Bayesian Statistics

Fabio Sigrist

ETH Zurich, Autumn Semester 2019

Today's topics

- Credible sets
- Bayesian asymptotics
- Likelihood principle
- Prior distributions: conjugate priors

Credible sets

Credible set

- Bayesian confidence sets are called **credible set**

- Specifically, a subset $C_x \subset \Theta$ with

$$P(\theta \in C_x \mid X = x) = \pi(C_x \mid x) \geq 1 - \alpha$$

is called a **$(1 - \alpha)$-credible set**

Clicker question

Comments

- x is fixed and θ is random

- In contrast, in frequentist statistics, θ is fixed and x is random, and we require

$$P_\theta(C_x \ni \theta) \geq 1 - \alpha \quad \forall \theta \in \Theta$$

- For one dimensional parameters, credible sets are also called **credible intervals**

Highest posterior density credible set

- Among the many $(1-\alpha)$-credible sets, the one minimizing the volume is particularly attractive

- This is obtained by taking C_x as a level set of the posterior, $C_x = L_{k_\alpha}$, where

$$L_k = \{\theta; \pi(\theta \mid x) \geq k\}, \quad k_\alpha = \sup\{k; \pi(L_k \mid x) \geq 1-\alpha\}$$

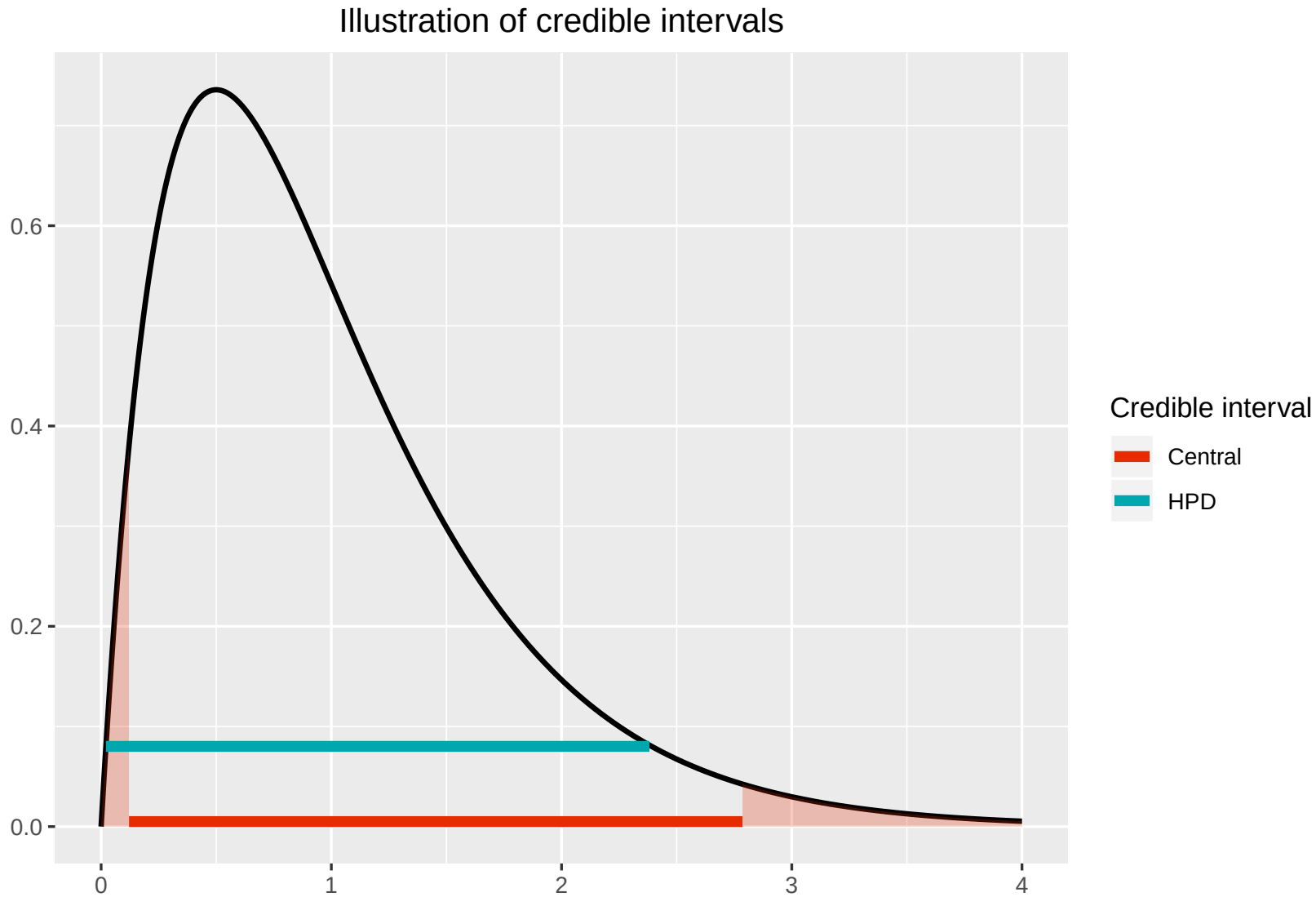
and it is called a **highest posterior density (HPD) credible set**

