Conditional probability

And

Bayes theorem

Conditional probablity

Given events E and F, often we are interested in statements like

if even E has occurred, then the probability of F is ...

Some examples:

- *Roll two dice:* what is the probability that the sum of faces is 6 *given that* the first face is 4?
- *Gene expressions:* What is the probability that gene A is switched off (e.g. down-regulated) given that gene B is also switched off?

This **conditional probability** can be derived following a similar construction:

- Repeat the experiment N times.
- Count the number of times event E occurs, $N(E)$, and the number of times **both** E and F occur jointly, $N(E \cap F)$. Hence $N(E) \leq N$
- The proportion of times that F occurs in this **reduced** space is

$$\frac{N(E \cap F)}{N(E)}$$

since E occurs at each one of them.

- Now note that the ratio above can be re-written as the ratio between two (unconditional) probabilities

$$\frac{N(E \cap F)}{N(E)} = \frac{N(E \cap F)/N}{N(E)/N}$$

- Then *the probability of* F, given that E has occurred should be defined as

$$\frac{P(E \cap F)}{P(E)}$$

The definition of Conditional Probability

The **conditional probability** of an event F, given that an event E has occurred, is defined as

$$P(F|E) = \frac{P(E \cap F)}{P(E)}$$

and is defined only if $P(E) > 0$.

Note that, if E has occurred, then

- $F|E$ is a point in the set $P(E \cap F)$
- E is the new sample space

it can be proved that the function $P(\cdot|\cdot)$ defyning a conditional probability also satisfies the three probability axioms.

Example. Roll a die

Let $A = \{score\ an\ even\ number\}$ and $B = \{score\ a\ number\ \geq\ 3\}$.

$$P(A) = \frac{1}{2}, \quad P(B) = \frac{2}{3}, \quad P(A \cap B) = \frac{1}{3}$$

because the intersection has only two elements, then

$P(A/B) = \frac{P(A \cap B)}{P(B)} = \frac{1/3}{2/3} = \frac{1}{2}$

$P(B/A) = \frac{P(B \cap A)}{P(A)} = \frac{1/3}{1/2} = \frac{2}{3}$

$$P(A/B) \neq P(B/A)$$

Conditional probabilities behave like ordinary probabilities

Standard results for probability extend to the conditional probability, such that conditional probabilities behave like ordinary probabilities.

For example, for events A and B

$$P(\bar{A}|B) = 1 - P(A|B)$$

In order to prove this, first decompose B as

$$B = (A \cap B) \cup (\bar{A} \cap B) \longrightarrow P(B) = P(A \cap B) + P(\bar{A} \cap B)$$

because they are mutually exclusive. Then

$$P(\bar{A} \cap B) = P(B) - P(A \cap B)$$

divide both sides by $P(B)$

$$P(\bar{A}|B) = 1 - \frac{P(A \cap B)}{P(B)} = 1 - P(A|B)$$

Joint probability

The definition of conditional probability also provides us with a working definition of joint probability, i.e. the probability that two events occur jointly.

The probability that both events E and F occur is

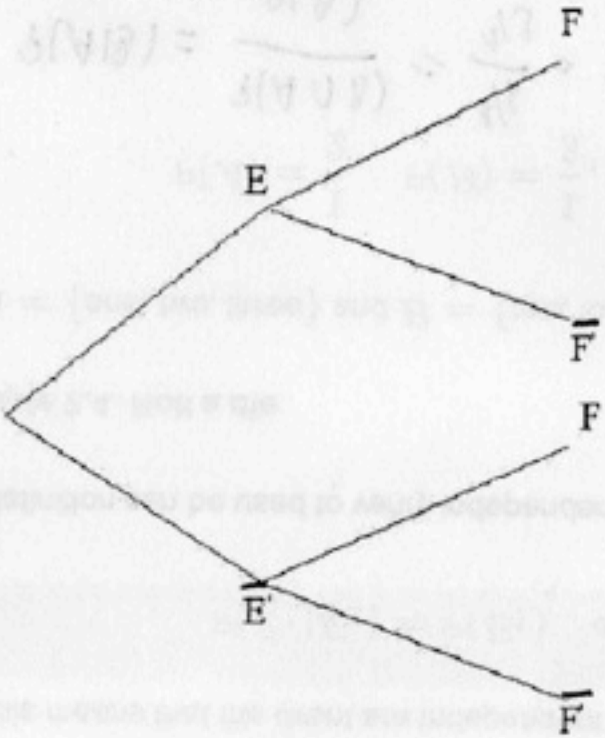
$$\mathrm{P}(E \cap F) = \mathrm{P}(F|E)\mathrm{P}(E)$$

which helps solve many probability problems.

2.7 The chain rule

The chain rule

- Given two events E and F, a simple way to think about joint and conditional probability is via a **probability tree**



$$P(E \cap F) = P(E) \times P(F \mid E)$$

$$P(E \cap \overline{F}) = P(E) \times P(\overline{F} \mid E)$$

$$P(\overline{E} \cap F) = P(\overline{E}) \times P(F \mid \overline{E})$$

$$P(\overline{E} \cap \overline{F}) = P(\overline{E}) \times P(\overline{F} \mid \overline{E})$$

More generally, given events $E_1, \ldots, E_k$, the probability of their intersection is given by

$$\mathsf{P}(E_1 \cap \ldots \cap E_k) = \mathsf{P}(E_1)\mathsf{P}(E_2|E_1)\mathsf{P}(E_3|E_1 \cap E_2)\ldots \\ \mathsf{P}(E_k|E_1 \cap E_2 \cap \ldots \cap E_{k-1})$$

which is called the **chain rule**.

