# Bayesian Statistics

EE209 - Eléments de Statistiques pour les Data Sciences

# An estimator that takes the form of a probability distribution...

In the context of estimation, we have seen so far:

Point Estimators: the MLE, the method of moments (MM) produce estimators that take the form of a single number (or vector).

Confidence Intervals: i.e., interval estimators that are specified by two numbers. (They can be generalized in higher dimension with confidence regions which are sets.)

A school of thought in statistics called Bayesian statistic considers that estimation being intrinsically uncertain, we should express our uncertainty about the estimated quantity by providing an estimator which is itself a probability distribution, but this time about the unknown parameter. Bayesians propose a simple methodology to compute these probabilities, which is based on treating unknown parameters as random variables + specifying an initial "a priori" distribution on these parameters and applying... the Bayes rule.

Frequentist statisticians, who represent the other main school of thought have traditionally had reservations about the Bayesian principles and have proposed another way producing an estimator taking the form of a probability distribution, the bootstrap, which we will not cover in this course.

Today, it is well accepted that both frequentist and Bayesian principle have merits.

# About notations

In this chapter we will simplify/losen some of the notations for legibility:

We will write

- $p(x)$ for $p_X(x)$ or $P_X(x)$: we will use the same notations for pmf and pdfs and drop the index indicating whose r.v. this is the pdf of. The choice of the letter in the argument will implicitly specify this.
- similarly, we will write $p(x|y)$ for $p_{X|Y}(x|y)$, $p(x,y)$ for $p_{(X,Y)}(x,y)$, and $p(y)$ for $p_Y(y)$. Note that this does not mean that we assume that $X$ and $Y$ have the same distribution... Indices, might reappear in ambiguous cases.
- for reason that will become clear, we will write $p(x|\theta)$ instead of the previous $p_\theta(x)$ or $p(x;\theta)$ to indicate the dependence on a parameter of the statistical model. This is the "Bayesian way" of writing this dependence.

# Bayesian estimation for a single observation

Bayesians treat the parameter $\theta$ as a **random variable**.

## A priori

The Bayesian has to specify an *a priori* distribution $p(\theta)$ for the model parameters $\theta$, which models her/his prior belief of the relative plausibility of different values of the parameter.

## A posteriori

The observation contributes through the likelihood: $p(x|\theta)$.

The *a posteriori* distribution on the parameters is then

$$p(\theta|x) = \frac{p(x|\theta)\, p(\theta)}{p(x)} \propto p(x|\theta)\, p(\theta) \qquad \text{with} \quad p(x) = \int p(x \mid \theta)\, p(\theta)\, d\theta$$

$\rightarrow$ The Bayesian estimator is therefore a probability distribution on the parameters.

This estimation procedure is called Bayesian inference.

# Bayesian inference for a sample of observations

Assume that we have a sample $\mathcal{D}_n = \{x_1, \ldots, x_n\}$ of observations that are i.i.d. from a distribution with pmf/pdf $p(x \mid \theta)$ in a statistical model $\mathcal{P}_\Theta$ for some value of $\theta$. The data $\mathcal{D}_n$ is often called the *evidence*.

As before, we use an a priori distribution (or prior distribution) $p(\theta)$ over $\Theta$.

Since for a given $\theta$ the data is i.i.d., the likelihood now takes the form:

$$p(\mathcal{D}_n \mid \theta) = p(x_1, \ldots, x_n \mid \theta) = p(x_1 \mid \theta) \ldots p(x_n \mid \theta)$$

The a posteriori distribution (or posterior distribution) is obtained again using Bayes' rule

$$p(\theta \mid \mathcal{D}_n) = \frac{p(\mathcal{D}_n \mid \theta)\, p(\theta)}{p(\mathcal{D}_n)} \qquad \text{with} \qquad p(\mathcal{D}_n) = \int p(\mathcal{D}_n \mid \theta)\, p(\theta)\, d\theta.$$

# The Beta distribution

A beta random variable $\theta \sim \text{Beta}(\alpha, \beta)$ is a random variable defined on the interval $[0, 1]$ and whose density takes form

$$p(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\, \theta^{\alpha-1}\, (1-\theta)^{\beta-1}\, 1_{\{0\leq\theta\leq 1\}}, \quad \text{for} \quad \alpha, \beta > 0.$$

$$\mathbb{E}[\theta] = \frac{\alpha}{\alpha+\beta} \qquad \text{and} \qquad \text{Var}(\theta) = \frac{\alpha}{\alpha+\beta}\frac{\beta}{\alpha+\beta}\frac{1}{1+\alpha+\beta}.$$

