Bayesian Statistics

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

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

May 5, 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 (2000).

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;

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

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

- 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

FGV EMAp

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: striclty 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 $\operatorname{Pr}(X = x \mid T(x), \theta)$ 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:

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 $\mathbf{x} = (x_1, x_2, \ldots)$. Inferences about θ should depend on τ only through $\mathbf{x}$.

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

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

$\square$

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

Recommended reading

- Robert (2007) Ch. 1;
- Next lecture: Robert (2007) Ch. 2 and * Schervish (2012) Ch.3;

Let F, G and $H \in \mathcal{S}$ be three (possibly overlapping) statements about the world. For example, consider the following statements about a person:

$$F = \{\text{votes for a left-wing candidate}\};$$
$$G = \{\text{is in the 10\% lower income bracket}\};$$
$$H = \{\text{lives in a large}\};$$

Definition 4 (Belief function)

For $A, B \in \mathcal{S}$, a belief function $\operatorname{Be} : \mathcal{S} \to \mathbb{R}$ *assigns numbers to statements such that* $\operatorname{Be}(A) < \operatorname{Be}(B)$ *implies one is more confident in B than in A.*

Belief functions: properties

It is useful to think of Be as **preferences over bets**:

- $\text{Be}(F) > \text{Be}(G)$ means we would bet on F being true over G being true;
- $\text{Be}(F \mid H) > \text{Be}(G \mid H)$ means that, **conditional** on knowing H to be true, we would bet on F over G;
- $\text{Be}(F \mid G) > \text{Be}(F \mid H)$ means that if we were forced to be on F, we would be prefer doing so if G were true than H.

Belief functions: axioms

In order for Be to be **coherent**, it must adhere to a certain set of properties/axioms. A self-sufficient collection is:

A1 (boundedness of complete [dis]belief):

$$\operatorname{Be}(\neg H \mid H) \leq \operatorname{Be}(F \mid H) \leq \operatorname{Be}(H \mid H),\ \forall\, F \in \mathcal{S};$$

A2 (monotonicity):

$$\operatorname{Be}(F \text{ or } G \mid H) \geq \max\left\{\operatorname{Be}(F \mid H), \operatorname{Be}(G \mid H)\right\};$$

A3 (sequentiality): There exists $f : \mathbb{R}^2 \to \mathbb{R}$ such that

$$\operatorname{Be}(F \text{ and } G \mid H) = f\left(\operatorname{Be}(G \mid H), \operatorname{Be}(F \mid G \text{ and } H)\right).$$

Probabilities can be beliefs!

Exercise 1 (Probabilities and beliefs)

Show that the axioms of belief functions map one-to-one to the axioms of probability:

P1. $0 \leq \Pr(E), \forall E \in \mathcal{S}$;

P2. $\Pr(\mathcal{S}) = 1$;

P3. *For any countable sequence of disjoint statements* $E_1, E_2, \ldots \in \mathcal{S}$ *we have*

$$\Pr\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} \Pr(E_i).$$

Hint: derive the consequences (e.g. monotonicity) of these axioms and compare them with the axioms of belief functions.

Definition 5 (Partition)

If $H = \{H_1, H_2, \ldots, H_k\}$, $H_i \in \mathcal{S}$, such that $H_i \cap H_j = \emptyset$ for all $i \neq j$ and $\bigcup_{k=1}^{K} = \mathcal{S}$, we say H is a partition of $\mathcal{S}$.

For any $H \in \mathcal{D}(\mathcal{S})$:

- **Total probability**: $\sum_{k=1}^{K} \Pr(H_k) = 1$;
- **Marginal probability**:

$$\Pr(E) = \sum_{k=1}^{K} = \Pr(E \cap H_k) = \sum_{k=1}^{K} \Pr(E \mid H_k) \Pr(H_k),$$

  for all $E \in \mathcal{S}$;
- Consequence $\implies$ Bayes's rule:

$$\Pr(H_j \mid E) = \frac{\Pr(E \mid H_j) \Pr(H_j)}{\sum_{k=1}^{K} \Pr(E \mid H_k) \Pr(H_k)}.$$

Independence

We will now state a central concept in probability theory and Statistics.

Definition 6 ((Conditional) Independence)

For any $F, G \in \mathcal{S}$, we say F and G are ***conditionally independent*** *given A if*

$$\Pr(F \cap G \mid A) = \Pr(F \mid A) \Pr(G \mid A).$$

Remark 1

If F and G are conditionally independent given A, then

$$\Pr(F \mid A \cap G) = \Pr(F \mid A).$$

Proof.

First, notice that the axioms P1-P3 imply $\Pr(F \cap G \mid A) = \Pr(G \mid A) \Pr(F \mid A \cap G)$. Now use conditional independence to write

$$\Pr(G \mid A) \Pr(F \mid A \cap G) = \Pr(F \cap G \mid A) = \Pr(F \mid A) \Pr(G \mid A),$$
$$\Pr(G \mid A) \Pr(F \mid A \cap G) = \Pr(F \mid A) \Pr(G \mid A).$$

Definition 7 (Exchangeable)

We say a sequence of random variables $\mathbf{Y} = \{Y_1, Y_2, \ldots, Y_n\}$ *are* ***exchangeable*** *if*

$$\Pr(Y_1, Y_2, \ldots Y_n) = \Pr(Y_{\xi_1}, Y_{\xi_2}, \ldots Y_{\xi_n}),$$

for all ***permutations*** ξ *of the labels of* $\mathbf{Y}$.

Example 4 (Uma vez Flamengo... continued)

Suppose we survey 12 people and record whether they cheer for Flamengo $Y_i = 1$ or not $Y_i = 0$, $i = 1, 2, \ldots, 12$. What value shoud we assign to :

- $p_1 := \Pr(1, 0, 0, 1, 0, 1, 1, 1, 1, 1, 1, 1)$;
- $p_2 := \Pr(1, 1, 0, 1, 0, 1, 1, 1, 1, 0, 1, 1)$;
- $p_3 := \Pr(1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0)$?

If your answer is $p_1 = p_2 = p_3$ then you are saying the Y_i are (at least partially) exchangeable!

An application of conditional independence

For $\theta \in (0,1)$, consider the following sequence of probability statements:

$$\begin{aligned}
\Pr(Y_{12} = 1 \mid \theta) &= \theta, \\
\Pr(Y_{12} = 1 \mid Y_1, \ldots Y_{11}, \theta) &= \theta, \\
\Pr(Y_{11} = 1 \mid Y_1, \ldots Y_{10}, Y_{12}, \theta) &= \theta.
\end{aligned}$$

These imply that the Y_i are conditionally independent and identically distributed (iid), and in particular:

$$\begin{aligned}
\Pr(Y_1 = y_1, \ldots, Y_{12} = y_{12} \mid \theta) &= \prod_{i=1}^{12} \theta^{y_i} (1-\theta)^{1-y_i}, \\
&= \theta^S (1-\theta)^{12-S},
\end{aligned}$$

with $S := \sum_{i=1}^{12} y_i$. Also, under a uniform prior,

$$\Pr(Y_1, \ldots Y_{12}) = \int_0^1 t^S (1-t)^{12-S} \pi(t)\, dt = \frac{(S+1)!(12-S+1)!}{13!} = \binom{13}{S+1}^{-1}.$$

Relaxing exchangeability (a bit)

Sometimes total symmetry can be a burden. We can relax this slightly by introducing the concept of **partial exchangeability**:

Definition 8 (Partially exchangeable)

Let $\boldsymbol{X} = \{X_1, \ldots, X_n\}$ *and* $\boldsymbol{X} = \{Y_1, \ldots, Y_m\}$ *be two sets of random variables. We say* $\boldsymbol{X}$ *and* $\boldsymbol{Y}$ *are* ***partially*** *exchangeable if*

$$\Pr(X_1, \ldots, X_n; Y_1, \ldots, Y_m) = \Pr\left(X_{\xi_1}, \ldots, X_{\xi_n}; Y_{\sigma_1}, \ldots, Y_{\sigma_m}\right),$$

for any two permutations $\boldsymbol{\xi}$ and $\boldsymbol{\sigma}$ of $1, \ldots, n$ and $1, \ldots, m$, respectively.

Example 5 (Uma vez Flamengo...continued)

To see how exchangeability can be relaxed into partial exchangeability, consider $\boldsymbol{X}$ and $\boldsymbol{Y}$ as observations coming from populations from Rio de Janeiro and Ceará, respectively. If the covariate "state" were deemed to not matter, then we would have complete exchangeability.