2.8 Contingency table

Suppose we have classified 161 individuals according to two criteria: gender and age (below or above 30). A convenient way to present the results is to use a **contingency table.**

- Select one person at random from a group with distribution represented in the contingency table below

	male	female	total
under 30	54	47	101
over 30	28	32	60
total	82	79	161

- Define the following two events and their probabilities:

$$E_1 = \{\textit{under 30}\} \qquad \mathrm{P}(E_1) = \frac{101}{161} = 0.627$$

$$E_2 = \{\textit{female}\} \qquad \mathrm{P}(E_2) = \frac{79}{161} = 0.490$$

The **joint probability** of the event $E_1 \cap E_2$ is given by

$$P(E_1 \cap E_2) = \frac{47}{161} = 0.291$$

2.9 Contingency table 2

Suppose we have classified 161 individuals according to two criteria: gender and age (below or above 30). A convenient way to present the results is to use a **contingency table.**

- Select one person at random from a group with distribution represented in the contingency table below

	male	female	total
under 30	54	47	101
over 30	28	32	60
total	82	79	161

- Define the following two events and their probabilities:

$$E_1 = \{\textit{under 30}\} \qquad P(E_1) = \frac{101}{161} = 0.627$$

$$E_2 = \{\textit{female}\} \qquad P(E_2) = \frac{79}{161} = 0.490$$

Suppose that E_2 has been observed first. The **conditional probability** that a randomly picked female has age under 30 is

$$P(E_1|E_2) = \frac{P(E_1 \cap E_2)}{P(E_2)} = \frac{47/161}{79/161} = 0.594$$

Independence

We can use the definition of joint probability to assess whether two events are independent, i.e. when the occurrence of one event *does not affect* the probability of occurrence of another event.

- Two events E_1 and E_2 are *independent* if

$$\mathsf{P}(E_1 \cap E_2) = \mathsf{P}(E_1|E_2)\mathsf{P}(E_2) = \mathsf{P}(E_1)\mathsf{P}(E_2)$$

or, alternatively

$$\mathsf{P}(E_1 \cap E_2) = \mathsf{P}(E_2|E_1)\mathsf{P}(E_1) = \mathsf{P}(E_2)\mathsf{P}(E_1)$$

- This means that the event are independent if

$$\mathsf{P}(E_1|E_2) = \mathsf{P}(E_1) \quad \text{or} \quad \mathsf{P}(E_2|E_1) = \mathsf{P}(E_2)$$

The definition can be used to verify independence between events.

Example: Roll a die

Let $A = \{one, two, three\}$ and $B = \{two, four\}$. Are A and B independent?

Example: Roll a die

Let $A = \{one, two, three\}$ and $B = \{two, four\}$. Are A and B independent?

$$P(A) = \frac{1}{2}, \quad P(B) = \frac{1}{3}, \quad P(A \cap B) = \frac{1}{6}$$

then

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{1/6}{1/3} = \frac{1}{2} = P(A)$$

$$P(B|A) = \frac{P \cap A}{P(A)} = \frac{1/6}{1/2} = \frac{1}{3} = P(B)$$

Thus we conclude that A and B are independent.

2.15 Two letter sequence

- A short pattern contains only two letters from a {G,B} alphabet
- What is the probability that **both letters are B** given that **at least one is B,** regardless of the order?

- Write down the sample space first

$$S = \{GG, GB, BG, BB\}$$

- The question requires to compute P($BB|$ one B at least)

2.16 Two letter sequence

Assuming equal probabilities, i.e. 1/4, we have

$$P(BB|one\, B\, at\, least) = P(BB|GB \cup BG \cup BB) = \frac{P(BB \cap (GB \cup BG \cup \cup BB))}{P(GB \cup BG \cup BB)}$$

$$= \frac{P(BB)}{P(GB \cup BG \cup BB)} = \frac{1/4}{3/4} = \frac{1}{3}$$

A little variation of this question requires to condition on the fact that the second letter is a B (now order matters):

$$P(BB|second\, letter\, is\, B) = P(BB|GB \cup BB) = \frac{P(BB \cap (GB \cup BB))}{P(GB \cup BB)}$$

2.18 DNA sequence

- Consider the sequence

$$ATAGTAGATACGCACCGAGGA$$

consisting of 21 letters from the alphabet $\{A, T, G, C\}$.

- If we wish to assess the probability of observing this sequence, we might start assuming that

$$\mathsf{P}(A) = p_A \quad \mathsf{P}(C) = p_C \quad \mathsf{P}(G) = p_G \quad \mathsf{P}(T) = p_T$$

for some suitable probabilities satisfying

$$0 \leq p_A, p_C, p_G, p_T \leq 1 \qquad p_A + p_C + p_G + p_T = 1$$

- Under the **independence assumption**, the probability

$$\mathsf{P}(\{ATAGTAGATACGCACCGAGGA\})$$

can be factorized into the product

$$p_A \times p_T \times p_A \times \ldots \times p_G \times p_A$$

which simplifies to

$$p_A^8 p_C^4 p_G^6 p_T^3$$

- In cases such as this one, the independence assumption is often unrealistic but simplifies calculations – we only need 4 probabilities here.

- The assumption of independence of events may not be correct – indeed it is often unrealistic. It is usually adopted to keep computations easy.
- In the sequence example, we may assume that having observed a given letter in the current position may influence the probability of observing the subsequent letter
- In this case, the required probability

$$\mathsf{P}(ATAGTAGATACGCACCGAGGA)$$

can be written as

$$\mathsf{P}(A) \times \mathsf{P}(T|A) \times \mathsf{P}(A|AT) \times \mathsf{P}(G|ATA) \times \ldots$$
$$\ldots \times \mathsf{P}(A|ATAGTAGATACGCACCGAGG)$$

- Here we need to define and compute 21 **conditional** probabilities, whereas under the independence assumption we only needed the 4 **unconditional** probabilities (one for each base).

In order to build a working system, we need to randomly pick three components out of 100 available components, some of which are known to be defective. If any of the selected component does not work, then the system also does now work.

