Extension of the LTP (4)

Assuming all conditional probabilities are defined in what follows, if C is in $\mathcal{F}$ then

$$P(A|C) = \sum_{j=1}^{k} P(B_j|C)\, P(A \mid B_j \cap C).$$

Definition

Events A, B are <u>independent</u> iff

$$P(A \cap B) = P(A) \cdot P(B), \quad \leftarrow \text{(freq)}$$

which (as long as $P(A) > 0$, $P(B) > 0$) is equivalent to

$$\left.\begin{aligned} P(A|B) &= P(A) \\ \text{and} \quad P(B|A) &= P(B) \end{aligned}\right\} \leftarrow \text{(Bayesian)}.$$

Consequences of the definition of independence

① If A and B are independent, then so are A and B^c, A^c and B, and A^c and B^c.

② Extension of the definition to more than 2 events:

Definition:

Given events $A_1, \dots, A_k$, they are (mutually) independent if, for every subset $A_{i_1}, \dots A_{i_j}$ of $(A_1, \dots, A_k)$ $(j = 2, \dots, k)$,

$$P(A_{i_1} \cap \dots \cap A_{i_j}) = P(A_{i_1}) \cdot \dots \cdot P(A_{i_j})$$

(Bayesian) Interpretation of independence

A, B independent $\overset{\text{iff}}{\Longleftrightarrow}$ information about A doesn't change the chances associated with B, and vice versa.

Definition

Another useful extension of independence

Events $\{A_1, \dots, A_k\}$ are conditionally independent given event B if for every subset $\{A_{i_1}, \dots, A_{i_j}\}$ of $\{A_1, \dots, A_k\}$ $(j = 2, \dots, k)$

product

$$P(A_{i_1} \cap \dots \cap A_{i_j} \mid B) = \prod_{\ell=1}^{j} P(A_{i_\ell} \mid B).$$

---

statistical Example Suppose that there is a machine that can take an ordinary coin and produce IID tosses of the coin with $P(\underset{\text{any one toss}}{H}) = \theta$, and θ can be set to any value in $[0,1]$ with a dial on the machine's control panel. Someone sets the dial to a θ value that's <u>unknown</u> to you and starts producing coin tosses $Y_1, Y_2, \dots$

Suppose the first 10 tosses come out

1	0	1	1	1	0	0	1	1	1	← "bits" (binary digits)
H	T	H	H	H	T	T	H	H	H	(7 H, 3 T).

(John Tukey)

$\boxed{Q:}$ Is there information in these first 10 tosses that helps you to predict Y_{11}?

A: Yes, definitely: it looks like θ is around $\frac{7}{10}$, so you would predict $I_{11} = H$. Thus I_{11} **depends** on $I_1, \dots, I_{10}$ probabilistically. Now, suppose instead that you <u>watched</u> the person with the machine set the dial to $\theta = 0.81$, so that θ is now <u>known</u> to you. The next 10 tosses come out H H H T H T H H H H (8 H, 2 T). **Q:** Is there information in these 10 tosses that helps you to predict the next toss?

A: <u>No</u>; you know that $\theta = 0.81$, so there's no information in any of the I_i

that helps you to predict any of the other I_j.

given θ, the I_i are $\boxed{\text{indep.}}$

Thus the I_i are unconditionally dependent but conditionally independent given θ.

(4 Aug 17)

~~Bayes's Theorem for events~~ (a finite partition)

Suppose that the events $B_1, \dots, B_k$ partition the sample space in such a way that $P(B_j) > 0$ for all $j = 1, \dots, k$. If A is an event with $P(A) > 0$, then for each $i = 1, \dots, k$

$$P(B_i \mid A) = \frac{P(B_i)\, P(A \mid B_i)}{P(A)}$$

and, by the LTP, this is

$$P(B_i \mid A) = \frac{P(B_i) \cdot P(A \mid B_i)}{\sum_{j=1}^{k} P(B_j) \cdot P(A \mid B_j)} .$$

How this theorem is used in Bayesian statistics: The B_i represent unknown states of the world: they're all possible — $P(B_i) > 0$ — and only one of them is true, but you don't know which one. A represents <u>data</u>: information that will help you identify the most probable B_i.

Before (a priori) the dataset A arrives, you have background information about the plausibility of the B_i that you can represent with prior probabilities $P(B_i)$. After (a posteriori) the dataset A arrives, you can use Bayes's Theorem to update your prior probabilities to posterior probabilities $P(B_i \mid A)$.

The probabilities $P(A \mid B_i)$ represent how likely the dataset A would be if B_i were the actual unknown state; this is often called likelihood information.

