

18.650 – Fundamentals of Statistics

5. Bayesian Statistics

Goals

So far, we have followed the *frequentist* approach (cf. meaning of a confidence interval).

An alternative is the **Bayesian approach**.

New concepts will come into play:

- ▶ prior and posterior distributions
- ▶ Bayes' formula
- ▶ Priors: improper, non informative
- ▶ Bayesian estimation: posterior mean, Maximum a posteriori (MAP)
- ▶ Bayesian confidence region

In a sense, Bayesian inference amounts to having a likelihood function $L_n(\theta)$ that is weighted by prior knowledge on what θ might be. This is useful in many applications.

The frequentist approach

- ▶ Assume a statistical model $(E, \{\mathbb{P}_\theta\}_{\theta \in \Theta})$.
- ▶ We assumed that the data X_1, \dots, X_n was drawn i.i.d from \mathbb{P}_{θ^*} for some unknown *fixed* θ^* .
- ▶ When we used the MLE for example, we looked at all possible $\theta \in \Theta$.
- ▶ Before seeing the data we did not prefer a choice of $\theta \in \Theta$ over another.

The Bayesian approach

- ▶ In many practical contexts, we have a *prior belief* about θ^*
- ▶ Using the data, we want to update that belief and transform it into a *posterior belief*.

The kiss example

- ▶ Let p be the proportion of couples that turn their head to the right
- ▶ Let $X_1, \dots, X_n \stackrel{i.i.d}{\sim} \text{Ber}(p)$.
- ▶ In the frequentist approach, we estimated p (using the MLE), we constructed some confidence interval for p , we did hypothesis testing (e.g., $H_0 : p = .5$ v.s. $H_1 : p \neq .5$).
- ▶ Before analyzing the data, we may believe that p is likely to be close to $1/2$.
- ▶ The Bayesian approach is a tool to update our prior belief using the data.

The kiss example

- ▶ Our prior belief about p can be quantified:
- ▶ E.g., we are 90% sure that p is between .4 and .6, 95% that it is between .3 and .8, etc...
- ▶ Hence, we can model our prior belief using a distribution for p , as if p was random.
- ▶ In reality, the true parameter is not random ! However, the Bayesian approach is a way of modeling our belief about the parameter by doing **as if** it was random.
- ▶ E.g., $p \sim \text{Beta}(a, b)$ (*Beta distribution*) It has pdf

$$f(x) = \frac{1}{K} x^{a-1} (1-x)^{b-1} \mathbb{I}(x \in [0, 1]), \quad K = \int_0^1 t^{a-1} (1-t)^{b-1} dt$$

- ▶ This distribution is called the prior distribution

The kiss example

- ▶ In our statistical experiment, X_1, \dots, X_n are assumed to be i.i.d. Bernoulli r.v. with parameter p **conditionally on** \mathcal{P} .
- ▶ After observing the available sample X_1, \dots, X_n , we can update our belief about p by taking its distribution conditionally on the data.
- ▶ The distribution of p conditionally on the data is called the *posterior distribution*
- ▶ Here, the posterior distribution is

$$\text{Beta}\left(a + \sum_{i=1}^n X_i, a + n - \sum_{i=1}^n X_i\right)$$

Clinical trials

Let us revisit our clinical trial example

- ▶ Pharmaceutical companies use hypothesis testing to test if a new drug is efficient.
- ▶ To do so, they administer a drug to a group of patients (test group) and a placebo to another group (control group).
- ▶ We consider testing a drug that is supposed to lower LDL (low-density lipoprotein), a.k.a "bad cholesterol" among patients with a high level of LDL (above 200 mg/dL)

Clinical trials

- ▶ Let $\Delta_d > 0$ denote the expected decrease of LDL level (in mg/dL) for a patient that has used the drug.
- ▶ Let $\Delta_c > 0$ denote the expected decrease of LDL level (in mg/dL) for a patient that has used the placebo.

