COMP2610 / COMP6261 - Information Theory

Lecture 3: Probability Theory and Bayes' Rule

Robert C. Williamson

Research School of Computer Science

Australian National University

July 30, 2018

Last time

- A general communication system
- Why do we need probability?

- Basics of probability theory

- Joint, marginal and conditional distributions

Review Exercise

Suppose I go through the records for $N = 1000$ students, checking their admission status, $A = \{0, 1\}$, and whether they are "brilliant" or not, $B = \{0, 1\}$

(Aside: "Brilliance" is a dodgy concept, and does not predict scientific achievement as well as persistence and combinatorial ability; see e.g. Dean Simonton, *Scientific Genius: A Psychology of Science*, Cambridge University Press, 2009; this is just a toy example!)

Say that the counts for admission and brilliance are

	$B = 0$	$B = 1$
$A = 0$	680	10
$A = 1$	220	90

Then:

$$p(A = 1, B = 0)$$
$$p(B = 1)$$
$$p(A = 0)$$
$$p(B = 1 | A = 1)$$
$$p(A = 0 | B = 0)$$

Review Exercise

Suppose I go through the records for $N = 1000$ students, checking their admission status, $A = \{0, 1\}$, and whether they are “brilliant” or not, $B = \{0, 1\}$

(Aside: “Brilliance” is a dodgy concept, and does not predict scientific achievement as well as persistence and combinatorial ability; see e.g. Dean Simonton, *Scientific Genius: A Psychology of Science*, Cambridge University Press, 2009; this is just a toy example!)

Say that the counts for admission and brilliance are

	$B = 0$	$B = 1$
$A = 0$	680	10
$A = 1$	220	90

Then:

$p(A = 1, B = 0)$	$220/1000$
$p(B = 1)$	
$p(A = 0)$	
$p(B = 1 \mid A = 1)$	
$p(A = 0 \mid B = 0)$	

Review Exercise

Suppose I go through the records for $N = 1000$ students, checking their admission status, $A = \{0, 1\}$, and whether they are "brilliant" or not, $B = \{0, 1\}$

(Aside: "Brilliance" is a dodgy concept, and does not predict scientific achievement as well as persistence and combinatorial ability; see e.g. Dean Simonton, *Scientific Genius: A Psychology of Science*, Cambridge University Press, 2009; this is just a toy example!)

Say that the counts for admission and brilliance are

	$B = 0$	$B = 1$
$A = 0$	680	10
$A = 1$	220	90

Then:

$p(A = 1, B = 0)$	$220/1000$
$p(B = 1)$	$100/1000$
$p(A = 0)$	
$p(B = 1 \vert A = 1)$	
$p(A = 0 \vert B = 0)$	

Review Exercise

Suppose I go through the records for $N = 1000$ students, checking their admission status, $A = \{0, 1\}$, and whether they are "brilliant" or not, $B = \{0, 1\}$

(Aside: "Brilliance" is a dodgy concept, and does not predict scientific achievement as well as persistence and combinatorial ability; see e.g. Dean Simonton, *Scientific Genius: A Psychology of Science*, Cambridge University Press, 2009; this is just a toy example!)

Say that the counts for admission and brilliance are

	$B = 0$	$B = 1$
$A = 0$	680	10
$A = 1$	220	90

Then:

$p(A = 1, B = 0)$	$220/1000$
$p(B = 1)$	$100/1000$
$p(A = 0)$	$690/1000$
$p(B = 1 \mid A = 1)$	
$p(A = 0 \mid B = 0)$	

Review Exercise

Suppose I go through the records for $N = 1000$ students, checking their admission status, $A = \{0, 1\}$, and whether they are "brilliant" or not, $B = \{0, 1\}$

(Aside: "Brilliance" is a dodgy concept, and does not predict scientific achievement as well as persistence and combinatorial ability; see e.g. Dean Simonton, *Scientific Genius: A Psychology of Science*, Cambridge University Press, 2009; this is just a toy example!)

Say that the counts for admission and brilliance are

	$B = 0$	$B = 1$
$A = 0$	680	10
$A = 1$	220	90

Then:

$p(A = 1, B = 0)$	$220/1000$
$p(B = 1)$	$100/1000$
$p(A = 0)$	$690/1000$
$p(B = 1 \mid A = 1)$	$90/310$
$p(A = 0 \mid B = 0)$	

Review Exercise

Suppose I go through the records for $N = 1000$ students, checking their admission status, $A = \{0, 1\}$, and whether they are "brilliant" or not, $B = \{0, 1\}$.

(Aside: "Brilliance" is a dodgy concept, and does not predict scientific achievement as well as persistence and combinatorial ability; see e.g. Dean Simonton, *Scientific Genius: A Psychology of Science*, Cambridge University Press, 2009; this is just a toy example!)

