STAT3016/4116/7016:
Introduction to Bayesian Data Analysis

RSFAS, College of Business and Economics, ANU

Review of probability and Exchangeability

Partitions

Definition: A collection of sets $\{H_1, ..., H_k\}$ is a partition of another set $\mathscr{H}$ if

1. The events are disjoint, $H_i \cap H_j = \emptyset$ for $i \neq j$
2. The union of the sets is $\mathscr{H}$, which we write as $\cup_{k=1}^{K} H_k = \mathscr{H}$.

In the context of identifying which of several statements is true, if $\mathscr{H}$ is the set of all possible truths and $\{H_1, ..., H_k\}$ is a partition of $\mathscr{H}$, then exactly one out of $\{H_1, ..., H_k\}$ contains the truth.

Example: let $\mathscr{H}$ be someone's number of children. Partitions could be
- $\{0, 1, 2, 3 \textit{ or more}\}$
- $\{0, 1, 2, 3, 4, 5, 6, ...\}$

Partitions and probability

$Pr(\mathscr{H}) = 1$ and let E be some event.

Rule of total probability: $\sum_{k=1}^{K} Pr(H_k) = 1$

Rule of marginal probability:

$$\begin{aligned} Pr(E) &= \sum_{k=1}^{K} Pr(E \cap H_k) \\ &= \sum_{k=1}^{K} Pr(E|H_k)Pr(H_k) \end{aligned}$$

Partitions and probability

Bayes' rule:

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

Partitions and probability

Example: A survey is conducted on a sample of males over 30 years of age. Information is collected on income and education. Let H_i denote the event that a randomly selected person from the sample is in the ith income quartile of the sample. What is $Pr(H_1)$?

Note that $\{H_1, H_2, H_3, H_4\}$ define a partition.

Let E be the event that a randomly sampled person from the survey has a college education. We are told:
$\{Pr(E|H_1), Pr(E|H_2), Pr(E|H_3), Pr(E|H_4)\} =$
$\{0.11, 0.19, 0.31, 0.53\}$

Why don't the above probabilities sum to 1?
Find $Pr(H_i|E)\ \forall_i$. Interpret your answers

Partitions and probability

In Bayesian inference, $\{H_1, ..., H_k\}$ often refer to disjoint hypotheses or states of nature, and E refers to the outcome of a survey, study or experiment. To compare hypotheses post-experimentally, we calculate:

$$\begin{aligned}
\frac{Pr(H_i|E)}{Pr(H_j|E)} &= \frac{Pr(E|H_i)Pr(H_i)/Pr(E)}{Pr(E|H_j)Pr(H_j)/Pr(E)} \\
&= \frac{Pr(E|H_i)Pr(H_i)}{Pr(E|H_j)Pr(H_j)} \\
&= \frac{Pr(E|H_i)}{Pr(E|H_j)} \times \frac{Pr(H_i)}{Pr(H_j)} \\
&= \text{“Bayes factor”} \times \text{“prior beliefs”}
\end{aligned}$$

Independence

Definition Two events F and G are conditionally independent given H if $Pr(F \cap G|H) = Pr(F|H)Pr(G|H)$.

Conditional independence implies $Pr(F|H \cap G) = Pr(F|H)$.

Random variables

A random variable is defined as an unknown numerical quantity about which we make probability statements. In Bayesian inference, the random variable could be the quantitative outcome of a survey or experiment, or a population parameter.

Let Y be a random variable and let $\mathscr{Y}$ be the set of all possible values of Y.

Discrete random variables: Y is a discrete random variable if the set of all possible outcomes is countable.

Examples: - number of children of a randomly sampled person
- number of years of education of a randomly sampled person.

Discrete random variables

We write $Pr(Y = y) = p(y)$ (*probability density function of Y*).

1. $0 \leq p(y) \leq 1$ for all $y \in \mathscr{Y}$
2. $\sum_{y \in \mathscr{Y}} p(y) = 1$

We use the pdf to make any general probability statements about Y. Example, $Pr(Y \in A) = \sum_{y \in A} p(y)$.