A statistically useful remark

Remark 2 (Exchangeability from conditional independence)

Take $\theta \sim \pi(\theta)$, i.e., represent uncertainty about θ using a probability distribution. If $\Pr(Y_1 = y_1, \ldots, Y_n = y_n \mid \theta) = \prod_{i=1}^{n} \Pr(Y_i = y_i \mid \theta)$, then $Y_1, \ldots, Y_n$ are exchangeable.

Proof.

Sketch: Use

- Marginalisation;
- Conditional independence;
- Commutativity of products in $\mathbb{R}$;
- Definition of exchangeability.

$\square$

A fabulous theorem!

Theorem 3 (De Finetti's theorem[2])

If $\Pr(Y_1, \ldots, Y_n) = \Pr\left(Y_{\xi_1}, \ldots, Y_{\xi_n}\right)$ *for all permutations* ξ *of* $1, \ldots, n$*, then*

$$\Pr(Y_1, \ldots, Y_n) = \Pr\left(Y_{\xi_1}, \ldots, Y_{\xi_n}\right) = \int_{\Theta} \Pr(Y_1, \ldots, Y_n \mid t)\, \pi(t)\, dt, \tag{6}$$

for some choice of triplet $\{\theta, \pi(\theta), f(y_i \mid \theta)\}$*, i.e., a parameter, a prior and a sampling model.*

See Proposition 4.3 in Bernardo and Smith (2000) for a proof outline. Here we shall prove the version from De Finetti (1931).

[2] Technically, the theorem stated here is more general than the representation theorem proven by De Finetti in his seminal memoir, which concerned binary variables only.

This theorem has a few important implications, namely:

- $\pi(\theta)$ represents our beliefs about $\lim_{n\to\infty} \sum_i (Y_i \leq c)/n$ for all $c \in \mathcal{Y}$;
- $\{ Y_1, \ldots, Y_n \mid \theta \text{ are i.i.d} \} + \{ \theta \sim \pi(\theta) \} \iff \{ Y_1, \ldots, Y_n \text{ are exchangeable for all } n \}$;
- If $Y_i \in \{0, 1\}$, we can also claim that:
  - ◇ If the Y_i are assumed to be independent, then they are distributed Bernoulli conditional on a random quantity θ;
  - ◇ θ has a prior measure $\Pi \in \mathcal{P}((0, 1))$;
  - ◇ By the strong law of large numbers (SLLN), $\theta = \lim_{n\to\infty}(\frac{1}{n}\sum_{i=1}^{n} Y_i)$, so Π can be interpreted as a "belief about the limiting relative frequency of 1's".

As the exchangeability results above clearly demonstrate, being able to use conditional independence is a handy tool. More specifically, knowing on what to condition so as to make things exchangeable is key to statistical analysis.

Idea 5 (Conditioning is the soul of Statistics[3])

Knowing on what to condition can be the difference between an unsolvable problem and a trivial one. When confronted with a statistical problem, always ask yourself "What do I know for sure?" and then "How can I create a conditional structure to include this information?".

[3] This idea is due to Joe Blitzstein, who did his PhD under no other than the great Persi Diaconis.

Recommended reading

- Hoff (2009) Ch. 2 and *Schervish (2012) Ch.1;
  - *Paper: Diaconis and Freedman (1980) explains why if n samples are taken from an exchangeable population of size $N \gg n$ without replacement, then the sample $Y_1, \ldots Y_n$ can be modelled as approximately exchangeable;
- Next lecture: Robert (2007) Ch. 3.

Priors: a curse and a blessing

- Priors are the main point of contention between Bayesians and non-Bayesians;
- As we shall see, there is usually no unique way of constructing a prior measure;
- Moreover, in many situations the choice of prior is not inconsequential.
- There is always a question of when to stop adding uncertainty...

Determination of priors: existence

It is usually quite hard to determine a (unique) prior even when substantial knowledge. Why? One reason is that a prior measure is guaranteed to exist only when there is a **coherent ordering** of the Borel sigma-algebra $\mathcal{B}(\Theta)$. This entails that the following axioms hold:

(A1) Total ordering: For all measurable $A, B \in \mathcal{B}(\Theta)$ one and only one of these can hold:

$$A \prec B, B \prec A \text{ or } A \sim B.$$

(A2) Transitivity: For measurable $A_1, A_2, B_1, B_2 \in \mathcal{B}(\Theta)$ such that $A_1 \cap A_2 = \emptyset = B_1 \cap B_2$ and $A_i \preceq B_i, i = 1, 2$ then the following holds:

- ⋄ $A_1 \cup A_2 \preceq B_1 \cup B_2$;
- ⋄ If $A_1 \prec B_1$ then $A_1 \cup A_2 \prec B_1 \cup B_2$;

(A3) For any measurable A, $\emptyset \preceq A$ and also $\emptyset \prec \Theta$;

(A4) Continuity: If $E_1 \supset E_2 \ldots$ is a decreasing sequence of measurable sets and B is such that $B \preceq E_i$ for all i, then

$$B \preceq \bigcap_{i=1}^{\infty} E_i.$$

Approximation I: marginalisation

One way to approach the problem of determining a prior measure is to consider the marginal distribution of the data:

$$m(x) = \int_\Theta f(x \mid \theta)\pi(\theta)\, d\theta. \tag{7}$$

In other words we are trying to solve an inverse problem in the form of an integral equation by placing restrictions on $m(x)$ and calibrating π to satisfy them.

Approximation II: moments

Another variation on the integral-equation-inverse-problem theme is to consider expectations of measurable functions. Suppose

$$E_\pi[g_k] := \int_\Theta g_k(t)\pi(t)\,dt = w_k. \tag{8}$$

For instance, if the analyst knows that $E_\pi[\theta] = \mu$ and $\operatorname{Var}_\pi(\theta) = \sigma^2$, then this restricts the class of functions in $\mathcal{L}_1(\Theta)$ that can be considered as prior density[4]. One can also consider *order statistics* by taking $g_k(x) = \mathbb{I}_{(-\infty, a_k]}(x)$.

[4] As we shall see in the coming lectures, π needs not be in $\mathcal{L}_1(\Theta)$, i.e., needs not be **proper**. But this "method-of-moments" approach is then complicated by lack of integrability.

Maximum entropy priors

The moments-based approach is not complete in the sense that it does not lead to a unique prior measure π.

Definition 9 (Entropy)

The entropy of a probability distribution P is defined as

$$H(P) := E_p[-\log p] = -\int_{\mathcal{X}} \log p(x) \, dP(x). \tag{9}$$

When θ has finite support, we get the familiar

$$H(P) = -\sum_i p(\theta_i) \log(p(\theta_i)).$$

We can leverage this concept in order to pick π.

Definition 10 (Maximum entropy prior)

Let $\mathcal{P}_r$ be a class of probability measures on $\mathcal{B}(\Theta)$. A maximum entropy prior in $\mathcal{P}_r$ is a distribution that satisfies

$$\underset{P \in \mathcal{P}_r}{\arg\max}\, H(P).$$

Finding a maxent prior

When Θ is finite, we can write

$$\pi^*(\theta_i) = \frac{\exp\left\{\sum_{k=1} \lambda_k g_k(\theta_i)\right\}}{\sum_j \exp\left\{\sum_{k=1} \lambda_k g_k(\theta_j)\right\}},$$

where the λ_k are Lagrange multipliers. In the uncountable case things are significantly more delicate, but under regularity conditions there exists a reference measure Π_0 such that

$$\begin{aligned} H_\Pi &= E_{\pi_0}\left[\log\left(\frac{\pi(\theta)}{\pi_0(\theta)}\right)\right], \\ &= \int_\Theta \log\left(\frac{\pi(\theta)}{\pi_0(\theta)}\right) \Pi_0(d\theta). \end{aligned}$$

Exercise 2 (Maximum entropy Beta prior)

Find the maximum entropy Beta distribution under the following constraints:

- $E[\theta] = 1/2$;
- $E[\theta] = 9/10$.

Hint: If P is a Beta distribution with parameters α and β, then

$$H_P = \log B(\alpha, \beta) - (\alpha - 1)\psi(\alpha) - (\beta - 1)\psi(\beta) + (\alpha + \beta - 2)\psi(\alpha + \beta),$$

where $B(x, y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$ is the Beta function and $\psi(x) = \frac{d}{dx}\log(\Gamma(x))$ is the digamma function.

In some situations, the “right” parametric family presents itself naturally.

