SF3961 GRADUATE COURSE IN STATISTICAL INFERENCE

HENRIK HULT

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, $X = (X_1, \ldots, X_n)$, 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 $X_1, \ldots, X_n$, each with density $f_{X_i|\Theta}(x_i \mid \theta)$ when $\Theta = \theta$. In this case $X = (X_1, \ldots, X_n)$ 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 $(X_1, \ldots, X_n) = (x_1, \ldots, x_n)$ 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

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

Date: Lecture notes 2015.

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 $X = (X_1, \ldots, X_n)$, 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 $\mu_{X|\Theta}$ on $\mathcal{B} \times \Omega$ such that

(i) for each $\theta \in \Omega$, $\mu_{X|\Theta}(\cdot \mid \theta)$ is a probability measure on $\mathcal{B}$.
(ii) for each $B \in \mathcal{B}$, $\mu_{X|\Theta}(B \mid \cdot)$ 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 $\mu_{X|\Theta}(\cdot \mid \theta)$ 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 $f_{X|\Theta}(x \mid \theta)$ (measureable $\mathcal{B} \times \tau$) with respect to a measure ν, that is $P_\theta \ll \nu$ and

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

then for fixed x, $\theta \mapsto f_{X|\Theta}(x \mid \theta)$ 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, $\mu_X\{x : \int_\Omega f_{X|\Theta}(x \mid \vartheta)\mu_\Theta(d\vartheta) = 0 \text{ or } \infty\} = 0$ 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 $X = (X_1, \ldots, X_n)$ and we have observed $X = x = (x_1, \ldots, x_n)$. To compute the probability of future events, the Bayesian methodology proposes to use

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

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

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

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

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

Lecture 2

3. Some common distributions in classical and Bayesian statistics

3.1. **Conjugate prior distributions.** In the Bayesian setting it is important to compute posterior distributions. This is not always an easy task. The main difficulty is to compute the normalizing constant in the denominator of Bayes theorem. However, for certain parametric families $\mathcal{P}_0 = \{P_\theta : \theta \in \Omega\}$ there are convenient choices of prior distributions. Particularly convenient is when the posterior belongs to the same family of distributions as the prior. Such families are called conjugate families.

Definition 1. Let $\mathcal{F}$ denote a class of probability densities $f(x \mid \theta)$. A class Π of prior distributions is a *conjugate family* for $\mathcal{F}$ if the posterior distribution is in the class Π for all $f \in \mathcal{F}$, all priors in Π, and all $x \in \mathcal{X}$.

(See Exercise 7.22, 7.23, 7.24 in Casella & Berger)

Example 4 (Casella & Berger, Example 7.2.14)**.** Let $X_1, \ldots, X_n$ be IID Ber(θ) given $\Theta = \theta$ and put $Y = \sum_{i=1}^n X_i$. Then $Y \sim \text{Bin}(n, \theta)$. Let the prior distribution be Beta(α, β). Then the posterior of Θ given $Y = y$ is Beta($y + \alpha, n - y + \beta$).

The joint density is

$$\begin{aligned} f_{Y,\Theta}(y, \theta) &= f_{Y|\Theta}(y \mid \theta) f_\Theta(\theta) \\ &= \binom{n}{y} \theta^y (1-\theta)^{n-y} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &= \binom{n}{y} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1}. \end{aligned}$$

The marginal density of Y is

$$\begin{aligned} f_Y(y) &= \int_0^1 f_{Y,\Theta}(y, \theta) d\theta \\ &= \binom{n}{y} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 \theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1} d\theta \\ &= \binom{n}{y} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(y+\alpha)\Gamma(n-y+\beta)}{\Gamma(n+\alpha+\beta)} \end{aligned}$$

(this distribution is known as the Beta-binomial distribution). The posterior is then computed as

$$f_{\Theta|Y}(\theta \mid y) = \frac{f_{Y,\Theta}(y, \theta)}{f_Y(y)} = \frac{\Gamma(n+\alpha+\beta)}{\Gamma(y+\alpha)\Gamma(n-y+\beta)} \theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1}.$$

This is the density of the Beta($y + \alpha, n - y + \beta$) distribution.

4. Exponential families

Exponential families of distributions are perhaps the most widely used family of distributions in statistics. It contains most of the common distributions that we know from undergraduate statistics.

Definition 2. A parametric family of distributions $\mathcal{P}_0 = \{P_\theta : \theta \in \Omega\}$ with parameter space Ω and conditional density $f_{X|\Theta}(x \mid \theta)$ with respect to a measure ν is called an exponential family if

$$f_{X|\Theta}(x \mid \theta) = c(\theta)h(x)\exp\Big\{\sum_{i=1}^{k}\pi_i(\theta)t_i(x)\Big\}$$

for some measurable functions $c, h, \pi_1, \ldots, \pi_k, t_1, \ldots, t_k$, and some integer k.

Example 5. If X are conditionally IID Exp(θ) given $\Theta = \theta$ then it follows that $f_{X|\Theta}(x \mid \theta) = \theta^{-n}\exp\{-\theta^{-1}\sum_{i=1}^{n} x_i\}$ so this is an one-dimensional exponential family with $c(\theta) = \theta^{-n}$, $h(x) = 1$, $\pi(\theta) = 1/\theta$, and $t(x) = x_1 + \cdots + x_n$.

Example 6. If X are conditionally IID Ber(θ), then with $m = x_1 + \cdots + x_n$, we have

$$f_{X|\Theta}(x \mid \theta) = \theta^m(1-\theta)^{n-m} = (1-\theta)^n\Big(\frac{\theta}{1-\theta}\Big)^m = (1-\theta)^n\exp\Big\{\log\Big(\frac{\theta}{1-\theta}\Big)m\Big\}.$$

so this is also a one-dimensional exponential family with $c(\theta) = (1-\theta)^n$, $h(x) = 1$, $\pi(\theta) = \log(\theta/(1-\theta))$, and $t(x) = x_1 + \cdots + x_n$.