Discrete random variables

Example: Binomial distribution

If $Y \sim Bin(n, \theta)$, then $Pr(Y = y|\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}$

For example, a player bets one one of the numbers 1 through 6. Three dice are then rolled, and if the number bet by the player appears i times, $i = 1, 2, 3$, then the player wins i units; on the other hand, if the number bet by the player does not appear on any of the dice, then the player loses 1 unit. Is this game fair to the player?

Example: Poisson distribution

If $Y \sim Pois(\theta)$, then $Pr(Y = y|\theta) = \theta^y e^{-\theta}/y!$

For example, the average number of births per Australian woman is 1.93 (based on 2012 data). What is the probability that a randomly sampled woman of child bearing age has 5 children??

Continuous random variables

Continuous random variables: Suppose that the sample space $\mathscr{Y}$ is roughly the set of all real numbers. Is $Pr(Y \leq 5) = \sum_{y \leq 5} p(y)$ defined?

We need to define the cumulative distribution function (cdf):

$$F(y) = Pr(Y \leq y)$$

($F(\infty) = ?; F(-\infty) = ??$) also $F(b) \leq F(a)$ if $b < a$.

For every continuous cdf F, there exists a positive function $p(y)$ such that

$$F(a) = \int_{-\infty}^{a} p(y)dy$$

1. $0 \leq p(y)$ for all $y \in \mathscr{Y}$
2. $\int_{-\infty}^{\infty} p(y)dy = 1$

So $Pr(Y \in A) = \int_{y \in A} p(y)dy$

(nb: we can think of integration as a generalisation of summation when the sample space is not countable)

Continuous random variables

Example: Normal distribution

If $Y \sim N(\mu, \sigma^2)$, $p(y|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} exp\left\{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right\}$

Suppose that a binary message - either 0 or 1 - must be transmitted by wire from location A to location B. However, the data sent over the wire are subject to a channel noise disturbance, so to reduce the possibility of error, the value 2 is sent over the wire when the message is 1 and the value -2 is sent when the message is 0. If $x = \pm 2$ is the value sent at location A, then R, the value received at location B is given by $R = x + N$, where N is the channel noise disturbance. When the message is received at location B the receiver decodes it according to the following rule:

- If $R \geq 0.5$ then 1 is concluded
- If $R < 0.5$ the 0 is concluded

Suppose N is a standard normal random variable. Find Pr(error | message is 1) and Pr(error | message is 0).

Descriptions of distributions

- $E[Y] = \sum_{y \in \mathscr{Y}} yp(y)$ if Y is discrete
- $E[Y] = \int_{y \in \mathscr{Y}} yp(y)dy$ if Y is continuous

$$Var[Y] = E[(Y - E[Y])^2] = E[Y^2] - [E[Y]]^2$$

$\alpha - \textit{quantile}$ is the value y_α such that $F(Y_\alpha) \equiv Pr(Y \leq y_\alpha) = \alpha$

Joint distributions - discrete

Discrete distributions:

$$p_{Y_1 Y_2}(y_1, y_2) = Pr(\{Y_1 = y_1\} \cap \{Y_2 = y_2\}), \text{ for } y_1 \in \mathscr{Y}_1, y_2 \in \mathscr{Y}_2$$

The marginal density can be computed from the joint density:
$P_{Y_1}(y_1) \equiv Pr(Y_1 = y) = \sum_{y_2 \in \mathscr{Y}_2} p_{Y_1, Y_2}(y_1, y_2)$

Also can derive the conditional density $p_{Y_2|Y_1}(y_2|y_1)$ from the joint and marginal density.

Can $p_{Y_1 Y_2}(y_1, y_2)$ be derived from $p_{Y_1}(y_1)$ and $p_{Y_2}(y_2)$?

