GRADUATE COURSE IN STATISTICAL INFERENCE

LECTURE NOTES

Practical matters

- Lecturers: Henrik Hult and Timo Koski
  `hult@kth.se` `tjtkoski@kth.se`
- Lecture notes: Published on course web site.
- Course book: *Statistical Inference 2nd Ed.*, G. Casella and R. Berger, Duxbury, 2002.
- Additional reading: *Theory of Statistics*, M. Schervish, Springer, 1995, and *Information Theory, Inference, and Learning Algorithms*, D. Mackay, Cambridge University Press, 2003,

Lecture 1

1. Mathematical introduction to statistics

In this section we have a look at the mathematical foundation of statistics. Throughout the course we will try to first give an elementary introduction, not using measure theoretic probability, in order to get a sense of what is going on. In the elementary approach one can work with discrete or continuous densities. We will use the notation of continuous densities, but these can just as well be replaced by discrete ones and integrals replaced by sums.

1.1. **Elementary introduction.** Usually one starts with a number of observed data, $X = (X_1, \dots, X_n)$, where X_i are random variables with values in $\mathbb{R}$. The distribution of X is unknown but we assume it has a density that depend on an unknown parameter. We write Θ for the unknown parameter and think of it as a random variable representing the uncertainty of its value and assume that it takes values in the parameter space Ω. When the value of Θ is θ we write P_θ for the conditional distribution of X given $\Theta = \theta$. One may assume that there is a single "true value" θ of the parameter but this value is unknown.

We write $f_{X|\Theta}(x \mid \theta)$ for the conditional probability density of X given $\Theta = \theta$. The density $f_{X|\Theta}(x \mid \theta)$ is the basis of classical statistics. If one observes $X = x$, then the function $\theta \mapsto f_{X|\Theta}(x \mid \theta)$ is called the *likelihood function* and is used for making inferences.

Example 1 (Independent observations)**.** Suppose we have n observations of independent random variables $X_1, \dots, X_n$, each with density $f_{X_i|\Theta}(x_i \mid \theta)$ when $\Theta = \theta$. In this case $X = (X_1, \dots, X_n)$ and $\mu_{X|\Theta}(\cdot \mid \theta)$ is a probability measure on $\mathbb{R}^n$ with density

$$f_{X|\Theta}(x, \theta) = \prod_{i=1}^{n} f_{X_i|\Theta}(x_i \mid \theta).$$

After observing $(X_1, \dots, X_n) = (x_1, \dots, x_n)$ the likelihood function $L(\theta)$ is the function $\theta \mapsto \prod_{i=1}^{n} f_{X_i|\Theta}(x_i \mid \theta)$.

Example 2 (Independent and identically distributed observations)**.** If we in addition suppose that the independent random variables are identically distributed, then $f_{X_i|\Theta}(x_i \mid \theta) = f_{X_1|\Theta}(x_i \mid \theta)$ and P_θ is a probability measure on $\mathbb{R}^n$ with density

$$f_{X|\Theta}(x, \theta) = \prod_{i=1}^{n} f_{X_1|\Theta}(x_i \mid \theta).$$

1.2. **General introduction.** Let us now take a look at the general setting. We will use measure theory based probability and essentially repeat the "elementary introduction". This will enable a general framework where we can study the elements of statistics. In this section we will be more rigorous with the mathematical details.

