Introduction to Bayesian Inference

Tathagata Basu

UE de Master 2, AOS1

Autumn 2022

Outline I

Statistical Inference

Statistical Inference

Statistical inference is concerned with drawing conclusions, from *random* numerical data, about quantities that are not observed.

For example, we may collect data to observe the average height of adults in France.

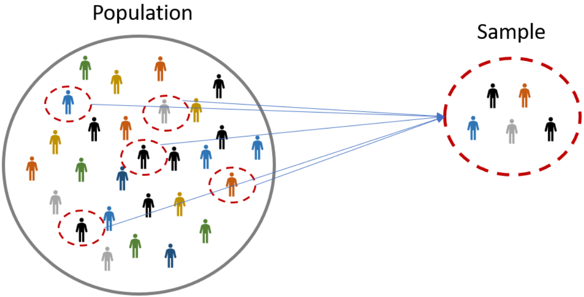
Statistical Inference

Statistical Inference

Statistical inference is concerned with drawing conclusions, from *random* numerical data, about quantities that are not observed.

For example, we may collect data to observe the average height of adults in France.

- However, it is not practical to observe the whole *population* of France.
- Instead, we collect a finite set of observations or *samples* from the population.



pc:medium.com

Inference Methods

Parametric Inference

Parametric methods are based on the assumption that the sample comes from a population that can be modelled by a probability distribution with fixed set of *parameters*.

For example, *likelihood-based approaches, Bayesian approaches.*

Inference Methods

Parametric Inference

Parametric methods are based on the assumption that the sample comes from a population that can be modelled by a probability distribution with fixed set of *parameters*.

For example, *likelihood-based approaches, Bayesian approaches.*

Non-parametric Inference

Non-parametric methods are used when we may not have any distributional assumption.

For example, *order statistics, quantiles.*

Likelihood

Let X denotes a random variable, so that we have an associated probability density function (probability mass function for discrete) $f_X(\cdot \mid \theta)$.

Likelihood

Let X denotes a random variable, so that we have an associated probability density function (probability mass function for discrete) $f_X(\cdot \mid \theta)$.

Now, let, $x_1, x_2, \cdots, x_n$ be n observations of X. Then the joint probability of the observed data is called *likelihood function* and is denoted by $\mathcal{L}(\theta \mid \tilde{x})$ so that

$$
\begin{aligned}
&\mathcal{L}(\theta \mid \tilde{x}) && (1)\\
&= f_X(\tilde{x} \mid \theta) && (2)\\
&= f_X(x_1 \mid x_2, \cdots, x_n, \theta) \cdot f_X(x_2 \mid x_3, \cdots, x_n, \theta) \cdots f_X(x_n \mid \theta) && (3)\\
&= \prod_{i=1}^{n} f_X(x_i \mid \theta) \quad \text{when } x_i\text{'s are independent.} && (4)
\end{aligned}
$$

Example

Let $x_1, x_2, \cdots, x_n$ are i.i.d. normally distributed variables with mean μ and variance σ^2.

Then the likelihood function is given by:

$$\mathcal{L}(\mu, \sigma^2 \mid \tilde{x}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right) \tag{5}$$

$$= \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\left(-\sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2}\right) \tag{6}$$

How can we estimate this μ and σ^2?

Maximum Likelihood Estimation (MLE)

We can estimate the parameter θ from the likelihood function by maximising it.

$$\hat{\theta} = \arg\max_{\theta} \mathcal{L}(\theta \mid \tilde{x}). \tag{7}$$

In many cases, we work with natural logarithm of the likelihood function and we denote it by $\ell(\theta \mid \tilde{x})$ so that

$$\ell(\theta \mid \tilde{x}) = \log\left(\mathcal{L}(\theta \mid \tilde{x})\right). \tag{8}$$

Necessary and Sufficient Conditions

Necessary condition: For p different parameters

$$\frac{\partial \ell}{\partial \theta_1} = \frac{\partial \ell}{\partial \theta_2} = \cdots = \frac{\partial \ell}{\partial \theta_p} = 0 \tag{9}$$

Necessary and Sufficient Conditions

Necessary condition: For p different parameters

$$\frac{\partial \ell}{\partial \theta_1} = \frac{\partial \ell}{\partial \theta_2} = \cdots = \frac{\partial \ell}{\partial \theta_p} = 0 \tag{9}$$