There are many other examples as the Normal, Poisson, Gamma, Beta distributions (see Casella & Berger, Exercise 3.28, p. 132).

Note that the function $c(\theta)$ can be thought of as a normalizing function to make $f_{X|\Theta}$ a probability density. It is necessary that

$$c(\theta) = \Big(\int_{\mathcal{X}} h(x)\exp\Big\{\sum_{i=1}^{k}\pi_i(\theta)t_i(x)\Big\}\nu(dx)\Big)^{-1}$$

so the dependence on θ comes through the vector $(\pi_1(\theta), \ldots, \pi_k(\theta))$ only. It is useful to have a name for this vector; it will be called the *natural parameter*.

Definition 3. For an exponential family the vector $\Pi = (\pi_1(\Theta), \ldots, \pi_k(\Theta))$ is called the *natural parameter* and

$$\Gamma = \Big\{\pi \in \mathbb{R}^k : \int_{\mathcal{X}} h(x)\exp\Big\{\sum_{i=1}^{k}\pi_i t_i(x)\Big\}\nu(dx) < \infty\Big\}$$

the *natural parameter space*.

When we deal with an exponential family it is convenient to use the notation $\Theta = (\Theta_1, \ldots, \Theta_k)$ for the natural parameter and Ω for the parameter space. Therefore we will often write

$$f_{X|\Theta}(x \mid \theta) = c(\theta)h(x)\exp\Big\{\sum_{i=1}^{n}\theta_i t_i(x)\Big\} \tag{4.1}$$

and Ω for the natural parameter space Γ and hope that this does not cause confusion.

For an example on how to write the normal distribution as an exponential family with its natural parametrisation see Examples 3.4.4 and 3.4.6, pp. 112-113 in Casella & Berger.

4.1. **Conjugate priors for exponential families.** Let us take a look at conjugate priors for exponential families. Suppose that the conditional distribution of $X = (X_1, \dots, X_n)$ given $\Theta = \theta$ forms a natural exponential family (4.1). We will look for a "natural" family of priors that serves as conjugate priors. If $f_\Theta(\theta)$ is a prior (w.r.t. a measure ρ on Ω) then the posterior has density

$$\begin{aligned} f_{\Theta|X}(\theta \mid x) &= \frac{f_{X|\Theta}(x \mid \theta) f_\Theta(\theta)}{\int_\Omega f_{X|\Theta}(x \mid \theta) f_\Theta(\theta) \rho(d\theta)} \\ &= \frac{c(\theta) e^{\sum_{i=1}^k \theta_i t_i(x)} f_\Theta(\theta)}{\int_\Omega c(\theta) e^{\sum_{i=1}^k \theta_i t_i(x)} f_\Theta(\theta) \rho(d\theta)}. \end{aligned}$$

Then a natural choice for the conjugate family is densities of the form

$$f_\Theta(\theta) = \frac{c(\theta)^\alpha e^{\sum_{i=1}^k \theta_i \beta_i}}{\int_\Omega c(\theta)^\alpha e^{\sum_{i=1}^k \theta_i \beta_i} \rho(d\theta)},$$

where $\alpha > 0$ and $\beta = (\beta_1, \dots, \beta_k)$.

Indeed, the posterior is then proportional to

$$c(\theta)^{\alpha+1} \exp\Big\{ \sum_{i=1}^k \theta_i (t_i(x) + \beta_i) \Big\}$$

which is of the same form as the prior (after putting in the right normalizing constant). Note that the posterior is an exponential family with natural parameter $\xi = t + \beta$ and representation

$$c'(\xi) h'(\theta) \exp\Big\{ \sum_{i=1}^k \xi_i \theta_i \Big\},$$

where $h'(\theta) = c(\theta)^{\alpha+1}$ and $c'(\xi)$ is the normalizing constant to make it a probability density.

Example 7. Take another look at the family of n iid Ber(p) variables (see Example 6. The natural parameter is $\theta = \log(p/1-p)$ and then $c(\theta) = (1-p)^n = (1+e^\theta)^{-n}$. Then the proposed conjugate prior is proportional to

$$c(\theta)^\alpha e^{\theta\beta} = (1-p)^{\alpha n} p^\beta (1-p)^{-\beta} = p^\beta (1-p)^{\alpha n - \beta}$$

which is a Beta($\beta + 1, \alpha n - \beta + 1$) distribution (when you put in the normalization). So again we see that Beta-distributions are conjugate priors for IID Bernoulli random variables (here α and β are not the same as in Example 4, though).

4.2. **Some properties of exponential families.** The random vector $T(X) = (t_1(X), \dots, t_k(X))$ is of great importance for exponential families. We can compute the distribution and density of T with respect to a measure ν'_T to be introduced.