So if we define $\theta_0 := \mathbb{E}[\theta]$ and $n_0 := \alpha + \beta$, we have

$$p(\theta) = \frac{\Gamma(n_0)}{\Gamma\big(n_0\theta_0\big)\,\Gamma\big(n_0(1-\theta_0)\big)}\, \theta^{n_0\theta_0-1}\, (1-\theta)^{n_0(1-\theta_0)-1}\, 1_{\{0\leq\theta\leq 1\}}.$$

$$\mathbb{E}[\theta] = \theta_0 \qquad \text{and} \qquad \text{Var}(\theta) = \frac{\theta_0(1-\theta_0)}{1+n_0}$$

# Bayesian inference for the Beta-Bernoulli model

Let $\mathcal{D}_n = \{x_1, \dots, x_n\}$, with $X_i \overset{iid}{\sim} \text{Ber}(\theta)$, so that we have $p(x_i \mid \theta) = \theta^{x_i}(1-\theta)^{1-x_i}$.

For the whole sample, the likelihood is $p(\mathcal{D}_n \mid \theta) = p(x_1 \mid \theta) \dots p(x_n \mid \theta) = \theta^N (1-\theta)^{n-N}$.

And we use the prior distribution $p(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}\, 1_{\{0\leq\theta\leq 1\}}$.

We can calculate the posterior distribution:

$$\begin{aligned} p(\theta \mid \mathcal{D}_n) &\propto p(\mathcal{D}_n \mid \theta)\, p(\theta) \\ &\propto \theta^N (n-\theta)^{n-N}\, \theta^{\alpha-1}(1-\theta)^{\beta-1}\, 1_{\{0\leq\theta\leq 1\}} \\ &\propto \theta^{N+\alpha-1}\, (1-\theta)^{n-N+\beta-1}\, 1_{\{0\leq\theta\leq 1\}} \\ &= \frac{\Gamma(n+\alpha+\beta)}{\Gamma(N+\alpha)\Gamma(n-N+\beta)}\, \theta^{N+\alpha-1}\, (1-\theta)^{n-N+\beta-1}\, 1_{\{0\leq\theta\leq 1\}} \end{aligned}$$

So $\theta \mid \mathcal{D}_n \sim \text{Beta}(N+\alpha, n-N+\beta)$.

# Posterior mean and posterior variance (in the Beta-Bernoulli model)

Based on the posterior distribution $\theta \mid \mathcal{D}_n \sim \text{Beta}(N + \alpha, n - N + \beta)$,

We can compute:

- the **posterior mean**

$$\theta_{\text{PM}} = \mathbb{E}[\,\theta \mid \mathcal{D}_n\,] = \frac{N + \alpha}{n + \alpha + \beta} = \frac{n}{n + \alpha + \beta}\,\frac{N}{n} + \frac{\alpha + \beta}{n + \alpha + \beta}\,\frac{\alpha}{\alpha + \beta}$$
$$= \frac{n}{n + n_0}\,\hat{\theta}_{\text{MLE}} + \frac{n_0}{n + n_0}\,\theta_0.$$

  $\theta_{\text{PM}}$ can be seen as a *point estimator* which, in this case (and others as we will see), is a convex combination of the MLE and the prior mean.

- the **posterior variance**

$$\text{Var}(\,\theta \mid \mathcal{D}_n\,) = \frac{\theta_{\text{PM}}\,(1 - \theta_{\text{PM}})}{1 + n + n_0},$$

  which is way to quantify how the posterior distribution is concentrated.

# The posterior mode aka the *maximum a posteriori* (MAP)

Another *point estimator* that can be derived from the whole posterior distribution, is the posterior mode which is usually called the MAP or *maximum a posteriori*.

$$\begin{aligned}\theta_{\mathsf{MAP}} &= \arg\max_{\theta} \log p(\theta \mid \mathcal{D}_n) = \arg\max_{\theta} \log\left(\frac{p(\mathcal{D}_n \mid \theta)}{}p(\theta)p(\mathcal{D}_n)\right)\\ &= \arg\max_{\theta} \log p(\mathcal{D}_n \mid \theta) + \log p(\theta).\end{aligned}$$

Note that $\theta_{\mathsf{MAP}}$ resembles the MLE because $\log p(\mathcal{D}_n \mid \theta) = \ell(\theta)$ is the log-likelihood.