See blackboard

- For one dimensional parameters, the **central credible interval** (c_l, c_u) is defined such that

$$P(\theta < c_l \mid X = x) = \alpha/2 \text{ and } P(\theta > c_u \mid X = x) = \alpha/2$$

Example of credible intervals



Bayesian asymptotics

Frequentist asymptotics

Consider X_i i.i.d. $\sim f(x \mid \theta)dx$ where $\theta \in \Theta$ is an open set in $\mathbb{R}^p$

Denote by

- $L_n(\theta) = \prod_{i=1}^n f(x_i \mid \theta)$ the likelihood function
- $\widehat{\theta}_n = \arg\max_\theta L_n(\theta)$ the maximum likelihood estimator (MLE)
- $I(\theta) = -\mathbb{E}_\theta \left(\frac{\partial^2}{\partial\theta\partial\theta^T} \log f(X_i \mid \theta) \right)$ the Fisher information

Frequentist asymptotics

The two main results in frequentist statistics are:*

1. $\widehat{\theta}_n \overset{\text{approx}}{\sim} \mathcal{N}\left(\theta_0, \frac{1}{n} I(\theta_0)^{-1}\right)$
2. $2\left(\log L_n(\widehat{\theta}_n) - \log L_n(\theta_0)\right) \xrightarrow{d} \chi^2_p$

- where θ_0 is the "true" parameter

*Under regularity conditions

Bayesian asymptotics

For any smooth prior which is strictly positive in a neighborhood of θ_0, we have*

$$\theta \mid (x_1, \ldots, x_n) \overset{\text{approx}}{\sim} \mathcal{N}\left(\widehat{\theta}_n, \frac{1}{n} I(\widehat{\theta}_n)^{-1}\right)$$

- The **influence of the prior disappears asymptotically and the posterior is concentrated in a $\sqrt{1/n}$ neighborhood of the MLE**

*Again under regularity conditions. This result is known as Bernstein-von Mises theorem

Likelihood principle

Recap: Bayesian vs. frequentist approach

- The **frequentist approach considers other values of the data that did not occur, but could have been obtained** (frequentist properties of estimators when data is drawn repeatedly from the same model)

- The **Bayesian approach considers only the data that were actually observed** (the posterior quantifies the uncertainty about the parameter conditional on the data)

There is an argument that favors the second approach

Conditionality principle

Conditionality principle:

If an experiment for inference about a parameter θ is chosen independently from a collection of different possible experiments, then any experiment not chosen is irrelevant to the inference.

Conditionality principle

Example: Estimation of an unknown concentration of a substance in a probe

Assume there are two labs:

1. 1st lab measures with high precision: standard deviation = 1. Probability that the lab is available is 0.5
2. 2nd lab measures with low precision: standard deviation = 10

- If the measurement was made by the imprecise lab $\Rightarrow$ the true value is within ± 19.6 of the result
- Arguing that because there was a 50% chance to have the analysis done in the precise lab and that therefore the standard deviation is $\sqrt{0.5 \cdot 1^2 + 0.5 \cdot 10^2} = 7.1$ is obviously not reasonable

Sufficiency principle

- A function $T(x)$ of the observation x is **sufficient** for a model $\{f(x|\theta); \theta \in \Theta)\}$ if the conditional distribution of x given $T(x) = t$ does not involve θ

- **Interpretation**: No other statistic that can be calculated from the same sample provides any additional information as to the value of the parameter

Clicker question

Sufficiency principle:

If there are two observations x and y such that $T(x) = T(y)$ for a sufficient statistic T, then any conclusion about θ should be the same for x and y.

Likelihood principle

Likelihood principle:

If there are two different experiments for inference about the same parameter θ and if the outcomes x and y from the two experiments are such that the likelihood functions differ only by a multiplicative constant, then the inference should be the same.

Birnbaum (1962) showed that
Conditionality + sufficiency principle $\Rightarrow$ likelihood principle

- **"Problem":** many frequentist tests and confidence intervals violate this principle

Example violation of likelihood principle

Assume we are interested in an **event with unknown probability *p* and test**

$$H_0 : \ p \leq 0.5$$
$$H_1 : \ p > 0.5$$

This can be done using **two different experiments**:

Exp. 1 Repeat the trial a fixed number n times and observing the random number X of trials where the event occurred

Exp. 2 Repeat the experiment a random number N times until the event has occurred a fixed number of times x

Example violation of likelihood principle

Exp. 1 $P_p(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$

Reject H_0 if $x > c_1 \quad \Leftrightarrow \quad x/n > c_1', \, c_1' = c_1 \cdot n$

Exp. 2 $P_p(N = n) = \binom{n-1}{x-1} p^x (1-p)^{n-x}$

Reject H_0 if $n < c_2 \quad \Leftrightarrow \quad x/n > x/c_2 =: c_2'$

- Both experiments have **likelihood functions $\propto p^x(1-p)^{n-x}$, i.e., they differ only by a multiplicative constant**
- The rejection regions in both experiments have the same form. However, **the boundary values c_1' and c_2' differ** in general because they are computed differently

Example violation of likelihood principle

Conclusions

- Frequentist tests can violate the likelihood principle
- Bayesian tests do not suffer from this drawback
- Point estimation by maximum likelihood does obey the likelihood principle

In defense of the frequentist approach

The following quote from Mosteller and Tukey (1968) justifies the frequentist approach:

"One hallmark of the statistically conscious investigator is his firm belief that however the survey, experiment or observational program actually turned out, it could have turned out somewhat differently. Holding such a belief and taking appropriate actions make effective use of data possible. We need not always ask explicitly "How much differentlys?", but we should be aware of such questions."

