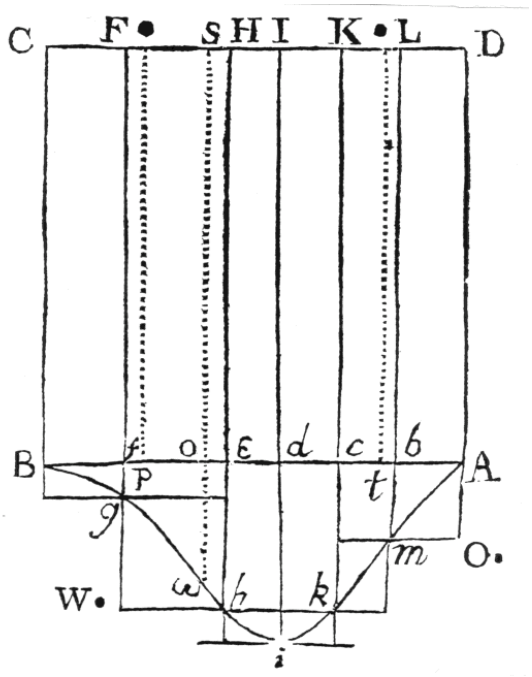
Bayesian Statistics



B.J.K. Kleijn

University of Amsterdam
Korteweg-de Vries institute for Mathematics

Spring 2009

Contents

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, Bayesian inference, model selection and applications. In addition, non-parametric Bayesian modelling and posterior asymptotic behaviour have received due attention and computational methods were presented.

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; these lecture notes provide a general introduction to this topic. 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.

The author wishes to thank Aad van der Vaart for his contributions to this course and these lecture notes, concerning primarily (but not exclusively) the chapter entitled 'Numerical methods in Bayesian statistics'. For corrections to the notes, the author thanks C. Muris, ...

Bas Kleijn, Amsterdam, January 2007

Chapter 1

Introduction

The goal of inferential statistics 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. Based on that assumption, one may proceed to estimate this distribution directly, or to give estimates of certain characteristic properties (like its mean, variance, *etcetera*). It is this straightforward assumption that underlies frequentist statistics and markedly distinguishes it from the Bayesian approach.

1.1 Frequentist statistics

Any frequentist inferential procedure relies on three basic ingredients: the data, a model and an estimation procedure. The *data* is a measurement or observation which we denote by Y, taking values in a corresponding samplespace.

Definition 1.1.1. *The samplespace for an observation Y is a measurable space $(\mathscr{Y}, \mathscr{B})$ (see definition A.1.1) containing all values that Y can take upon measurement.*

Measurements and data can take any form, ranging from categorical data (sometimes referred to as nominal data where the samplespace is simply a (usually finite) set of points or labels with no further mathematical structure), ordinal data (sometimes called ranked data, where the samplespace is endowed with an total ordering), to interval data (where in addition to having an ordering, the samplespace allows one to compare differences or distances between points), to 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 *i.i.d.*) sample). The data Y may even take its values in a space of functions or in other infinite-dimensional spaces.

The samplespace $\mathscr{Y}$ is assumed to be a measurable space to enable the consideration of probability measures on $\mathscr{Y}$, 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 $P_0 : \mathscr{B} \to [0,1]$ on the samplespace $\mathscr{Y}$ representing the "true distribution of the data":

$$Y \sim P_0 \tag{1.1}$$

Hence from the frequentist perspective, inferential statistics revolves around the central question: "What is P_0?", which may be considered in parts by questions like, "What is the mean of P_0?", "What are the higher moments 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 1.1.2. *A statistical model $\mathscr{P}$ is a collection of probability measures $P : \mathscr{B} \to [0,1]$ on the samplespace $(\mathscr{Y}, \mathscr{B})$.*

The model $\mathscr{P}$ 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. Often, we describe the model in terms of probability densities rather than distributions.

Definition 1.1.3. *If there exists a σ-finite measure $\mu : \mathscr{B} \to [0,\infty]$ such that for all $P \in \mathscr{P}$, $P \ll \mu$, we say that the model is dominated.*

The Radon-Nikodym theorem (see theorem A.4.2) guarantees that we may represent a dominated model $\mathscr{P}$ in terms of probability density functions $p = dP/d\mu : \mathscr{Y} \to \mathbb{R}$. Note that the dominating measure may not be unique and hence, that the representation of $\mathscr{P}$ in terms of densities depends on the particular choice of dominating measure μ. A common way of representing a model is a description in terms of a parameterization.

