Bayesian Statistics

Fabio Sigrist

ETH Zurich, Autumn Semester 2019

Today's topics

- Point estimation & decision theory
- Testing and Bayes factor

Summary of important quantities in Bayesian statistics

- **Likelihood**: $f(x|\theta)$
- **Prior**: $\pi(\theta)$
- **Posterior**: $\pi(\theta|x) = \frac{\pi(\theta)f(x|\theta)}{f(x)} \propto \pi(\theta)f(x \mid \theta)$
- **Marginal likelihood (=prior predictive density)**: $f(x) = \int f(x \mid \theta)\pi(\theta)d\theta$
- **Posterior predictive density:** $f(y \mid x) = \int f(y \mid \theta, x)\pi(\theta|x)d\theta$

Point estimation and decision theory

Bayesian point estimates

A Bayesian point estimate summarizes the posterior distribution in a number. The following estimates for the location are often used:

- **Posterior mean**

$$\widehat{\theta} = \mathbb{E}(\theta \mid x) = \int_{-\infty}^{\infty} \theta \pi(\theta \mid x) d\theta$$

- **Posterior median**: solution of the equation

$$\int_{-\infty}^{\widehat{\theta}} \pi(\theta) f(x \mid \theta) d\theta = \frac{1}{2} \int_{-\infty}^{\infty} \pi(\theta) f(x \mid \theta) d\theta$$

- **Posterior mode**

$$\widehat{\theta} = \arg\max_{\theta} \pi(\theta \mid x) = \arg\max_{\theta} \left(\log \pi(\theta) + \log f(x \mid \theta) \right)$$

  - **Comment**: If the prior $\pi(\theta)$ is uniform $\pi(\theta) \propto 1$, the posterior mode equals the maximum likelihood estimator

Bayesian decision theory

Bayesian decision theory provides a unified approach for the above point estimates:

1. Choose a **loss function**

$$L : \Theta \times \Theta \to [0, \infty)$$

2. Minimize the **posterior risk** :

$$\widehat{\theta} = \arg\min_{T} \rho(T(x), \pi)$$

$$\rho(T(x), \pi) = \mathbb{E}(L(T(X), \theta) \mid x) = \int_{\Theta} L(T(x), \theta)\pi(\theta \mid x) d\theta$$

- $L(T(x), \theta)$ is the loss if the true value is θ and the estimate is $T(x)$

Bayesian decision theory

We have the following relationship between loss functions and Bayesian point estimates:

- If $L(T, \theta) = (T - \theta)^2$, we obtain the posterior mean
- If $L(T, \theta) = |T - \theta|$, we obtain the posterior median
- If $L(T, \theta) = 1_{[-\varepsilon,\varepsilon]^c}(T - \theta)$ and we let ε go to zero, we obtain the posterior mode

Bayesian decision theory

Comments

- The posterior risk $\rho(T(x), \pi)$ is the expected loss under the posterior
  - It is obtained by integrating the loss function over the posterior of the parameter θ
  - It depends on the data x but not on the parameter θ

- We call an estimator T that minimizes the posterior risk a **Bayes estimator**

Comparison to frequentist approach

In **frequentist decision theory**

1. One also choose a loss function $L : \Theta \times \Theta \to [0, \infty)$
2. And considers the **frequentist risk**

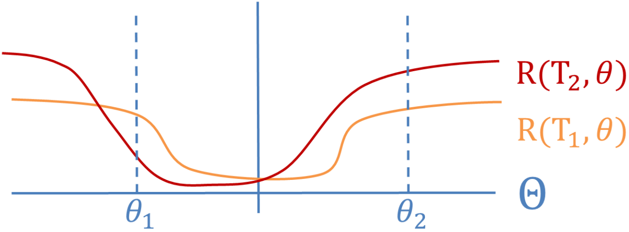
$$R(T, \theta) = \mathbb{E}_\theta(L(T(X), \theta)) = \int_{\mathbf{X}} L(T(x), \theta) f(x \mid \theta) dx$$

- The frequentist risk is obtained by integrating the loss function over the data x
- It depends on the parameter θ but not on the data

- How can we minimize the frequentist risk? There are the following approaches:
  1. Simultaneously for all θ? $\to$ not possible (*see next slide*)
  2. Minimax
  3. Minimize weighted risk
  4. Admissibility

Frequentist decision theory: minimize risk simultaneously for all θ?