Sufficient condition: Let

$$H(\theta) = \begin{bmatrix} \frac{\partial^2 \ell}{\partial \theta_1^2} & \frac{\partial^2 \ell}{\partial \theta_1 \partial \theta_2} & \cdot & \frac{\partial^2 \ell}{\partial \theta_1 \partial \theta_p} \\ \frac{\partial^2 \ell}{\partial \theta_2 \partial \theta_1} & \frac{\partial^2 \ell}{\partial \theta_2^2} & \cdots & \frac{\partial^2 \ell}{\partial \theta_2 \partial \theta_p} \\ \vdots & & & \\ \frac{\partial^2 \ell}{\partial \theta_p \partial \theta_1} & \frac{\partial^2 \ell}{\partial \theta_p \partial \theta_2} & \cdots & \frac{\partial^2 \ell}{\partial \theta_p^2} \end{bmatrix}. \tag{10}$$

Then $H(\hat{\theta})$ has to be negative (semi)definite.

Example

Let $x_1, x_2, \cdots, x_n$ are i.i.d. normally distributed variables with mean μ and variance σ^2. Then the likelihood function is given by:

$$\mathcal{L}(\mu, \sigma^2 \mid \tilde{x}) = \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\left(-\sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2}\right) \tag{11}$$

Then, necessary condition for MLE gives us

$$\frac{\partial \ell(\mu, \sigma^2 \mid \tilde{x})}{\partial \mu} = \frac{2n(\overline{x} - \mu)}{2\sigma^2} \tag{12}$$

$$\frac{\partial \ell(\mu, \sigma^2 \mid \tilde{x})}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i - \mu)^2 \tag{13}$$

where, $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.
We can show that $\hat{\mu} = \overline{x}$ and $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$.

Bayes' Rule

For any two event A and B, we have

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}. \tag{14}$$

Bayes' Rule

For any two event A and B, we have

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}. \tag{14}$$

Similarly for two continuous random variable X and Y

$$f_{X|Y}(x \mid y) = \frac{f_{Y|X}(y \mid x)f_X(x)}{f_Y(y)}. \tag{15}$$

Then using law of total probability we have

$$f_{X|Y}(x \mid y) = \frac{f_{Y|X}(y \mid x)f_X(x)}{\int_x f_{Y|X}(y \mid x)f_X(x)dx}. \tag{16}$$

Bayesian Inference

Let, $x_1, x_2, \cdots, x_n$ be observations of a random variable with p.d.f $f_X(x \mid \theta)$. Let θ be our parameter of interest. Then

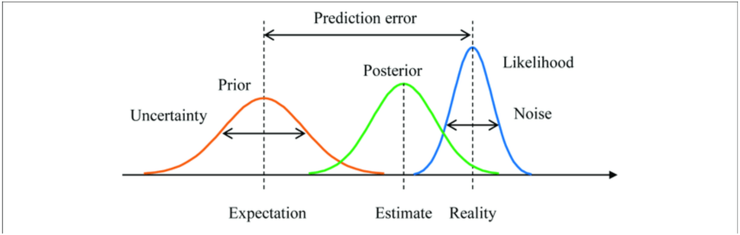
$$f_{\theta|X}(\theta \mid \tilde{x}) = \frac{f_{X|\theta}(\tilde{x} \mid \theta)\pi(\theta)}{\int_\theta f_{X|\theta}(\tilde{x} \mid \theta)\pi(\theta)d\theta} = \frac{\mathcal{L}(\theta \mid \tilde{x})\pi(\theta)}{\int_\theta \mathcal{L}(\theta \mid \tilde{x})\pi(\theta)d\theta}. \tag{17}$$

- $f_{\theta|X}(\theta \mid \tilde{x}) \equiv$ Posterior distribution
- $\pi(\theta) \equiv$ Prior distribution
- $\int_\theta \mathcal{L}(\theta \mid \tilde{x})\pi(\theta)d\theta \equiv$ Marginal likelihood or model evidence

We can simply write

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}. \tag{18}$$

Visualisation



pc:researchgate.net

Choice of priors I

- Subjective Priors Subjective priors are usually used to incorporate one's subjective belief about the modelling parameter. Subjective priors are often elicitation-based and allow us to gather information from previous analysis.
- Prior Predictive Before the data x is observed, we can look into the distribution of this unknown but observable data x, which is given by:

$$f_X(x) = \int_\theta f_{X|\theta}(x \mid \theta)\pi(\theta)d\theta \tag{19}$$

where $f_{X|\theta}(x \mid \theta)$ refers to our sampling distribution of some observable quantity x and $\pi(\theta)$ refers to our prior on the parameter θ. We call this distribution $f_X(x)$ the prior predictive distribution.