What is the probability of building a working system if we know that there are 10 faulty components?

- Call A_i the event that occur when component i is among those that are fully functional, where $i = 1, 2, 3$. Therefore

$$\text{P(system works)} = \text{P}(A_1 \cap A_2 \cap A_3)$$

- Using the chain rule, this can be written as

$$P(A_1 \cap A_2 \cap A_3) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2)$$

- Since 10 components are faulty,

$$P(\overline{A_1}) = \frac{10}{100} \quad \text{or} \quad P(A_1) = \frac{90}{100}$$

- If component 1 is among the functional ones, then component 2 will be one of the remaining 99, 89 of which are working, therefore

$$P(A_2|A_1) = \frac{89}{99}$$

- Similarly,

$$P(A_3|A_1 \cap A_2) = \frac{88}{98}$$

and the required probability is 0.726.

Often we need to compute conditional probabilities involving more than just one event, e.g. the probability and events A and B occur, given that C has occurred.

Example 2.13.

Show that

$$\mathsf{P}(A \cap B|C) = \mathsf{P}(A|B \cap C)\mathsf{P}(B|C)$$

Using the definition of conditional probability, we obtain:

$$\begin{aligned}\mathsf{P}(A|B \cap C)\mathsf{P}(B|C) &= \frac{\mathsf{P}(A \cap B \cap C)}{\mathsf{P}(B \cap C)} \frac{\mathsf{P}(B \cap C)}{\mathsf{P}(C)} \\ &= \frac{\mathsf{P}(A \cap B \cap C)}{\mathsf{P}(C)} = \mathsf{P}(A \cap B|C)\end{aligned}$$

Independence for more than two events

The events E_1, E_2 and E_3 are called *mutually independent* if they are independent *in pairs*, that is

$$\mathsf{P}(E_i \cap E_j) = \mathsf{P}(E_i)\mathsf{P}(E_j) \quad \forall i \neq j$$

and

$$\mathsf{P}(E_1 \cap E_2 \cap E_3) = \mathsf{P}(E_1)\mathsf{P}(E_2)\mathsf{P}(E_3)$$

- Note that three events may be independent *in pairs* but not be independent.

- The independence of n events can be defined inductively. Suppose we have defined independence of k events for every $k < n$. Then the events $E_1, \ldots, E_n$ are independent if any $k < n$ of them are independent and

$$\mathsf{P}(E_1 \cap E_2 \cdots \cap E_n) = \mathsf{P}(E_1)\mathsf{P}(E_2) \cdots \mathsf{P}(E_n)$$

2.26 Law of the total probability

- Suppose that F and $\bar{F}$ form a **partition** of the sample space

- Given an event E, we can write

$$E = (E \cap F) \cup (E \cap \bar{F})$$

(you may want to draw a Vann diagram to check this result)

- Note that, by construction, $(E \cap F)$ and $(E \cap \bar{F})$ are **mutually exclusive**

- Applying the definition of joint probability,

$$\begin{aligned} \mathsf{P}(E) &= \mathsf{P}(E \cap F) + \mathsf{P}(E \cap \bar{F}) \\ &= \mathsf{P}(E|F)P(F) + \mathsf{P}(E|\bar{F})\mathsf{P}(\bar{F}) \\ &= \mathsf{P}(E|F)P(F) + \mathsf{P}(E|\bar{F})(1 - \mathsf{P}(F)) \end{aligned}$$

- Note how $\mathsf{P}(E)$ has been expressed as a **weighted average of conditional probabilities** with weights given by the probabilities of the conditioning event

2.27 Probability table

With two events E and F we have:

	E	$\overline{E}$	Sum
F	$P(E \cap F)$	$P(\overline{E} \cap F)$	$P(F)$
$\overline{F}$	$P(E \cap \overline{F})$	$P(\overline{E} \cap \overline{F})$	$P(\overline{F})$
Sum	$P(E)$	$P(\overline{E})$	

Summing in the **columns**:

$$\mathrm{P}(E \cap F) + \mathrm{P}(E \cap \overline{F}) = \mathrm{P}(E)$$

$$\mathrm{P}(\overline{E} \cap F) + \mathrm{P}(\overline{E} \cap \overline{F}) = \mathrm{P}(\overline{E})$$

Summing in the **rows**:

$$\mathrm{P}(E \cap F) + \mathrm{P}(\overline{E} \cap F) = \mathrm{P}(F)$$

$$\mathrm{P}(E \cap \overline{F}) + \mathrm{P}(\overline{E} \cap \overline{F}) = \mathrm{P}(\overline{F})$$

A simple way to prove these results is by using Venn diagrams.

- More generally, assume that events $F_1, F_2, \cdots, F_n$ form a partition of the sample space S, i.e.

$$S = \bigcup_{i=1}^{n} F_i$$

and $F_i \cap F_j = \emptyset$ for all $i \neq j$.

- Then an event E in S can be expressed as

$$E = \cup_{i=1}^{n}(E \cap F_i)$$

- Using the fact that events $(E \cap F_i)$ are mutually exclusive,

$$\begin{aligned} \mathsf{P}(E) &= \mathsf{P}(\cup_i^n (E \cap F_i)) \\ &= \sum_i \mathsf{P}(E \cap F_i) \\ &= \sum_i \mathsf{P}(E|F_i)\mathsf{P}(F_i) \end{aligned}$$

- Also, if the event $G \subseteq S$ is such that $\mathsf{P}(G) > 0$, the conditional probability of E given G can be written as

$$\mathsf{P}(E|G) = \sum_i \mathsf{P}(E|F_i \cap G)\mathsf{P}(F_i|G)$$

2.28 General case: the law of total probability

2.29 High-throughput genotyping machine

A biotech company uses 3 high-throughput genotyping machines, say X, Y and Z to process a certain number of arrays.