Say that the counts for admission and brilliance are

	$B = 0$	$B = 1$
$A = 0$	680	10
$A = 1$	220	90

Then:

$p(A = 1, B = 0)$	$220/1000$
$p(B = 1)$	$100/1000$
$p(A = 0)$	$690/1000$
$p(B = 1 \mid A = 1)$	$90/310$
$p(A = 0 \mid B = 0)$	$680/900$

Review Exercise

Suppose I go through the records for $N = 1000$ students, checking their admission status, $A = \{0, 1\}$, and whether they are "brilliant" or not, $B = \{0, 1\}$.

(Aside: "Brilliance" is a dodgy concept, and does not predict scientific achievement as well as persistence and combinatorial ability; see e.g. Dean Simonton, *Scientific Genius: A Psychology of Science*, Cambridge University Press, 2009; this is just a toy example!)

Say that the counts for admission and brilliance are

	$B = 0$	$B = 1$
$A = 0$	680	10
$A = 1$	220	90

Then:

$p(A = 1, B = 0)$	$220/1000$	Joint
$p(B = 1)$	$100/1000$	
$p(A = 0)$	$690/1000$	
$p(B = 1 \mid A = 1)$	$90/310$	
$p(A = 0 \mid B = 0)$	$680/900$	

Review Exercise

Suppose I go through the records for $N = 1000$ students, checking their admission status, $A = \{0, 1\}$, and whether they are "brilliant" or not, $B = \{0, 1\}$

(Aside: "Brilliance" is a dodgy concept, and does not predict scientific achievement as well as persistence and combinatorial ability; see e.g. Dean Simonton, *Scientific Genius: A Psychology of Science*, Cambridge University Press, 2009; this is just a toy example!)

Say that the counts for admission and brilliance are

	$B = 0$	$B = 1$
$A = 0$	680	10
$A = 1$	220	90

Then:

$p(A = 1, B = 0)$	$220/1000$	Joint
$p(B = 1)$	$100/1000$	Marginal
$p(A = 0)$	$690/1000$	
$p(B = 1 \mid A = 1)$	$90/310$	
$p(A = 0 \mid B = 0)$	$680/900$	

Review Exercise

Suppose I go through the records for $N = 1000$ students, checking their admission status, $A = \{0, 1\}$, and whether they are "brilliant" or not, $B = \{0, 1\}$

(Aside: "Brilliance" is a dodgy concept, and does not predict scientific achievement as well as persistence and combinatorial ability; see e.g. Dean Simonton, *Scientific Genius: A Psychology of Science*, Cambridge University Press, 2009; this is just a toy example!)

Say that the counts for admission and brilliance are

	$B = 0$	$B = 1$
$A = 0$	680	10
$A = 1$	220	90

Then:

$p(A = 1, B = 0)$	$220/1000$	Joint
$p(B = 1)$	$100/1000$	Marginal
$p(A = 0)$	$690/1000$	Marginal
$p(B = 1 \mid A = 1)$	$90/310$	
$p(A = 0 \mid B = 0)$	$680/900$	

Review Exercise

Suppose I go through the records for $N = 1000$ students, checking their admission status, $A = \{0, 1\}$, and whether they are "brilliant" or not, $B = \{0, 1\}$.

(Aside: "Brilliance" is a dodgy concept, and does not predict scientific achievement as well as persistence and combinatorial ability; see e.g. Dean Simonton, *Scientific Genius: A Psychology of Science*, Cambridge University Press, 2009; this is just a toy example!)

Say that the counts for admission and brilliance are

	$B = 0$	$B = 1$
$A = 0$	680	10
$A = 1$	220	90

Then:

$p(A = 1, B = 0)$	$220/1000$	Joint
$p(B = 1)$	$100/1000$	Marginal
$p(A = 0)$	$690/1000$	Marginal
$p(B = 1 \mid A = 1)$	$90/310$	Conditional
$p(A = 0 \mid B = 0)$	$680/900$	

Review Exercise

Suppose I go through the records for $N = 1000$ students, checking their admission status, $A = \{0, 1\}$, and whether they are “brilliant” or not, $B = \{0, 1\}$

(Aside: “Brilliance” is a dodgy concept, and does not predict scientific achievement as well as persistence and combinatorial ability; see e.g. Dean Simonton, *Scientific Genius: A Psychology of Science*, Cambridge University Press, 2009; this is just a toy example!)

Say that the counts for admission and brilliance are

	$B = 0$	$B = 1$
$A = 0$	680	10
$A = 1$	220	90

Then:

$p(A = 1, B = 0)$	$220/1000$	Joint
$p(B = 1)$	$100/1000$	Marginal
$p(A = 0)$	$690/1000$	Marginal
$p(B = 1 \mid A = 1)$	$90/310$	Conditional
$p(A = 0 \mid B = 0)$	$680/900$	Conditional

This time

- More on joint, marginal and conditional distributions

- When can we say that X, Y do not influence each other?