For Bayesian estimation in the Bernoulli model with a Beta prior we get

$$\log p(\theta \mid \mathcal{D}_n) = (N+\alpha-1)\log\theta + (n-N+\beta-1)\log(1-\theta) + \mathsf{cst}$$

if $\theta \in [0,1]$ and $-\infty$ if $\theta \notin [0,1]$.
We thus find

$$\theta_{\mathsf{MAP}} = \frac{N+\alpha-1}{n+\alpha+\beta-2} \qquad \text{so that for } \alpha=\beta=1 \quad \theta_{\mathsf{MAP}} = \hat{\theta}_{\mathsf{MLE}}.$$

# Bayesian inference for the mean of a Gaussian, assuming that $\sigma^2$ is known

Let $\mathcal{D}_n = \{x_1, \ldots, x_n\}$, with $x_i \mid \mu \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2), \quad \mu \sim \mathcal{N}(\mu_0, \tau^2)$.

$$\begin{aligned}
p(\mu \mid \mathcal{D}_n) &\propto p(\mathcal{D}_n \mid \mu)\, p(\mu) \\
&\propto \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2} \frac{1}{(2\pi\tau^2)^{1/2}} e^{-\frac{1}{2\tau^2}(\mu-\mu_0)^2} \\
&\propto \exp\left(-\tfrac{1}{2\sigma^2}\sum_{i=1}^n (\mu - x_i)^2 - \tfrac{1}{2\tau^2}(\mu-\mu_0)^2\right) \\
&\propto \exp\left\{-\tfrac{1}{2}\left(\left[\tfrac{n}{\sigma^2} + \tfrac{1}{\tau^2}\right]\mu^2 - 2\left[\tfrac{n\bar{x}}{\sigma^2} + \tfrac{\mu_0}{\tau^2}\right]\mu + \ldots\right)\right\} \\
&\propto \exp\left\{-\tfrac{1}{2}\left[\tfrac{n\tau^2+\sigma^2}{\tau^2\sigma^2}\right]\left(\mu^2 - 2\left[\tfrac{n\tau^2\,\bar{x}+\sigma^2\,\mu_0}{n\tau^2+\sigma^2}\right]\mu + \ldots\right)\right\}
\end{aligned}$$

$\mu \mid \mathcal{D}_n \sim \mathcal{N}(\mu_{\text{post}}, \tau^2_{\text{post}})$ with

$$\mu_{\text{post}} = \frac{n\tau^2}{n\tau^2+\sigma^2}\,\bar{x} + \frac{\sigma^2}{n\tau^2+\sigma^2}\mu_0, \qquad \tau^2_{\text{post}} = \frac{\sigma^2}{n + \frac{\sigma^2}{\tau^2}}.$$

# Posterior distribution for $\mu$, assuming that $\sigma^2$ is known

$$\mu \mid \mathcal{D}_n \sim \mathcal{N}(\mu_{\text{post}}, \tau^2_{\text{post}}) \quad \text{with} \quad \mu_{\text{post}} = \frac{n\tau^2}{n\tau^2 + \sigma^2}\,\bar{x} + \frac{\sigma^2}{n\tau^2 + \sigma^2}\mu_0, \qquad \tau^2_{\text{post}} = \frac{\sigma^2}{n + \frac{\sigma^2}{\tau^2}}.$$

Of course, the parameters of the posterior are the **posterior mean** and **posterior variance**:

$$\mu_{\text{PM}} = \mathbb{E}[\mu \mid \mathcal{D}_n] = \mu_{\text{post}} \qquad \text{and} \qquad \text{Var}(\mu \mid \mathcal{D}_n) = \tau^2_{\text{post}}.$$

For the MAP, since the mode of a Gaussian distribution is also its mean we have

$$\theta_{\text{MAP}} = \theta_{\text{PM}}.$$

# Bayesian inference for a sample of observations

Given an i.i.d. sample $x_1, \ldots, x_n$, in frequentist statistics, we have

$$p_\theta(x_1, \ldots, x_n) = p_\theta(x_1)\, p_\theta(x_2)\, \ldots p_\theta(x_n)$$

Similarly, in Bayesian terms, we have

$$p(x_1, \ldots, x_n \mid \theta) = p(x_1|\theta)\, p(x_2|\theta)\, \ldots\, p(x_n|\theta). \qquad (*)$$

But since $\theta$ is a random variable, the above equation is **not equivalent to**

$$p(x_1, \ldots, x_n) = p(x_1)\, p(x_2)\, \ldots\, p(x_n).$$

Equation $(*)$ is true if

$X_1, \ldots, X_n$ are "independent **given** $\theta$ " which is different than $X_1, \ldots, X_n$ are independent.

