

Chapter 1-1: Introduction to Bayesian

1-Bayes Theorem:

Consider a statistical experiment E , with a sample space S of possible outcomes and let $\{B_1, \dots, B_k\}$ comprise a partition of S .

Let $\{p(A); A \subseteq S\}$ denote a probability distribution defined on all event in S .

Then for any events A and B in S , with $p(A) > 0$,

$$p(B|A) = p(A \cap B)/P(A)$$

denotes the conditional probability that B occurs given that A is known to occur.

Bayes Theorem is then

$$p(B_i|A) = \frac{p(A|B_i)p(B_i)}{p(A)}, \quad i = 1, \dots, k$$

where

$$p(A) = \sum_{i=1}^k p(A|B_i)p(B_i)$$

Example:

Consider a disease that is thought to occur in a proportion $\theta = 0.01$ of the population and suppose that a physician observes that out of patients with the disease, 99% possess symptom Z (e.g., a positive result on a blood test) possesses a high propensity for the disease. However, for a randomly chosen person in the population, let S and D , respectively, denote the events that person has the symptom and the disease. Then

$$p(D) = \theta = 0.01 \text{ and } p(S|D) = 0.99$$

therefore, Bayes' Theorem with $k = 2$, $B_1 = D$, and $B_2 = D^c$ tells us that

$$\begin{aligned}
 p(\text{disease}|\text{symptom}) &= p(D|S) \\
 &= \frac{p(S|D)p(D)}{p(S|D)p(D) + P(S|D^c)p(D^c)} \\
 &= \frac{0.99 * 0.01}{0.99 * 0.01 + 0.99 * p(S|D^c)} \\
 &= \frac{1}{1 + 100p(S|D^c)}
 \end{aligned}$$

If $p(S|D^c) = 1/10$, then $p(D|S) = 1/11$.

$$\begin{aligned}
 p(\text{disease}|\text{no - symptom}) &= p(D|S^c) \\
 &= \frac{p(S^c|D)p(D)}{p(S^c|D)p(D) + P(S^c|D^c)p(D^c)} \\
 &= \frac{0.01 * 0.01}{0.01 * 0.01 + 0.99 * \{1 - p(S|D^c)\}} \\
 &= \frac{1}{9901 + 9900p(S|D^c)}
 \end{aligned}$$

So if the symptom is absent, there is still a small probability that the patient possesses the disease.

Example AIDS/HIV

(Please read the paper of **Breslow (1990a): Biostatistics and Bayes (with discussion)** in *Statistical Science*, 5, 269-98

Bayesian inference:

When we have a probability statements noted as $p(\theta|y)$

Bayesian inference is the statistical inference based on the evaluation of the procedure used to estimate θ over possible y values.

Bayes rule:

$$\begin{aligned}
 p(\theta, y) &= p(\theta)p(y|\theta) \\
 p(\theta|y) &= \frac{p(\theta, y)}{p(y)} = \frac{p(\theta)p(y|\theta)}{p(y)}
 \end{aligned}$$

where

$$p(y) = \sum_{\theta} p(\theta)p(y|\theta) \text{ (for discrete values) or}$$

$$p(y) = \int p(\theta)p(y|\theta)d\theta \text{ for continuous values}$$

so

$$p(\theta|y) \propto p(\theta)p(y|\theta)$$

2-Exchangeability:

y_1, \dots, y_n are exchangeable (or permutable) if the joint probability density

$$p(y_1, \dots, y_n)$$

is invariant under any permutation of the indexes or the suffices.

(Please read the paper of **Lindley and Novick**, 1981: **The role of exchangeability in inferences**, *annals of statistics*, 9, 45-58)

It is generally useful to model data from an exchangeable distribution as independently and identically (*i.i.d*) given some unknown parameter vector θ with distribution $p(\theta)$.

3- Hierarchical model:

In hierarchical model, it is possible to speak of exchangeability at each level of units.

Example :

Two medical treatments are applied, in separate randomized experiments, to patients in several different cities.