Example 6 (Eliciting Beta distributions)

Let $x_i \sim \text{Binomial}(n_i, p_i)$ be the number of Flamengo supporters out of n_i people surveyed. Over the years, the average of p_i has been 0.70 with variance 0.1. If we assume $p_i \sim \text{Beta}(\alpha, \beta)$ we can elicit an informative distribution based on historical data by solving the system of equations

$$E[\theta] = \frac{\alpha}{\alpha + \beta} = 0.7,$$

$$\text{Var}(\theta) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)} = 0.1.$$

Parametric approximations: difficulties

Other times we may have a hard time narrowing down the prior to a specific parametric family. Consider the following example.

Example 7 (Normal or Cauchy?)

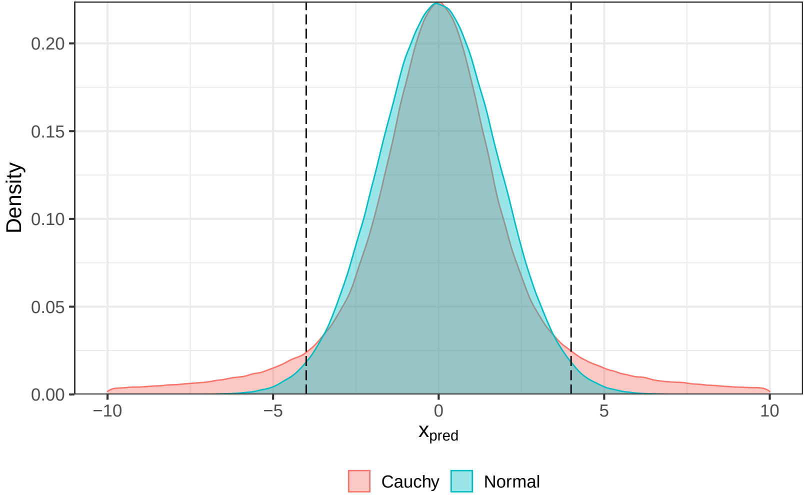
Suppose $x_i \sim \text{Normal}(\theta, 1)$ and we are informed that $\Pr(\theta \leq -1) = 1/4$, $\Pr(\theta \leq 0) = 1/2$ and $\Pr(\theta \leq 1) = 3/4$. Seems like plenty of information. It can be shown that

$$\pi_1(\theta) = \frac{1}{\sqrt{2\pi 2.19}} \exp\left(-\frac{\theta^2}{2 \times 2.19}\right) \text{ (Normal)},$$
$$\pi_2(\theta) = \frac{1}{\pi(1 + x^2)} \text{ (Cauchy)},$$

both satisfy the requirements. Unfortunately, under quadratic loss we get $\delta_1(4) = 2.75$ and $\delta_2(4) = 3.76$ and differences are exacerbated for $|x| \geq 4$.

Why, though?

Remember the marginal approach? It is illuminating in this case. Heres $m(x)$:



Prior predictive distributions of x under Normal and Cauchy priors.

Conjugacy is a central concept in Bayesian statistics. It provides a functional view of the prior-posterior mechanic that emphasises tractability over coherence.

Definition 11 (Conjugate)

A family $\mathcal{F}$ of distributions on Θ is called ***conjugate*** *or closed under sampling for a likelihood $f(x \mid \theta)$ if, for every $\pi \in \mathcal{F}$, $p(\theta \mid x) \in \mathcal{F}$.*

Arguments for using conjugate priors

- "Form-preservation": in a limited-information setting it makes sense that $p(\theta \mid x)$ and $\pi(\theta)$ lie on the same family, since the information in x might not be enough to change the structure of the model, just its parameters;
- Simplicity: when you do not know a whole lot, it makes sense to KISS[5];
- Sequential learning: since $\mathcal{F}$ is closed under sampling, one can update a sequence of posteriors $p_i(\theta \mid x_1, \ldots, x_i)$ as data comes in.

[5] Keep it simple, stupid!

Exponential families

The exponential family of distributions is a cornerstone of statistical practice, underlying many often-used models. Here are a few useful definitions.

Definition 12 ((Natural) Exponential family)

Let μ be a σ-finite measure on $\mathcal{X}$ and let Θ be a non-empty set serving as the parameter space. Let $C : \Theta \to (0, \infty)$ and $h : \Theta \to (0, \infty)$ and let $R : \Theta \times \mathcal{X} \to \mathbb{R}^k$ and $T : \Theta \times \mathcal{X} \to \mathbb{R}^k$. The family of distributions with density

$$f(x \mid \theta) = C(\theta)h(x) \exp\left(R(\theta) \cdot T(x)\right)$$

w.r.t. μ is called an ***exponential family***. *Moreover, if $R(\theta) = \theta$, the family is said to be* ***natural***.

Definition 13 (Regular exponential family)

We say a natural exponential family $f(x \mid \theta)$ is ***regular*** *if the natural parameter space*

$$N := \left\{\theta : \int_{\mathcal{X}} \exp(\theta \cdot x)h(x)\, d\mu(x) < \infty\right\}, \tag{10}$$

is an open set of the same dimension as the closure of the convex hull of $\operatorname{supp}(\mu)$.

There is an intimate link between sufficiency (i.e. the existence of sufficient statistics) and conjugacy. The following is a staple of Bayesian theory.

Theorem 4 (Pitman-Koopman-Darmois)

If a family of distributions $f(\cdot \mid \theta)$ whose support does not depend on θ is such that, for a sample size large enough, there exists a sufficient statistic of fixed dimension, then $f(\cdot \mid \theta)$ is an exponential family.

The support condition is not a complete deal breaker, however:

Remark 3 (Quasi-exponential)

The $\mathsf{Uniform}(-\theta, \theta)$ *and* $\mathsf{Pareto}(\theta, \alpha)$ *families are called quasi-exponential due to the fact that there do exist sufficient statistics of fixed dimension for these families, even though their supports depend on θ.*

Conjugacy in the exponential family

I hope you are convinced of the utility of the exponential family by now. It would be nice to have an automated way to deduce a conjugate prior for $f(x \mid \theta)$ when it is in the exponential family. This is exactly what the next result gives us.

Remark 4 (Conjugate prior for the exponential family)

A conjugate family for $f(x \mid \theta)$ is given by

$$\pi(\theta \mid \mu, \lambda) = K(\mu, \lambda) \exp\left(\theta \cdot \mu - \lambda g(\theta)\right), \tag{11}$$

such that the posterior is given by $p(\theta \mid \mu + x, \lambda + 1)$.

Please do note that (11) is only a valid density when $\lambda > 0$ and μ/λ belongs to the interior of the natural space parameter. Then, it is a σ-finite measure. See Diaconis and Ylvisaker (1979) for more details.

Conjugacy: common families

Table 3.3.1. *Natural conjugate priors for some common exponential families*

$f(x\|\theta)$	$\pi(\theta)$	$\pi(\theta\|x)$
Normal $\mathcal{N}(\theta, \sigma^2)$	Normal $\mathcal{N}(\mu, \tau^2)$	$\mathcal{N}(\varrho(\sigma^2\mu + \tau^2 x), \varrho\sigma^2\tau^2)$ $\varrho^{-1} = \sigma^2 + \tau^2$
Poisson $\mathcal{P}(\theta)$	Gamma $\mathcal{G}(\alpha, \beta)$	$\mathcal{G}(\alpha + x, \beta + 1)$
Gamma $\mathcal{G}(\nu, \theta)$	Gamma $\mathcal{G}(\alpha, \beta)$	$\mathcal{G}(\alpha + \nu, \beta + x)$
Binomial $\mathcal{B}(n, \theta)$	Beta $\mathcal{B}e(\alpha, \beta)$	$\mathcal{B}e(\alpha + x, \beta + n - x)$
Negative Binomial $\mathcal{N}eg(m, \theta)$	Beta $\mathcal{B}e(\alpha, \beta)$	$\mathcal{B}e(\alpha + m, \beta + x)$
Multinomial $\mathcal{M}_k(\theta_1, \ldots, \theta_k)$	Dirichlet $\mathcal{D}(\alpha_1, \ldots, \alpha_k)$	$\mathcal{D}(\alpha_1 + x_1, \ldots, \alpha_k + x_k)$
Normal $\mathcal{N}(\mu, 1/\theta)$	Gamma $\mathcal{G}a(\alpha, \beta)$	$\mathcal{G}(\alpha + 0.5, \beta + (\mu - x)^2/2)$

Taken from Robert (2007), page 121.