Joint distributions - continuous

Given a joint continuous cdf
$F_{Y_1 Y_2}(a, b) \equiv Pr(\{Y_1 \leq a\} \cap \{Y_2 \leq b\})$, then there is a function $p_{Y_1 Y_2}$

$$F_{Y_1 Y_2}(a, b) = \int_{-\infty}^{a} \int_{-\infty}^{b} p_{Y_1 Y_2}(y_1, y_2) dy_2 dy_1$$

and as in the discrete case

- $p_{Y_1}(y_1) = \int_{-\infty}^{\infty} p_{Y_1 Y_2}(y_1, y_2) dy_2$
- $p_{Y_2|Y_1}(y_2|y_1) = p_{Y_1 Y_2}(y_1, y_2)/p_{Y_1}(y_1)$

Joint distributions - continuous

Example: suppose that the joint density of X and Y is given by

$$f(x,y) = \begin{cases} \frac{e^{-x/y}e^{-y}}{y} & 0 < x < \infty, 0 < y < \infty \\ 0 & \textit{otherwise} \end{cases}$$

Find $Pr(X > 1 | Y = y)$

Joint distributions - mixed continuous and discrete

Let Y_1 be discrete and let Y_2 be continuous with joint density $p_{Y_1 Y_2}(y_1, y_2) = p_{Y_1}(y_1) \times p_{Y_2|Y_1}(y_2|y_1)$.

Find the expression for $Pr(Y_1 \in A, Y_2 \in B)$.

Joint distributions -Bayes' rule and parameter estimation

Let θ denote a population parameter and let y denote our observed data. Bayesian estimation requires derivation of $p(\theta|y)$. Using Bayes' rule

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

Suppose θ_a and θ_b are two possible numerical values for θ. The relative posterior probability is:

$$\frac{p(\theta_a|y)}{p(\theta_b|y)} = \frac{p(\theta_a)p(y|\theta_a)}{p(\theta_b)p(y|\theta_b)}$$

Note, we do not need to compute $p(y) = \int_\Theta p(\theta)p(y|\theta)d\theta$, the proportionality constant. That is $p(\theta|y) \propto p(\theta)p(y|\theta)$.

Independent random variables

Under independence the joint density is given by

$$p(y_1, ..., y_n|\theta) = p_{Y_n}(y_n|\theta) \times ... \times p_{Y_n}(y_n|\theta) \prod_{i=1}^{n} p_{Y_i}(y_i|\theta)$$

Suppose $Y_1, ...Y_n$ are random variables generated by some common process and $Y_1, ..., Y_n|\theta \overset{\text{iid}}{\sim} p(y|\theta)$. This means

$$p(y_1, ..., y_n|\theta) = \prod_{i=1}^{n} p(y_i|\theta)$$

Exchangeability

Frequentist statistics relies heavily on the concept of *independence*, and hypothetical repeated sampling under conditions that are as close to independent and identically distributed as possible, given a parameter

Example: Hospital specific mortality rates. Suppose I'm interested in the quality of care administered by Hospital A. I collect medical record data over the time period 2010-2013 for 400 patients, and record whether 30 days after admission, the patient is alive (Y=0) or dead (Y=1).

iid. before the study (2010-2013)

From a frequentist perspective and the vantage point of 2009, let Y_i denote the unknown death outcome of patient i. To model the uncertainty in the outcome we assume:

$$Y_i \overset{\text{iid}}{\sim} \text{Bern}(\theta)$$

Exchangeability

Suppose I went through the records in chronological order. After going through 2010 data, if I find 95% of patients died, this is going to change my predictions on the death outcomes of future years . Under the Bayesian framework, I can modify my beliefs on θ. I haven't actually observed θ and I wish to quantify my uncertainty on θ.

Contrast this to the frequentist approach where the model is used to describe the process of observing a repeatable event (that is, random sampling under 'iid' conditions).

Exchangeability

marginal distribution unconditional on parameters.

Exchangeability is a more general concept than focusing on 'iid' observations. To be exchangeble, the y_i's must have a (marginal) distribution for the vector y that is the same regardless of the order in which the observations are written down.

Definition: Let $p(y_1, ..., y_n)$ be the joint density of $Y_1, ..., Y_n$. If $p(y_1, ..., y_n) = p(y_{\pi_1}, ..., y_{\pi_n})$ for all permutations of π of $\{1, ..., n\}$, then $Y_1, ..., Y_n$ are exchangeable.