Then we can suppose that:

- the patients within each city are exchangeable and
- the results from different cities are themselves exchangeable

4-Some important terminology in probability:

Joint distribution of θ_1 and θ_2

$$p(\theta_1, \theta_2) = p(\theta_1) p(\theta_2|\theta_1)$$

we can write also the joint distribution of θ_1, θ_2 , and θ_3 as a product of the marginal and the conditional

$$p(\theta_1, \theta_2, \theta_3) = p(\theta_1|\theta_2, \theta_3)p(\theta_2|\theta_3)p(\theta_3)$$

Posterior Distribution

$$p(\theta|y) = \frac{p(\theta, y)}{p(y)}$$

Conditional distribution of θ_1 given θ_2 is

$$p(\theta_1|\theta_2)$$

Marginal distribution

$$p(\theta_1) = \int p(\theta_1, \theta_2) d\theta_2$$

5-Prediction:

1-The distribution of an unknown but observable y is (a marginal distribution)

$$\begin{aligned} p(y) &= \int p(y, \theta) d\theta \text{ (using marginals)} \\ &= \int p(\theta) p(y|\theta) d\theta \text{ (using Bayes rule)} \end{aligned}$$

Here, this marginal distribution is called: *prior predictive distribution*

2-When we observe y , we can then predict an unknown observable, \tilde{y} (future) as the following:

$$\begin{aligned}
p(\text{future}|\text{past}) &= p(\tilde{y}|y) = \int p(\tilde{y}, \theta|y) d\theta \\
&= \int p(\tilde{y}|\theta, y) p(\theta|y) d\theta \text{ (using Bayes rule)} \\
&= \int p(\tilde{y}|\theta, y) p(\theta|y) d\theta \\
&= \int p(\tilde{y}|\theta) p(\theta|y) d\theta
\end{aligned}$$

here \tilde{y} and y are conditionally independent given θ

For the above quantity, sometime the integration is not straightforward, so an approximation is needed as we will see in the coming chapter and then Gibbs sampler or any Markov chain will be used to generate the sample.

This distribution possesses mean vector

$$E(\tilde{y}|y) = E_{\theta|y}[E(\tilde{y}|\theta, y)]$$

and covariance matrix

$$\text{cov}(\tilde{y}|y) = E_{\theta|y}[\text{cov}(\tilde{y}|\theta, y)] + \text{cov}_{\theta|y}[E(\tilde{y}|\theta, y)]$$

6- Likelihood and Posterior:

Using Bayes Rule gives a relation between y and $p(\theta|y)$ through $p(y|\theta)$ by this:

$$p(\theta|y) \propto p(\theta)p(y|\theta)$$

(what the constant of proportionality is?)

When y is fixed, and when $p(y|\theta)$ is regarded as a function of θ , it is called the **likelihood function**.

Likelihood principle:

For a given sample of data, if $p(y|\theta_1)$ and $p(y|\theta_2)$ have the same likelihood function, it yields to the same inference about θ .

7- Likelihood and odds ratio:

Posterior odds ratio is defined as:

$$odds = \frac{p(\theta_1|y)}{p(\theta_2|y)}$$

where $p(\theta_i|y)$ is the posterior density evaluated at the point θ_i under a given model.

Odds ratio becomes important when combined with Bayes rule:

$$\frac{p(\theta_1|y)}{p(\theta_2|y)} = \frac{p(\theta_1)p(y|\theta_1)/p(y)}{p(\theta_2)p(y|\theta_2)/p(y)} = \frac{p(\theta_1)p(y|\theta_1)}{p(\theta_2)p(y|\theta_2)}$$

8- Example:

Male has X or Y

Female has X or X

Hemophilia is a disease gene carried on the X-Chromosomes

X-chromo: $\begin{cases} \text{male who inherits the gene that causes the disease on X is affected} \\ \text{female who carries the gene on one of her two X is not affected.} \end{cases}$

Prior:

The unknown quantity of interest θ , which is the state of the woman (infected or not) has a prior probability as the following:

father(Y-X)	mother (X-X _{affected})
woman(X-X)	brother (Y-X _{affected})

so the state of the woman θ has two values: the woman is a carrier of the gene or not which means:

$$\theta = 1 \text{ or } \theta = 0 \text{ (the present)}$$

so

$$p(\theta = 1) = p(\theta = 0) = 1/2 \text{ (the present)}$$

Suppose that the data are the woman's sons status which are not affected (**predict the future**).