Definition 1.1.4. *A model $\mathscr{P}$ is parameterized with parameter space Θ, if there exists a surjective map $\Theta \to \mathscr{P} : \theta \mapsto P_\theta$, called the parameterization of $\mathscr{P}$.*

Surjectivity of the parameterization is imposed so that for all $P \in \mathscr{P}$, there exists a $\theta \in \Theta$ such that $P_\theta = P$: unless surjectivity is required the parameterization may describe $\mathscr{P}$ only partially. Also of importance is the following property.

Definition 1.1.5. *A parameterization of a statistical model $\mathscr{P}$ is said to be identifiable, if the map $\Theta \to \mathscr{P} : \theta \mapsto P_\theta$ 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 θ_1 and θ_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_θ, identifiability is a first requirement. Other common conditions on the map

$\theta \mapsto P_\theta$ are continuity (with respect to a suitable (often metric) topology on the model), differentiability (which may involve technical subtleties in case Θ is infinite-dimensional) and other smoothness conditions.

Remark 1.1.1. *Although strictly speaking ambivalent, it is commonplace to refer to both $\mathscr{P}$ and the parameterizing space Θ 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 Θ and $\mathscr{P}$.*

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

Definition 1.1.6. *A model $\mathscr{P}$ is said to be well-specified if it contains the true distribution of the data P_0, i.e.*

$$P_0 \in \mathscr{P}. \tag{1.2}$$

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

Clearly if $\mathscr{P}$ is parameterized by Θ, (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 θ_0 is unique.

Notwithstanding the fact that there may be inherent restrictions on the possible distributions for Y (like guaranteed positivity of the measurement outcome, or symmetries in the problem), the model we use in a statistical procedure constitutes a *choice* rather than a given: presented 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. However, since P_0 is unknown, (1.2) has the status of an assumption on the unknown quantity of interest P_0 and may, as such, be hard to justify depending on the comprehensiveness of $\mathscr{P}$. When choosing the model, two considerations compete: on the one hand, small models are easy to handle mathematically and parameters are usually clearly interpretable, 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). In this respect the most important distinction is made in terms of the dimension of the model.

Definition 1.1.7. *A model $\mathscr{P}$ is said to be parametric of dimension d, if there exists an identifiable parameterization $\Theta \to \mathscr{P} : \theta \mapsto P_\theta$, where $\Theta \subset \mathbb{R}^d$ with non-empty interior $\mathring{\Theta} \neq \emptyset$.*

The requirement regarding the interior of Θ in definition 1.1.7 ensures that the dimension d really concerns Θ and not just the dimension of the space $\mathbb{R}^d$ of which Θ forms a subset.

Example 1.1.1. *The normal model for a single, real measurement Y, is the collection of all normal distributions on $\mathbb{R}$, i.e.*

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

where the parameterizing space Θ 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:

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

Definition 1.1.8. *If an infinite-dimensional space Θ is needed to parameterize $\mathscr{P}$, then $\mathscr{P}$ is called a non-parametric model.*

For instance, the model consisting of *all* probability measures on $(\mathscr{Y}, \mathscr{B})$ (sometimes referred to as the full non-parametric model) is non-parametric unless the samplespace contains a finite number of points. Note that if the full non-parametric model is used, (1.2) holds trivially.

Example 1.1.2. *Let $\mathscr{Y}$ be a finite set containing $n \geq 1$ points $y_1, y_2, \ldots, y_n$ and let $\mathscr{B}$ be the power-set $2^{\mathscr{Y}}$ of $\mathscr{Y}$. Any probability measure $P : \mathscr{B} \to [0, 1]$ on $(\mathscr{Y}, \mathscr{B})$ is absolutely continuous with respect to the counting measure on $\mathscr{Y}$ (see example A.2.1). The density of P with respect to the counting measure is a map $p : \mathscr{Y} \to \mathbb{R}$ 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

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

This leads to an identifiable parameterization $S_n \to \mathscr{P} : p \mapsto P$ of the full non-parametric model on $(\mathscr{Y}, \mathscr{B})$, 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:

$$\Big\{ (p_1, \ldots, p_{n-1}) \in \mathbb{R}^{n-1} \,:\, p_i \geq 0, \sum_{i=1}^{n-1} p_i \leq 1 \Big\} \to S_n \,:$$

$$(p_1, \ldots, p_{n-1}) \mapsto \Big(p_1, \ldots, p_{n-1}, 1 - \sum_{i=1}^{n-1} p_i \Big).$$

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 $\mathscr{P}$ as prototypical.

