Bayesian Statistics

Mathieu Ribatet—Full Professor of Statistics

1. Introduction

Some references

[1] M. K. Cowles. *Applied Bayesian Statistics with R and OpenBugs Examples*. Springer Texts in Statistics. Springer-Verlag, 2013.

[2] J. A. Hartigan. *Bayes Theory*. Springer Series in Statistics. Springer-Verlag, 1983.

[3] C.P. Robert. *The Bayesian Choice: A Decision-theoretic Motivation*. Springer Texts in Statistics. Springer-Verlag, 2007.

Statistical model (parametric)

Definition 1. A parametric family of functions $\{f(x;\theta) : x \in E, \theta \in \Theta\}$ is a statistical model if, for all $\theta \in \Theta$, $x \mapsto f(x;\theta)$ is probabilty density function on E. The set Θ is the parameter space and E observational space.

Remark. A statistical model is parametric whenever $\dim(\Theta) < \infty$.

Example 1. The family

$$\left\{ f(x;\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} \right) : x \in \mathbb{R}, \theta = (\mu, \sigma^2) \in \mathbb{R} \times (0, \infty) \right\}$$

is a statistical model, that of Gaussian distributions.

Frequentist statistics

- Let $\{g(x;\theta)\colon x \in \mathbf{X}, \psi \in \Psi\}$ be a statistical model.
- Frequentist statistics suppose that there exists a true parameter ψ_*, from which data are drawn, i.e.,

$$(X_1, \ldots, X_n) \sim g(\cdot; \psi_*).$$

- Ignoring the true model we consider a proposal model $\{f(x;\theta)\colon x \in E, \theta \in \Theta\}$ and fit this model using an estimator for θ_0, often denoted $\hat{\theta}$, with some desirable properties, e.g.,

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{\mathscr{L}} N(0, \Sigma), \qquad n \to \infty.$$

Frequentist statistics

- Let $\{g(x;\theta)\colon x \in \mathbf{X}, \psi \in \Psi\}$ be a statistical model.
- Frequentist statistics suppose that there exists a true parameter ψ_*, from which data are drawn, i.e.,

$$(X_1, \ldots, X_n) \sim g(\cdot; \psi_*).$$

- Ignoring the true model we consider a proposal model $\{f(x;\theta)\colon x \in E, \theta \in \Theta\}$ and fit this model using an estimator for θ_0, often denoted $\hat{\theta}$, with some desirable properties, e.g.,

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{\mathscr{L}} N(0, \Sigma), \qquad n \to \infty.$$

Remark. Often we suppose that the proposal statistical model contains the true one. However we can generate this setting and this is known as misspecified models.

Point estimate and precision

- The estimator $\hat{\theta}$ yields a point estimation for θ_0
- Usualy we rely on asymptotic behavior as the one above to get confidence intervals, e.g.,

$$\left[\hat{\theta} - z_{0.975}\text{std.err}(\hat{\theta}), \hat{\theta} + z_{0.975}\text{std.err}(\hat{\theta})\right],$$

where

$$z_{0.975} = \Phi^{-1}(0.975), \qquad \text{std.err}(\hat{\theta}) = \sqrt{\text{Var}(\hat{\theta})}.$$

Point estimate and precision

- The estimator $\hat{\theta}$ yields a point estimation for θ_0
- Usualy we rely on asymptotic behavior as the one above to get confidence intervals, e.g.,

$$\left[\hat{\theta} - z_{0.975}\text{std.err}(\hat{\theta}), \hat{\theta} + z_{0.975}\text{std.err}(\hat{\theta})\right],$$

where

$$z_{0.975} = \Phi^{-1}(0.975), \qquad \text{std.err}(\hat{\theta}) = \sqrt{\text{Var}(\hat{\theta})}.$$

☞ To sum up, $\hat{\theta}$ is a random variable for which, most often, the asymptotic distribution is fully characterized.

2. Basic notions

Foundation of Bayesian statistics

Idea: Treat the parameter θ as a random variable!

Foundation of Bayesian statistics

Idea: Treat the parameter θ as a random variable!

- We thus need to assume is generated from a distribution, i.e., $\theta \sim \pi$. The latter distribution is called prior distribution.
- The prior distribution encodes our knowledage / ignorance about θ_* prior to have a look at the data.

Idea: Treat the parameter θ as a random variable!

- We thus need to assume is generated from a distribution, i.e., $\theta \sim \pi$. The latter distribution is called prior distribution.
- The prior distribution encodes our knowledage / ignorance about θ_* prior to have a look at the data.