This makes sense if you think of observations from a biased coin with unknown bias. The concept of **conditional independence** is a complicated concept and we will not explore it further, but it important to remember that the observations are independent only "given $\theta$."

# Consecutive updates of the posterior

$$\begin{aligned} p(\theta \mid \mathcal{D}_n)\, p(\mathcal{D}_n) = p(\mathcal{D}_n \mid \theta)\, p(\theta) &= p(x_1 \mid \theta) \ldots p(x_n \mid \theta)\, p(\theta) \\ &= p(x_n \mid \theta)\, p(\mathcal{D}_{n-1} \mid \theta)\, p(\theta) \\ &= p(x_n \mid \theta)\, p(\theta \mid \mathcal{D}_{n-1})\, p(\mathcal{D}_{n-1}). \end{aligned}$$

But $p(\mathcal{D}_n) = p(x_1, \ldots, x_n) = p(x_n \mid \mathcal{D}_{n-1})\, p(\mathcal{D}_{n-1})$. So

$$p(\theta \mid \mathcal{D}_n) = p(x_n \mid \theta)\, p(\theta \mid \mathcal{D}_{n-1}) \frac{p(\mathcal{D}_{n-1})}{p(\mathcal{D}_n)} = \frac{p(x_n \mid \theta)\, p(\theta \mid \mathcal{D}_{n-1})}{p(x_n \mid \mathcal{D}_{n-1})}.$$

And

$$\int p(x_n \mid \theta)\, p(\theta \mid \mathcal{D}_{n-1})\, d\theta = \int p(x_n \mid \theta) \frac{p(\mathcal{D}_{n-1} \mid \theta)\, p(\theta)}{p(\mathcal{D}_{n-1})} d\theta = \frac{p(\mathcal{D}_n)}{p(\mathcal{D}_{n-1})} = p(x_n \mid \mathcal{D}_{n-1}).$$

Finally $$\boxed{p(\theta \mid \mathcal{D}_n) = \frac{p(x_n \mid \theta)\, p(\theta \mid \mathcal{D}_{n-1})}{p(x_n \mid \mathcal{D}_{n-1})} = \frac{p(x_n \mid \theta)\, p(\theta \mid \mathcal{D}_{n-1})}{\int p(x_n \mid \theta)\, p(\theta \mid \mathcal{D}_{n-1})\, d\theta}}.$$

# The posterior is the new prior after having seen data...

If we apply the formula we established

$$p(\theta \mid \mathcal{D}_n) = \frac{p(x_n \mid \theta)\, p(\theta \mid \mathcal{D}_{n-1})}{p(x_n \mid \mathcal{D}_{n-1})}.$$

to data arriving one by one, we get

$$p(\theta \mid x_1) = \frac{p(x_1 \mid \theta)\, p(\theta)}{p(x_1)} \qquad \text{and} \qquad p(\theta \mid x_1, x_2) = \frac{p(x_2 \mid \theta)\, p(\theta \mid x_1)}{p(x_2 \mid x_1)} \qquad \text{etc}$$

So the posterior after seeing $x_1$ becomes the new prior before seeing $x_2$ and if we apply Bayes rule, we obtain a new posterior which is the same as if we had observed $x_1$ and $x_2$ at the same time !

The posterior acts as a memory of what we have learned, which can be updated when new information arrives.

# The Dirichlet distribution

We say that $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$ follows the Dirichlet distribution and note

$$\boldsymbol{\theta} \sim \mathrm{Dir}(\boldsymbol{\alpha})$$

for $\boldsymbol{\theta}$ in the simplex $\triangle_K = \{\boldsymbol{u} \in \mathbb{R}_+^K \mid \sum_{k=1}^K u_k = 1\}$ and if it has the density