Conjugacy: drawbacks

Conjugate modelling is certainly useful, but has its fair share of pitfalls.

Arguments against using conjugate priors

- Conjugate priors are restrictive *a priori*: in many settings, specially in high dimensions, the set of conjugate priors that retain tractability is so limited so as to not be able to encode all prior information available;
- Conjugate priors are not truly subjective: they limit the analyst's input to picking values for the hyperparameters;
- Conjugate priors are restrictive *a posteriori*: you are stuck with a given structure forever, no matter how much data you run into.

The principle of insufficient reason

Also called principle of indifference by Keynes[6].

> *...if there is no known reason for predicating of our subject one rather than another of several alternatives, then relatively to such knowledge the assertions of each of these alternatives have an equal probability." (Keynes, 1921, Ch4 pg. 52-53).*

The idea dates back to Laplace and even Bayes himself and usually leads to

$$\pi(\theta) \propto 1.$$

[6] John Maynard Keynes (1883–1946) was an English economist.

In many applications we might want some sort of *invariance* in our prior model.

Definition 14 (Invariant model)

A statistical model is said to be ***invariant*** *(or closed) under the action of a group* $\mathcal{G}$ *if* $\forall g \in \mathcal{G}\ \exists \theta^\star \in \mathbf{\Theta}$ *such that* $y = g(x)$ *is distributed with density* $f(y \mid \theta^\star)$*, denoting* $\theta^\star = \bar{g}(\theta)$*.*

Consider two types of invariance

- *Translation* invariance: A model $f(x-\theta)$ such that $x - x_0$ has a distribution in the same family for every x_0 leads to

$$\pi(\theta) = \pi(\theta - \theta_0),\ \forall\, \theta_0 \in \mathbf{\Theta}.$$

- *Scale* invariance: Similarly, a model of the form $\sigma^{-1}f(x/\sigma)$, $\sigma > 0$ is *scale-invariant* and leads to

$$\pi(A/c) = \pi(A),$$

for any measurable A.

Jeffreys's prior

One can try to build a prior that captures only the essential structural information about the problem by deriving an invariant distribution from the Fisher information:

$$I(\theta) = E\left[\left(\frac{\partial \log f(X \mid \theta)}{\partial \theta}\right)^2\right].$$

Under regularity conditions, we can usually also write

$$I(\theta) = -E\left[\frac{\partial^2 \log f(X \mid \theta)}{\partial \theta^2}\right].$$

Jeffreys showed that

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

is invariant. There are straightforward generalisations when θ is multidimensional.

Jeffreys's priors: examples

A good exercise is to show that

- If $x \sim \text{Normal}(0, \theta)$, $\pi_J(\theta) \propto 1/\theta^2$;
- If $x \sim \text{Normal}_d(\theta, \boldsymbol{I}_d)$, $\pi_J(\theta) \propto 1$;
- If $x \sim \text{Binomial}(n, \theta)$, $\pi_J(\theta) \equiv \text{Beta}(1/2, 1/2)$;
- If $f(x \mid \theta) = h(x) \exp\left(\theta \cdot x - \psi(\theta)\right)$, then

$$\pi_J(\theta) \propto \sqrt{\prod_{i=1}^{k} \psi''(\theta)}.$$

Beware!

One important caveat of Jeffreys's priors is that they violate the Likelihood Principle. To see why, consider the following exercise.

Exercise 3 (Poisson process)

Suppose one is interested in estimating the rate, θ, of a Poisson process:

$$Y(t) \sim \text{Poisson}(t\theta).$$

There are two possible experimental designs:

a) *Fix a number n of events to be observed and record the time X to observe them, or;*
b) *Wait a fixed amount of time, t, and count the number Y of occurences of the event of interest. Show that*

$$a)\ I_X(\theta) = \frac{n}{\theta^2},$$
$$b)\ I_Y(\theta) = \frac{t}{\theta}.$$

Which conclusions can we draw from this example?

See also example 3.5.7 in Robert (2007).

Jeffreys's approach can sometimes lead to marginalisation paradoxes and calibration issues (see exercise 4.47 in Robert (2007)). Bernardo (1979) proposes a modification that avoids these difficulties by explicit separating parameters in *nuisance* and *interest*. It works like this: take $f(x \mid \theta)$, with $\theta = (\theta_1, \theta_2)$ and let θ_1 be the parameter of interest. We must first compute[7]

$$\tilde{f}(x \mid \theta_1) = \int_{\boldsymbol{\Theta}_2} f(x \mid \theta_1, t_2)\pi(t_2 \mid \theta_1)\, dt_2,$$

and then compute the Jeffreys's prior associated with this marginalised likelihood. Notice that this entails first deriving $\pi(\theta_2 \mid \theta_1)$.

[7] Notice that this need not be well-defined. One common way of dealing with difficulties is to integrate on a sequence of measurable compact sets and take the limit.

Reference priors: example

Suppose we have $x_{ij} \sim \text{Normal}(\mu_i, \sigma^2)$, $i = 1, \ldots, n$, $j = 1, 2$ and consider making inferences about $\boldsymbol{\theta} = (\sigma^2, \boldsymbol{\mu})$. Here $\theta_1 = \sigma$ is a nuisance parameter and we're interested in the location $\boldsymbol{\theta}_2 = \boldsymbol{\mu}$. The Jeffreys's prior is

$$\pi_J(\boldsymbol{\theta}) \propto 1/\sigma^{n+1},$$

leading to a Bayes estimator under quadratic loss:

$$\hat{\sigma}_J := E[\sigma^2 \mid \boldsymbol{x}] = \frac{\sum_{i=1}^{n}(x_{i1} - x_{i2})^2}{4n - 4},$$

which in not consistent. The reference approach give $\pi(\theta_1 \mid \boldsymbol{\theta_2})$ as a flat prior - because $\boldsymbol{\theta_2}$ is a location parameter. Marginalising the likelihood against this flat density over $(0, \infty)$ gives $\pi_R(\sigma^2) \propto 1/\sigma^2$, leading to

$$\hat{\sigma}_R = \frac{\sum_{i=1}^{n}(x_{i1} - x_{i2})^2}{2n - 4},$$

which is consistent. Phew!

If you are a bit greedy and want to please Greeks and Troyans, you might also try to construct your prior so that it attains good frequency properties. One such way is to construct **matching priors**:

Definition 15 (Matching prior)

*We say $\pi(\theta)$ is a **matching prior** for a confidence level α if it is constructed in such a way that*

$$\Pr(g(\theta) \in C_x \mid x) = \frac{1}{m(x)} \int_{C_x} f(x \mid t)\pi(t)\, dt = 1 - \alpha,$$

holds for a given confidence set $C_x(\alpha)$ for $g(\theta)$.

In other words, if the posterior matches the confidence set. It can be shown that, in unidimensional families, the Jeffreys's prior gives

$$\Pr(\theta \leq k_\alpha(x)) = 1 - \alpha + O(n^{-1}),$$

where $C_x = (-\infty, k_\alpha(x))$ is a one-sided confidence interval.

Prior classes

Robert (2007) gives a classification of priors in classes:

i) Conjugate classes:

$$\Gamma_C = \{\pi \in \mathcal{F} : p \in \mathcal{F}\},$$

ii) Determined moment(s) classes:

$$\Gamma_M = \{\pi : a_i \leq E_\pi[\theta] \leq b_i, i = 1, \ldots, k\},$$

iii) Neighbourhood (or ϵ-contamination) classes:

$$\Gamma_{\epsilon,\mathcal{Q}} = \{\pi = (1-\epsilon)\pi_0 + q, q \in \mathcal{Q}\},$$

iv) Underspecified classes:

$$\Gamma_U = \{\pi : \int_{I_i} \pi(t)\, dt \leq \mu_i, i = 1, \ldots, k\},$$

v) Ratio of density classes:

$$\Gamma_R = \{\pi : L(\theta) \leq \pi(\theta) \leq U(\theta)\}.$$

Prior sensitivity analysis

General recommendations about building priors:

- Check the **observable consequences** of your priors :what kinds of data does this produce?
- Check the inferential consequences of your priors: how do my estimators change under different priors?
- Make sure you know what your restrictions do to the tail of your prior;
- It is usually a good idea to understand what the prior **does** to the model, as opposed to only which values θ can plausibly take;
- Sometimes it may be useful to think of priors as *penalisations* that **regularise** inference.

Recommended reading

- Robert (2007) Ch. 3;
- Next lecture: Robert (2007) Ch. 3.6, Seaman III et al. (2012), Gelman et al. (2017) and Simpson et al. (2017).