To start let us take an underlying probability space $(S, \mathcal{A}, \mu)$. S is the abstract space of outcomes, $\mathcal{A}$ is a σ-field, and μ a probability measure. We will often use the notation Pr to denote the underlying probability measure μ. Suppose we do an experiment where the collected data takes values in the sample space $\mathcal{X}$ which has a σ-field $\mathcal{B}$. This space is denoted $(\mathcal{X}, \mathcal{B})$. The observed data is denoted by X, where $X : S \to \mathcal{X}$ is a random variable (i.e. it is a measurable mapping). We will use the term *random variable* in a general sense. That is $\mathcal{X}$ could be a general space. Often $\mathcal{X}$ will be some familiar space, for instance, X could be a vector of random variables $X = (X_1, \dots, X_n)$, in which case $\mathcal{X} = \mathbb{R}^n$, but it may also be a continuous stochastic process, in which case $\mathcal{X}$ is the space $C[0, 1]$ of continuous functions. The distribution of X (which is a probability measure on $\mathcal{B}$) is unknown but we assume that it belongs to $\mathcal{P}_0$ which is a parametric family of probability measures (probability distributions) on $\mathcal{B}$. The probability measures in the family $\mathcal{P}_0$ are indexed by a parameter θ taking values in the space Ω with σ-field τ. It is assumed that the parametric family $\mathcal{P}_0$ can be written as $\mathcal{P}_0 = \{P_\theta : \theta \in \Omega\}$. We assume that (i) for each $\theta \in \Omega$, $P_\theta(\cdot)$ is a probability measure on $\mathcal{B}$ and (ii) for each $B \in \mathcal{B}$ the function $\theta \mapsto P_\theta(B)$ is a measurable function on Ω.

Reminder: Let X and Θ be random variables on a probability space $(S, \mathcal{A}, \mu)$. Recall that (a version of) the conditional distribution of X given Θ is a mapping $\mu_{X\mid\Theta}$ on $\mathcal{B} \times \Omega$ such that

(i) for each $\theta \in \Omega$, $\mu_{X\mid\Theta}(\cdot \mid \theta)$ is a probability measure on $\mathcal{B}$.
(ii) for each $B \in \mathcal{B}$, $\mu_{X\mid\Theta}(B \mid \cdot)$ is a measurable function on Ω.

Note that we have defined P_θ to be a conditional distribution. An alternative is to represent the uncertainty of the parameter Θ as a random variable, i.e. as a measurable mapping from S to Ω. The joint distribution of (X, Θ) is then a probability measure on $\mathcal{B} \times \tau$ given by

$$\mu_{X,\Theta}(B) = \Pr(s : (X(s), \Theta(s)) \in B), \qquad B \in \mathcal{B} \times \tau.$$

Then one can define the conditional distribution of X given $\Theta = \theta$ and write P_θ for the conditional distribution of X given $\Theta = \theta$. Correspondingly we write E_θ for the expected value under P_θ. We will also use the notation $\mu_{X\mid\Theta}(\cdot \mid \theta)$ to denote the conditional distribution P_θ.

It should be noted that in the classical setup it is sufficient start directly with the family $\mathcal{P}_0$ without first defining Θ as a random variable. For instance, the classical paradigm never use the joint distribution of (X, Θ) or the marginal distribution of X or Θ. However, to fit the classical and the Bayesian into the same framework we will think about Θ as a random variable and P_θ as the conditional distribution of X given $\Theta = \theta$.

If, for each $\theta \in \Omega$, P_θ has a density $f_{X\mid\Theta}(x \mid \theta)$ (measureable $\mathcal{B} \times \tau$) with respect to a measure ν, that is $P_\theta \ll \nu$ and

$$f_{X\mid\Theta}(x \mid \theta) = \frac{dP_\theta}{d\nu}(x), \quad \text{for each } \theta \in \Omega,$$

then for fixed x, $\theta \mapsto f_{X\mid\Theta}(x \mid \theta)$ is called the *likelihood function* and is denoted $L(\theta)$. Usually the reference measure ν will be Lebesgue measure or counting measure but it can be more general.

Reminder: Recall that a measure μ is absolutely continuous with respect to a measure ν, written $\mu \ll \nu$, if $\nu(B) = 0$ implies $\mu(B) = 0$ and in that case the Radon-Nikodym Theorem guarantees the existence of a density $f(x) = \frac{d\mu}{d\nu}(x)$ such that

$$\int h(x)\mu(dx) = \int h(x)f(x)\nu(dx)$$

for each integrable function h.