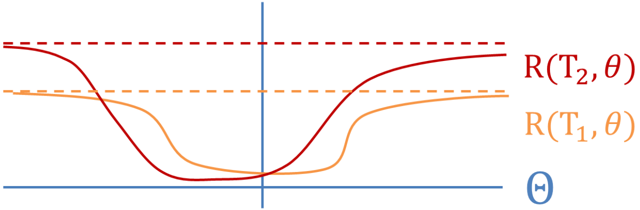
- It is usually not possible to find an estimator that minimizes the frequentist risk for all θ
- **Example**



$$R(T_2, \theta_1) < R(T_1, \theta_1) \quad \text{and} \quad R(T_2, \theta_2) > R(T_1, \theta_1).$$

Frequentist decision theory: minimax approach

- **Minimax** approach: choose estimator T which minimizes the maximal risk: $\sup_{\theta \in \Theta} R(T, \theta)$
- **Example**



$\Rightarrow$ choose T_1

Frequentist decision theory: weighted risk approach

- A less conservative approach is to choose the estimator T which minimizes the **weighted risk**

$$R(T, w) = \int_{\Theta} R(T, \theta) w(\theta) d\theta$$

  $R(T, w)$ is also called the **Bayes risk**

- One can show that the frequentist estimator which minimizes the weighted risk is the Bayes estimator with prior $\pi(\theta) = w(\theta)$ which minimizes the posterior risk (see next slide)

Weighted risk and Bayes estimator

Theorem

Assume that $\int w(\theta)d\theta = 1$ *and choose* w *as the prior for* θ*. If*

$$T(x) = \arg\min_T \rho(T(x), w) = \arg\min_T \mathbb{E}(L(T,\theta) \mid x)$$

is well defined for almost all x *with respect to the prior predictive distribution* $f(x) = \int f(x \mid \theta)w(\theta)d\theta$ *, then* T *minimizes the weighted risk* $R(T, w)$*. Any other minimizer* T' *is almost surely equal to* T*.*

Proof: see blackboard

Conclusion: Even if you are a frequentist, the posterior can be considered as a technical device to compute the estimator which minimizes the weighted risk

Frequentist decision theory: admissibility approach

- An estimator T is called **admissible** if no other estimator T' exists which is uniformly better than T. I.e., if $R(T', \theta) \leq R(T, \theta)$ for all θ, then we must have $R(T', \theta) = R(T, \theta)$ for all θ
- One can show that a Bayes estimator is admissible if the frequentist risk is continuous in θ for any estimator with finite risk and if the prior density is strictly positive everywhere
- There is a large literature showing that under certain conditions any admissible estimator is a limit (in a sense to be made precise) of Bayes estimators

Comment on bias of Bayes estimator

- Bayes estimators are often biased due to the influence of the prior

  See blackboard for example

- However, also from a frequentist point of view, this is not a major disadvantage since modern frequentist statistics tends to deemphasize unbiasedness

Clicker question

Testing and Bayes factor

Comment on bias of Bayes estimator

- In Bayesian statistics, we can make statements about "the probability that the null hypothesis is true" or "the probability that θ belongs to some interval"
- In frequentist statistics, such statements have no meaning, and one has to be very careful if one wants to explain the meaning of a *p*-value or a confidence interval in words

Bayesian testing

Goal: test a null hypothesis $\theta \in \Theta_0 \subset \Theta$ against the alternative $\theta \in \Theta_1 = \Theta_0^c$

- **Posterior probability of the null hypothesis** is given by

$$\pi(\Theta_0 \mid x) = \int_{\Theta_0} \pi(\theta \mid x) d\theta = \frac{\int_{\Theta_0} f(x \mid \theta)\pi(\theta) d\theta}{\int_{\Theta} f(x \mid \theta)\pi(\theta) d\theta}$$

Clicker question

Bayesian testing

- Quantify the loss in case of an error of the first kind as a_1 and in case of an error of the second kind as a_2

- Posterior expected loss of a test $\varphi : \mathbf{X} \to \{0, 1\}^*$ equals

$$\begin{cases} a_2(1 - \pi(\Theta_0 \mid x)) & \text{if } \varphi(x) = 0 \\ a_1\pi(\Theta_0 \mid x) & \text{if } \varphi(x) = 1 \end{cases}$$

- Posterior expected loss is minimized if

$$\varphi(x) = \begin{cases} 0 & \text{if } \pi(\Theta_0 \mid x) > a_2/(a_1 + a_2) \\ 1 & \text{if } \pi(\Theta_0 \mid x) < a_2/(a_1 + a_2) \end{cases}$$

See blackboard

* "1" means reject null hypothesis, "0" means do not reject null hypothesis

Bayes factor

Instead of $\pi(\Theta_0 \mid x)$, Bayesians often consider the **Bayes factor**:

$$B(x) = \frac{\pi(\Theta_0 \mid x)}{\pi(\Theta_1 \mid x)} \frac{\pi(\Theta_1)}{\pi(\Theta_0)} = \frac{\underbrace{\frac{\pi(\Theta_0 \mid x)}{\pi(\Theta_1 \mid x)}}_{\text{"Posterior odds"}}}{\underbrace{\frac{\pi(\Theta_0)}{\pi(\Theta_1)}}_{\text{"Prior odds"}}}$$

- The Bayes factor tells us how the prior odds are modified to obtain the posterior odds:

$$\frac{\pi(\Theta_0 \mid x)}{\pi(\Theta_1 \mid x)} = B(x) \frac{\pi(\Theta_0)}{\pi(\Theta_1)}$$