Example 2. For our Gaussian model we may assume

$$\pi(\theta) = \pi(\mu) \times \pi(\sigma^2) = N(\mu_0, \tau) \times \mathsf{InvGamma}(\alpha, \beta).$$

Idea: Treat the parameter θ as a random variable!

- We thus need to assume is generated from a distribution, i.e., $\theta \sim \pi$. The latter distribution is called prior distribution.
- The prior distribution encodes our knowledage / ignorance about θ_* prior to have a look at the data.

Example 2. For our Gaussian model we may assume

$$\pi(\theta) = \pi(\mu) \times \pi(\sigma^2) = N(\mu_0, \tau) \times \text{InvGamma}(\alpha, \beta).$$

Definition 2. Parameters of the prior distribution, in our previous example μ_0, τ, α et β, are called hyper–parameters.
It's up to the statistician to set values for these hyper–parameters. The latter are not fitted but held fixed!

Joint and posterior distributions

- Since θ is random and, given a statistical model $\{f(x \mid \theta) \colon x \in E, \theta \in \Theta\}$, compute the joint distribution of θ and x, i.e.,

$$\pi(x, \theta) = f(x \mid \theta)\pi(\theta)$$

- However, the joint distribution is rarely useful in a Bayesian framework, and focus is rather put on the posterior distribution

Definition 3. The posterior distribution is the distribution whose density is

$$\pi(\theta \mid x) = \frac{f(x \mid \theta)\pi(\theta)}{\int_\Theta f(x \mid \theta)\pi(\theta)\mathrm{d}\theta}.$$

Joint and posterior distributions

- Since θ is random and, given a statistical model $\{f(x \mid \theta) \colon x \in E, \theta \in \Theta\}$, compute the joint distribution of θ and x, i.e.,

$$\pi(x, \theta) = f(x \mid \theta)\pi(\theta)$$

- However, the joint distribution is rarely useful in a Bayesian framework, and focus is rather put on the posterior distribution

Definition 3. The posterior distribution is the distribution whose density is

$$\pi(\theta \mid x) = \frac{f(x \mid \theta)\pi(\theta)}{\int_\Theta f(x \mid \theta)\pi(\theta)\mathrm{d}\theta}.$$

⚠ Beware, we will always use $f(x \mid \theta)$ to refer to, depending on the situation, either the probability density/mass function or the likelihood evaluated at $x \in \mathbb{R}$.

Marginal distribution and normalizing constant

$$\pi(\theta \mid x) = \frac{f(x \mid \theta)\pi(\theta)}{\int_\Theta f(x \mid \theta)\pi(\theta)\mathrm{d}\theta}.$$

- Denominator is the marginal distribution for x—that may be denoted by $m(x)$.
- It is a normalizing constant for the posterior distribution and, as so, is independent of θ.

Marginal distribution and normalizing constant

$$\pi(\theta \mid x) = \frac{f(x \mid \theta)\pi(\theta)}{\int_\Theta f(x \mid \theta)\pi(\theta)\mathrm{d}\theta}.$$

- Denominator is the marginal distribution for x—that may be denoted by $m(x)$.
- It is a normalizing constant for the posterior distribution and, as so, is independent of θ.

☞ Hence, most often, we will work up to normalizing constants, i.e.,

$$\pi(\theta \mid x) = \frac{f(x \mid \theta)\pi(\theta)}{\int_\Theta f(x \mid \theta)\pi(\theta)\mathrm{d}\theta} \propto f(x \mid \theta)\pi(\theta).$$

By the was, why Bayesian statistics?

- Recall Bayes theorem

$$\Pr(Y \in A \mid X \in B) = \frac{\Pr(X \in B \mid Y \in A)\Pr(Y \in A)}{\Pr(X \in B)}.$$

By the was, why Bayesian statistics?

- Recall Bayes theorem

$$\Pr(Y \in A \mid X \in B) = \frac{\Pr(X \in B \mid Y \in A)\Pr(Y \in A)}{\Pr(X \in B)}.$$

- Since θ is a random variable, we can set $Y = \theta$ to get

$$\Pr(\theta \in A \mid X \in B) = \frac{\Pr(X \in B \mid \theta \in A)\Pr(\theta \in A)}{\Pr(X \in B)}.$$

By the was, why Bayesian statistics?