Quantity of interest: $\theta := \Delta_d - \Delta_c$

In practice we have a prior belief on θ . For example,

- ▶ $\theta \sim \text{Unif}([100, 200])$
- ▶ $\theta \sim \text{Exp}(100)$
- ▶ $\theta \sim \mathcal{N}(100, 300),$
- ▶ ...

Prior and posterior

- ▶ Consider a probability distribution on a parameter space Θ with some pdf $\underline{\pi(\cdot)}$: the *prior distribution*.
- ▶ Let X_1, \dots, X_n be a sample of n random variables.
- ▶ Denote by $L_n(\cdot|\theta)$ the joint pdf of X_1, \dots, X_n conditionally on θ , where $\theta \sim \pi$.
- ▶ **Remark:** $L_n(X_1, \dots, X_n|\theta)$ is the *likelihood* used in the frequentist approach.
- ▶ The conditional distribution of θ given X_1, \dots, X_n is called the *posterior distribution*. Denote by $\pi(\cdot|X_1, \dots, X_n)$ its pdf.

Bayes' formula

- Bayes' formula states that:

$$\pi(\theta|X_1, \dots, X_n) \propto \pi(\theta)L_n(X_1, \dots, X_n|\theta), \quad \forall \theta \in \Theta.$$

proportional
up to constant
that does not depend
on θ

- The constant does not depend on θ :

$$\pi(\theta|X_1, \dots, X_n) = \frac{\pi(\theta)L_n(X_1, \dots, X_n|\theta)}{\int_{\Theta} \pi(\theta)L_n(X_1, \dots, X_n|\theta)d\theta}, \quad \forall \theta \in \Theta.$$

Bernoulli experiment with a Beta prior

In the Kiss example:

- $p \sim \text{Beta}(a, a)$:

$$\pi(p) \propto p^{a-1} (1-p)^{a-1}, p \in (0, 1)$$

- Given $p, X_1, \dots, X_n \stackrel{i.i.d.}{\sim} \text{Ber}(p)$, so

$$L_n(X_1, \dots, X_n | p) = p^{\sum_{i=1}^n \hat{x}_i} (1-p)^{n - \sum_{i=1}^n \hat{x}_i}$$

- Hence,

$$\pi(p | X_1, \dots, X_n) \propto p^{a-1 + \sum_{i=1}^n \hat{x}_i} (1-p)^{a-1 + n - \sum_{i=1}^n \hat{x}_i}.$$

- The posterior distribution is

$$\text{Beta}\left(a + \sum_{i=1}^n \hat{x}_i, a + n - \sum_{i=1}^n \hat{x}_i\right)$$

Conjugate prior.

Non informative priors

- We can still use a Bayesian approach if we have no prior information about the parameter. How to pick prior π ?
- Good candidate: $\pi(\theta) \propto 1$, i.e., constant pdf on Θ .
- If Θ is bounded, this is the *uniform* prior on Θ .
- If Θ is unbounded, this does not define a proper pdf on Θ !
- An *improper prior* on Θ is a measurable, nonnegative function $\pi(\cdot)$ defined on Θ that is not integrable. *Improper iff $\int \pi(\theta) d\theta = \infty$*
- In general, one can still define a posterior distribution using an improper prior, using Bayes' formula.

Examples

- If $p \sim \mathcal{U}(0, 1)$ and given $p, X_1, \dots, X_n \stackrel{i.i.d.}{\sim} \text{Ber}(p)$:

$$\pi(p|X_1, \dots, X_n) \propto p^{\sum_{i=1}^n X_i} (1-p)^{n - \sum_{i=1}^n X_i} = L_n(X_1, \dots, X_n | p)$$

i.e., the posterior distribution is

$$\text{Beta}\left(1 + \sum_{i=1}^n X_i, n - \sum_{i=1}^n X_i\right)$$

- If $\pi(\theta) = 1, \forall \theta \in \mathbb{R}$ and given $X_1, \dots, X_n | \theta \stackrel{i.i.d.}{\sim} \mathcal{N}(\theta, 1)$:

$$\pi(\theta|X_1, \dots, X_n) \propto \exp\left(-\frac{1}{2} \sum_{i=1}^n (X_i - \theta)^2\right)$$

i.e., the posterior distribution is

$$\mathcal{N}\left(\bar{X}_n, \frac{1}{n}\right)$$

Jeffreys' prior

► Jeffreys prior:

$$\pi_J(\theta) \propto \sqrt{\det I(\theta)}$$

where $I(\theta)$ is the Fisher information matrix of the statistical model associated with X_1, \dots, X_n in the frequentist approach (provided it exists).

