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 $X_1, \dots, X_n$ be IID $\text{Ber}(\theta)$ given $\Theta = \theta$ and put $Y = \sum_{i=1}^n X_i$. Then $Y \sim \text{Bin}(n, \theta)$. Let the prior distribution be $\text{Beta}(\alpha, \beta)$. Then the posterior of Θ given $Y = y$ is $\text{Beta}(y + \alpha, n - y + \beta)$.

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 $\text{Beta}(y + \alpha, n - y + \beta)$ 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 $\mathrm{Exp}(\theta)$ 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*

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

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 $\theta_1 = (\theta_{11}, \dots, \theta_{1k})$ and $\theta_2 = (\theta_{21}, \dots, \theta_{2k})$ 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,

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