- Recall Bayes theorem

$$\Pr(Y \in A \mid X \in B) = \frac{\Pr(X \in B \mid Y \in A)\Pr(Y \in A)}{\Pr(X \in B)}.$$

- Since θ is a random variable, we can set $Y = \theta$ to get

$$\Pr(\theta \in A \mid X \in B) = \frac{\Pr(X \in B \mid \theta \in A)\Pr(\theta \in A)}{\Pr(X \in B)}.$$

- Hence using our previous notations we get in terms of p.d.f.

$$\pi(\theta \mid x) = \frac{f(x \mid \theta)\pi(\theta)}{\int_\Theta f(x \mid \theta)\pi(\theta)\mathrm{d}\theta}.$$

Sleep

- Let X be the number of students among n that will sleep during today's lecture.
- Focus is on the unknown probability

$$p = \Pr(\text{a student will sleep}).$$

- What prior distribution we may use? What are the joint and posterior distributions?

Numerical illustration

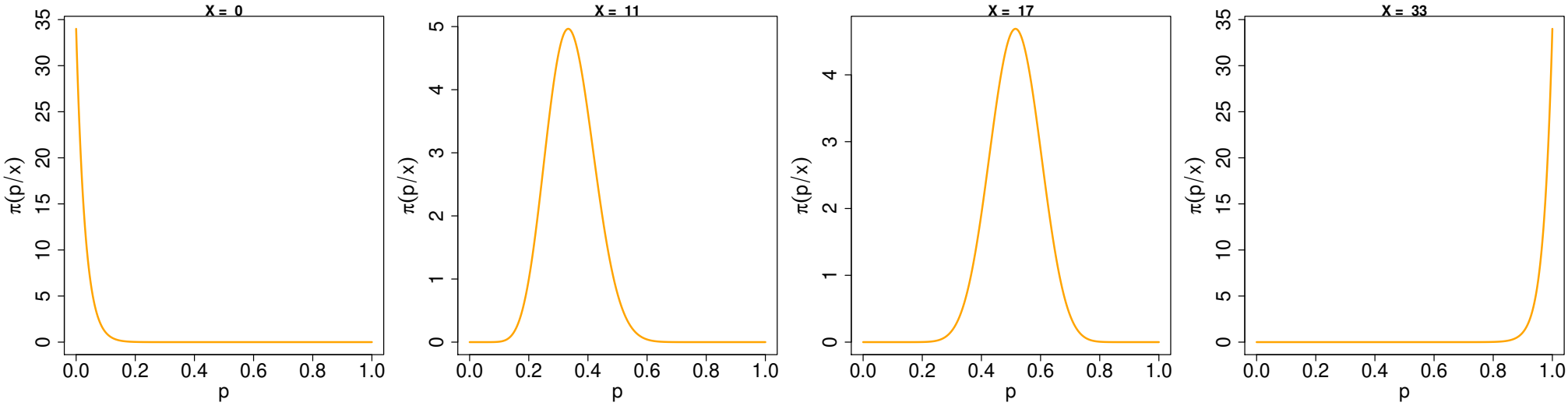

Figure 1: *Plot of the posterior distribution* $\pi(p \mid X)$ *as* X *changes and where* $n = 33$.

3. Prior distribution

Definition 4. A family of distributions $\mathscr{F}$ defined on Θ is said conjugate for the statistical model $\{f(x \mid \theta) : x \in E, \theta \in \Theta\}$ if, for all $\pi \in \mathscr{F}$, the posterior distribution

$$\pi(\theta \mid x) \propto f(x \mid \theta)\pi(\theta)$$

belongs to $\mathscr{F}$.

Conjugate priors

Definition 4. A family of distributions $\mathscr{F}$ defined on Θ is said conjugate for the statistical model $\{f(x \mid \theta) : x \in E, \theta \in \Theta\}$ if, for all $\pi \in \mathscr{F}$, the posterior distribution

$$\pi(\theta \mid x) \propto f(x \mid \theta)\pi(\theta)$$

belongs to $\mathscr{F}$.

☞ The use of conjugate priors is often a "trick" to get explicit posterior distributions.

Conjugate priors

Definition 4. A family of distributions $\mathscr{F}$ defined on Θ is said conjugate for the statistical model $\{f(x \mid \theta) : x \in E, \theta \in \Theta\}$ if, for all $\pi \in \mathscr{F}$, the posterior distribution

$$\pi(\theta \mid x) \propto f(x \mid \theta)\pi(\theta)$$

belongs to $\mathscr{F}$.

☞ The use of conjugate priors is often a "trick" to get explicit posterior distributions.

⏸ We will see later, in the advanced Bayesian statistics course, that it will be very useful for MCMC algorithms.

Sleep 2

Consider our sleep example again but now suppose that the prior distribution is Beta(α, β), i.e., whose density is

$$f(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)} 1_{\{0 \leq x \leq 1\}},$$

where $\alpha > 0$, $\beta > 0$ and $B(\cdot, \cdot)$ is the Beta function.

- The posterior distribution is therefore...