► In the previous examples:

- Bernoulli experiment: $\pi_J(p) \propto \frac{1}{\sqrt{p(1-p)}}, p \in (0, 1)$: the prior is Beta($\frac{1}{2}, \frac{1}{2}$).
- Gaussian experiment: $\pi_J(\theta) \propto 1, \theta \in \mathbb{R}$ is an improper prior.

Jeffreys' prior

- ▶ Jeffreys prior satisfies a **reparametrization invariance principle**:
If η is a reparametrization of θ (i.e., $\eta = \phi(\theta)$ for some one-to-one map ϕ), then the pdf $\tilde{\pi}(\cdot)$ of η satisfies:

$$\tilde{\pi}(\eta) \propto \sqrt{\det \tilde{I}(\eta)},$$

where $\tilde{I}(\eta)$ is the Fisher information of the statistical model parametrized by η instead of θ .

Bayesian confidence regions

- ▶ For $\alpha \in (0, 1)$, a Bayesian confidence region with level α is a random subset \mathcal{R} of the parameter space Θ , which depends on the sample X_1, \dots, X_n , such that:

$$\mathbb{P}[\theta \in \mathcal{R} | X_1, \dots, X_n] = 1 - \alpha$$

- ▶ Note that \mathcal{R} depends on the prior $\pi(\cdot)$.
- ▶ "Bayesian confidence region" and "confidence interval" are two **distinct** notions.

Bayesian estimation

- ▶ The Bayesian framework can also be used to estimate the true underlying parameter (hence, in a frequentist approach).
- ▶ In this case, the prior distribution does not reflect a prior belief: It is just an artificial tool used in order to define a new class of estimators.
- ▶ **Back to the frequentist approach:** The sample X_1, \dots, X_n is associated with a statistical model $(E, (\mathbb{P}_\theta)_{\theta \in \Theta})$.
- ▶ Define a prior (that can be improper) with pdf π on the parameter space Θ .
- ▶ Compute the posterior pdf $\pi(\cdot | X_1, \dots, X_n)$ associated with π .

Bayesian estimation

- Bayes estimator:

$$\hat{\theta}^{(\pi)} = \int_{\Theta} \theta \pi(\theta | X_1, \dots, X_n) d\theta$$

This is the *posterior mean*.

- The Bayesian estimator depends on the choice of the prior distribution π (hence the superscript π).
- Another popular choice is the point that maximizes the posterior distribution, provided it is unique. It is called the MAP (maximum a posteriori):

$$\hat{\theta}^{\text{MAP}} = \operatorname{argmax}_{\theta \in \Theta} \frac{\pi(\theta | X_1, \dots, X_n)}{L_n(X_1, \dots, X_n | \theta) \pi(\theta)}$$

Bayesian estimation

- In the previous examples:
 - Kiss example with prior $\text{Beta}(a, a)$ ($a > 0$):

$$\hat{p}^{(\pi)} = \frac{a + \sum_{i=1}^n X_i}{2a + n} = \frac{a/n + \bar{X}_n}{2a/n + 1}.$$

In particular, for $a = 1/2$ (Jeffreys prior),

$$\hat{p}^{(\pi_J)} = \frac{1/(2n) + \bar{X}_n}{1/n + 1}.$$

- Gaussian example with Jeffrey's prior: $\hat{\theta}^{(\pi_J)} = \bar{X}_n$.
- In each of these examples, the Bayes estimator is consistent and asymptotically normal.
- In general, the asymptotic properties of the Bayes estimator do not depend on the choice of the prior.