Definition 1.1.9. *A point-estimator (or estimator) is a map $\hat{P} : \mathscr{Y} \to \mathscr{P}$, representing our "best guess" $\hat{P}(Y) \in \mathscr{P}$ for P_0 based on the data Y (and other known quantities).*

Note that a point-estimator is a *statistic*, *i.e.* a quantity that depends only on the data (and possibly on other known information): 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 with outcome $Y = y$. Also note that a point-estimator is a stochastic quantity: $\hat{P}(Y)$ depends on Y and is hence random with its own distribution on $\mathscr{P}$ (as soon as a σ-algebra on $\mathscr{P}$ is established with respect to which $\hat{P}$ is measurable). Upon measurement of Y resulting in a realisation $Y = y$, the estimator $\hat{P}(y)$ is a definite point in $\mathscr{P}$.

Remark 1.1.2. *Obviously, many other quantities may be estimated as well and the definition of a point-estimator given above is too narrow in that sense. Firstly, if the model is parameterized, one may define a point-estimator $\hat{\theta} : \mathscr{Y} \to \Theta$ for θ_0, from which we obtain $\hat{P}(Y) = P_{\hat{\theta}(Y)}$ as an estimator for P_0. If the model is identifiable, estimation of θ_0 in Θ is equivalent to estimation of P_0 in $\mathscr{P}$. But if the dimension d of the model is greater than one, we may choose to estimate only one component of θ (called the parameter of interest) and disregard other components (called nuisance parameters). More generally, we may choose to estimate certain properties of P_0, for example its expectation, variance or quantiles, rather than P_0 itself. As an example, consider a model $\mathscr{P}$ consisting of distributions on $\mathbb{R}$ with finite expectation and define the linear functional $e : \mathscr{P} \to \mathbb{R}$ by $e(P) = PX$. Suppose that we are interested in the expectation $e_0 = e(P_0)$ of the true distribution. Obviously, based on an estimator $\hat{P}(Y)$ for P_0 we may define an estimator*

$$\widehat{e}(Y) = \int_{\mathscr{Y}} y \, d[\hat{P}(Y)](y) \tag{1.3}$$

to estimate e_0. But in many cases, direct estimation of the property of interest of P_0 can be done more efficiently than through $\hat{P}$.

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,

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

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 $\hat{P}(Y) = \mathbb{P}_n$ and $\mathbb{P}_n \in \mathscr{P}$.) The law of large numbers guarantees that $\mathbb{P}_n X$ converges to e_0 almost-surely as $n \to \infty$, and 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) if the variance of X under P_0 is finite. (More on the behaviour of estimators in the limit of large sample-size n can be found in chapter 4.) Many parameterizations $\theta \mapsto P_\theta$ are such that parameters coincide with expectations: for instance in the normal model, the parameter μ coincides with

the expectation, so that we may estimate

$$\hat{\mu}(Y) = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

Often, other properties of P_0 can also be related to expectations: for example, if $X \in \mathbb{R}$, the probabilities $F_0(s) = P_0(X \leq s) = P_0 1\{X \leq s\}$ 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, $\mathbb{P}_n g(X)$. (With regard to the estimator $\hat{F}$, the convergence $\hat{F}(s) \to F_0(s)$ does not only hold for all $s \in \mathbb{R}$ but even uniform in s, i.e. $\sup_{s\in\mathbb{R}} |\hat{F}(s) - F_0(s)| \to 0$, c.f. the Glivenko-Cantelli theorem.)

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 2.4 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 1.1.9).