$y = 1$ for affected son or $y = 0$ for unaffected son respectively. The two items of independent data generates

Likelihood:

$$p(y_1 = 0, y_2 = 0 | \theta = 1) = (0.5)(0.5) = 0.25$$

$$p(y_1 = 0, y_2 = 0 | \theta = 0) = (1)(1) = 1$$

Posterior distribution:

$$\begin{aligned} p(\theta = 1 | y) &= \frac{p(y | \theta = 1)p(\theta = 1)}{p(y | \theta = 1)p(\theta = 1) + p(y | \theta = 0)p(\theta = 0)} \\ &= \frac{(0.25)(0.5)}{(0.25)(0.5) + (1.0)(0.5)} = 0.20 \end{aligned}$$

Prior odds:

$$\frac{p(\theta_1)}{p(\theta_2)} = 0.5/0.5 = 1$$

posterior odds:

$$\frac{p(y | \theta_1 = 1)}{p(y | \theta_2 = 0)} = 0.25/1 = 0.25$$

so the posterior odds is 0.25.

9-Some useful results from probability theory:

if we are dealing with hypothesis or models we may use this notation:

$$p(\theta, y | H) = p(\theta | H)p(y | \theta, H)$$

We use also $E(\cdot)$ for mean and $var(\cdot)$ for variance:

$$E(u) = \int u p(u) du$$

$$var(u) = \int (u - E(u))^2 p(u) du$$

if we are dealing with a vector, then the covariance matrix is defined as:

$$\text{var}(u) = \int (u - E(u))(u - E(u))^t p(u) du$$

Mean and variance of conditional distribution:

$$E(u) = E_v(E_u(u|v))$$

why?

$$\begin{aligned} E(u) &= E(u) = \int u p(u) du = \int \int u p(u, v) du dv \\ &= \int \int u p(u|v) p(v) du dv \\ &= \int \int u p(u|v) du p(v) dv \\ &= \int E(u|v) p(v) dv = E(E(u|v)) \end{aligned}$$

we can do the same to prove that:

$$\text{var}(u) = E(\text{var}(u|v)) + \text{var}(E(u|v))$$

10- Transformation of variables:

suppose that

$p_u(u)$ is the density of the vector u

suppose that

$v = f(u)$ where v has the same number of component as u

if p_u has discrete distribution and f is one to one function, then

$$p_v(v) = p_u(f^{-1}(v))$$

if f is many to one function, then

$$p_v(v) = \sum_u p_u(f^{-1}(v))$$

if p_u is a continuous distribution and f is one to one function, then

$$p_v(v) = |J| p_u(f^{-1}(v))$$

where $|J|$ is the determinant of the Jacobian of the transformation $u = f^{-1}(v)$.

if p_u is a continuous distribution and f is many to one function, then

$$p_v(v) = \int |J| p_u(f^{-1}(v)) du$$

where $|J|$ is the determinant of the Jacobian of the transformation $u = f^{-1}(v)$.

What is a Jacobian?

It is the square matrix of partial derivatives with dimension given by the number of components of u :

example : $u = (u_1, u_2)$

and $v = (v_1, v_2)$

$$\begin{pmatrix} \frac{\partial u_1}{\partial v_1} & \frac{\partial u_1}{\partial v_2} \\ \frac{\partial u_2}{\partial v_1} & \frac{\partial u_2}{\partial v_2} \end{pmatrix}$$

$$v = f(u) = \text{logit}(u) = \log \frac{u}{1-u}$$

then

$$u = f^{-1}(v) = \frac{e^v}{1+e^v} \text{ and } |J| = \left| \frac{\partial u}{\partial v} \right| = \left| \frac{\partial \left(\frac{e^v}{1+e^v} \right)}{\partial v} \right| = \frac{e^v}{(1+e^v)^2}$$

if u is uniform $u \sim U[0, 1]$ then

$$p(v) = p_v(v) = |J| p_u(f^{-1}(v)) = \frac{e^v}{(1+e^v)^2}$$

if u is $\sim N(0, 1)$ then

$$p(v) = |J| p_u(f^{-1}(v)) = \frac{1}{\sqrt{2\pi}} \frac{e^v}{(1+e^v)^2} \exp\left\{-\frac{1}{2} \frac{(e^v)^2}{(1+e^v)^4}\right\}$$

Some important transformations:

Interval	Transformation
$(0, +\infty)$	$\rightarrow (-\infty, +\infty)$ $\log(u)$ to symmetrize
$(0, 1)$	$\rightarrow (-\infty, +\infty)$ $\text{logit}(u)$
$(-\infty, +\infty)$	$\rightarrow u^\alpha$ ($0 < \alpha < 1$) to shrink spread data

Example:

Bayes example

Billiard ball W rolled on a line of length one, with a *uniform* probability of stopping anywhere: W stops at p .