The maximum *a posteriori* (MAP) estimator

Definition 16 (Maximum *a posteriori*)

The posterior mode or maximum a posteriori (MAP) estimator of a parameter θ *is given by*

$$\delta_{\pi}^{MAP}(x) := \arg\max_{\theta \in \Theta} p(\theta \mid x). \tag{12}$$

Example 8 (MAP for the binomial case)

Suppose $x \sim \text{Binomial}(n, p)$. Now consider the following three priors for p:

- $\pi_0(p) = \frac{\sqrt{p(1-p)}}{B(1/2,1/2)}$ [Jeffreys];
- $\pi_1(p) = 1$ [Beta(1,1)/Uniform];
- $\pi_2(p) = (p(1-p))^{-1}$ [Haldane (1932)].

These lead to

- $\delta_0^{\text{MAP}}(x) = \max\{(x - 1/2)/(n-1), 0\}$;
- $\delta_1^{\text{MAP}}(x) = x/n$;
- $\delta_2^{\text{MAP}}(x) = \max\{(x-1)/(n-2), 0\}$.

We end this discussion with the following warning:

Idea 6 (Marginalise, not maximise)

Bayesian approaches to estimation and prediction usually focus on marginalisation rather than optmisation. This is because, following the Likelihood Principle, all of the information available about the unknowns is contained in the posterior distribution, and thus all inferences must be made using this probability measure, usually by finding suitable expectations of functionals of interest.

In particular, for higher dimensions, **concentration of measure**[8] ensures that the posterior mode has less and less relevance as a summary, at least so far as the barycentre of the distribution is concerned.

[8] See these excellent notes by Terence Tao:https://terrytao.wordpress.com/2010/01/03/254a-notes-1-concentration-of-measure/ .

A central quantity in the evaluation of Bayesian estimators is

$$E_p\left[(\delta_\pi - h(\theta))^2\right] = E_\pi\left[(\delta_\pi - h(\theta))^2 \mid x\right] \tag{13}$$

for measurable h.

Example 9 (Bayes versus frequentist risk)

Take $x \sim \text{Binomial}(n, \theta)$ with n known and place a Jeffreys's prior on θ. Consider the MLE: $\delta_1(x) = x/n$. It can be shown that:

$$E_\pi\left[(\delta_1 - \theta)^2 \mid x\right] = \left(\frac{x - n/2}{n(n+1)}\right)^2 + \frac{(x+1/2)(n-x+1/2)}{(n+1)^2(n+2)}.$$

Moreover,

$$\max_{\theta \in (0,1)} E_\pi\left[(\delta_1 - \theta)^2 \mid x\right] = [4(n+2)]^{-1},$$

and

$$\max_{\theta \in (0,1)} E_\theta\left[(\delta_1 - \theta)^2\right] = [4n]^{-1}.$$

A brief aside about prediction

Prediction is an important inferential task and is somewhat related to the previous discussion on precision. Consider predicting a quantity z **conditional** on data x. For that we need $g(z \mid x, \theta)$, $f(x \mid \theta)$ and $\pi(\theta)$. Then,

$$g_\pi(z \mid x) = \int_\Theta g(z \mid x, t) p(t \mid x)\, dt \tag{14}$$

encodes all of the information brought by the posterior about z. A special case is i.i.d prediction:

$$g(\tilde{x} \mid x) = \int_\Theta f(\tilde{x} \mid t) p(t \mid x)\, dt \tag{15}$$

is the posterior predictive of the new data $\tilde{x}$.

Idea 7 (Calibrated priors for prediction)

The prior, π, can be constructed so as to minimise error in a prediction task.

A neat trick

Computing expectations all the time means we have to become familiar with a few tricks to facilitate obtaining approximate answers.

Example 10 (Mixture representation of the Student-t)

Take $x \sim \text{Normal}_p(\theta, \boldsymbol{I}_p)$ and put $\theta \sim \text{Student-t}_p(\alpha, 0, \tau^2\boldsymbol{I}_p)$. Then $p(\theta \mid x)$ does not have a closed-form normalising constant and computing the Bayes estimator under quadratic loss is a chore. However, we can use the representation

$$\begin{aligned} \theta \mid z &\sim \text{Normal}_p(0, \tau^2 z\boldsymbol{I}_p), \\ z &\sim \text{InverseGamma}(\alpha/2, \alpha/2), \end{aligned}$$

to get

$$\theta \mid x, z \sim \text{Normal}_p\left(\frac{x}{1+\tau^2 z}, \frac{\tau^2 z}{1+\tau^2 z}\boldsymbol{I}_p\right)$$

Thus, the Bayes estimator $\delta_\pi(x) = \int_0^\infty E_\pi[\theta \mid x, z]p(z \mid x)\, dz$ can be computed with a single integral for any dimension p.

Conjugacy is handy!

Table 4.2.1. *The Bayes estimators of the parameter θ under quadratic loss for conjugate distributions in the usual exponential families.*

Distribution	Conjugate prior	Posterior mean
Normal $\mathcal{N}(\theta, \sigma^2)$	Normal $\mathcal{N}(\mu, \tau^2)$	$\dfrac{\mu\sigma^2 + \tau^2 x}{\sigma^2 + \tau^2}$
Poisson $\mathcal{P}(\theta)$	Gamma $\mathcal{G}(\alpha, \beta)$	$\dfrac{\alpha + x}{\beta + 1}$
Gamma $\mathcal{G}(\nu, \theta)$	Gamma $\mathcal{G}(\alpha, \beta)$	$\dfrac{\alpha + \nu}{\beta + x}$
Binomial $\mathcal{B}(n, \theta)$	Beta $\mathcal{B}e(\alpha, \beta)$	$\dfrac{\alpha + x}{\alpha + \beta + n}$
Negative binomial $\mathcal{N}eg(n, \theta)$	Beta $\mathcal{B}e(\alpha, \beta)$	$\dfrac{\alpha + n}{\alpha + \beta + x + n}$
Multinomial $\mathcal{M}_k(n; \theta_1, \ldots, \theta_k)$	Dirichlet $\mathcal{D}(\alpha_1, \ldots, \alpha_k)$	$\dfrac{\alpha_i + x_i}{\left(\sum_j \alpha_j\right) + n}$
Normal $\mathcal{N}(\mu, 1/\theta)$	Gamma $\mathcal{G}(\alpha/2, \beta/2)$	$\dfrac{\alpha + 1}{\beta + (\mu - x)^2}$

[9] Taken from Robert (2007).

A worked example

We will stretch our Bayesian muscles with the next problem.

Exercise 4 (Inference for the rate of a Gamma)

Let $x \sim \text{Gamma}(\nu, \theta)$ *with* $\nu > 0$ *known. A natural choice of prior is* $\theta \sim \text{Gamma}(\alpha, \beta)$*. Find the Bayes estimator under*

$$L_1(\delta, \theta) = \left(\delta - \frac{1}{\theta}\right)^2,$$

and the scale-invariant loss

$$L_2(\delta, \theta) = \theta^2 \left(\delta - \frac{1}{\theta}\right)^2$$

Hint: If $X \sim \text{Gamma}(\alpha, \beta)$, $Y = 1/X \sim \text{InverseGamma}(\alpha, \beta)$ *and* $E[Y^k] = \frac{\beta^k}{(\alpha-1)\cdots(\alpha-k)}$

A quick note on quadratic loss

Exercise 4 is a special case of the general situation where

$$L(\delta, \theta) = w(\theta)||\delta - \theta||_{\boldsymbol{G}}^2,$$

for $\boldsymbol{G}$ a $p \times p$ non-negative symmetric matrix. In this case, we get

$$\delta_{\pi} = \frac{E_p[w(\theta)\theta]}{E_p[w(\theta)]}.$$

Please **note** that there is no universal justification for quadratic loss other than (sometimes leading to increased) mathematical tractability

Since the loss function, $L(\delta(x), \theta)$ is usually measurable w.r.t the posterior, it can be estimated much the same way as other functionals. In particular, if you are feeling particularly eclectic, you can always constructed π such that

$$E\left[E_p[L(\delta_\pi(x), \theta]\right] \geq R(\delta_\pi(x), \theta), \theta \in \boldsymbol{\Theta},$$

i.e. that the estimated loss never underestimates the error resulting from the use of δ_π, at least in the long run. This is called **frequentist validity**.

The following problem is described by Jeffreys as originating with Jerzy Neyman[10].