- Idea of the Bayes factor: partially eliminate the influence of the prior and give more weight to the data

Dependence on prior of Bayes factor

- If $\Theta = \{\theta_0, \theta_1\}$, the Bayes factor is independent of the prior and equal to the likelihood ratio

$$B(x) = f(x \mid \theta_0)/f(x \mid \theta_1)$$

- For **composite hypotheses**, the Bayes factor still depends on the prior

Bayes factor for composite hypotheses

- If
$$\pi_0(\theta) = \frac{\pi(\theta)1_{\Theta_0}(\theta)}{\pi(\Theta_0)}, \quad \pi_1(\theta) = \frac{\pi(\theta)1_{\Theta_1}(\theta)}{\pi(\Theta_1)}$$
denote the conditional priors under the null and the alternative, respectively, then
$$B(x) = \frac{\int_\Theta f(x \mid \theta)\pi_0(\theta)d\theta}{\int_\Theta f(x \mid \theta)\pi_1(\theta)d\theta} = \frac{f(x \mid \theta \in \Theta_0)}{f(x \mid \theta \in \Theta_1)}$$
- Conclusion: the Bayes factor can be seen as a Bayesian likelihood ratio

Decisions based on Bayes factors

- $1 \geq B(x) \geq \frac{1}{3}$ is considered as **weak**
- $\frac{1}{3} \geq B(x) \geq 0.1$ is considered as **substantial**
- $0.1 \geq B(x) \geq 0.01$ is considered as **strong**
- $0.01 \geq B(x)$ is considered as **decisive**

... **evidence** against the null hypothesis according to Jeffreys (1961)

Bayesian testing: point null hypothesis

- In many applications the **null hypothesis** consists of a subset with **Lebesgue measure zero**
  - E.g., for $\mathcal{N}(\mu, \sigma^2)$-observations $\Theta_0 = \{(\mu, \sigma^2); \mu = \mu_0\}$
- If we choose a prior that has a density w.r.t. the Lebesgue measure, the prior and the posterior give zero probability to the null hypothesis $\Rightarrow$ no need to collect data

Bayesian testing: point null hypothesis

- We need to choose a prior which assigns to Θ_0 a probability strictly between 0 and 1

  This can be achieved by a mixture

  $$\pi(d\theta) = \rho_0 \pi_0(d\theta) + (1 - \rho_0)\pi_1(\theta)d\theta$$

  where π_0 is a distribution which is concentrated on Θ_0 and ρ_0 is the prior probability of Θ_0

- With such a prior, the posterior probability of Θ_0 is

$$\pi(\Theta_0 \mid x) = \frac{\rho_0 \int_{\Theta_0} f(x \mid \theta)\pi_0(d\theta)}{\rho_0 \int_{\Theta_0} f(x \mid \theta)\pi_0(d\theta) + (1 - \rho_0) \int_{\Theta} f(x \mid \theta)\pi_1(\theta)d\theta}$$

P-value vs. posterior probability

- In frequentist statistics, the ***p*-value** is taken as a measure of evidence against the null hypothesis
- However, the *p*-value is not the same as the **posterior probability of the null hypothesis**

  How close are these two values?

P-value vs. posterior probability

- Consider $\Theta_0 = \{\theta_0\}$. Then for $\rho_0 = \frac{1}{2}$

$$\pi(\Theta_0 \mid x) = \frac{f(x \mid \theta_0)}{f(x \mid \theta_0) + \int_\Theta f(x \mid \theta)\pi_1(\theta)d\theta}$$

- This depends on the chosen prior for the alternative. But we have a (conservative) lower bound:

$$\inf_{\pi_1} \pi(\Theta_0 \mid x) = \frac{f(x \mid \theta_0)}{f(x \mid \theta_0) + \sup_\theta f(x \mid \theta)}$$

- Further, if one assumes θ to be scalar and restricts π_1 to the class $\mathcal{S}$ of symmetric unimodal densities, then one can show that

$$\inf_{\pi_1 \in \mathcal{S}} \pi(\Theta_0 \mid x) = \frac{f(x \mid \theta_0)}{f(x \mid \theta_0) + \sup_c \frac{1}{2c} \int_{\theta_0 - c}^{\theta_0 + c} f(x \mid \theta)d\theta}$$

P-value vs. posterior probability

Example:* null hypothesis $\mu = \mu_0$ for i.i.d normal observations with mean μ and known variance, $\rho_0 = \frac{1}{2}$

p-value	0.10	0.05	0.01	0.001
$\inf_{\pi_1} \pi(\Theta_0 \mid x)$	0.205	0.128	0.035	0.004
$\inf_{\pi_1 \in \mathcal{S}} \pi(\Theta_0 \mid x)$	0.392	0.290	0.109	0.018

Conclusions

- Posterior probabilities can be substantially larger than *p*-values
- *p*-values can be misleading measures of evidence against the null hypothesis

*Source: Tables 4 and 6 in Berger and Selke, JASA 82 (1987)