Illustration

$$f(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}, \quad \mathbb{E}[X] = \frac{\alpha}{\alpha+\beta}, \quad \text{Var}[X] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

- $\alpha \gg \beta \Rightarrow$ puts more weight close to 1, i.e., bad teacher
- $\alpha \ll \beta \Rightarrow$ puts more weight close to 0, i.e., good teacher

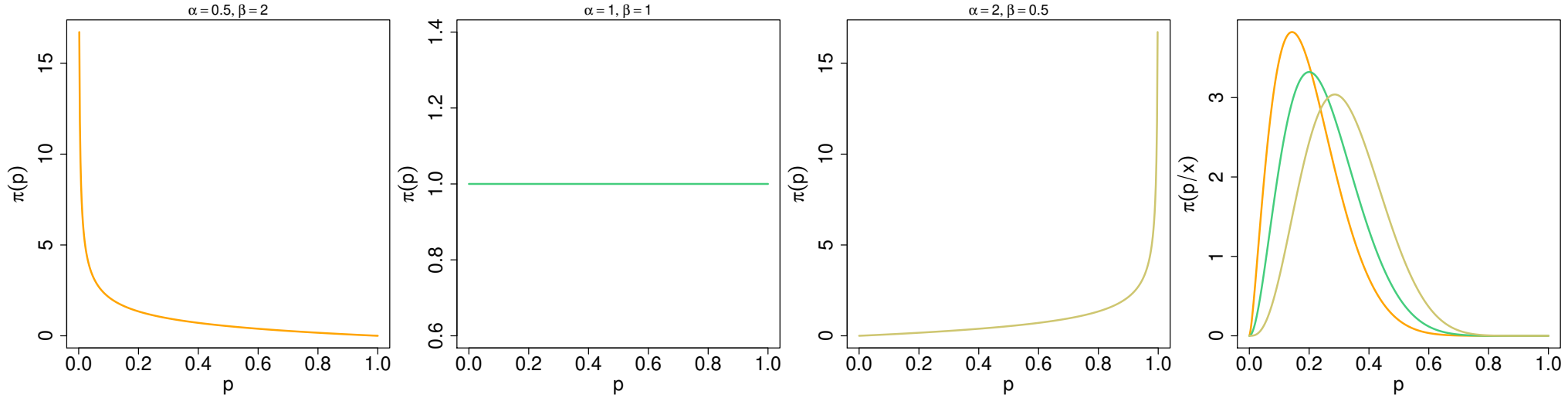

Figure 2: *Effect of the prior distribution on the posterior distribution ($n = 10$ and $x = 2$).*

Teasing. . .

- We will see that in advanced Bayesian statistics that as $n \to \infty$ the impact of $\pi(\theta)$ will (most often) be negligible
- For now we will just "prove it" numerically

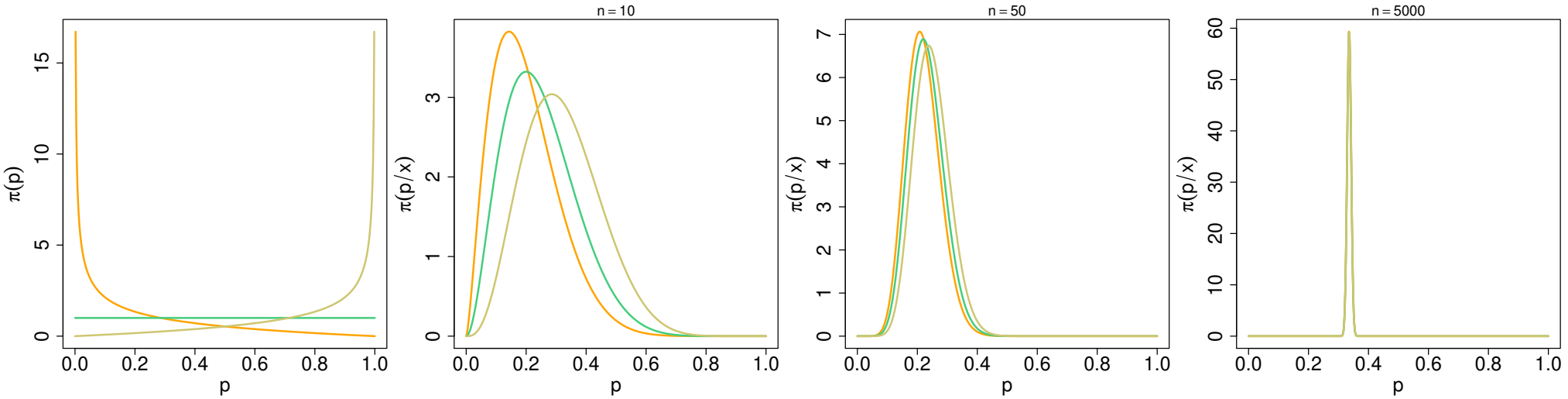

Figure 3: *Changes in the posterior distribution as the sample size n increases ($p_* = 1/3$).*