- What, if anything, does $p(X = x|Y = y)$ tell us about $p(Y = y|X = x)$?

This time

- More on joint, marginal and conditional distributions

- When can we say that X, Y do not influence each other?

- What, if anything, does $p(X = x|Y = y)$ tell us about $p(Y = y|X = x)$?

Philosophically related to “How do we know / learn about the world?”

This time

- More on joint, marginal and conditional distributions

- When can we say that X, Y do not influence each other?

- What, if anything, does $p(X = x|Y = y)$ tell us about $p(Y = y|X = x)$?

Philosophically related to "How do we know / learn about the world?"

I am *not* providing a general answer; but keep it in mind!

Outline

1. More on Joint, Marginal and Conditional Distributions
2. Statistical Independence
3. Bayes' Theorem
4. Wrapping up

1. More on Joint, Marginal and Conditional Distributions

2. Statistical Independence

3. Bayes' Theorem

4. Wrapping up

Document Modelling Example

Suppose we have a large document of English text, represented as a sequence of characters:

$$x_1 x_2 x_3 \dots x_N$$

- e.g. `hello␣how␣are␣you`

Treat each consecutive pair of characters as the outcome of "random variables" X, Y, i.e.

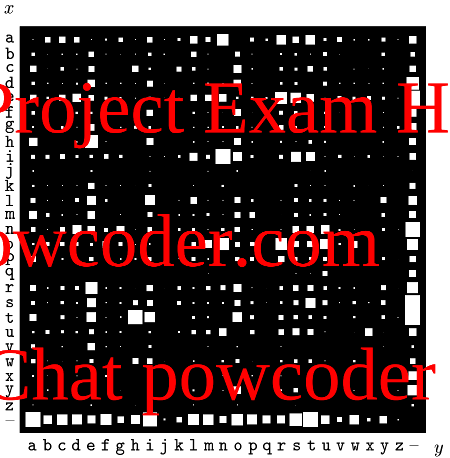

```
X = 'h', Y = 'e'

X = 'e', Y = 'l'

X = 'l', Y = 'l'
```

$$\vdots$$

Document Modelling: Marginal and Joint Distributions

i	a_i	p_i
1	a	0.0575
2	b	0.0128
3	c	0.0263
4	d	0.0285
5	e	0.0913
6	f	0.0173
7	g	0.0133
8	h	0.0313
9	i	0.0599
10	j	0.0006
11	k	0.0084
12	l	0.0335
13	m	0.0235
14	n	0.0596
15	o	0.0689
16	p	0.0192
17	q	0.0008
18	r	0.0508
19	s	0.0567
20	t	0.0706
21	u	0.0334
22	v	0.0069
23	w	0.0119
24	x	0.0073
25	y	0.0164
26	z	0.0007
27	–	0.1928

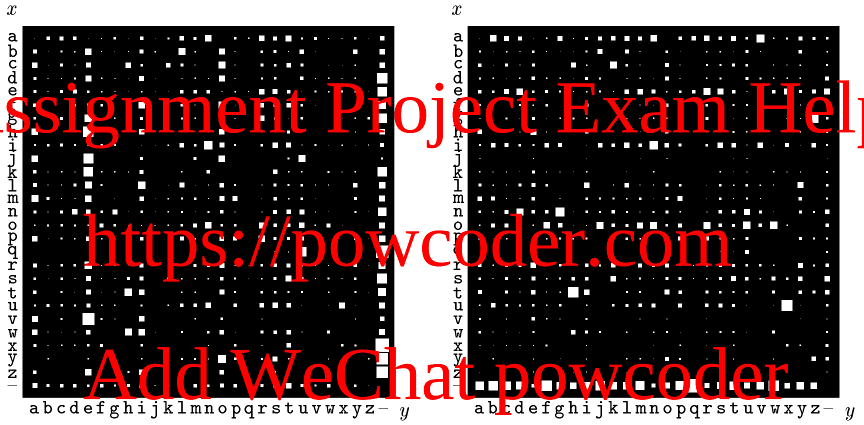
Unigram / Monogram



Bigram

Marginal and joint distributions for English alphabet, estimated from the "FAQ manual for Linux". Figure from Mackay (ITILA, 2003); areas of squares proportional to probability (the right way to do it!).

Document Modelling: Conditional Distributions



(a) $P(y \mid x)$ (b) $P(x \mid y)$

Conditional distributions for English alphabet, estimated from the "FAQ manual for Linux". Are these distributions "symmetric"? Figure from Mackay (ITILA, 2003)

$P(X = x|Y = y) = P(Y = y|X = x)$? $P(X = x|Y = y) = P(X = y|Y = x)$?.