Suppose that:

1. Machine X processes 50% of the arrays with a genotypying error rate of 3%
2. Machine Y processes 30% of the arrays with a genotypying error rate of 4%
3. Machine Z processes 20% of the arrays with a genotypying error rate of 5%

Compute the probability that a randomly selected array is erroneous

Let D denote the event that an array is erroneous.

By the law of total probability

2.30 High-througput genotyping machine

A biotech company uses 3 high-throughput genotyping machines, say X, Y and Z to process a certain number of arrays.

Suppose that:

1. Machine X processes 50% of the arrays with a genotypying error rate of 3%
2. Machine Y processes 30% of the arrays with a genotypying error rate of 4%
3. Machine Z processes 20% of the arrays with a genotypying error rate of 5%

Compute the probability that a randomly selected array is erroneous

Let D denote the event that an array is erroneous.

By the law of total probability

$$P(D) = \underbrace{P(D/X)\,P(X)}_{P(D\cap X)} + \underbrace{P(D/Y)\,P(Y)}_{P(D\cap Y)} + \underbrace{P(D/Z)\,P(Z)}_{P(D\cap Z)}$$

$$= 0.03\cdot 0.5 + 0.04\cdot 0.3 + 0.05\cdot 0.2 = 0.037$$

- Given two events E and F, the joint probabilities can be written as

$$\mathrm{P}(E \cap F) = \mathrm{P}(E|F)\mathrm{P}(F)$$

and

$$\mathrm{P}(E \cap F) = \mathrm{P}(F|E)\mathrm{P}(E)$$

- Equating the right hand sides of the equations we have

$$\mathrm{P}(E|F)\mathrm{P}(F) = \mathrm{P}(F|E)\mathrm{P}(E)$$

- Assuming that $\mathrm{P}(F) > 0$ and solving for $\mathrm{P}(E|F)$ we obtain a result known as the **Bayes' rule**:

$$\mathrm{P}(E|F) = \frac{\mathrm{P}(F|E)\mathrm{P}(E)}{\mathrm{P}(F)}$$

This is an *very important* result because in general

$$\mathrm{P}(E|F) \neq \mathrm{P}(F|E)$$

- Note how the conditional probability $\mathrm{P}(E|F)$ can be interpreted as a *re-scaled* version of $\mathrm{P}(E)$. The result for $\mathrm{P}(F|E)$ is similar

- Suppose, like before, that the events $F_1, \ldots, F_n$ form a **partition** of the sample space S
- Given an event E, using the law of total probability, we can then write its probability as

$$\mathsf{P}(E) = \sum_i \mathsf{P}(E|F_i)\mathsf{P}(F_i)$$

which assumes knowledge of the conditional probabilities $\mathsf{P}(E|F_i)$ and unconditional probabilities $\mathsf{P}(F_i)$

- Using Bayes' theorem, we have

$$\mathsf{P}(F_i|E) = \frac{\mathsf{P}(E|F_i)\mathsf{P}(F_i)}{\mathsf{P}(E)}$$

assuming that $\mathsf{P}(E) > 0$.

2.33 Bayes theorem

- Suppose, like before, that the events $F_1, \ldots, F_n$ form a **partition** of the sample space S

- Given an event E, using the law of total probability, we can then write its probability as

$$\mathsf{P}(E) = \sum_i \mathsf{P}(E|F_i)\mathsf{P}(F_i)$$

which assumes knowledge of the conditional probabilities $\mathsf{P}(E|F_i)$ and unconditional probabilities $\mathsf{P}(F_i)$

- Using Bayes' theorem, we have

$$\mathsf{P}(F_i|E) = \frac{\mathsf{P}(E|F_i)\mathsf{P}(F_i)}{\mathsf{P}(E)}$$

assuming that $\mathsf{P}(E) > 0$.

- The general expression is therefore given by

$$\mathsf{P}(F_i|E) = \frac{\mathsf{P}(E|F_i)\mathsf{P}(F_i)}{\sum_i \mathsf{P}(E|F_i)\mathsf{P}(F_i)}$$

- Note that

$$\sum_{i=1}^{n} \mathsf{P}(F_i|E) = 1$$

2.29 Arrays example contd.

Suppose an errneous array is found among the arrays processed by the company.

What is the probability that it was processed by each one of the three machines?

We seek $P(X|D)$, $P(Y|D)$ and $P(Z|D)$.

Earlier we found $P(D) = 0.037$

2.29 Arrays example contd.

Suppose an errneous array is found among the arrays processed by the company.

What is the probability that it was processed by each one of the three machines?

We seek $P(X|D)$, $P(Y|D)$ and $P(Z|D)$.

Earlier we found $P(D) = 0.037$

We can compute the required probabilities by applying the Bayes rule:

$$P(X|D) = \frac{P(D|X)P(X)}{P(D)} = \frac{0.03(0.5)}{0.037} = 0.4054$$

$$P(Y|D) = \frac{P(D|Y)P(Y)}{P(D)} = \frac{0.04(0.3)}{0.037} = 0.3243$$

$$P(Z|D) = \frac{P(D|Z)P(Z)}{P(D)} = \frac{0.05(0.2)}{0.037} = 0.2703$$

Note that $P(X|D) + P(Y|D) + P(Z|D) = 1$

2.35 Diagnostic test

- A diagnostic test has probability 0.95 of giving correct diagnosis. Incidence of disease in the population is 0.005. What is the probability that a person with a positive test result has the disease?

- First, introduce some notation

$$D = \{ \text{ has disease } \} \quad \text{and} \quad R = \{ \text{ positive test } \}$$

- We know that $\mathsf{P}(R \mid D) = 0.95$, $\mathsf{P}(D) = 0.005$

- The required probability is

$$\mathsf{P}(D \mid R) = \frac{\mathsf{P}(R \mid D)\,\mathsf{P}(D)}{\mathsf{P}(R)}$$

2.35 Diagnostic test

- A diagnostic test has probability 0.95 of giving correct diagnosis. Incidence of disease in the population is 0.005. What is the probability that a person with a positive test result has the disease?

