Bayesian Statistics

Luiz Max de Carvalho[lmax.fgv@gmail.com]

PhD-level course
School of Applied Mathematics (EMAp/FGV), Rio de Janeiro.

March 24, 2021

Welcome!

- This is a 60-hour, PhD-level course on Bayesian inference.
- We have 11 planned weeks. Reading material is posted at https://github.com/maxbiostat/BayesianStatisticsCourse/
- Assessment will be done via a written exam (70%) and an assignment (30%);
- Tenets:
  - Respect the instructor and your classmates;
  - Read before class;
  - Engage in the discussion;
  - Don't be afraid to ask/disagree.
- Books are
  - Robert (2007);
  - Hoff (2009);
  - Bernardo and Smith (2009).

Bayes's Theorem

What do

$$\Pr(A \mid B) = \frac{\Pr(B \mid A)\Pr(A)}{\Pr(B)}, \tag{1}$$

and

$$\Pr(A_i \mid B) = \frac{\Pr(B \mid A)\Pr(A)}{\sum_{i=1}^{n} \Pr(B \mid A_i)\Pr(A_i)}, \tag{2}$$

and

$$p(\theta \mid \boldsymbol{y}) = \frac{l(\boldsymbol{y} \mid \theta)\pi(\theta)}{\int_{\Theta} l(\boldsymbol{y} \mid t)\pi(t)\,dt}, \tag{3}$$

and

$$p(\theta \mid \boldsymbol{y}) = \frac{l(\boldsymbol{y} \mid \theta)\pi(\theta)}{m(\boldsymbol{y})}, \tag{4}$$

all have in common? In this course, we will find out how to use Bayes's rule in order to draw statistical inferences in a coherent and mathematically sound way.

Bayesian Statistics is a complete approach

Our whole paradigm revolves around the posterior:

$$p(\theta \mid \boldsymbol{x}) \propto l(\theta \mid \boldsymbol{x})\pi(\theta).$$

Within the Bayesian paradigm, you are able to

- Perform point and interval inference about unknown quantities;

$$\delta(\boldsymbol{x}) = E_p[\theta] := \int_\Theta t p(t \mid \boldsymbol{x})\, dt,$$

$$\Pr(a \leq \theta \leq b) = 0.95 = \int_a^b p(t \mid \boldsymbol{x})\, dt;$$

- Compare models:

$$\mathrm{BF}_{12} = \frac{\Pr(M_1 \mid \boldsymbol{x})}{\Pr(M_2 \mid \boldsymbol{x})} = \frac{\Pr(\boldsymbol{x} \mid M_1)\Pr(M_1)}{\Pr(\boldsymbol{x} \mid M_2)\Pr(M_2)};$$

- Make predictions: $g(\tilde{x} \mid \boldsymbol{x}) := \int_\Theta f(\tilde{x} \mid t) p(t \mid \boldsymbol{x})\, dt$;
- Make decisions: $E_p[U(r)]$.

Statistical model: informal definition

Stuff you say at the bar:

Definition 1 (Statistical model: informal)

DeGroot, def 7.1.1, pp. 377 A statistical model consists in identifying the random variables of interest (observable and potentially observable), the specification of the joint distribution of these variables and the identification of parameters (θ) that index this joint distribution. Sometimes it is also convenient to assum that the parameters are themselves random variables, but then one needs to specify a joint distribution for θ also.

Statistical model: formal definition

Stuff you say in a Lecture:

Definition 2 (Statistical model: formal)

McCullagh, 2002. Let $\mathcal{X}$ be an arbitrary sample space, Θ a non-empty set and $\mathcal{P}(\mathcal{X})$ the set of all probability distributions on $\mathcal{X}$, i.e. $P : \Theta \to [0, \infty)$, $P \in \mathcal{P}$. A parametric statistical model is a function $P : \Theta \to \mathcal{P}(\mathcal{X})$, that associates each point $\theta \in \Theta$ to a probability distribution P_θ over $\mathcal{X}$.

Examples:

- Put $\mathcal{X} = \mathbb{R}$ and $\Theta = (-\infty, \infty) \times (0, \infty)$. We say P is a *normal* (or *Gaussian*) statistical model[1] if for every $\theta = \{\mu, \sigma^2\} \in \Theta$,

$$P_\theta(x) \equiv \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \; x \in \mathbb{R}.$$

- Put $\mathcal{X} = \mathbb{N} \cup \{0\}$ and $\Theta = (0, \infty)$. P is a Poisson statistical model if, for $\lambda \in \Theta$,

$$P_\lambda(k) \equiv \frac{e^{-\lambda}\lambda^k}{k!}, \; k = 0, 1, \ldots$$

[1] Note the abuse of notation: stricly speaking, P_θ is a probability **measure** and not a *density* as we have presented it here.

Principle I: the sufficiency principle

Sufficiency plays a central role in all of Statistics.

Definition 3 (Sufficient statistic)

*Let $x \sim f(x \mid \theta)$. We say $T : \mathcal{X} \to \mathbb{R}$ is a **sufficient statistic** for the parameter θ if $\Pr(X = x \mid T(x), \theta)$ is independent of θ.*

This is the basis for a cornerstone of Statistics,

Theorem 1 (Factorisation theorem)

Under mild regularity conditions, we can write:

$$f(x \mid \theta) = g(T(x) \mid \theta)h(x \mid T(x)).$$

We can now state

Idea 1 (Sufficiency principle (SP))

For $x, y \in \mathcal{X}$, if T is sufficient for θ and $T(x) = T(y)$, then x and y should lead to the same inferences about θ.

Principle II: the Likelihood principle

The Likelihood Principle (LP) is a key concept in Statistics, of particular Bayesian Statistics.