Recap: Sum and Product Rules

Sum rule:

$$p(X = x_i) = \sum_j p(X = x_i, Y = y_j)$$

Product rule:

$$p(X = x_i, Y = y_j) = p(Y = y_j | X = x_i) p(X = x_i)$$

Relating the Marginal, Conditional and Joint

Suppose we knew $p(X = x, Y = y)$ for all values of x, y. Could we compute all of $p(X = x|Y = y)$, $p(X = x)$ and $p(Y = y)$?

Relating the Marginal, Conditional and Joint

Suppose we knew $p(X = x, Y = y)$ for all values of x, y. Could we compute all of $p(X = x | Y = y)$, $p(X = x)$ and $p(Y = y)$? Yes.

Relating the Marginal, Conditional and Joint

Suppose we knew $p(X = x, Y = y)$ for all values of x, y. Could we compute all of $p(X = x|Y = y)$, $p(X = x)$ and $p(Y = y)$? Yes.

Now suppose we knew $p(X = x)$ and $p(Y = y)$ for all values of x, y. Could we compute $p(X = x, Y = y)$ or $p(X = x|Y = y)$?

Relating the Marginal, Conditional and Joint

Suppose we knew $p(X = x, Y = y)$ for all values of x, y. Could we compute all of $p(X = x|Y = y)$, $p(X = x)$ and $p(Y = y)$? Yes.

Now suppose we knew $p(X = x)$ and $p(Y = y)$ for all values of x, y. Could we compute $p(X = x, Y = y)$ or $p(X = x|Y = y)$? No.

Relating the Marginal, Conditional and Joint

Suppose we knew $p(X = x, Y = y)$ for all values of x, y. Could we compute all of $p(X = x|Y = y)$, $p(X = x)$ and $p(Y = y)$? Yes.

Now suppose we knew $p(X = x)$ and $p(Y = y)$ for all values of x, y. Could we compute $p(X = x, Y = y)$ or $p(X = x|Y = y)$? No.

Relating the Marginal, Conditional and Joint

Suppose we knew $p(X = x, Y = y)$ for all values of x, y. Could we compute all of $p(X = x|Y = y)$, $p(X = x)$ and $p(Y = y)$? Yes.

Now suppose we knew $p(X = x)$ and $p(Y = y)$ for all values of x, y. Could we compute $p(X = x, Y = y)$ or $p(X = x|Y = y)$? No.

The difference in answers above is of great significance

Relating the Marginal, Conditional and Joint

Suppose we knew $p(X = x, Y = y)$ for all values of x, y. Could we compute all of $p(X = x|Y = y)$, $p(X = x)$ and $p(Y = y)$? Yes.

Now suppose we knew $p(X = x)$ and $p(Y = y)$ for all values of x, y. Could we compute $p(X = x, Y = y)$ or $p(X = x|Y = y)$? No.

The difference in answers above is of great significance

	$B = 0$	$B = 1$
$A = 0$	680	10
$A = 1$	220	90

	$B = 0$	$B = 1$
$A = 0$	640	50
$A = 1$	260	50

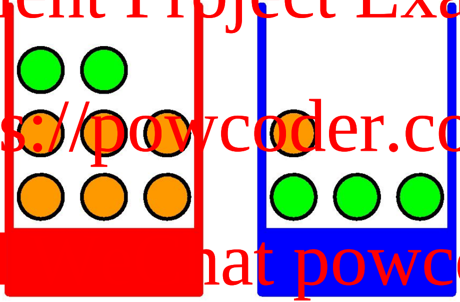
These have the same marginals, but different joint distributions

Joint as the "Master" Distribution

In general, there can be many consistent joint distributions for a given set of marginal distributions

The joint distribution is the "master" source of information about the dependence

Recall: Fruit-Box Experiment



Suppose that both boxes (red and blue) contain the same proportion of apples and oranges.

If fruit is selected uniformly at random from each box:

$$p(F = \mathtt{a} | B = \mathtt{r}) = p(F = \mathtt{a} | B = \mathtt{b}) \quad (= p(F = \mathtt{a}))$$
$$p(F = \mathtt{o} | B = \mathtt{r}) = p(F = \mathtt{o} | B = \mathtt{b}) \quad (= p(F = \mathtt{o}))$$

Statistical Independence

Suppose that both boxes (red and blue) contain the same proportion of apples and oranges.

If fruit is selected uniformly at random from each box:

$$p(F = \mathtt{a}|B = \mathtt{r}) = p(F = \mathtt{a}|B = \mathtt{b}) \quad (= p(F = \mathtt{a}))$$
$$p(F = \mathtt{o}|B = \mathtt{r}) = p(F = \mathtt{o}|B = \mathtt{b}) \quad (= p(F = \mathtt{o}))$$

The probability of selecting an apple (or an orange) is independent of the box that is chosen.

We may study the properties of F and B separately: this often simplifies analysis

Statistical Independence: Definition

Definition: Independent Variables

Two variables X and Y are statistically independent, denoted $X \perp\!\!\!\perp Y$, if and only if their joint distribution *factorizes* into the product of their marginals:

$$X \perp\!\!\!\perp Y \leftrightarrow p(X, Y) = p(X)p(Y)$$

This definition generalises to more than two variables.

