our decision.

**Definition 2.49.** Let \( \mathcal{P} = \{P_\theta : \theta \in \Theta\} \) be a parametrized model for data \( Y \). Formulate two hypotheses \( H_0 \) and \( H_1 \) for \( \theta \) based on the two-set partition \( \{\Theta_0, \Theta_1\} \) of the model \( \mathcal{P} \):

\[
H_0 : \quad \theta_0 \in \Theta_0, \quad H_1 : \quad \theta_0 \in \Theta_1.
\]

A test function \( \phi \) is a map \( \phi : \mathcal{Y} \to [0, 1] \) used to formulate the following procedure called a randomized test: given a realization \( Y = y \) we reject \( H_0 \) with probability \( \phi(y) \). The power function associated with the test function \( \phi \) is given by \( \pi : \Theta \to [0, 1] : \theta \mapsto P_\theta \phi \).

Note that if we use the test function \( \phi(Y) = 1\{T(Y) \in K\} \), the randomized test reduces to the original (non-random) procedure of rejecting \( H_0 \) if \( T(Y) \in K \). Clearly the probability for Type-I error when using the randomized procedure equals \( \pi(\theta) \) (for \( \theta \in \Theta_0 \)) and the probability for Type-II error equals \( 1 - \pi(\theta) \) (for \( \theta \in \Theta_1 \)). Now return to the question regarding optimal Neyman-Pearson tests: we saw that the existence of an optimal test cannot be guaranteed as long as we look at composite hypotheses. However, if both null and alternative hypothesis are simple, no such ambiguity is possible and a (randomized) optimal test exists by the famed Neyman-Pearson lemma [62].

**Lemma 2.50.** Suppose the model is \( \mathcal{P} = \{P_{\theta_0}, P_{\theta_1}\} \) and write \( p_{\theta_0} : \mathcal{Y} \to \mathbb{R} \) and \( p_{\theta_1} : \mathcal{Y} \to \mathbb{R} \) for the densities of \( P_{\theta_0} \) and \( P_{\theta_1} \) relative to some \( \sigma \)-finite measure \( \mu \). Choose a significance level \( \alpha \in (0, 1) \) and consider a test of the form,

\[
\phi(y) = \begin{cases} 
1 & \text{if } p_{\theta_1}(y) > c p_{\theta_0}(y) \\
\gamma(x) & \text{if } p_{\theta_1}(y) = c p_{\theta_0}(y) \\
0 & \text{if } p_{\theta_1}(y) < c p_{\theta_0}(y) 
\end{cases}, \quad (2.28)
\]

where the measurable function \( \gamma : \mathcal{Y} \to [0, 1] \) and the constant \( c \in [0, \infty] \) form a solution to the equation:

\[
P_{\theta_0} \phi = \alpha.
\]

The following two assertions concern the hypotheses,

\[
H_0 : \quad \theta = \theta_0, \quad H_1 : \quad \theta = \theta_1.
\]

(i.) If a test of the form (2.28) has significance level \( \alpha \) then it is most powerful among all tests of level \( \alpha \).

(ii.) If a test \( \phi' \) is most powerful, then \( \phi' \) is of the form (2.28) for some \( \gamma(x) \) and some \( c \), almost-surely with respect to both \( P_{\theta_0} \) and \( P_{\theta_1} \).
Proof A proof can be found in Lehmann and Cassela (2005) [62]. □

The lemma is often used in conjunction with some condition on the model (or its likelihood function) to extend this point-vs-point version to composite hypotheses which are more interesting from a practical point of view.

Example 2.51. Suppose that we consider a random variable $X$ drawn from a normal distribution $N(\theta, 1)$ where $\theta \in \Theta = \{-1, 1\}$. Fixing a significance level $\alpha \in (0, 1)$, we consider a test of the form (2.28) for the hypotheses,

$$H_0 : \theta = -1, \quad H_1 : \theta = +1.$$  

(2.29)

A simple calculation shows that $p_{+1}(X)/p_{-1}(X) = e^{2X}$, so, with $\Phi$ denoting the distribution function for the standard normal distribution,

$$P_{-1}\phi(X) = P_{-1}(p_{+1}(X) > c p_{-1}(X)) = P_{-1}(e^{2X} > c)$$

$$= P_{-1}(X > \frac{1}{2} \log c) = 1 - \Phi(\frac{1}{2} \log c + 1),$$

where we have used that $X$ is distributed continuously (so that the middle term in (2.28) does not play a role and any $\gamma$ will do), and that $P_{-1}(X \leq x) = \Phi(x + 1)$. So to find $c$, we solve $1 - \Phi(\frac{1}{2} \log c + 1) = \alpha$, so that the Neyman-Pearson procedure for testing the hypotheses (2.29) has the form,

(i) if $X > x_{1-\alpha} - 1$, we reject $H_0$, and,

(ii) if $X \leq x_{1-\alpha} - 1$, we do not see enough evidence in the data to reject $H_0$, at significance level $\alpha$.

2.4.3 Test sequences and minimax optimality

Two points remain, the first being an asymptotic perspective on testing: just like we often study limits of sequences of estimators and conditions for their optimality, we are interested also in sequences of tests.

Definition 2.52. Let $\mathcal{P}$ be a model for an i.i.d. sample $X_1, X_2 \ldots$ taking values in $\mathcal{X}$ and assume that the true distribution of the data lies in the model, $(X_1, \ldots, X_n) \sim P^*_n$ and $P_0 \in \mathcal{P}$. Formulate two hypotheses $H_0$ and $H_1$ by introducing a two-set partition $\{\Theta_0, \Theta_1\}$ of the parameter space $\Theta$:

$H_0 : \theta_0 \in \Theta_0, \quad H_1 : \theta_0 \in \Theta_1.$

A test sequence $(\phi_n)_{n \geq 1}$ is a sequence of statistics $\phi_n : \mathcal{X}^n \rightarrow [0, 1]$, (for all $n \geq 1$). For every $n \geq 1$, we reject $H_0$ with probability $\phi_n(X_1, \ldots, X_n)$. The power function sequence of the test sequence $(\phi_n)$ as a map $\pi_n : \Theta \rightarrow [0, 1]$ on the model is defined by:

$$\pi_n(\theta) = P^*_{\theta} \phi_n.$$
Like in definition 2.47, the quality of the test sequence depends on the behaviour of the power sequence on $\Theta_0$ and $\Theta_1$ respectively. If we are interested exclusively in rejection of the null hypothesis, we could reason like in definition 2.47 and set a significance level $\alpha$ to select only those test sequences that satisfy $\sup_{\theta \in \Theta_0} P_\theta^n(P_\theta) \leq \alpha$. Subsequently, we prefer test sequences that have high power on the alternative in the limit $n \to \infty$.

**Definition 2.53.** If we have two test sequences ($\phi_n$) and ($\psi_n$) and a point $\theta \in \Theta_1$ such that

$$\lim_{n \to \infty} P_\theta^n \phi_n \geq \lim_{n \to \infty} P_\theta^n \psi_n,$$

then ($\phi_n$) is said to be asymptotically more powerful than ($\psi_n$) at $\theta$. If (2.30) holds for all $\theta \in \Theta_1$, the test sequence ($\phi_n$) is said to be uniformly asymptotically more powerful than ($\psi_n$). If one can show that this holds for all test sequences ($\psi_n$), then ($\phi_n$) is said to be uniformly asymptotically most powerful.

Note, however, that the above ordering of test sequences is not complete: it is quite possible that ($\phi_n$) is asymptotically more powerful than ($\psi_n$) on a subset of $\Theta_1$, whereas on its complement in $\Theta_1$, ($\psi_n$) is asymptotically more powerful. As a result, the existence of uniformly asymptotically most powerful test sequences can be problematic.

The second point that remains is to stress that the Neyman-Pearson criterion for tests is entirely geared at minimizing the probability of Type-I error (rejecting $H_0$ when in fact $H_0$ contains the true distribution). As a result the procedure can only lead to one definite conclusion, rejection of the null hypothesis. More natural is a framework that treats null and alternative hypotheses symmetrically: we look for tests or test sequences that indicate whether the true value of the parameter resides in $\Theta_0$ or in $\Theta_1$. As a further generalization, we let $\Theta_0$ and $\Theta_1$ be arbitrary disjoint model subsets, rather than requiring that $\Theta_0$ and $\Theta_1$ partition the model. To formulate things in the asymptotic context immediately, we consider test sequences ($\phi_n$) such that the quantity,

$$\sup_{\theta \in \Theta_0} P_\theta^n \phi_n + \sup_{\theta \in \Theta_1} P_\theta^n (1 - \phi_n),$$

is “small” in the limit $n \to \infty$, possibly quantified by introduction of a significance level pertaining to both type-I and type-II errors simultaneously. Tests of this sort are used extensively in the formulation of Bayesian limit theorems.

Optimality of test sequences obtains as follows: we say that ($\phi_n$) is minimax optimal if,

$$\sup_{\theta \in \Theta_0} P_\theta^n \phi_n + \sup_{\theta \in \Theta_1} P_\theta^n (1 - \phi_n) = \inf_{\psi} \left( \sup_{\theta \in \Theta_0} P_\theta^n \psi + \sup_{\theta \in \Theta_1} P_\theta^n (1 - \psi) \right),$$

for every $n \geq 1$. The infimum is taken over all measurable functions $\psi : X^n \to [0,1]$. Such tests exist as a consequence of the so-called minimax theorem (see theorem 2.65 in the next section): under certain convexity, continuity and compactness conditions, we can guarantee that:

$$\inf_{\psi} \left( \sup_{\theta \in \Theta_0} P_\theta^n \psi + \sup_{\theta \in \Theta_1} P_\theta^n (1 - \psi) \right) = \sup_{\theta \in \Theta_0} \sup_{\theta' \in \Theta_1} \inf_{\psi} \left( P_\theta^n \psi + P_{\theta'}^n (1 - \psi) \right).$$
Bayesian basics

and, furthermore, that there exist tests \((\phi_n)\) satisfying (2.32). The difference between left and right sides in (2.33) is that for every \(n \geq 1\) and all pairs \((\theta, \theta') \in \Theta_0 \times \Theta_1\) the test \(\psi\) on the right-hand side can be chosen equal to the (optimal) test based on the likelihood ratio \(p_\theta/p_{\theta'}(X_1, \ldots, X_n)\). After taking suprema and upper bounding, we prove the existence of a so-called minimax test sequence with explicitly known upper bound to the testing power.

**Lemma 2.54.** Let \(B, V \subset \mathcal{P}\) be model subsets such that their convex hulls \(\to B\) and \(\to V\) are separated by non-zero Hellinger distance:

\[
H(\to B, \to V) = \inf_{P \in \to B} \inf_{Q \in \to V} H(P, Q) > 0.
\] (2.34)

Then, for every \(n \geq 1\), there exists a test function \(\phi_n\) such that,

\[
\sup_{P \in B} P^n \phi_n + \sup_{Q \in V} Q^n (1 - \phi_n) \leq e^{-\frac{1}{4}nH(B,V)^2}.
\]

This lemma plays a central role in Bayesian limit theorems. Note that it does not refer to any underlying distribution \(P_0\), it is a lemma that applies in Bayesian and frequentist approaches alike.

2.4.4 Bayes factors

Bayesian hypothesis testing is formulated in a far more straightforward fashion than frequentist methods based on the Neyman-Pearson approach. It treats null and alternative hypotheses symmetrically. The two hypotheses \(H_0\) and \(H_1\) correspond to a two-set partition \(\{\Theta_0, \Theta_1\}\) of the model \(\Theta\) and for each of the parts, we have both posterior and prior probabilities. Based on the proportions between those, we shall decide which hypothesis is the more likely one. We make the following definitions.

**Definition 2.55.** Let \((\Theta, \mathcal{G})\) a measurable space parameterizing a model \(\Theta \to \mathcal{P}: \theta \mapsto P_\theta\) for data \(Y \in \mathcal{Y}\), with prior \(\Pi: \mathcal{G} \to [0, 1]\). Let \(\{\Theta_0, \Theta_1\}\) be a measurable partition of \(\Theta\) such that \(\Pi(\Theta_0) > 0\) and \(\Pi(\Theta_1) > 0\). The prior and posterior odds ratios are defined by \(\Pi(\Theta_0)/\Pi(\Theta_1)\) and \(\Pi(\Theta_0|Y)/\Pi(\Theta_1|Y)\) respectively. The Bayes factor in favour of \(\Theta_0\) is defined to be

\[
B = \frac{\Pi(\Theta_0|Y)}{\Pi(\Theta_1|Y)} \frac{\Pi(\Theta_1)}{\Pi(\Theta_0)}.
\]

When doing Bayesian hypothesis testing, we have a choice of which ratio to use and that choice will correspond directly with a choice for subjectivist or objectivist philosophies. In the subjectivist’s view, the posterior odds ratio has a clear interpretation: if

\[
\Pi(\Theta_0|Y) \frac{\Pi(\Theta_1)}{\Pi(\Theta_0)} > 1,
\]

then the probability of \(\theta \in \Theta_0\) is greater than the probability of \(\theta \in \Theta_1\) and, hence, the subjectivist decides to adopt \(H_0\) rather than \(H_1\). If, on the other hand, the Bayes factor in favour of \(\Theta_0\) is smaller than 1, the subjectivist decides to adopt \(H_1\) rather than \(H_0\). (Note
the difference with Neyman-Pearson frequentists who may reject \( H_0 \) but not adopt it.) The objectivist would object to this practice, saying that the relative prior weights of \( \Theta_0 \) and \( \Theta_1 \) can introduce a heavy bias in favour of one or the other in this approach (upon which the subjectivist would answer that that is exactly what he had in mind). Therefore, the objectivist would prefer to use a criterion that is less dependent on the prior weights of \( \Theta_0 \) and \( \Theta_1 \). We look at a very simple example to illustrate the point.

**Example 2.56.** Let \( \Theta \) be a dominated model that consists of only two points, \( \theta_0 \) and \( \theta_1 \) and let \( \Theta_0 = \{ \theta_0 \} \), \( \Theta_1 = \{ \theta_1 \} \), corresponding to simple null and alternative hypotheses \( H_0 \), \( H_1 \). Denote the prior by \( \Pi \) and assume that both \( \Pi(\{\theta_0\}) > 0 \) and \( \Pi(\{\theta_1\}) > 0 \). By Bayes rule, the posterior weights of \( \Theta_0 \) and \( \Theta_1 \) are

\[
\Pi(\vartheta \in \Theta_i | Y) = \frac{p_{\theta_i}(Y)\Pi(\Theta_i)}{p_{\theta_0}(Y)\Pi(\Theta_0) + p_{\theta_1}(Y)\Pi(\Theta_1)},
\]

for \( i = 0, 1 \). Therefore, the posterior odds ratio takes the form:

\[
\frac{\Pi(\vartheta \in \Theta_0 | Y)}{\Pi(\vartheta \in \Theta_1 | Y)} = \frac{p_{\theta_0}(Y)\Pi(\Theta_0)}{p_{\theta_1}(Y)\Pi(\Theta_1)},
\]

and the Bayes factor equals the likelihood ratio:

\[
B = \frac{p_{\theta_0}(Y)}{p_{\theta_1}(Y)}.
\]

(Note the similarity with the Neyman-Pearson and optimal minimax tests for \( \theta_0 \) versus \( \theta_1 \), in that Bayes factors is defined in terms of the likelihood ratio \( p_{\theta_0}/p_{\theta_1}(Y) \).) We see that the Bayes factor does not depend on the prior weights assigned to \( \Theta_0 \) and \( \Theta_1 \) (in this simple example), but the posterior odds ratio does. Indeed, suppose we stack the prior odds heavily in favour of \( \Theta_0 \), by choosing \( \Pi(\Theta_0) = 1 - \epsilon \) and \( \Pi(\Theta_1) = \epsilon \) (for some small \( \epsilon > 0 \)). Even if the likelihood ratio \( p_{\theta_0}(Y)/p_{\theta_1}(Y) \) is much smaller than one (but greater than \( \epsilon/1 - \epsilon \)), the subjectivist’s criterion favours \( H_0 \). In that case, the data clearly advocates hypothesis \( H_1 \) but the prior odds force adoption of \( H_0 \). The Bayes factor \( B \) equals the likelihood ratio (in this example), so it does not suffer from the bias imposed on the posterior odds.

The objectivist prefers the Bayes factor to make a choice between two hypotheses: if \( B > 1 \) the objectivist adopts \( H_0 \) rather than \( H_1 \); if, on the other hand, \( B < 1 \), then the objectivist adopts \( H_1 \) rather than \( H_0 \). In example 2.56 the Bayes factor is independent of the choice of the prior. In general, the Bayes factor is not completely independent of the prior, but it does not depend on the relative prior weights of \( \Theta_0 \) and \( \Theta_1 \).

**Lemma 2.57.** Let \( (\Theta, \mathscr{F}) \) a measurable space parameterizing a model \( \Theta \to \mathscr{P} : \theta \to P_\theta \) for data \( Y \in \mathscr{Y} \), with prior \( \Pi : \mathscr{F} \to [0, 1] \). Let \( \{ \Theta_0, \Theta_1 \} \) be a partition of \( \Theta \) such that \( \Pi(\Theta_0) > 0 \) and \( \Pi(\Theta_1) > 0 \). Then the Bayes factor \( B \) in favour of \( \Theta_0 \) does not depend on the prior odds ratio.
Proof For any prior such that $\Pi(\Theta_0) > 0$ and $\Pi(\Theta_1) > 0$,

$$\Pi(A) = \Pi(A|\Theta_0) \Pi(\Theta_0) + \Pi(A|\Theta_1) \Pi(\Theta_1),$$

(2.35)

for all $A \in \mathcal{G}$. In other words, $\Pi$ is decomposed as a convex combination of two probability measures on $\Theta_0$ and $\Theta_1$ respectively. The Bayes factor is then rewritten (see lemma A.32):

$$B = \frac{\Pi(\Theta_0|Y) \Pi(\Theta_1)}{\Pi(\Theta_1|Y) \Pi(\Theta_0)} = \frac{\Pi(Y|\Theta_0)}{\Pi(Y|\Theta_1)},$$

where, in a dominated model,

$$\Pi(Y|\Theta_i) = \int_{\Theta_i} p_\theta(Y) \, d\Pi(\theta|\Theta_i),$$

for $i = 0, 1$. In terms of the decomposition (2.35), $B$ depends on $\Pi(\cdot|\Theta_0)$ and $\Pi(\cdot|\Theta_1)$, but not on $\Pi(\Theta_0)$ and $\Pi(\Theta_1)$.

□

So the difference between Bayes factors and posterior odds is exactly the bias introduced by non-zero prior odds; as such, it represents directly the difference between objectivist and subjectivist Bayesian philosophies.

Example 2.58. Consider a data vector $Y = (X_1, \ldots, X_n)$ modelled as an i.i.d. sample from a uniform distribution $U[\theta, \theta + 1]$, with $\theta \in \Theta = [-1, 1]$. We formulate hypotheses,

$$H_0 : \theta \geq 0, \quad H_1 : \theta < 0.$$

and, to show how prior odds influence posterior odds but not Bayes factors, we use a prior with a Lebesgue density of the form,

$$\pi(\theta) = \lambda \{\theta < 0\} + (1 - \lambda) \{\theta \geq 0\},$$

for some $0 < \lambda < 1$ (where we note that $\lambda = 0$ or $\lambda = 1$ would not be valid choices). Consequently, the prior odds for $\Theta_0$ versus $\Theta_1$ are $1 - 1/\lambda$. The likelihood is given by,

$$L_n(\theta; X_1, \ldots, X_n) = \prod_{i=1}^n \{\theta \leq X_i \leq \theta + 1\},$$

and the posterior density (relative to the Lebesgue measure on $\Theta = [-1, 1]$) is proportional to,

$$\pi(\theta|X_1, \ldots, X_n) \propto \lambda \{\theta < 0\} \{\theta \leq X_{(1)}\} \{X_{(n)} \leq \theta + 1\}
+ (1 - \lambda) \{\theta \geq 0\} \{\theta \leq X_{(1)}\} \{X_{(n)} \leq \theta + 1\},$$

where $X_{(1)}$ and $X_{(n)}$ denote first and last order statistics of the sample respectively. To calculate the posterior odds we do not need the normalization factor in the posterior and we see immediately that,

$$\frac{\Pi(\theta \geq 0 \mid X_1, \ldots, X_n)}{\Pi(\theta < 0 \mid X_1, \ldots, X_n)} = \frac{1 - \lambda}{\lambda} \int_0^1 1\{X_{(n)} - 1 \leq \theta \leq X_{(1)}\} \, d\theta$$

Note the proportionality to the prior odds: the Bayes factor $B$ is equal to only the latter fraction in the expression on the right-hand side of the above display and is insensitive to the subjective choice for $\lambda$. 
To conclude this section we make the following important remark.

**Remark 2.59.** The condition that both $\Theta_0$ and $\Theta_1$ receive prior mass strictly above zero is important since Bayes factors and odds ratios are based on conditioning of $\vartheta$. Bayesian hypothesis testing is sensible only if both $\Theta_0$ and $\Theta_1$ receive non-zero prior mass. This remark plays a role particularly when comparing a simple null hypothesis to an alternative, as illustrated in exercise 2.17.

### 2.5 Decision theory and classification

Many practical problems require that we make an observation and based on the outcome, make a decision of some kind. For instance when looking for the diagnosis for a patient, a doctor will observe variables like the patient's temperature, blood-pressure and appearance, in addition to the results of chemical and physical scans to come to a decision regarding the affliction the patient is probably suffering from. Another example concerns the financial markets, where past stock- and option-prices are considered by analysts to decide whether to buy or sell stocks and derivatives. In a chemical plant, regulation of a chemical process amounts to a succession of decisions to control and optimize conditions, based on the measurement of thermo-dynamical quantities such as concentrations of chemicals involved in the reaction. In this section, we look at problems of this nature, first from a frequentist perspective and then with the Bayesian approach.

#### 2.5.1 Decision theory

Practical problems like those described above usually involve optimality criteria that are prescribed by the context of the problem itself: for example, when a doctor makes the wrong diagnosis for a patient suffering from cancer the consequences can be most serious, whereas the misdiagnosis of a case of influenza is usually no more than unfortunate. In any useful statistical procedure meant to assist in medical diagnosis, such differences should be reflected in the decision-making procedure. That is certainly not the case for the methods that we have discussed thus far. Up to this point, we have used optimality criteria of a more general nature, like the accuracy of an estimation procedure, coverage probabilities for confidence intervals or the probability of Type-I and type-II errors in a testing procedure. The distinction lies in the nature of the optimality criteria: so far we have practiced what is called *statistical inference*, in which optimality is formulated entirely in terms of the stochastic description of the data. For that reason, it is sometimes said that statistical inference limits itself to those questions that “summarize the data”. By contrast, *statistical decision theory* formalizes the criteria for optimality by adopting the use of a so-called loss-function to quantify the consequences of wrong decisions in a way prescribed by the context of the statistical problem.
In statistical decision theory the nomenclature is slightly different from that introduced earlier. We consider a system that is in an unknown state $\theta \in \Theta$, where $\Theta$ is called the state-space. The observation $Y$ takes its values in the samplespace $\mathcal{Y}$, a measurable space with $\sigma$-algebra $\mathcal{B}$. The observation is stochastic, its distribution $P_0 : \mathcal{B} \to [0,1]$ being dependent on the state $\theta$ of the system. The observation does not reveal the state of the system completely or with certainty. Based on the outcome $Y = y$ of the observation, we take a decision $a \in \mathcal{A}$ (or perform an action $a$, as some prefer to say), where $\mathcal{A}$ is the called the decision-space. For each state $\theta$ of the system there may be an optimal or prescribed decision, but since observation of $Y$ does not give us the state $\theta$ of the system with certainty, the decision is stochastic and may be wrong. The goal of statistical decision theory is to arrive at a rule that decides in the best possible way given only the data $Y$.

The above does not add anything new to the approach we were already following: aside from the names, the concepts introduced here are those used in the usual problem of statistically estimating $a \in \mathcal{A}$. Decision theory distinguishes itself through its definition of optimality in terms of a so-called loss-function.

**Definition 2.60.** Any lower-bounded function $L : \Theta \times \mathcal{A} \to \mathbb{R}$ may serve as a loss-function. The utility-function is $-L : \Theta \times \mathcal{A} \to \mathbb{R}$.

(Although statisticians talk about loss-functions, people in applied fields often prefer to talk of utility-functions, which is why the above definition is given both in a positive and a negative version.) The interpretation of the loss-function is the following: if a particular decision $a$ is taken while the state of the system is $\theta$, then a loss $L(\theta, a)$ is incurred which can be either positive (loss) or negative (profit). To illustrate, in systems where observation of the state is direct (i.e. $Y = \theta$) and non-stochastic, the optimal decision $a(\theta)$ given the state $\theta$ is the value of $a$ that minimizes the loss $L(\theta, a)$. However, the problem we have set is more complicated because the state $\theta$ is unknown and can not be measured directly. All we have is the observation $Y$.

**Definition 2.61.** Let $\mathcal{A}$ be a measurable space with $\sigma$-algebra $\mathcal{H}$. A measurable $\delta : \mathcal{Y} \to \mathcal{A}$ is called a decision rule.

A decision-rule is an automated procedure to arrive at a decision $\delta(y)$, given that the observation is $Y = y$. We denote the collection of all decision rules under consideration by $\Delta$. Clearly our goal will be to find decision rules in $\Delta$ that “minimize the loss” in an appropriate sense. The above basic ingredients of decision-theoretic problems play a role in both the frequentist and Bayesian analysis. We consider the frequentist approach first and then look at decision theory from a Bayesian perspective.

In frequentist decision theory we assume that $Y \sim P_{\theta_0}$ for some state $\theta_0 \in \Theta$ and we analyze the expectation of the loss.
Definition 2.62. The risk-function $R : \Theta \times \Delta \rightarrow \mathbb{R}$ is defined as the expected loss under $Y \sim P_{\theta}$ when using $\delta$,

$$R(\theta, \delta) = \int L(\theta, \delta(Y)) \, dP_\theta.$$  

(2.36)

Of interest to the frequentist is only the expected loss under the true distribution $Y \sim P_{\theta_0}$. But since $\theta_0$ is unknown, we are forced to consider all values of $\theta$, i.e. look at the risk-function $\theta \mapsto R(\theta, \delta)$ for each decision rule $\delta$.

Definition 2.63. Let the state-space $\Theta$, states $P_\theta$, $(\theta \in \Theta)$, decision space $\mathcal{A}$ and loss $L$ be given. Choose $\delta_1, \delta_2 \in \Delta$. The decision rule $\delta_1$ is $R$-better than $\delta_2$, if

$$\forall \theta \in \Theta : R(\theta, \delta_1) < R(\theta, \delta_2).$$  

(2.37)

A decision rule $\delta$ is admissible if there exists no $\delta' \in \Delta$ that is $R$-better than $\delta$ (and inadmissible if such a $\delta'$ does exist).

It is clear that the definition of $R$-better decision-rules is intended to order decision rules: if the risk-function associated with a decision-rule is relatively small, then that decision rule is preferable. Note, however, that the ordering we impose by definition 2.63 may be partial rather than complete: pairs $\delta_1, \delta_2$ of decision rules may exist such that neither $\delta_1$ nor $\delta_2$ is $R$-better than the other. This is due to the fact that $\delta_1$ may perform better (in the sense that $R(\theta, \delta_1) \leq R(\theta, \delta_2)$) for values of $\theta$ in some $\Theta_1 \subset \Theta$, while $\delta_2$ performs better in $\Theta_2 = \Theta \setminus \Theta_1$, resulting in a situation where (2.37) is true for neither. For that reason, it is important to find a way to compare risks (and thereby decision rules) in a $\theta$-independent way and thus arrive at a complete ordering among decision rules. This motivates the following definition.

Definition 2.64. (Minimax decision principle) Let the state-space $\Theta$, states $P_\theta$, $(\theta \in \Theta)$, decision space $\mathcal{A}$ and loss $L$ be given. The function

$$\Delta \rightarrow \mathbb{R} : \delta \mapsto \sup_{\theta \in \Theta} R(\theta, \delta)$$

is called the minimax risk. Let $\delta_1, \delta_2 \in \Delta$ be given. The decision rule $\delta_1$ is minimax-preferred to $\delta_2$, if

$$\sup_{\theta \in \Theta} R(\theta, \delta_1) < \sup_{\theta \in \Theta} R(\theta, \delta_2).$$

If $\delta^M \in \Delta$ minimizes $\delta \mapsto \sup_{\theta} R(\theta, \delta)$ then $\delta^M$ is called a minimax decision-rule.

One of the foundations of statistical decision theory is the so-called minimax theorem which guarantees the existence of minimax decision rules under very general conditions.

Theorem 2.65. Assume that $\Delta$ and $\Theta$ are convex sets. Furthermore, assume that the map $\delta \mapsto R(\theta, \delta)$ is convex on $\Delta$ for every $\theta \in \Theta$ and that the map $\theta \mapsto R(\theta, \delta)$ is concave on $\Theta$ for every $\delta$. Finally, the topology on $\Delta$ is such that $\Delta$ is compact and $\delta \mapsto R(\theta, \delta)$ is continuous for all $\theta$. Then there exists a minimax decision rule $\delta^M$, i.e.

$$\sup_{\theta \in \Theta} R(\theta, \delta^M) = \inf_{\delta \in \Delta} \sup_{\theta \in \Theta} R(\theta, \delta) = \sup_{\theta \in \Theta} \inf_{\delta \in \Delta} R(\theta, \delta).$$  

(2.38)
Proof. See Strasser (1985) [85], p. 239.

Since many loss-functions used in practice satisfy the convexity requirements, the minimax theorem has broad applicability in statistical decision theory and many other fields, particularly econometrics.

Note that the minimax theorem holds only for convex $\Delta$. In other words, if we want to guarantee the existence of a minimax-optimal decision rule, we are forced to consider convex combinations of decision rules: as an example, suppose we entertain the possibility of using one of two different decision rules $\delta_1, \delta_2 : \mathcal{Y} \to \mathcal{A}$. After the data is realised as $Y = y$, $\delta_1$ and $\delta_2$ give rise to two decisions $\delta_1(y), \delta_2(y)$, which may differ. In that case, we flip a coin to decide which decision to use: the extra stochastic element introduced by the coin-flip has then randomized our decision rule. To formalize this and illustrate the connection with convexity, consider a measurable space $(\Omega, \mathcal{F})$ with data $Y : \Omega \to \mathcal{Y}$ and a decision rule $\delta : \Omega \to \text{co}(\mathcal{A})$.

The decision rule $\delta$ is a randomized decision rule whenever $\sigma(\delta)$ is not a subset of $\sigma(Y)$, i.e. $\delta$ is not just a function of $Y$. The product space $\mathcal{Y} \times \{0, 1\}$ endowed with the product $\sigma$-algebra may serve as the measurable space $(\Omega, \mathcal{F})$ with $\delta : \Omega \to \text{co}(\mathcal{A})$ defined by,

$$(y, c) \mapsto \delta(y, c) = c \delta_1(Y) + (1 - c) \delta_2(y),$$

for all $y \in \mathcal{Y}$ and $c \in \{0, 1\}$. Perhaps a bit counterintuitively (but certainly in accordance with the fact that minimization over a larger set produces a lower infimum), in some decision problems the minimax risk associated with such randomized decision rules lies strictly below the minimax risks of both non-randomized decision rules.