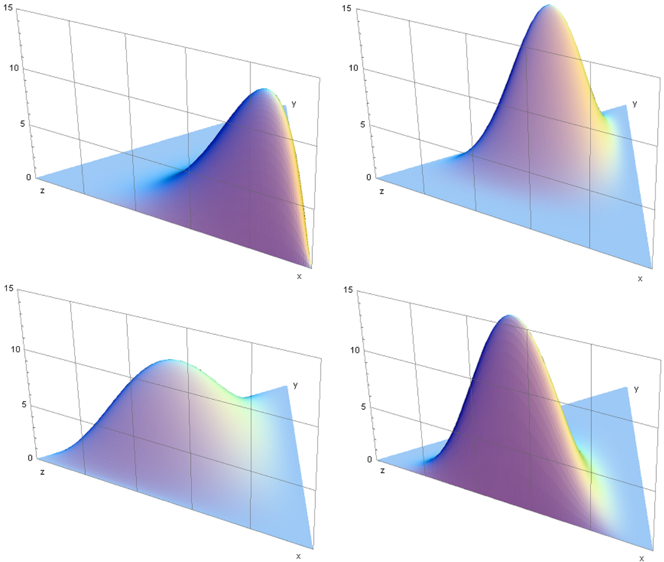
$$p(\boldsymbol{\theta}; \boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\prod_{k=1}^K \Gamma(\alpha_k)}\, \theta_1^{\alpha_1 - 1} \ldots \theta_{K-1}^{\alpha_{K-1}-1} \big(1 - \textstyle\sum_{j=1}^{K-1} \theta_j\big)^{\alpha_K - 1}\, 1_{\{\forall 1 \le j \le K-1,\ 0 \le \theta_j \le 1\}}.$$

where

$$\alpha_0 = \sum_k \alpha_k \quad \text{and} \quad \Gamma(x) := \int_0^\infty t^{x-1} e^{-t} dt.$$

In fact to be rigorous, we should write $p(\theta_1, \ldots, \theta_{K-1}; \boldsymbol{\alpha})$ but it is more convenient to write $p(\boldsymbol{\theta}; \boldsymbol{\alpha})$ and keep in mind that $\theta_K = 1 - \sum_{j=1}^{K-1} \theta_j$.

# Dirichlet distribution II



# Bayesian estimation of a multinomial random variable

Let $\mathcal{D}_n$ be an i.i.d. sample $\boldsymbol{z}_1, \dots, \boldsymbol{z}_n$ with

- $\boldsymbol{z}_i = (z_{i1}, \dots, z_{iK}) \sim \mathcal{M}(1, \boldsymbol{\theta})$ following a "multinoulli" r.v. variable
- $\boldsymbol{\theta} \sim \mathrm{Dir}(\boldsymbol{\alpha})$, i.e. $\boldsymbol{\theta}$ follows a Dirichlet prior.

$$p(\boldsymbol{\theta}) \propto \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \qquad \text{and} \qquad p(\boldsymbol{z}_i | \boldsymbol{\theta}) = \prod_{k=1}^{K} \theta_k^{z_{ik}}$$

We have

$$p(\boldsymbol{\theta} | \boldsymbol{z}_1, \dots, \boldsymbol{z}_n) = \frac{p(\boldsymbol{\theta}) \prod_i p(\boldsymbol{z}_i | \boldsymbol{\theta})}{p(\boldsymbol{z}_1, \dots, \boldsymbol{z}_n)} \propto \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \prod_{i=1}^{n} \prod_{k=1}^{K} \theta_k^{z_{ik}} \propto \prod_k \theta_k^{\sum_{i=1}^n z_{ik} + \alpha_k - 1}$$

So that $(\boldsymbol{\theta} \mid \mathcal{D}_n) \sim \mathrm{Dir}(\alpha_1 + N_1, \dots, \alpha_K + N_K)$ with $N_k = \sum_{i=1}^n z_{ik}$.

# Posterior Mean and Posterior Variance in the Dirichlet-Multinomial model

If $\boldsymbol{\theta} \sim \mathrm{Dir}(\alpha_1, \ldots, \alpha_K)$ then

$$\theta_k^{\text{prior}} = \mathbb{E}[\theta_k] = \frac{\alpha_k}{\alpha_0} \quad , \quad \mathrm{Var}(\theta_k) = \frac{\alpha_k(\alpha_0 - \alpha_k)}{\alpha_0^2(\alpha_0 + 1)} \quad \text{with} \quad \alpha_0 = \sum_k \alpha_k .$$

We have shown that $(\boldsymbol{\theta} \mid \mathcal{D}_n) \sim \mathrm{Dir}(\alpha_1 + N_1, \ldots, \alpha_K + N_K)$

