3 The inference problem

Inference theory deals with the study of the laws and rules for drawing conclusions for entities in mathematical models of a reality, that are based on data or observations of that same reality.

Thus there is a real world W which in some sense we do not know anything about. To understand W and in some sense even to be able to discuss W one constructs models of different kinds, for instance mathematical models. It is also possible to make other kinds of models, e.g. linguistic models. So in this sense we can view mathematics as a language with which we describe the thing we call the real world.

Thus we can make a mathematical model M of the real world W, in which there is a definition of for instance the concepts "effect", "interaction", "independence" and so on. The problem of making adequate models is to a certain extent not formalized from a mathematical point of view, and usually draws quite heavily from the theoretical basis in other sciences. Mainly the mathematical modeling of real world events is something which is often seen as separate from mathematics, e.g. as in physics, economy, biology or medicine. These notes will not deal with the problem of how to find good models, and it will not deal will the problem of model fit or the evaluation of the models. We will mainly deal with the problem of, having found or being given a useful model or class of models, how to derive properties for solutions to equations in the mathematical model M.

Having obtained a model, and having made observations of entities in the mathematical model, i.e. having gathered data D, it is of interest to look at the data D and draw conclusions or make inference about unknown parameters in the model M based on the data. Inference theory is concerned with the laws for that project.

Thus we have the real world W, which is modeled with a mathematical model M which contains entities of which we can gather observations or data D. The models we use are stochastic, i.e. they contain entities which can be modeled as random entities, with the use of tools from measure theory and probability theory.

Depending on the level of mathematical modeling one can distinguish between three main levels of laws of drawing inference.

1. *Data analysis.* On this level the mathematical modeling consists of labeling things which are possible to observe in the real world, and assuming that some are stochastic in nature. Usually no or very little is assumed on possible relations between different random observables. The most common assumptions that are made are that a particular random observable is discrete or continuous, possibly with an assumption on the range. The analysis consists of making tabulations and frequency plots, such as histogram plots. This form of analysis is usually called *descriptive statistics.* A refined form of this is *data mining.*

2. *Frequentist Inference theory.* The mathematical model consists of a probability model, containing unknown parameters that are modeled as non random. We then want to draw conclusions about the parameters that are based on the gathered data. The unknown parameters can be finite vectors, or functions or

measures or functionals, basically any kind of mathematical objects. A basic study is how to make point estimates or confidence intervals for or tests of the parameters.

3. *Bayesian Inference theory.* This is a refinement and a generalization of the frequentist theory in which the unknown parameters are allowed to be random. One is then interested in drawing conclusions on the distribution of the unknown parameters that are based on the data.

We will in these lecture notes treat only the problems in Inference theory, so parts 2 and 3.

Depending on whether the parameters are finite dimensional vectors, or infinite dimensional vectors or functions, the problems are called parametric or nonparametric inference problems. We will only treat parametric problems, and leave the nonparametric part for later courses.

In parametric inference, the mathematical model in it's simplest form consists of providing a distribution function $F = F_\theta$ that is indexed by a finite dimensional vector θ assumed (i.e. restricted) to lie in some pre specified set Ω. The data consists of observations $x = (x_1, \dots, x_n)$ of a random vector $X = (X_1, \dots, X_n)$ with distribution F_θ. Based on the values of x we want to draw a conclusion about θ or more generally about some known function $g(\theta)$.

The simplest types of data are vectors $X = (X_1, \dots, X_n)$ in which the X_i are independent and all have the same distribution function. In real life situations one often does not have independent data. One can then use stochastic process theory, and model the sequence as satisfying various assumptions such as the Markov property or the stationarity property. Inference for stationary *dependent data* will be treated in a later course.

We will only treat finite sample properties of estimators and test, i.e. inference based on a finite sample $x = (x_1, \dots, x_n)$. Especially when doing nonparametric inference many of the results that are possible to obtain are *asymptotic*, i.e. say something about the obtained estimators and tests as the sample sizes grows, in the sense of $n \to \infty$. This will be done in a later course.