Example 2.66. (Decision theoretic $L_2$-estimation) The decision-theoretic approach can also be used to formulate estimation problems in a generalized way, if we choose the decision space $\mathcal{A}$ equal to the state-space $\Theta = \mathbb{R}$. Let $Y \sim N(\theta_0, 1)$ for some unknown $\theta_0 \in \Theta$. Choose $L : \Theta \times \Theta \to \mathbb{R}$ equal to the quadratic difference

$$L(\theta, a) = (\theta - a)^2,$$

a choice referred to as an $L_2$-loss (or squared-error loss). Consider the decision-space

$$\Delta = \{\delta_c : \mathcal{Y} \to \mathcal{A} : \delta_c(y) = cy, \ c \geq 0\}.$$

Note that $\Delta$ plays the role of a family of estimators for $\theta_0$ here. The risk-function takes the form:

$$R(\theta, \delta_c) = \int L(\theta, \delta_c(Y)) \, dP_\theta = \int_{\mathbb{R}} (\theta - cy)^2 \, dN(\theta, 1)(y)$$

$$= \int_{\mathbb{R}} (c(\theta - y) + (1 - c)\theta)^2 \, dN(\theta, 1)(y)$$

$$= \int_{\mathbb{R}} \left(c^2(y - \theta)^2 + 2c(1 - c)\theta(y - y) + (1 - c)^2\theta^2\right) \, dN(\theta, 1)(y)$$

$$= c^2 + (1 - c)^2\theta^2.$$
It follows that $\delta_1$ is $R$-better than all $\delta_c$ for $c > 1$, so that for all $c > 1$, $\delta_c$ is inadmissible. If we had restricted $c$ to be greater than or equal to 1, $\delta_1$ would have been admissible. However, since $c$ may lie in $[0, 1)$ as well, admissibility in the uniform sense of (2.37) does not apply to any $\delta_c$. To see this, note that $R(\theta, \delta_1) = 1$ for all $\theta$, whereas for $c < 1$ and some $\theta > c/(1-c)$, $R(0, \delta_c) < 1 < R(\theta, \delta_c)$. Therefore, there is no admissible decision rule in $\Delta$.

The minimax criterion does give rise to a preference. However, in order to guarantee its existence, we need to bound (or rather, compactify) the parameter space: let $M > 0$ be given and assume that $\Theta = [-M, M]$. The minimax risk for $\delta_c$ is given by

$$\sup_{\theta \in \Theta} R(\theta, \delta_c) = c^2 + (1 - c)^2 M^2,$$

which is minimal iff $c = M^2/(1 + M^2)$, i.e. the (unique) minimax decision rule for this problem (or, since we are using decision theory to estimate a parameter in this case, the minimax estimator in $\Delta$ with respect to $L_2$-loss) is therefore,

$$\delta^M(Y) = \frac{M^2}{1 + M^2} Y.$$

Note that if we let $M \to \infty$, this estimator for $\theta$ converges to the MLE for said problem.

As demonstrated in the above example, uniform admissibility of a decision rule (c.f. (2.37)) is hard to achieve, but in many such cases a minimax decision rule does exist. One important remark concerning the use the minimax decision principle remains: considering (2.38), we see that the minimax principle chooses the decision rule that minimizes the maximum of the risk $R(\cdot, \delta)$ over $\Theta$. As such, the minimax criterion takes into account only the worst-case scenario and prefers decision rules that perform well under those conditions. In practical problems, that means that the minimax principle tends to take a rather pessimistic perspective on decision problems.

### 2.5.2 Bayesian decision theory

Bayesian decision theory presents a more balanced perspective because instead of maximizing the risk function over $\Theta$, the Bayesian has the prior to integrate over $\Theta$. Optimization of the resulting integral takes into account more than just the worst case, so that the resulting decision rule is based on a less pessimistic perspective than the minimax decision rule.

**Definition 2.67.** Let the state-space $\Theta$, states $P_\theta$, $(\theta \in \Theta)$, decision space $\mathcal{A}$ and loss $L$ be given. In addition, assume that $\Theta$ is a measurable space with $\sigma$-algebra $\mathcal{A}$ and prior $\Pi : \mathcal{A} \to \mathbb{R}$. The function

$$r(\Pi, \delta) = \int_\Theta R(\theta, \delta) d\Pi(\theta),$$

(2.39)

is called the Bayesian risk function. Let $\delta_1, \delta_2 \in \Delta$ be given. The decision rule $\delta_1$ is Bayes-preferred to $\delta_2$, if

$$r(\Pi, \delta_1) < r(\Pi, \delta_2).$$
If \( \delta^\Pi \in \Delta \) minimizes \( \delta \mapsto r(\Pi, \delta) \), i.e.

\[
r(\Pi, \delta^\Pi) = \inf_{\delta \in \Delta} r(\Pi, \delta).
\]

then \( \delta^\Pi \) is called a Bayes rule. The quantity \( r(\Pi, \delta^\Pi) \) is called the Bayes risk.

**Lemma 2.68.** Let \( Y \in \mathcal{Y} \) denote data in a decision theoretic problem with state space \( \Theta \), decision space \( \mathcal{A} \) and loss \( L : \Theta \times \mathcal{A} \to \mathbb{R} \). For any prior \( \Pi \) and all decision rules \( \delta : \mathcal{Y} \to \mathcal{A} \),

\[
r(\Pi, \delta) \leq \sup_{\theta \in \Theta} R(\theta, \delta),
\]

i.e. the Bayesian risk is always upper bounded by the minimax risk.

The proof of this lemma follows from the fact that the minimax risk is an upper bound for the integrand in the Bayesian risk function.

**Example 2.69.** (continuation of example 2.66) Let \( \Theta = \mathbb{R} \) and \( Y \sim N(\theta_0, 1) \) for some unknown \( \theta_0 \in \Theta \). Choose the loss-function \( L : \Theta \times \mathcal{A} \to \mathbb{R} \) and the decision space \( \Delta \) as in example 2.66. We choose a prior \( \Pi = N(0, \tau^2) \) (for some \( \tau > 0 \)) on \( \Theta \). Then the Bayesian risk function is given by:

\[
r(\Pi, \delta_c) = \int_\Theta R(\theta, \delta_c) \, d\Pi(\theta) = \int_\mathbb{R} (c^2 + (1-c)^2 \theta^2) \, dN(0, \tau^2)(\theta)
\]

which is minimal iff \( c = \frac{\tau^2}{1 + \tau^2} \). The (unique) Bayes rule for this problem and corresponding Bayes risk are therefore,

\[
\delta^\Pi(Y) = \frac{\tau^2}{1 + \tau^2} Y, \quad r(\Pi, \delta^\Pi) = \frac{\tau^2}{1 + \tau^2}.
\]

In the Bayesian case, there is no need for a compact parameter space \( \Theta \), since we do not maximize but integrate over \( \Theta \).

In the above example, we could find the Bayes rule by straightforward optimization of the Bayesian risk function, because the class \( \Delta \) was rather restricted. If we extend the class \( \Delta \) to contain all non-randomized decision rules, the problem of finding the Bayes rule seems to be far more complicated at first glance. However, as we shall see in theorem 2.6, the following definition turns out to be the solution to this question.

**Definition 2.70.** (The conditional Bayes decision principle) Let the state-space \( \Theta \), states \( P_\theta \), \( (\theta \in \Theta) \), decision space \( \mathcal{A} \) and loss \( L \) be given. In addition, assume that \( \Theta \) is a measurable space with \( \sigma \)-algebra \( \mathcal{G} \) and prior \( \Pi : \mathcal{G} \to \mathbb{R} \). We define the decision rule \( \delta^* : \mathcal{Y} \to \mathcal{A} \) to be such that for all \( y \in \mathcal{Y} \),

\[
\int_\Theta L(\theta, \delta^*(y)) \, d\Pi(\theta \mid Y = y) = \inf_{a \in \mathcal{A}} \int_\Theta L(\theta, a) \, d\Pi(\theta \mid Y = y),
\]

i.e. point-wise for every \( y \), the decision rule \( \delta^*(y) \) minimizes the posterior expected loss.
The above defines the decision rule \( \delta^* \) implicitly as a point-wise minimizer, which raises the usual questions concerning existence and uniqueness, of which little can be said in any generality. However, if the existence of \( \delta^* \) is established, it is optimal.

**Theorem 2.71.** Let the state-space \( \Theta \), states \( P_\theta \), \( (\theta \in \Theta) \), decision space \( A \) and loss \( L \) be given. In addition, assume that \( \Theta \) is a measurable space with \( \sigma \)-algebra \( \mathcal{G} \) and prior \( \Pi : \mathcal{G} \to \mathbb{R} \). Assume that there exists a \( \sigma \)-finite measure \( \mu : \mathcal{B} \to \mathbb{R} \) such that \( P_\theta \ll \mu \) for all \( \theta \in \Theta \). If the decision rule \( \delta^* : \mathcal{Y} \to A \) is well-defined, then \( \delta^* \) is a Bayes rule.

**Proof** Denote the class of all decision rules for this problem by \( \Delta \) throughout the proof. We start by rewriting the Bayesian risk function for a decision rule \( \delta : \mathcal{Y} \to A \).

\[
r(\Pi, \delta) = \int_{\Omega} R(\theta, \delta) d\Pi(\theta) = \int_{\mathcal{Y}} \int_{\Theta} L(\theta, \delta(y)) dP_\theta(y) d\Pi(\theta)
\]

\[
= \int_{\mathcal{Y}} \int_{\Theta} L(\theta, \delta(y)) p_\theta(y) d\Pi(\theta) d\mu(y)
\]

\[
= \int_{\mathcal{Y}} \left( \int_{\Theta} p_\theta(y) d\Pi(\theta) \right) \int_{\Theta} L(\theta, \delta(y)) d\Pi(\theta|Y = y) d\mu(y).
\]

where we use definitions (2.39) and (2.36), the Radon-Nikodym theorem (see theorem A.26), Fubini’s theorem (see theorem A.25) and the definition of the posterior, c.f. (2.12). Using the prior predictive distribution (2.3), we rewrite the Bayesian risk function further:

\[
r(\Pi, \delta) = \int_{\mathcal{Y}} \int_{\Theta} L(\theta, \delta(y)) d\Pi(\theta|Y = y) dP^\Pi(y).
\]

By assumption, the conditional Bayes decision rule \( \delta^* \) exists. Since \( \delta^* \) satisfies (2.41) point-wise for all \( y \in \mathcal{Y} \), we have

\[
\int_{\Theta} L(\theta, \delta^*(y)) d\Pi(\theta|Y = y) = \inf_{\delta \in \Delta} \int_{\Theta} L(\theta, \delta(y)) d\Pi(\theta|Y = y).
\]

Substituting this in (2.42), we obtain

\[
r(\Pi, \delta^*) = \int_{\mathcal{Y}} \inf_{\delta \in \Delta} \int_{\Theta} L(\theta, \delta(y)) d\Pi(\theta|Y = y) dP^\Pi(y)
\]

\[
\leq \inf_{\delta \in \Delta} \int_{\mathcal{Y}} \int_{\Theta} L(\theta, \delta(y)) d\Pi(\theta|Y = y) dP^\Pi(y)
\]

\[
= \inf_{\delta \in \Delta} r(\Pi, \delta).
\]

which proves that \( \delta^* \) is a Bayes rule. \( \square \)

To conclude, it is noted that randomization of the decision is not needed when optimizing with respect to the Bayes risk. The conditional Bayes decision rule is non-randomized and optimal.

### 2.5.3 Frequentist versus Bayesian classification

Many decision-theoretic questions take the form of a classification problem: under consideration is a population \( \Omega \) of objects that each belong to one of a finite number of classes
$\mathcal{A} = \{1, 2, \ldots, L\}$. The class $K$ of the object is the unknown quantity of interest. Observing a vector $Y$ of features of the object, the goal is to classify the object, i.e. estimate which class it belongs to. We formalize the problem in decision-theoretic terms: the population is a probability space $(\Omega, \mathcal{F}, P)$; both the feature vector and the class of the object are random variables, $Y : \Omega \rightarrow \mathcal{Y}$ and $K : \Omega \rightarrow \mathcal{A}$ respectively. The state-space in a classification problem equals the decision space $\mathcal{A}$: the class can be viewed as a “state” in the sense that the distribution $P_{Y|K=k}$ of $Y$ given the class $K = k$ depends on $k$. Based on the feature vector $Y$, we decide to classify in class $\delta(Y)$, i.e. the decision rule (or classifier, as it is usually referred to in the context of classification problems) maps features to classes by means of a map $\delta : \mathcal{Y} \rightarrow \mathcal{A}$. A classifier $\delta$ can be viewed equivalently as a finite partition of the feature-space $\mathcal{Y}$: for every $k \in \mathcal{A}$, we define $\mathcal{Y}_k = \{y \in \mathcal{Y} : \delta(y) = k\}$ and note that if $k \neq l$, then $\mathcal{Y}_k \cap \mathcal{Y}_l = \emptyset$ and $\mathcal{Y}_1 \cup \mathcal{Y}_2 \cup \ldots \cup \mathcal{Y}_L = \mathcal{Y}$. The partition of the feature space is such that if $Y = y \in \mathcal{Y}_k$ for certain $k \in \mathcal{A}$, then we classify the object in class $k$.

Depending on the context of the classification problem, a loss-function $L : \mathcal{A} \times \mathcal{A} \rightarrow \mathbb{R}$ is defined (see the examples in the introduction to this section, e.g. the example on medical diagnosis). Without context, the loss function in a classification problem can be chosen as follows

$$L(k, l) = 1_{\{k \neq l\}}.$$ i.e. we incur a loss equal to one for each misclassification. Using the minimax decision principle, we look for a classifier $\delta^M : \mathcal{Y} \rightarrow \mathcal{A}$ that minimizes:

$$\delta \mapsto \sup_{k \in \mathcal{A}} \int_{\mathcal{Y}} L(k, \delta(y)) \, dP(y | K = k) = \sup_{k \in \mathcal{A}} P(\delta(Y) \neq k \mid K = k),$$

i.e. the minimax decision principle prescribes that we minimize the probability of misclassification uniformly over all classes.

In a Bayesian context, we need a prior on the state-space, which equals $\mathcal{A}$ in classification problems. Note that if known (or estimable), the marginal probability distribution for $K$ is to be used as the prior for the state $k$, in accordance with definition 2.2. In practical problems, frequencies of occurrence for the classes $\{1, \ldots, L\}$ in $\Omega$ are often available or easily estimable; in the absence of information on the marginal distribution of $K$ equal prior weights can be assigned. Here, we assume that the probabilities $P(K = k)$ are known and use them to define the prior density with respect to the counting measure on the (finite) space $\mathcal{A}$:

$$\pi(k) = P(K = k).$$

The Bayes rule $\delta^* : \mathcal{Y} \rightarrow \mathcal{A}$ for this classification problem is defined to be as the minimizer of

$$\delta \mapsto \sum_{k \in \mathcal{A}} L(k, \delta(y)) \, d\Pi(k | Y = y) = \sum_{k=1}^{L} \Pi(\delta(y) \neq K \mid Y = y)$$
for every $y \in \mathcal{Y}$. According to theorem 2.71, the classifier $\delta^*$ minimizes the Bayes risk, which in this situation is given by:

$$r(\Pi, \delta) = \sum_{k \in \mathcal{A}} R(k, \delta) \pi(k) = \sum_{k \in \mathcal{A}} \int_{\mathcal{Y}} L(k, \delta(y)) \, dP(y | K = k) \, \pi(k) = \sum_{k \in \mathcal{A}} P(k \neq \delta(Y) \mid K = k) P(K = k) = P(K \neq \delta(Y)).$$

Summarizing, the Bayes rule $\delta^*$ minimizes the overall probability of misclassification, i.e. without referring to the class of the object. (Compare this with the minimax classifier.)

Readers interested in the statistics of classification and its applications are encouraged to read B. Ripley’s “Pattern recognition and neural networks” (1996) [77].

To close the chapter, the following remark is in order: when we started our comparison of frequentist and Bayesian methods, we highlighted the conflict in philosophy. However, now that we have seen some of the differences in more detail by considering estimation, confidence sets, testing and decision theory in both schools, we can be far more specific. Statistical problems can be solved in both schools; whether one chooses for a Bayesian or frequentist solution is usually not determined by deeply felt belief in either philosophy but by much more practical considerations. Perhaps the classification example of this subsection illustrates this point most clearly: if one is concerned about correct classification for objects in the most difficult class, one should opt for the minimax decision rule. If, on the other hand, one wants to minimize the overall misclassification probability (disregarding misclassification per class), one should choose to adopt the conditional Bayes decision rule. In other words, depending on the risk to be minimized (minimax risk and Bayes risk are different!) one arrives at different classifiers.

An argument in favour of Bayesian methods is the control that the statistician can gain over a statistical procedure by choice of a suitable prior. Particularly, a statistical procedure may benefit from a well-chosen form of bias, expressed through the prior. Examples include the subjective bias intended in example 1.22. To give another example of this nature, Bayesian methods are popular in forensic statistics because the freedom to choose a prior leaves room to incorporate background information and common-sense. However, from a more technical point of view, bias may be required for regularization purposes (like a penalty, in frequentist terms, see remark 2.32). Prior bias may even be guided in a data-dependent way, e.g. when we employ empirical Bayesian methods (see section 3.5), to optimize the procedure.

Another reason to use one or the other may be computational advantages or useful theoretical results that exist for one school but have no analog in the other. Philosophical preference should not play a role in the choice for a statistical procedure, practicality should (and usually does).
2.6 Exercises

Exercise 2.1. Calibration

A physicist prepares for measurement of a physical quantity $Z$ in his laboratory. To that end, he installs a measurement apparatus that will give him an outcome of the form $Y = Z + e$ where $e$ is a measurement error due to the inaccuracy of the apparatus, assumed to be stochastically independent of $Z$. Note that if the expectation of $e$ equals zero, long-run sample averages converge to the expectation of $Z$; if $Pe \neq 0$, on the other hand, averaging does not cancel out the resulting bias. The manufacturer of the apparatus says that $e$ is normally distributed with known variance $\sigma^2 > 0$. The mean $\theta$ of this normal distribution depends on the way the apparatus is installed and thus requires calibration. The following questions pertain to the calibration procedure.

The physicist decides to conduct the following steps to calibrate his measurement. First, he makes certain that the apparatus receives no input signal, $Z = 0$. Then he repeats measurement of $Y$, generating an i.i.d. sample of size $n$, which amounts to an i.i.d. sample from the distribution of $e$ used to estimate the unknown mean $\theta$. The physicist expects that $Ee$ lies close to zero.

a. Explain why, from a subjectivist point of view, the choice $\theta \sim N(0, \tau^2)$ forms a suitable prior in this situation. Explain the role of the parameter $\tau^2 > 0$.

b. With the choice of prior as in part a., calculate the posterior density for $\theta$.

c. Interpret the influence of $\tau^2$ on the posterior, taking into account your answer under part a. (Hint: take limits $\tau^2 \downarrow 0$ and $\tau^2 \uparrow \infty$ in the expression you have found under b.)

d. What is the influence of the sample size $n$? Show that the particular choice of the constant $\tau^2$ becomes irrelevant in the large-sample limit $n \to \infty$.

Exercise 2.2. Let $X_1, \ldots, X_n$ be an i.i.d. sample from the uniform distribution $U[0, \theta]$, with unknown parameter $\theta \in \Theta = (1, \infty)$. As a prior for $\theta$, choose the Pareto distribution with exponent $\alpha > 0$. Calculate the posterior density for $\theta$ with respect to the Lebesgue measure on $[0, \infty)$.

Exercise 2.3. Let $X_1, \ldots, X_n$ be an i.i.d. sample from the Poisson distribution $\text{Poisson}(\lambda)$, with unknown parameter $\lambda > 0$. As a prior for $\lambda$, let $\lambda \sim \Gamma(2,1)$. Calculate the posterior density for $\lambda$ with respect to the Lebesgue measure on $[0, \infty)$.

Exercise 2.4. Let $X_1, \ldots, X_n$ be an i.i.d. sample from a binomial distribution $\text{Bin}(n, \theta)$, with known $n \geq 1$ and unknown $\theta \in \Theta = [0, 1]$. As a prior for $\theta$, use a beta distribution, $\theta \sim \beta(2,2)$. Calculate the posterior density for $\theta$ with respect to the Lebesgue measure on $[0, 1]$.

Exercise 2.5. Let $X_1, \ldots, X_n$ be an i.i.d. sample from a normal distribution $N(0, \sigma^2)$, with unknown $\sigma^2 > 0$. We define the prior for the variance $\sigma^2$ implicitly, by stating that the inverse
$1/\sigma^2$ is distributed according to a $\Gamma(\alpha, \beta)$ distribution. Calculate the posterior density for $\sigma^2$ with respect to the Lebesgue measure on $[0, \infty)$.

**Exercise 2.6.** Let the measurement $Y \sim P_0$ be given. Assume that the model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ is dominated but possibly misspecified. Let $\Pi$ denote a prior distribution on $\Theta$. Show that the posterior distribution is $P_0$-almost-surely equal to the prior distribution iff the likelihood is $\Pi \times P_0$-almost-surely constant (as a function of $(\theta, y) \in \Theta \times \mathcal{Y}$). Explain the result of example 2.12 in this context.

**Exercise 2.7.** For this exercise, we introduce two concepts. Firstly, a locally compact space is a topological space in which every point has an open neighbourhood with compact closure. Secondly, a Borel measure $\mu$ is called a Radon measure if $\mu$ is locally finite (i.e. every point has a neighbourhood of finite $\mu$-measure) and inner regular (i.e. every measurable set can be approximated in $\mu$-measure by compact subsets: for every Borel measurable $G$, $\mu(G) = \sup\{\mu(K) : K \subset G, K \text{ compact}\}$). Suppose $\Theta$ is a locally compact space and a probability space $(\Theta, \mathcal{F}, \Pi)$ with $\sigma$-algebra $\mathcal{F}$ that contains the Borel $\sigma$-algebra. Show that if $\Pi$ is a Radon measure, $\Pi(S) = 1$ (with $S$ as in (2.10)).

**Exercise 2.8.** In the model of exercise 2.2, calculate the maximum-likelihood estimator, the posterior mean and the maximum-a-posteriori estimator.

**Exercise 2.9.** In the model of exercise 2.4, calculate the maximum-likelihood estimator, the posterior mean and the maximum-a-posteriori estimator.

**Exercise 2.10.** In the model of exercise 2.5, calculate the maximum-likelihood estimator, the posterior mean and the maximum-a-posteriori estimator.

**Exercise 2.11.** Consider the following questions in the context of exercise 2.3.

a. Calculate the maximum-likelihood estimator and the maximum-a-posteriori estimator for $\lambda \in (0, \infty)$.

b. Let $n \to \infty$ both in the MLE and MAP estimator and conclude that the difference vanishes in the limit, $P_\lambda$-almost-surely.

c. Following remark 2.32, explain the difference between ML and MAP estimators exclusively in terms of the prior.

d. Consider and discuss the choice of prior $\lambda \sim \Gamma(2,1)$ twice, once in a qualitative, subjectivist Bayesian fashion, and once following the frequentist interpretation of the log-prior-density.

**Exercise 2.12.** Let $Y \sim P_0$ denote the data and $\mathcal{P}$ a model with metric $d$. Suppose that $\mathcal{P}$ is endowed with a prior defined on the Borel $\sigma$-algebra induced by the metric topology. Assume that $P_0 \ll P^{\Pi}$ and that $\mathcal{P}$ is compact. The following questions pertain to the small-ball estimators defined in definition 2.29 and remark 2.30. We assume that the posterior
Bayesian basics

distribution is such that for all $\epsilon > 0$ and all $P \in \mathcal{P}$, the (topological) boundary of the ball $B_d(P, \epsilon)$ receives mass equal to zero: $\Pi(\partial B_d(P, \epsilon)|Y) = 0$, $P_0$-almost surely.

a. Show that, for any $p \in (1/2, 1)$ and large enough $\epsilon > 0$, the small-ball estimator $\hat{P}$ of exists, $P_0$-almost-surely.

b. Show that for any two measurable model subsets $A, B \subset \mathcal{P}$,
\[
|\Pi(A|Y) - \Pi(B|Y)| \leq \Pi(A \cup B|Y) - \Pi(A \cap B|Y),
\]
$P_0$-almost-surely.

c. Show that for every $\epsilon > 0$, the map $P \mapsto \Pi(B_d(P, \epsilon)|Y)$ is continuous, $P_0$-almost-surely.

d. Show that for every $\epsilon > 0$, the small-ball estimator of definition 2.29 exists.

e. Let some $p \in (1/2, 1)$ be given. Suppose that $\epsilon > 0$ denotes some radius for which there exists a ball $B_d(P, \epsilon) \subset \mathcal{P}$ of posterior probability greater than or equal to $p$. Show that, if both $\hat{P}_1$ and $\hat{P}_2$ are centre points of such balls, then $d(\hat{P}_1, \hat{P}_2) < 2\epsilon$, $P_0$-almost-surely.

Exercise 2.13. Let $Y_n = (X_1, \ldots, X_n)$ be an i.i.d. sample from the normal distribution $N(\mu, \sigma^2)$ for certain $\mu \in \mathbb{R}$, $\sigma^2 > 0$. Show that the sample average is distributed according to the normal distribution,
\[
P_nX \sim N(\mu, \sigma_n^2),
\]
with variance $\sigma_n^2 = \sigma^2/n$.

Exercise 2.14. Let $Y$ be normally distributed with known variance $\sigma^2 > 0$ and unknown location $\theta$. As a prior for $\theta$, choose $\Pi = N(0, \tau^2)$. Let $\alpha \in (0, 1)$ be given. Using the posterior density with respect to the Lebesgue measure, express the level-$\alpha$ HPD-credible set in terms of $Y$, $\sigma^2$, $\tau^2$ and quantiles of the standard normal distribution. Consider the limit $\tau^2 \to \infty$ and compare with level-$\alpha$ confidence intervals centred on the ML estimate for $\theta$.

Exercise 2.15. Let $Y \sim \text{Bin}(n; p)$ for known $n \geq 1$ and unknown $p \in (0, 1)$. As a prior for $p$, choose $\Pi = \text{Beta}(\frac{1}{2}, \frac{1}{2})$. Calculate the posterior distribution for the parameter $p$. Using the Lebesgue measure on $(0, 1)$ to define the posterior density, give a level-$\alpha$ (non-HPD!) credible interval for $p$ in terms of $Y$, $n$ and the quantiles of beta-distributions.

Exercise 2.16. Let $\Theta$ be a subset of $\mathbb{R}$ and let $\Theta \to \mathcal{P} : \theta \mapsto P_\theta$ describe an identifiable parametrization of the model $\mathcal{P}$ for an i.i.d. sample $X_1, \ldots, X_n$, and assume that there exists a $\theta_0 \in \Theta$ such that $P_{\theta_0}$ is the marginal distribution for each of the $X_i$. Let $\theta$ and $\theta'$ with $\theta' > \theta$ from $\Theta$ be given and consider the hypotheses,
\[
H_0 : \theta_0 = \theta, \quad H_1 : \theta_0 = \theta',
\]
Given a significance level $\alpha \in (0, 1)$, write down the Neyman-Pearson test for $H_0$ versus $H_1$ (see lemma 2.50), in each of the following cases,
a. for all $\theta \in [0, 1]$, $P_\theta = \text{Bernoulli}(\theta)$;

b. for all $\theta \in (0, \infty)$, $P_\theta = \text{Poisson}(\theta)$;

c. for all $\theta \in [0, 1]$, $P_\theta = \text{Bin}(\theta, k)$ for some known integer $k \geq 1$.

**Exercise 2.17.** Consider a dominated model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ for data $Y$, where $\Theta \subset \mathbb{R}$ is an interval. For certain $\theta_0 \in \Theta$, consider the simple null-hypothesis and alternative:

$$H_0 : \theta = \theta_0, \quad H_1 : \theta \neq \theta_0.$$ 

Show that if the prior $\Pi$ is absolutely continuous with respect to the Lebesgue measure on $\Theta$, then the posterior odds ratio in favour of the hypothesis $H_0$ equals zero.

[Remark: conclude that calculation of posterior odds ratios makes sense only if both hypotheses receive non-zero prior mass. Otherwise, the statistical question we ask is rendered invalid ex ante by our beliefs concerning $\theta$, as expressed through the choice of the prior. (See example 2.12.)]

**Exercise 2.18.** Let $X_1, \ldots, X_n$ be an i.i.d. sample from a binomial distribution $\text{Bin}(\theta, k)$, for some known integer $k \geq 1$ and an unknown parameter $\theta \in \Theta = [0, 1]$. Let the prior $\Pi$ for $\theta$ be a Beta distribution $B(\alpha, \beta)$, with certain parameters $\alpha, \beta > 1$.

a. Calculate the posterior distribution for $\theta$.

b. Write down the equations that determine the two end-points of the HPD credible interval (based on the density of the posterior relative to Lebesgue measure on $\Theta$), for given credible level $\delta \in (0, 1)$.

Consider the hypotheses,

$$H_0 : \quad \theta \leq \frac{1}{2}, \quad H_1 : \quad \theta > \frac{1}{2}.$$ 

c. Give the prior odds, posterior odds and Bayes factor for the hypotheses $H_0$ and $H_1$.

**Exercise 2.19.** PRISONER’S DILEMMA

Two men have been arrested on the suspicion of burglary and are held in separate cells awaiting interrogation. The prisoners have been told that burglary carries a maximum sentence of $x$ years. However, if they confess, their prison terms are reduced to $y$ years (where $0 < y < x$).

If one of them confesses and the other does not, the first receives a sentence of $y$ years while the other is sentenced to $x$ years. If he keeps his mouth shut but his partner talks, he gets the maximum sentence. If he talks, he will always receive a sentence of
y years and the other prisoner receives y or x years depending on whether he confessed or not himself. To talk or not to talk, that is the question.

There is no data in this problem, so we set θ equal to 1 or 0, depending on whether the other prisoner talks or not. Our prisoner can decide to talk (t = 1) or not (t = 0). The loss function L(θ, t) equals the prison term for our prisoner. In the absence of data, risk and loss are equal.

a. Calculate the minimax risk for both t = 0 and t = 1. Argue that the minimax-optimal decision for our prisoner is to confess.

As argued in section 2.5, the minimax decision can be overly pessimistic. In the above, it assumes that the other prisoner will talk and chooses t accordingly.

The Bayesian perspective balances matters depending on the chance that the other prisoner will confess when interrogated. This chance finds its way into the formalism as a prior for the trustworthiness of the other prisoner. Let p ∈ [0, 1] be the probability that the other prisoner confesses, i.e. Π(θ = 1) = p and Π(θ = 0) = 1 − p.

b. Calculate the Bayes risks for t = 0 and t = 1 in terms of x, y, and p. Argue that the Bayes decision rule for our prisoner is as follows: if y/x > p then our prisoner does not confess, if y/x < p, the prisoner confesses. If y/x = p, the Bayes decision criterion does not have a preference.

So, depending on the degree to which our prisoner trusts his associate and the ratio of prison terms, the Bayesian draws his conclusion. The latter is certainly more sophisticated and perhaps more realistic, but it requires that our prisoner quantifies his trust in his partner in the form of a prior Bernoulli(p) distribution.
Chapter 3

Choice of the prior

Bayesian procedures have been the object of much criticism, often focusing on the choice of the prior as an undesirable source of ambiguity. The answer of the subjectivist that the prior represents the “belief” of the statistician or “expert knowledge” pertaining to the measurement elevates this ambiguity to a matter of principle, thus setting the stage for a heated debate between “pure” Bayesians and “pure” frequentists concerning the philosophical merits of either school within statistics. As said, the issue is complicated further by the fact that the Bayesian procedure does not refer to the “true” distribution $P_0$ for the observation (see section 2.1), providing another point of fundamental philosophical disagreement for the fanatically pure to lock horns over. Leaving the philosophical argumentation to others, we shall try to discuss the choice of a prior at a more conventional, practical level.

In this chapter, we look at the choice of the prior from various points of view: in section 3.1, we consider the priors that emphasize the subjectivist’s prior “belief”. In section 3.2 we construct priors with the express purpose not to emphasize any part of the model, as advocated by objectivist Bayesians. Because it is often desirable (and always computationally advantageous) to have closed-form expressions for posterior distributions, so-called conjugacy of families of distributions over the model is considered in section 3.3. Hierarchical prior construction and Bayesian modelling are the subject of section 3.4 and the practice of choosing priors by frequentist means (commonly known as empirical Bayes) forms the subject of section 3.5. Special attention goes to the Dirichlet distributions of section 3.6 because they describe distributions on spaces of probability measures (rather than their parametrizing spaces). As will become clear in the course of the chapter, the choice of a “good” prior is highly dependent on the model under consideration, as well as on the purpose of the analysis.