Second ball O then rolled n times under the same assumptions. y denotes the number of times the ball O stopped on the left of W

Given y what inference can we make on p ?

Modern translation:

Derive the posterior distribution of p given y when

$$p \sim U[0, 1] \text{ and } y \sim B(n, p)$$

Since

$$p(Y = y|p) = \binom{n}{y} p^y (1 - p)^{n-y}$$

Remember: when calculating the posterior, we use the ratio of the joint over the marginal

$$p(a < p < b \text{ and } Y = y) = \int_a^b \binom{n}{y} p^y (1 - p)^{n-y} dp$$

using cond dist definition

and

$$p(Y = y) = \int_0^1 \binom{n}{y} p^y (1-p)^{n-y} dp \text{ using marginal definition}$$

then

$$\begin{aligned} p(a < p < b | Y = y) &= \frac{p(a < p < b \text{ and } Y = y)}{p(Y = y)} \\ &= \frac{\int_a^b \binom{n}{y} p^y (1-p)^{n-y} dp}{\int_0^1 \binom{n}{y} p^y (1-p)^{n-y} dp} \\ &= \frac{\int_a^b \binom{n}{y} p^y (1-p)^{n-y} dp}{B(y+1, n-y+1)} \end{aligned}$$

Example 2:

Consider

$$y \sim N(\theta, 1) \text{ and } \theta \sim N(0, 10)$$

$$\begin{aligned} p(\theta|y) &\propto p(y|\theta)p(\theta) \propto \exp\left(-\frac{(y-\theta)^2}{2} - \frac{\theta^2}{20}\right) \\ &\propto \exp\left(-\left(\frac{11\theta^2}{20} - \theta y\right)\right) \\ &\propto \exp\left(-\left(\frac{11}{20}\left\{\theta - \frac{10y}{11}\right\}^2\right)\right) \end{aligned}$$

and

$$\theta|y \sim N\left(\frac{10}{11}y, \frac{10}{11}\right)$$

Natural confidence region

$$\begin{aligned} C &= \{\theta; \pi(\theta|y) > k\} \\ &= \left\{\theta; \left|\theta - \frac{10}{11}y\right| > k'\right\} \end{aligned}$$

Highest Posterior density (HPD) region

11- Introduction to simulation

Definition: Simulating is generating a

sample from a probability distribution even when the density function cannot be explicitly integrated.

Probability density function duality histogram of a set of random draws from the distribution.

Advantages of simulation:

1. Extract some Characteristics of the sample: moments, average, median, percentiles... *To estimate the 95th percentile of a sample of size m , we simulate first the sample and we use the 95Lth order statistics.*
2. Extrem values (small or large) might not noticed if estimates are generated in analytic form.

Sampling using inverse cumulative distribution

Cumulative distribution function is defined as

$$\begin{aligned} cdf &= F(a) = p(v \leq a) = u \\ &= \begin{cases} \sum_{v \leq a} p(v) & \text{if } p \text{ is discrete} \\ \int_{-\infty}^a p(v) dv & \text{if } p \text{ is continuous} \end{cases} \end{aligned}$$

Remember, u is a probability and a is a percentile.

When simulating, we generate values (a_1, \dots, a_k) which have a certain distribution p , so here the inverse cdf can be used to obtain random samples (a_1, \dots, a_k) from the distribution p :

1. draw a random value, U , from the uniform distribution on $[0, 1]$. U is a probability

2. calculate

$$v = F^{-1}(U)$$

3. The value v will be a random draw from p and easy to calculate as long as $F^{-1}(U)$ is simple. For discrete distribution, F^{-1} can simply be tabulated.

Example1:

$$v \sim \text{expon}(\lambda)$$

$$p(v|\lambda) = \lambda e^{-\lambda v}, v > 0, E(v) = \frac{1}{\lambda}, \text{var}(v) = \frac{1}{\lambda^2}$$

$$\begin{aligned} \text{cdf} = F(a) &= p(v \leq a) = \int_{-\infty}^a p(v) dv \\ &= \int_{-\infty}^a \lambda e^{-\lambda v} dv = \int_0^a \lambda e^{-\lambda v} dv \\ &= [e^{-\lambda v}]_0^a = 1 - e^{-\lambda a} \end{aligned}$$

If we simulate $U \sim [0, 1]$, then

$$F^{-1}(U) = v = -\log(1 - U)/\lambda$$

Conclusion: we can obtain random draws from the exponential distribution as $-\log(1 - U)/\lambda$.