- First, introduce some notation

$$D = \{\text{ has disease }\} \quad \text{and} \quad R = \{\text{ positive test }\}$$

- We know that $\mathsf{P}(R \mid D) = 0.95$, $\mathsf{P}(D) = 0.005$

- The required probability is

$$\mathsf{P}(D \mid R) = \frac{\mathsf{P}(R \mid D)\,\mathsf{P}(D)}{\mathsf{P}(R)}$$

- A direct application of the **total probability theorem** gives

$$P(R) = P(R \cap D) + P(R \cap \bar{D}) =$$

$$= P(R|D)P(D) + P(R|\bar{D})P(\bar{D})$$

$$= (0.95 \times 0.005) + (0.05 \times 0.995) = \frac{19}{218} = 0.087$$

- The required probability is

$$\mathsf{P}(D \mid R) = \frac{(0.95)(0.005)}{0.087} = 0.0545$$

using the **Bayes' rule**

RANDOM VARIABLES AND THEIR DISTIBUTIONS

A random variable is **discrete** if the set $\mathbb{X}$ is of form

$$\mathbb{X} = \{x_1, x_2, ..., x_n\} \quad \text{or} \quad \mathbb{X} = \{x_1, x_2, ...\}$$

that is, a finite or at most a countably infinite number of values

- A discrete random variable is used to describe the outcomes of experiments that involve **counting** or **classification**, e.g.
  - number of males and females in this classroom
  - number of up-regulated genes
  - number of letters in a sequence
  - and so on

(non countable)

A random variable is **continuous** if the set $\mathbb{X}$ is of the form

$$\mathbb{X} = \bigcup_i \{x : a_i \leq x \leq b_i\}$$

for real numbers a_i, b_i, that is, the union of *intervals* in $\mathbb{R}$.

- A continuous random variable is used to describe the outcomes of experiments that involve **continuous measurements**, e.g.
  - height of students in this classroom
  - peak intensity in a mass spectrum
  - pixel intensity in a digital image
  - and so on

2.42 Probability distributions

- Consider the experiment that consists of tossing three fair coins (or, what is the same, a fair coin three times) and looking at all faces.
- Define the random variable

$$X = \{\ \textbf{number of heads observed in all the three tosses}\ \}$$

- The sample space S consists of 8 possible outcomes. All outcomes and corresponding values of X are given in the table below:

s	HHH	HHT	HTH	THH	HTT	THT	TTH	TTT
x	3	2	2	2	1	1	1	0

and notice that $\mathbb{X} = \{0, 1, 2, 3\}$

- Assuming that all eight sample points in S have equal probability, the **probability distribution** of X can be described by the following table

x	$\mathsf{P}(X = x)$
0	1/8
1	3/8
2	3/8
3	1/8
sum	1

- For instance $\mathsf{P}(X = 1) = \mathsf{P}(\{HTT, THT, TTH\}) = 3/8$
- Note that $\sum_{i=0}^{3} \mathsf{P}(X = i) = 1$

Probability mass function

- The probability distribution of a **discrete** random variable X is described by the function

$$p_X(x) = \mathsf{P}(X = x) = \mathsf{P}(\{s : X(s) = x\}) \quad x \in \mathbb{X}$$

called **probability mass function** or **p.m.f.**

- The p.m.f. is a function that exhibits the following **two properties**:

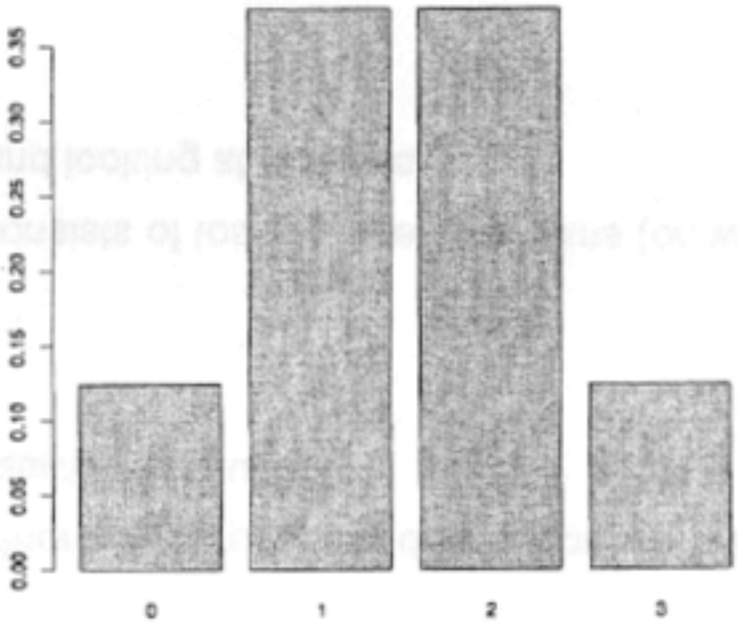
$$i.\ p_X(x_i) \geq 0 \quad \text{for all } x_i$$

$$ii.\ \sum_i p_X(x_i) = 1$$

2.45 Example. Three coins

Each outcome has an associated probability mass which can be visualized–the plot represents the probability distribution.

x	$P_X(X = x)$
0	1/8
1	3/8
2	3/8
3	1/8



Cumulative distribution function

The **cumulative distribution function** or **c.d.f.** or simply distribution function $F(\cdot)$ of the random variable X is defined by

$$F_X(x) = \mathsf{P}(X \leq x) = \mathsf{P}(\{s : X(s) \leq x\})$$

and is defined for all values of x

Any probabilistic aspects concerning a random variable X can be studied using its c.d.f. $F_X(x)$.