Since the Bayesian school has taken up an interest in non-parametric statistics only relatively recently, most (if not all) of the material presented in the first sections of this chapter applies only to parametric models. To find a suitable prior for a non-parametric model
Choice of the prior can be surprisingly complicated. Not only does the formulation involve topological aspects that do not play a role in parametric models, but also the properties of the posterior may be surprisingly different from those encountered in parametric models! A natural prior on fully non-parametric models is considered in subsection 3.6.2, extending the finite-dimensional Dirichlet distributions of subsection 3.6.1.

3.1 Subjective priors

As was explained in chapters 1 and 2, all statistical procedures require the statistician to make certain choices, e.g. for model and method of inference. The subjectivist chooses the model as a collection of stochastic explanations of the data that he finds “reasonable”, based on criteria no different from those frequentists and objectivist Bayesians would use.

3.1.1 Motivation for the subjectivist approach

Bayesians then proceed to choose a prior, in such a manner that the support of this prior is not essentially smaller than the model itself. But even when the support of the prior is fixed, there is a large collection of possible priors left to be considered, each leading to a different posterior distribution. The objectivist Bayesian will choose from those possibilities a prior that is “homogeneous” (in a suitable sense), in the hope of achieving unbiased inference. The subjectivist, however, chooses his prior such as to emphasize parts of the model that he believes in stronger than others, thereby introducing a bias in his inferential procedure explicitly. Such a prior is called a subjective prior, or informative prior. The reason for this approach is best explained by examples like 1.22 which demonstrate that intuitive statistical reasoning is not free of bias either.

Subjectivity finds its mathematical expression when high prior “belief” is translated into “relatively large” amounts of assigned prior mass to certain regions of the model. However, there is no clear rule directing the exact fashion in which prior mass is to be distributed. From a mathematical perspective this is a rather serious shortcoming, because it leaves us without a precise definition of the subjectivist approach. Often the subjectivist will have a reasonably precise idea about his “beliefs” at the roughest level (e.g. concerning partitions of the model into a few subsets) but none at more detailed levels. When the parameter space $\Theta$ is unbounded this lack of detail becomes acute, given that the tail of the prior is hard to fix by subjective reasoning, yet highly influential for the inferential conclusions based on it. In practice, a subjectivist will often choose his prior without mathematical precision. He considers the problem, interprets the parameters in his model and chooses a prior to reflect all the (background) information at his disposition, ultimately filling in remaining details in an ad-hoc manner. It is worthwhile to mention that studies have been conducted focused on the ability of people to make a realistic guess at a probability distribution: they have shown that without specific training or practice, people tend to be overconfident in their assessment,
assigning too much mass to possibilities they deem most likely and too little to others [1]. A tentative conclusion might be, that people tend to formulate their “beliefs” on a deterministic basis and deviate from that point of view only slightly (or, too little) when asked to give a realistic probabilistic perspective. (For more concerning the intricacies of choosing subjective prior distributions, see Berger (1985) [8].)

**Remark 3.1.** For this reason, it is imperative that a subjectivist prior is always reported alongside inferential conclusions based upon it! Reporting methods is important in any statistical setting, but if chosen methods lead to express bias, explanation is even more important. Indeed, not only the prior but also the reasoning leading to its choice should be reported, because in a subjectivist setting, the motivation for the choice of a certain prior (and not any other) is part of the analysis rather than an external consideration.

### 3.1.2 Methods for the construction of subjective priors

If the model $\Theta$ is one-dimensional and the parameter $\theta$ has a clear interpretation, it is often not exceedingly difficult to find a reasonable prior $\Pi$ expressing the subjectivist’s “belief” concerning the value of $\theta$.

**Example 3.2.** If one measures the speed of light in vacuo $c$ (a physical constant, approximately equal to 299792458 m/s), the experiment will be subject to random perturbations outside the control of the experimenter. For example, imperfection of the vacuum in the experimental equipment, small errors in timing devices, electronic noise and countless other factors may influence the resulting measured speed $Y$. We model the perturbations collectively as a normally distributed error $e \sim N(0, \sigma^2)$ where $\sigma$ is known as a characteristic of the experimental setup. The measured speed is modelled as $Y = c + e$, i.e. the model $\mathcal{P} = \{N(c, \sigma^2) : c > 0\}$ is used to infer on $c$. Based on experiments in the past (most famous is the Michelson-Morley experiment (1887)), the experimenter knows that $c$ has a value close to $3 \cdot 10^8$ m/s, so he chooses his prior to reflect this: a normal distribution located at $300000000$ m/s with a standard deviation of (say) $1000000$ m/s will do. The latter choice is arbitrary, just like the choice for a normal location model over other families.

The situation changes when the parameter has a higher dimension, $\Theta \subset \mathbb{R}^d$: first of all, interpretability of each of the $d$ components of $\theta = (\theta_1, \theta_2, \ldots, \theta_d)$ can be far from straightforward, so that concepts like prior “belief” or “expert knowledge” become inadequate guidelines for the choice of a prior. Additionally, the choice for a prior in higher-dimensional models also involves choices concerning the dependence structure between parameters!

**Remark 3.3.** Often, subjectivist inference employs exceedingly simple, parametric models for the sake of interpretability of the parameter (and to be able to choose a prior accordingly). Most frequentists would object to such choices for their obvious lack of realism, since they
view the data as being generated by a “true, underlying distribution”, usually assumed to be an element of the model. However, the subjectivist philosophy does not involve the ambition to be strictly realistic and calls for interpretability instead: to the subjectivist, inference is a personal rather than a universal matter. As such, the preference for simple parametric models is a matter of subjective interpretation rather than an assumption concerning reality or realistic distributions for the data.

When confronted with the question which subjective prior to use on a higher-dimensional model, it is often of help to define the prior in several steps based on a choice for the dependence structure between various components of the parameter. Suppose that the subjectivist can formulate a reasonable distribution for the first component $\theta_1$, if he has some definite values for all other components $\theta_2, \ldots, \theta_d$. Then this distribution is $\Pi_{\theta_1|\theta_2, \ldots, \theta_d}$, the subjectivist prior distribution of $\theta_1$, conditional on $\theta_2, \ldots, \theta_d$. Next suppose that a reasonable subjective prior for the second component may be found, conditional on $\theta_3, \ldots, \theta_d$ but independent of $\theta_1$. This amounts to specification of the conditional distribution $\Pi_{\theta_2|\theta_3, \ldots, \theta_d}$. If we continue like this, eventually defining the marginal prior for the last component $\theta_d$, we have found a prior for the full parameter $\theta$, because for all $A_1, \ldots, A_d \in \mathcal{B}$,

$$
\Pi(\theta_1 \in A_1, \ldots, \theta_d \in A_d) = \Pi(\theta_1 \in A_1|\theta_2 \in A_2, \ldots, \theta_d \in A_d) \Pi(\theta_2 \in A_2|\theta_3 \in A_3, \ldots, \theta_d \in A_d) \times \ldots \times \Pi(\theta_{d-1} \in A_{d-1}|\theta_d \in A_d) \Pi(\theta_d \in A_d).
$$

Because prior beliefs may be more easily expressed when imagining a situation where other parameters have fixed values, one eventually succeeds in defining the prior for the high-dimensional model.

**Example 3.4.** Suppose that we measure pairs $(W_i, L_i)$, $1 \leq i \leq n$, where $W_i$ is the weight and $L_i$ is the length of the $i$-th draw from an i.i.d. sample of monkeys from a certain population. As our model for $(W, L)$, we choose model distributions that are products of Gamma distributions with shape parameter $k = 3$: $(W, L) \sim \Gamma(w/3, 3) \times \Gamma(l/3, 3)$, so that $(W, L)$ given $(w, l)$ has expectation $(w, l)$. To choose a prior for the parameter $(w, l)$, we note that a relatively tall monkey will also be relatively heavy. We express this through the specification of the prior: we assume that weight $w$ and length $l$ approximately follow the relationship $w = K l^\alpha$, for some known $K > 0$ and a known exponent $\alpha$. In that case, given the length $l$, we specify (for some choice $\lambda > 0$),

$$
w|l \sim \Gamma(\lambda^{-1} K l^\alpha, \lambda),
$$

so that the conditional prior expectation for $w|l$ is $K l^\alpha$. For $l$, we choose a marginal prior,

$$
l \sim \Gamma(\lambda^{-1} \ell, \lambda),
$$

where $\ell$ is an approximate mean length for a monkey from the population. The prior variances of $l$ and $w|l$ are equal to $\ell^2/\lambda$ and $K^2 l^{2\alpha}/\lambda$ respectively, i.e. $\lambda$ has the interpretation of being inversely proportional to uncertainty expressed in the prior (higher values of $\lambda$ bias the prior (and hence also the posterior) more to the prior expectation $(\ell, K l^\alpha)$).
The construction indicated here is reminiscent of that of a so-called hyperprior, which is discussed in section 3.3. The difference is that components of \( \theta \) all occur in the definition of model distributions \( P_\theta \), whereas hyperparameters do not. Note that it is important to choose a parametrization of the model in which the independence between \( \theta_i \) and \((\theta_1, \ldots, \theta_{i-1})\), given \((\theta_{i+1}, \ldots, \theta_d)\), is plausible for all \( i \geq 1 \).

In certain situations, the subjectivist has more factual information at his disposal when defining the prior for his analysis. In particular, if a probability distribution on the model reflecting the subjectivist’s “beliefs” can be found by other statistical means, it can be used as a prior. Suppose the statistician is planning to measure a quantity \( Y \) and infer on a model \( P \); suppose also that this experiment repeats or extends an earlier analysis. From the earlier analysis, the statistician may have obtained a posterior distribution on \( P \). For the new experiment, this posterior may serve as a prior.

**Example 3.5.** Let \( \Theta \to \mathcal{P} : \theta \mapsto P_\theta \) be a parametrized model for i.i.d. data \( X_1, X_2, \ldots, X_n \) with prior measure \( \Pi_1 : \mathcal{G} \to [0,1] \). Let the model be dominated (see definition 1.3), so that the posterior \( \Pi_1(\cdot \mid X_1, \ldots, X_n) \) satisfies (2.13). Suppose that this experiment has been conducted, with the sample realised as \((X_1, X_2, \ldots, X_n) = (x_1, x_2, \ldots, x_n)\). Next, consider a new, independent experiment in which a quantity \( X_{n+1} \) is measured (with the same model). As a prior \( \Pi_2 \) for the new experiment, we use the (realised) posterior of the earlier experiment, i.e. for all \( G \in \mathcal{G} \),

\[
\Pi_2(G) = \Pi_1(G \mid X_1 = x_1, \ldots, X_n = x_n).
\]

The posterior for the second experiment then satisfies:

\[
d\Pi_2(\theta \mid X_{n+1}) = \frac{p_\theta(X_{n+1}) d\Pi_1(\theta \mid X_1 = x_1, \ldots, X_n = x_n)}{\int_\Theta p_\theta(X_{n+1}) d\Pi_1(\theta \mid X_1 = x_1, \ldots, X_n = x_n)}
\]

\[
= \frac{p_\theta(X_{n+1}) \prod_{i=1}^n p_\theta(x_i) d\Pi_1(\theta)}{\int_\Theta p_\theta(X_{n+1}) \prod_{j=1}^n p_\theta(x_j) d\Pi_1(\theta)}
\]

The latter form is comparable to the posterior that would have been obtained if we had conducted a single experiment with an i.i.d. sample \( X_1, X_2, \ldots, X_{n+1} \) of size \( n+1 \) and prior \( \Pi_1 \).

In that case, the posterior would have been of the form:

\[
\Pi(\cdot \mid X_1, \ldots, X_{n+1}) = \frac{\prod_{i=1}^{n+1} p_\theta(X_i) d\Pi_1(\theta)}{\int_\Theta \prod_{j=1}^{n+1} p_\theta(X_j) d\Pi_1(\theta)},
\]

i.e. the only difference is the fact that the posterior \( \Pi_1(\cdot \mid X_1 = x_1, \ldots, X_n = x_n) \) is realised. As such, we may interpret independent consecutive experiments as a single, interrupted experiment and the posterior \( \Pi_1(\cdot \mid X_1, \ldots, X_n) \) can be viewed as an intermediate result.
Clearly, there are other ways to obtain a distribution on the model that can be used as an informative prior. One example is the distribution that is obtained when a previously obtained frequentist estimator $\hat{\theta}$ for $\theta$ is subject to a procedure called the bootstrap. Although the bootstrap gives rise to a distribution that is interpreted (in the frequentist sense) as the distribution of the estimator $\hat{\theta}$ rather than $\theta$ itself, a subjectivist may reason that the estimator provides him with the “expert knowledge” on $\theta$ that he needs to define a prior on $\Theta$. (For more on bootstrap methods, see Efron and Tibshirani (1993) [32].)

3.2 Non-informative priors

Objectivist Bayesians agree with frequentists that the “beliefs” of the statistician analyzing a given measurement should play a minimal role in the methodology. Obviously, the model choice already introduces a bias, but rather than embrace this necessity and expand upon it like subjectivists, they seek to keep the remainder of the procedure unbiased. In particular, they aim to use priors that do not introduce additional information (in the form of prior “belief”) in the procedure. Subjectivists introduce their “belief” by concentrating prior mass in certain regions of the model; correspondingly, objectivists prefer priors that are “homogeneous” in an appropriate sense.

3.2.1 Uniform priors

At first glance, one may be inclined to argue that a prior is objective (or non-informative) if it is uniform over the parameter space: if we are inferring on parameter $\theta \in [-1,1]$ and we do not want to favour any part of the model over any other, we would choose a prior of the form, $(A \in \mathcal{B})$,

$$\Pi(A) = \frac{1}{2} \mu(A), \quad (3.3)$$

where $\mu$ denotes the Lebesgue measure on $[-1,1]$. Attempts to minimize the amount of subjectivity introduced by the prior therefore focus on uniformity (argumentation that departs from the Shannon entropy in discrete probability spaces reaches the same conclusion (see, for example, Ghosh and Ramamoorthi (2003) [42], p. 47)). The original references on Bayesian methods (e.g. Bayes (1763) [4], Laplace (1774) [59]) use uniform priors as well. But there are several problems with this approach: first of all, one must wonder how to extend such reasoning when $\theta \in \mathbb{R}$ (or any other unbounded subset of $\mathbb{R}$). In that case, $\mu(\Theta) = \infty$ and we can not normalize $\Pi$ to be a probability measure! Any attempt to extend $\Pi$ to such unbounded models as a probability measure would eventually lead to inhomogeneity, i.e. go at the expense of the unbiasedness of the procedure.

The compromise some objectivists are willing to make, is to relinquish the interpretation that subjectivists give to the prior: they do not express any prior “degree of belief” in $A \in \mathcal{B}$ through the subjectivist statement that the (prior) probability of finding $\vartheta \in A$ equals $\Pi(A)$. Although they maintain the Bayesian interpretation of the posterior, they view the prior as
a mathematical definition rather than a philosophical concept. Then, the following definition can be made without further reservations.

**Definition 3.6.** Given a model \((\Theta, \mathcal{G})\), a prior measure \(\Pi : \mathcal{G} \to \bar{\mathbb{R}}\) such that \(\Pi(\Theta) = \infty\) is called an improper prior.

Note that the normalization factor \(\frac{1}{2}\) in (3.3) cancels in the expression for the posterior, c.f. (2.6): any finite multiple of a (finite) prior is equivalent to the original prior as far as the posterior is concerned. However, this argument does not extend to the improper case: integrability problems or other infinities may ruin the procedure, even to the point where the posterior measure becomes infinite or ill-defined. So not just the philosophical foundation of the Bayesian approach is lost, mathematical integrity of the procedure can no longer be guaranteed either! When confronted with an improper prior, the entire procedure must be checked for potential problems. In particular, one must verify that the posterior is a well-defined probability measure.

**Remark 3.7.** Throughout these notes, whenever we refer to a prior measure, it is implied that this measure is a probability measure unless stated otherwise.

But even if one is willing to accept that objectivity of the prior requires that we restrict attention to models on which “uniform” probability measures exist (e.g. with \(\Theta\) a bounded subset of \(\mathbb{R}^d\)), a more fundamental problem exists: the very notion of uniformity is dependent on the parametrization of the model! To see this we look at a model that can be parametrized in two ways and we consider the way in which uniformity as seen in one parametrization manifests itself in the other parametrization. Suppose that we have a \(d\)-dimensional parametric model \(\mathcal{P}\) with two different parametrizations, on \(\Theta_1 \subset \mathbb{R}^d\) and \(\Theta_2 \subset \mathbb{R}^d\) respectively,

\[
\phi_1 : \Theta_1 \rightarrow \mathcal{P}, \quad \phi_2 : \Theta_2 \rightarrow \mathcal{P}
\]

both of which are bijective. Assume that \(\mathcal{P}\) has a topology and is endowed with the corresponding Borel \(\sigma\)-algebra \(\mathcal{G}\); let \(\phi_1\) and \(\phi_2\) be continuous and assume that their inverses \(\phi_1^{-1}\) and \(\phi_2^{-1}\) are continuous as well. Assuming that \(\Theta_1\) is bounded, we consider the uniform prior \(\Pi_1\) on \(\Theta_1\), i.e. the normalized Lebesgue measure on \(\Theta_1\), i.e. for all \(A \in \mathcal{B}_1\),

\[
\Pi_1(A) = \mu(\Theta_1)^{-1}\mu(A),
\]

This induces a prior \(\Pi'_1\) on \(\mathcal{P}\): for all \(B \in \mathcal{G}\),

\[
\Pi'_1(B) = (\Pi_1 \circ \phi_1^{-1})(B).
\]

In turn, this induces a prior \(\Pi''_1\) on \(\Theta_2\): for all \(C \in \mathcal{B}_2\),

\[
\Pi''_1(C) = (\Pi'_1 \circ (\phi_2^{-1})^{-1})(C) = (\Pi'_1 \circ \phi_2)(C) = (\Pi_1 \circ (\phi_1^{-1} \circ \phi_2))(C).
\]

Even though \(\Pi_1\) is uniform, generically \(\Pi''_1\) is not, because, effectively, we are mapping (a subset of) \(\mathbb{R}^d\) to \(\mathbb{R}^d\) by \(\phi_2^{-1} \circ \phi_1 : \Theta_1 \rightarrow \Theta_2\). (Such re-coordinatizations are used extensively in
differential geometry, where a manifold can be parametrized in various ways by sets of maps called charts.

**Example 3.8.** Consider the model \( \mathcal{P} \) of all normal distributions centred on the origin with unknown variance between 0 and 1. We may parametrize this model in many different ways, but we consider only the following two:

\[
\phi_1 : (0, 1) \to \mathcal{P} : \tau \mapsto N(0, \tau), \quad \phi_2 : (0, 1) \to \mathcal{P} : \sigma \mapsto N(0, \sigma^2). \tag{3.6}
\]

Although used more commonly than \( \phi_1 \), parametrization \( \phi_2 \) is not special in any sense: both parametrizations describe exactly the same model. Now, suppose that we choose to endow the first parametrization with a uniform prior \( \Pi_1 \), equal to the Lebesgue measure \( \mu \) on \((0, 1)\). By (3.5), this induces a prior on \( \mathcal{P} \). Let us now see what this prior looks like if we consider \( \mathcal{P} \) parametrized by \( \sigma \): for any constant \( C \in (0, 1) \) the point \( N(0, C) \) in \( \mathcal{P} \) is the image of \( \tau = C \) and \( \sigma = \sqrt{C} \), so the relation between \( \tau \) and corresponding \( \sigma \) is given by

\[
\tau(\sigma) = (\phi_2^{-1} \circ \phi_1)(\sigma) = \sigma^2.
\]

Since \( \Pi_1 \) equals the Lebesgue measure, we find that the density of \( \Pi'_1 \) with respect to the Lebesgue measure equals:

\[
\pi'_1(\sigma) = \pi_1(\tau(\sigma)) \frac{d\tau}{d\sigma}(\sigma) = 2\sigma.
\]

This density is non-constant and we see that \( \Pi'_1 \) is non-uniform. In a subjectivist sense, the prior \( \Pi'_1 \) places higher prior “belief” on values of \( \sigma \) close to 1 than on values close to 0.

From the above argument and example 3.8, we see that uniformity of the prior is entirely dependent on the parametrization: what we call “uniform” in one parametrization, may be highly non-uniform in another. Consequently, what is deemed “objective” in one parametrization may turn out to be highly subjective in another.

What matters is the model \( \mathcal{P} \), not its parametrization in terms of one parameter or another! The parametrization is a mere choice made by the statistician analyzing the problem. Therefore, any statistical concept that depends on the parametrization is flawed from the outset. Through \( \mathcal{P} \) and only through \( \mathcal{P} \) do the parameters \( \sigma \) and \( \tau \) have any bearing on (the law of) the observation in example 3.8. If we could define what is meant by uniformity on the model \( \mathcal{P} \) itself, instead of on its parametrizing spaces, one would obtain a viable way to formalize objectivity. But spaces of probability measures do not have an intrinsic notion of uniformity (like translation-invariance of Lebesgue measure on \( \mathbb{R}^d \), or more generally, left-invariance of the Haar measure on locally compact topological groups).

### 3.2.2 Jeffreys prior and reference priors

Once it is clear that uniformity on any parametrizing space does not have intrinsic meaning in the model \( \mathcal{P} \), the very definition of objectivity in terms of uniformity of the prior is void. A
subjectivist can use any parametrization to formulate his prejudice (note that the subjectivist uses relative prior weights rather than deviations from uniformity to express his prior “belief”), but an objectivist has to define his notion of “objectivity” regardless of the parametrization used. Therefore, the emphasis is shifted: instead of looking for uniform priors, we look for priors that are well-defined on \( P \) and declare them objective. For differentiable parametric models, a construction from Riemannian geometry can be used to define a parameterisation-independent prior (see Jeffreys (1946), (1961) [48, 49] if we interpret the Fisher information as a Riemannian metric on the model (as first proposed by Rao (1945) [75] and extended by Efron (1975) [31]; for an overview, see Amari (1990) [2]) and use the square-root of its determinant as a density with respect to the Lebesgue measure.

**Definition 3.9.** Let \( \Theta \subset \mathbb{R}^d \) be open and let \( \mathcal{P} \) be a dominated model with identifiable, differentiable parametrization \( \Theta \to \mathcal{P} \). Assume that for every \( \theta \in \Theta \), the score-function \( \dot{\ell}_\theta \) is square-integrable with respect to \( P_\theta \). Then Jeffreys prior \( \Pi \) has the square root of the determinant of the Fisher information \( I_\theta = P_\theta \dot{\ell}_\theta \dot{\ell}_\theta^T \) as its density with respect to the Lebesgue measure on \( \Theta \):

\[
d\Pi(\theta) = \sqrt{\det(I_\theta)} \, d\theta.
\]  

(3.7)

Although the expression for Jeffreys prior has the appearance of being parametrization-dependent, the form (3.7) of this prior is the same in any parametrization (a property referred to sometimes as (coordinate-)covariance). In other words, no matter which parametrization we use to calculate \( \Pi \) in \((c.f. (3.7))\), the induced measure \( \Pi' \) on \( \mathcal{P} \) is always the same one. As such, Jeffreys prior is a measure defined on \( \mathcal{P} \) rather than a parametrization-dependent measure.

**Example 3.10.** We calculate the density of Jeffreys prior in the normal model of example 3.8. The score-function with respect to the parameter \( \sigma \) in parametrization \( \phi_2 \) of \( \mathcal{P} \) is given by:

\[
\dot{\ell}_\sigma(X) = \frac{1}{\sigma} \left( \frac{X^2}{\sigma^2} - 1 \right).
\]

The Fisher information (which is a \( 1 \times 1 \)-matrix in this case), is then given by:

\[
I_\sigma = P_\sigma \dot{\ell}_\sigma \dot{\ell}_\sigma = \frac{1}{\sigma^2} P_\sigma \left( \frac{X^2}{\sigma^2} - 1 \right)^2 = \frac{2}{\sigma^2}
\]

Therefore, the density for Jeffreys prior \( \Pi \) takes the form

\[
d\Pi(\sigma) = \frac{\sqrt{2}}{\sigma} \, d\sigma,
\]

for all \( \sigma \in \Theta_2 = (0, 1) \). A similar calculation using the parametrization \( \phi_1 \) shows that, in terms of the parameter \( \tau \), Jeffreys prior takes the form:

\[
d\Pi(\tau) = \frac{1}{\sqrt{2\tau}} \, d\tau,
\]

for all \( \tau \in \Theta_1 = (0, 1) \). That both densities give rise to the same measure on \( \mathcal{P} \) is the assertion of the following lemma.
Lemma 3.11. (Parameterization-independence of Jeffreys prior)
Consider the situation of (3.4) and assume that the parametrizations \( \phi_1 \) and \( \phi_2 \) satisfy the conditions of definition 3.9. In addition, we require that the map \( \phi_1^{-1} \circ \phi_2 : \Theta_2 \rightarrow \Theta_1 \) is differentiable. Then the densities (3.7), calculated in coordinates \( \phi_1 \) and \( \phi_2 \) induce the same measure on \( \mathcal{P} \), Jeffreys prior.

Proof Since the Fisher information can be written as:
\[
I_{\theta_1} = P_{\theta_1}(\dot{\ell}_{\theta_1}^T \dot{\ell}_{\theta_1}),
\]
and the score \( \dot{\ell}_{\theta_1}(X) \) is defined as the derivative of \( \theta_1 \mapsto \log p_{\theta_1}(X) \) with respect to \( \theta_1 \), a change of parametrization \( \theta_2(\theta_2) = (\phi_1^{-1} \circ \phi_2)(\theta_2) \) induces a transformation of the form
\[
I_{\theta_2} = S_{1,2}(\theta_2) I_{\theta_1(\theta_2)} S_{1,2}(\theta_2)^T,
\]
on the Fisher information matrix, where \( S_{1,2}(\theta_2) \) is the total derivative matrix of \( \theta_2 \mapsto \theta_1(\theta_2) \) in the point \( \theta_2 \) of the model. Therefore,
\[
\sqrt{\det I_{\theta_2}} d\theta_2 = \sqrt{\det(S_{1,2}(\theta_2) I_{\theta_1(\theta_2)} S_{1,2}(\theta_2)^T) d\theta_2} = \sqrt{\det(I_{\theta_1(\theta_2)})^2 \det(I_{\theta_1(\theta_2)})} d\theta_2
\]
i.e. the form of the density is such that reparametrization leads exactly to the Jacobian for the transformation of \( d\theta_2 \) to \( d\theta_1 \).

Ultimately, the above construction derives from the fact that the Fisher information \( I_{\theta} \) (or in fact, any other positive-definite symmetric matrix-valued function on the model, e.g. the Hessian of a twice-differentiable, convex function) can be viewed as a Riemann metric on the “manifold” \( \mathcal{P} \). The construction of a measure with Lebesgue density (3.7) is then a standard construction in differential geometry.

Example 3.12. To continue with the normal model of examples 3.8 and 3.10, we note that \( \sigma(\tau) = \sqrt{\tau} \), so that \( d\sigma/d\tau(\tau) = 1/2\sqrt{\tau} \). As a result,
\[
\sqrt{\det I_{\theta_2}} d\theta_2 = \frac{\sqrt{2}}{\sigma} d\sigma = \frac{\sqrt{2}}{\sigma(\tau)} d\tau = \frac{\sqrt{2}}{\sqrt{\tau}} \frac{1}{2\sqrt{\tau}} d\tau = \frac{1}{2\tau} d\tau = \sqrt{\det(I_{\theta_1})} d\theta_1,
\]
which verifies the assertion of lemma 3.11 explicitly.

Other constructions and criteria for the construction of non-informative priors exist: currently very popular is the use of so-called reference priors, as introduced in Lindley (1956) [69] and rediscovered in Bernardo (1979) [12] (see also Berger and Bernardo (1992) [9]). By defining principle, a reference prior is required to maximize the Kullback-Leibler divergence between prior and posterior. To motivate this condition, we have to look at information theory, from which the Kullback-Leibler divergence has emerged as one (popular but by no means unique) way to quantify the notion of the “amount of information” contained in a probability
Conjugate families

distribution. Sometimes called the Shannon entropy, the Kullback-Leibler divergence of a distribution $P$ with respect to the counting measure in discrete probability spaces,

$$S(P) = \sum_{\omega \in \Omega} p(\omega) \log(p(\omega)),$$

can be presented as such convincingly (see Bolzmann (1895, 1898) [22], Shannon (1948) [82]). For lack of a default dominating measure, the argument does not extend formally to continuous probability spaces but is generalized nevertheless. A reference prior $\Pi$ on a dominated, parametrized model $\Theta \to \mathcal{P} : \theta \mapsto P_{\theta}$ for an observation $Y$ is to be chosen such that the Lindley entropy,

$$S_L = \int \int \log\left(\frac{\pi(\theta|Y = y)}{\pi(\theta)}\right) d\Pi(\theta|Y = y) dP^\Pi(y),$$

is maximized. Note that this definition does not depend on the specific parametrization, since the defining property is parametrization independent. Usually, the derivation of a reference prior [12] is performed in the limit where the posterior becomes asymptotically normal, in accordance with the Bernstein-von Mises theorem. Jeffreys prior emerges as a special case of a reference prior.

For an overview of various objective methods of constructing priors, the reader is referred to Kass and Wasserman (1995) [51]. When using non-informative priors, however, the following general warning should be heeded.

Remark 3.13. In many models, non-informative priors, including Jeffries prior and reference priors, are improper.

3.3 Conjugate families

In this section, we consider a type of prior choice that is motivated primarily by mathematical convenience, rather than philosophy or statistical utility. Recall that if we model the data with normal distributions of known variance but unknown location $\theta$ and we supply $\theta$ with a normal prior, then the posterior for $\theta$ is again a normal distribution. Since the calculation of the posterior is tractable, any choice for the parameters of the normal prior can immediately be updated to values for location and variance of the normal posterior upon observation of $Y = y$. Not only does this signify ease of manipulation in calculations with the posterior, it also reduces the computational burden dramatically since numerical integration or simulation from the posterior is no longer necessary.

3.3.1 Basic definitions and some examples

The subject of this section revolves around the following definition.

Definition 3.14. Let $(\mathcal{P}, \mathcal{A})$ be a measurable model for an observation $Y \in \mathcal{Y}$. Let $M$ denote a collection of probability distributions on $(\mathcal{P}, \mathcal{A})$. The set $M$ is called a conjugate
family for the model $\mathcal{P}$, if the posterior based on a prior from $M$ again lies in $M$:

$$\Pi \in M \Rightarrow \Pi(\cdot|Y = y) \in M,$$

(3.8)

for all almost all $y \in \mathcal{Y}$.

(Like before, the phrase “almost all” in the above definition refers to the prior predictive distribution for Bayesians, and to the true $P_0$ for frequentists.) Such structure was first proposed by Raiffa and Schlaifer (1961) [74]. Their method for the prior choice is usually classified as objectivist because it does not rely on subjectivist notions and is motivated without reference to outside factors.

**Remark 3.15.** Often in the literature, a prior is refered to as a “conjugate prior” if the posterior is of the same form. This is somewhat misleading, since it is the family $M$ that is closed under conditioning on the data $Y$, a property that depends on the model and on $M$ itself, but not on the particular $\Pi \in M$.

**Example 3.16.** Consider an experiment in which we observe $n$ independent Bernoulli trials and consider the total number of successes, $Y \sim \text{Bin}(n,p)$ with unknown parameter $p \in [0,1]$, $P_p(Y = k) = \binom{n}{k} p^k (1-p)^{n-k}$.

For the parameter $p$ we choose a prior $p \sim \text{Beta}(\alpha, \beta)$ from the Beta-family, for some $\alpha, \beta > 0$,

$$d\Pi(p) = B(\alpha, \beta) p^{\alpha-1} (1-p)^{\beta-1} dp,$$

where $B(\alpha, \beta) = \Gamma(\alpha + \beta) / (\Gamma(\alpha) \Gamma(\beta))$ normalizes $\Pi$. Then the posterior density with respect to the Lebesgue measure on $[0,1]$ is proportional to:

$$d\Pi(p|Y) \propto p^Y (1-p)^{n-Y} p^{\alpha-1} (1-p)^{\beta-1} dp = p^{\alpha+Y-1} (1-p)^{\beta+n-Y-1} dp,$$

We conclude that the posterior again lies in the Beta-family, with parameters equal to a data-amended version of those of the prior, as follows:

$$\Pi(\cdot|Y) = \text{Beta}(\alpha + Y, \beta + n - Y).$$

So the family of Beta-distributions is a conjugate family for the binomial model. Depending on the available amount of prior information on $\theta$, the prior’s parameters may be chosen on subjective grounds (see figure 2.1 for graphs of the densities of Beta-distributions for various parameter values). However, in the absence thereof, the parameters $\alpha, \beta$ suffer from the same ambiguity that plagues the parameter $\tau^2$ featuring in the example with which we opened this section.