Improper distributions

Definition 5. A measure μ on E is improper if it is σ-finite (but non finite) but, however, there exists a partition $\{E_n : n \in I\}$, $I \subseteq \mathbb{N}$, at most enumerable, of E such that

$$\mu(E_n) < \infty, \qquad \forall n \in I.$$

Example 3. The following measures defined on $(0, \infty)$

$$\mathrm{d}\mu_1(x) = \mathrm{d}x, \qquad \mathrm{d}\mu_2(x) = x^{-2}\mathrm{d}x$$

are improper distributions.

Use of improper distributions

- One can use improper prior distributions...

Use of improper distributions

- One can use improper prior distributions...
- ...but we have to check that

$$m(x) = \int_{\Theta} f(x \mid \theta)\pi(\theta)\mathrm{d}\theta < \infty,$$

so that the posterior distribution is well defined

$$\pi(\theta \mid x) = \frac{f(x \mid \theta)\pi(\theta)}{m(x)}.$$

Gaussian case

Example 4. Having observed a n-sample $\mathbf{x} = (x_1, \ldots, x_n)$ of iid random variables. Consider the Gaussian model $N(\mu, 1)$ and take as prior distribution $\pi(\mu) \propto 1_{\{\mu \in \mathbb{R}\}}$. What is $m(\mathbf{x})$?

Non informative priors

- We see that the posterior distribution depends on the prior distribution.
- It is weak point of Bayesian statistics and often criticized by purely frequentist statisticians

Non informative priors

- We see that the posterior distribution depends on the prior distribution.
- It is weak point of Bayesian statistics and often criticized by purely frequentist statisticians
- Wouldn't it be nice to "annihilate" the effect of prior distribution?

Non informative priors

- We see that the posterior distribution depends on the prior distribution.
- It is weak point of Bayesian statistics and often criticized by purely frequentist statisticians
- Wouldn't it be nice to "annihilate" the effect of prior distribution?
- This is the (unreachable) aim of using non informative prior distributions for which
  - we encode our total ignorance about θ;
  - it doesn't influence the posterior distribution.
- For this lecture we will focus on the two main non informative prior distributions: Laplace and Jeffreys prior distributions.

Non informative priors

- We see that the posterior distribution depends on the prior distribution.
- It is weak point of Bayesian statistics and often criticized by purely frequentist statisticians
- Wouldn't it be nice to "annihilate" the effect of prior distribution?
- This is the (unreachable) aim of using non informative prior distributions for which
  - we encode our total ignorance about θ;
  - it doesn't influence the posterior distribution.
- For this lecture we will focus on the two main non informative prior distributions: Laplace and Jeffreys prior distributions.

☞ The notion of non informative prior distribution is controversial.

Laplace prior distributions

Definition 6. Laplace priors place as prior distribution $\pi(\theta) \propto 1_{\{\theta \in \Theta\}}$, i.e., the uniform distribution over the set Θ provided that $\int_\Theta \pi(\theta) \mathrm{d}\theta < \infty$. If not, it is an improper uniform distribution on Θ, i.e., the underlying measure is Lebesgue.

Laplace prior distributions

Definition 6. Laplace priors place as prior distribution $\pi(\theta) \propto 1_{\{\theta \in \Theta\}}$, i.e., the uniform distribution over the set Θ provided that $\int_\Theta \pi(\theta) \mathrm{d}\theta < \infty$. If not, it is an improper uniform distribution on Θ, i.e., the underlying measure is Lebesgue.

☞ Laplace prior distribution has some limitations:

- ☐ it could lead to improper posterior distribution and Bayesian inference is hopeless
- ☐ it is not invariant to reparemetrization as shown by the example below

Example 5. Consider the Exponential(λ) model, $\lambda > 0$, and the following reparametrization $\lambda = t^{-1}(\theta)$, $\theta \in \mathbb{R}$, with $t^{-1}(x) = \exp(x)$. First case we have

$$\pi_1(\lambda) \propto 1_{\{\lambda > 0\}},$$

while in the second, using chain rule,

$$\pi_2(\theta) \propto 1 \Longrightarrow \pi_2(\lambda) = t'(\lambda)\pi_2\{t(\lambda)\} \propto \lambda^{-1} 1_{\{\lambda > 0\}}.$$