Thus we will treat parametric problems for independent data and we will discuss finite sample properties, in these notes. The main parametric families of distributions treated are the location and scale group families and the exponential families.

Example 1 (Methods for nonparametric inference.) *Density estimation problems, under smoothness assumptions or order assumptions. Regression problems under smoothness or order assumptions. Survival analysis for medical data.* □

Example 2 (Inference for dependent data.) *Inference for stochastic processes. Markov processes, ARMA time series. Inference for extremal events. Fatigue of materials. Inference in economic time series. Linkage analysis in statistical genetics. Microarray analysis.* □

Example 3 (Asymptotics) *Limit properties for (the two main processes in inference theory) the empirical process and the partial sum process. Optimal rates for estimation problems.* □

3.1 Estimation

Assume we are given an observation x of the random vector X taking it's values in the outcome space $\mathcal{X}$, which has a distribution P_θ, assumed to have a known form with an unknown parameter θ. The parameter θ is assumed to lie in a subset $\Omega \subset \mathbb{R}^s$, with $s < \infty$. The estimation problem consists of finding an estimate of the estimand $g(\theta)$, where g is a known real valued function defined on Ω,

$$g : \Omega \to \mathbb{R}.$$

Thus we restrict ourselves to looking at estimation of real-valued entities only.

Definition 1 *An estimator is a real valued function δ defined on $\mathcal{X}$,*

$$\delta : \mathcal{X} \to \mathbb{R}.$$

□

We want the estimator value $d = \delta(x)$ to be close to the estimand $g(\theta)$. A function satisfying

$$\begin{aligned} L(\theta, d) &\geq 0 \text{ for all } \theta, d, \\ L(\theta, g(\theta)) &= 0 \text{ for all } \theta, \end{aligned}$$

is called a loss function. It is used to measure the error made in the estimation problem when estimating the value $g(\theta)$ of the estimand with the value d of the estimator, based on the outcome x of X. Examples of measures of closeness are

$$\begin{aligned} L_1(\theta, d) &= |g(\theta) - d|^p, \\ L_2(\theta, d) &= 1\{|g(\theta) - d| > c\}, \end{aligned}$$

for a fixed p and a fixed c, respectively.

Viewing the estimator $\delta(X)$ as a random variable and the loss function $L(g(\theta), \delta(X))$ as a random variable, it is possible to calculate the average loss, with respect to P_θ, this gives the so called risk function

$$R(\theta, \delta) = E_\theta(L(\theta, \delta(X))) = \int_{\mathcal{X}} L(\theta, \delta(x))\, dP_\theta(x).$$

Note that the risk function depends on the loss function L, on the distribution P_θ of the random variable (and thus on θ) and on the estimator δ, i.e. on the rule (function) that specifies how to use the data to draw conclusions for $g(\theta)$.

Given an estimation problem $(\{P_\theta : \theta \in \Omega\}, g(\theta))$, one can ask if it is possible to find an estimator δ that minimizes the risk for all $\theta \in \Omega$? To answer this assume the loss function satisfies $L(\theta, d) > 0$ whenever $d \neq g(\theta)$, and let $\delta(x) = g(\theta_0)$. Then, since δ is a constant estimator, $R(\theta, \delta) = L(\theta, g(\theta_0))$ which is zero for $\theta = \theta_0$, and strictly positive for $\theta \neq \theta_0$. And of course for $\theta_1 \neq \theta_0$ the estimator $\delta_1(x) = g(\theta_1)$ has risk $R(\theta_1, \delta_1) = 0$ and $R(\theta, \delta_1) > 0$ when $\theta \neq \theta_1$. Thus the two estimators δ and δ_1 are overlapping in the sense that neither is uniformly better than the other. This in fact shows that *there is no uniformly best estimator*, i.e. there is no estimator that minimizes the risk (pointwise at θ) for all θ.

There are two possibilities to come to terms with this problem. One is to make a restriction of the possible estimators, so that one rules out anomalies as above. For instance one can demand that the estimator is unbiased, i.e. that the estimator satisfies

$$E_\theta(\delta(X)) \quad = \quad g(\theta),$$

for every $\theta \in \Omega$, or that the estimator satisfies and "equivariance" property. Then in certain situations, or for certain classes of distributions, it is possible to find an estimator that is uniformly best under the restriction.