Example 3.16 indicates a strategy to find conjugate families for a given parametrized, dominated model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$. We view densities $y \mapsto p_\theta(y)$ as functions of the outcome $Y = y$ foremost, but they are functions of the parameter $\theta$ as well and their dependence $\theta \mapsto p_\theta(y)$ determines which prior densities $\theta \mapsto \pi(\theta)$ preserve their functional form when multiplied by the likelihood $p_\theta(Y)$ to yield the posterior density.
3.3.2 Exponential families

Although we shall encounter an example of a conjugate family for a non-parametric model in section 3.6, conjugate families are, by and large, part of parametric statistics. Many of these families are so-called exponential families, for which conjugate families of priors can be found readily.

**Definition 3.17.** A dominated collection of probability measures $\mathcal{P} = \{ P_\theta : \theta \in \Theta \}$ is called a $k$-parameter exponential family, if there exists a $k \geq 1$ such that for all $\theta \in \Theta$,

$$p_\theta(x) = \exp\left( \sum_{i=1}^{k} \eta_i(\theta) T_i(x) - B(\theta) \right) h(x),$$  \hspace{1cm} (3.9)

where $h$ and $T_i$, $i = 1, \ldots, k$, are statistics and $B$, $\eta_i$, $i = 1, \ldots, k$ are real-valued functions on $\Theta$.

Any exponential family can be parametrized such that the exponent in (3.9) is linear in the parameter: by the mapping $\Theta \to H : \eta_i = \eta_i(\theta)$ (a bijection if the original parametrization is identifiable), taking $\Theta$ into $H = \eta(\theta)$ and $B$ into $A(\eta) = B(\theta(\eta))$, any exponential family can be rewritten in its so-called canonical form.

**Definition 3.18.** An exponential family $\mathcal{P} = \{ P_\eta : \eta \in H \}$, $H \subset \mathbb{R}^k$ is said to be in its canonical representation, if

$$p_\eta(x) = \exp\left( \sum_{i=1}^{k} \eta_i T_i(x) - A(\eta) \right) h(x).$$ \hspace{1cm} (3.10)

In addition, $\mathcal{P}$ is said to be of full rank if the interior of $H \subset \mathbb{R}^k$ is non-void, i.e. $H \neq \emptyset$.

Although parametric, exponential families are both versatile modelling tools and mathematically tractable; many common models, like the Bernoulli-, normal-, binomial-, Gamma-, Poisson-models, *etcetera*, can be rewritten in the form (3.9). To give an example of a type of parameter that cannot be accommodated in an exponential family, consider models in which the support of model distributions is parameter-dependent, like the family of all uniform distributions on $\mathbb{R}$, or the parameter that describes the domain offset in the Pareto-model. Their statistical practicality stems primarily from the fact that for an exponential family of full rank, the statistics $T_i$, $i = 1, \ldots, k$ are sufficient and complete, enabling the use of the Lehmann-Scheffé theorem for minimal-variance unbiased estimation (see, for instance, Lehmann and Casella (1998) [61]). Their versatility can be understood in many ways, *e.g.* by the Pitman-Koopman-Darmois theorem (see, Jeffreys (1961) [49]), which says that a family of distributions whose support does not depend on the parameter, is exponential, if and only if in the models describing its *i.i.d.* samples, there exist sufficient statistics whose dimension remains bounded asymptotically (i.e. as we let the sample size diverge to infinity).
Example 3.19. The model of all normal distributions \( \mathcal{P} = \{N(\mu, \sigma^2) : \theta \in \mathbb{R}, \sigma^2 > 0\} \) on \( \mathbb{R} \) forms an exponential family. To see this, write \( \theta = (\mu, \sigma^2) \in \Theta = \mathbb{R} \times (0, \infty) \) and rewrite the usual parametrization in the form (3.9), as follows,

\[
p_{\mu, \sigma^2}(x) = \frac{1}{(2\pi)^{1/2}} \frac{1}{\sigma} \exp \left( -\frac{(x - \mu)^2}{2\sigma^2} \right) = \exp \left( -\frac{x^2}{2\sigma^2} + \frac{x\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2} - \frac{1}{2} \log(2\pi) - \frac{1}{2} \log \sigma^2 \right),
\]

and, comparing with (3.9), we read off,

\[
\eta_1(\theta) = \frac{\theta_1}{\theta_2}, \quad \eta_2(\theta) = -\frac{1}{2\theta_2}, \quad B(\theta) = \frac{\theta_1^2}{2\theta_2} + \frac{1}{2} \log \theta_2,
\]

\[
T_1(x) = x, \quad T_2(x) = x^2, \quad h(x) = (2\pi)^{-1/2}.
\]

The map \( \eta : \Theta \rightarrow H : \theta \mapsto (\eta_1, \eta_2)(\theta) \) takes the original parameter into the canonical parameter \( \eta \in H = \mathbb{R} \times (-\infty, 0) \). Note that the inverse of \( \eta \) takes the form,

\[
(\theta_1, \theta_2)(\eta) = \left( -\frac{\eta_1}{2\eta_2}, -\frac{1}{2\eta_2} \right),
\]

from which we deduce that,

\[
A(\eta) = B(\theta(\eta)) = -\frac{\eta_1^2}{4\eta_2} + \frac{1}{2} \log \left( -\frac{1}{2\eta_2} \right),
\]

for the new normalization. Expressed in these new parameters \( \eta \), the density takes the form (3.10). Note that \( H = \mathbb{R} \times (-\infty, 0) \) has non-empty interior, so the normal model is an exponential family of full rank. In case we had started with the model \( \mathcal{P} = \{N(\theta, \theta) : \theta > 0\} \), for example, the analysis would have been largely analogous; however, the latter \( \mathcal{P} \) is not of full rank.

Presently our interest lies with the following theorem which says that if a model \( \mathcal{P} \) constitutes an exponential family, there exists a conjugate family of priors for \( \mathcal{P} \).

**Theorem 3.20.** Let \( \mathcal{P} \) be a model that can be written as an exponential family, c.f. definition 3.17. Then there exists a parametrization of \( \mathcal{P} \) of the form (3.10) and the family of distributions \( \Pi_{\mu, \lambda} \), defined by Lebesgue probability densities

\[
\pi_{\mu, \lambda}(\eta) = K(\mu, \lambda) \exp \left( \sum_{i=1}^{k} \eta_i \mu_i - \lambda A(\eta) \right), \tag{3.11}
\]

(where \( \mu \in \mathbb{R}^k \) and \( \lambda \in \mathbb{R} \) are such that \( 0 < K(\mu, \lambda) < \infty \)), is a conjugate family for \( \mathcal{P} \).

**Proof** It follows from the argument preceding definition 3.18 that \( \mathcal{P} \) can be parametrized as in (3.10). Choosing a prior on \( H \) of the form (3.11), we find that the posterior again takes the form (3.11),

\[
\pi(\eta|X) \propto \exp \left( \sum_{i=1}^{k} \eta_i (\mu_i + T_i(X)) - (\lambda + 1) A(\eta) \right)
\]
Hierarchical priors

(the factor \( h(X) \) arises both in numerator and denominator of (2.6) and is \( \eta \)-independent, so that it cancels). The data-amended versions of the parameters \( \mu \) and \( \lambda \) that emerge from the posterior are therefore given by:

\[
(\mu + T(X), \lambda + 1),
\]

and we conclude that the distributions \( \Pi_{\mu,\lambda} \) form a conjugate family for \( \mathcal{P} \).

From a frequentist perspective, it is worth noting the import of the factorization theorem, which says that the parameter-dependent factor in the likelihood is a function of the data only through the sufficient statistic. Since the posterior is a function of the likelihood, in which data-dependent factors that do not depend on the parameter can be cancelled between numerator and denominator, the posterior is a function of the data \( Y \) only through the sufficient statistic \( T(Y) \). Specifying to the case of exponential families, we note the following conclusion regarding \( L_2 \)-optimality.

**Lemma 3.21.** If \( \mathcal{P} \) is an exponential family of full rank, \( T \) is not just sufficient but also complete for \( \mathcal{P} \), and any point-estimator \( \hat{\theta} \) based on the posterior (e.g. the posterior mean, see definition 2.20) that is unbiased and quadratically integrable (for all \( \theta \in \Theta \)), is optimal in the sense that, for all \( \theta \in \Theta \),

\[
P_\theta(\hat{\theta} - \theta)^2 \leq P_\theta(\hat{\eta} - \theta)^2,
\]

for any other unbiased estimator \( \hat{\eta} \) for \( \theta \).

We omit the proof because it is a direct consequence of the famous Lehmann-Scheffé theorem (see, e.g., Lehmann and Casella (1998) [61]).

### 3.4 Hierarchical priors

Consider again the problem of estimating the mean of a single, normally distributed observation \( Y \) with known variance. The model consists of all normal distributions \( P_\theta = N(\theta, \sigma^2) \), where \( \theta \in \mathbb{R} \) is unknown and \( \sigma^2 > 0 \) is known. Imposing a normal prior on the parameter \( \theta \), \( \Pi = N(0, \tau^2) \), for some choice of \( \tau^2 > 0 \), we easily calculate the posterior distribution,

\[
\Pi(\theta \in A | Y) = N\left(\frac{\tau^2}{\sigma^2 + \tau^2} Y, \frac{\sigma^2 \tau^2}{\sigma^2 + \tau^2}\right)(A),
\]

for every \( A \in \mathcal{B} \). The posterior mean, a point-estimator for \( \theta \), is then given by,

\[
\hat{\theta}(Y) = \frac{\tau^2}{\sigma^2 + \tau^2} Y.
\]

As long as expert knowledge is available in some sufficiently strong form, the subjectivist choice for a certain value of \( \tau^2 \) is well-motivated. However in situations where no prior belief or information on the parameter \( \theta \) is available, or if the parameter itself does not have a clear interpretation, the subjectivist has no answer. Yet a choice for \( \tau^2 \) is required! One may express ignorance about \( \tau^2 \) either by choosing a prior on objectivist grounds, or by considering more and more homogeneous (but still normal) priors by means of the limit \( \tau \to \infty \).
Remark 3.22. From a statistical perspective, however, there exists a better answer to the question regarding $\tau^2$: if $\tau$ is not known, why not estimate its value from the data!

In this section and the next, we consider this answer from the Bayesian and from the frequentist’s angle respectively, giving rise to procedures known as hierarchical Bayesian modelling and empirical Bayesian estimation.

3.4.1 Hyperparameters and hyperpriors

Next, we turn to the Bayesian answer to remark 3.15 which said that parameters of the prior (e.g. $\tau^2$) are to be estimated themselves. Recall that the Bayesian views a parameter to be estimated as just another random variable in the probability model. In case we want to estimate the parameter for a family of priors, then that parameter is to be included in the probability space from the start. Going back to the example with which we started this section, this means that we still use normal distributions $P_\theta = N(\theta, \sigma^2)$ to model the uncertainty in the data $Y$, supply $\theta \in \mathbb{R}$ with a prior $\Pi_1 = N(0, \tau^2)$ and then proceed to choose a another prior $\Pi_2$ for $\tau^2 \in (0, \infty)$:

$$Y|\theta, \tau^2 = Y|\theta \sim P_\theta = N(\theta, \sigma^2), \quad \theta|\tau^2 \sim \Pi_1 = N(0, \tau^2), \quad \tau^2 \sim \Pi_2,$$

Note that the parameter $\tau^2$ has no direct bearing on the model distributions: conditional on $\theta$, $Y$ is independent of $\tau^2$. In a sense, the hierarchical Bayesian approach to prior choice combines subjective and objective philosophies: whereas the subjectivist will make a definite, informed choice for $\tau^2$ and the objectivist will keep himself as uncommitted as possible by striving for uniformity, the choice for a hierarchical prior expresses uncertainty about the value of $\tau^2$ in the form of a probability distribution $\Pi_2$. As such, the hierarchical Bayesian approach allows for intermediate prior choices: if $\Pi_2$ is chosen highly concentrated around one point in the model, resembling a degenerate measure, the procedure will be close to subjective; if $\Pi_2$ is spread widely and is far from degenerate, the procedure will be less biased and closer to objective. Besides interpolating between objective and subjective prior choices, the flexibility gained through introduction of $\Pi_2$ offers a much wider freedom of modelling. In particular, we may add several levels of modelled parameter uncertainty to build up a hierarchy of priors for parameters of priors. Such structures are used to express detailed subjectivist beliefs, much in the way graphical models are used to build intricate dependency structures for observed data. of Bishop (2006) [20]). The origins of the hierarchical approach go back, at least, to Lindley and Smith (1972) [70].

Definition 3.23. Let the data $Y$ be random in $(\mathcal{Y}, \mathcal{B})$. A hierarchical Bayesian model for $Y$ consists of a collection of probability measures $\mathcal{P} = \{P_\theta : \theta \in \Theta_0\}$, with $(\Theta_0, \mathcal{G}_0)$ measurable and endowed with a prior $\Pi : \mathcal{G}_0 \to [0, 1]$ built up in the following way: for some $k \geq 1$, we introduce measurable spaces $(\Theta_i, \mathcal{G}_i), i = 1, 2, \ldots, k$ and conditional priors

$$\mathcal{G}_i \times \Theta_{i+1} \to [0, 1] : (G_i, \theta_{i+1}) \mapsto \Pi_i(G|\theta_{i+1}),$$
Hierarchical priors

for \( i = 1, \ldots, k - 1 \) and a marginal \( \Pi_k : \mathcal{G}_k \to [0,1] \) on \( \Theta_k \). The prior for the original parameter \( \theta \) is then defined by,

\[
\Pi(\theta \in G) = \int_{\Theta_1 \times \cdots \times \Theta_k} \Pi_0(\theta \in G|\theta_1) d\Pi(\theta_1|\theta_2) \cdots d\Pi(\theta_{k-1}|\theta_k) d\Pi_k(\theta_k),
\]

(3.12)

for all \( G \in \mathcal{G}_0 \). The parameters \( \theta_1, \ldots, \theta_k \) and the priors \( \Pi_1, \ldots, \Pi_2 \) are called hyperparameters and their hyperpriors.

Definition 3.23 is very close to the general Bayesian model that incorporates all parameters \( (\theta, \theta_1, \ldots, \theta_k) \) as modelling parameters. What distinguishes hierarchical modelling from the general situation is the dependence structure imposed on the parameters. The parameter \( \theta \) is distinct from the hyperparameters by the fact that conditional on \( \theta \), the data \( Y \) is independent of all hyperparameters \( \theta_1, \ldots, \theta_k \). This distinction is repeated at higher levels in the hierarchy, i.e. levels are separate from one another through the conditional independence of \( \theta_i|\theta_{i+1} \) from \( \theta_{i+2}, \ldots, \theta_k \).

Remark 3.24. The hierarchy indicated in definition 3.23 inherently loses interpretability as we ascend in level. One may be able to give a viable interpretation to the parameter \( \theta \) and to the hyperparameter \( \theta_1 \), but higher-level parameters \( \theta_2, \theta_3, \ldots \) become harder and harder to understand heuristically. Since the interpretation of the hierarchy requires a subjective motivation of the hyperpriors, interpretability of each level is imperative, or left as a non-informative choice. In practice, Bayesian hierarchical models are rarely more than two levels deep (\( k = 2 \)) and the last hyperprior \( \Pi_k \) is often chosen by objective criteria.

3.4.2 Hierarchical prior construction in an example

This subsection considers a more-or-less practical example of a hierarchical prior in some detail.

Example 3.25. We observe the number \( Y \) of surviving offspring from a bird’s litter and aim to estimate the number of eggs the bird has laid: the bird lays \( N \geq 0 \) eggs, distributed according to a Poisson distribution with parameter \( \lambda > 0 \). For the particular species of bird in question, the Poisson rate \( \lambda \) is not known exactly: the uncertainty in \( \lambda \) can be modelled in many ways; here we choose to model it by a Gamma-distribution \( \Gamma(\alpha, \beta) \), where \( \alpha \) and \( \beta \) are chosen to reflect our imprecise knowledge of \( \lambda \) as well as possible. Each of the eggs then comes out, producing a viable chick with known probability \( p \in [0, 1] \), independently. Hence, the total number \( Y \) of surviving chicks from the litter is distributed according to a binomial distribution, conditional on \( N \),

\[
Y|N \sim \text{Bin}(N, p), \quad N|\lambda \sim \text{Poisson}(\lambda), \quad \lambda \sim \Gamma(\alpha, \beta).
\]

The posterior distribution is now obtained as follows: conditional on \( N = n \), the probability of finding \( Y = k \) is binomial,

\[
P(Y = k|N = n) = \binom{n}{k} p^k (1-p)^{n-k},
\]
so Bayes’s rule tells us that the posterior is given by:

\[ P(N = n | Y = k) = \frac{P(N = n)}{P(Y = k)} \binom{n}{k} p^k (1-p)^{n-k}. \]

Since \( \sum_{n \geq 0} P(N = n | Y = k) = 1 \) for every \( k \), the marginal \( P(Y = k) \) (viz. the denominator or normalization factor for the posterior given \( Y = k \)) can be read off once we have the expression for the numerator. We therefore concentrate on the marginal for \( N = n \), \( n \geq 0 \):

\[ P(N = n) = \frac{\Gamma(\alpha + n)}{\Gamma(\alpha)} \frac{1}{n!} \frac{1}{\beta^\alpha} \left( \frac{\beta}{\beta + 1} \right)^{\alpha + n} = \frac{1}{n!} \left( \frac{\beta}{\beta + 1} \right)^n \frac{1}{(\beta + 1)^\alpha} \prod_{l=1}^n (\alpha + l - 1) \quad (3.13) \]

Although not in keeping with the subjective argumentation we insist on in the introduction to this example, for simplicity we consider \( \alpha = \beta = 1 \) and find that in that case,

\[ P(N = n) = (1/2)^{n+1}. \]

The posterior for \( N = n \) given \( Y = k \) then takes the form:

\[ P(N = n | Y = k) = \frac{\binom{n}{k} p^k (1-p)^{n-k}}{\sum_{m \geq 0} \binom{m}{k} p^k (1-p)^{m-k}}. \]

The eventual form of the posterior illustrates that the hierarchy contributes only to the construction of the prior: in case we choose \( \alpha = \beta = 1 \), the posterior we find from the hierarchical approach does not differ from the posterior that we would have found if we had started from the model that incorporates a geometric prior for \( N \),

\[ Y | N \sim \text{Bin}(N, p), \quad N \sim \text{Geo}(1/2). \]

Indeed, even if we leave \( \alpha \) and \( \beta \) free, the marginal distribution for \( N \) we found in (3.13) is none other than the prior (3.12) for this problem.

The hierarchical approach to prior construction allows for greater freedom and a more solid foundation to motivate the choice for certain prior over other possibilities. This point is all the more significant in light of remark 3.1: the motivation of a subjectivist choice for the prior is part of the statistical analysis rather than an external aspect of the procedure. Hierarchical Bayesian modelling helps to refine and justify motivations for subjectivist priors.
3.5 Empirical prior choice

More unexpected is the frequentist perspective on remark 3.15: if $\tau^2$ is an unknown, point-estimate it first and then perform the Bayesian analysis with this point-estimate as a “plug-in” for the unknown $\tau^2$. Critical notes can be placed with the philosophical foundations for this practice, since it appears to combine the methods of two contradictory schools of statistics. Be that as it may, the method is used routinely based on its practicality: eventually, the justification comes from the subjectivist who does not reject frequentist methods to obtain expert knowledge on his parameters, as required in his own paradigm.

Remark 3.26. Good statistical practice dictates that one may not “peek” at the data to decide which statistical method to use for the analysis of the same data. The rationale behind this dictum is that pre-knowledge of the data could bias the analysis. If we take this point strictly, the choice for a prior (read, the point-estimate for $\tau^2$) should not be made on the basis of the same data $Y$ that is to be used later to derive the posterior for $\theta$. If one has two independent realisations of the data, one can be used to choose the prior, (here, by a point-estimate for $\tau^2$) and the other to condition on, in the posterior.

Yet the above “rule” cannot be taken too strictly. Any statistician (and common sense) will tell you that it is crucial for the statistical analysis that one first obtains a certain feeling for the statistical problem by inspection of the data, before making decisions on how to analyse it (to see this point driven to the extreme, read, e.g. Tukey (1977) [86]). Ideally, one would make those decisions based on a sample of the data that is independent of the data used in the analysis proper. This precaution is often omitted, however: for example, it is common practice to use “plug-in” parameters based on the sample $Y$ whenever the need arises, possibly leading to a bias in the subsequent analysis of the same data $Y$ (unless the “plug-in” estimator is independent of all other estimators used, of course).

There are many different ways in which the idea of a prior chosen by frequentist methods is applied, all of which go under the name empirical Bayes. Following Berger [8], we note two types of statistical questions that are especially well suited for application. When we analyse data pertaining to an individual from a larger population and it is reasonable to assume that the prior can be inferred from the population, then one may estimate parameters like $\tau^2$ above from population data and use the estimates in the prior for the individual.

A more sophisticated form of applying empirical Bayesian methods is the use of “plug-in” hyperparameters, estimated by maximum-likelihood applied to the prior predictive distribution (see definition 2.4). Recall that the marginal distribution of the data in the subjectivist Bayesian formulation (c.f. section 2.1) predicts how the data is distributed. This prediction may be reversed to decide which value for the hyperparameter leads to the best explanation of the observed data, where our notion of “best” is based on the likelihood principle.

Denote the data by $Y$ and assume that it takes its values in a measurable space $(\mathcal{Y}, \mathcal{B})$. Denote the model by $\mathcal{P} = \{P_\theta : \theta \in \Theta_0\}$. Consider a family of priors parametrized by a
hyperparameter $\eta \in H$, $\{\Pi_\eta : \eta \in H\}$. For every $\eta$, the prior predictive distribution $P_\eta$ is given by:

$$P_\eta(A) = \int_{\Theta} P_\theta(A) d\Pi_\eta(\theta),$$

for all $A \in \mathcal{B}$, i.e. we obtain a new model for the observation $Y$, given by $\mathcal{M} = \{P_\eta : \eta \in H\}$, contained in the convex hull of the original model $\text{co}(\mathcal{P})$ (under some mild conditions, see lemma 2.23). Note that this new model is parametrized by the hyperparameter; hence if we close our eyes to the rest of the problem and we follow the maximum-likelihood procedure for estimation of $\eta$ in this new model, we find the value of the hyperparameter that best explains the observation $Y$. Assuming that the model $\mathcal{M}$ is dominated, with densities $\{p_\eta : \eta \in H\}$, the maximum-likelihood estimate is found as the point $\hat{\eta}(Y) \in H$ such that

$$p_\eta(Y) = \sup_{\eta \in H} p_\eta(Y).$$

under the assumptions of existence and uniqueness, by the usual methods for maximum-likelihood estimation.

**Definition 3.27.** The estimator $\hat{\eta}(Y)$ is called the ML-II estimator, provided it exists and is unique.

**Remark 3.28.** There is one caveat that applies to the ML-II approach: in case the data $Y$ consists of an i.i.d.-distributed sample, the prior predictive distribution describes the sample as exchangeable, but not i.i.d.! Hence, comparison of prior predictive distributions with the data suffer from the objection raised in remark 2.16. The frequentist who assumes that the true, underlying distribution $P_n^0$ of the sample is i.i.d., has to keep in mind that the ML-II model is misspecified.

### 3.5.1 Model selection with empirical methods

A situation where empirical Bayes is often used, is in model selection: suppose that there are several models $\mathcal{M}_1, \mathcal{M}_2, \ldots$ with priors $\Pi_1, \Pi_2, \ldots$, each of which may serve as a reasonable explanation of the data, depending on an unknown parameter $K \in \{1, 2, \ldots\}$. The choice to use model-prior pair $k$ in the determination of the posterior can only be made after observation (or estimation) of $K$. If $K$ is estimated by frequentist methods, the resulting procedure belongs to the realm of the empirical Bayes methods.

**Example 3.29.** Consider the situation where we are provided with a specimen from a population that is divided into an unknown number of classes. Assume that all we know about the classes is that they occur with equal probabilities in the population. The particular class of our specimen remains unobserved. We perform a real-valued measurement $Y$ on the specimen, which is normally distributed with known variance $\sigma^2$ and an unknown mean $\mu_k \in \mathbb{R}$ that depends on the class $k$. Then $Y$ is distributed according to a discrete mixture of normal
distributions of the form,

\[ Y \sim P_{K; \mu_1, \ldots, \mu_K} = \sum_{k=1}^{K} P_{Y|K=k} P(K = k) = \frac{1}{K} \sum_{k=1}^{K} N(\mu_k, 1), \]

where \( \mu = (\mu_1, \ldots, \mu_K) \in \mathbb{R}^K \) are unknown. For every \( K \geq 1 \), we have a model of the form,

\[ \mathcal{P}_K = \{ P_{K; \mu_1, \ldots, \mu_K} : (\mu_1, \ldots, \mu_K) \in \mathbb{R}^K \} \]

Each of these models can be endowed with a prior \( \Pi_K \) on \( \mathbb{R}^K \), for example, by declaring \( \mu_1, \ldots, \mu_K \) independent and marginally distributed standard normal:

\[ \mu \sim \Pi_K = N(0, I_K). \]

At this point, a Bayesian would choose a hyperprior \( \Pi_2 \) for the discrete hyperparameter \( K \geq 1 \) and proceed to calculate the posterior using all models \( \mathcal{P}_K \), weighed by the prior masses \( \Pi_2(K = k) \) for all \( k \geq 1 \). Alternatively, the Bayesian can use Bayes factors to make a decision as to which value of \( K \) to use, reducing the analysis to a selected, or estimated value for \( K \).

Here, we concentrate on the frequentist approach. The frequentist also aims to select one of the models \( \mathcal{P}_K \): in the empirical Bayes approach, we “point-estimate” which model-prior combination we shall be using to analyse the data, from the choices \( (\mathcal{P}_K, \Pi_K) \), \( K \geq 1 \). In such a case, inspection of the data may reveal which number of classes is most appropriate, if one observes clearly separated peaks in the observations, in accordance with the second point made in remark 3.26. Otherwise, frequentist methods exist to estimate \( K \), for instance from a larger population of specimens. After we have an estimate \( \hat{K} \) for \( K \), we are in a position to calculate the posterior for \( \mu \) based on \( (\mathcal{P}_{\hat{K}}, \Pi_{\hat{K}}) \).

There are two remarks to be made with regard to the estimation of \( K \) from a larger population of specimens: first of all, maximization of the likelihood will always lead to a number of classes in the order of the sample size, simply because the largest number of classes offers the most freedom and hence always provides the best fit to the data. A similar phenomenon arises in regression, where it is called over-fitting, if we allow regression polynomials of arbitrary degree: the MLE will fit the data perfectly by choosing a polynomial of degree in the order of the sample size. Therefore in such questions of model selection, penalized likelihood criteria are employed which favour low-dimensional models over high-dimensional ones, i.e. smaller choices for \( K \) over larger ones. Note that it is not clear, neither intuitively nor mathematically, how the penalty should depend on \( K \), nor which proportionality between penalty and likelihood is appropriate. Well-known standard choices come in the form of the so-called AIC and BIC criteria for model selection [81]: the AIC criterion argues for maximization of the likelihood and subsequent subtraction of twice the dimension of the \( k \)’th parameter space, motivated from information theory and large sample sizes. The BIC criterion weighs the penalty by the logarithm of the sample size, motivated from the large sample limit for the posterior.
The Bayesian faces the same problem when he chooses a prior for \( K \): if he assigns too much prior weight to the higher-dimensional models, his estimators (or, equivalently, the bulk of the resulting posterior’s mass) will get the chance to “run off” to infinity with growing samplesize, indicating inconsistency from over-fitting. Indeed, the correspondence between the frequentist’s necessity for a penalty in maximum-likelihood methods on the one hand, and the Bayesian’s need for a prior expressing sufficient bias for the lower-dimensional model choices on the other, is explained in remark 2.32.

On another sidenote: it is crucial in the example above that all classes are represented in equal proportions. Otherwise identifiability and testability problems arise and persist even after we decide to exclude from the model the vectors \( \mu \) which have \( \mu_i = \mu_j \) for some \( i \neq j \). If one imagines the situation where the number of observations is of the same order as the number of classes, this should come as no surprise.

3.6 Dirichlet priors

The construction of priors on non-parametric models is far from trivial. Broadly, there are two mathematical reasons for this: whereas the usual norm topology on \( \mathbb{R}^k \) is unique (in the sense that all other norm topologies are equivalent, see [71]), infinite-dimensional vector spaces support many different norm topologies and various other topologies besides. Similarly, whereas on \( \mathbb{R}^k \) the (unique shift-invariant) Lebesgue measure provides a solid foundation for the definition of models in terms of densities, no such default uniform dominating measure exists in infinite-dimensional spaces.

Nevertheless examples of probability measures on infinite-dimensional spaces exist. Some of the constructions and properties of the measures they result in, are discussed in detail in Ghosh and Ramamoorthi (2003) [42]. In this section, we look at a class of priors first proposed by Ferguson (1973) [34], which have become known as Dirichlet process priors.

3.6.1 Dirichlet distributions

Let \( \mathcal{X} = \{1, 2, \ldots, k\} \) (with its powerset \( 2^\mathcal{X} \) as a \( \sigma \)-algebra) and consider the collection \( M(\mathcal{X}) \) of all probability measures on \( \mathcal{X} \). Every \( P \in M(\mathcal{X}) \) has a density \( p : \mathcal{X} \to [0, 1] \) (with respect to the counting measure on \( \mathcal{X} \)) and we denote \( p_i = p(i) = P(\{i\}) \), so that for every \( A \in 2^\mathcal{X} \),

\[
P(A) = \sum_{i \in A} p_i.
\]

Therefore, the space \( M(\mathcal{X}) \) can be parametrized as follows,

\[
M(\mathcal{X}) = \{ P : 2^\mathcal{X} \to [0, 1] : \sum_{i=1}^{k} p_i = 1, p_i \geq 0, (1 \leq i \leq k) \},
\]

and is in bijective correspondence with the simplex in \( \mathbb{R}^k \). For reasons to be discussed shortly, we consider the following family of distributions on \( M(\mathcal{X}) \).
Definition 3.30. (Finite-dimensional Dirichlet distribution)

Let $\alpha = (\alpha_1, \ldots, \alpha_k)$ with $\alpha_i > 0$ for all $1 \leq i \leq k$. A stochastic vector $p = (p_1, \ldots, p_k)$ is said to have Dirichlet distribution $D_\alpha$ with parameter $\alpha$, if the density $\pi$ for $p$ satisfies:

$$\pi(p) = \frac{\Gamma(\sum_{i=1}^k \alpha_i)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_k)} p_1^{\alpha_1-1} \cdots p_k^{\alpha_k-1} (1 - \sum_{i=1}^{k-1} p_i)^{\alpha_k-1}$$

If $\alpha_i = 0$ for some $i, 1 \leq \alpha_i \leq k$, then we set $D_\alpha(p_i = 0) = 1$ marginally and we treat the remaining components of $p$ as $(k-1)$-dimensional.

As an example, consider the case where $k = 2$ (so that $p_2 = 1 - p_1$): in that case, the density of the Dirichlet distribution takes the form:

$$\pi(p_1, p_2) = \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1) \Gamma(\alpha_2)} p_1^{\alpha_1-1} (1 - p_1)^{\alpha_2-1},$$

i.e. $p_1$ has a Beta distribution $B(\alpha_1, \alpha_2)$. (Examples of graphs of Beta densities with $\alpha_1 = k+1$, $\alpha_2 = n - k + 1$ for various integer values of $k$ are depicted in figure 2.1). We also note the following two well-known facts on the Dirichlet distribution (proofs can be found in [42]).