2. Bayesian statistics

2.1. **Elementary Bayesian statistics.** In the Bayesian paradigm it is assumed that Θ is a random variable and some prior knowledge of the parameter Θ is available. The information about Θ is put into the model by specifying the *prior distribution* with density $f_\Theta(\theta)$. The densities $f_{X|\Theta}(x \mid \theta)$ and $f_\Theta(\theta)$ can be combined to obtain the joint density of (X, Θ) given by

$$f_{X,\Theta}(x, \theta) = f_{X|\Theta}(x \mid \theta) f_\Theta(\theta).$$

Once the joint density is specified we can also derive the marginal density of X

$$f_X(x) = \int_\Omega f_{X|\Theta}(x \mid \theta) f_\Theta(\theta) d\theta.$$

An important ingredient in the Bayesian paradigm is the *posterior distribution* given the observation $X = x$. Its density is given by Bayes' theorem as

$$f_{\Theta|X}(\theta \mid x) = \frac{f_{X,\Theta}(x,\theta)}{f_X(x)} = \frac{f_{X|\Theta}(x \mid \theta) f_\Theta(\theta)}{\int_\Omega f_{X|\Theta}(x \mid \theta) f_\Theta(\theta) d\theta}.$$

The posterior distribution of Θ given $X = x$ can be thought of as the updated beliefs about Θ after taking into accound the observation $X = x$.

In Bayesian statistics all inference in based on the posterior distribution. Note that the difference from classical statistics is that the posterior density is just the likelihood function multiplied by the prior density and then normalized to become a probability distribution.

2.2. **General framework.** The general Bayesian setup is similar to the classical case. We consider the data X and the parameter Θ as random variables. The joint distribution of (X, Θ) is denoted by $\mu_{X,\Theta}$. It is specified by choosing the marginal distribution μ_Θ of Θ, called the *prior distribution* and the conditional distribution $\mu_{X|\Theta}(\cdot \mid \theta)$, which we also denote by P_θ, from a parametric family $\mathcal{P}_0 = \{P_\theta : \theta \in \Omega\}$. Once the prior distribution and the conditional distribution are specified the joint distribution is given by

$$\mu_{X,\Theta}(B \times A) = \int_A \mu_{X|\Theta}(B \mid \theta) \mu_\Theta(d\theta).$$

The it is easy to derive the marginal distributions of X and Θ as

$$\mu_X(B) = \int_{\mathcal{X}\times\Omega} I_B(x) \mu_{X,\Theta}(dx, d\theta),$$

$$\mu_\Theta(A) = \int_{\mathcal{X}\times\Omega} I_A(\theta) \mu_{X,\Theta}(dx, d\theta).$$

If, for each θ, P_θ (or which is the same $\mu_{X|\Theta}(\cdot \mid \theta)$) has a density $f_{X|\Theta}(x \mid \theta)$ w.r.t. a measure ν, then we can write

$$P_\theta(B) = \mu_{X|\Theta}(B \mid \theta) = \int_B f_{X|\Theta}(x \mid \theta)\nu(dx).$$

Using Fubini's theorem the marginal distribution of X can be written as

$$\mu_X(B) = \int_\Omega \int_B f_{X|\Theta}(x \mid \theta)\nu(dx)\mu_\Theta(d\theta) = \int_B \Big[\int_\Omega f_{X|\Theta}(x \mid \theta)\mu_\Theta(d\theta)\Big]\nu(dx)$$

and we see that the density of μ_X w.r.t. ν is

$$f_X(x) = \int_\Omega f_{X|\Theta}(x \mid \theta)\mu_\Theta(d\theta).$$

If, in addition, μ_Θ has a density f_Θ w.r.t. a measure ρ on τ (recall that τ is the σ-field on the parameter space Ω) then the marginal density of X w.r.t. ν becomes

$$f_X(x) = \int_\Omega f_{X|\Theta}(x \mid \theta)f_\Theta(\theta)\rho(d\theta).$$

2.3. **Posterior distribution.** Once we have observed the data $X = x$, we can use Bayes' theorem to write down the conditional distribution of Θ given $X = x$. This distribution is called the *posterior distribution* and is of central importance in Bayesian statistics. Here is a general version of Bayes' theorem.