Idea 2 (Likelihood Principle)

The information brought by an observation $x \in \mathcal{X}$ about a parameter $\theta \in \Theta$ is ***completely*** *contained in the likelihood function $l(\theta \mid x) \propto f(x \mid \theta)$.*

Example 1 (Uma vez Flamengo...)

Principle II: the Likelihood principle

Suppose a pollster is interested in estimating the fraction θ of football fans that cheer for Clube de Regatas do Flamengo (CRF). They survey $n = 12$ people and get $x = 9$ supporters and $y = 3$ "antis". Consider the following two designs:

i) Survey 12 people and record the number of supporters;

ii) Survey until they get $y = 3$.

The likelihoods for both surveys are, respectively,

$$x \sim \text{Binomial}(n, \theta) \implies l_1(\theta \mid x, n) = \binom{n}{x}\theta^x(1-\theta)^{n-x},$$

$$n \sim \text{NegativeBinomial}(y, 1-\theta) \implies l2(\theta \mid n, y) = \binom{n-1}{y-1}y(1-\theta)^{n-y}\theta^y,$$

hence

$$l_1(\theta) \propto l_2(\theta) \propto \theta^3(1-\theta)^9.$$

Therefore, we say that these two experiments bring exactly the same information about θ.

A generalised version of the LP can be stated as follows:

Principle II: the Likelihood principle

Theorem 2 (Likelihood Proportionality Theorem (Gonçalves and Franklin, 2019))

Let Θ be a nonempty set and $\mathcal{P} = \{P_\theta ; \theta \in \Theta\}$ be a family of probability measures on $(\Omega, \mathcal{A})$ and ν_1 and ν_2 be σ-finite measures on $(\Omega, \mathcal{A})$. Suppose $P \ll \nu_1$ and $P \ll \nu_2$ for all $P \in \mathcal{P}$. Then there exists a measurable set $A \in \mathcal{A}$ such that $P_\theta(A) = 1$ for all $\theta \in \Theta$ and there exist $f_{1,\theta} \in \left[\frac{dP_\theta}{d\nu_1}\right]$ and $f_{2,\theta} \in \left[\frac{dP_\theta}{d\nu_2}\right]$ and a measurable function h such that

$$f_{1,\theta}(\omega) = h(\omega) f_{2,\theta}(\omega), \forall\, \theta \in \Theta\, \forall\, \omega \in A.$$

Principle III: stopping rule principle

A subject of contention between inference paradigms is the role of stopping rules in the inferences drawn.

Idea 3 (Stopping rule principle (SRP))

Let τ be a stopping rule directing a series of experiments $\mathcal{E}_1, \mathcal{E}_2, \ldots$, which generates data $\boldsymbol{x} = (x_1, x_2, \ldots)$. Inferences about θ should depend on τ only through $\boldsymbol{x}$.

Example 3 (Finite stopping rules)

Suppose experiment $\mathcal{E}_i$ leads to the observation of $x_i \sim f(x_i \mid \theta)$ and let $\mathcal{A}_i \subset \mathcal{X}_1 \times \ldots \times \mathcal{X}_i$ be a sequence of events. Define

$$\tau := \inf \{n : (x_1, \ldots, x_n) \in \mathcal{A}_n\}.$$

It can be shown that $\Pr(\tau < \infty) = 1$ (exercise 1.20 BC).

Principle IV: the conditionality principle

We will now state one of the main ingredients of the derivation of the LP. The Conditionality Principle (CP) is a statement about the permissible inferences from randomised experiments.

Idea 4 (Conditionality Principle)

Let $\mathcal{E}_1$ and $\mathcal{E}_2$ be two experiments about θ. Let $Z \sim \text{Bernoulli}(p)$ and

- *If $Z = 1$, perform $\mathcal{E}_1$ to generate $x_1 \sim f_1(x_1 \mid \theta)$;*
- *If $Z = 0$ perform $\mathcal{E}_2$ to generate $x_2 \sim f_2(x_2 \mid \theta)$.*

Inferences about θ should depend ***only*** *on the selected experiment, $\mathcal{E}_i$.*

Deriving the Likelihood Principle

Birnbaum (1962) showed that the simpler and mostly uncontroversial Sufficiency and Conditionality principles lead to the Likelihood Principle.

Theorem 2 (Birnbaum's theorem(Birnbaum, 1962))

$$\mathrm{SP} + \mathrm{CP} \implies \mathrm{LP}\,. \tag{5}$$

Proof.

Sketch:

- Define a function $\mathrm{EV}(\mathcal{E}, x)$ to quantify the evidence about θ brought by data x from experiment $\mathcal{E}$ and consider a randomised experiment $\mathcal{E}^*$ in which $\mathcal{E}_1$ and $\mathcal{E}_2$ are performed with probability p;
- Show that CP implies $\mathrm{EV}(\mathcal{E}^*, (j, x_j)) = \mathrm{EV}(\mathcal{E}_j, x_j), j = 1, 2$;
- Show that SP implies $\mathrm{EV}(\mathcal{E}^*, (1, x_1)) = \mathrm{EV}(\mathcal{E}^*, (2, x_2))$ when

$$l(\theta \mid x_1) = cl(\theta \mid x_2).$$

□

See Robert (2007), pg.18 for a complete proof.

References

Bernardo, J. M. and Smith, A. F. (2009). *Bayesian theory*, volume 405. John Wiley & Sons.

Birnbaum, A. (1962). On the foundations of statistical inference. *Journal of the American Statistical Association*, 57(298):269–306.

Gonçalves, F. B. and Franklin, P. (2019). On the definition of likelihood function. *arXiv preprint arXiv:1906.10733*.

Hoff, P. D. (2009). *A first course in Bayesian statistical methods*, volume 580. Springer.

Robert, C. (2007). *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer Science & Business Media.