Lemma 3.31. (Gamma-representation of $D_\alpha$)

If $Z_1, \ldots, Z_k$ are independent and each marginally distributed according to a $\Gamma$-distribution with parameter $\alpha_i$, i.e.

$$Z_i \sim \Gamma(\alpha_i),$$

for all $1 \leq i \leq k$, then the normalized vector

$$\left( \frac{Z_1}{S}, \ldots, \frac{Z_k}{S} \right) \sim D_\alpha,$$

(3.14)

with $S = \sum_{i=1}^k Z_i$.

Lemma 3.31 shows that we may think of a $D_\alpha$-distributed vector as being composed of $k$ independent, $\Gamma$-distributed components, normalized to form a probability distribution, through division by $S$ in (3.14). This division should be viewed as an $L_1$-projection from the positive cone in $\mathbb{R}^k$ onto the $k-1$-dimensional simplex. The following property can also be viewed as a statement on the effect of a projection on a distribution, this time from the simplex in $\mathbb{R}^k$ to lower-dimensional simplices. It is this property (related to a property called infinite divisibility of the Dirichlet distribution) that motivates the choice for the Dirichlet distribution made by definition 3.30.

Lemma 3.32. Let $\mathcal{X}$ be a finite pointset. If the density $p : \mathcal{X} \rightarrow [0, 1]$ of a distribution $P$ is itself distributed according to a Dirichlet distribution with parameter $\alpha$, $p \sim D_\alpha$, then for any partition $\{A_1, \ldots, A_m\}$ of $\mathcal{X}$, the vector of probabilities $(P(A_1), P(A_2), \ldots, P(A_m))$ has a Dirichlet distribution again,

$$(P(A_1), P(A_2), \ldots, P(A_m)) \sim D_{\alpha'},$$
where the parameter \( \alpha' \) is given by:

\[
(\alpha'_1, \ldots, \alpha'_m) = \left( \sum_{l \in A_1} \alpha_l, \ldots, \sum_{l \in A_m} \alpha_l \right).
\]

(3.15)

The identification (3.15) in lemma 3.32 suggests that we adopt a slightly different perspective on the definition of the Dirichlet distribution: we view \( \alpha \) as a finite measure on \( \mathcal{X} \), so that \( P \sim D_\alpha \), if and only if, for every partition \((A_1, \ldots, A_m)\),

\[
(P(A_1), \ldots, P(A_m)) \sim D(\alpha(A_1), \ldots, \alpha(A_m)).
\]

(3.16)

Property (3.16) serves as the point of departure of the generalization to the non-parametric model, because it does not depend on the finite nature of \( \mathcal{X} \).

**Definition 3.33.** Let \( \mathcal{X} \) be a finite pointset; denote the collection of all probability measures on \( \mathcal{X} \) by \( M(\mathcal{X}) \). The Dirichlet family \( \mathcal{D}(\mathcal{X}) \) is defined to be the collection of all Dirichlet distributions on \( M(\mathcal{X}) \), i.e. \( \mathcal{D}(\mathcal{X}) \) consists of all \( D_\alpha \) with \( \alpha \) a finite measure on \( \mathcal{X} \).

The following property of the Dirichlet distribution describes two independent Dirichlet-distributed quantities in convex combination, which form a new Dirichlet-distributed quantity if mixed by means of an (independent) Beta-distributed parameter.

**Lemma 3.34.** Let \( \mathcal{X} \) be a finite pointset and let \( \alpha_1, \alpha_2 \) be two measures on \((\mathcal{X}, 2^{\mathcal{X}})\). Let \((P_1, P_2)\) be independent and marginally distributed as

\[
P_1 \sim D_{\alpha_1}, \quad P_2 \sim D_{\alpha_2}.
\]

Furthermore, let \( \lambda \) be independent of \( P_1, P_2 \) and marginally distributed according to \( \lambda \sim B(\alpha_1(\mathcal{X}), \alpha_2(\mathcal{X})) \). Then the convex combination \( \lambda P_1 + (1 - \lambda) P_2 \) again has a Dirichlet distribution with base measure \( \alpha_1 + \alpha_2 \):

\[
\lambda P_1 + (1 - \lambda) P_2 \sim D_{\alpha_1 + \alpha_2}.
\]

Many other properties of the Dirichlet distribution could be considered here, most notably the so-called *tail-free property* and *neutrality to the right* (see [42]). We do not provide details because both are rather technical and we do not use them in following chapters, but the reader should be aware of their existence because some authors use them extensively.

A most important property of the family of Dirichlet distributions is its conjugacy for the full non-parametric model.

**Theorem 3.35.** Let \( \mathcal{X} \) be a finite pointset; let \( X_1, \ldots, X_n \) denote an i.i.d. sample of observations taking values in \( \mathcal{X} \). The Dirichlet family \( \mathcal{D}(\mathcal{X}) \) is a conjugate family: if the prior equals \( D_\alpha \), the posterior equals \( D_{\alpha + nP_n} \).
Proof Since $\mathcal{X}$ is finite ($\#(\mathcal{X}) = k$), $M(\mathcal{X})$ is dominated (by the counting measure), so the posterior can be written as in (2.13). The likelihood takes the form:

$$P \mapsto \prod_{i=1}^{n} p(X_i) = \prod_{l=1}^{k} p_l^{n_l},$$

where $n_l = \#\{X_i = l : 1 \leq i \leq n\}$. Multiplying by the prior density for $\Pi = D_\alpha$, we find that the posterior density is proportional to,

$$\pi(p_1, \ldots, p_k | X_1, \ldots, X_n) \propto \pi(p_1, \ldots, p_k) \prod_{i=1}^{n} p_{X_i} \propto \prod_{l=1}^{k} p_l^{n_l} \left(1 - \sum_{i=1}^{k-1} p_i\right)^{\alpha_k - 1} = \prod_{l=1}^{k-1} p_l^{\alpha_l + n_l - 1} \left(1 - \sum_{i=1}^{k-1} p_i\right)^{\alpha_k},$$

which is again a Dirichlet density, but with changed base measure $\alpha$. Since the posterior is a probability distribution, we know that the normalization factor follows suit. Noting that we may view $n_l$ as the density of the measure $n \mathbb{P}_n$ since

$$n_l = \sum_{i=1}^{n} 1\{X_i = l\} = n \mathbb{P}_n 1\{X = l\},$$

we complete the argument. □

3.6.2 The Dirichlet process

Next we consider the Dirichlet process prior, a probability measure on the full non-parametric model for a measurable space $(\mathcal{X}, \mathcal{B})$. For the sake of simplicity, we assume that $\mathcal{X} = \mathbb{R}$ and $\mathcal{B}$ is the Borel $\sigma$-algebra on $\mathbb{R}$. We denote the collection of all probability measures on $(\mathbb{R}, \mathcal{B})$ by $M(\mathbb{R}, \mathcal{B})$. We consider the collection of random quantities \(\{P(A) : A \in \mathcal{B}\}\) and impose two straightforward conditions on its finite-dimensional marginals. The Kolmogorov existence theorem (see theorem A.30) then guarantees existence of a stochastic process with finitely additive sample path $P : \mathcal{B} \rightarrow [0, 1]$. Said straightforward conditions are satisfied if we choose the finite-dimensional marginal distributions to be (finite-dimensional) Dirichlet distributions (3.16). Also by this choice, $\sigma$-additivity of $P$ can be guaranteed. The resulting process on the space of all probability measures on $(\mathcal{X}, \mathcal{B})$ is called the called the Dirichlet process and the associated probability measure $\Pi$ is called the Dirichlet process prior.

**Theorem 3.36. (Existence of the Dirichlet process)**

Given a finite measure $\alpha$ on $(\mathbb{R}, \mathcal{B})$, there exists a probability measure $D_\alpha$ on $M(\mathbb{R}, \mathcal{B})$ (called the Dirichlet process prior with parameter $\alpha$) such that for $P \sim D_\alpha$ and every $\mathcal{B}$-measurable partition $(B_1, \ldots, B_k)$ of $\mathbb{R}$,

$$\left(P(B_1), \ldots, P(B_k)\right) \sim D_{(\alpha(B_1), \ldots, \alpha(B_k))}.$$

(3.17)
Proof Let \( k \geq 1 \) and \( A_1, \ldots, A_k \in \mathcal{B} \) be given. Through the indicators \( 1_{A_i} \) for these sets, we define \( 2^k \) new sets
\[
1_{B_{\nu_1 \ldots \nu_k}} = \prod_{i=1}^{k} 1_{A_i}^\nu_i (1 - 1_{A_i})^{1-\nu_i},
\]
where \( \nu_1, \ldots, \nu_k \in \{0,1\} \). Then the collection \( \{B_{\nu_1 \ldots \nu_k} : \nu_i \in \{0,1\}, 1 \leq i \leq k\} \) forms a partition of \( \mathbb{R} \). For the \( P \)-probabilities corresponding to this partition, we assume finite-dimensional marginals
\[
(P(B_{\nu_1 \ldots \nu_k}) : \nu_i \in \{0,1\}, 1 \leq i \leq k) \sim \Pi_{B_{\nu_1 \ldots \nu_k} : \nu_i \in \{0,1\}, 1 \leq i \leq k}.
\]
The distribution of the vector \((P(A_1), \ldots, P(A_k))\) then follows from the definition:
\[
P(A_i) = \sum_{\{i: \nu_i = 1\}} P(B_{\nu_1 \ldots \nu_k}),
\]
for all \( 1 \leq i \leq k \). This defines marginal distributions for all finite subsets of \( \mathcal{B} \), as needed in theorem A.30. To define the underlying probability space \((\Omega, \mathcal{F}, \Pi)\) we now impose two conditions.

(F1) With \( \Pi \)-probability one, the empty set has \( P \)-measure zero:
\[
\Pi(P(\emptyset) = 0) = 1.
\]

(F2) Let \( k, k' \geq 1 \) be given. If \((B_1, \ldots, B_k)\) is a partition and \((B'_1, \ldots, B'_{k'})\) a refinement thereof, with
\[
B_1 = \bigcup_{i=1}^{r_1} B'_i, \quad \ldots, \quad B_k = \bigcup_{i=r_{k-1}+1}^{k'} B'_i,
\]
(for certain \( r_1 < \ldots < r_{k-1} \)), then we have the following equality in distribution:
\[
\mathcal{L}\left(\sum_{i=1}^{r_1} P(B'_i), \ldots, \sum_{i=r_{k-1}+1}^{k'} P(B'_i)\right) = \mathcal{L}(P(B_1), \ldots, P(B_k)).
\]
Condition (F1) ensures that if \((A_1, \ldots, A_k)\) is itself a partition of \( \mathbb{R} \), the above construction does not lead to a contradiction. Condition (F2) ensures finite additivity of \( P \) with prior probability one, \( i.e. \) for any \( A, B, C \in \mathcal{B} \) such that \( A \cap B = \emptyset \) and \( A \cup B = C \),
\[
\Pi\left( P(A) + P(B) = P(C) \right) = 1.
\]
Ferguson (1973,1974) [34, 35] has shown that conditions (F1) and (F2) imply that Kolmogorov’s consistency conditions (K1) and (K2) (see section A.4) are satisfied. As we have seen in the first part of this section, if we impose the Dirichlet distribution:
\[
(P(B_{\nu_1 \ldots \nu_k}) : \nu_i \in \{0,1\}, 1 \leq i \leq k) \sim D_{\{\alpha(B_{\nu_1 \ldots \nu_k}) : \nu_i \in \{0,1\}, 1 \leq i \leq k\}},
\]
and \( \alpha \) is a measure on \( \mathcal{B} \), condition (F2) is satisfied. Combining all of this, we conclude that there exists a probability space \((\Omega, \mathcal{F}, \Pi)\) on which the stochastic process \( \{P(A) : A \in \mathcal{B}\} \)
can be represented with finite dimensional marginals c.f. (3.19). Lemma 3.37 shows that \( \Pi(P \in M(\mathbb{R}, \mathcal{B})) = 1 \), completing the proof.

The last line in the above proof may require some further explanation: \( P \) is merely the sample-path of our stochastic process. The notation \( P(A) \) suggests that \( P \) is a probability measure, but all we have shown up to that point, is that (F1) and (F2) imply that \( P \) is a finitely additive set-function such that:

\[
\Pi(P(B) \in [0, 1]) = 1,
\]

with \( \Pi \)-probability equal to one. What remains to be demonstrated is \( \Pi \)-almost-sure \( \sigma \)-additivity of \( P \).

**Lemma 3.37.** If \( \Pi \) is a Dirichlet process prior \( D_\alpha \) on \( M(\mathcal{B}) \),

\[
\Pi\left( P \text{ is } \sigma\text{-additive} \right) = 1.
\]

**Proof** Let \((A_n)_{n \geq 1}\) be a sequence in \( \mathcal{B} \) that decreases to \( \emptyset \). Since \( \alpha \) is \( \sigma \)-additive, \( \alpha(A_n) \to \alpha(\emptyset) = 0 \). Therefore, there exists a subsequence \((A_{n_j})_{j \geq 1}\) such that \( \sum_j \alpha(A_{n_j}) < \infty \). For fixed \( \epsilon > 0 \), using Markov’s inequality first,

\[
\sum_{j \geq 1} \Pi(P(A_{n_j}) > \epsilon) \leq \sum_{j \geq 1} \frac{1}{\epsilon} \int P(A_{n_j}) d\Pi(P) = \frac{1}{\epsilon} \sum_{j \geq 1} \frac{\alpha(A_{n_j})}{\alpha(\mathbb{R})} < \infty,
\]

according to lemma 3.38. From the Borel-Cantelli lemma (see lemma A.13), we see that

\[
\Pi(\limsup_{j \to \infty} \{P(A_{n_j}) > \epsilon\}) = \Pi\left( \bigcap_{j \geq 1} \bigcup_{j \geq J} \{P(A_{n_j}) > \epsilon\} \right) = 0,
\]

which shows that \( \lim_j P(A_{n_j}) = 0 \), \( \Pi \)-almost-surely. Since, by \( \Pi \)-almost-sure finite additivity of \( P \),

\[
\Pi\left( P(A_n) \geq P(A_{n+1}) \geq \ldots \right) = 1,
\]

we conclude that \( \lim_n P(A_n) = 0 \), \( \Pi \)-almost-surely. By the continuity theorem for measures (see theorem A.10 and the proof in [54], theorem 3.2), \( P \) is \( \sigma \)-additive \( \Pi \)-almost-surely.

The proof of lemma 3.37 makes use of the following lemma, which establishes the basic properties of the Dirichlet process prior.

**Lemma 3.38.** Let \( \alpha \) be a finite measure on \( (\mathbb{R}, \mathcal{B}) \) and let \( \{P(A) : A \in \mathcal{B}\} \) be the associated Dirichlet process with distribution \( D_\alpha \). Let \( B \in \mathcal{B} \) be given.

(i) If \( \alpha(B) = 0 \), then \( P(B) = 0 \), \( \Pi \)-a.s.

(ii) If \( \alpha(B) > 0 \), then \( P(B) > 0 \), \( \Pi \)-a.s.

(iii) The expectation of \( P \) under \( D_\alpha \) is given by

\[
\int P(B) dD_\alpha(P) = \frac{\alpha(B)}{\alpha(\mathbb{R})},
\]
Proof Let $B \in \mathcal{B}$ be given. Consider the partition $(B_1, B_2)$ of $\mathbb{R}$, where $B_1 = B$, $B_2 = \mathbb{R} \setminus B$. According to (3.17),

$$(P(B_1), P(B_2)) \sim D_{(\alpha(B), \alpha(\mathbb{R}) - \alpha(B))},$$

so that $P(B) \sim B(\alpha(B), \alpha(\mathbb{R}) - \alpha(B))$. Stated properties then follow from the properties of the Beta-distribution.

This concludes the proof for the existence of Dirichlet processes and the associated priors.

One may then wonder what is the nature of the prior we have constructed. As it turns out, the Dirichlet process prior has some remarkable properties.

Lemma 3.39. (Support of the Dirichlet process prior)

Consider $M(\mathbb{R}, \mathcal{B})$, endowed with the topology of weak convergence. Let $\alpha$ be a finite measure on $(\mathbb{R}, \mathcal{B})$. The support of $D_\alpha$ is given by

$$M_\alpha(\mathbb{R}, \mathcal{B}) = \{ P \in M(\mathbb{R}, \mathcal{B}) : \text{supp}(P) \subset \text{supp}(\alpha) \}.$$

In fact, we can be more precise, as shown in the following lemma.

Lemma 3.40. Let $\alpha$ be a finite measure on $(\mathbb{R}, \mathcal{B})$ and let $\{ P(A) : A \in \mathcal{B} \}$ be the associated Dirichlet process with distribution $D_\alpha$. Let $Q \in M(\mathbb{R}, \mathcal{B})$ be such that $Q \ll \alpha$. Then, for any $m \geq 1$ and $A_1, \ldots, A_m \in \mathcal{B}$ and $\epsilon > 0$,

$$D_\alpha \{ P \in M(\mathbb{R}, \mathcal{B}) : |P(A_i) - Q(A_i)| < \epsilon, 1 \leq i \leq m \} > 0.$$

Proof The proof of this lemma can be found in [42], theorem 3.2.4.

So if we endow $M(\mathbb{R}, \mathcal{B})$ with the (slightly stronger) topology of pointwise convergence (see definition A.41), the support of $D_\alpha$ remains large, consisting of all $P \in M(\mathbb{R}, \mathcal{B})$ that are dominated by $\alpha$.

The following property reveals a most remarkable characterization of Dirichlet process priors: the subset $D(\mathbb{R}, \mathcal{B})$ of all finite convex combinations of Dirac measures (see example A.12) receives prior mass equal to one.

Lemma 3.41. Let $\alpha$ be a finite measure on $(\mathbb{R}, \mathcal{B})$ and let $\{ P(A) : A \in \mathcal{B} \}$ be the associated Dirichlet process with distribution $D_\alpha$. Then,

$$D_\alpha \{ P \in D(\mathbb{R}, \mathcal{B}) \} = 1.$$

Proof The proof of this lemma can be found in [42], theorem 3.2.3.

The above phenomenon leads to problems with support or convergence in stronger topologies (like total variation or Hellinger topologies) and with regard to the Kullback-Leibler criteria that govern asymptotic concentration of posterior mass. Generalizing this statement somewhat, we may infer from the above that the Dirichlet process prior is not suited to (direct) estimation of densities. Although clearly dense enough in $M(\mathbb{R}, \mathcal{B})$ in the topology of weak...
convergence, the set $D(\mathbb{R}, \mathcal{B})$ may be rather sparse in stronger topologies! (Notwithstanding the fact that mixture models with a Dirichlet process prior for the mixing distribution can be (minimax) optimal for the estimation of mixture densities [41].)

**Lemma 3.42.** Let $\alpha$ be a finite measure on $(\mathbb{R}, \mathcal{B})$ and let $\{P(A) : A \in \mathcal{B}\}$ be the associated Dirichlet process with distribution $D_{\alpha}$. Let $g : \mathbb{R} \to \mathbb{R}$ be non-negative and Borel-measurable. Then,

$$\int_{\mathbb{R}} g(x) d\alpha(x) < \infty \iff \int_{\mathbb{R}} g(x) dP(x) < \infty, \quad (D_{\alpha} - a.s.).$$

**Proof** Add proof! □

Perhaps the most important result of this section is the fact that the family of Dirichlet process priors on $M(\mathbb{R}, \mathcal{B})$ is a conjugate family for the full, non-parametric model on $(\mathbb{R}, \mathcal{B})$, as stated in the following theorem.

**Theorem 3.43.** Let $X_1, X_2, \ldots$ be an i.i.d. sample of observations in $\mathbb{R}$. Let $\alpha$ be a finite measure on $(\mathbb{R}, \mathcal{B})$ with associated Dirichlet process prior $\Pi = D_{\alpha}$. For any measurable $C \subset M(\mathbb{R}, \mathcal{B})$,

$$\Pi(P \in C \mid X_1, \ldots, X_n) = D_{\alpha + nP_n}(C),$$

i.e. the posterior is again a Dirichlet process distribution, with base measure $\alpha + nP_n$.

**Proof** The proof is a direct consequence of theorem 3.35 and the fact that equality of two measures on a generating ring implies equality on the whole $\sigma$-algebra. (Cylindersets generate the relevant $\sigma$-algebra and for cylindersets, theorem 3.35 asserts equality.) □

**Example 3.44.** Let $X_1, X_2, \ldots$ be an i.i.d. sample of observations in $\mathbb{R}$. Let $\alpha$ be a finite measure on $(\mathbb{R}, \mathcal{B})$ with associated Dirichlet process prior $\Pi = D_{\alpha}$. Let $B \in \mathcal{B}$ be given. The expectation of $P(B)$ under the prior distribution equals,

$$\int P(B) dD_{\alpha}(P) = \frac{\alpha(B)}{\alpha(\mathbb{R})}, \quad (3.20)$$

the measure of $B$ under $\alpha$ normalized to be a probability measure (which we denote by $P_\alpha(B)$).

The posterior mean (see definition 2.20), is then given by:

$$\int P(B) d\Pi(P \mid X_1, \ldots, X_n) = \int P(B) dD_{\alpha + nP_n}(P) = \frac{(\alpha + nP_n)(B)}{(\alpha + nP_n)(\mathbb{R})} \frac{\alpha(\mathbb{R})}{\alpha(\mathbb{R}) + n} P_\alpha(B) + \frac{n}{\alpha(\mathbb{R}) + n} P_n(B),$$

$P_0^0$-almost-surely. Defining $\lambda_n = \alpha(\mathbb{R})/(\alpha(\mathbb{R}) + n)$, we see that the posterior mean $\hat{P}_n$ can be viewed as a convex combination of the prior mean distribution and the empirical distributions,

$$\hat{P}_n = \lambda_n P_\alpha + (1 - \lambda_n)P_n,$$

$P_0^0$-almost-surely. As a result, we see that

$$\|\hat{P}_n - P_n\|_{TV} = \lambda_n\|P_\alpha - P_n\| \leq \lambda_n,$$
Choice of the prior

$P_0^n$-almost-surely. Since $\lambda_n \to 0$ as $n \to \infty$, the difference between the sequence of posterior means $(\hat{P}_n)_{n \geq 1}$ and the sequence of empirical measures $(P_n)_{n \geq 1}$ converges to zero in total variation as we let the sample size grow to infinity. Generalizing likelihood methods to non-dominated models, Dvoretzky, Kiefer and Wolfowitz (1956) [30] have shown that the empirical distribution $P_n$ can be viewed as the non-parametric maximum-likelihood estimator (usually abbreviated NPMLE). This establishes (an almost-sure form of) consistency for the posterior mean, in the sense that it has the same point of convergence as the NPMLE.

Remark 3.45. The above example provides the subjectivist with a guideline for the choice of the base measure $\alpha$. More particularly, equality (3.20) says that the prior predictive distribution equals the (normalized) base measure $\alpha$. In view of the fact that subjectivists should choose the prior to reflect their prior "beliefs", $\alpha$ should therefore be chosen such that it assigns relatively high mass to sets $B \in \mathcal{B}$ that are believed to be probable.

3.7 Exercises

Exercise 3.1. A proper Jeffreys prior

Let $X$ be a random variable, distributed $\text{Bin}(n;p)$ for known $n$ and unknown $p \in (0,1)$. Calculate Jeffreys prior for this model, identify a standard family of probability distributions that this prior would belong to, if it were normalized as a probability distribution.

Exercise 3.2. Jeffreys and uniform priors

Let $\mathcal{P}$ be a model parametrized according to some mapping $\Theta \to \mathcal{P} : \theta \mapsto P_\theta$. Assuming differentiability of this map, Jeffreys prior $\Pi$ takes the form (3.7). In other parametrizations, the form of this expression remains the same, but the actual dependence on the parameter changes. This makes it possible that there exists another parametrization of $\mathcal{P}$ such that Jeffreys prior is equal to the uniform prior. We shall explore this possibility below.

For each of the following models in their 'standard' parametrizations $\theta \mapsto P_\theta$, find a parameter $\eta \in H$, $\eta = \eta(\theta)$, such that the Fisher information $I_\eta$, expressed in terms of $\eta$, is constant.

a. Find $\eta$ for the model of all Poisson distributions.

b. In the cases $\alpha = 1, 2, 3$, find $\eta$ for the model $\mathcal{P}$ consisting of all $\text{Gamma}(\alpha, \theta)$-distributions, with $\theta \in (0, \infty)$.

c. Find $\eta$ for the model $\mathcal{P}$ of all $\text{Bin}(n; \theta)$ distributions, where $n$ is known and $\theta \in (0,1)$.

Note that if the Fisher information $I_\eta$ is constant, Jeffreys prior is uniform. Therefore, if a parametrization like above exists and $H$ is unbounded, Jeffreys prior is improper (in all parametrizations).

Exercise 3.3. Optimality of unbiased Bayesian point estimators

Let $\mathcal{P}$ be a dominated, parametric model, parametrized identifiably by $\Theta \to \mathcal{P} : \theta \mapsto P_\theta$, for
some $\Theta \subset \mathbb{R}^k$. Assume that $(X_1, \ldots, X_n) \in \mathcal{X}^n$ form an i.i.d. sample from a distribution $P_0 = P_{\theta_0} \in \mathcal{P}$, for some $\theta_0 \in \Theta$. Let $\Pi$ be a prior on $\Theta$ and denote the posterior by $\Pi(\cdot | X_1, \ldots, X_n)$. Assume that $T: \mathcal{X}^n \to \mathbb{R}^m$ is a sufficient statistic for the model $\mathcal{P}$.

a. Use the factorization theorem to show that the posterior depends on the data only through the sufficient statistic $T(X_1, \ldots, X_n)$.

b. Let $\hat{\theta}_n : \mathcal{X}^n \to \Theta$ denote a point-estimator derived from the posterior. Use a. above to argue that there exists a function $\tilde{\theta}_n : \mathbb{R}^m \to \Theta$, such that,

$$\hat{\theta}_n(X_1, \ldots, X_n) = \tilde{\theta}_n(T(X_1, \ldots, X_n)).$$

Bayesian point-estimators share this property with other point-estimators that are derived from the likelihood function, like the maximum-likelihood estimator and penalized versions thereof.

Next, assume that $T$ is complete, that $P_n^0(\hat{\theta}_n)^2 < \infty$ and that $\hat{\theta}_n$ is unbiased, i.e. $P_n^0\hat{\theta}_n = \theta_0$.

c. Apply the Lehmann-Scheffé theorem to prove that, for any other unbiased estimator $\hat{\theta}'_n : \mathcal{X}^n \to \Theta$,

$$P_n^0(\hat{\theta}_n - \theta_0)^2 \leq P_n^0(\hat{\theta}'_n - \theta_0)^2.$$

The message of this exercise is, that Bayesian point-estimators that happen to be unbiased and quadratically integrable, are automatically $L_2$-optimal in the class of all unbiased estimators for $\theta$. They share this remarkable property with maximum-likelihood estimators.

**Exercise 3.4. Conjugate model-prior pairs**

In this exercise, conjugate model-prior pairs $(\mathcal{P}, \Pi)$ are provided. In each case, we denote the parameter we wish to estimate by $\theta$ and assume that other parameters have known values. Let $X$ denote a single observation.

In each case, derive the posterior distribution to prove conjugacy and identify the $X$-dependent transformation of parameters that takes prior into posterior.

a. $X|\theta \sim N(\theta, \sigma^2)$ and $\theta \sim N(\mu, \tau^2)$.

b. $X|\theta \sim \text{Poisson}(\theta)$ and $\theta \sim \Gamma(\alpha, \beta)$.

c. $X|\theta \sim \Gamma(\rho, \theta)$ and $\theta \sim \Gamma(\alpha, \beta)$.

d. $X|\theta \sim \text{Bin}(n; \theta)$ and $\theta \sim \beta(\alpha, \beta)$.

e. $X|\theta \sim N(\mu, \theta^{-1})$ and $\theta \sim \Gamma(\alpha, \beta)$.

f. $X|\theta_1, \ldots, \theta_k \sim M(n; \theta_1, \ldots, \theta_k)$ and $\theta \sim D_\alpha$, where $M$ denotes the multinomial distribution for $n$ observations drawn from $k$ classes with probabilities $\theta_1, \ldots, \theta_k$ and $D_\alpha$ is a Dirichlet distribution on the simplex in $\mathbb{R}^k$ (see definition 3.30).
Exercise 3.5. In this exercise, we generalize the setup of example 3.25 to multinomial rather than binomial context. Let \( k \geq 1 \) be known. Consider an observed random variable \( Y \) and an unobserved \( N = 1, 2, \ldots, \) such that, conditionally on \( N, Y \) is distributed multinomially over \( k \) classes, while \( N \) has a Poisson distribution with hyperparameter \( \lambda > 0, \)

\[
Y|N \sim M_k(N; p_1, p_2, \ldots, p_k), \quad N \sim \text{Poisson}(\lambda). 
\]

Determine the prior predictive distribution of \( Y \), as a function of the hyperparameter \( \lambda \).

Exercise 3.6. Let \( X_1, \ldots, X_n \) form an i.i.d. sample from a Poisson distribution \( \text{Poisson}(\theta) \) with unknown \( \theta > 0 \). As a family of possible priors for the Bayesian analysis of this data, consider exponential distributions \( \theta \sim \Pi_\lambda = \text{Exp}(\lambda) \), where \( \lambda > 0 \) is a hyperparameter.

a. Calculate the prior predictive distribution for \( X \).

b. Give the ML-II estimate \( \hat{\lambda} \) for \( \lambda \).

c. With the estimated hyperparameter, give the posterior distribution \( \theta|X_1, \ldots, X_n \).

d. Calculate the posterior mean. Compare its data-dependence to that of the posterior mean we would have obtained if we had not made an empirical choice for the hyperparameter, but a fixed choice.

Exercise 3.7. Let \( X_1, \ldots, X_n \) form an i.i.d. sample from a binomial distribution \( \text{Bin}(N; p) \), for known \( N \) and unknown \( p \in [0, 1] \). For the parameter \( p \) we take a prior \( p \sim \beta(\alpha, \beta) \) with hyperparameters \( \alpha, \beta > 0 \).

a. Show that the family of \( \beta \)-distributions is conjugate for binomial data.

b. Using (standard expressions for) the expectation and variance of \( \beta \)-distributions, give the posterior mean and variance in terms of the original \( \alpha \) and \( \beta \) chosen for the prior and the data.

c. Calculate the prior predictive distribution and discuss the steps one would perform in the ML-II procedure to estimate \( p \).
Chapter 4

The Bernstein-von Mises theorem

Given an infinite i.i.d. sample $X_1, X_2, \ldots$ drawn from $P_0$ and a model $\mathcal{P}$, an estimation procedure prescribes a sequence of estimates $\hat{P}_n \in \mathcal{P}$, each calculated using only the first $n$ observations. More generally, any statistical procedure can be indexed by the size $n$ of the sample used to calculate it, leading to sequences of (parameter) estimates, tests, confidence regions, etcetera. Properties of such sequences reflect the behaviour of the estimation procedure with growing sample-size. An intuitively reasonable requirement of any estimation procedure is a property known as consistency: the sequence $\hat{P}_n$ approaches the true distribution $P_0$ to within arbitrary precision if the sample on which the estimation is based is made large enough. Similarly, samples of arbitrarily large size should enable one to test with power arbitrarily close to one and define arbitrarily small confidence regions. Further analysis of a consistent sequence $\hat{P}_n$ concerns the (suitably rescaled) distribution of the estimator-sequence around its point of convergence. The mathematical formulation of these concepts is based on so-called limit theorems, which describe the behaviour of an estimation procedure in the limit that the number of measurements goes to infinity.

The study of the asymptotic regime of an estimation procedure is interesting for two reasons. First of all, asymptotic results provide approximations to exact values; finite-sample calculations often become intractable, even in relatively simple (parametric) situations. However, the analysis of the large-sample limit is often still possible when finite-sample are hard to carry out exactly. The answer obtained in the large-sample limit may then be used as an approximation to the finite-sample answer (asymptotic confidence intervals are a good example). Secondly, if we have several possible estimation procedures available for a certain problem, asymptotic behaviour provides us with the means to compare their performance for large samples. Of course, the first performance criterion is consistency. To choose between two consistent procedures, one may consider rate of convergence and properties of the limit distribution characterising the degree of concentration (like asymptotic variance or asymptotic risk).
In this chapter, we consider the asymptotic behaviour of the Bayesian procedure in cases where the model is parametric and depends on the parameter in a smooth fashion. Certainly the most useful reference for this subject is Le Cam and Yang (1990) [68] and a version of the Bernstein-von Mises theorem can also be found in Le Cam (1986) [66]. A more readable version of the argument can be found in chapter 10 of van der Vaart (1998) [87].