## Posterior mean

$$\theta_{\mathrm{PM},k} := \mathbb{E}[\theta_k | \mathcal{D}_n] = \frac{\alpha_k + N_k}{\alpha_0 + n} = \frac{\alpha_0}{\alpha_0 + n} \frac{\alpha_k}{\alpha_0} + \frac{n}{\alpha_0 + n} \frac{N_k}{n} .$$

## Posterior variance

$$\mathrm{Var}(\, \theta_k \mid \mathcal{D}_n \,) = \frac{\theta_{\mathrm{PM},k} \, (1 - \theta_{\mathrm{PM},k})}{\alpha_0 + n + 1}$$

# Conjugate priors

A family of prior distribution $\mathcal{P}_A = \{p_\alpha(\theta) \mid \alpha \in A\}$

is said to be **conjugate** to a model $\mathcal{P}_\Theta$, if, for a sample

$$X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} p(\cdot \mid \theta) \qquad \text{with} \qquad p(\cdot \mid \theta) \in \mathcal{P}_\Theta,$$

the distribution $q$ defined by $q(\theta) := p(\theta|x_1, \dots, x_n) = \dfrac{p_\alpha(\theta) \prod_i p(x_i \mid \theta)}{\int p_\alpha(\theta) \prod_i p_\theta(x_i) d\theta}$

is such that $q \in \mathcal{P}_A$.

Note that if we could also have used Bayesian notations $p(\theta \mid \alpha)$ instead of $p_\alpha(\theta)$.

# Conjugate families

| Likelihood | Conjugate prior |
|---|---|
| Bernoulli/Binomial | Beta |
| Multinomial | Dirichlet |
| Poisson | Gamma |
| Normal with fixed $\sigma^2$ | Normal |
| Normal with fixed $\mu$ | Inverse gamma |
| Normal multivar. Normal with fixed $\Sigma$ | Normal |
| multivar. Normal with fixed $\Sigma$ with fixed $\boldsymbol{\mu}$ | Inverse Wishart |
| Exponential | Gamma |

In all the examples that we have considered (Bernoulli, Gaussian and Multinomial model) we have used the conjugate prior for convenience each time. This is not necessary but then the integrals cannot be computed analytically in general...

# Posterior expectations and the predictive distribution

The principle of Bayesian estimation is that the prior and posterior distribution model the *uncertainty* that we have in the estimation process. As a consequence, one should always integrate over the uncertainty. So the final estimate for a function $f(\theta)$ is

$$\mathbb{E}[f(\theta) \mid \mathcal{D}_n] = \int f(\theta)\, p(\theta|\mathcal{D}_n)\, d\theta.$$

Of course the posterior mean is a particular example.
In particular the predictive distribution is the pmf/pdf of a new observation $x'$ from the *model* given the *evidence* provided by the data $\mathcal{D}_n = \{\mathbf{x}_1, \dots, x_n\}$

$$\mathbb{E}\big[p(x'|\theta) \mid \mathcal{D}_n\big] = p(x'|x_1, \dots, x_n) = \int p(x'|\theta)\, p(\theta|x_1, \dots, x_n)\, d\theta.$$

We will not discuss the calculations of the predictive distribution further in this course.

# Bayesian inference for the precision of a Gaussian

Reminder: $X \sim \Gamma(k, \beta)$ then $p(x) = \frac{\beta^k}{\Gamma(k)} x^{k-1} \exp\big(-\beta x\big)$

By definition the *precision* is $\lambda = \sigma^{-2}$

Let $\mathcal{D}_n = \{x_1, \ldots, x_n\}$, with $x_i \,|\, \lambda \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \lambda^{-1})$, $\quad \lambda \sim \Gamma(\frac{n_0}{2}, \frac{n_0 \sigma_0^2}{2})$.