Definition 7. Consider the statistical model $\{f(x;\theta)\colon x \in E, \theta \in \Theta\}$. The Fisher information is given by

$$I(\theta) = \mathbb{E}\left[\{\nabla_\theta \ln f(X;\theta)\}^\top \nabla_\theta \ln f(X;\theta)\right], \qquad X \sim f(\cdot;\theta).$$

Under some regularity conditions (inversion of integral and derivation signs), we have

$$I(\theta) = -\mathbb{E}\left[\nabla^2_\theta \ln f(X;\theta)\right].$$

Jeyffreys prior distributions

Definition 7. Consider the statistical model $\{f(x;\theta)\colon x \in E, \theta \in \Theta\}$. The Fisher information is given by

$$I(\theta) = \mathbb{E}\left[\{\nabla_\theta \ln f(X;\theta)\}^\top \nabla_\theta \ln f(X;\theta)\right], \qquad X \sim f(\cdot;\theta).$$

Under some regularity conditions (inversion of integral and derivation signs), we have

$$I(\theta) = -\mathbb{E}\left[\nabla^2_\theta \ln f(X;\theta)\right].$$

Definition 8. Jeffreys prior distribution is given by $\pi(\theta) \propto |I(\theta)|^{1/2}$ where $|A|$ corresponds is the determinant of the matrix A.

- it is invariant to reparametrization
- it could lead to improper prior distributions
- it is usually not recommended in high dimensional spaces $\dim \Theta \geq 1$.

Jeyffreys prior distributions

Definition 7. Consider the statistical model $\{f(x;\theta)\colon x \in E, \theta \in \Theta\}$. The Fisher information is given by

$$I(\theta) = \mathbb{E}\left[\{\nabla_\theta \ln f(X;\theta)\}^\top \nabla_\theta \ln f(X;\theta)\right], \qquad X \sim f(\cdot;\theta).$$

Under some regularity conditions (inversion of integral and derivation signs), we have

$$I(\theta) = -\mathbb{E}\left[\nabla^2_\theta \ln f(X;\theta)\right].$$

Definition 8. Jeffreys prior distribution is given by $\pi(\theta) \propto |I(\theta)|^{1/2}$ where $|A|$ corresponds is the determinant of the matrix A.

- ☐ it is invariant to reparametrization
- ☐ it could lead to improper prior distributions
- ☐ it is usually not recommended in high dimensional spaces $\dim \Theta \geq 1$.

⏸ We will see during labs how to handle it...

Example 6. It can be shown (check it) that Jeyffreys prior for the Exponential(λ) model is

$$\pi_1(\lambda) \propto \lambda^{-1} 1_{\{\lambda>0\}},$$

while for the parametrization Exponential(ψ) where $\psi = \log \lambda$, it is

$$\pi_2(\psi) \propto 1.$$

Using the chain rule as before we have, as stated,

$$\pi_2(\lambda) = (\log \lambda)' \pi_2(\log \lambda) \propto \lambda^{-1} 1_{\{\lambda>0\}}.$$

Sleep 3

- Since your are still not asleep, consider one more time our sleep example. . .
- What is the Jeffreys prior distribution for the parameter p?

4. Bayesian Inference

Motivation

- In a Bayesian framework, the posterior distribution $\pi(\theta \mid x)$ is all we need for inference.
- However it is not usual to get a density as a result.
- It is therefore useful to compute statistical summary of $\pi(\theta \mid x)$, i.e., similar to point estimation in a frequentist framework.
- Typical choices for summary statistics are:
  - posterior mean, i.e., $\mathbb{E}_\pi[\theta \mid x]$
  - posterior median, i.e., $\inf\{m\colon \int \pi(\theta \mid x) 1_{\{\theta \leq m\}} \mathrm{d}m$;
  - maximum a posterior (MAP), i.e., $\arg\max_\theta \pi(\theta \mid x)$;
  - a given posterior quantile of order p.
- We can also focus on credible regions.

Motivation

- In a Bayesian framework, the posterior distribution $\pi(\theta \mid x)$ is all we need for inference.
- However it is not usual to get a density as a result.
- It is therefore useful to compute statistical summary of $\pi(\theta \mid x)$, i.e., similar to point estimation in a frequentist framework.
- Typical choices for summary statistics are:
  - posterior mean, i.e., $\mathbb{E}_\pi[\theta \mid x]$
  - posterior median, i.e., $\inf\{m\colon \int \pi(\theta \mid x) 1_{\{\theta \leq m\}} \mathrm{d}m$;
  - maximum a posterior (MAP), i.e., $\arg\max_\theta \pi(\theta \mid x)$;
  - a given posterior quantile of order p.
- We can also focus on credible regions.