4.1 Efficiency in smooth parametric models

In this section we consider estimation in smooth, parametric models and formulate a theorem that characterizes efficiency, a notion of optimality in the class of all so-called regular estimators. The aim of this lecture is to prepare for the Bernstein-von Mises theorem, which asserts that posterior distributions in smooth parametric models tend to follow efficient estimators and lead to efficient inferential conclusions.

Recall from chapter 2 that at the conceptual level, the Bayesian posterior plays a role that is analogous to that of the frequentist sampling distribution: it is a distribution on the model or on its parameter space, supposedly informative at the inferential level. In the following, we shall not be too strict in Bayesian, subjectivist orthodoxy and interpret the posterior as a frequentist device, in a role very close conceptually to that of the sampling distribution of an estimator used above.

From that perspective it is natural to wonder whether credible sets and confidence sets have anything to do with each other. Since they are conceptually so close, could it be that they are close also mathematically, at least, when explained as frequentist devices? In this section we discuss efficiency of (point-)estimation culminating in the so-called convolution theorem, which provides a notion of asymptotic optimality at the inferential level (for the class of all regular estimators). In the next lecture, we switch to the Bayesian leg of the story and show that a correspondence between confidence sets and credible sets does exist asymptotically in smooth, parametric models.

The so-called Bernstein-von Mises theorem (see theorem 4.11 below [68]) not only demonstrates asymptotic equivalence of credible sets and confidence sets, it also shows that the relevant sets are optimal in the sense that they are associated with so-called efficient estimators. Essential to the development of the optimality theory are two concepts: differentiability of the model and regularity of the estimator. Combined, these two properties lead to a notion of optimality comparable to estimators that achieve optimality within the family of unbiased estimators in the Cramér-Rao sense.

4.1.1 Optimality in smooth, parametric estimation problems

The concept of efficiency has its origin in Fisher’s 1920’s claim of asymptotic optimality of the maximum-likelihood estimator in differentiable parametric models. Here, optimality of the ML estimate means that they are consistent, achieve optimal $n^{-1/2}$ rate of convergence
and possessed a asymptotic sampling distribution of minimal variance. In 1930’s and –40’s, Fisher’s ideas on optimality in differentiable models were sharpened and elaborated upon. To illustrate, consider the following classical result from M-estimation (which can be found as theorem 5.23 in [87]).

**Theorem 4.1.** Let \( \Theta \) be open in \( \mathbb{R}^k \) and assume that \( \mathcal{P} \) is characterized by densities \( p_\theta : \mathcal{X} \rightarrow \mathbb{R} \) such that \( \theta \mapsto \log p_\theta(x) \) is differentiable at \( \theta_0 \) for all \( x \in \mathcal{X} \), with derivative \( \ell_\theta(x) \). Assume that there exists a function \( I : \mathcal{X} \rightarrow \mathbb{R} \) such that \( I \) is differentiable at \( \theta_0 \) for all \( \theta_0 \), and possesses a asymptotic sampling distribution of minimal variance. In 1930’s and –40’s, Fisher’s ideas on optimality in differentiable models were sharpened and elaborated upon. To illustrate, consider the following classical result from M-estimation (which can be found as theorem 5.23 in [87]).

For all \( \theta_1, \theta_2 \) in an open neighbourhood of \( \theta_0 \). Furthermore, assume that \( \theta \mapsto P_\theta \log p_\theta \) has a second-order Taylor expansion around \( \theta_0 \) of the form,

\[
P_\theta \log p_\theta = P_{\theta_0} \log p_{\theta_0} + \frac{1}{2}(\theta - \theta_0)^T I_{\theta_0}(\theta - \theta_0) + o(||\theta - \theta_0||^2),
\]

with non-singular \( I_{\theta_0} \). If \( (\hat{\theta}_n) \) is a sequence satisfying,

\[
\mathbb{P}_n \log p_{\hat{\theta}_n} \geq \sup_{\theta \in \Theta} \mathbb{P}_n \log p_{\theta} - o_{P_{\theta_0}}(n^{-1}),
\]

such that \( \hat{\theta}_n \xrightarrow{\theta_0} \theta_0 \), then the estimator sequence is asymptotically linear,

\[
n^{1/2}(\hat{\theta}_n - \theta_0) = n^{-1/2} \sum_{i=1}^n I_{\theta_0}^{-1} \ell_{\theta_0}(X_i) + o_{P_{\theta_0}}(1)
\]

In particular, \( n^{1/2}(\hat{\theta}_n - \theta_0) \xrightarrow{\theta_0} \mathcal{N}(0, I_{\theta_0}^{-1}) \).

The last assertion of theorem 4.1 says that the (near-)maximum-likelihood estimators \( (\hat{\theta}_n) \) are asymptotically consistent, converge at rate \( n^{-1/2} \) and have the inverse Fisher information \( I_{\theta_0}^{-1} \) as the covariance matrix for their (normal) limit distribution. At this stage of the discussion, we do not have an argument to show that this asymptotic behaviour is in any sense optimal. Nevertheless, let us take the opportunity to illustrate briefly how asymptotic behaviour translates into inference on \( \theta \) by considering associated asymptotic confidence sets.

An **asymptotic confidence set** is an approximate confidence set that is derived not from exact sampling distributions, but from approximations implied by limit distributions, e.g. from \( n^{1/2}(\hat{\theta}_n - \theta_0) \xrightarrow{\theta_0} \mathcal{N}(0, I_{\theta_0}^{-1}) \) in the above example. To demonstrate, first suppose that the model is one-dimensional and satisfies the conditions of theorem 4.1. Denoting quantiles of the standard normal distribution by \( \xi_\alpha \), we see from the last assertion of the theorem that:

\[
P_{\theta_0}^{\alpha} \left(-\xi_\alpha I_{\theta_0}^{1/2} \leq n^{1/2}(\hat{\theta}_n - \theta_0) \leq \xi_\alpha I_{\theta_0}^{1/2}\right) \rightarrow 1 - 2\alpha,
\]

If the Fisher information were known, this would give rise immediately to a confidence interval: the above display implies that,

\[
\left[\hat{\theta}_n - n^{-1/2} \xi_\alpha I_{\theta_0}^{1/2}, \hat{\theta}_n + n^{-1/2} \xi_\alpha I_{\theta_0}^{1/2}\right]
\]
has coverage probability $1 - 2\alpha$. Since the Fisher information is not known exactly, we substitute an estimator for it, for example the sample variance $S_n^2$, to arrive at a *studentized* version of the above, which has the same asymptotic coverage and can therefore be used as an asymptotic confidence interval. But we could also have chosen to “plug in” the estimator $\hat{\theta}_n$ for $\theta_0$ in the expression for the Fisher information to arrive at an estimate $I_{\hat{\theta}_n}$. To generalize to higher-dimensional $\Theta \subset \mathbb{R}^k$, recall that if $Z$ has a $k$-dimensional multivariate normal distribution $N_k(0, \Sigma)$, then $Z^T \Sigma^{-1} Z$ possess a $\chi^2$-distribution with $k$ degrees of freedom. Denoting quantiles of the $\chi^2$-distribution with $k$ degrees of freedom by $\chi^2_{k,\alpha}$, we find that ellipsoids of the form,

$$C_{\alpha}(X_1, \ldots, X_n) = \{ \theta \in \Theta : n(\theta - \hat{\theta}_n)^T I_{\hat{\theta}_n} (\theta - \hat{\theta}_n) \leq \chi^2_{k,\alpha} \},$$

have coverage probabilities converging to $1 - \alpha$ and are therefore asymptotic confidence sets.

### 4.1.2 Regularity and efficiency

Theorem 4.1 requires a rather large number of smoothness properties of the model: logdensities are required to be differentiable and Lipschitz and the Kullback-Leibler divergence must display a second-order expansion with non-singular second derivative matrix. These sufficient conditions are there to guarantee that the ML estimator displays a property known as *regularity* and the conditions listed are usually referred to as “regularity conditions”. The prominence of regularity in the context of optimality questions was not fully appreciated until J. Hodges discoved an estimator that displayed a property now known as *superefficiency* in 1951.

**Example 4.2.** Suppose that we estimate a parameter $\theta \in \Theta = \mathbb{R}$ with an estimator sequence $(T_n)$, satisfying limiting behaviour described by $n^{1/2}(T_n - \theta) \overset{\theta}{\rightsquigarrow} L_\theta$ for some laws $L_\theta$, for all $\theta \in \Theta$. In addition, we define a so-called shrinkage estimator $S_n$, by $S_n = T_n$ as long as $|T_n| \geq n^{-1/4}$ and $S_n = 0$ otherwise. The name shrinkage estimator arises because any realization of $T_n$ that is close enough to 0 is shrunk to 0 fully. One shows quite easily that $S_n$ has the same asymptotic behaviour as $T_n$ as long as $\theta \neq 0$, i.e. $n^{1/2}(S_n - \theta) \overset{\theta}{\rightsquigarrow} L_\theta$ if $\theta \neq 0$. By contrast, if $\theta = 0$, $\epsilon_n(S_n - 0) \overset{0}{\rightsquigarrow} 0$ for any rate sequence $\epsilon_n$. In other words, the asymptotic quality of $S_n$ is as good as that of $T_n$, or strictly better if $\theta = 0$! (NB: Superefficiency does come at a price, paid in terms of the behaviour of risk functions in neighbourhoods of the point of shrinkage. Furthermore, in one dimensional parametric models, superefficiency can be achieved on a subset of Lebesgue measure no greater than zero. In models of dimension 3 or higher, this restriction does not apply, as demonstrated by the James-Stein phenomenon.)

So at certain points in the parameter space, Hodges’ shrinkage estimators and other superefficient estimators outperform the MLE and other estimators like it asymptotically, while doing equally well for all other points. Superefficiency indicated that Fisher’s 1920’s claim was false without further refinement and that a comprehensive understanding of optimality in differentiable estimation problems remained elusive.
To resolve the issue and arrive at a sound theory of asymptotic optimality of estimation in differentiable models, we have to introduce two concepts. The first is a concise notion of smoothness that describes local behaviour of likelihood products directly in terms of score functions. The “local” aspect of the definition stems from the $n$-dependent re-coordinatization in terms of the local parameter $h = n^{1/2}(\theta - \theta_0)$. (In the following we assume that the sample is i.i.d., although usually the definition is extended to more general models for the data.)

**Definition 4.3.** (Local asymptotic normality, [64])

Let $\Theta \subset \mathbb{R}^k$ be open, parametrizing a model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ that is dominated by a $\sigma$-finite measure with densities $p_\theta$. The model is said to be locally asymptotically normal (LAN) at $\theta_0$ if, for any converging sequence $h_n \to h$:

$$\log \prod_{i=1}^n \frac{P_{\theta_0 + h_n}}{P_{\theta_0}}(X_i) = h^T \Gamma_{n,\theta_0} - \frac{1}{2} h^T I_{\theta_0} h + o_{P_{\theta_0}}(1),$$

(4.2)

for random vectors $\Gamma_{n,\theta_0}$ such that $\Gamma_{n,\theta_0} \overset{\mathcal{D}}{\to} N_k(0, I_{\theta_0})$.

For example, if the data is distributed i.i.d., differentiability of the log-density $\theta \mapsto \log p_\theta(x)$ at $\theta_0$ for every $x$ (with score $\ell_\theta(x) = (d/d\theta) \log p_\theta(x)$) implies that the model is LAN at $\theta_0$ with,

$$\Gamma_{n,\theta_0} = n^{-1/2} \sum_{i=1}^n \ell_{\theta_0}(X_i).$$

But local asymptotic normality can be achieved under weaker conditions; best known is the following property, best described as Hadamard differentiability of square-roots of model densities relative to the $L_2(P_{\theta_0})$ norm.

**Definition 4.4.** Let $\Theta \subset \mathbb{R}^k$ be open. A model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ that is dominated by a $\sigma$-finite measure $\mu$ with densities $p_\theta$ is said to be differentiable in quadratic mean (DQM) at $\theta_0 \in \Theta$, if there exists a score function $\ell_{\theta_0} \in L_2(P_{\theta_0})$ such that:

$$\int \left( p_{\theta_0 + h}^{1/2} - p_{\theta_0}^{1/2} - \frac{1}{2} h^T \ell_{\theta_0} p_{\theta_0}^{1/2} \right)^2 d\mu = o(\|h\|^2),$$

as $h \to 0$.

Theorem 7.2 in [87] shows that a model that is DQM at $\theta_0$ is LAN at $\theta_0$. However, in many situations, it is quite straightforward to demonstrate the LAN property directly.

The second concept is a property that characterizes the class of estimators over which optimality is achieved, in particular excluding Hodges’ shrinkage estimators (and all other examples of superefficiency, as becomes clear below). To prepare the definition heuristically, note that, given Hodges’ counterexample, it is not enough to have estimators with pointwise convergence to limit laws; we must restrict the behaviour of estimators over $(n^{-1/2})$-neighbourhoods rather than allow the type of wild variations that make superefficiency possible.
Definition 4.5. Let $\Theta \subset \mathbb{R}^k$ be open. An estimator sequence $(T_n)$ for the parameter $\theta$ is said to be regular at $\theta$ if, for all $h \in \mathbb{R}^k$,

$$n^{1/2} \left( T_n - \left( \theta + n^{-1/2} h \right) \right) \xrightarrow{L} \mathcal{L}_h, \; (\text{under } P_{\theta + n^{-1/2} h}),$$

i.e. with a limit law independent of $h$.

So regularity describes the property that convergence of the estimator to a limit law is insensitive to perturbation of the parameter of $n$-dependent size $n^{-1/2}h$. The two properties covered above come together through the following theorem (see theorems 7.10, 8.3 and 8.4 in [87]), which formulate the foundation for the convolution theorem that follows.

Theorem 4.6. Let $\Theta \subset \mathbb{R}^k$ be open; let $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ be LAN at $\theta_0$ with non-singular Fisher information $I_{\theta_0}$. Let $(T_n)$ be regular estimators in the “localized models” $\{P_{\theta_0 + n^{-1/2} h} : h \in \mathbb{R}^k\}$. Then there exists a (randomized) statistic $T$ in the normal location model $\{N_k(h, I_{\theta_0}^{-1}) : h \in \mathbb{R}^k\}$ such that $T - h \sim \mathcal{L}_{\theta_0}$ for all $h \in \mathbb{R}^k$.

Theorem 4.6 provides every regular estimator sequence with a limit in the form of a statistic in a very simple model in which the only parameter is the location of a normal model. The theorem states that, within the class of regular estimates, unbiased so they satisfy the Cramér-Rao bound in the limiting model $\{N_k(h, I_{\theta_0}^{-1}) : h \in \mathbb{R}^k\}$.

Theorem 4.7. (Convolution theorem ([44]))

Let $\Theta \subset \mathbb{R}^k$ be open and let $\{P_\theta : \theta \in \Theta\}$ be LAN at $\theta_0$ with non-singular Fisher information $I_{\theta_0}$. Let $(T_n)$ be a regular estimator sequence with limit distribution $\mathcal{L}_{\theta_0}$. Then there exists a probability distribution $M_{\theta_0}$ such that,

$$L_{\theta_0} = N_k(0, I_{\theta_0}^{-1}) \ast M_{\theta_0},$$

In particular, if $\mathcal{L}_{\theta_0}$ has a covariance matrix $\Sigma_{\theta_0}$, then $\Sigma_{\theta_0} \geq I_{\theta_0}^{-1}$.

The occurrence of the inverse Fisher information is no coincidence: the estimators $T$ are unbiased so they satisfy the Cramér-Rao bound in the limiting model $\{N_k(h, I_{\theta_0}^{-1}) : h \in \mathbb{R}^k\}$. Convolution of $N_k(0, I_{\theta_0}^{-1})$ with any distribution $M$ raises its variance unless $M$ is degenerate: the last assertion of the convolution theorem says that, within the class of regular estimates, asymptotic variance is lower-bounded by the inverse Fisher information. A regular estimator that is optimal in this sense, is called best-regular. Anderson’s lemma broadens the notion of optimality, in the sense that best-regular estimators outperform other regular estimators with respect to a large family of loss functions.
Definition 4.8. A loss-function is any $\ell : \mathbb{R}^k \to [0, \infty)$; a subconvex loss-function is a loss function such that the level sets $\{x \in \mathbb{R}^k : \ell(x) \leq c\}$ are closed, convex and symmetric around the origin.

Examples of subconvex loss-functions are many and include, for example, the common choices $\ell(x) = \|x\|^p$, $p \geq 1$.

Lemma 4.9. (Anderson’s lemma)
For any $k \geq 1$, any subconvex loss function $\ell : \mathbb{R}^k \to [0, \infty)$, any probability distribution $M$ on $\mathbb{R}^k$ and any $k$-dimensional covariance matrix $\Sigma$,

$$\int \ell dN_k(0, \Sigma) \leq \int \ell d(N_k(0, \Sigma) \ast M).$$

(A proof of Anderson’s lemma can be found, for instance, in [47].) Finally, we mention the following equivalence, which characterizes efficiency concisely in terms of a weakly converging sequence.

Lemma 4.10. In a LAN model, estimators $(T_n)$ for $\theta$ are best-regular if and only if the $(T_n)$ are asymptotically linear, i.e. for all $\theta$ in the model,

$$n^{1/2}(T_n - \theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} I_{\theta}^{-1}\hat{\ell}_{\theta}(X_i) + o_P(1). \quad (4.3)$$

The random sequence of differences on the r.h.s. of (4.3) is denoted by $\Delta_{n,\theta_0}$ below.

Coming back to theorem 4.1, we see that under stated conditions, a consistent sequence of MLE’s ($\hat{\theta}_n$) is best-regular, finally giving substance to Fisher’s 1920’s claim. Referring to the discussion on confidence sets with which we opened this lecture, we now know that in a LAN model, confidence sets of the form (4.1) based on best-regular estimators ($\hat{\theta}_n$) satisfy a similar notion of optimality: according to the the convolution theorem, the asymptotic sampling distributions of best-regular estimator sequences are all the same and sharpest among asymptotic sampling distributions for regular estimators. The question remains if we can somehow identify confidence sets and credible intervals; in the next chapter, that identification is made asymptotically.

4.2 Le Cam’s Bernstein-von Mises theorem

To address the question of efficiency in smooth parametric models from a Bayesian perspective, we turn to the Bernstein-von Mises theorem. In the literature many different versions of the theorem exist, varying both in (stringency of) conditions and (strength or) form of the assertion. We following Le Cam and Yang (1990) [68] (For later reference, define a parametric prior to be thick at $\theta_0$, if it has a Lebesgue density that is continuous and strictly positive at $\theta_0$.)
The Bernstein-von Mises theorem has a Lebesgue-density given by:

\[ \pi_n(h \mid X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} p_{\theta_0 + h/\sqrt{n}}(X_i) \pi(\theta_0 + h/\sqrt{n}) / \int \prod_{i=1}^{n} p_{\theta_0 + h'/\sqrt{n}}(X_i) \pi(\theta_0 + h'/\sqrt{n}) \, dh', \]

Then the posterior for the local parameter \( H = \sqrt{n}(\theta - \theta_0) \) has a Lebesgue-density given by:

\[ P_{\theta_0}^n - \text{almost-surely.} \]

Continuity of \( \pi \) at \( \theta_0 \) implies that (up to an \( n \)-dependent proportionality constant), \( \pi(\theta_0 + h/\sqrt{n}) \) converges to the constant \( \pi(\theta_0) \), which is strictly positive by assumption. This makes it plausible that upon substitution of the likelihood expansion (4.2), the posterior density converges to:

\[ \int \prod_{i=1}^{n} p_{\theta_0 + h'/\sqrt{n}}(X_i) \, dh' \approx \frac{e^{h^T \Delta_n, \theta_0} - \frac{1}{2}h^T I_{\theta_0}^{-1}h}{\int e^{h'^T \Delta_n, \theta_0} - \frac{1}{2}h'^T I_{\theta_0}^{-1}h' \, dh'} \to \frac{dN(h, I_{\theta_0}^{-1})(\Delta)}{\int dN(h', I_{\theta_0}^{-1})(\Delta) \, dh'} \]

(in a suitable sense with respect to \( P_{\theta_0}^n \)). Here \( \Delta \) is an observation in the normal limit model \( \{N(h, I_{\theta_0}^{-1}) : h \in \mathbb{R}^k\} \). The l.h.s. of the last display equals \( dN(\Delta, I_{\theta_0}^{-1})(h) \) and is the posterior based on a sample consisting only of \( \Delta \) and the (improper) Lebesgue prior for the limit model.

**Theorem 4.11.** (Bernstein-Von Mises, Le Cam and Yang (1990) [68])

Assume that \( \Theta \subset \mathbb{R}^k \) is open and that the model \( \mathcal{P} = \{P_\theta : \theta \in \Theta\} \) is identifiable and dominated. Suppose \( X_1, X_2, \ldots \) forms an i.i.d. sample from \( P_{\theta_0} \) for some \( \theta_0 \in \Theta \). Assume that the model is locally asymptotically normal at \( \theta_0 \) with non-singular Fisher information \( I_{\theta_0} \).

Furthermore, suppose that, the prior \( \Pi_{\Theta} \) is thick at \( \theta_0 \) and that for every \( \epsilon > 0 \), there exists a test sequence \( (\phi_n) \) such that,

\[ P_{\theta_0}^n \phi_n \to 0, \quad \sup_{\|\theta - \theta_0\| > \epsilon} P_{\theta}^n(1 - \phi_n) \to 0. \]

Then the posterior distributions converge in total variation,

\[ \sup_B \left| \Pi(\theta \in B \mid X_1, \ldots, X_n) - N_{\theta_n, (nI_{\theta_0})^{-1}}(B) \right| \to 0, \]

in \( P_{\theta_0} \)-probability, where \( (\theta_n) \) denotes any best-regular estimator sequence.

Since the total-variational distance \( \|N(\mu, \Sigma) - N(\nu, \Sigma)\| \) is bounded by a multiple of \( \|\mu - \nu\| \), we find that the assertion of the Bernstein-Von-Mises theorem can also be formulated with \( \sqrt{n}(\tilde{\theta}_n - \theta_0) \) replacing \( \Delta_n, \theta_0 \). Using the invariance of total-variation under rescaling and shifts, this leads to the conclusion that:

\[ \left\| \Pi_{\theta \mid X_1, \ldots, X_n} - N(\tilde{\theta}_n, n^{-1}I_{\theta_0}^{-1}) \right\| \to 0, \]

for any best-regular estimator sequence.
for any best-regular estimator-sequence $\tilde{\theta}_n$. In particular, according to theorem 4.1 and the lemma 4.3, the maximum-likelihood estimator is best-regular under smoothness conditions on the (log-)likelihood. This serves to motivate the often-heard statement that “Bayesian statistics coincides with the maximum-likelihood estimator asymptotically”. In figure 4.2, Bernstein-von Mises-type of convergence of the posterior is demonstrated with a graphical/numerical example. Also displayed in figure 4.2 are the so-called MAP estimator (the

![Figure 4.1](image)

**Figure 4.1** Convergence of the posterior density. The samples used for calculation of the posterior distributions consist of $n$ observations; the model consists of all normal distributions with mean between $-1$ and 2 and variance 1 and has a polynomial prior, shown in the first ($n=0$) graph. For all sample sizes, the maximum a posteriori and maximum likelihood estimators are indicated by a vertical line and a dashed vertical line respectively. (From Kleijn (2003))

location of maximal posterior density, a popular point-estimator derived from the posterior) and the ML estimator. It is noted that, here, the MLE is efficient so it forms a possible centring sequence for the limiting sequence of normal distributions in the assertion of the Bernstein-Von Mises theorem. Furthermore, it is noted that the posterior concentrates more and more sharply, reflecting the $n^{-1}$-proportionality of the variance of its limiting sequence of normals. It is perhaps a bit surprising in figure 4.2 to see limiting normality obtain already at such relatively low values of the sample size $n$. It cannot be excluded that, in this case, that is a manifestation the normality of the underlying model, but onset of normality of the posterior appears to happen at unexpectedly low values of $n$ also in other smooth, paramet-
ric models. It suggests that asymptotic conclusions based on the Bernstein-Von Mises limit accrue validity fairly rapidly, for $n$ in the order of several hundred to several thousand $i.i.d.$ replications of the observation.

The uniformity in the assertion of the Bernstein-Von Mises theorem over model subsets $B$ implies that it holds also for model subsets that are random. In particular, given some $0 < \alpha < 1$, it is noted that the smallest sets $C_\alpha(X_1, \ldots, X_n)$ such that,

$$N_{\hat{\theta}_n, (nI_{\theta_0})^{-1}}(C_\alpha(X_1, \ldots, X_n)) \geq 1 - \alpha,$$

are ellipsoids of the form (4.1). Since posterior coverage of $C_\alpha$ converges to the $l.h.s.$ in the above display, in accordance with the Bernstein-Von Mises limit, we see that the $C_\alpha$ are asymptotic credible sets of posterior coverage $1 - \alpha$. Conversely, any sequence $(C_n(X_1, \ldots, X_n))$ of (data-dependent) credible sets of coverage $1 - \alpha$, is also a sequence of sets that have asymptotic confidence level $1 - \alpha$ (where we use that $\hat{\theta}_n$ is best-regular). This gives rise to an identification in smooth, parametric models between inference based on frequentist best-regular point-estimators and inference based on Bayesian posteriors. In a practical sense, it eliminates the need to estimate $\theta$ and the Fisher information $I_\theta$ at $\theta$ to arrive at asymptotic confidence sets, if we have an approximation of the posterior distribution of high enough quality (e.g. from MCMC simulation), if we know that the Bernstein-Von Mises theorem holds. In high dimensional parametric models, maximization of the likelihood may be much more costly computationally than generation of a suitable MCMC approximation. As a consequence, the Bernstein-Von Mises theorem has an immediate practical implication of some significance.

Before we continue with the proof of the Bernstein-Von Mises theorem, let is briefly reflect on its conditions: local asymptotic normality and non-singularity of the associated Fisher information are minimal smoothness conditions. They also arise in theorem 4.1 and form the backdrop for any discussion of efficiency. More significant is the required existence of a “consistent” test sequence: what is required is that, asymptotically, we can distinguish $P_\theta$ from any complement of a $\theta$-neighbourhood around $\theta_0$ uniformly. One should compare this condition with that of consistency of near-maximizers of the likelihood in theorem 4.1. Apparently, if such a global (rather than $n^{-1/2}$-sized local) consistency guarantee can not be given, likelihood-based methods like ML or Bayesian estimation cannot be trusted to give rise to asymptotic normality (in their respective forms). In the next section, we shall divide the Bernstein-Von Mises theorem in two parts, with a requirement of local $n^{-1/2}$-sized consistency for the posterior in between. In a separate lemma, we show that a score-test can fill in the gap between local and global consistency.

4.2.1 Proof of the Bernstein-von Mises theorem

We prove the assertion of the Bernstein-Von Mises theorem using a smoothness property that is slightly stronger than necessary, because we shall need a similar formulation in the semiparametric case.
Definition 4.12. We say that a parametric model $\mathcal{P}$ is stochastically LAN at $\theta_0$, if the LAN property of definition 4.3 is satisfied for every random sequence $(h_n)$ that is bounded in probability, i.e. for all $h_n = O_P(1)$:

$$\log \prod_{i=1}^{n} \frac{p_{\theta_0 + n^{-1/2}h_n}(X_i)}{p_{\theta_0}} \Gamma_{n, \theta_0} = \frac{1}{2} h_n^T I_{\theta_0} h_n + o_{P_0}(1),$$  \hfill (4.4) 

for random vectors $\Gamma_{n, \theta_0}$ such that $\Gamma_{n, \theta_0} \Rightarrow N_k(0, I_{\theta_0})$.

Theorem 4.13. Let the sample $X_1, X_2, \ldots$ be distributed i.i.d. $P_0$. Let $\Theta \subset \mathbb{R}^k$ be open, let $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ be stochastically LAN at $\theta_0$ with non-singular Fisher information $I_{\theta_0}$ and let the prior $\Pi$ on $\Theta$ be thick. Furthermore, assume that for every sequence of balls $(K_n) \subset \mathbb{R}^d$ with radii $M_n \to \infty$, we have:

$$\Pi_n(h \in K_n \mid X_1, \ldots, X_n) \xrightarrow{P_0} 1.$$  \hfill (4.5) 

Then the sequence of posteriors converges to a sequence of normal distributions in total variation:

$$\sup_B \left| \Pi_n(h \in B \mid X_1, \ldots, X_n) - N_{\Delta_{n, \theta_0}, I_{\theta_0}^{-1}}(B) \right| \xrightarrow{P_0} 0.$$  \hfill (4.6) 

Proof The proof is split into two parts: in the first part, we prove the assertion conditional on a compact neighbourhood $K$ of 0 in $\Theta$, while in the second part we use this to prove (4.6). Throughout the proof we denote the posterior for $h$ given $X_1, X_2, \ldots, X_n$ by $\Pi_n$ and the normal distribution $N_{\Delta_{n, \theta_0}, I_{\theta_0}^{-1}}$ by $\Phi_n$ (for the definition of $\Delta_{n, \theta_0}$, see lemma 4.10). For $K \subset \mathbb{R}^k$, conditional versions are denoted $\Pi_n^K$ and $\Phi_n^K$ respectively (assuming that $\Pi_n(K) > 0$ and $\Phi_n(K) > 0$, of course).