Statistical Independence: Definition

Definition: Independent Variables

Two variables X and Y are statistically independent, denoted $X \perp\!\!\!\perp Y$, if and only if their joint distribution *factorizes* into the product of their marginals:

$$X \perp\!\!\!\perp Y \leftrightarrow p(X, Y) = p(X)p(Y)$$

This definition generalises to more than two variables.

Are the variables in the language example statistically independent?

A Note on Notation

When we write

$$p(X, Y) = p(X)p(Y)$$

we have not specified the outcomes of X, Y explicitly

This statement is a shorthand for

$$p(X = x, Y = y) = p(X = x)p(Y = y)$$

for every possible x and y

This notation is sometimes called implied universality

Conditional independence

We may also consider random variables that are conditionally independent given some other variable.

Definition: Conditionally Independent Variables

Two variables X and Y are conditionally independent given Z, denoted $X \perp\!\!\!\perp Y|Z$, if and only if

$$p(X, Y|Z) = p(X|Z)p(Y|Z)$$

Intuitively, Z is a common cause for X and Y.

Example: X = whether I have a cold

Y = whether I have a headache

Z = whether I have the flu

Revisiting the Product Rule

The product rule tells us:

$$p(X, Y) = p(Y|X)p(X)$$

This can equivalently be interpreted as a *definition* of conditional probability:

$$p(Y|X) = \frac{p(X, Y)}{p(X)}$$

Can we use these to relate $p(X|Y)$ and $p(Y|X)$?

Posterior Inference:

Example 1 (Mackay, 2003)

- Dicksy Sick had a test for a rare disease
  - Only 1% people of Dicksy's background have the disease

Posterior Inference:

Example 1 (Mackay, 2003)

- Dicksy Sick had a test for a rare disease
  - Only 1% people of Dicksy's background have the disease
- The test simply classifies a person as having the disease, or not

Posterior Inference:

Example 1 (Mackay, 2003)

- Dicksy Sick had a test for a rare disease
  - Only 1% people of Dicksy's background have the disease
- The test simply classifies a person as having the disease, or not

- The test is reliable, but not infallible
  - It correctly identifies a sick individual 95% of the time
    $p(\text{identifies sick} \mid \text{sick}) = 95\%$.

  - It correctly identifies a healthy individual 96% of the time
    $p(\text{identifies healthy} \mid \text{healthy}) = 96\%$.

Posterior Inference:

Example 1 (Mackay, 2003)

- Dicksy Sick had a test for a rare disease
  - Only 1% people of Dicksy's background have the disease
- The test simply classifies a person as having the disease, or not
- The test is reliable, but not infallible
  - It correctly identifies a sick individual 95% of the time
    $p(\text{identifies sick} \mid \text{sick}) = 95\%$.
  - It correctly identifies a healthy individual 96% of the time
    $p(\text{identifies healthy} \mid \text{healthy}) = 96\%$.
- Dicksy has tested positive (apparently sick)

Posterior Inference:

Example 1 (Mackay, 2003)

- Dicksy Sick had a test for a rare disease
  - Only 1% people of Dicksy's background have the disease
- The test simply classifies a person as having the disease, or not
- The test is reliable, but not infallible
  - It correctly identifies a sick individual 95% of the time $p(\text{identifies sick} \mid \text{sick}) = 95\%$.
  - It correctly identifies a healthy individual 96% of the time $p(\text{identifies healthy} \mid \text{healthy}) = 96\%$.
- Dicksy has tested positive (apparently sick)
- What is the probability of Dicksy having the disease?

Let $D \in \{0, 1\}$ denote whether Dicksy has the disease, and $T \in \{0, 1\}$ the outcome of the test:

Posterior Inference:

Example 1: Formalization

Let $D \in \{0, 1\}$ denote whether Dicksy has the disease, and $T \in \{0, 1\}$ the outcome of the test:

$$p(D = 1) = 0.01 \qquad p(D = 0) = 0.99$$

$$p(T = 1|D = 1) = 0.95 \qquad p(T = 1|D = 0) = 0.04$$

$$p(T = 0|D = 1) = 0.05 \qquad p(T = 0|D = 0) = 0.96$$

Posterior Inference:

Example 1: Formalization

Let $D \in \{0, 1\}$ denote whether Dicksy has the disease, and $T \in \{0, 1\}$ the outcome of the test:

$$
\begin{aligned}
p(D = 1) &= 0.01 & p(D = 0) &= 0.99 \\
p(T = 1|D = 1) &= 0.95 & p(T = 1|D = 0) &= 0.04 \\
p(T = 0|D = 1) &= 0.05 & p(T = 0|D = 0) &= 0.96
\end{aligned}
$$

We need to compute $p(D = 1|T = 1)$, *the probability of Dicksy having the disease given that the test has resulted positive.*

Posterior Inference:

Example 1: Solution

$$p(D = 1 | T = 1) = \frac{p(D = 1, T = 1)}{p(T = 1)}$$ Def. conditional prob.