Suppose $P_\theta \ll \nu$ for all θ with density $f_{X|\Theta}$ as above. Let us write

$$g(\theta, T(x)) = c(\theta) \exp\Big\{ \sum_{i=1}^k \theta_i t_i(x) \Big\},$$

so $f_{X|\Theta}(x \mid \theta) = h(x) g(\theta, T(x))$. Write $\mathcal{T}$ for the space where T takes its values and $\mathcal{C}$ for the σ-field on $\mathcal{T}$. Introduce the measure $\nu'(B) = \int_B h(x) \nu(dx)$ for $B \in \mathcal{B}$

and $\nu'_T(C) = \nu' \circ T^{-1}(C)$ for $C \in \mathcal{C}$. Then we see that

$$\begin{aligned}\mu_{T|\Theta}(C \mid \theta) &= \mu_{X|\Theta}(T^{-1}C \mid \theta) \\ &= \int_{T^{-1}C} f_{X|\Theta}(x \mid \theta)\nu(dx) \\ &= \int_{T^{-1}C} g(\theta, T(x))\nu'(dx) \\ &= \int_C g(\theta, t)\nu'_T(dt).\end{aligned}$$

Hence, $\mu_{T|\Theta}(\cdot \mid \theta)$ has a density $g(\theta, t)$ with respect to ν'_T. This is nothing but rewriting the density of an exponential family but it truns out to be useful when studying properties of an exponential family.

In concrete situations one may identify what ν'_T is. Here is an example.

Example 8. Consider the exponential family of n IID Exp(θ) random variables as in Example 5. Then $t(x) = x_1 + \cdots + x_n$ and $T = t(X)$ has $\Gamma(n, \theta)$ distribution. Thus, T has density w.r.t. Lebesgue measure which is

$$f_{T|\Theta}(t \mid \theta) = \theta^{-n} e^{-t/\theta} \frac{t^{n-1}}{\Gamma(n)}.$$

In this case we can identify $c(\theta) = \theta^{-n}$ and hence $g(\theta, t) = \theta^{-n} e^{-t/\theta}$ and ν'_T must have density $t^{n-1}/\Gamma(n)$ w.r.t. Lebesgue measure. This is also possible to verify another way. Since $\nu'_T(B) = \nu(T^{-1}B)$ where ν is Lebesgue measure we see that

$$\begin{aligned}\nu'_T([0, t]) &= \nu\{x \in [0, \infty)^n : 0 \le x_1 + \cdots + x_n \le t\} \\ &= \int_{0 \le x_1 + \cdots + x_n \le t} dx_1 \dots dx_n = t^n/n!.\end{aligned}$$

Differentiating this w.r.t. t gives the density $t^{n-1}/\Gamma(n)$ with respect to Lebesgue measure (Recall that $\Gamma(n) = (n-1)!$).

Theorem 2. *The moment generating function $M_T(u)$ of T for an exponential family is given by*

$$M_T(u) = M_T(u_1, \dots, u_k) = \frac{c(\theta)}{c(u + \theta)}.$$

Proof. Since

$$\begin{aligned}c(\theta) &= \Big(\int_{\mathcal{X}} h(x) \exp\Big\{ \sum_{i=1}^k \theta_i t_i(x) \Big\} \nu(dx) \Big)^{-1} \\ &= \Big(\int_{\mathcal{T}} \exp\Big\{ \sum_{i=1}^k \theta_i t_i \Big\} \nu'_T(dt) \Big)^{-1}\end{aligned}$$

it follows that

$$M_T(u) = E_\theta\Big[\exp\Big\{ \sum_{i=1}^k u_i T_i \Big\} \Big] = \int_{\mathcal{T}} \exp\Big\{ \sum_{i=1}^k u_i t_i \Big\} c(\theta) \exp\Big\{ \sum_{i=1}^k \theta_i t_i \Big\} \nu'_T(dt) = \frac{c(\theta)}{c(u + \theta)}.$$

□

Hence, whenever θ is in the interior of the parameter space all moments of T are finite and can be computed. We call the function

$$\kappa(\theta) = -\log c(\theta)$$

the *cumulant function*. The cumulant function is useful to compute moments of $T(X)$.

Theorem 3. *For an exponential family with cumulant function κ we have*

$$E_\theta[T_i] = \frac{\partial}{\partial \theta_i}\kappa(\theta),$$
$$\operatorname{cov}_\theta(T_i, T_j) = \frac{\partial^2}{\partial\theta_i\partial\theta_j}\kappa(\theta).$$

Proof. For the mean we have

$$\begin{aligned} E_\theta[T_i] &= \frac{\partial}{\partial u_i} M_T(u)\Big|_{u=0} \\ &= \frac{\partial}{\partial u_i}\frac{c(\theta)}{c(u+\theta)}\Big|_{u=0} \\ &= -\frac{\frac{\partial}{\partial\theta_i}c(\theta)}{c(\theta)} \\ &= -\frac{\partial}{\partial\theta_i}\log c(\theta). \end{aligned}$$

The proof for the covariance is similar. □

Theorem 4. *The natural parameter space Ω of an exponential family is convex and $1/c(\theta)$ is a convex function.*

Proof. Let $\theta_1 = (\theta_{11}, \ldots, \theta_{1k})$ and $\theta_2 = (\theta_{21}, \ldots, \theta_{2k})$ be points in Ω and $\lambda \in (0,1)$. Then, since the exponential function is convex

$$\begin{aligned} \frac{1}{c(\lambda\theta_1 + (1-\lambda)\theta_2)} &= \int_{\mathcal{X}} h(x)\exp\{\sum_{i=1}^n [\lambda\theta_{1i} + (1-\lambda)\theta_{2i}]t_i(x)\}\nu(dx) \\ &\le \int_{\mathcal{X}} h(x)[\lambda\exp\{\sum_{i=1}^n \theta_{1i}t_i(x)\} + (1-\lambda)\exp\{\sum_{i=1}^n \theta_{2i}t_i(x)\}\nu(dx) \\ &= \lambda\frac{1}{c(\theta_1)} + (1-\lambda)\frac{1}{c(\theta_2)}. \end{aligned}$$

Hence, $1/c$ is convex. Since $\theta \in \Omega$ if $1/c(\theta) < \infty$ it follows also that $\lambda\theta_1 + (1-\lambda)\theta_2 \in \Omega$. Thus, Ω is convex. □