- In the previous experiment we had three tosses of a fair coin and the random variable X counted the number of observed heads.
- Remember that

$$p_X(x) = P(X = x) \quad \text{and} \quad F_X(x) = \mathrm{P}(X \leq x)$$

- In simple experiments such as this one, the **cumulative distribution** can also be represented in a table

x	$p_X(x)$	$F_X(x)$
0	1/8	1/8
1	3/8	1/2
2	3/8	7/8
3	1/8	1

- What is the value and meaning of $F_X(2)$?

More precisely, the relationship between p_X and F_X is obtained by noting that, if

$$x_1 \leq x_2 \leq \ldots \leq x_n \ldots$$

then

$$\mathrm{P}(X \leq x_i) = \mathrm{P}(X = x_1) + \ldots + \mathrm{P}(X = x_i)$$

and therefore

$$F_X(x) = \sum_{x_i \leq x} p_X(x_i)$$

If we know the distribution function, we can derive the probability mass function by noting that

$$p_X(x_1) = F_X(x_1)$$
$$p_X(x_i) = F_X(x_i) - F_X(x_{i-1}) \text{ for } i \geq 2$$

Notice how

- We calculate F_X from p_X by **summation**
- We calculate p_X from F_X by **differencing.**

We can then use the distribution function to compute specific probabilities, for instance

$$\mathrm{P}(a < X \leq b) = F_X(b) - F_X(a) \quad \text{for any} \quad a < b$$

2.48 Relationship between probability mass and distribution functions

2.49 Answering other probability questions using the distribution function

Remember that

$$P(a < X \le b) = F_X(b) - F_X(a)$$

We can compute other probabilities by noticing that

$$\begin{aligned} P(a \le X \le b) &= P(\{X = a\} \cup \{a < X \le b\}) \\ &= P(X = a) + P(a < X \le b) \\ &= P(X = a) + F_X(b) - F_x(a) \end{aligned}$$

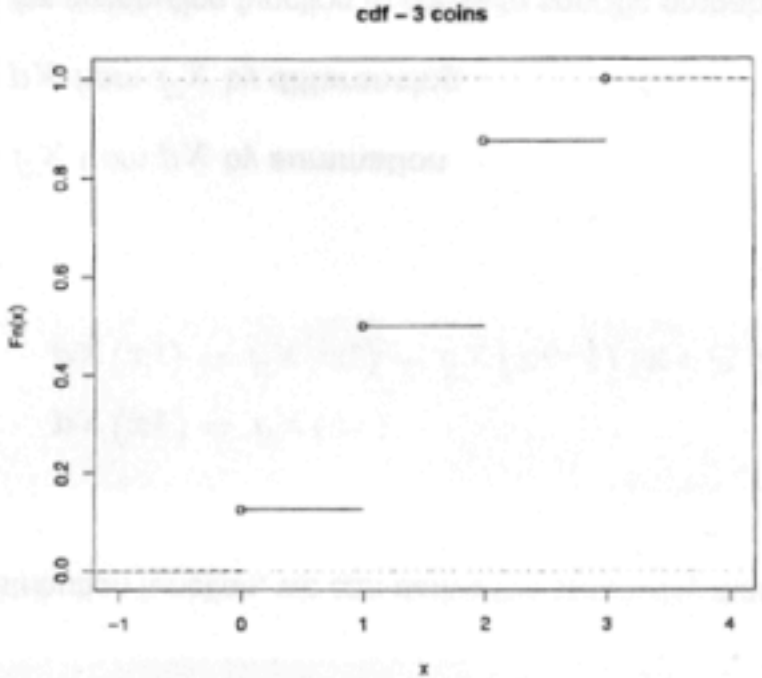
and

$$\begin{aligned} P(a < X < b) &= P(a < X \le b) - P(X = b) \\ &= F_X(b) - F_X(a) - P(X = b) \end{aligned}$$

2.50 Visualization

The distribution function for discrete random variables looks like a step-function. For instance, with reference to the previous example (three coins) where

$\mathbb{X} = \{0, 1, 2, 3\}$



Notice here:

- F_X can be computed for **all values of** x. For instance

$$F_X(2.5) = \mathsf{P}(X \leq 2.5) = \mathsf{P}(X = 0, 1, \text{ or } 2) = 7/8$$

- F_X has **jumps** at the values of $x_i \in \mathbb{X}$ and the size of the jump at x_i is equal to $\mathsf{P}(X = x_i)$
- F_X is **constant between jumps**
- $F_X = 0$ for $x < 0$ and $F_X = 1$ for $x \geq 3$ (in this example)

The cumulative distribution function of a discrete random variable is a **step-function** with the following **general properties**

- $0 \leq F_X(x) \leq 1$
- $\lim_{x \to -\infty} F(x) = F(-\infty) = \mathsf{P}(\emptyset) = 0$
- $\lim_{x \to \infty} F(x) = F(\infty) = \mathsf{P}(S) = 1$
- F_X is **discontinuous**, with jumps at some x
- The size of the jump at x is equal to $\mathsf{P}(X = x)$.
- **Right-continuity**: at the jump points, F_X takes the value at the top of the jump (i.e. the function is continuous when a point is approaching from the right)

$$\lim_{h \to 0+} F_X(x+h) = F_X(x)$$

- It is **non-decreasing**, i.e.

$$F_X(x_1) \leq F_X(x_2) \text{ if } x_1 \leq x_2$$

2.53 How to specify a probability distribution?

- We have established that a probability distribution is a **function** that assigns probabilities to the possible values of a random variable.
- So far we have looked at examples where the the p.m.f. and c.d.f. were derived by direct inspection of the entire sample space.
- When specifying a probability distribution, two aspects need to be considered:
  (a) The **range** of the random variable (that is, the values of the random variable which have positive probability)
  (b) The method via which the probabilities are assigned to different values in the range – typically this is achieved by means of a **function or formula.** We need to find a function or formula via which probabilities of form

$$\mathsf{P}(X = x) = \mathsf{P}(\{s : X(s) = s\})$$

can be calculated for each x in a suitable range $\mathbb{X}$.