☞ We will see in the advances Bayesian lecture (or not depending on time) that these choices are justified from a decision theory point of view.

Reminder: Confidence intervals (frequentist)

Definition 9. A confidence interval of level α for some unknown quantity $f(\theta_0)$ is a random interval I_α such that $\Pr\{f(\theta_0) \in I_\alpha\} = \alpha$.

Example 7. Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ where σ^2 is considered as known. Then

$$I = \left[\bar{X}_n - 1.96\frac{\sigma}{\sqrt{n}}, \bar{X}_n + 1.96\frac{\sigma}{\sqrt{n}}\right],$$

is a confidence interval fo μ with level 95%.

Reminder: Confidence intervals (frequentist)

Definition 9. A confidence interval of level α for some unknown quantity $f(\theta_0)$ is a random interval I_α such that $\Pr\{f(\theta_0) \in I_\alpha\} = \alpha$.

Example 7. Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ where σ^2 is considered as known. Then

$$I = \left[\bar{X}_n - 1.96\frac{\sigma}{\sqrt{n}}, \bar{X}_n + 1.96\frac{\sigma}{\sqrt{n}}\right],$$

is a confidence interval fo μ with level 95%.

☞ Remind that given a sample, we get a realization of this random interval. If we were able to get K independent copies of our original interval, the level $\alpha\%$ will correspond to the proportion of time the true parameter value θ_0 will belong these confidence interval as $K \to infty$.

Definition 10. For a given prior distribution π, a set $C_x \subset \Theta$ is a set α–credible if

$$\Pr_\pi(\theta \in C_x \mid x) \geq \alpha.$$

Definition 10. For a given prior distribution π, a set $C_x \subset \Theta$ is a set α–credible if

$$\Pr_\pi(\theta \in C_x \mid x) \geq \alpha.$$

❓ Do you really get the following probability?

$$\Pr_\pi(\theta \in C_x \mid x) = \ldots?$$

Definition 10. For a given prior distribution π, a set $C_x \subset \Theta$ is a set α–credible if

$$\Pr_\pi(\theta \in C_x \mid x) \geq \alpha.$$

❓ Do you really get the following probability?

$$\Pr_\pi(\theta \in C_x \mid x) = \int_\Theta 1_{\{\theta \in C_x\}} \dots$$

Definition 10. For a given prior distribution π, a set $C_x \subset \Theta$ is a set α–credible if

$$\Pr_\pi(\theta \in C_x \mid x) \geq \alpha.$$

❓ Do you really get the following probability?

$$\Pr_\pi(\theta \in C_x \mid x) = \int_\Theta 1_{\{\theta \in C_x\}} \pi(\theta \mid x) \mathrm{d}\theta.$$

Credible intervals

First we typically focus on the univariate case, i.e., $\theta \in \mathbb{R}$, or we take the j–component θ_j of θ.

Definition 11. For a given prior distribution π, an interval $I_x \subset \mathbb{R}$ is a credible interval of level level α if

$$\Pr_\pi(\theta \in I_x \mid x) = \alpha.$$

Credible intervals

First we typically focus on the univariate case, i.e., $\theta \in \mathbb{R}$, or we take the j–component θ_j of θ.

Definition 11. For a given prior distribution π, an interval $I_x \subset \mathbb{R}$ is a credible interval of level level α if

$$\Pr_\pi(\theta \in I_x \mid x) = \alpha.$$

☞ Often we typically use symmetric credible intervals, i.e.,

$$I_x = \left[q_\pi\left(\frac{1-\alpha}{2}, x\right), q_\pi\left(\frac{\alpha}{2}, x\right) \right],$$

where

$$q_\pi(p, x) = \inf\left\{u \in \mathbb{R} \colon \Pr_\pi\left(\theta \leq u \mid x\right) \geq 1 - \alpha\right\}.$$