4.3. **Exponential tilting.** Let X be a random variable with moment generating function $M(u) = E[e^{uX}] < \infty$. Then the probability distribution given by

$$P_u(B) = \frac{E[e^{uX}I\{X \in B\}]}{M(u)},$$

is called an *exponentially tilted* distribution. If X has a density f w.r.t. a measure ν then P_u has density w.r.t. ν given by

$$f_u(y) = \frac{e^{uy}f(y)}{M(u)}.$$

Now, if $f(y)$ is the density of a natural exponential family, $f(y) = c(\theta)h(y)\exp\{\theta y\}$, then the density of the exponentially tilted distribution is

$$f_u(y) = \frac{e^{uy}f(y)}{M(u)} = \frac{c(\theta)h(y)\exp\{(\theta+u)y\}}{c(\theta)/c(\theta+u)} = c(\theta+u)h(y)\exp\{(\theta+u)y\}.$$

Hence, for an exponential family, exponential tilting by u is identical to shifting the parameter by u.

This also suggests how to construct exponential families; start with a probability distribution μ with density f and consider the family of all exponential tilts. This forms an exponential family. Indeed, if we tilt f by θ the resulting density is

$$f_\theta(x) = \frac{1}{M(\theta)}f(y)\exp\{\theta y\},$$

so putting $c(\theta) = 1/M(\theta)$ and $h(y) = f(y)$ yields the representation of a natural exponential family.

4.4. **Curved exponential family.** Consider for example the family $\{N(\theta,\theta^2); \theta \in \mathbb{R}\}$. Is this an exponential family? Let us check.

The density is given by

$$\begin{aligned} f_{X|\Theta}(x \mid \theta) &= \frac{1}{\sqrt{2\pi}\theta}\exp\Big\{-\frac{1}{2\theta^2}(x-\theta)^2\Big\} \\ &= \frac{1}{\sqrt{2\pi}\theta}\exp\Big\{-\frac{1}{2}\Big\}\exp\Big\{-\frac{x^2}{2\theta^2}+\frac{x}{\theta}\Big\}. \end{aligned}$$

This is an exponential family with $\pi_1(\theta) = 1/(2\theta^2)$ and $\pi_2(\theta) = 1/\theta$. Hence, the natural parameter $\pi = (\pi_1, \pi_2)$ can only take values on a curve. Such a family will be called a curved exponential family.

Definition 4. A parametric family of distributions $\mathcal{P}_0 = \{P_\theta : \theta \in \Omega\}$ with parameter space Ω is called a *curved exponential family* if it is an exponential family, i.e.

$$f_{X|\Theta}(x \mid \theta) = c(\theta)h(x)\exp\Big\{\sum_{i=1}^{k}\pi_i(\theta)t_i(x)\Big\},$$

and the dimension d of the vector θ satisfies $d < k$.

If $d = k$ the family is called a *full exponential family.*

5. Location-scale families

In the last section we saw that exponential families are generated by starting with a particular density and then considering the family of all exponential tilts. In this section we will see what happens if we instead of exponential tilts simply shift and scale the random variable, i.e. we do linear transformations.

Exercise: Let X have a probability density f. Consider $Y = \sigma X + \mu$ for some $\sigma > 0$ and $\mu \in \mathbb{R}$. What is the density of Y?

Theorem 5. *Let f be a probability density and μ and $\sigma > 0$ be constants. Then*

$$g(x \mid \mu, \sigma) = \frac{1}{\sigma}f\Big(\frac{x-\mu}{\sigma}\Big)dx$$

is a probability density.

Proof. Casella and Berger p. 116.. □

Definition 5. Let f be a probability density.

(i) The family of probability densities $\{f(x-\mu);\mu\in\mathbb{R}\}$ is a *location family* with *location parameter* μ.
(ii) The family of probability densities $\{f(x/\sigma)/\sigma;\sigma>0\}$ is a *scale family* with *scale parameter* σ.
(iii) The family of probability densities $\{f((x-\mu)/\sigma)/\sigma;\mu\in\mathbb{R},\sigma>0\}$ is a *location-scale family* with *location parameter* μ and *scale parameter* σ.

Example 9. The family of normal distributions $N(\mu,\sigma)$ is a location-scale family. Indeed, with φ being the standard normal density,

$$\varphi_{\mu,\sigma}(x)=\frac{1}{\sigma\sqrt{2\pi}}\exp\{-(x-\mu)^2/(2\sigma^2)\}=\frac{1}{\sigma}\varphi((x-\mu)/\sigma))$$

Before getting deeper into the fundamentals of statistics we take a look at some distributions that appear frequently in statistics. These distributions will provide us with examples throughout the course.

Additional material that you probably know...

5.1. **Normal, Chi-squared, t, and F.** Here we look at some common distributions and their relationship. From elementary statistics courses it is known that the sample mean

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

of IID random variables $X_1, \dots, X_n$ can be used to estimate the expected value EX_i and that the sample variance

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

is used to estimate the variance $\mathrm{Var}(X_i)$.

The distribution of $\bar{X}_n$ and S is important in the construction of confidence intervals and hypothesis tests. The most popular situation is when $X_1, \dots, X_n$ are IID $\mathrm{N}(\mu, \sigma^2)$. The following result may be familiar.

