18.650 – Fundamentals of Statistics

5. Bayesian Statistics

Goals

So far, we have followed the *frequentist* approach (cf. meaning of a confidence interval). start with a model(iid, distribution), than use some methods.

An alternative is the **Bayesian approach**. when i can only do the experiment once

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 $X_1, \ldots, X_n$ was drawn i.i.d from $\mathbb{P}_{\theta^*}$ 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 $X_1, \ldots, X_n \overset{i.i.d}{\sim} \text{Ber}(p)$.
- 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: finding priori is hardest
- 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

on support of [0, 1]

$$f(x) = \frac{1}{K} x^{a-1}(1-x)^{b-1}\mathbb{1}(x \in [0,1]), \quad \text{normalization factor } K = \int_0^1 t^{a-1}(1-t)^{b-1}\,\mathrm{d}t$$

- This distribution is called the prior distribution

The kiss example

- In our statistical experiment, $X_1, \ldots, X_n$ are assumed to be i.i.d. Bernoulli r.v. with parameter p **conditionally on** p. (r.v.)

- After observing the available sample $X_1, \ldots, X_n$, 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

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

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 $\Delta_{\mathrm{d}} > 0$ denote the expected decrease of LDL level (in mg/dL) for a patient that has used the drug.
- Let $\Delta_{\mathrm{c}} > 0$ denote the expected decrease of LDL level (in mg/dL) for a patient that has used the placebo.

Quantity of interest: $\theta := \Delta_{\mathrm{d}} - \Delta_{\mathrm{c}}$.

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

- $\theta \sim \mathsf{Unif}([100, 200])$
- $\theta \sim \mathsf{Exp}(100^{-1})$
- $\theta \sim \mathcal{N}(100, 300)$,
- ...

Prior and posterior

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

Bayes' formula

- Bayes' formula states that:

proportional up to constant that does not depend on θ

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

Total probability and expectation theorems

$$p_X(x) = \sum_y p_Y(y)p_{X|Y}(x\,|\,y) \qquad f_X(x) = \int_{-\infty}^{\infty} f_Y(y) f_{X|Y}(x\,|\,y)\,dy \quad \text{Thm.}$$

- The constant does not depend on θ:

normalized term, normalize to 1

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

$$\begin{aligned} p_{X,Y}(x,y) &= p_X(x)\,p_{Y|X}(y\,|\,x) \\ &= p_Y(y)\,p_{X|Y}(x\,|\,y) \end{aligned} \qquad \begin{aligned} f_{X,Y}(x,y) &= f_X(x)\,f_{Y|X}(y\,|\,x) \\ &= f_Y(y)\,f_{X|Y}(x\,|\,y) \end{aligned}$$

$$p_{X|Y}(x\,|\,y) = \frac{p_X(x)\,p_{Y|X}(y\,|\,x)}{p_Y(y)} \qquad f_{X|Y}(x\,|\,y) = \frac{f_X(x)\,f_{Y|X}(y\,|\,x)}{f_Y(y)}$$

posterior

$$p_Y(y) = \sum_{x'} p_X(x')\,p_{Y|X}(y\,|\,x') \qquad f_Y(y) = \int f_X(x')\,f_{Y|X}(y\,|\,x')\,dx'$$

Bernoulli experiment with a Beta prior

In the Kiss example:

- $p \sim \text{Beta}(a, a)$: 只要normalized term 不 depends on parameters，就可以这样写

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

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

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

- Hence,

$$\pi(p|X_1, \ldots, X_n) \propto p^{a-1+\sum_{i=1}^n X_i}(1-p)^{a-1+n-\sum_{i=1}^n X_i}.$$

- The posterior distribution is

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

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. π improper iff $\int \pi(\theta)d\theta = \infty$

比如pi(theta) = 1

- In general, one can still define a posterior distribution using an improper prior, using Bayes' formula.

Examples

Beta(1, 1)

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

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

i.e., the posterior distribution is

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

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

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

i.e., the posterior distribution is

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

$\theta \sim N(\mu, \sigma^2)$

pdf: $\pi(\theta) \propto \exp(-\frac{1}{2}\frac{(\theta-\mu)^2}{\sigma^2})$

$L(X_1 \ldots X_n|\theta) \propto \exp(-\frac{1}{2}\sum_{i=1}^{n}(X_i - \bar{X}_n + \bar{X}_n - \theta)^2)$

$\Rightarrow -\frac{1}{2}\sum_{i=1}^{n}[(X_i - \bar{X}_n)^2 + (\bar{X}_n - \theta)^2 + \underbrace{2(X_i - \bar{X}_n)}_{0}\underbrace{(\bar{X}_n - \theta)}_{\text{cst}}]$

posterior: $\pi(\theta|X_1 \ldots X_n) \propto \underbrace{\exp(-\frac{1}{2}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2)}_{\text{cst. w.r.t. } \theta} \cdot \exp(-\frac{n}{2}(\bar{X}_n - \theta)^2)$

$\propto \exp(-\frac{n}{2}(\bar{X}_n - \theta)^2)$

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

Jeffreys' prior

- **Jeffreys prior:**

non-informative
选择一个depends on你选的统计模型的prior $\pi_J(\theta) \propto \sqrt{\det I(\theta)}$

det is a way to collapse all the fisher information to one number

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

Fisher information characterizes what points(in the statsmodel you choose) are easy to estiamte
Fisher information is high, and MLE's variance is low, it's easy to estimate.
So points with high Fisher information are easy to estimate, which to be put more prior.

- 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 $X_1, \ldots, X_n$, such that:

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

R depends on prior

- Note that $\mathcal{R}$ depends on the prior $\pi(\cdot)$.

- "Bayesian confidence region" and "confidence interval" are two **distinct** notions.

infinite experiments

这里的randomness来自于每一个data是r.v.
因此不同的experiments会有不同的结果

这里的randomness不来自于每一个数据，因为condition on data
randomness只来自于参数的posterior。
而posterior又来自于prior，也就是randomness来自于你放进prior中的。
如果开始你很确定，那么data改变不了什么
整个过程来自于prior到posterior的过程。

Bayesian estimation

no true prior
parameters were drawn from prior

- 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 $X_1, \ldots, X_n$ 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 $\pi(\cdot|X_1, \ldots, X_n)$ associated with π.

Bayesian estimation

- *Bayes estimator:*

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

  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):

$$\hat{\theta}^{\mathrm{MAP}} = \operatorname*{argmax}_{\theta \in \Theta} \pi(\theta \mid X_1, \ldots, X_n) = L_n(X_1, \ldots, X_n \mid \theta)\, \pi(\theta)$$

MLE

weighted by prior

Bayesian estimation

- In the previous examples:
  - Kiss example with prior 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),

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

    - 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.

因为n>inf, 数据太多了，直到可以忽略掉prior