- The functions used to specify these probabilities are just real-valued functions of a single real argument, similar to polynomial, exponential, logarithmic, etc. – for instance

$$f(x) = e^{x} \qquad \text{or} \qquad f(x) = 6x^3 + 3x^2 + 2x - 5$$

However, functions specifying probability distributions must exhibit **certain properties.**

2.54 Example.

Suppose we are given $A = \{2^i : i = 0, 1, 2, \cdots\}$

Consider the function $p(1) = \frac{3}{4}$ and $p(2^i) = \left(\frac{1}{5}\right)^i, i > 0$

Does this function define a probability mass function?

- We need to verify the two main properties, that $p(x)$ is positive for every x, and that $\sum_{x \in A} p(x) = 1$

- The first property is easily verified, so let us check the second one. First, write

$$\sum_{x \in A} p(x) = p(1) + \sum_{i=1}^{\infty} p(2^i) = \frac{3}{4} + \sum_{i=1}^{\infty} \left(\frac{1}{5}\right)^i = \frac{3}{4} + \frac{1}{5} \sum_{i=0}^{\infty} \left(\frac{1}{5}\right)^i$$

Given that

$$\sum_{i \geq 0}^{\infty} \left(\frac{1}{5}\right)^i = \frac{1}{1 - \frac{1}{5}} = \frac{5}{4}$$

We verify that

$$\sum_{x \in A} p(x) = \frac{3}{4} + \frac{1}{5}\frac{5}{4} = 1$$

2.55 More examples

Let X be a discrete random variable with probability mass function

$$p_X(x) = kx \qquad \mathbb{X} = \{1, \ldots, 5\}$$

that is, X takes values $1, 2, \ldots, 5$ with probabilities $k, 2k, \ldots, 5k$.

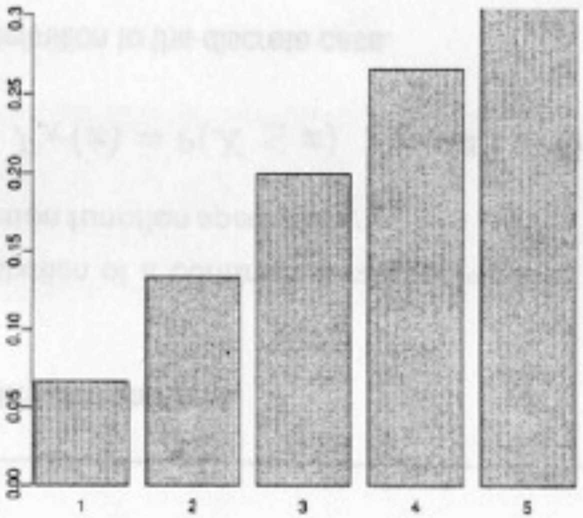
What value of k makes this function a probability mass function?

- We know that all probabilities must sum up to 1, therefore

$$1 = \sum_x p_X(x) = k + 2k + \ldots + 5k = k(1 + \ldots + 5)$$

$$= \frac{5 \cdot 6}{2} k = 15k \Rightarrow k = \frac{1}{15}$$

- When k is known, we can draw the complete probability distribution for this random variable



2.56 Previous example contnd.

- For instance, what is the probability that X is greater than 3?

$$P(X>3) = 1 - P(X \leq 3)$$
$$= 1 - F_X(3) = 1 - (k + 2k + 3k) = \frac{3}{5}$$

- What is the probability that X is greater than 1 and less or equal 2

$$P(1 < X \leq 2) = F_X(2) - F_X(1)$$
$$= (k + 2k) - k = \frac{2}{15}$$

- What is the cumulative distribution function?

x	$p_X(x)$	$F_X(x)$
1	0.067	0.067
2	0.134	0.200
3	0.200	0.400
4	0.267	0.667
4	0.334	1

which allows the compute the probabilities above by reading off $F_X(x)$ directly from the table.

The probability distribution of a **continuous** random variable X is defined by the **cumulative distribution function** specified by

$$F_X(x) = \mathrm{P}(X \leq x) \qquad \text{for all } x \in \mathbb{R}$$

That is, an *identical* definition to the discrete case.

The continuous c.d.f. F_X must exhibit the same properties as for the discrete c.d.f., except the **right-continuity** which is now replaced by **continuity**:

- $0 \leq F_X(x) \leq 1$
- $\lim_{x \to -\infty} F(x) = F(-\infty) = \mathrm{P}(\emptyset) = 0$
- $\lim_{x \to \infty} F(x) = F(\infty) = \mathrm{P}(S) = 1$
- $F_X(x)$ is **continuous**, i.e.

$$\lim_{h \to 0} F_X(x+h) = F_X(x)$$

- It is **non-decreasing**, i.e.

$$F_X(x_1) \leq F_X(x_2) \text{ if } x_1 \leq x_2$$

2.58 Probability density function

- Associated with a **continuous random variable** X and its c.d.f. F_X there is another function called the **probability density function or p.d.f.** $f_X(x)$.
- The density function is a function defined as

$$\frac{d}{dx}F_X(x) = f_X(x)$$

and

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt \quad \text{for all } x$$

Notice the analogy with the discrete case, but here

- We calculate F_X from f_X by **integration**
- We calculate f_X from F_X by **differentiation**

A density function $f_X(x)$ must exhibit the following properties:

- $f_X(x) \geq 0 \quad$ for $x \in \mathbb{X}$
- $\int_{-\infty}^{\infty} f(x)dx = 1$