If $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2$, then the Gaussian likelihood takes the form

$$p(\mathcal{D}_n|\lambda) = \big(\tfrac{\lambda}{2\pi}\big)^{\frac{n}{2}} \exp\big(-\lambda \tfrac{n\hat{\sigma}^2}{2}\big) \quad \text{and} \quad p(\lambda) = \frac{\big(\frac{n_0 \sigma_0^2}{2}\big)^{\frac{n_0}{2}}}{\Gamma\big(\frac{n_0}{2}\big)} \lambda^{\frac{n_0}{2}-1} \exp\big(-\lambda \tfrac{n_0 \sigma_0^2}{2}\big)$$

$$p(\lambda|\mathcal{D}_n) \propto \lambda^{\frac{n_0+n}{2}-1} \exp\big(-\lambda\big[\tfrac{n_0\sigma_0^2}{2} + \tfrac{n\hat{\sigma}^2}{2}\big]\big)$$

So that $\lambda|\mathcal{D}_n \sim \Gamma\big(\frac{n_0+n}{2}, \frac{n_0\sigma_0^2 + n\hat{\sigma}^2}{2}\big)$

$$\mathbb{E}[\lambda|\mathcal{D}_n] = \frac{n_0 + n}{n_0 \sigma_0^2 + n\hat{\sigma}^2} = \left(\frac{n_0}{n_0+n}\sigma_0^2 + \frac{n}{n_0+n}\hat{\sigma}^2\right)^{-1}, \qquad \text{Var}(\lambda|\mathcal{D}_n) = \frac{n_0+n}{(n_0\sigma_0^2 + n\hat{\sigma}^2)^2}.$$

# Improper priors

For the Gaussian model with prior $\mu \sim \mathcal{N}(\mu_0, \tau^2)$ we found $\mu \mid \mathcal{D}_n \sim \mathcal{N}(\mu_{\text{post}}, \tau^2_{\text{post}})$ with

$$\mu_{\text{post}} = \frac{n\tau^2}{n\tau^2 + \sigma^2}\,\bar{x} + \frac{\sigma^2}{n\tau^2 + \sigma^2}\mu_0, \qquad \tau^2_{\text{post}} = \frac{\sigma^2}{n + \frac{\sigma^2}{\tau^2}}.$$

If we don't have a strong prior on $\mu$ it is natural to let $\tau^2 \to +\infty$. In that case the prior is not a probability distribution anymore, but the prior on $\mu$ is uniform over all real numbers! This is an example of an improper prior. The posterior distribution is still well defined with

$$\mu_{\text{post}} = \bar{x}, \qquad \tau^2_{\text{post}} = \frac{\sigma^2}{n}.$$

In that case $\mu_{\text{post}}$ is unbiased (and coincidentally $\text{Var}(\theta \mid \mathcal{D}_n) = \text{Var}(\bar{X})$).

For the Beta-Bernoulli model we found $\theta \mid \mathcal{D}_n \sim \text{Beta}(N + \alpha, n - N + \beta)$, and the posterior mean was $\theta_{\text{PM}} = \frac{N+\alpha}{n+\alpha+\beta}$. Letting $\alpha \to 0$ and $\beta \to 0$ corresponds to using an improper prior $\text{Beta}(0, 0)$ and we again obtain an unbiased posterior mean $\theta_{\text{PM}} = \frac{N}{n}$.

# Frequentist Probability *vs* Bayesian probability

For *frequentists*, the concept of probability is grounded in the law of large numbers. For them, it only makes sense to talk about the probability of some event, if this event occurs in a random experiment that can (a least theoretically) be repeated indefinitely, so that the probability can be defined as the limiting frequency of occurence of the event.

*Examples:* Probability that a coin falls on heads, probability of winning at the lottery, probability that an isotope disintegrates in less than 1 sec, probability that my friend is upset on Mondays.

For *Bayesians*, probability distributions can be used as well to express a *belief* about possible outcomes or truth. And this belief can be updated based on a likelihood (consisting of frequentist or Bayesian probabilities) based on Bayes rules.

*Examples:* Probability that the universe is finite, probability that my friend is upset today, probability that the missing mass in the universe is larger than $x$.

# Some remarks on Bayesian methods

Subjective vs objective priors. A difficultly encountered in practice is how to choose the prior. *Subjective Bayesians* argue that the prior should reflect their prior knowlegde and beliefs. *Objective Bayesians* seeks ways to choose priors that are neutral or optimal in some sense (improper priors, Jeffreys' priors, etc).

Bayesian uncertainty vs frequentist uncertainty. Bayesians consider data as fixed/given and the uncertainty on $\theta$ is obtained by combining "prior belief" encoded in the a priori distribution with the likelihood of the different values of the parameters, whereas for frequentist the uncertainty on $\theta$ is the uncertainty of $\hat{\theta}$ which depends on the distribution of the data.

Assuming that the model is correct. The logic of Bayesian inference requires that the data really comes from the likelihood, otherwise we cannot really apply Bayes' rule.