Theorem 1 (Bayes' theorem)**.** *Suppose there is a measure ν on $\mathcal{B}$ such that $P_\theta \ll \nu$ for each $\theta \in \Omega$ and let $f_{X|\Theta}(x \mid \theta)$ be the density. Let $\mu_{\Theta|X}(\cdot \mid x)$ be the conditional distribution of Θ given $X = x$. Then $\mu_{\Theta|X}(\cdot \mid x) \ll \mu_\Theta$ μ_X-a.s. and*

$$\frac{d\mu_{\Theta|X}}{d\mu_\Theta}(\theta \mid x) = \frac{f_{X|\Theta}(x \mid \theta)}{\int_\Omega f_{X|\Theta}(x \mid \vartheta)\mu_\Theta(d\vartheta)}$$

for those x such that the denominator is neither 0 nor ∞. Moreover, $\mu_X\{x : \int_\Omega f_{X|\Theta}(x \mid \vartheta)\mu_\Theta(d\vartheta) = 0$ or $\infty\} = 0$ and $\mu_{\Theta|X}$ can be artbitrarily defined on this set.

Remark 1. If the prior distribution μ_Θ has density f_Θ wrt a measure ρ on τ and P_θ has density $f_{X|\Theta}(\cdot \mid \theta)$ wrt ν on $\mathcal{B}$, then the posterior distribution of Θ given $X = x$ has a density (wrt ρ) given by

$$f_{\Theta|X}(\theta \mid x) = \frac{f_{X,\Theta}(x,\theta)}{\int_\Omega f_{X,\Theta}(x,\theta)\rho(d\theta)} = \frac{f_{X|\Theta}(x \mid \theta)f_\Theta(\theta)}{\int_\Omega f_{X|\Theta}(x \mid \theta)f_\Theta(\theta)\rho(d\theta)}.$$

Density proof. Suppose that all relevant densities exists and that densities are w.r.t. Lebesgue measure. Then Bayes' theorem simply says that

$$f_{\Theta|X}(\theta \mid x) = \frac{f_{X|\Theta}(x \mid \theta)f_\Theta(\theta)}{\int_\Omega f_{X|\Theta}(x \mid \theta)f_\Theta(\theta)d\theta}. \tag{2.1}$$

This is just a consequence of the "elementary" definition of conditional density as $f_{X|\Theta}(x \mid \theta) = f_{X,\Theta}(x,\theta)/f_\Theta(\theta)$. We need to watch out that we do not plug in values of x where $f_X(x) = 0$ or ∞ in (2.1), but that should not be big a concern since if C is the set of those values, then we must have $\Pr(X \in C) = 0$. □

For the sake of completeness, here is a formal proof in the general case.

Proof. Let us start with the second claim. Write

$$C_0 = \{x : \int_\Omega f_{X|\Theta}(x \mid \vartheta)\mu_\Theta(d\vartheta) = 0\}$$
$$C_\infty = \{x : \int_\Omega f_{X|\Theta}(x \mid \vartheta)\mu_\Theta(d\vartheta) = \infty\}$$

and note that

$$\mu_X(C_0) = \int_{C_0}\int_\Omega f_{X|\Theta}(x \mid \vartheta)\mu_\Theta(d\vartheta)\nu(dx) = 0.$$

For C_∞ we have

$$\infty > \mu_X(C_\infty) = \int_{C_\infty}\int_\Omega f_{X|\Theta}(x \mid \vartheta)\mu_\Theta(d\vartheta)\nu(dx).$$

Hence, we must have $\nu(C_\infty) = 0$ and then it follows that $\mu_X(C_\infty) = 0$.