Lemma 1. *Let $X_1, \dots, X_n$ be IID $N(\mu, \sigma^2)$. Then, $\bar{X}$ and S^2 are independent and*

$$\bar{X} \sim N(\mu, \sigma^2/n), \tag{5.1}$$

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1), \tag{5.2}$$

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1). \tag{5.3}$$

Moreover, if $\tilde{X}_1, \dots, \tilde{X}_m$ is IID $N(\tilde{\mu}, \tilde{\sigma}^2)$ and independent of $X_1, \dots, X_n$, then

$$\frac{S^2}{\sigma^2} \cdot \frac{\tilde{\sigma}^2}{\tilde{S}^2} \sim F(n-1, m-1). \tag{5.4}$$

It is a **good exercise** to prove the above lemma. If you get stuck, Section 5.3 in Casella & Berger contains the proof.

As a reminder we will show how Lemma 1 is used in undergraduate statistics. Suppose we have a sample $X = (X_1, \dots, X_n)$ that have IID $\mathrm{N}(\mu, \sigma^2)$ distribution.

5.1.1. *Confidence interval for μ with σ known.* If we estimate μ by $\bar{X}_n$ and σ is known, then we can use (5.1) to derive a $(1-\alpha)$-confidence interval for μ of the form $\bar{X}_n \pm \frac{\sigma}{\sqrt{n}} z_{\alpha/2}$, where z_α is such that $\Phi(z_\alpha) = 1 - \alpha$. Indeed,

$$P_\mu\Big(\bar{X}_n - \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \le \mu \le \bar{X}_n + \frac{\sigma}{\sqrt{n}} z_{\alpha/2}\Big) = P_\mu\Big(-z_{\alpha/2} \le \frac{\bar{X}_n - \mu}{\sigma\sqrt{n}} \le z_{\alpha/2}\Big) = 1 - \alpha.$$

5.1.2. *Confidence interval for μ with σ unknown.* If we estimate μ by $\bar{X}_n$ and σ is unknown, then we can estimate σ by S and use (5.3) to derive a $(1-\alpha)$-confidence interval for μ of the form $\bar{X}_n \pm \frac{S}{\sqrt{n}} t_{\alpha/2}$, where t_α is such that $t(z_\alpha) = 1 - \alpha$ and $t(x)$ is the cdf of the t-distribution with $n-1$ degrees of freedom. Indeed,

$$P_{\mu,\sigma}\Big(\bar{X}_n - \frac{S}{\sqrt{n}} t_{\alpha/2} \le \mu \le \bar{X}_n + \frac{S}{\sqrt{n}} t_{\alpha/2}\Big) = P_{\mu,\sigma}\Big(-t_{\alpha/2} \le \frac{\bar{X}_n - \mu}{S\sqrt{n}} \le t_{\alpha/2}\Big) = 1 - \alpha.$$

Lecture 3

6. De Finetti's theorem

In this lecture we discuss how one can replace the common assumption that there exists a parameter Θ such that the observations $X_1, \ldots, X_n$ are conditionally IID given $\Theta = \theta$, by something which (at first) looks weaker and is called exchangeability. We do the analysis for Bernoulli random variables. All the references in this lecture refers to Schervish "Theory of Statistics".

6.1. **Exchangeability.** First we need to know what exchangeable is. The interpretation of an exchangeable collection of random variables is that it does not matter in which order we collected the data.

Definition 6. (i) A finite set $X_1, \ldots, X_n$ of random elements is *exchangeable* if $(X_1, \ldots, X_n) \stackrel{\mathrm{d}}{=} (X_{\pi(1)}, \ldots, X_{\pi(n)})$ for every permutation π of the numbers $1, \ldots, n$.

(ii) An infinite collection is exchangeable if every finite subcollection is exchangeable.

Note that if $X_1, \ldots, X_n$ is exchangeable, then the X_i's are identically distributed.

Example 10. A finite or infinite collection of IID random variables is exchangeable. Indeed, for any permutation π of $\{1, \ldots, n\}$

$$\begin{aligned}\Pr(X_1 \le x_1, \ldots, X_n \le x_n) &= \prod_{i=1}^{n} \Pr(X_i \le x_i) = \prod_{i=1}^{n} \Pr(X_{\pi(i)} \le x_i) \\ &= \Pr(X_{\pi(1)} \le x_1, \ldots, X_{\pi(n)} \le x_n).\end{aligned}$$

Since this holds for any n we conclude that also an infinite collection if IID random variables is exchangeable.

Example 11. A finite or infinite collection $\{X_n\}$ of random variables that are IID conditioned on some random variable Y is exchangeable. Indeed, for any $n \ge 1$ and any permutation π of $\{1, \ldots, n\}$ we have

$$\begin{aligned}P(X_1 \le x_1, \ldots, X_n \le x_n) &= E[P(X_1, \le x_1, \ldots, X_n \le x_n \mid Y)] \\ &= E[P(X_{\pi(1)}, \le x_1, \ldots, X_{\pi(n)} \le x_n \mid Y)] \\ &= P(X_{\pi(1)}, \le x_1, \ldots, X_{\pi(n)} \le x_n).\end{aligned}$$

This is the "typical" form of exchangeability. Sometimes exchangeability is not so obvious.