Choice of priors II

- Objective Priors Objective prior is an alternative method for describing a prior where we usually use objective source of information about the modelling parameter such as parameter support or sign of the modelling parameter. We often consider these priors as non-informative priors/uninformative priors as they do not posses any other descriptive information.
- Improper Priors Improper priors can also be classified as objective priors. However, improper priors may not integrate to 1. To give some intuition, we can consider an unbounded parameter, then a uniform distribution will result to an improper prior.
- We also have conjugate priors which is used the most because of convenience.

Conjugate Priors

In Bayesian inference, if the posterior distribution $f_{\theta|X}(\theta \mid \tilde{x})$ is in the same probability distribution family as the prior probability distribution $\pi(\theta)$, then the prior and posterior are then called conjugate distributions, and the prior is called a conjugate prior for the likelihood function $\mathcal{L}(\theta \mid \tilde{x})$.
For example, Inverse-gamma distribution is a conjugate prior for the variance of normal distribution

Exponential Family

Let $\theta := (\theta_1, \cdots, \theta_p)$ be a vector of parameters. Then the exponential family of distributions is defined by:

$$f(x \mid \theta) = h(x) \exp\left(\sum_{i=1}^{p} a_i(\theta) T_i(x) - b(\theta)\right) \tag{20}$$

where h, a, T and b are fixed functions for each probability distribution.

Example

In case of a normal distribution, the probability density function is given by:

$$\begin{aligned} f(x \mid \mu, \sigma^2) &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) && (21) \\ &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2 - 2x\mu + \mu^2}{2\sigma^2} - \ln\sigma\right) && (22) \\ &= \frac{1}{\sqrt{2\pi}} \exp\left(\left(-\frac{1}{2\sigma^2}, \frac{\mu}{\sigma^2}\right) \cdot (x^2, x)^T - \frac{\mu^2}{2\sigma^2} - \ln\sigma\right). && (23) \end{aligned}$$

Therefore, $h(x) := \frac{1}{\sqrt{2\pi}}$, $a(\mu, \sigma^2) := \left(-\frac{1}{2\sigma^2}, \frac{\mu}{\sigma^2}\right)$, $T(x) := (x^2, x)$ and $b(\mu, \sigma^2) := \left(\frac{\mu^2}{2\sigma^2} + \ln\sigma\right)$.

Jeffrey's Prior

In Bayesian inference, Jeffrey's prior is an objective prior distribution for a parameter space. Its probability density function is proportional to the square root of the determinant of the Fisher information matrix ($I(\theta)$).

For log-likelihood $\ell(\theta \mid \tilde{x})$, the Fisher information matrix is given by:

$$I(\theta) = \mathbb{E}\left[\left(\frac{\partial \ell(\theta \mid \tilde{x})}{\partial \theta}\right)^2 \mid \theta\right] \tag{24}$$

under regularity conditions

$$= -\mathbb{E}\left[\frac{\partial^2 \ell(\theta \mid \tilde{x})}{\partial \theta^2} \mid \theta\right]. \tag{25}$$

Estimation

- Posterior Mean The most common and convenient way to learn from the posterior distribution is to check the posterior mean given by:

$$\mathbb{E}(\theta \mid X) = \int_{\theta} \theta f_{\theta|X}(\theta \mid \tilde{x}) d\theta. \tag{26}$$

- Posterior Mode Besides posterior mean, we sometimes look for the maximum a posteriori (MAP) estimates. That is we look for the value that achieves greatest posterior density. We look for MAP in the following way:

$$\theta_{\text{MAP}} = \arg\max_{\theta} f_{\theta|X}(\theta \mid \tilde{x}). \tag{27}$$