- Note that, for continuous random variables,

$$\mathsf{P}(a < X \leq b) = F_X(b) - F_X(a) \to 0 \quad \text{as } b \to a$$

- Hence for each x we must have

$$\mathsf{P}(X = x) = 0$$

if X is continuous

- Therefore, for a continuous random variable,

$$f_X(x) \neq \mathsf{P}(X = x)$$

- We must use F_X to specify the probability distribution initially
- In some cases it is often easier to think of the **shape** of a continuous distribution, which is described by the density function f_X.
- For instance, when we think of the *normal distribution* as a bell-shaped distribution, we are referring to its density

2.60 Example

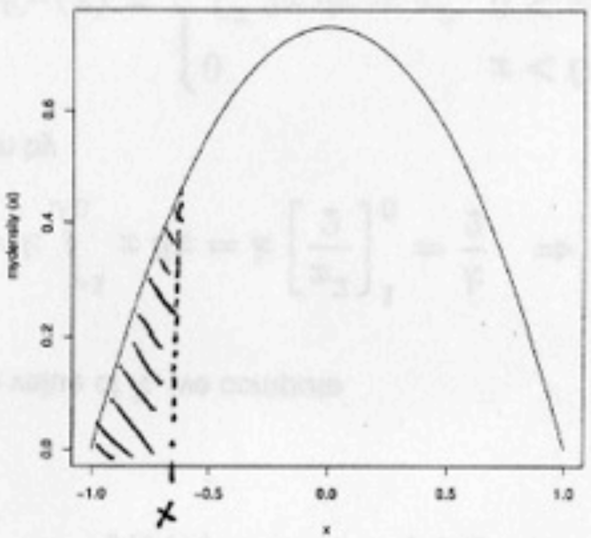
We are given a function $f(x) = k(1 - x^2)$ on $[-1, 1]$

What is the value of k that makes $f(x)$ a probability density function?

We procede as before,

$$1 = k\int_{-1}^{1}(1-x^2)dx = k\left[x - \frac{x^3}{3}\right]_{-1}^{1} = \frac{4k}{3} \quad \Rightarrow \quad k = \frac{3}{4}$$

We can then sketch the probability density



- And compute specific probabilities, for instance:

$$\mathrm{P}(\text{non-negative outcome}) = \mathrm{P}(X \geq 0) = \int_0^1 k(1-x^2)\,dx = \frac{1}{2}$$

$$\mathrm{P}(-1/2 \leq X \leq 1/2) = \int_{-1/2}^{1/2} k(1-x^2)\,dx = \frac{11}{16}$$

Derive the distributon function using the density function provided in the previous example.

2.61 Previous example continued

- By direct application of the definition

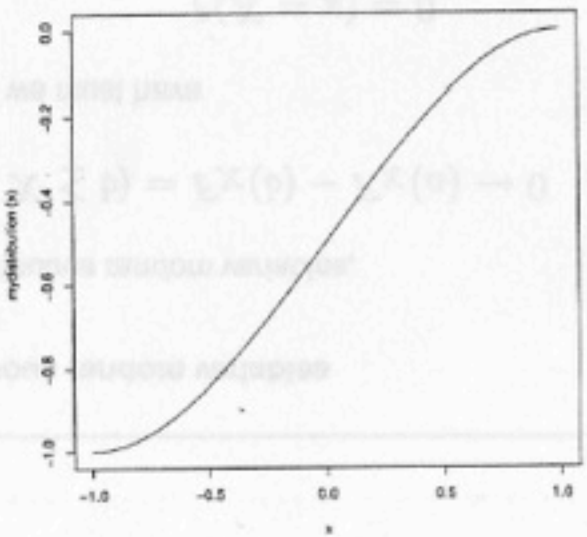
$$F(x) = \int_{-\infty}^{x} f_X(s)ds$$

$$= \int_{-1}^{x} k(1 - s^2)ds$$

$$= k\left[s - \frac{1}{3}s^3\right]_{-1}^{x}$$

$$= \frac{3}{4}\left(x - \frac{1}{3}x^3 - \frac{2}{3}\right) \quad \text{for } -1 \le x \le 1$$

- $F(x) = 0 \quad$ for $x < -1$
- $F(x) = 1 \quad$ for $x > 1$



Let X be a continuous r.v. with pdf

$$f_X(x) = kx \quad \text{on } [0,1]$$

- To determine the value of k, we compute

$$1 = k\int_0^1 x\,dx = k\left[\frac{x^2}{2}\right]_0^1 = \frac{k}{2} \quad \Rightarrow \quad k = 2$$

- The c.d.f. is given by

$$F_X(x) = \begin{cases} 0 & x < 0 \\ \int_0^x 2x\,dx = x^2 & 0 \le x < 1 \\ \int_0^1 2x\,dx = 1 & x \ge 1 \end{cases}$$

- And we may want to compute

$$\mathrm{P}\left(\frac{1}{4} < X \le 2\right) = F_X(2) - F_X\left(\frac{1}{4}\right) = 1 - \left(\frac{1}{4}\right)^2 = \frac{15}{16}$$

Discrete random variables

- A *discrete* random variable X takes on at most a *countable* number of possible values.
- The *probability mass function* $p_X(x)$ gives the probability of observing all values of X, namely $\mathrm{P}(X = x)$ for all possible x.
- The probability distribution of X is specified by the *cumulative distribution function* $F_X(x) = \mathrm{P}(X \leq x)$
- Alternatively, we say that a random variable X is *discrete* if $F_X(x)$ is a *step function* of x.

Continuous random variables

- A *continuous* random variable X takes values over an interval
- The probability distribution of X is specified by the *cumulative distribution function* $F_X(x) = \mathrm{P}(X \leq x)$
- Alternatively, we say that a random variable X is *continuous* if $F_X(x)$ is a *continuous function* of x