Example 12 (Polya's urn, Example 1.15 p. 9)**.** Suppose we have balls of k different colors (denoted $1, \ldots, k$) and an urn containing u_i balls of color i, $i = 1, \ldots, k$. We assume $u = \sum_{i=1}^{k} u_i > 0$, so the urn is nonempty. We draw balls according to the following rule. Draw a ball at random (uniformly among all the balls in all urns) and let X_1 be the color of that ball. Replace the ball and add another ball with the same color. Then we repeat the previous step and record X_2 as the color for the second ball etc. We claim that the sequence $\{X_n\}$ is exchangeable. To see this, let $\mathcal{X} = \{1, \ldots, k\}$ (the colors), $n > 0$ an integer, and $j_1, \ldots, j_n$ elements of

$\mathcal{X}$. We want to compute $\Pr(X_1 = j_1, \dots, X_n = j_n)$. For $i = 1, \dots, k$ we denote by $c_i(j_1, \dots, j_n)$ the number of occasions of i among $j_1, \dots, j_n$. Formally

$$c_i(j_1, \dots, j_n) = \sum_{t=1}^{n} I_{\{i\}}(j_t).$$

With the notation $(a)_t = a(a-1)\dots(a-t+1)$ for integer $t \geq 1$ and $(a)_0 = 1$ we claim that

$$\Pr(X_1 = j_1, \dots, X_n = j_n) = \frac{\prod_{i=1}^{k}(u_i + c_i(j_1, \dots, j_n) - 1)_{c_i(j_1,\dots,j_n)}}{(u+n-1)_n}.$$

If the claim is true then we see that $\{X_1, \dots, X_n\}$ are exchangeable for each n because $c_i(j_1, \dots, j_n)$ does not change if we apply a permutation to $j_1, \dots, j_n$. We can prove the claim by induction. For $n = 1$ this is $\Pr(X_1 = j_1) = u_{j_1}/u$, which is true. Suppose it holds for n. Then

$$\begin{aligned}
&\Pr(X_1 = j_1, \dots, X_{n+1} = j_{n+1}) \\
&\quad = \Pr(X_1 = j_1, \dots, X_n = j_n) \Pr(X_{n+1} = j_{n+1} \mid X_1 = j_1, \dots, X_n = j_n) \\
&\quad = \Pr(X_1 = j_1, \dots, X_n = j_n) \frac{u_{j_{n+1}} + c_{j_{n+1}}(j_1, \dots, j_{n+1}) - 1}{u+n}.
\end{aligned}$$

Plug in the expression for $\Pr(X_1 = j_1, \dots, X_n = j_n)$ and note that

$$\begin{aligned}
c_i(j_1, \dots, j_{n+1}) &= c_i(j_1, \dots, j_n), \quad i \neq j_{n+1}, \\
c_{j_{n+1}}(j_1, \dots, j_n) &= c_{j_{n+1}}(j_1, \dots, j_{n+1}) - 1.
\end{aligned}$$

The claim follows.

6.2. **Throwing the thumbtack.** So far our basic setup is that X represents the data and Θ the parameter. We also assume that the distribution of X given $\Theta = \theta$ belongs to $\mathcal{P}_0 = \{P_\theta : \theta \in \Omega\}$.

In many cases one has a potentially infinite collection of random variables $\{X_n\}$ and stop the experiment after some finite number of observations so that $X = (X_1, \dots, X_n)$. It is common to assume "there exists a parameter Θ such that $\{X_i\}$ are conditionally IID given $\Theta = \theta$". We shall see how the existence of such a parameter can be deduced from exchangeability of the sequence.

We start with an example as the setting for our discussion. Suppose we throw a thumbtack repeatedly and we set $X_i = 1$ if it lands "point up" and $X_i = 0$ if it lands "point down". We could assume that there is a parameter Θ such that conditionally on $\Theta = \theta$ $\{X_n\}$ are IID $\mathrm{Ber}(\theta)$. The parameter space is $\Omega = [0, 1]$. In the classical setting the outcome of Θ can be considered as the "true value" of the parameter. It is fixed but unknown. We expect that the "true value" can be interpreted as the long-term frequency of 1's. A way to formulate this is, $\Theta = \lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} X_i$ a.s. Let us see how this works in detail.

Proposition 1 (c.f. Lemma 1.61 p. 35)**.** *If $\{X_n\}$ is a sequence of bounded random variables and there exists a random quantity Θ such that conditional on $\Theta = \theta$ $\{X_n\}$ are IID with mean $c(\theta)$, then*

$$n^{-1} S_n = \frac{1}{n} \sum_{i=1}^{n} X_i \to \Theta, \quad \Pr - a.s.$$

Proof. Without loss of generality we can assume that $E[X_i \mid \Theta] = 0$. Indeed, $n^{-1}S_n \to c(\Theta)$ a.s. is equivalent to $n^{-1}\sum_{i=1}^n (X_i - c(\Theta)) \to 0$ so putting $Y_i = X_i - c(\Theta)$ we see that it is sufficient to consider the case $E[X_i \mid \Theta] = 0$.

The idea is to use the first Borel-Cantelli lemma. That is, if we show

$$\sum_{n=1}^{\infty} \Pr(|n^{-1}S_n| > \varepsilon) < \infty \tag{6.1}$$

for each $\varepsilon > 0$, then by the first Borel-Cantelli lemma $\Pr(|n^{-1}S_n| > \varepsilon \quad i.o) = 0$ $\forall \varepsilon > 0$. In particular the event $\cap_{k=1}^{\infty}\{|n^{-1}S_n| > k^{-1} \quad i.o.\}^c$ has probability 1 and $n^{-1}S_n \to 0$ on this event. Let us show (6.1). Using Chebyshev's inequality we can write

$$\begin{aligned}
\Pr(|n^{-1}S_n| > \varepsilon) &\le \Pr(n^{-4}(\sum_{i=1}^n X_i)^4 > \varepsilon^4) \\
&\le \varepsilon^{-4} n^{-4} E[(\sum_{i=1}^n X_i)^4] \\
&= \varepsilon^{-4} n^{-4} E[(\sum_{i_1=1}^n \sum_{i_2=1}^n \sum_{i_3=1}^n \sum_{i_4=1}^n E[X_{i_1} X_{i_2} X_{i_3} X_{i_4} \mid \Theta] \\
&= \varepsilon^{-4} n^{-4} E[(\sum_{i=1}^n \mathrm{E}[X_i^4 \mid \Theta] + 3\sum_{i=1}^n \sum_{j=1}^n E[X_i^2 X_j^2 \mid \Theta]] \\
&\le \varepsilon^{-4} n^{-4} [nM^4 + 3n(n-1)M^4] \\
&\le \frac{4M^4}{n^2}.
\end{aligned}$$