Conjugate priors

Conjugate priors

Definition

A parametric family $\mathcal{P}_\Xi = \{\pi_\xi(\theta); \xi \in \Xi\}$, $\Xi \subset \mathbb{R}^q$, of prior densities is called **conjugate** *for the model $\{f(x \mid \theta); \theta \in \Theta\}$ if, for any $\pi \in \mathcal{P}_\Xi$ and any x, $\pi(\theta \mid x)$ is again in $\mathcal{P}_\Xi$.*

- I.e., for any $\xi \in \Xi$ and x there must be a $\xi' = \xi'(\xi, x)$ such that

$$\pi_\xi(\theta) f(x \mid \theta) \propto \pi_{\xi'}(\theta)$$

- Computing the posterior means determining $\xi'(\xi, x)$

Conjugate priors

- The following two conditions guarantee that $\mathcal{P}_\Xi$ is conjugate:

  1. For any x, there is a $\xi(x) \in \Xi$ such that $f(x \mid \theta) \propto \pi_{\xi(x)}(\theta)$

  2. For any pair $\xi_1, \xi_2 \in \Xi$, there is a $\xi_3 \in \Xi$ such that $\pi_{\xi_1}(\theta)\pi_{\xi_2}(\theta) \propto \pi_{\xi_3}(\theta)$

  Example: see blackboard

- A class of conjugate priors $\mathcal{P}_\Xi$ remains conjugate under repeated sampling

  See blackboard

Conjugate priors

If $\mathcal{P}_\Xi$ is conjugate for $f(x \mid \theta)$, for arbitrary, but fixed ξ_0, we can write

$$\prod_{i=1}^{n} f(x_i \mid \theta) = \frac{\pi_{\xi_n(x_1,\ldots,x_n)}(\theta)}{\pi_{\xi_0}(\theta)} f_n(x_1, \ldots, x_n)$$

where

- f_n is the prior predictive density of $X_1, \ldots, X_n$
- ξ_n maps n-tupels of observed values to Ξ

- ξ_n is a sufficient statistic whose dimension is independent of n

  See blackboard

Conjugate priors & exponential families

- One can show that if the set $\{x; f(x \mid \theta) > 0\}$ does not depend on θ, exponential families are the only class of distributions which allow for sufficient statistics whose dimension is independent of n

- An **exponential family** has densities of the following form

$$f(x \mid \theta) = \exp(c_1(\theta)T_1(x) + \ldots c_q(\theta)T_q(x) + d(\theta))h(x)$$

The conjugate family consists then of densities

$$\pi_\xi(\theta) \propto \exp(c_1(\theta)\xi_1 + \ldots c_q(\theta)\xi_q + d(\theta)\xi_{q+1})$$

Conjugate priors & exponential families

Examples of exponential families and their conjugate prior distributions.

Model	Prior	Posterior
Binomial(n, θ)	Beta(α, β)	Beta$(\alpha + x, \beta + n - x)$
Multinomial $(n, \theta_1, \ldots, \theta_k)$	Dirichlet$(\alpha_1, \ldots, \alpha_k)$	Dirichlet$(\alpha_1 + x_1, \ldots, \alpha_k + x_k)$
i.i.d. Poisson(θ)	Gamma(γ, λ)	Gamma$(\gamma + \sum_i x_i, \lambda + n)$
i.i.d. Normal$(\mu, \frac{1}{\tau})$ $\theta = (\mu, \tau)$	Normal$(\mu_0, \frac{1}{n_0 \tau}) \times$ Gamma(γ, λ)	Normal$(\frac{n}{n+n_0} \bar{x} + \frac{n_0}{n+n_0} \mu_0, \frac{1}{(n+n_0)\tau}) \times$ Gamma$(\gamma + \frac{n}{2}, \lambda + \frac{1}{2} \sum_i (x_i - \bar{x})^2 + \frac{n n_0}{2(n+n_0)} (\bar{x} - \mu_0)^2)$
Uniform$(0, \theta)$	Pareto(α, σ)	Pareto$(\alpha + n, \max(\sigma, x_1, \ldots, x_n))$

See exercises and blackboard for details

Hyperparameters

- Conjugate priors have again parameters called **hyperparameters**
- How to choose these hyperparameters?
  - Often, one of the hyperparameters can be regarded as a ”hypothetical sample size” of the prior
  - The other parameters usually are related to a location parameter of the prior