Exercise 5 (The tramcar problem)

A person travelling in a a foreign country has to change trains at a junction, and goes into the town, the existence of which they have only just heard. They have no idea of its size. The first thing they see is a tramcar numbered 100. *Assuming tramcars are numbered consecutively from* 1 *onwards, what could one infer about the number* N *of tramcars in this town?*

[10] Jerzy Neyman (1894-1981) was a Polish-American statistician, known for with work with Egon Pearson (1895-1980) on the foundations of the null hypothesis significance testing (NHST) framework.

Recommended reading

- Robert (2007), Ch4.
- Next lecture: Robert (2007) Ch. 5.

The duality between estimation and testing

Similarly to the frequentist case, in Bayesian inference there is an intimate relationship between testing hypotheses an estimating measurable functions of the parameters.

Definition 17 (Test)

Consider a statistical model $f(x \mid \theta)$ with $\theta \in \Theta$. Given $\Theta_0 \subset \Theta$, a test consists in answering the question of whether

$$H_0 : \theta \in \Theta_0$$

is true. We call H_0 the null hypothesis and Θ_0 can often be a point, i.e. $\Theta_0 = \{\theta_0\}$.

Notice that $\mathbb{I}_{\Theta_0}(\theta)$ is measurable and thus we can define, for instance

$$L_1(\theta, \varphi) = \begin{cases} 1, \varphi = \mathbb{I}_{\Theta_0}(\theta), \\ 0, \text{ otherwise}, \end{cases}$$

which in turn leads to

$$\varphi_1 = \begin{cases} 1, \Pr(\theta \in \Theta_0 \mid x) > \Pr(\theta \in \Theta_0^c \mid x), \\ 0, \text{ otherwise}. \end{cases}$$

A refinement

The loss function just seen can be refined to

$$L_2(\theta, \varphi) = \begin{cases} 0, \varphi = \mathbb{I}_{\Theta_0}(\theta), \\ a_0, \theta \in \Theta_0, \varphi = 0 \\ a_1, \theta \in \Theta_0^c, \varphi = 1. \end{cases}$$

Under this loss, we have

$$\varphi_2 = \begin{cases} 1, \Pr(\theta \in \Theta_0 \mid x) > a_1/(a_0 + a_1), \\ 0, \text{ otherwise.} \end{cases}$$

Example

Example 11 (*One* Normal test)

Take, for example, $x \sim \text{Normal}(\theta, \sigma^2)$, with $\theta \sim \text{Normal}(\mu_0, \tau^2)$. This implies $\theta \mid x \sim \text{Normal}(\mu(x), \omega^2)$, where

$$\mu(x) = \frac{\sigma^2\mu_0 + \tau^2 x}{\sigma^2 + \tau^2}; \omega^2 = \frac{\sigma^2\tau^2}{\sigma^2 + \tau^2}.$$

To test $H_0 : \theta < 0$, we can compute

$$\begin{aligned}\Pr(\theta < 0 \mid x) &= \Pr\left(\frac{\theta - \mu(x)}{\omega} < \frac{\mu(x)}{\omega}\right), \\ &= \Phi\left(\frac{-\mu(x)}{\omega}\right).\end{aligned}$$

This means that if z_{a_0,a_1} is such that $\Phi(z_{a_0,a_1}) = a_1/(a_0 + a_1)$, we can accept H_0 if

$$\mu(x) < -z_{a_0,a_1}\omega.$$

A central tool in Bayesian testing is the **Bayes factor** – see Kass and Raftery (1995) for a review and guide for interpretation.

Definition 18 (Bayes factor)

The Bayes factor is the ratio of posterior odds and the prior odds over the null and the alternative:

$$\begin{aligned} B_{01}^{\pi}(x) &= \frac{\Pr(\theta \in \Theta_0 \mid x)}{\Pr(\theta \in \Theta_1 \mid x)} \bigg/ \frac{\Pr(\theta \in \Theta_0)}{\Pr(\theta \in \Theta_1)}, \\ &= \frac{\Pr(\theta \in \Theta_0 \mid x) \cdot \Pr(\theta \in \Theta_1)}{\Pr(\theta \in \Theta_1 \mid x) \cdot \Pr(\theta \in \Theta_0)}. \end{aligned}$$

Remark 5

When $\Theta_0 = \{\theta_0\}$ and $\Theta_1 = \{\theta_1\}$ the Bayes factor simplifies to

$$r_{01}(x) = \frac{f(x \mid \theta_0)}{f(x \mid \theta_1)},$$

also known as the ***likelihood ratio****.*

A few more considerations on the Bayes factor

The Bayes factor can also be written as

$$B_{01}^{\pi}(x) = \frac{\int_{\Theta_0} f(x \mid t)\pi_0(t)\,dt}{\int_{\Theta_1} f(x \mid t)\pi_1(t)\,dt} = \frac{m_0(x)}{m_1(x)},$$

where π_0 and π_1 are the prior distributions under each hypothesis. Also, if $\hat{\theta}_0$ and $\hat{\theta}_1$ are the MLE under each hypothesis, by making π_0 and π_1 Dirac masses at $\hat{\theta}_0$ and $\hat{\theta}_1$, respectively, we recover

$$R(x) = \frac{\sup_{\theta \in \Theta_0} f(x \mid \theta)}{\sup_{\theta \in \Theta_1} f(x \mid \theta)} \tag{16}$$

Exercise 6 (Bayesian justifcation of LRT)

Does (16) offer a Bayesian justifcation for likelihood ratios?

Hypotheses of the form $H_i : \theta \in \{\theta_i\}$, called point-null hypotheses, are hard to deal with from a probabilistic point of view.

Remark 6 (Point-null hypotheses under continuous priors)

Point-null cannot be tested under continuous prior distributions. More generally, if either H_0 or H_1 are ***impossible*** *a priori, then no amount of data can change that belief.*

Idea 8 (Cromwell's law[11])

In general, one not assign probability zero to events that are not logically or physically demonstrably impossible. Or, more eloquently, as Oliver Cromwell writes to the General Assembly of the Church of Scotland on 3 August 1650:

> *I beseech you, in the bowels of Christ, think it possible that you may be mistaken.*

[11] This idea is attributed to British statistician Dennis Lindley (1923-2013), one of the founders of modern Bayesian theory.

Point-null hypotheses: modification of the prior

Testing point-null hypotheses involves a **modification of the prior** If $H_0 : \theta \in \{\theta_0\}$ we can write $\rho_0 = \Pr(\theta = \theta_0)$ and then

$$\tilde{\pi}(\theta) = \rho_0 \mathbb{I}_{\Theta_0}(\theta) + (1 - \rho_0)\pi_1(\theta),$$

is our new prior, where π_1 is the distribution with density $g_1(\theta) \propto \pi(\theta)\mathbb{I}_{\Theta_1}(\theta)$ with respect to the dominating measure on Θ_1. This gives a posterior probability

$$\tilde{\pi}(\Theta_0 \mid x) = \frac{f(x \mid \theta_0)\rho_0}{f(x \mid \theta_0)\rho_0 + (1 - \rho_0)m_1(x)}.$$

where $m_1(x) = \int_{\Theta_1} f(x \mid t) g_1(t)\, dt$. It can be shown that

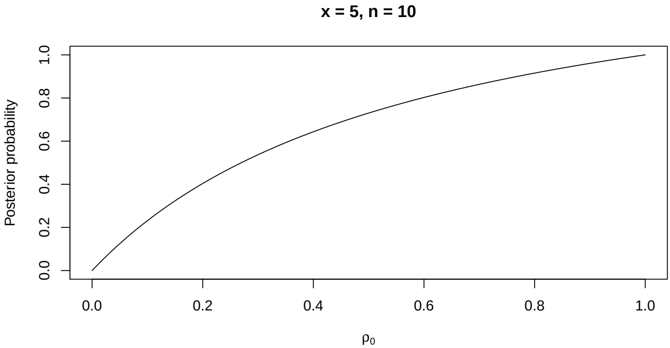
$$\tilde{\pi}(\Theta_0 \mid x) = \left[1 + \frac{1 - \rho_0}{\rho_0} \frac{1}{B_{01}^{\pi}(x)}\right]^{-1},$$

which makes clear the relationship between posterior probabilities and Bayes factors.