where we use M as the upper bound for X_i. We see that

$$\sum_{n=1}^{\infty} \Pr(|n^{-1}S_n| > \varepsilon) \le \sum_{n=1}^{\infty} \frac{4M^4}{n^2} < \infty$$

which completes the proof. □

So far everything seems fine, but how do we motivate the assumption that "there exists Θ such that given $\Theta = \theta$ the $\{X_n\}$ are conditionally IID Ber(θ)"? One could argue that this assumption is strong and hard to verify. The point of replacing this "(heavy) probabilistic assumption" with exchangeability is that exchangeability reflects only the symmetry of the problem (it does not matter in which order we throw the thumbtack). The beauty of it all is that then it will follow from DeFinetti's theorem that there exists a parameter Θ such that conditional on $\Theta = \theta$ the sequence $\{X_n\}$ is IID Ber(θ) and the limit $n^{-1}\sum_{i=1}^n X_i \to \Theta$ holds Pr-a.s. Thus, instead of assuming a lot of structure from the beginning, i.e. "assume there exists Θ such that variables are conditionally IID given Θ", it is sufficient to assume exchangeability and the existence of a parameter Θ with the right properties will follow from DeFinetti's representation theorem.

6.3. **DeFinetti's representation theorem for Bernoulli random variables.** In the Bernoulli case DeFinetti's representation theorem is as follows.

Theorem 6 (c.f. Theorem 1.47 p. 26)**.** *Let $\{X_n\}$ be an infinite sequence of Bernoulli random variables. $\{X_n\}$ is exchangeable if and only if there exists a random variable Θ with vales in $[0,1]$ such that, conditional on $\Theta=\theta$, $\{X_n\}$ are IID Ber(θ). Furthermore, if $\{X_n\}$ is exchangeable, then the distribution of Θ is unique and $n^{-1}\sum_{i=1}^n X_i \to \Theta$ Pr-a.s.*

Consider Example 1.52, p. 29.

The difficult part of the proof is the "only if" direction. We have to show that the desired Θ exists. The idea is to show that $\{X_n\}$ exchangeable implies that the limit $\lim_n n^{-1}\sum_{i=1}^n X_i$ exists, and use this limit as our Θ. Thus, we need something like a strong law of large numbers for exchangeable random variables. The general way to find such a SLLN is to use martingale theory (se Theorem 1.62, p. 36). We will use the more elementary version

Theorem 7 (c.f. Theorem 1.59, p. 34)**.** *Let $\{X_n\}$ be exchangeable and real-valued. Suppose $E[X_i^2]=\mu_2<\infty$ and $E[X_iX_j]=m_2<\infty$ for each $i\neq j$. Put $Y_n=n^{-1}\sum_{i=1}^n X_i$. Then the subsequence $\{Y_{8^n}\}$ converges* Pr-*a.s.*

This more elementary version only gives us convergence along a subsequence, but it will be sufficient for our purpose. For those not afraid of martingale theory the general SLLN (Theorem 1.62 in the book) is very nice.

Proof. The way to prove that the subsequence converges is to show that it is almost surely a Cauchy sequence. By Chebychev's inequality we have for $m>n$

$$\begin{aligned}\Pr(|Y_m-Y_n|\geq c) &\leq c^{-2}E[(Y_m-Y_n)^2]\\ &= c^{-2}(E[Y_m^2]+E[Y_n^2]-2E[Y_mY_n])\\ &= c^{-2}(m^{-2}[m\mu_2+m(m-1)m_2]+n^{-2}[n\mu_2+n(n-1)m_2]\\ &\quad -2m^{-1}n^{-1}[n\mu_2+n(n-1)m_2+n(m-n)m_2])\\ &= c^{-2}(n^{-1}-m^{-1})(\mu_2-m_2)\\ &< n^{-1}c^{-2}(\mu_2-m_2).\end{aligned}$$

Consider now the subsequence $Z_k=Y_{8^k}$ and let $A_k=\{s:|Z_{k+1}-Z_k|\geq 2^{-k}\}$. If we take $c=2^{-k}$ we see that

$$\Pr(A_k)=\Pr(|Y_{8^{k+1}}-Y_{8^k}|\geq 2^{-k})<8^{-k}2^{2k}(\mu_2-m_2)=2^{-k}(\mu_2-m_2).$$

Put $A=\{A_k \quad i.o.\}=\cap_{n=1}^\infty\cup_{k=n}^\infty A_k$. Then, by the first Borel-Cantelli lemma, since $\Pr(A_k)$ is summable, $\Pr(A)=0$. We need only to show that for $s\in A^c$ $\{Z_k(s)\}$ is a Cuachy sequence. Take $\varepsilon>0$. We have to show that for each $s\in A^c$ there is an $N_{\varepsilon,s}$ such that $|Z_m(s)-Z_n(s)|<\varepsilon$ for $m>n\geq N_{\varepsilon,s}$. Write $A^c=\cup_{n=1}^\infty\cap_{k=n}^\infty A_k^c$. Then, for each $s\in A^c$ there is a c_s (depend on s) such that $s\in\cap_{k=c_s}^\infty A_k^c$. If $m>n\geq c_s$ if follows that

$$|Z_m(s)-Z_n(s)|\leq\sum_{i=n}^m|Z_{i+1}(s)-Z_i(s)|<2^{-n+1}\leq 2^{-c_s+1}.$$

Hence, we may take $N_{\varepsilon,s}>1+\max\{c_s,-\log_2\varepsilon\}$. This completes the proof. □