$P(A)$ (the denominator) does not depend on the B_i, and can therefore be regarded as a normalizing constant, put into Bayes's theorem to make all the $P(B_i \mid A)$ add up to 1. Thus

$$P(B_i \mid A) = \frac{P(B_i)\, P(A \mid B_i)}{P(A)}$$

is interpreted as

$$\begin{pmatrix} \text{posterior} \\ \text{information} \end{pmatrix} = \frac{\begin{pmatrix} \text{prior} \\ \text{information} \end{pmatrix} \cdot \begin{pmatrix} \text{(data) likelihood} \\ \text{information} \end{pmatrix}}{\begin{pmatrix} \text{normalizing} \\ \text{constant} \end{pmatrix}}.$$

Random variables and their distributions

Example: Tay-Sachs Disease

T = T-S baby
N = not

$\mathcal{S}$	Y
NNNNN	0
TNNNN	1
NTNNN	1
NNTNN	1
NNNTN	1
NNNNT	1
TTNNN	2
TNTNN	2
TNNTN	2
TNNNT	2
NTTNN	2
NTNTN	2
NTNNT	2
NNTTN	2
NNTNT	2
NNNTT	2
⋮	⋮
TTTTT	5

← # of T-S babies = Y

Definition

Given a sample space $\mathcal{S}$ for an experiment $\mathcal{E}$, a (real-valued) <u>random variable</u> (rv) is a function from the non-weird collection $\mathcal{C}$ of subsets of $\mathcal{S}$ to the real number line $\mathbb{R}$.

In the T-S case study, the elements s of $\mathcal{S}$ look like NNNTN and the rv Y counts how many Ts they contain

For instance, $\mathcal{I}(TNNTN) = 2$ and $\mathcal{I}(NNNTT) = (a)(so)/2$ (i.e., $\mathcal{I}$ ignores the order of the children). We can use the following notation to simplify things.

Notation $\quad P(\underbrace{\mathcal{I} = y}_{\text{T/F proposition}}) \triangleq P(\underbrace{\{s : \mathcal{I}(s) = y\}}_{\text{set}})$

For example, $P(\mathcal{I} = 1) = P(\{s \in \mathcal{S} : \mathcal{I}(s) = 1\})$
$= P(\{TNNNN, NTNNN, NNTNN, NNNTN, NNNNT\})$. In general the values a random variable takes on could be just about anything, but in this course all of our rvs will be real-valued. In the T-S case study the rv $\mathcal{I}$ can only take on the values $0, 1, \ldots, 5$.

(7 Aug 17)

y	$P(Y = y)$
0	0.237
1	0.396
2	0.264
3	0.088
4	0.015
5	0.001

(see p. 27)

You can see that a rv Y is completely specified by two things: the values it can take on, and the probabilities for those values.

Definition/ The (probability) distribution of a random variable Y is the collection of all probabilities of the form $P(Y \in A)$ for all sets A of real numbers in the non-weird collection $\mathcal{C}_{\mathbb{R}}$ of subsets of the real number line $\mathbb{R}$.

The rv Y in the T-S case study has a finite set of possible values —

this is true of some, but not all, rvs.

Definition A random variable Y has a <u>discrete distribution</u>, or equivalently Y is a <u>discrete</u> rv, if the set of (distinct) possible values of Y is finite or at most countably infinite; rvs for which the set of possible values is uncountable are called <u>continuous random variables</u>.

Examples ① The rv $X = \begin{cases} 1 & \text{if } Y > 0 \\ 0 & \text{otherwise} \end{cases}$

(with Y = # T-s baby) is discrete, taking on only the values $\{0, 1\}$ – such rvs are called <u>dichotomous</u> or <u>binary</u>.

{yes, no} {1, 0}

(2) Imagine a scale for weighing things (65)
that has a dial you can set to specify
how many significant figures (sigfigs) of precision
you want. Buy a "1 pound" package of
butter at your favorite market and weigh it.

possible weights (ounces)
16
16.0
15.99
15.9930
15.9928
⋮

If there's no conceptual limit to the number of sigfigs you could get, a rv $Y = \begin{pmatrix} \text{the actual (true)} \\ \text{weight of the package} \end{pmatrix}$ should be modeled as continuous, having values (e.g.) on $(0, \infty)$, the positive part of $\mathbb{R}$. Reality check: Infinite precision is impossible in practice;

every measurement you ever make is in actuality discrete, but it's useful to regard rvs that are <u>conceptually</u> continuous (i.e., no limit in principle to the precision of measurement) as continuous.

(23 Apr 19) $\boxed{\text{Definition}}$ Given a <u>discrete</u> rv Y, the <u>probability</u> (mass) <u>function</u> (pmf or pf) of Y is the function f_Y that keeps track of the probabilities associated with Y: $f_Y(y) = P(Y = y)$.

The set $\{ y : f_Y(y) > 0 \}$ is called the <u>support</u> of (the distribution of) Y.

(DS is almost unique in using "pf"; nearly everybody talks about the pmf.)