Example 1.1.3. *To illustrate what we mean by "desirable properties", note the following. When estimating P_0 one may decide to use an estimator $\hat{P}(Y)$ because it has the property that it is close to the true distribution of Y in total variation (see appendix A, definition A.2.1). To make this statement more specific, the property that make such an estimator $\hat{P}$ attractive is that there exists a small constant $\epsilon > 0$ and a (small) significance level $0 < \alpha < 1$, such that for all $P \in \mathscr{P}$,*

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

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

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

Definition 1.1.10. *Suppose that the model $\mathscr{P}$ is dominated by a σ-finite measure μ. The likelihood principle says that one should pick $\hat{P} \in \mathscr{P}$ as an estimator for the distribution P_0 of Y such that*

$$\hat{p}(Y) = \sup_{P\in\mathscr{P}} p(Y).$$

thus defining the maximum-likelihood estimator (or MLE) $\hat{P}(Y)$ for P_0.

Remark 1.1.3. *Note that $\hat{P}$ does not depend on the particular dominating measure μ.*

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 $\mathscr{P}$ 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 1.1.4. *If $P_\cdot : \Theta \to \mathscr{P}$ parameterizes $\mathscr{P}$, the above is extended to the maximum-likelihood estimator $\hat{\theta}(Y)$ for θ_0, when we note that $\sup_{\theta\in\Theta} p_\theta(Y) = \sup_{P\in\mathscr{P}} p(Y)$.*

The above is only a very brief and rather abstract overview of the basic framework of frequentist statistics, highlighting the central premise that a P_0 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 2.2.7 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 1.2.1).

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 1.1.4. *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 2.4.

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*", (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 θ. This point of view 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 2.1.2 in these lecture notes). An element P_θ of the model is interpreted simply as the distribution of Y *given* the parameter value θ, *i.e.* as the conditional distribution of $Y|\theta$. The joint distribution of (Y, θ) then follows upon specification of the marginal distribution of θ on Θ, which is called the *prior*. Based on the joint distribution for the data Y and the parameters θ, straightforward conditioning on Y gives rise to a distribution for the parameters $\theta|Y$ called the *posterior* distribution on the model Θ. 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 θ are more "likely" than others, even before any measurements are done. The model Θ 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 [74].

Example 1.2.1. *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 analyse 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:

$$H_0: \quad \theta_0 = \tfrac{1}{2}, \qquad H_1: \quad \theta_0 > \tfrac{1}{2}.$$

The total number of successes out of ten trials is a sufficient statistic for θ 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 the prior such as to 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 θ 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 θ. 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 a 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 1.2.1. 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 make a choice for a certain point-estimator. 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 2.2.5), one can demonstrate that the log-prior-density can be viewed as a penalty term in a penalized maximum-likelihood procedure, *c.f.* remark 2.2.7.) 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 dissolves the inherent ambiguity haunting the subjectivist interpretation of statistical results.

Note, however, that the derivation of expression (2.7) (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 (2.7) as a *definition* rather than a derived form. This has the consequence that some basic properties implicit by derivation in the Bayesian framework, have to be imposed as conditions in the hybrid perspective (see remark 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 (see chapter 4), 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 PZ. 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 $\mathbb{P}_n X$, is defined $\mathbb{P}_n X = n^{-1} \sum_{i=1}^n X_i$ (where it is assumed that X is P_0-integrable); the *empirical process* $\mathbb{G}_n$ is defined as $\mathbb{G}_n X = n^{1/2}(\mathbb{P}_n - P_0)X$ (where it is assumed that $P_0(X - P_0 X)^2 < \infty$). The distribution function of the standard normal distribution is denoted $\Phi : \mathbb{R} \to [0, 1]$. The transpose of a vector $\ell \in \mathbb{R}^d$ is denoted ℓ^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 $Y \in \mathscr{Y}$ be a random variable with unknown distribution P_0. Let $\mathscr{P}$ be a model for Y, dominated by a σ-finite measure μ. Assume that the maximum-likelihood estimator $\hat{P}(Y)$ (see definition 1.1.10) is well-defined, P_0-almost-surely.*
Show that if ν is a σ-finite measure dominating μ and we calculate the likelihood using ν-densities, then the associated MLE is equal to $\hat{P}(Y)$. Conclude that the MLE does not depend on the dominating measure used, c.f. remark 1.1.3.

Exercise 1.2. *In the three experiments of example 1.2.1, give the Neyman-Person test for hypotheses H_0 and H_1 at level $\alpha \in (0, 1)$. Calculate the p-value of the realization of 10 successes and 0 failures (in 10 Bernoulli trials according to H_0).*