Posterior Inference:

Example 1: Solution

$$
\begin{aligned}
p(D = 1 \mid T = 1) &= \frac{p(D = 1, T = 1)}{p(T = 1)} && \text{Def. conditional prob.} \\
&= \frac{p(T = 1, D = 1)}{p(T = 1)} && \text{Symmetry}
\end{aligned}
$$

Posterior Inference:

Example 1: Solution

$$
\begin{aligned}
p(D = 1 | T = 1) &= \frac{p(D = 1, T = 1)}{p(T = 1)} && \text{Def. conditional prob.} \\
&= \frac{p(T = 1, D = 1)}{p(T = 1)} && \text{Symmetry} \\
&= \frac{p(T = 1 | D = 1) p(D = 1)}{p(T = 1)} && \text{Product rule}
\end{aligned}
$$

Posterior Inference:

Example 1: Solution

$$
\begin{aligned}
p(D=1|T=1) &= \frac{p(D=1,T=1)}{p(T=1)} && \text{Def. conditional prob.} \\
&= \frac{p(T=1,D=1)}{p(T=1)} && \text{Symmetry} \\
&= \frac{p(T=1|D=1)p(D=1)}{p(T=1)} && \text{Product rule} \\
&= \frac{p(T=1|D=1)p(D=1)}{\sum_d p(T=1|D=d)p(D=d)} && \text{Sum rule}
\end{aligned}
$$

Posterior Inference:

Example 1: Solution

$$
\begin{aligned}
p(D=1|T=1) &= \frac{p(D=1, T=1)}{p(T=1)} && \text{Def. conditional prob.} \\
&= \frac{p(T=1, D=1)}{p(T=1)} && \text{Symmetry} \\
&= \frac{p(T=1|D=1)p(D=1)}{p(T=1)} && \text{Product rule} \\
&= \frac{p(T=1|D=1)p(D=1)}{\sum_d p(T=1|D=d)p(D=d)} && \text{Sum rule} \\
&= \frac{p(T=1|D=1)p(D=1)}{p(T=1|D=1)p(D=1)+p(T=1|D=0)p(D=0)}
\end{aligned}
$$

Posterior Inference:

Example 1: Solution

$$
\begin{aligned}
p(D=1|T=1) &= \frac{p(D=1,T=1)}{p(T=1)} && \text{Def. conditional prob.} \\
&= \frac{p(T=1,D=1)}{p(T=1)} && \text{Symmetry} \\
&= \frac{p(T=1|D=1)p(D=1)}{p(T=1)} && \text{Product rule} \\
&= \frac{p(T=1|D=1)p(D=1)}{\sum_d p(T=1|D=d)p(D=d)} && \text{Sum rule} \\
&= \frac{p(T=1|D=1)p(D=1)}{p(T=1|D=1)p(D=1)+p(T=1|D=0)p(D=0)} \\
&\approx 0.19.
\end{aligned}
$$

Posterior Inference:

Example 1: Solution

$$
\begin{aligned}
p(D=1|T=1) &= \frac{p(D=1, T=1)}{p(T=1)} && \text{Def. conditional prob.} \\
&= \frac{p(T=1, D=1)}{p(T=1)} && \text{Symmetry} \\
&= \frac{p(T=1|D=1)p(D=1)}{p(T=1)} && \text{Product rule} \\
&= \frac{p(T=1|D=1)p(D=1)}{\sum_d p(T=1|D=d)p(D=d)} && \text{Sum rule} \\
&= \frac{p(T=1|D=1)p(D=1)}{p(T=1|D=1)p(D=1) + p(T=1|D=0)p(D=0)} \\
&\approx 0.19.
\end{aligned}
$$

Despite testing positive and the high accuracy of the test, the probability of Dicksy having the disease is only 0.19!

Why is the Probability So Low?

A “Natural Frequency” Approach

In 100 people, only 1 is expected to have the disease ($p(D = 1) = 0.01$)

This sick person will most likely test positive ($p(T = 1|D = 1) = 0.95$)

But around 4 healthy people are expected to be wrongly flagged as sick ($p(T = 1|D = 0) = 0.04$, and $0.04 \times 99 \approx 4$)

So when the test is positive, the chance of being sick is $\approx 1/5$

Why is the Probability So Low?