Example

Consider $x \sim \text{Binomial}(n, p)$ and consider testing $H_0 : p = 1/2$ against $H_1 : p \neq 1/2$. Taking $g_1(p) = 1$, we have

$$\tilde{\pi}(\Theta_0 \mid x) = \left[1 + \frac{1 - \rho_0}{\rho_0} 2^n B(x + 1, n - x + 1)\right]^{-1}.$$



Testing with improper priors

Idea 9 (Bayesian hypothesis testing with improper priors)

No. Just... No.

See DeGroot (1973) for the many reasons why this is just a bad idea. If you insist, please see Section 5.2.5 in Robert (2007) and references therein.

An interesting little paradox

Idea 10 (The Jeffreys-Lindley paradox)

Consider $x \sim \text{Normal}(\theta, \sigma^2)$ *with* σ^2 *known and suppose we are interested in testing* $H_0 : \theta = \theta_0$ *against* $H_1 : \theta \neq \theta_0$. *We can summarise the data using the sample mean* $\bar{x}$ *and then compute* $t_n = \sqrt{n}(\bar{x} - \theta_0)/\sigma$. *Employing a conjugate prior* $\theta \sim \text{Normal}(\mu_0, \sigma^2)$, *the Bayes factor is*

$$B_{01}(\boldsymbol{x}) = \sqrt{1+n}\exp\left(-\frac{nt_n^2}{2(1+n)}\right),$$

which goes to infinity with n, while the p-value:

$$p(t_n) = 1 - 2\Phi(|t_n|),$$

is constant in n. In practice this means that, for instance $t_n = 1.96$ *and* $n = 16,818$, *we have 95% frequentist confidence that* $\theta \neq \theta_0$ *whilst* ***at the same time*** *having 95% belief that* $\theta = \theta_0$.

Another look at principled Bayesian testing

Before we were doing

$$L_3(\theta, \varphi) = |\varphi - \mathbb{I}_{\Theta_0}(\theta)|.$$

But considering a strictly convex loss such as the quadratic loss

$$L_4(\theta, \varphi) = \left(\varphi - \mathbb{I}_{\Theta_0}(\theta)\right)^2,$$

leads to better (more adaptable) estimators in general. For instance, the Bayes estimator under L_4 is

$$\varphi_\pi(x) = \Pr(\theta \in \Theta_0 \mid x).$$

After all of this work, we are finally ready to define credibility regions, the main object in Bayesian interval estimation.

Definition 19 (Credibility region)

For a prior π, a set C_x is called an α-credible set if

$$\Pr(\theta \in C_x \mid x) \geq 1 - \alpha.$$

We call C_x a highest posterior density (HPD) α-credible region if

$$\{\theta : p(\theta \mid x) > k_\alpha\} \subset C_x \subset \{\theta : p(\theta \mid x) \geq k_\alpha\},$$

subject to the restriction that

$$\Pr(\theta \in C_x^\alpha) \geq 1 - \alpha.$$

A couple remarks

Credibility regions have a few desirable properties that make them quite attractive as "interval" estimates.

Remark 7 (No randomisation)

One nice feature of credibility regions for discrete distributions is that, contrary to the frequentist approach, no randomisation is needed to attain a certain level α.

Also,

Remark 8 (Improper priors and credibility regions)

In principle, the use of improper priors poses no problem for the derivation of credibility regions.

Credibility regions: Example I

Sometimes we will be able to provide Bayesian justifcation for frequentist confidence regions/intervals.

Example 12 (Credibility intervals for the variance in the Normal)

Consider $\boldsymbol{x} = \{x_1, \ldots, x_n\}$, $x_i \sim \text{Normal}(\theta, \sigma^2)$, with both parameters unknown. Consider

$$\pi(\theta, \sigma^2) \propto \frac{1}{\sigma^2}.$$

Make $s^2 = \sum_{i=1}^{n} (x - \bar{x})^2$. It can be shown that $p(\sigma^2 \mid s^2) \equiv \text{Gamma}(\sigma^2; (n-1)/2, s^2/2)$. In particular, this implies

$$\frac{s^2}{\sigma^2} \mid \bar{x} \sim \text{Chi-square}(n-1),$$

which the attentive student will notice leads to the same solution as the classical confidence approach.

Example 13 (HPD for the normal mean)

Consider again the setting of example 12. Define $\bar{s}^2 = s^2/(n-1)$ and take $t = F^{-1}_{\text{Student}}(\alpha; n-1)$. The classical "T" interval,

$$C_t(\bar{x}, \bar{s}^2) = \left(\bar{x} - t\sqrt{\frac{\bar{s}^2}{n}}, \bar{x} + t\sqrt{\frac{\bar{s}^2}{n}}\right),$$

is a HPD region under the Jeffreys's prior. Again, we can show that

$$\sqrt{n}\frac{\theta - \bar{x}}{\sqrt{\bar{s}^2}} \mid \bar{x}, \sqrt{\bar{s}^2} \sim \text{Student-t}(n-1).$$

A little decision theory can't hurt... Or can it?

Consider the loss

$$L_1(C, \theta) = \text{vol}(C) + (1 - \mathbb{I}_C(\theta))a,$$

which leads to the risk

$$R(C_x, \theta) = E[\text{vol}(C_x)] + \Pr(\theta \notin C_x).$$

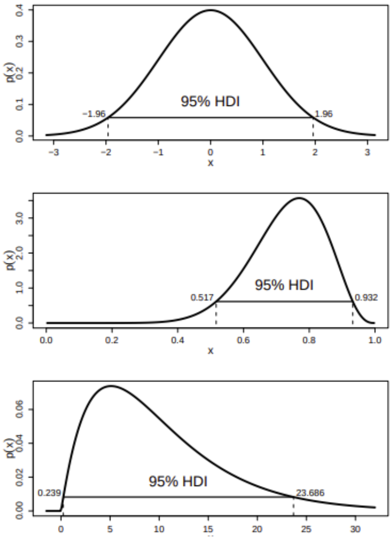
Under this loss, the interval in Example 13 is dominated by

$$C_t'(\bar{x}, \bar{s}^2) = \begin{cases} C_t(\bar{x}, \bar{s}^2), \sqrt{\bar{s}^2} < \sqrt{n}c/(2t), \\ \{\bar{x}\}, \text{ otherwise,} \end{cases}$$

which is a bit weird – why?
Now, consider what happens under a *rational loss*

$$L_k(C, \theta) = \frac{\text{vol}(C)}{\text{vol}(C) + k} + (1 - \mathbb{I}_C(\theta)), k > 0.$$

HPD (or HDI in one dimension)



Recommended reading

- Robert (2007), Ch. 5.
- Next lecture: Robert (2007) Ch. 7.

Bayesian model selection: testing all over again

Model choice (or selection) is a **major** topic within any school of inference: it is how scientists make decisions about competing theories/hypotheses in light of data. One can associate a set of models $\boldsymbol{\mathcal{M}} = \{\mathcal{M}_1, \ldots \mathcal{M}_n\}$ with a set of indices I such that $\mu \in I$ we want to estimate the posterior distribution of the indicator function $\mathbb{I}_{\boldsymbol{\Theta}_\mu}(\theta)$.
Recall that estimating indicator functions over $\boldsymbol{\Theta}$ was the fundamental mechanic of Bayesian testing. In the setting of Bayesian model selection (BMS), we have something of the form

$$\mathcal{M}_i : x \sim f_i(x \mid \theta_i), \theta_i \in \boldsymbol{\Theta}_i, i \in I.$$

M-completeness

A key step in model selection is to identify in which regime the analyst finds themselves in.

Definition 20 (M-open, M-closed, M-complete)

Model selection can be categorised in three settings:

- ***M-closed**: a situation where the true data-generating model is one of $\mathcal{M}_i \in \boldsymbol{\mathcal{M}}$, even though it is most often unknown to the analyst;*
- ***M-complete**: a situation where the true model exists and is out of the model set $\boldsymbol{\mathcal{M}}$. We nevertheless want to select one of the models in the set due to computational or mathematical tractability reasons.*
- ***M-open**: a situation in which we know the true data-generating model is not in $\boldsymbol{\mathcal{M}}$ and we have no idea what it looks like.*

See Bernardo and Smith (2000) and Yao et al. (2018).