Illustration

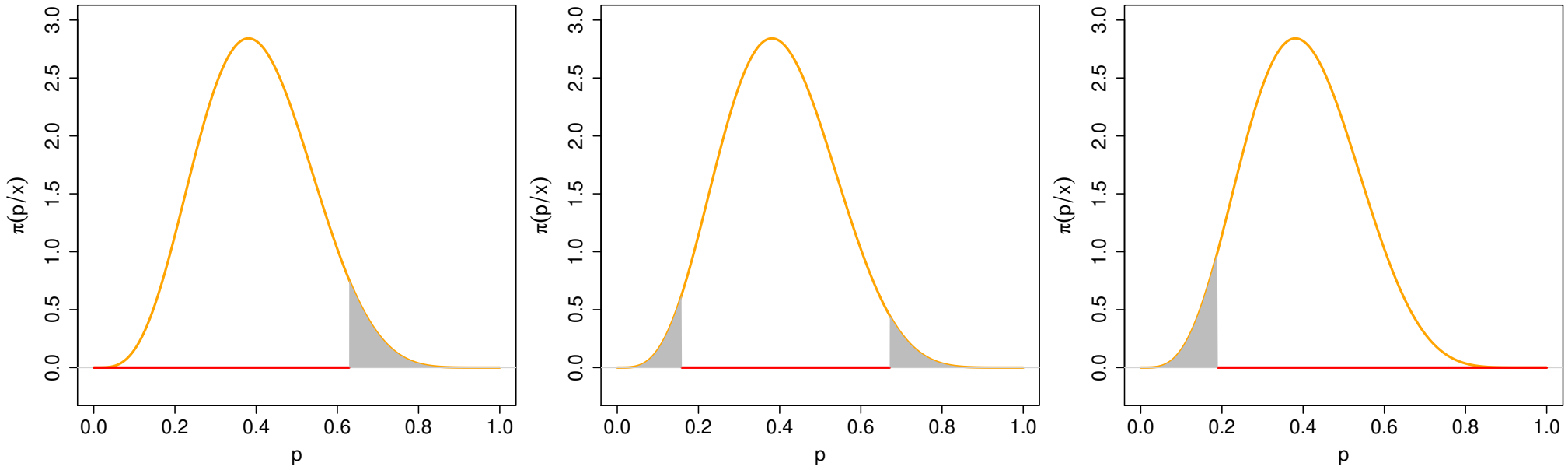

Figure 4: *Three different credible intervals (red intervals) all of them being of level 95%, i.e., the shaded areas are equal to 95%.*

Confidence vs credible intervals

$$\Pr(\theta_0 \in I) = \alpha$$

- I is random
- For a future sample, the probability that θ_0 lies in I is α.

$$\Pr_\pi(\theta \in I \mid x) = \alpha.$$

- θ is random
- Having observed x, thee probality that θ_0 lies in I is α.

High Posterior Density regions (HPD)

Definition 12. A region α–credible C_x is a HPD region α–credible if

$$C_x = \{\theta \in \Theta \colon \pi(\theta \mid x) \geq u_\alpha\}\,, \qquad \text{(region with highest density)}.$$

Figure 5: *Click me!*

High Posterior Density regions (HPD)

Definition 12. A region α–credible C_x is a HPD region α–credible if

$$C_x = \{\theta \in \Theta \colon \pi(\theta \mid x) \geq u_\alpha\}\,, \qquad \text{(region with highest density)}.$$

Figure 5: *Click me!*

☞ HPD regions can be non-connex!

Predictive posterior distribution

- We aim at predicting a future observation x_*.
- In a frequentist framework, we often use as predictor $\mathbb{E}[X]$, $X \sim f(\cdot; \hat{\theta})$, $\hat{\theta}$ is an estimateur of θ.
- However it doesn't take into account for the estimation uncertainty on θ.
- The Bayesian framework naturally take into account uncertainties.

Definition 13. The predictive posterior distribution has density

$$\pi(x_* \mid x) = \int f(x_* \mid \theta, x)\pi(\theta \mid x)\mathrm{d}\theta.$$

Predictive posterior distribution

- We aim at predicting a future observation x_*.
- In a frequentist framework, we often use as predictor $\mathbb{E}[X]$, $X \sim f(\cdot; \hat{\theta})$, $\hat{\theta}$ is an estimateur of θ.
- However it doesn't take into account for the estimation uncertainty on θ.
- The Bayesian framework naturally take into account uncertainties.

Definition 13. The predictive posterior distribution has density

$$\pi(x_* \mid x) = \int f(x_* \mid \theta, x)\pi(\theta \mid x)\mathrm{d}\theta.$$

☞ We will thus a Bayesian predictor based on the above distribution such as

$$\hat{x}_* = \mathbb{E}(X_*) = \int x_*\pi(x_* \mid x)\mathrm{d}x_*,$$

where $X_* \sim \pi(x_* \mid x)$.

Sommeil 4

A new student is just attending my lecture for the first time—he extended is year off. Will he get asleep?

What we did not cover?

- Bayesian statistics became increasingly popular since 2000—implementation on desktop computer were possible.
- In this course we just stick to a very simple framework
- Fortunately, next semester you will attend the Advanced Bayesian course where the following topics will be covered:
  - Monte Carlo Markov chain algorithms including Gibbs sampler
  - Hierarchical models
  - Directed acyclic graph
  - Modelling on complex models