To prove the claim for the Radon-Nikodym density observe that for $B \in \mathcal{B}$ and $A \in \tau$ we have on one hand

$$\begin{aligned}\Pr(X \in B, \Theta \in A) &= \mu_{X,\Theta}(B \times A)\\ &= \int_B \mu_{\Theta|X}(A \mid x)\mu_X(dx)\\ &= \int_B \Big[\mu_{\Theta|X}(A \mid x)\int_\Omega f_{X|\Theta}(x \mid \vartheta)\mu_\Theta(d\vartheta)\Big]\nu(dx).\end{aligned}$$

On the other hand we have by Fubini's theorem

$$\begin{aligned}\Pr(X \in B, \Theta \in A) = \int_A \mu_{X|\Theta}(B \mid \vartheta)\mu_\Theta(d\vartheta) &= \int_A \Big[\int_B f_{X|\Theta}(x \mid \vartheta)\nu(dx)\Big]\mu_\Theta(d\vartheta)\\ &= \int_B \Big[\int_A f_{X|\Theta}(x \mid \vartheta)\mu_\Theta(d\vartheta)\Big]\nu(dx).\end{aligned}$$

Combining these two we see that ν-a.e. (and hence μ_X-a.s.)

$$\mu_{\Theta|X}(A \mid x) = \frac{\int_A f_{X|\Theta}(x \mid \vartheta)\mu_\Theta(d\vartheta)}{\int_\Omega f_{X|\Theta}(x \mid \vartheta)\mu_\Theta(d\vartheta)} = \int_A \frac{f_{X|\Theta}(x \mid \theta)}{\int_\Omega f_{X|\Theta}(x \mid \vartheta)\mu_\Theta(d\vartheta)}\mu_\Theta(d\theta).$$

In particular $\mu_{\Theta|X}(\cdot \mid x) \ll \mu_\Theta$ μ_X-a.s. and the Radon-Nikodym density is the desired one. □

Remark 2. Generalized prior distributions.

2.4. **Posterior predictive distribution of future values.** Suppose that $X = (X_1, \dots, X_n)$ and we have observed $X = x = (x_1, \dots, x_n)$. To compute the probability of future events, the Bayesian methodology proposes to use

$$\begin{aligned}&\Pr(X_{n+1} \in A_1, \dots, X_{n+k} \in A_k \mid X = x)\\ &\quad = \int_\Omega \Pr(X_{n+1} \in A_1, \dots, X_{n+k} \in A_k \mid \Theta = \theta, X = x)\mu_{\Theta|X}(d\theta \mid x).\end{aligned}$$

This distribution is called the *posterior predictive distribution of future values.*

Example 3. If the X_i's are assumed conditionally IID given $\Theta = \theta$ then the posterior predictive distribution of future values is given by

$$\Pr(X_{n+1} \in A_1, \dots, X_{n+k} \in A_k \mid X = x) = \int_\Omega \prod_{i=1}^{k} \mu_{X_i|\Theta}(A_i \mid \theta) \mu_{\Theta|X}(d\theta \mid x).$$

If $\mu_{X_1|\Theta}(\cdot \mid \theta)$ has a density $f_{X_1|\Theta}(x \mid \theta)$ wrt a measure ν then the posterior predictive distribution has density

$$\begin{aligned} &f_{X_{n+1},\dots,X_{n+k}|X_1,\dots,X_n}(x_{n+1}, \dots, x_{n+k} \mid x_1, \dots, x_n) \\ &= \int_\Omega \prod_{i=1}^{k} f_{X_1|\Theta}(x_{n+i} \mid \theta) \mu_{\Theta|X}(d\theta \mid x_1, \dots, x_n). \end{aligned}$$