A “Natural Frequency” Approach

In 100 people, only 1 is expected to have the disease ($p(D=1)=0.01$)

This sick person will most likely test positive ($p(T=1|D=1)=0.95$)

But around 4 healthy people are expected to be wrongly flagged as sick ($p(T=1|D=0)=0.04$, and $0.04 \times 99 \approx 4$)

So when the test is positive, the chance of being sick is $\approx 1/5$

(Aside: If you can correctly perform the calculation on the previous slide, you are doing better than most medical doctors! See Gird Gigerenzer and Adrian Edwards, Simple tools for understanding risks: from innumeracy to insight, *British Medical Journal*, 327(7417), 741–744, 27 September 2003; Gird Gigerenzer, *Reckoning with risk: Learning to live with uncertainty*, Penguin, 2002.

Moral of the story — if you get sick, don’t delegate conditional probability computations to your doctor!)

Bayes' Theorem

We have implicitly used the following (at first glance remarkable) fact:

Bayes' Theorem:

$$p(Z|X) = \frac{p(Z,X)}{p(X)} = \frac{p(X,Z)}{p(X)} = \frac{p(X|Z)p(Z)}{p(X)} = \frac{p(X|Z)p(Z)}{\sum_{Z'} p(X|Z')p(Z')}$$

If we can express what knowledge of X (test) tells us about Z (disease), then we can express what knowledge of Z tells us about X

The Bayesian Inference Framework

Bayesian Inference

Bayesian inference provides a mathematical framework explaining how to change our (prior) beliefs in the light of new evidence.

$$\underbrace{p(Z|X)}_{\text{posterior}} = \frac{\overbrace{p(X|Z)}^{\text{likelihood}} \times \overbrace{p(Z)}^{\text{prior}}}{\underbrace{p(X)}_{\text{evidence}}}$$

Prior: Belief that someone is sick

Likelihood: Probability of testing positive given you are sick

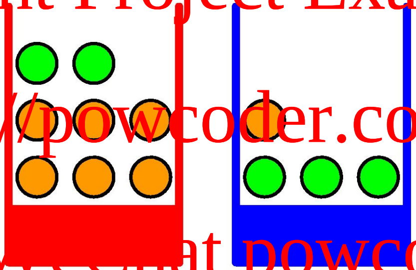
Posterior: Probability of being sick given you test positive

Posterior Inference:

Example 2 (Bishop, 2006)

Recall our fruit-box example:

- The proportion of oranges and apples are given by

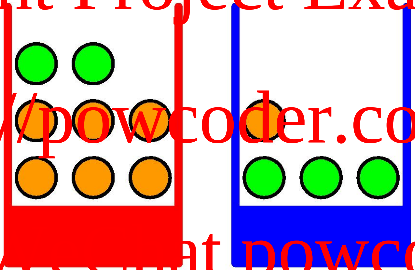


Posterior Inference:

Example 2 (Bishop, 2006)

Recall our fruit-box example:

- The proportion of oranges and apples are given by



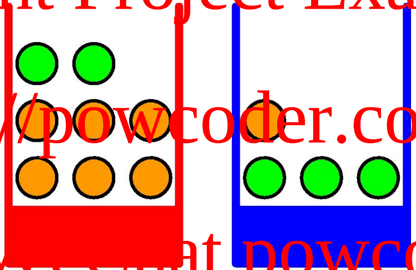
- Someone told us that in a previous experiment they ended up picking up the red box 40% of the time and the blue box 60% of the time.

Posterior Inference:

Example 2 (Bishop, 2006)

Recall our fruit-box example:

- The proportion of oranges and apples are given by



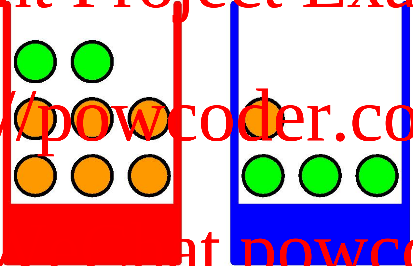
- Someone told us that in a previous experiment they ended up picking up the red box 40% of the time and the blue box 60% of the time.
- A piece of fruit has been picked up and it turned out to be an orange.

Posterior Inference:

Example 2 (Bishop, 2006)

Recall our fruit-box example:

- The proportion of oranges and apples are given by



- Someone told us that in a previous experiment they ended up picking up the red box 40% of the time and the blue box 60% of the time.
- A piece of fruit has been picked up and it turned out to be an orange.
- What is the probability that it came from the red box?

Posterior Inference:

Example 2: Formalization

Let $B \in \{r, b\}$ denote the selected box and $F \in \{a, o\}$ the selected fruit.