Let $K \subset \Theta$ be a ball centered on 0. For every open neighbourhood $U \subset \Theta$ of $\theta_0$, $\theta_0 + n^{-1/2} K \subset U$ for large enough $n$. Since $\theta_0$ is an internal point of $\Theta$, we can define, for large enough $n$, the random functions $f_n : K \times K \to \mathbb{R}$ by:

$$f_n(g, h) = \left(1 - \frac{\phi_n(h) s_n(g)}{\phi_n(g) s_n(h)} \frac{\pi_n(g)}{\pi_n(h)}\right)^+,$$

where $\phi_n : K \to \mathbb{R}$ is the Lebesgue density of the (randomly located) distribution $N_{\Delta_{n, \theta_0}, I_{\theta_0}^{-1}}$, $\pi_n : K \to \mathbb{R}$ is the Lebesgue density of the prior for the centred and rescaled parameter $h$ and $s_n : K \to \mathbb{R}$ equals the likelihood product:

$$s_n(h) = \prod_{i=1}^{n} \frac{p_{\theta_0 + h/\sqrt{n}}(X_i)}{p_{\theta_0}}.$$ 

Since the model is stochastically LAN by assumption, we have for every random sequence $(h_n) \subset K$:

$$\log s_n(h_n) = h_n G_{\theta_0} \ell_{\theta_0}^* - \frac{1}{2} h_n^T V_{\theta_0} h_n + o_{P_0}(1),$$

$$\log \phi_n(h_n) = -\frac{1}{2} (h_n - \Delta_{n, \theta_0})^T V_{\theta_0} (h_n - \Delta_{n, \theta_0}) + \text{const}.$$
For any two sequences \((h_n), (g_n) \subset K\), \(\pi_n(g_n)/\pi_n(h_n) \to 1\) as \(n \to \infty\). Combining this with the above display we see that:

\[
\log \frac{\phi_n(h_n)}{\phi_n(g_n)} - \frac{s_n(h_n)}{s_n(g_n)} = -h_n \mathbb{E}_n l_{\theta^*} + \frac{1}{2} \mathbb{E}_n V_{\theta^*} - \frac{1}{2} g_n^T V_{\theta^*} g_n + o_P(1)
\]

as \(n \to \infty\). Since \(x \mapsto (1-e^x)_+\) is continuous on \(\mathbb{R}\), we conclude that for every pair of random sequences \((g_n, h_n) \subset K \times K\):

\[
f_n(g_n, h_n) \xrightarrow{P_0} 0, \quad (n \to \infty).
\]

For fixed, large enough \(n\), \(P_0^n\)-almost-sure continuity of \((g, h) \mapsto \log s_n(g)/s_n(h)\) on \(K \times K\) is guaranteed by the stochastic LAN-condition. Each of the locations \(\Delta_{n, \theta_0}\) for \(\Phi_n\) is tight, so \((g, h) \mapsto \phi_n(g)/\phi_n(h)\) is continuous on all of \(K \times K\) \(P_0^n\)-almost-surely. Continuity (in a neighbourhood of \(\theta_0\)) and positivity of the prior density guarantee that this holds for \((g, h) \mapsto \pi_n(g)/\pi_n(h)\) as well. We conclude that for large enough \(n\), the random functions \(f_n\) are continuous on \(K \times K\), \(P_0^n\)-almost-surely. Application of lemma 4.15 then leads to the conclusion that,

\[
\sup_{g, h \in K} f_n(g, h) \xrightarrow{P_0} 0, \quad (n \to \infty).
\]

Since \(K\) contains a neighbourhood of 0, \(\Phi_n(K) > 0\) is guaranteed. Let \(\Xi_n\) denote the event that \(\Pi_n(K) > 0\). Let \(\eta > 0\) be given and based on that, define the events:

\[
\Omega_n = \{\omega : \sup_{g, h \in K} f_n(g, h) \leq \eta\}.
\]

Consider the expression (recall that the total-variation norm is bounded by 2):

\[
P_0^n \|\Pi^K_n - \Phi^K_n\| 1_{\Xi_n} \leq P_0^n \|\Pi^K_n - \Phi^K_n\| 1_{\Omega_n \cap \Xi_n} + 2P_0^n (\Xi_n \setminus \Omega_n).
\]

As a result of (4.7) the latter term is \(o(1)\) as \(n \to \infty\). The remaining term on the \(r.h.s.\) can be calculated as follows:

\[
\frac{1}{2} P_0^n \|\Pi^K_n - \Phi^K_n\| 1_{\Omega_n \cap \Xi_n} = P_0^n \int \left(1 - \frac{d\Phi^K_n}{d\Pi^K_n}\right) + d\Pi^K_n 1_{\Omega_n \cap \Xi_n}
\]

\[
= P_0^n \int_K \left(1 - \phi^K_n(h) \int_K \frac{s_n(g)\pi_n(g)}{s_n(h)\pi_n(h)} dg\right) + d\Pi^K_n 1_{\Omega_n \cap \Xi_n}
\]

\[
= P_0^n \int_K \left(1 - \int_K \frac{s_n(g)\pi_n(g)\phi^K_n(h)}{s_n(h)\pi_n(h)\phi^K_n(g)} d\Phi^K_n(g)\right) + d\Pi^K_n 1_{\Omega_n \cap \Xi_n}.
\]

Note that for all \(g, h \in K\) we have \(\phi^K_n(h)/\phi^K_n(g) = \phi_n(h)/\phi_n(g)\) since, on \(K\), \(\phi^K_n\) differs from \(\phi_n\) only by a normalisation factor. We use Jensen’s inequality (with respect to the
Le Cam’s Bernstein-von Mises theorem

\( \Phi_n^K \)-expectation) for the (convex) function \( x \mapsto (1 - x)_+ \) to derive:

\[
\frac{1}{2} P_0^n \left\| \Pi_n^K - \Phi_n^K \right\|_{L_1, n} \leq P_0^n \left( \left( 1 - \frac{s_n(g) \pi_n(g) \phi_n(h)}{s_n(h) \pi_n(h) \phi_n(g)} \right) + d \Phi_n^K(g) d \Pi_n^K(h) 1_{\Omega_n \cap \Xi_n} \right)
\]

\[
\leq P_0^n \left( \sup_{g, h \in K} f_n(g, h) 1_{\Omega_n \cap \Xi_n} d \Phi_n^K(g) d \Pi_n^K(h) \right) \leq \eta.
\]

Combination with (4.8) shows that for all compact \( K \subset \mathbb{R}^d \) containing a neighbourhood of 0,

\[
P_0^n \left\| \Pi_n^K - \Phi_n^K \right\|_{L_1, n} \to 0.
\]

Now let \( (K_m) \) be a sequence of balls centred at 0 with radii \( M_m \to \infty \). For each \( m \geq 1 \), the above display holds, so if we choose a sequence of balls \( (K_n) \) that traverses the sequence \( K_m \) slowly enough, convergence to zero can still be guaranteed. Moreover, the corresponding events \( \Xi_n = \{ \omega : \Pi_n(K_n) > 0 \} \) satisfy \( P_0^n(\Xi_n) \to 1 \) as a result of (4.5). We conclude that there exists a sequence of radii \( (M_n) \) such that \( M_n \to \infty \) and

\[
P_0^n \left\| \Pi_n^K - \Phi_n^K \right\|_{L_1, n} \to 0,
\]

(4.9)

(where it is understood that the conditional probabilities on the l.h.s. are well-defined on sets of probability growing to one). Combining (4.5) and lemma 4.17, we then use lemma 4.16 to conclude that:

\[
P_0^n \left\| \Pi_n - \Phi_n \right\|_{L_1, n} \to 0,
\]

which implies (4.6).

Aside from a slightly stronger smoothness property in the form of the stochastic LAN condition, theorem 4.13 appears to require more than theorem 4.11, in the sense that it requires posterior consistency at rate \( n^{-1/2} \) rather than the (fixed) tests for consistency. The following lemma shows that, assuming smoothness, the latter condition is enough to satisfy the former. Its proof is based on the construction of a score test that fills in the “gap” left between the fixed-alternative tests and the growing alternative \( \| \theta - \theta_0 \| \geq n^{-1/2} M_n \). However, the proof is long and detailed and it does not have a semiparametric analog. For that reason the proof is given only in the form of a reference.

**Lemma 4.14.** Assume that \( \Theta \subset \mathbb{R}^k \) is open and that the model \( \mathcal{P} = \{ P_\theta : \theta \in \Theta \} \) is identifiable and dominated. Assume that the model is locally asymptotically normal at \( \theta_0 \) with non-singular Fisher information \( I_{\theta_0} \) and that the prior is thick at \( \theta_0 \). Furthermore, suppose that there exists a test sequence \( (\phi_n) \) such that,

\[
P_0^n(\phi_n) \to 0, \quad \sup_{\| \theta - \theta_0 \| \geq \epsilon} P_\theta^n(1 - \phi_n) \to 0.
\]

Then the posterior converges at rate \( n^{-1/2} \), i.e. for every sequence \( M_n \to \infty \),

\[
\Pi \left( \| \theta - \theta_0 \| \geq n^{-1/2} M_n \mid X_1, \ldots, X_n \right) \xrightarrow{P_\theta} 0.
\]

**Proof** This lemma is a well-specified version of theorem 2.2 in [55], which incorporates theorem 2.3 therein, also found as lemma 10.3 in [87].
4.2.2 Three subsidiary lemmas

The proof of theorem 4.13 also makes use of the following three lemmas which are of a more general character than lemma 4.14.

Lemma 4.15. Let \((f_n)\) be a sequence of random functions \(K \to \mathbb{R}\), where \(K\) is compact. Assume that for large enough \(n \geq 1\), \(f_n\) is continuous \(P_0\)-almost-surely. Then the following are equivalent:

(i) Uniform convergence in probability:

\[
\sup_{h \in K} |f_n(h)| \xrightarrow{P_0} 0, \quad (n \to \infty),
\]

(ii) For any random sequence \((h_n) \subset K\):

\[
f_n(h_n) \xrightarrow{P_0} 0, \quad (n \to \infty),
\]

Proof ((ii)⇒(i), by contradiction.) Assume that there exist \(\delta, \epsilon > 0\) such that:

\[
\limsup_{n \to \infty} P_0 \left( \sup_{h \in K} |f_n(h)| > \delta \right) = \epsilon.
\]

Since the functions \(f_n\) are continuous \(P_0\)-almost-surely, there exists (with \(P_0\)-probability one) a sequence \((\tilde{h}_n)\) such that for every \(n \geq 1\), \(\tilde{h}_n \in K\) and

\[
|f_n(\tilde{h}_n)| = \sup_{h \in K} |f_n(h)|.
\]

Consequently, for this particular random sequence in \(K\), we have:

\[
\limsup_{n \to \infty} P_0 \left( |f_n(\tilde{h}_n)| > \delta \right) = \epsilon > 0.
\]

which contradicts (ii). ((i)⇒(ii).) Given a random sequence \((h_n) \subset K\), and for every \(\delta > 0\),

\[
P_0 \left( \sup_{h \in K} |f_n(h)| > \delta \right) \geq P_0 \left( |f_n(h_n)| > \delta \right).
\]

Given (i), the l.h.s. converges to zero and hence so does the r.h.s.. \(\square\)

The next lemma shows that given two sequences of probability measures, a sequence of balls that grows fast enough can be used conditionally to calculate the difference in total-variational distance, even when the sequences consist of random measures.

Lemma 4.16. Let \((\Pi_n)\) and \((\Phi_n)\) be two sequences of random probability measures on \(\mathbb{R}^k\). Let \((K_n)\) be a sequence of subsets of \(\mathbb{R}^k\) such that

\[
\Pi_n(\mathbb{R}^k \setminus K_n) \xrightarrow{P_0} 0, \quad \Phi_n(\mathbb{R}^k \setminus K_n) \xrightarrow{P_0} 0.
\]

Then

\[
\|\Pi_n - \Phi_n\| - \|\Pi_n^{K_n} - \Phi_n^{K_n}\| \xrightarrow{P_0} 0.
\]
**Proof** Let $K$, a measurable subset of $\mathbb{R}^k$ and $n \geq 1$ be given and assume that $\Pi_n(K) > 0$ and $\Phi_n(K) > 0$. Then for any measurable $B \subset \mathbb{R}^k$ we have:

$$\left| \Pi_n(B) - \Pi^K_n(B) \right| = \left| \Pi_n(B) - \frac{\Pi_n(B \cap K)}{\Pi_n(K)} \right|$$

$$= \left| \Pi_n(B \cap (\mathbb{R}^k \setminus K)) + (1 - \Pi_n(K)^{-1}) \Pi_n(B \cap K) \right|$$

$$\leq \Pi_n(B \cap (\mathbb{R}^k \setminus K)) + \Pi_n(\mathbb{R}^k \setminus K) \Pi^K_n(B) \leq 2\Pi_n(\mathbb{R}^k \setminus K).$$

and hence also:

$$\left| (\Pi_n(B) - \Pi^K_n(B)) - (\Phi_n(B) - \Phi^K_n(B)) \right| \leq 2(\Pi_n(\mathbb{R}^k \setminus K) + \Phi_n(\mathbb{R}^k \setminus K)). \quad (4.12)$$

As a result of the triangle inequality, we then find that the difference in total-variation distances between $\Pi_n$ and $\Phi_n$ on the one hand and $\Pi^K_n$ and $\Phi^K_n$ on the other is bounded above by the expression on the right in the above display (which is independent of $B$).

Define $A_n, B_n$ to be the events that $\Pi_n(K_n) > 0$, $\Phi_n(K_n) > 0$ respectively. On $\Xi_n = A_n \cap B_n$, $\Pi^K_n$ and $\Phi^K_n$ are well-defined probability measures. Assumption (4.10) guarantees that $P^0_n(\Xi_n)$ converges to 1. Restricting attention to the event $\Xi_n$ in the above upon substitution of the sequence $(K_n)$ and using (4.10) for the limit of (4.12) we find (4.11), where it is understood that the conditional probabilities on the l.h.s. are well-defined with probability growing to 1.

To apply the above lemma in the concluding steps of the proof of theorem 4.13, rate conditions for both posterior and limiting normal sequences are needed. The rate condition (4.5) for the posterior is assumed and the following lemma demonstrates that its analog for the sequence of normals is satisfied when the sequence of centre points $\Delta_{n,0}$ is uniformly tight.

**Lemma 4.17.** Let $K_n$ be a sequence of balls centred on the origin with radii $M_n \to \infty$. Let $(\Phi_n)$ be a sequence of normal distributions (with fixed covariance matrix $V$) located respectively at the (random) points $(\Delta_n) \subset \mathbb{R}^k$. If the sequence $\Delta_n$ is uniformly tight, then:

$$\Phi_n(\mathbb{R}^k \setminus K_n) = N_{\Delta_n,V}(\mathbb{R}^k \setminus K_n) \to P^0 \to 0.$$

**Proof** Let $\delta > 0$ be given. Uniform tightness of the sequence $(\Delta_n)$ implies the existence of a constant $L > 0$ such that:

$$\sup_{n \geq 1} P^0_n(\|\Delta_n\| \geq L) \leq \delta.$$

For all $n \geq 1$, call $A_n = \{\|\Delta_n\| \geq L\}$. Let $\mu \in \mathbb{R}^k$ be given. Since $N(\mu, V)$ is tight, for every given $\epsilon > 0$, there exists a constant $L'$ such that $N_{\mu,V}(B(\mu, L')) \geq 1 - \epsilon$ (where $B(\mu, L')$ defines a ball of radius $L'$ around the point $\mu$). Assuming that $\mu \leq L$, $B(\mu, L') \subset B(0, L + L')$ so that with $M = L + L'$, $N_{\mu,V}(B(0, M)) \geq 1 - \epsilon$ for all $\mu$ such that $\|\mu\| \leq L$. Choose $N \geq 1$ such that $M_n \geq M$ for all $n \geq N$. Let $n \geq N$ be given. Then:

$$P^0_n(\Phi_n(\mathbb{R}^k \setminus B(0, M_n)) > \epsilon) \leq P^0_n(A_n) + P^0_n(\{\Phi_n(\mathbb{R}^k \setminus B(0, M_n)) > \epsilon\} \cap A^c_n)$$

$$\leq \delta + P^0_n(\{N_{\Delta_n,V}(B(0, M_n)) > \epsilon\} \cap A^c_n) \quad (4.13)$$
Note that on the complement of $A_n$, $\|\Delta_n\| < L$, so:

$$N_{\Delta_n,V}(B(0,M_n)^c) \leq 1 - N_{\Delta_n,V}(B(0,M)) \leq 1 - \inf_{\|\mu\| \leq L} N_{\mu,V}(B(0,M)) \leq \epsilon,$$

and we conclude that the last term on the r.h.s. of (4.13) equals zero. $\square$

4.3 Exercises

Exercise 4.1. Assume that $n^{1/2}(\hat{\theta}_n - \theta_0) \sim N(0,I_{\theta_0}^{-1})$. Show that the ellipsoids (4.1) are of minimal Lebesgue measure among all subsets of coverage $1 - \alpha$.

Exercise 4.2. Consider Hodges’ estimators $S_n$ of example 4.2. Show that, for any rate sequence $(\epsilon_n)$, $\epsilon_n \downarrow 0$, $\epsilon_n(S_n - 0) \overset{p}{\to} 0$.

Exercise 4.3. Let $\Theta = (0, \infty)$ and let $\mathcal{P} = \{P_{\theta} : \theta \in \Theta\}$ be the model of Poisson distributions $P_{\theta}$ with means $\theta$. Let the data be an i.i.d. sample from $P_{\theta_0}$ for some $\theta_0 \in \Theta$. Show that this model is LAN for all $\theta$.

Exercise 4.4. Let $\Theta = \mathbb{R}$ and let $\mathcal{P} = \{P_{\theta} : \theta \in \Theta\}$ be the model of normal distributions $N(\theta,1)$ of unit variance with means $\theta$. Let the data be an i.i.d. sample from $P_{\theta_0}$ for some $\theta_0 \in \Theta$. Show that this model is LAN for all $\theta$.

Exercise 4.5. Let $f$ be a Lebesgue density on $\mathbb{R}$ that is symmetric around the origin. Define the model $\mathcal{P} = \{P_{\mu,\sigma} : \mu \in \mathbb{R}, \sigma \in (0, \infty)\}$ by densities $f_{\mu,\sigma}(x) = \sigma^{-1}f((x - \mu)/\sigma)$. Show that the Fisher information matrix is diagonal.

Exercise 4.6. Let $P$ and $Q$ be probability measures on a measurable space $(\mathcal{X}, \mathcal{B})$,

a. Show that there exists a $\sigma$-finite measure $\mu$ such that $P,Q \ll \mu$.

b. Using Radon-Nikodym derivatives $p = dP/d\mu$ and $q = dQ/d\mu$, prove that,

$$\sup_{B \in \mathcal{B}} |P(B) - Q(B)| = \int |p - q| \, d\mu.$$

c. Show that, for any sequence $(Q_n)$ of probability measures on $(\mathcal{X}, \mathcal{B})$, there exists a probability measure $P$ that dominates all $Q_n$, $(n \geq 1)$.

d. Use the completeness of $L_1(\mathcal{X}, \mathcal{B}, \mu)$ to show that the metric space $M(\mathcal{X}, \mathcal{B})$ of all probability measures on $(\mathcal{X}, \mathcal{B})$ is complete in the topology of total variation.

Exercise 4.7. Let $\Theta = (0, \infty)$ and $\mathcal{P} = \{N(0,\theta^2) : \theta \in \Theta\}$. Let $\Pi$ be a thick prior on $\Theta$. Show that this model satisfies the conditions of the Bernstein-von Mises theorem 4.11. Find the problematic range of parameter values in this model. (Hint: calculate the Fisher information, find a problematic limit for it and describe the effect on the limiting sequence of normal distributions for parameter values close to the problematic limit.)
Exercise 4.8. Approximation in measure from within by compact subsets has a deep background in analysis. Central is the notion of a Radon measure. Given a Hausdorff topological space $\Theta$, a Radon measure $\Pi$ is a Borel measure that is locally finite (meaning that any $\theta \in \Theta$ has a neighbourhood $U$ such that $\Pi(U) < \infty$) and inner regular (meaning that for any measurable subset $S \subset \Theta$ and any $\epsilon > 0$, there exists a compact $K \subset S$ such that $\mu(S \setminus K) < \epsilon$). Show that any probability measure on a Polish space is Radon. (NB: This statement can be generalized to continuous images of Polish spaces, known as Suslin spaces.)

Exercise 4.9. Prove the following: for $\theta \in \Theta = \mathbb{R}$, let $F_\theta(x) = (1 - e^{-(x-\theta)}) \vee 0$ be the standard exponential distribution function located at $\theta$. Assume that $X_1, X_2, \ldots$ form an i.i.d. sample from $F_{\theta_0}$, for some $\theta_0$. Let $\Pi$ be a thick prior on $\Theta$. Then the associated posterior distribution satisfies, with $h = n(\theta - \theta_0)$,

$$\sup_A \left| \Pi_n \left( h \in A \mid X_1, \ldots, X_n \right) - \text{Exp}^{-n}(\hat{\theta}_n - \theta_0)(A) \right| \xrightarrow{\theta_0 \to 0} 0,$$

where $\hat{\theta}_n = X_{(1)}$ is the maximum likelihood estimate for $\theta_0$ and $\text{Exp}_a^-$ denotes the standard negative exponential distribution located at $a$. (NB: This is an example of an irregular estimation problem: clearly the model does not depend on $\theta$ in a differentiable way. Inspection of the assertion shows that the rate of convergence is $n^{-1}$ rather than $n^{-1/2}$, the rate of convergence in regular situations. In addition, the limiting shape of the posterior is not normal but exponential.)
Appendix A

Measure theory

In this appendix we collect some important notions from measure theory. The goal is not a self-contained presentation but rather to establish the basic definitions and theorems from the theory for reference in the main text. As such, the presentation omits certain existence theorems and many of the proofs of other theorems (although references are given). The focus is strongly on finite (e.g. probability-)measures, in places at the expense of generality. Some background in elementary set-theory and analysis is required. As a comprehensive reference we note Kingman and Taylor (1966) [54], alternatives being Dudley (1989) [29] and Billingsley (1986) [15].

A.1 Sets and sigma-algebras

It is assumed that the reader is familiar with the following notions in set theory: set, subset, empty set, union, intersection, complement, set difference and disjointness. Let \( \Omega \) be a set. The powerset \( 2^\Omega \) is the collection of all subsets of \( \Omega \). A partition of \( \Omega \) is an \( \mathcal{A} \subset 2^\Omega \) such that \( \Omega = \bigcup_{A \in \mathcal{A}} A \) and \( A \cap A' = \emptyset \) for any \( A, A' \in \mathcal{A} \) such that \( A \neq A' \). Let \( (A_n) \) be a sequence of subsets of \( \Omega \). We say that \( (A_n) \) is monotone decreasing (resp. monotone increasing) if \( A_{n+1} \subseteq A_n \) (resp. \( A_n \subseteq A_{n+1} \)) for all \( n \geq 1 \). A monotone decreasing (resp. increasing) sequence \( (A_n) \) has a set-theoretic limit \( \lim A_n \) defined as \( \cap_{n \geq 1} A_n \) (resp. \( \cup_{n \geq 1} A_n \)). For any sequence of subsets \( (A_n) \), the sequence \( (\bigcup_{m \geq n} A_m)_{n \geq 1} \) (resp. \( (\bigcap_{m \geq n} A_m)_{n \geq 1} \)) is monotone decreasing (resp. increasing) and, accordingly, for any sequence \( (A_n) \) we define

\[
\limsup A_n = \cap_{n \geq 1} \bigcup_{m \geq n} A_m, \quad \liminf A_n = \bigcup_{n \geq 1} \cap_{m \geq n} A_m.
\]

The sequence \( (A_n) \) is said to converge, if \( \limsup A_n = \liminf A_n \).

**Definition A.1.** Let \( \Omega \) be a set. A collection \( \mathcal{F} \) of subsets of \( \Omega \) is called a \( \sigma \)-algebra, if \( \mathcal{F} \) has the following properties.

111
1. $\emptyset \in \mathcal{F}$,

2. If $A \in \mathcal{F}$, then $\Omega \setminus A \in \mathcal{F}$,

3. If $(A_n) \subset \mathcal{F}$, then $\cup_{n \geq 1} A_n \in \mathcal{F}$.

**Definition A.2.** A measurable space $(\Omega, \mathcal{F})$ consists of a set $\Omega$ and a $\sigma$-algebra $\mathcal{F}$ of subsets of $\Omega$.

A subset $A$ of a measurable space $(\Omega, \mathcal{F})$ is called measurable if $A \in \mathcal{F}$. It can be shown that a $\sigma$-algebra is a monotone class which means that if $(A_n) \subset \mathcal{F}$ is a monotone sequence, then $\lim A_n \in \mathcal{F}$.

**Definition A.3.** Let $\Omega$ be a non-empty set and let $\mathcal{C}$ be a collection of subsets of $\Omega$. The $\sigma$-algebra generated by $\mathcal{C}$, denoted $\sigma(\mathcal{C})$ is the smallest $\sigma$-algebra that contains $\mathcal{C}$.

**Lemma A.4.** Let $\Omega$ be a non-empty set and let $\mathcal{C}$ be a collection of subsets of $\Omega$. Then,

$$\sigma(\mathcal{C}) = \bigcap\{\Sigma \subset 2^\Omega : \mathcal{C} \subset \Sigma, \Sigma \text{ is a } \sigma\text{-algebra}\}$$

**Example A.5.** Let $(X, \mathcal{T})$ be a topological space. The Borel $\sigma$-algebra is the $\sigma$-algebra $\sigma(\mathcal{T})$ generated by the topology. The Borel $\sigma$-algebra on $\mathbb{R}^n$ is denoted $\mathcal{B}(\mathbb{R}^n)$.

### A.2 Measures

From here on, let $(\Omega, \mathcal{F})$ denote a measurable space. A set-function $\nu$ is any mapping $\mathcal{F} \to \mathbb{R}$.

**Definition A.6.** A set-function $\nu : \mathcal{F} \to \mathbb{R}$ is said to be additive if, for any $k \geq 1$ and any $\mathcal{A} = \{A_1, \ldots, A_k\} \subset \mathcal{F}$ such that $A_i \cap A_j = \emptyset$ for all $1 \leq i < j \leq k$,

$$\nu\left(\bigcup_{i=1}^{k} A_i\right) = \sum_{i=1}^{k} \nu(A_i).$$

A set-function $\nu$ is said to be countably additive (or $\sigma$-additive) if the above holds for any countable $\mathcal{A}$ like above.

**Definition A.7.** Let $(\mathcal{Y}, \mathcal{B})$ be a measurable space. Let $\mathcal{C}$ denote the collection of all finite partitions $\bar{B} = \{B_1, \ldots, B_n\}$ of $\mathcal{Y}$. Given a set-function $\nu : \mathcal{B} \to \mathbb{R}$, the total-variation norm of $\nu$ is defined:

$$\|\nu\|_{TV} = \sup_{B \in \mathcal{C}} \sum_{i=1}^{n} |\nu(B_i)|. \tag{A.1}$$

A set-function $\nu$ is said to be finite if its total variation is finite. A $\sigma$-additive set-function $\nu : \mathcal{F} \to \mathbb{R}$ is said to be $\sigma$-finite if there exists a measurable countable partition $(A_n)$ of $\Omega$ such that $\nu(A_n) < \infty$ for all $n \geq 1$.

Often this expression simplifies $\|\nu\|_{TV} = \nu(\mathcal{Y})$ whenever $\nu$ is a measure due to non-negativity.
Definition A.8. Given a measurable space $(\Omega, \mathcal{F})$, a set-function $\mu : \mathcal{F} \to \mathbb{R}$ is a signed measure if $\mu$ is countably additive and $\mu$ is a measure if $\mu$ is countably additive and $\mu \geq 0$. A measure with a Borel $\sigma$-algebra for a domain is called a Borel measure. If $\mu$ is a measure, $(\Omega, \mathcal{F}, \mu)$ is called a measure space. If $\mu$ is finite (resp. $\sigma$-finite), $(\Omega, \mathcal{F}, \mu)$ is called a finite (resp. $\sigma$-finite) measure space. If the total variation $\mu(\Omega) = 1$, $(\Omega, \mathcal{F}, \mu)$ is called a probability space. In the latter case, $\mu$ is often denoted with $P$ and called a probability measure.

The so-called Hahn-Jordan decomposition asserts that every signed measure can be written as the difference of two measures.

Lemma A.9. Let $(\Omega, \mathcal{F})$ be a measurable space. The collection of all finite signed measures on $\mathcal{F}$ forms a linear space $\text{ca}(\Omega, \mathcal{F})$ and total variation is a norm on this space.

As a result of $\sigma$-additivity, measures display a form of ‘continuity’ expressed by the following theorem.

Theorem A.10. Let $(\Omega, \mathcal{F})$ be a measurable space with measure $\mu : \mathcal{F} \to [0, \infty]$. Then,

(i) for any monotone decreasing sequence $(F_n)_{n \geq 1}$ in $\mathcal{F}$ such that $\mu(F_n) < \infty$ for some $n$,

$$\lim_{n \to \infty} \mu(F_n) = \mu\left(\bigcap_{n=1}^{\infty} F_n\right), \quad (A.2)$$

(ii) for any monotone increasing sequence $(G_n)_{n \geq 1}$ in $\mathcal{F}$,

$$\lim_{n \to \infty} \mu(G_n) = \mu\left(\bigcup_{n=1}^{\infty} G_n\right). \quad (A.3)$$

Theorem A.10 is sometimes referred to as the continuity theorem for measures, because if we view $\cap_n F_n$ as the monotone limit $\lim F_n$, (A.2) can be read as $\lim_n \mu(F_n) = \mu(\lim_n F_n)$, expressing continuity from below. Similarly, (A.3) expresses continuity from above. Note that theorem A.10 does not guarantee continuity for arbitrary sequences in $\mathcal{F}$. It should also be noted that theorem A.10 is presented here in simplified form: the full theorem states that continuity from below is equivalent to $\sigma$-additivity of $\mu$ (for a more comprehensive formulation and a proof of theorem A.10, see [54], theorem 3.2).

Example A.11. Let $\Omega$ be a discrete set and let $\mathcal{F}$ be the powerset $2^\Omega$ of $\Omega$, i.e. $\mathcal{F}$ is the collection of all subsets of $\Omega$. The counting measure $n : \mathcal{F} \to [0, \infty]$ on $(\Omega, \mathcal{F})$ is defined simply to count the number $n(F)$ of points in $F \subset \Omega$. If $\Omega$ contains a finite number of points, $n$ is a finite measure; if $\Omega$ contains a (countably) infinite number of points, $n$ is $\sigma$-finite. The counting measure is $\sigma$-additive.

Example A.12. We consider $\mathbb{R}$ with any $\sigma$-algebra $\mathcal{F}$, let $x \in \mathbb{R}$ be given and define the measure $\delta_x : \mathcal{F} \to [0, 1]$ by

$$\delta_x(A) = \mathbf{1}\{x \in A\},$$
for any \( A \in \mathcal{F} \). The probability measure \( \delta_x \) is called the Dirac measure (or delta measure, or atomic measure) degenerate at \( x \) and it concentrates all its mass in the point \( x \). Clearly, \( \delta_x \) is finite and \( \sigma \)-additive. Convex combinations of Dirac measures, i.e. measures of the form

\[
P = \sum_{j=1}^{m} \alpha_j \delta_{x_j},
\]

for some \( m \geq 1 \) with \( \alpha_1, \ldots, \alpha_m \) such that \( \alpha_j \geq 0 \) and \( \sum_{j=1}^{m} \alpha_j = 1 \), can be used as a statistical model for an observation \( X \) that take values in a discrete (but unknown) subset \( \{x_1, \ldots, x_m\} \) of \( \mathbb{R} \). The resulting model (which we denote \( D(\mathbb{R}, \mathcal{B}) \) for reference) is not dominated.

Often, one has a sequence of events \((A_n)_{n \geq 1}\) and one is interested in the probability of a limiting event \( A \), for example the event that \( A_n \) occurs infinitely often. The following lemmas pertain to this situation.

**Lemma A.13.** (First Borel-Cantelli lemma)
Let \((\Omega, \mathcal{F}, P)\) be a probability space and let \((A_n)_{n \geq 1} \subset \mathcal{F}\) be given and denote \( A = \limsup A_n \). If \( \sum_{n \geq 1} P(A_n) < \infty \), then \( P(A) = 0 \).

In the above lemma, the sequence \((A_n)_{n \geq 1}\) is general. To draw the converse conclusion, the sequence needs to exist of independent events: \( A, B \in \mathcal{F} \) are said to be independent under \( P \) if \( P(A \cap B) = P(A) P(B) \).

**Lemma A.14.** (Second Borel-Cantelli lemma)
Let \((\Omega, \mathcal{F}, P)\) be a probability space and let \((A_n)_{n \geq 1} \subset \mathcal{F}\) be independent and denote \( A = \limsup A_n \). If

\[
\sum_{n \geq 1} P(A_n) = \infty,
\]

then \( P(A) = 1 \).

Together, the Borel-Cantelli lemmas assert that for a sequence of independent events \((A_n)_{n \geq 1}\), \( P(A) \) equals zero or one, according as \( \sum_{n} P(A_n) \) converges or diverges. As such, this corollary is known as a zero-one law.

To conclude this section, we consider a property of random vectors called exchangeability.

**Definition A.15.** A random vector \((X_1, \ldots, X_n) \in \mathbb{R}^n\) with distribution \( P_n \) is said to be exchangeable, if, for any permutation \( \pi \in S_n \), the random vector \((X_{\pi(1)}, \ldots, X_{\pi(n)})\) also has distribution \( P_n \).

This property is a generalization of i.i.d.-ness: note that if \((X_1, \ldots, X_n) \sim P_0^n\) then \((X_1, \ldots, X_n)\) is exchangeable. The converse does not hold but exchangeable distributions can be characterized in terms of i.i.d. distributions, as the following result demonstrates.
Theorem A.16. (De Finetti's theorem) The random vector \((X_1, \ldots, X_n) \sim P_n\) is exchangeable if and only if there exists a (unique) probability measure \(\Pi\) on the collection \(\mathcal{M}(\mathbb{R})\) of all distributions on \(\mathbb{R}\) such that,

\[
P(A) = \int_{\mathcal{M}(\mathbb{R})} \prod_{i=1}^{n} P(A_i) \, d\Pi(P),
\]

for all \(A_1, \ldots, A_n \in \mathcal{B}(\mathbb{R})\).