Actually not assuming

Exchangeability as a Bayesian concept is parallel to the frequentist concept of independence.

independence

Exchangeability

Exercise 1: Suppose we make the following assumption on our data

$$y_1, \ldots, y_n | \theta \overset{\text{iid}}{\sim} \text{Bern}(\theta)$$

with

$$\theta \sim \text{beta}(a, b)$$

So the y_i's are conditionally iid. Show that

1. The y_i's are not unconditionally independent
2. Show that the sequence $y_1, \ldots, y_n$ is exchangeable

2. $f(y_1, \cdots, y_n) = f(n)$
$= \int f(y|\theta) f(\theta) d\theta$
$= \int \prod_{i=1}^{n} f(y_i|\theta) f(\theta) d\theta$
$= \int \prod_{i=1}^{n} f(y_{\pi_i}|\theta) f(\theta) d\theta$ ✡
$= f(y_{\pi_1}, \ldots, y_{\pi_n})$

1.
1. Show $\text{Cov}(y_i, y_j) \neq 0$, $i \neq j$ (unconditional)
$= \text{Cov}(E(y_i|\theta) \cdot E(y_j|\theta)) + E(\text{Cov}(y_i, y_j|\theta))$ ← $EX = E(E(X|Y))$
$= \text{Cov}(\theta, \theta) + 0$
$= \text{Var}\,\theta \quad (> 0)$
$\neq 0$ b/c $\theta \sim \text{Beta}(a, b)$

Exchangeability

IID implies exchangeability but not the converse.

Exchangeable sequences are identically distributed, just not necessarily independent. Exchangeability is therefore a broader concept than IID. While frequentist inference makes heavy use of IID, Bayesian inference uses IID much less often, and more commonly uses exchangeability.

Exchangeability and de Finetti's representation theorem

If $\{y_1, ..., y_n\}$ is an exchangeable sequence of real-valued random quantities, then any finite subset of them is a random sample of some model $p(y_i|\theta)$ and there exists a prior distribution $p(\theta)$ which has to describe the initially available information about the parameter which labels the model, hence requiring a Bayesian approach.

In other words, exchangeable random variables can always be generated by introducing some parameter and using the Bayesian conditional 'iid' formulation.

In short, exchangeability is a justification for Bayesian inference.

di Finetti's theorem

Let $Y_i \in \mathscr{Y}$ for all $i \in \{1, 2, ...\}$. Suppose that, for any n, our belief model for $Y_1, ..., Y_n$ is exchangeable. Then our model can be written as

$$p(y_1, ..., y_n) = \int \left\{ \prod_1^n p(y_i|\theta) \right\} p(\theta) d\theta$$

for some parameter θ, and some sampling model $p(y|\theta)$.

In short: de Finetti's theorem states that an infinite exchangeable sequence is distributed as a mixture of conditional IID sequences.

Alternatively, think of the theorem as the representation of an infinite sequence of exchangeable random variables by a mixture of independent random variables.

di Finetti's theorem

With exchangeability and di Finetti's theorem, we can easily objectively update our beliefs on θ as more data is collected. This is difficult to do if we consider θ to be fixed and $y_1, ..., y_n$ are IID draws from some distribution with unknown but fixed parameter θ.

Exchangeability

Exercise 2: Suppose in a random sample of 1000 transportation workers, all were found to be on drugs. You are told that exactly 10% of workers in the transport industry are on drugs and you truly believe the model. Predict the outcome of the 1001st individual who will now be sampled using a:

- non-Bayesian statistical analysis
- Bayesian analysis

① 0% (sample has no implication on next individual)
OR
② 100% (based on the sample)

Do your conclusions differ between the two approaches? How so?

$\Pr(\tilde{y} \mid y_1, \ldots, y_n)$ posterior predictive probability
can do this via simulation
$= E(\theta \mid y_1, \ldots, y_n)$ or analytically