Proof of Theorem 6. First the "if" direction. If there is Θ such that $\{X_n\}$ are IID Ber(θ) given $\Theta=\theta$ then $\{X_n\}$ are exchangeable (as in Example 11). For the "only

if" direction we use Theorem 7 to conclude that $8^{-n}S_{8^n}$ converges Pr-a.s. Let $S_0 \subset S$ be the set where the limit exists, $\Pr(S_0) = 1$. For $s \in S_0$ we let $\Theta(s)$ be the limit, and for $s \notin S_0$ take $\Theta(s) = 1/2$. μ_Θ denotes the distribution of Θ. The important step is to prove

$$\Pr(X_1 = j_1, \dots, X_k = j_k, \Theta \in C) = \int_C \theta^y (1-\theta)^{k-y} \mu_\Theta(d\theta) \tag{6.2}$$

for each $k \geq 1$, $j_1, \dots, j_k \in \{0, 1\}$ and Borel set $C \subset [0, 1]$, where $y = j_1 + \cdots + j_k$. If this relation holds, then we see that $X_1, \dots, X_k$ are conditionally IID given $\Theta = \theta$. To prove (6.2) we introduce $Z_n = I_C(\Theta)(8^{-n}S_{8^n})^y(1 - 8^{-n}S_{8^n})^{k-y}$ and $Z = I_C(\Theta)\Theta^y(1-\Theta)^{k-y}$. Since $n^{-1}S_n \to \Theta$ Pr-a.s. it follows that $Z_n \to Z$ Pr-a.s. By the bounded convergence theorem $E[Z_n] \to \mathrm{E}[Z]$. Now we can identify $\mathrm{E}[Z]$ as the RHS of (6.2). It remains to show that $E[Z_n]$ converges to the LHS of (6.2). We write, with $m = 8^n$,

$$\begin{aligned}
Z_n &= I_C(\Theta)(m^{-1}S_m)^y(1 - m^{-1}S_m)^{k-y} \\
&= I_C(\Theta)m^{-k}(\sum_{i=1}^m X_i)^y(\sum_{i=1}^m [1 - X_i])^{k-y} \\
&= I_C(\Theta)m^{-k}\sum_{i_1=1}^m \cdots \sum_{i_y=1}^m X_{i_1} \dots X_{i_y} \sum_{i_{y+1}=1}^m \cdots \sum_{i_k=1}^m (1 - X_{i_{y+1}}) \dots (1 - X_{i_k}) \\
&= I_C(\Theta)m^{-k}\sum_{i_1=1}^m \cdots \sum_{i_k=1}^m X_{i_1} \dots X_{i_y}(1 - X_{i_{y+1}}) \dots (1 - X_k) \\
&= I_C(\Theta)m^{-k}\sum_{i_1=1}^m \cdots \sum_{i_k=1}^m I_{\{j_1\}}(X_{i_1}) \dots I_{\{j_k\}}(X_{i_k}) \\
&= I_C(\Theta)m^{-k}\sum_{i_1 \neq \cdots \neq i_k} I_{\{j_1\}}(X_{i_1}) \dots I_{\{j_k\}}(X_{i_k}) \\
&\quad + I_C(\Theta)m^{-k}\sum_{\text{at least two } i_t \text{ equal}} I_{\{j_1\}}(X_{i_1}) \dots I_{\{j_k\}}(X_{i_k}).
\end{aligned}$$

For the first sum we have

$$\begin{aligned}
&E[I_C(\Theta)m^{-k}\sum_{i_1 \neq \cdots \neq i_k} I_{\{j_1\}}(X_{i_1}) \dots I_{\{j_k\}}(X_{i_k})] \\
&\quad = m^{-k}\sum_{i_1 \neq \cdots \neq i_k} E[I_C(\Theta)I_{\{j_1\}}(X_{i_1}) \dots I_{\{j_k\}}(X_{i_k})] \\
&\quad = \frac{m!}{m^k(m-k)!}E[I_C(\Theta)I_{\{j_1\}}(X_{i_1}) \dots I_{\{j_k\}}(X_{i_k})] \\
&\quad = \frac{m!}{m^k(m-k)!}\Pr(\Theta \in C, X_1 = j_1, \dots, X_k = j_k) \\
&\quad \to \Pr(\Theta \in C, X_1 = j_1, \dots, X_k = j_k)
\end{aligned}$$

as $m \to \infty$. In the last equality we used exchangeability. The second sum has $m^k - m!/(m-k)!$ terms, each bounded between 0 and 1 and since $m^{-k}(m^k - m!/(m-k)! \to 0$ the second sum converges to 0. This proves (6.2). To see that the distribution of Θ is unique, suppose ν_Θ is a distribution that satisfies (6.2). Then,

summing up over all $j_1, \dots, j_k$ we see that

$$\begin{aligned}\nu_\Theta(C) &= \sum_{j_1=0}^{1} \cdots \sum_{j_k=0}^{1} \int_C \theta^y (1-\theta)^{k-y} \nu_\Theta(d\theta) \\ &= \sum_{j_1=0}^{1} \cdots \sum_{j_k=0}^{1} \int_C \theta^y (1-\theta)^{k-y} \mu_\Theta(d\theta) = \mu_\Theta(C),\end{aligned}$$

for all Borel sets $C \subset [0,1]$. Hence $\nu_\Theta = \mu_\Theta$. To see that $n^{-1}S_n \to \Theta$ Pr-a.s. we can apply Proposition 1. Note that Theorem 7 is not enough because it only gives convergence along a particular subsequence (alternatively we could apply Theorem 1.62 in the book). $\square$