BMS: example I

Suppose one has $x \in \mathbb{N} \cup \{0\}$, which measures, say, the number of eggs Balerion The Black Dread has laid in five consecutive breeding seasons. One can conjure up

$$\mathcal{M}_1 : x \sim \text{Poisson}(\lambda), \lambda > 0,$$

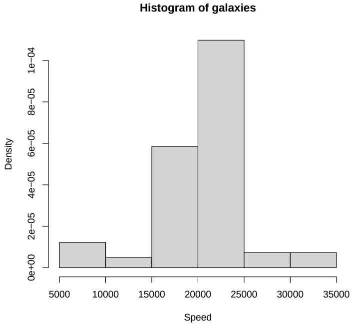
or, if feeling fancy,

$$\mathcal{M}_2 : x \sim \text{Negative-binomial}(\lambda, \phi), \lambda, \phi > 0.$$

Notice that, under $\mathcal{M}_2$, $E[X] = \lambda$ and $\text{Var}(X) = \lambda(1 + \lambda/\phi)$. What happens as $\phi \to \infty$?

BMS: example II

Take the famous Galaxy data set:



A now classical model is a Gaussian mixture:

$$\mathcal{M}_i : v_j \sim \sum_{l=1}^{i} p_{il} \cdot \mathrm{Normal}(v_j; \mu_{li}, \sigma^2_{li}).$$

BMS: example III

Consider the data:

Table 7.1.1. *Orange tree circumferences (in millimeters) against time (in days) for 5 trees. (Source: Gelfand (1996)).*

	tree number				
time	1	2	3	4	5
118	30	33	30	32	30
484	58	69	51	62	49
664	87	111	75	112	81
1004	115	156	108	167	125
1231	120	172	115	179	142
1372	142	203	139	209	174
1582	145	203	140	214	177

Amongst the models we can consider,

$$\mathcal{M}_1 : y_{it} \sim \text{Normal}(\beta_{10} + b_{1i}, \sigma_1^2),$$
$$\mathcal{M}_2 : y_{it} \sim \text{Normal}(\beta_{20} + \beta_{21} T_t + b_{2i}, \sigma_2^2),$$
$$\mathcal{M}_3 : y_{it} \sim \text{Normal}\left(\frac{\beta_{30}}{1 + \beta_{31}\exp\left(\beta_{32} T_t\right)}, \sigma_3^2\right),$$
$$\mathcal{M}_4 : y_{it} \sim \text{Normal}\left(\frac{\beta_{40} + b_{4i}}{1 + \beta_{41}\exp\left(\beta_{42} T_t\right)}, \sigma_4^2\right).$$

Step 0: priors

First, let us look at a convenient representation of model space:

$$\boldsymbol{\Theta} = \bigcup_{i \in I} \{i\} \times \boldsymbol{\Theta}_i.$$

Now, to each $\mathcal{M}_i$, we associate a prior $\pi_i(\theta_i)$ on each subspace and, by Bayes' theorem we get

$$\begin{aligned}
\Pr(\mathcal{M}_i \mid x) &= \Pr(\mu = i \mid x), \\
&= \frac{w_i \int_{\Theta_i} f_i(x \mid t_i)\pi_i(t_i)\, dt_i}{\sum_j w_j \int_{\Theta_j} f_j(x \mid t_j)\pi_j(t_j)\, dt_j},
\end{aligned}$$

where the w_i are the **prior probabilities** for each model.

An intuitive predictive

A nice consequence of the formulation we just saw is that the predictive distribution looks quite intuitive:

$$
\begin{aligned}
p(\tilde{x} \mid \boldsymbol{x}) &= \sum_j w_j \int_{\Theta_j} f_j(\tilde{x} \mid t_j) f_j(\mathbf{x} \mid t_j) \pi_j(t_j)\, dt_j, \\
&= \sum_j \Pr(\mathcal{M}_j \mid \boldsymbol{x}) m_j(\tilde{x}). \qquad (17)
\end{aligned}
$$

Hello, my old friend

Here, Bayes factors also play a central role:

$$\mathrm{BF}_{12} = \frac{\Pr(\mathcal{M}_1 \mid x)}{\Pr(\mathcal{M}_2 \mid x)} \Big/ \frac{\Pr(\mathcal{M}_1)}{\Pr(\mathcal{M}_2)},$$
$$= \frac{w_1' \cdot w_2}{w_2' \cdot w_1},$$

with $w_i' := \Pr(\mathcal{M}_1 \mid x)$.

Model averaging

What if we simply **refuse** to select one model? We can write

$$
\begin{aligned}
p(\tilde{x} \mid \mathbf{x}) &= \int_{\Theta} f(\tilde{x} \mid t) f(\mathbf{x} \mid t) \pi(t)\, dt, \\
&= \sum_j \int_{\Theta_j} f_j(\tilde{x} \mid t_j) g(j, t_j \mid \mathbf{x})\, dt_j, \\
&= \sum_j p(\mathcal{M}_j \mid \mathbf{x}) \int_{\Theta_j} f_j(\tilde{x} \mid t_j) p(t_j \mid \mathbf{x})\, dt_j, \\
&= \sum_j w_j' \int_{\Theta_j} f_j(\tilde{x} \mid t_j) p(t_j \mid \mathbf{x})\, dt_j. \qquad (18)
\end{aligned}
$$

which is another version of the expression in (17).

Model checking

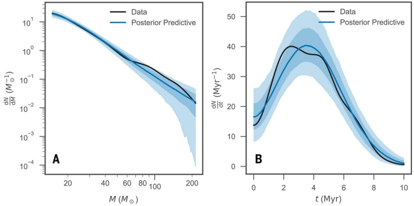
Modern Bayesian inference not only allows for, but actively encourages model interrogation and checking.

- The central idea of **Leave-one-out cross-validation (LOO)** is to estimate the *expected log pointwise predictive density for a new dataset*, elpd:

$$\text{elpd} = \sum_{i=1}^{n} \int m(\tilde{x}_i) \log p(\tilde{x}_i \mid \boldsymbol{x})\, d\tilde{x}_i.$$

  See Vehtari et al. (2017).

- With **Posterior predictive checks (PPCs)** we wish to compare functions of the observed data, $f(\boldsymbol{x})$ with functions of the predictive distribution, $f(\tilde{\boldsymbol{x}})$.



  See Berkhof et al. (2000) and Gabry et al. (2019).

Recommended reading

- Robert (2007), Ch. 7.
- Next lecture: Schervish (1995) Ch. 7.4.

References

Berkhof, J., Van Mechelen, I., and Hoijtink, H. (2000). Posterior predictive checks: Principles and discussion. *Computational Statistics*, 15(3):337–354.

Bernardo, J. M. (1979). Reference posterior distributions for bayesian inference. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(2):113–128.

Bernardo, J. M. and Smith, A. F. (2000). *Bayesian Theory*. John Wiley & Sons.

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

De Finetti, B. (1931). Funzione caratteristica di un fenomeno aleatorio. In *Atti della R Academia Nazionale dei Lincei*, volume 4, pages 251–299.

DeGroot, M. H. (1973). Doing what comes naturally: Interpreting a tail area as a posterior probability or as a likelihood ratio. *Journal of the American Statistical Association*, 68(344):966–969.

Diaconis, P. and Freedman, D. (1980). Finite exchangeable sequences. *The Annals of Probability*, pages 745–764.

Diaconis, P. and Ylvisaker, D. (1979). Conjugate priors for exponential families. *The Annals of Statistics*, pages 269–281.

References

Gabry, J., Simpson, D., Vehtari, A., Betancourt, M., and Gelman, A. (2019). Visualization in bayesian workflow. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 182(2):389–402.

Gelman, A., Simpson, D., and Betancourt, M. (2017). The prior can often only be understood in the context of the likelihood. *Entropy*, 19(10):555.

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

Haldane, J. B. S. (1932). A note on inverse probability. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 28, pages 55–61. Cambridge University Press.

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

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the american statistical association*, 90(430):773–795.

Keynes, J. M. (1921). *A Treatise on Probability*. Macmillan and Company, limited.

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

References

Schervish, M. J. (1995). *Theory of statistics.* Springer Science & Business Media.

Schervish, M. J. (2012). *Theory of Statistics.* Springer Science & Business Media.

Seaman III, J. W., Seaman Jr, J. W., and Stamey, J. D. (2012). Hidden dangers of specifying noninformative priors. *The American Statistician*, 66(2):77–84.

Simpson, D., Rue, H., Riebler, A., Martins, T. G., and Sørbye, S. H. (2017). Penalising model component complexity: A principled, practical approach to constructing priors. *Statistical science*, pages 1–28.

Vehtari, A., Gelman, A., and Gabry, J. (2017). Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and computing*, 27(5):1413–1432.

Yao, Y., Vehtari, A., Simpson, D., Gelman, A., et al. (2018). Using stacking to average bayesian predictive distributions (with discussion). *Bayesian Analysis*, 13(3):917–1007.