Example2:

$$v \sim \text{Cauchy}(\lambda)$$

$$p(v|\lambda) = \frac{1}{\pi} [1 + (v - \lambda)^2]^{-1}$$

$$\begin{aligned}
cdf = F(a) = p(v \leq a) &= \int_{-\infty}^a p(v) dv \\
&= \int_{-\infty}^a \frac{1}{\pi} [1 + (v - \lambda)^2]^{-1} dv = \frac{1}{\pi} \int_{-\infty}^{a-\lambda} \frac{1}{1 + z^2} dz \\
&= \frac{1}{\pi} [\arctan]_{-\infty}^{a-\lambda} = \frac{1}{\pi} [\arctan(a - \lambda) + \frac{1}{2}\pi]
\end{aligned}$$

If we simulate $U \sim [0, 1]$, then

$$F^{-1}(U) = v = \text{tg}(\pi U - \frac{1}{2}\pi) + \lambda$$

Remark:

Usually, the simulation are important when we are dealing with posterior distribution of a model of parameter θ , or from the posterior predictive distribution of unknown observables \tilde{y} .

Example:

If we simulate 1000 draws for $\theta = (\theta_1, \dots, \theta_k)$, we can calculate:

- estimation of θ_1/θ_4 by just calculating the ratio of the 1000 values simulated.
- estimation of $p(\tilde{y} > e^{\theta_1})$ by calculating the proportion within the 1000 for which this inequality is true.
- $p(\theta < a) = 0.025$ and $p(\theta > b) = 0.025$ by calculating the 25th and 976th order statistics.

12-Some important integral to know:

(a) Integration by Parts:

Integration by Parts (called also, Integration by Fubini's theorem) are useful in

many contexts. Here we record a few of the most useful ones.

Proposition:

Suppose that the left-continuous function f and the right-continuous function g are nondecreasing functions. Then for any $a \leq b$

$$f_+(b)g(b) - f(a)g_-(a) = \int_{[a,b]} f dg + \int_{[a,b]} g df$$

and

$$f(b)V(b) - f(a)g(a) = \int_{(a,b]} f dg + \int_{[a,b)} g df$$

where

$$f_+(x) = \lim_{y \rightarrow x} f(y) \text{ and } g_-(x) = \lim_{y \rightarrow x} g(y)$$

(b) Example: an integration we will need in chapter 2:

$$I = \int_0^{+\infty} \sigma^{-a} e^{-\frac{x}{2\sigma^2}} d\sigma^2 \tag{1.1}$$

if we put $z = \frac{x}{2\sigma^2}$ then

$$I = -\left(\frac{x}{2}\right)^{-\frac{a}{2}+1} \int_0^{+\infty} z^{\frac{a}{2}-2} e^{-z} dz \tag{1.2}$$

We have

$$\begin{aligned} & \int_0^{+\infty} z^{\frac{a}{2}-2} e^{-z} dz \tag{1.3} \\ &= \left[z^{\frac{a}{2}-3} e^{-z} \right]_0^{+\infty} + \int_0^{+\infty} z^{\frac{a}{2}-3} e^{-z} dz = 0 + \int_0^{+\infty} z^{\frac{a}{2}-3} e^{-z} dz \\ & \text{because } \underset{z \rightarrow +\infty}{z e^{-z}} \rightarrow 0 \text{ and } \underset{z \rightarrow 0}{z e^{-z}} \sim 0 \\ &= \dots = \int_0^{+\infty} z^{(\frac{a}{2}-\frac{a}{2})} e^{-z} dz = [e^{-z}]_0^{+\infty} = -1 \end{aligned}$$

So we have

$$\begin{aligned} I &= -\left(\frac{x}{2}\right)^{-\frac{a}{2}+1} \int_0^{+\infty} z^{\frac{a}{2}-2} e^{-z} dz \tag{1.4} \\ &= \left(\frac{x}{2}\right)^{-\frac{a}{2}+1} \end{aligned}$$