Another possibility is to find estimators that minimize the risk in some "average" sense, instead of looking at pointwise risk, i.e. for each θ. This is done by using some global measure of the risk over all possible $\theta \in \Omega$. One example of a global measure is the weighted or Bayes risk

$$\int_\Omega R(\theta, \delta)\, d\Lambda(\theta),$$

where Λ is a probability measure on Ω. Then the corresponding estimator

$$\delta_\Lambda \quad = \quad \text{argmin}_\delta \int_\Omega R(\theta, \delta)\, d\Lambda(\theta)$$

is called the Bayes estimator. Another example of a global risk measure is the maximum risk

$$\sup_{\theta \in \Omega} R(\theta, \delta).$$

The estimator

$$\delta^* \quad = \quad \text{argmin}_\delta \sup_{\theta \in \Omega} R(\theta, \delta)$$

is called the minimax estimator.

There are two classes of distributions that we will study: Group families and exponential families.

3.2 Location/scale families (group families)

Assume U is a real valued random variable with a fixed known distribution function F. Assume a is an arbitrary real number and $b > 0$ is an arbitrary real positive number, and define the random variable

$$X = a + bU.$$

Then

$$\begin{aligned} P(X \leq x) &= P(a + bU \leq x) \\ &= F(\frac{x-a}{b}). \end{aligned}$$

Assuming that U has density f, the r.v. X has density

$$f(x) = \frac{1}{b} f(\frac{x-a}{b}).$$

The set $\mathcal{P} = \{F(\frac{x-a}{b}) : a \in \mathbb{R}, b \in \mathbb{R}^+\}$ is called a location-scale family of distributions. Note that F is assume to be known, whereas a, b are assumed unknown. Thus $P_\theta := F_\theta := F_{a,b}$ where $F_{a,b}(x) = F((x-a)/b)$ and $\Omega = \mathbb{R} \times \mathbb{R}_+$, and we have a two-parameter parametric inference problem

Now let $g_{a,b}$ be the transformation on $\mathbb{R}$

$$g_{a,b}(x) = a + bx,$$

for $a \in \mathbb{R}, b \in \mathbb{R}^+$ fixed, and let $\mathcal{G} = \{g_{a,b} : a \in \mathbb{R}, b \in \mathbb{R}^+\}$ be the set of such transformations. Define the operation $\times : \mathcal{G} \times \mathcal{G} \to \mathcal{G}$ by

$$(g_1 \times g_2)(x) = g_1(g_2(x)).$$

Then $\mathcal{G}$ is a group under $\times$, i.e.

(i) $\mathcal{G}$ is closed under $\times$,

(ii) every $g \in \mathcal{G}$ has an inverse,

(iii) there is an identity element in $\mathcal{G}$.

In fact, the first assertion follows since composition of two linear maps is linear, the second follows since the inverse to $g_{a,b}$ is $g_{-a/b,1/b}$ and the third since the identity element is $g_{0,1}$.

Example 4 *The Normal, Exponential and Uniform distributions are all location/scale,*

$$\begin{aligned} \mathcal{P} &= \{f(x) = \frac{1}{\sqrt{2\pi}b} e^{-\frac{(x-a)^2}{2b^2}} : a \in \mathbb{R}, b \in \mathbb{R}^+\}, \\ \mathcal{P} &= \{f(x) = \frac{1}{b} e^{-\frac{x-a}{b}} : a \in \mathbb{R}, b \in \mathbb{R}^+\}, \\ \mathcal{P} &= \{f(x) = \frac{1}{b} 1\{a - \frac{b}{2} < x < a + \frac{b}{2}\} : a \in \mathbb{R}, b \in \mathbb{R}^+\}. \end{aligned}$$

□

Special cases of location-scale families are location families and scale families

$$\{F(x-a) : a \in \mathbb{R}\},$$
$$\{F(x/b) : b \in \mathbb{R}^+\},$$

respectively.