A.3 Measurability, random variables and integration

In this section we consider random variables and their expectation values. Throughout this section, let \((\Omega, \mathcal{F}, P)\) denote a probability space.

Definition A.17. Given a map \(X : A \to B\) and a subset \(C \subset B\), the pre-image of \(C\) under \(X\), is defined as,

\[
X^{-1}(C) = \{a \in A : X(a) \in C\} \subset A.
\]

Given two measurable spaces \((\Omega, \mathcal{F})\) and \((\mathcal{X}, \mathcal{B})\), a map \(X : \Omega \to \mathcal{X}\) is called measurable if, for all \(B \in \mathcal{B}\), \(X^{-1}(B) \in \mathcal{F}\).

Essentially, measurability makes it possible to speak of “the probability that \(X\) lies in \(B\)’:

\[
P(X \in B) = P(\{\omega \in \Omega : X(\omega) \in B\}),
\]

is well-defined only if \(X^{-1}(B)\) belongs to the domain of \(P\). Specializing to real-valued measurable maps, it follows from elementary manipulation of set-limits that suprema of sequences of measurable maps are again measurable. This statement can be framed in the following central theorem in measure theory.

Theorem A.18. (Monotone class theorem) For every \(n \geq 1\), let \(f_n : \Omega \to \mathbb{R}\) be measurable and assume that \(f_{n+1}(\omega) \geq f_n(\omega)\) for all \(n \geq 1\) and \(\omega \in \Omega\). Then \(f(\omega) = \lim_{n \to \infty} f_n(\omega)\) defines a measurable map \(f : \Omega \to \mathbb{R}\).

This means that the set of all measurable \(f : \Omega \to \mathbb{R}\) forms what is know as a monotone class, an ordered set that is closed for limits over monotone sequences. Although measurability is preserved under linear combination, the space of all measurable \(f : \Omega \to \mathbb{R}\) is not a linear space because if, for some \(\omega \in \Omega\), \(f(\omega) = \infty\) and \(g(\omega) = -\infty\), then \((f + g)(\omega) = \infty - \infty\) is ill-defined. No such problems arise when we restrict to the set of all measurable \(f \geq 0\), which form a cone. Restriction to measurable \(f : \Omega \to \mathbb{R}\), on the other hand, invalidates the monotone class theorem so that does not lead to a viable approach.

Definition A.19. Let \((\Omega, \mathcal{F}, P)\) be a probability space. A random variable is a measurable map \(X : \Omega \to \mathbb{R}\) with the property that \(P(|X| = \infty) = 0\). Therefore, every random variable can be represented by a real-valued \(X' : \Omega \to \mathbb{R}\), up to null-sets of \(P\), i.e. \(P(X = X') = 1\).
Note that random variables do not form a monotone class (take $f_n = n$), but they do form a linear space. To define expectations (integrals with respect to $P$), we extend by monotone limit starting from the following definition.

**Definition A.20.** A measurable map $f : \Omega \rightarrow \mathbb{R}$ is called simple if there exists a $k \geq 1$, a $k$-set partition $A_1, \ldots, A_k$ of $\Omega$ and $a_1, \ldots, a_k \in \mathbb{R}$ such that,

$$f(\omega) = \sum_{i=1}^{k} a_i 1_{A_i}(\omega).$$

The integral of a simple $f$ with respect to $P$ is defined as,

$$\int f \, dP = \sum_{i=1}^{k} a_i P(A_i).$$

A straightforward construction shows that for every measurable $f \geq 0$, there exists an increasing sequence $(f_n)$ of non-negative, simple functions such that $f_n(\omega) \uparrow f(\omega)$ for all $\omega \in \Omega$. By the monotony of $(f_n)$, this defines an integral for every non-negative, measurable $f$,

$$\int f \, dP = \lim_{n \to \infty} \int f_n \, dP,$$

(after one demonstrates that the l.h.s. does not depend on the particular $(f_n)$ we choose to approximate $f$). Extension to real-valued measurable functions that take on negative values as well is done by treating negative $f_-$ and non-negative $f_+$ parts of $f$ separately. Extension to $\mathbb{R}^d$ with $d > 1$ proceeds component-wise. The most important result in integration theory is the following elementary theorem.

**Theorem A.21.** (Monotone convergence) Let $(f_n)$ be a monotone sequence of measurable maps $\Omega \rightarrow \mathbb{R}$. Then $\lim_n \int f_n \, dP = \int (\lim_n f_n) \, dP$.

Before we can state Fatou’s lemma and the dominated convergence theorem, we define integrability of measurable maps.

**Definition A.22.** Let $(\Omega, \mathcal{F}, P)$ be a probability space. A real-valued measurable function $f : \Omega \rightarrow \mathbb{R}$ is said to be integrable with respect to $P$ if

$$\int_{\Omega} |f| \, dP < \infty. \quad (A.4)$$

It follows immediately from the definition that an integrable $f$ is a random variable. Note that any sequence of measurable $f_n$ is dominated by the sequence $(\sup_{m \geq n} f_m)$. By the monotone class theorem the suprema are measurable and the resulting sequence of maps is monotone decreasing.

**Lemma A.23.** (Fatou’s lemma) Let $f_n : \Omega \rightarrow \mathbb{R}$ be a sequence of measurable maps such that $f_n \leq g$, $P$-almost-surely for all $n \geq 1$, for some $P$-integrable $g : \Omega \rightarrow \mathbb{R}$. Then,

$$\limsup_{n \to \infty} \int f_n \, dP \leq \int (\limsup_{n \to \infty} f_n) \, dP.$$
An obvious extension provides an inequality for the limes inferior. When combined, the lim sup and lim inf versions of Fatou’s lemma imply the following result, known as Lebesgue’s (dominated convergence) theorem.

**Theorem A.24.** *(Dominated convergence)* Let \( f_n : \Omega \to \mathbb{R} \) be a sequence of measurable maps such that \( \lim_n f_n : \Omega \to \mathbb{R} \) exists and \( |f_n| \leq g \), \( P \)-almost-surely for all \( n \geq 1 \), for some \( P \)-integrable \( g : \Omega \to \mathbb{R} \). Then,

\[
\lim_{n \to \infty} \int f_n \, dP = \int (\lim_{n \to \infty} f_n) \, dP.
\]

For any two probability spaces \((\Omega_1, \mathcal{F}_1, P_1)\) and \((\Omega_2, \mathcal{F}_2, P_2)\), the set \( \Omega_1 \times \Omega_2 \) can be endowed with the \( \sigma \)-algebra generated by products of the form \( A_1 \times A_2 \) where \( A_1 \in \mathcal{F}_1 \), \( A_2 \in \mathcal{F}_2 \), which is called the product \( \sigma \)-algebra, denoted \( \mathcal{F} = \sigma(\mathcal{F}_1 \times \mathcal{F}_2) \) and a product measure \( P = P_1 \times P_2 \), to arrive at a probability space \((\Omega, \mathcal{F}, P)\), for which the following elementary theorem on the interchangability of integrals applies.

**Theorem A.25.** *(Fubini’s theorem)* Let \((\Omega_1, \mathcal{F}_1, P_1)\) and \((\Omega_2, \mathcal{F}_2, P_2)\) be probability spaces and denote their product by \((\Omega, \mathcal{F}, P)\). For any non-negative, \( \mathcal{F} \)-measurable \( f : \Omega \to \mathbb{R} \) and any \( \omega_1 \in \Omega_1 \), \( f(\omega_1, \cdot) : \Omega_2 \to \mathbb{R} \) is \( \mathcal{F}_2 \)-measurable. Furthermore, for any \( A_1 \in \mathcal{F}_1 \) and \( A_2 \in \mathcal{F}_2 \),

\[
\int_{A_1 \times A_2} f(\omega_1, \omega_2) \, dP(\omega) = \int_{A_1} \left( \int_{A_2} f(\omega_1, \omega_2) \, dP_2(\omega_2) \right) dP_1(\omega_1)
\]

Another central result from integration theory forms the foundation for the probability density we associate with many distributions.

**Theorem A.26.** *(Radon-Nikodym theorem)* Let \((\Omega, \mathcal{F})\) be a measurable space and let \( \mu, \nu : \mathcal{F} \to [0, \infty] \) be two \( \sigma \)-finite measures on \((\Omega, \mathcal{F})\). There exists a unique decomposition

\[
\mu = \mu_{\parallel} + \mu_{\perp},
\]

such that \( \nu_{\parallel} \ll \nu \) and \( \mu_{\perp} \) and \( \nu \) are mutually singular. Furthermore, there exists a finite-valued, \( \mathcal{F} \)-measurable function \( f : \Omega \to \mathbb{R} \) such that for all \( F \in \mathcal{F} \),

\[
\mu_{\parallel}(F) = \int_F f \, d\nu.
\]

(A.5)

The function \( f \) is \( \nu \)-almost-everywhere unique.

The function \( f : \Omega \to \mathbb{R} \) in the above theorem is called the Radon-Nikodym derivative of \( \mu \) with respect to \( \nu \). If \( \mu \) is a probability distribution, then \( f \) is called the (probability) density for \( \mu \) with respect to \( \nu \). The assertion that \( f \) is “\( \nu \)-almost-everywhere unique” means that if there exists a measurable function \( g : \Omega \to \mathbb{R} \) such that (A.5) holds with \( g \) replacing \( f \), then \( f = g \), (\( \nu \)-a.e.), i.e. \( f \) and \( g \) may differ only on a set of \( \nu \)-measure equal to zero. Through a construction involving increasing sequences of simple functions, we see that the Radon-Nikodym theorem has the following implication.
Corollary A.27. Assume that the conditions for the Radon-Nikodym theorem are satisfied. Let $X : \Omega \rightarrow [0, \infty]$ be measurable and $\mu$-integrable. Then the product $Xf$ is $\nu$-integrable and

$$\int X \, d\mu = \int Xf \, d\nu.$$ 

Remark A.28. Integrability is not a necessary condition here, but the statement of the corollary becomes rather less transparent if we indulge in generalization.

A.4 Existence of stochastic processes

A stochastic processes have the following broad definition.

Definition A.29. Let $(\Omega, \mathcal{F}, P)$ be a probability space, let $T$ be an arbitrary set. A collection of $\mathcal{F}$-measurable random variables $\{X_t : \Omega \rightarrow \mathbb{R} : t \in T\}$ is called a stochastic process indexed by $T$.

The problem with the above definition is the requirement that there exists an underlying probability space: typically, one approaches a problem that requires the use of stochastic processes by proposing a collection of random quantities $\{X_t : t \in T\}$. The guarantee that an underlying probability space $(\Omega, \mathcal{F}, P)$ exists on which all $X_t$ can be realised as random variables is then lacking so that we have not defined the stochastic process properly yet. Kolmogorov’s existence theorem provides an explicit construction of $(\Omega, \mathcal{F}, P)$. Clearly, if the $X_t$ take their values in a measurable space space $(\mathcal{X}, \mathcal{B})$, the obvious choice for $\Omega$ is the collection $\mathcal{X}^T$ in which the process takes its values. The question remains how to characterize $P$ and its domain $\mathcal{F}$. Kolmogorov’s solution here is to assume that for any finite subset $S = \{t_1, \ldots, t_k\} \subset T$, the distribution $P_{t_1,\ldots,t_k}$ of the $k$-dimensional stochastic vector $(X_{t_1}, \ldots, X_{t_k})$ is given. Since the distributions $P_{t_1,\ldots,t_k}$ are as yet unrelated and given for all finite subsets of $T$, consistency requirements are implicit if they are to serve as marginals to the probability distribution $P$: if two finite subsets $S_1, S_2 \subset T$ satisfy $S_1 \subset S_2$, then the distribution of $\{X_t : t \in S_1\}$ should be marginal to that of $\{X_t : t \in S_2\}$. Similarly, permutation of the components of the stochastic vector in the above display should be reflected in the respective distributions as well. The requirements for consistency are formulated in two requirements called Kolmogorov’s consistency conditions:

(K1) Let $k \geq 1$ and $\{t_1, \ldots, t_{k+1}\} \subset T$ be given. For any $C \in \sigma(\mathcal{B}^k)$,

$$P_{t_1,\ldots,t_k}(C) = P_{t_1,\ldots,t_{k+1}}(C \times \mathcal{F}),$$

(K2) Let $k \geq 1$, $\{t_1, \ldots, t_k\} \subset T$ and a permutation $\pi$ of $k$ elements be given. For any $A_1, \ldots, A_k \in \mathcal{B}$,

$$P_{t_{\pi(1)}\ldots t_{\pi(k)}}(A_1 \times \ldots \times A_k) = P_{t_1,\ldots,t_k}(A_{\pi^{-1}(1)} \times \ldots \times A_{\pi^{-1}(k)}).$$
Theorem A.30. (Kolmogorov’s existence theorem)
Let a collection of random quantities \( \{X_t : t \in T \} \) be given. Suppose that for any \( k \geq 1 \) and all \( t_1, \ldots, t_k \in T \), the finite-dimensional marginal distributions
\[
(X_{t_1}, \ldots, X_{t_k}) \sim P_{t_1 \ldots t_k},
\]
are defined and satisfy conditions (K1) and (K2). Then there exists a probability space \((\Omega, \mathcal{F}, P)\) and a stochastic process \( \{X_t : \Omega \rightarrow \mathcal{X} : t \in T \} \) such that all distributions of the form (A.6) are marginal to \( P \).

Kolmogorov’s approach to the definition and characterization of stochastic processes in terms of finite-dimensional marginals turns out to be of great practical value: it allows one to restrict attention to finite-dimensional marginal distributions when characterising the process. The drawback of the construction becomes apparent only upon closer inspection of the \( \sigma \)-algebra \( \mathcal{F} \): \( \mathcal{F} \) is the \( \sigma \)-algebra generated by the cylinder sets, which implies that measurability of events restricting an uncountable number of \( X_t \)’s simultaneously can not be guaranteed! For instance, if \( T = [0, \infty) \) and \( \mathcal{X} = \mathbb{R} \), the probability that sample-paths of the process are continuous,
\[
P(\Omega \rightarrow \mathbb{R} : t \mapsto X_t \text{ is continuous}),
\]
may be ill-defined because it involves an uncountable number of \( t \)’s. This is the ever-recurring trade-off between generality and strength of a mathematical result: Kolmogorov’s existence theorem always works but it does not give rise to a comfortably ‘large’ domain for the resulting \( P : \mathcal{F} \rightarrow [0, 1] \).

A.5 Conditional distributions

In this section, we consider conditioning of probability measures. In first instance, we consider straightforward conditioning on events and illustrate Bayes’s rule, but we also cover conditioning on \( \sigma \)-algebras and random variables, to arrive at the posterior distribution and Bayes’s rule for densities.

Definition A.31. Let \((\Omega, \mathcal{F}, P)\) be a probability space and let \( B \in \mathcal{F} \) be such that \( P(B) > 0 \). For any \( A \in \mathcal{F} \), the conditional probability of the event \( A \) given event \( B \) is defined:
\[
P(A|B) = \frac{P(A \cap B)}{P(B)},
\]
\[
(A.7)
\]
Conditional probability given \( B \) describes a set-function on \( \mathcal{F} \) and one easily checks that this set-function is a measure. The conditional probability measure \( P(\cdot | B) : \mathcal{F} \rightarrow [0, 1] \) can be viewed as the restriction of \( P \) to \( \mathcal{F} \)-measurable subsets of \( B \), normalized to be a probability measure. Definition A.31 gives rise to a relation between \( P(A|B) \) and \( P(B|A) \) (in case both \( P(A) > 0 \) and \( P(B) > 0 \), of course), which is called Bayes’s Rule.
Lemma A.32. (Bayes’s Rule)

Let \((\Omega, \mathcal{F}, P)\) be a probability space and let \(A, B \in \mathcal{F}\) be such that \(P(A) > 0, P(B) > 0\). Then

\[ P(A|B) P(B) = P(B|A) P(A). \]

However, only being able to condition on events \(B\) of non-zero probability is too restrictive. Moreover, \(B\) above is a definite event; it is desirable also to be able to discuss probabilities conditional on events that have not been measured yet, i.e. to condition on a whole \(\sigma\)-algebra of events like \(B\) above.

Definition A.33. Let \((\Omega, \mathcal{F}, P)\) be a probability space, let \(\mathcal{C}\) be a sub-\(\sigma\)-algebra of \(\mathcal{F}\) and let \(X\) be a real-valued \(P\)-integrable random variable. The conditional expectation of \(X\) given \(\mathcal{C}\) is any \(\mathcal{C}\)-measurable random variable \(E[X|\mathcal{C}] : \Omega \to \mathbb{R}\) such that,

\[ \int_C X \, dP = \int_C E[X|\mathcal{C}] \, dP, \]

for all \(C \in \mathcal{C}\).

The condition that \(X\) be \(P\)-integrable is sufficient for existence and uniqueness of \(E[X|\mathcal{C}]\) \(P\)-almost-surely, the proof being an application of the Radon-Nikodym theorem (see theorem 10.1.1 in Dudley (1989)). So conditional expectations are not unique but if we have two different random variables \(e_1\) and \(e_2\) satisfying the defining conditions for \(E[X|\mathcal{C}]\), then \(e_1 = e_2\), \(P\)-almost-surely. Often, the \(\sigma\)-algebra \(\mathcal{C}\) is the \(\sigma\)-algebra \(\sigma(Z)\) generated by another random variable \(Z\). In that case we denote the conditional expectation by \(E[X|Z]\) and realizations are denoted \(E[X|Z = z]\).

Definition A.34. Let \((\Omega, \mathcal{F}, P)\) be a probability space and let \(\mathcal{C}\) be a sub-\(\sigma\)-algebra of \(\mathcal{F}\). Furthermore, let \((\mathcal{Y}, \mathcal{B})\) be a measurable space and let \(Y : \Omega \to \mathcal{Y}\) be a random variable. The conditional distribution of \(Y\) given \(\mathcal{C}\) is defined as follows:

\[ P_{Y|\mathcal{C}}(A, \omega) = E[1_{\{Y \in A\}}|\mathcal{C}](\omega), \]

\(P\)-almost-surely.

Although seemingly innocuous, the fact that conditional expectations are defined only \(P\)-almost-surely poses a rather subtle problem: for every \(A \in \mathcal{B}\) there exists an \(A\)-dependent null-set on which \(P_{Y|\mathcal{C}}(A, \cdot)\) is not defined. This is not a problem if we are interested only in \(A\) (or in a countable number of sets). But usually, we wish to view \(P_{Y|\mathcal{C}}\) as a probability measure, that is to say, it must be well-defined as a map on the \(\sigma\)-algebra \(\mathcal{B}\) almost-surely. Since most \(\sigma\)-algebras are uncountable, there is no guarantee that the corresponding union of exceptional null-sets has measure zero as well. This means that definition A.34 is not sufficient for our purposes: the property that the conditional distribution is well-defined almost-surely as a map is called regularity.
**Definition A.35.** If $\pi : \mathcal{B} \times \Omega \to [0, 1]$ is such that,

1. for every $B \in \mathcal{B}$, $\omega \mapsto \pi(G, \omega)$ is $\mathcal{C}$-measurable,
2. there is an $E \in \mathcal{C}$ with $P(E) = 0$ such that for all $\omega \in \Omega \setminus E$, $B \mapsto \pi(B, \omega)$ is a probability measure,
3. for all $C \in \mathcal{C}$,
   $$\int_C \pi(B, \omega) dP(\omega) = P(B \cap C),$$

then $\pi$ is said to be a regular conditional distribution.

The existence of a regular conditional probability cannot be guaranteed without further conditions on the underlying probability space.

**Definition A.36.** A topological space $(S, \mathcal{T})$ is said to be a Polish space if $\mathcal{T}$ is metrizable, complete and separable.

Polish spaces appear in many subjects in measure theory: the existence of a countable, dense subset in a metric setting allows formulation of properties that display a form of continuity in terms of a countable number of metric balls covering the whole space. In this manner Polish spaces allow countable formulations for properties that involve uncountable collections of subsets otherwise, in correspondence with countability restrictions arising from measure theory. Such a transition occurs most notably in a theorem that guarantees the existence of regular conditional distributions.

**Theorem A.37.** Let $(\Omega, \mathcal{F}, P)$ be a probability space and let $\mathcal{Y}$ be a Polish space with Borel $\sigma$-algebra $\mathcal{B}$. If $Y : \Omega \to \mathcal{Y}$ is a random variable taking values in $\mathcal{Y}$, there exists a regular conditional distribution $\Pi_{Y|\mathcal{C}} : \mathcal{B} \times \Omega \to [0, 1]$.

**Proof** For a proof of this theorem, the reader is referred to Dudley (1989) [29, theorem 10.2.2).

In Bayesian context we can be more specific regarding the sub-$\sigma$-algebra $\mathcal{C}$ since $\Omega = \mathcal{X} \times \Theta$. To condition on $\theta$ we consider $\mathcal{C} = \{\mathcal{X} \times G : G \in \mathcal{F}\}$. Due to this special choice for $\mathcal{C}$, $\mathcal{C}$-measurability implies that $\Pi_{Y|\mathcal{C}}(\cdot, (y, \theta))$ is a $\mathcal{C}$-measurable function of $\theta$ which we denote $\Pi_{Y|\theta : \mathcal{B} \times \Theta \to [0, 1]}$. Similarly, since the data $Y$ usually takes its values in $\mathbb{R}$ or $\mathbb{R}^n$ (and certainly in a Polish space), the existence of a regular version of the posterior $\Pi_{\theta|Y : \mathcal{B} \times \mathcal{Y} \to [0, 1]}$ is guaranteed.

A.6 Convergence in spaces of probability measures

Let $M(\mathbb{R}, \mathcal{B})$ denote the space of all probability measures on $\mathbb{R}$ with Borel $\sigma$-algebra $\mathcal{B}$. 
Definition A.38. (topology of weak convergence)
Let \((Q_n)_{n \geq 1}\) and \(Q\) in \(M(\mathbb{R}, \mathcal{B})\) be given. Denote the set of points in \(\mathbb{R}\) where \(\mathbb{R} \to [0, 1]: t \mapsto Q(-\infty, t]\) is continuous by \(C\). We say that \(Q_n\) converges weakly to \(Q\) if, for all \(t \in C\), \(Q_n(-\infty, t]\) \(\to Q(-\infty, t]\).

Weak convergence has several equivalent definitions. The following lemma, known as the Portmanteau lemma (from the French word for coat-rack),

Lemma A.39. Let \((Q_n)_{n \geq 1}\) and \(Q\) in \(M(\mathbb{R}, \mathcal{B})\) be given. The following are equivalent:

(i) \(Q_n\) converges weakly to \(Q\).

(ii) For every bounded, continuous \(f: \mathbb{R} \to \mathbb{R}\), \(Q_n f \to Qf\).

(iii) For every bounded, Lipschitz \(g: \mathbb{R} \to \mathbb{R}\), \(Q_n g \to Qg\).

(iv) For all non-negative, continuous \(h: \mathbb{R} \to \mathbb{R}\), \(\liminf_{n \to \infty} Q_n f \geq Qf\).

(v) For every open set \(F \subset \mathbb{R}\), \(\liminf_{n \to \infty} Q_n(F) \geq Q(F)\).

(vi) For every closed set \(G \subset \mathbb{R}\), \(\limsup_{n \to \infty} Q_n(G) \leq Q(G)\).

(vii) For every Borel set \(B\) such that \(Q(\partial B) = 0\), \(Q_n(B) \to Q(B)\).

In (vii) above, \(\partial B\) denotes the boundary of \(B\), which is defined as the closure of \(B\) minus the interior of \(B\).

Lemma A.40. When endowed with the topology of weak convergence, the space \(M(\mathbb{R}, \mathcal{B})\) is Polish, i.e. complete, separable and metric.

Definition A.41. (topology of pointwise convergence)
Let \((Q_n)_{n \geq 1}\) and \(Q\) in \(M(\mathbb{R}, \mathcal{B})\) be given. We say that \(Q_n\) converges pointwise to \(Q\) if, for all \(B \in \mathcal{B}\), \(Q_n(B) \to Q(B)\).

Definition A.42. (topology of total variation)
Let \((Q_n)_{n \geq 1}\) and \(Q\) in \(M(\mathbb{R}, \mathcal{B})\) be given. We say that \(Q_n\) converges in total variation to \(Q\) if,

\[
\sup_{B \in \mathcal{B}} |Q_n(B) - Q(B)| \to 0.
\]

In exercise 4.6, it is shown that this distance can also be calculated as the \(L_1\)-difference between densities for \(Q_n\) and \(Q\).

Lemma A.43. When endowed with the topology of total variation, the space \(M(\mathbb{R}, \mathcal{B})\) is a Polish subspace of the Banach space of all signed measures on \((\mathbb{R}, \mathcal{B})\).
Bibliography


BIBLIOGRAPHY


## Index

- **$R$-better**, 49  
- **$\sigma$-algebra**, 111  
  - Borel, 112  
  - generated, 112  
- **$p$-value**, 39  
- **i.i.d.**, 1  

- **action**, 48  
- **admissible**, 49  
- **alternative**, *see* alternative hypothesis 38  
  - **hypothesis**, 38  

- **Bayes factor**, 44  
- **Bayes’s billiard**, 20  
- **Bayes’s Rule**, 18  
- **belief**, 12  
- **bootstrap**, 66  

- **central limit theorem**, 40  
- **classification**, 38, 53  
- **classifier**, 54  
- **complement**, 111  
- **conditional distribution**, 16, 120  
  - **regular**, 121  
- **conditional expectation**, 120  
- **conditional independence**, 24  
- **conditional probability**, 119  
- **confidence level**, 33, *see* level, confidence 33  
- **confidence set**, 33  
  - **asymptotic**, 35  
- **conjugate family**, 72, 89  
- **consistency**, 14  
- **consistency conditions**, 118  

- **continuity theorem**, 113  
- **convergence in total variation**, 122  
- **convex hull**  
  - **closed**, 28  
- **counting measure**, 3, 5, 113  
- **credible interval**, 36  
- **credible region**, 36  
- **credible set**, 36  
  - **asymptotic**, 36  
  - **HPD−**, 37  
- **critical set**, 38, 39  

- **data**, 1  
  - **categorical**, 1  
  - **functional**, 1  
  - **interval**, 1  
  - **nominal**, 1  
  - **ordinal**, 1  
  - **ranked**, 1  
  - **ratio**, 1  

- **data distribution**  
  - **Bayesian**, 17  
  - **frequentist**, 2, 24  

- **decision**, 48  
- **decision principle**  
  - **minimax**, 49  

- **decision rule**, 48  
  - **Bayes**, 52  
  - **minimax**, 49  
  - **randomized**, 50  

- **decision-space**, 48  
- **density**, *see* probability density 117
Dirichlet distribution, 83
Dirichlet family, 84
Dirichlet process, 85
disintegration, 18
disjointness, 111
distribution
  unimodal, 27
empirical Bayes, 79
empirical expectation, 13
empirical process, 13
empty set, 111
entropy
  Lindley, 71
  Shannon, 71
estimate, 6
estimator, 6
  $M$-, 32
  formal Bayes, 31
  MAP, 31
  maximum-a-posteriori, 31
  maximum-likelihood, 8, 13
minimax, 51
non-parametric ML, 90
penalized maximum-likelihood, 32
small-ball, 31
unbiased, 91
exchangability, 114
exchangeability, 25
expectation
  empirical, 7
expected loss, 30
exponential family, 73
  canonical representation, 73
  of full rank, 73
feature vector, 54
Hahn-Jordan decomposition, 113
hyperparameter, 77
hyperprior, 65, 77
hypothesis, 38
  composite, 38
  simple, 38
identifiability, 3
inadmissible, 49
independence, 114
inference, 47
infinite divisibility, 83
integrability, 116
integral, 116
intersection, 111
lemma
  Fatou’s, 116
  First Borel-Cantelli, 114
  Second Borel-Cantelli, 114
level, 36, 39
  confidence, 33
  significance, 39
likelihood, 8
likelihood principle, 8
limit
  set-theoretic, 111
limit distribution, 7
local compactness, 57
location, 27
location family, 38
loss, 30, see loss-function
  loss-function, 31, 48
  $L_2$-, 50
measurability, 115
  of a subset, 112
measure, 113
  atomic, 114
  Borel, 113
  delta, 114
  Dirac, 114
  inner regular, 57
Radon, 57
signed, 113
measure space, 113
minimax optimality, 43
minimax theorem, 43
misclassification, 54
ML-II estimator, 80
MLE, see estimator, maximum-likelihood8
model, 2
Bayesian, 17, 21
dimension, 4
dominated, 2
full, 2
full non-parametric, 2
hierarchical Bayes, 76
identifiable, 3
mis-specified, 4
non-parametric, 5
normal, 5
parameterized, 3
parametric, 4
well-specified, 4
model distributions
Bayesian, 16
frequentist, 2
model selection, 80, 81
monotone class, 112, 115
monotony
set-theoretic, 111
norm
total variation, 28
total-variation, 2, 8, 112
NPMLE, see nn-parametric MLE90
null
hypothesis, 38
odds ratio
posterior, 44
prior, 44
optimality criteria, 8
over-fitting, 81
parameter
nuisance, 7
of interest, 7
parameter space, 3
partition, 111
point-estimator, see estimator6
pointwise convergence, 122
Polish space, 17, 121
Portmanteau lemma, 122
posterior, 10, 17
posterior mean, 29
posterior median, 30
posterior mode, 31
power function, 39
sequence, 42
power-set, 5
powerset, 111, 113
pre-image, 115
predictive distribution
posterior, 17, 27
prior, 17
preferred
Bayes, 51
minimax, 49
prior, 10, 17
conjugate, 72
Dirichlet process, 19
improper, 67
informative, 62
Jeffreys, 69
non-informative, 66
objective, 66
reference, 70
subjective, 62
probability density, 117
probability density function, 2
probability space, 113
Radon-Nikodym derivative, 117
random variable, 115
randomization, 41
randomized test, 41
rate of convergence, 7, 14
regularity, 120
risk
Bayes, 52
minimax, 49
risk function
Bayesian, 51
sample-average, 7, 13
samplespace, 1, 48
sampling distribution, 34
set, 111
set difference, 111
set-function, 112
σ-additive, 112
σ-finite, 112
additive, 112
countably additive, 112
finite, 112
significance level, see level, see level, significance
simple function, 116
simplex, 5
state, 48
state-space, 48
statistic, 6, 34
statistical decision theory, 47
statistics
inferential, 47
stochastic process, 118
studentization, 35
subset, 111
support, 22
test
asymptotic, 40
uniformly asymptotically most powerful, 43
uniformly most powerful, 39
test function, 41
test sequence, 42
minimax, 44
test statistic, 38
test-statistic, 39
theorem
Bernstein-von Mises, 36
central limit, 7
De Finetti’s, 115
dominated convergence, 117
Fubini’s, 117
minimax, 49
monotone class, 115
monotone convergence, 116
Radon-Nikodym, 117
type-I error, 39
type-II error, 39
union, 111
utility, see utility-function
utility-function, 48
version
posterior, 18
weak convergence, 122
zero-one law, 114
The figure on the front cover originates from Bayes (1763), *An essay towards solving a problem in the doctrine of chances*, (see [4] in the bibliography), and depicts what is nowadays known as Bayes’ Billiard. To demonstrate the uses of conditional probabilities and Bayes’ Rule, Bayes came up with the following example: one white ball and \( n \) red balls are placed on a billiard table of length normalized to 1, at independent, uniformly distributed positions. Conditional on the distance \( X \) of the white ball to one end of the table, the probability of finding exactly \( k \) of the \( n \) red balls closer to that end, is easily seen to be:

\[
P(k \mid X = x) = \frac{n!}{k!(n-k)!} x^k (1 - x)^{n-k}.
\]

One finds the probability that \( k \) red balls are closer than the white, by integrating with respect to the position of the white ball:

\[
P(k) = \frac{1}{n+1}.
\]

Application of Bayes’ Rule then gives rise to a Beta-distribution \( B(k+1, n-k+1) \) for the position of the white ball conditional on the number \( k \) of red balls that are closer. The density:

\[
\beta_{k+1,n-k+1}(x) = \frac{(n+1)!}{k!(n-k)!} x^k (1 - x)^{n-k},
\]

for this Beta-distribution is the curve drawn at the bottom of the billiard in the illustration. (See example 2.11)