Posterior Inference:

Example 2: Formalization

Let $B \in \{r, b\}$ denote the selected box and $F \in \{a, o\}$ the selected fruit.

$$p(B = r) = 4/10 \qquad p(B = b) = 6/10$$

$$p(F = a|B = r) = 1/4 \qquad p(F = o|B = r) = 3/4$$

$$p(F = a|B = b) = 3/4 \qquad p(F = o|B = b) = 1/4$$

Posterior Inference:

Example 2: Formalization

Let $B \in \{r, b\}$ denote the selected box and $F \in \{a, o\}$ the selected fruit.

$$p(B = r) = 4/10 \qquad p(B = b) = 6/10$$

$$p(F = a|B = r) = 1/4 \qquad p(F = o|B = r) = 3/4$$

$$p(F = a|B = b) = 3/4 \qquad p(F = o|B = b) = 1/4$$

We need to compute $p(B = r|F = o)$, *the probability that a picked up orange came from the red box.*

Posterior Inference:
Example 2: Solution

We simply use Bayes' rule:

Posterior Inference:

Example 2: Solution

We simply use Bayes' rule:

$$p(B = r|F = o) = \frac{p(F = o|B = r)p(B = r)}{p(F = o)}$$

Posterior Inference:

Example 2: Solution

We simply use Bayes' rule:

$$p(B = r|F = o) = \frac{p(F = o|B = r)p(B = r)}{p(F = o)}$$
$$= \frac{p(F = o|B = r)p(B = r)}{p(F = o|B = r)p(B = r) + p(F = o|B = b)p(B = b)}$$

Posterior Inference:

Example 2: Solution

We simply use Bayes' rule:

$$p(B = r|F = o) = \frac{p(F = o|B = r)p(B = r)}{p(F = o)}$$

$$= \frac{p(F = o|B = r)p(B = r)}{p(F = o|B = r)p(B = r) + p(F = o|B = b)p(B = b)}$$

$$= \frac{2}{3}$$

Posterior Inference:

Example 2: Solution

We simply use Bayes' rule:

$$
\begin{aligned}
p(B = r|F = o) &= \frac{p(F = o|B = r)p(B = r)}{p(F = o)} \\
&= \frac{p(F = o|B = r)p(B = r)}{p(F = o|B = r)p(B = r) + p(F = o|B = b)p(B = b)} \\
&= \frac{2}{3}
\end{aligned}
$$

and therefore $p(B = b|F = o) = 1/3$.

Posterior Inference:

Example 2: Interpretation of the Solution

- If we hadn't been told any information about the fruit picked, the blue box is more likely to be selected than the red box

Posterior Inference:

Example 2: Interpretation of the Solution

- If we hadn't been told any information about the fruit picked, the blue box is more likely to be selected than the red box
  - *A priori* we have $p(B = r) = 4/10$ and $p(B = b) = 6/10$

Posterior Inference:

Example 2: Interpretation of the Solution

- If we hadn't been told any information about the fruit picked, the blue box is more likely to be selected than the red box
  - *A priori* we have $p(B = r) = 4/10$ and $p(B = b) = 6/10$

- Once we get new information that an orange has been picked, this increases the probability of the selected box being the red one

Posterior Inference:

Example 2: Interpretation of the Solution

- If we hadn't been told any information about the fruit picked, the blue box is more likely to be selected than the red box
  - *A priori* we have $p(B = r) = 4/10$ and $p(B = b) = 6/10$

- Once we get new information that an orange has been picked, this increases the probability of the selected box being the red one
  - *Because the red box contains more oranges than the blue box*

Posterior Inference:

Example 2: Interpretation of the Solution

- If we hadn't been told any information about the fruit picked, the blue box is more likely to be selected than the red box
  - *A priori* we have $p(B = r) = 4/10$ and $p(B = b) = 6/10$

- Once we get new information that an orange has been picked, this increases the probability of the selected box being the red one
  - *Because the red box contains more oranges than the blue box*
- In fact, the proportion of oranges is so much higher in the red box that this is strong evidence that the orange came from it

Posterior Inference:

Example 2: Interpretation of the Solution

- If we hadn't been told any information about the fruit picked, the blue box is more likely to be selected than the red box
  - *A priori* we have $p(B = r) = 4/10$ and $p(B = b) = 6/10$

- Once we get new information that an orange has been picked, this increases the probability of the selected box being the red one
  - *Because the red box contains more oranges than the blue box*

- In fact, the proportion of oranges is so much higher in the red box that this is strong evidence that the orange came from it
  - *So after picking up the orange the red box is much more likely to have been selected than the blue one*

Summary

- Recap on joint, marginal and conditional distributions

- Interpretation of conditional probability

- Statistical Independence
- Bayes' rule: combination of prior, likelihood to get a posterior

- Reading: Mackay § 2.1, § 2.2 and § 2.3

Homework Exercise

Suppose we know that random variables X, Y satisfy

$$p(X|Y) = p(Y|X)$$

What can you conclude about the relationship between X and Y?

If X and Y are independent, does that imply $p(X|Y) = p(Y|X)$?

Repeat the above questions for the statement

$$\frac{p(X|Y)}{p(Y|X)} = \frac{p(X)}{p(Y)}$$

Next time

- More examples on Bayes' theorem:
  - Eating hamburgers
  - Detecting terrorists
  - The Monty Hall problem
  - Document modelling
- Are there notions of probability beyond frequency counting?