Chapter 1

PROBABILITY

1.1 Set Theory Definitions

A **SET** is a collection of different elements e.g. A=set of scores on a dice={1,2,3,4,5,6}.

The **SIZE** n(A) of a set A is the number of elements it contains e.g. n(A) = 6.

An **EXPERIMENT** is a controlled process by which observations are obtained. Some experiments that can be 'independently and identically repeated', e.g. tossing a dice

The set of all possible outcomes of an experiment is called the SAMPLE SPACE S

e.g.	**EXPERIMENT**	**SAMPLE SPACE S**
i)	toss a dice and record the score	{1,2,3,4,5,6}
ii)	toss two coins and record the sequence of results	{HH, HT, TH, TT}
iii)	toss three coins and record the number of heads	{0,1,2,3}

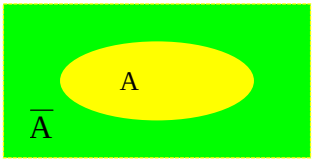
A subset A of the sample space S is called an **EVENT**.

A subset A with a **SINGLE** event is called a **SIMPLE EVENT**

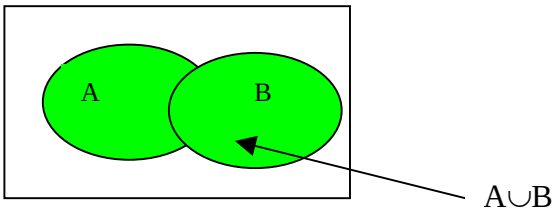
An EVENT is said to **OCCUR** if any of the elements in A occurs e.g. toss a dice and record the score S=(1,2,3,4,5,6) and let A=event of an even score=(2,4,6), then A is said to have occurred if the result of tossing the dice is 2 or 4 or 6

SET NOTATION

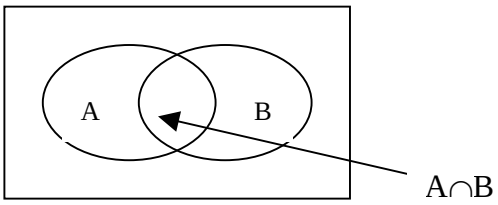
$\overline{A}$, the **COMPLEMENT** of A contains all elements of S not in A e.g. S = {1,2,3,4,5,6}, A = {2,4,6} then $\overline{A}$ = {1, 3, 5)



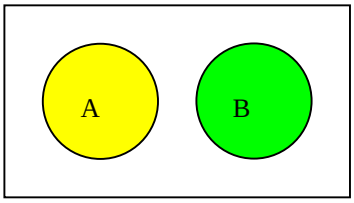
$A \cup B$, the UNION of two sets A and B contains all elements that lie in A or B or both, $A \cup B$ is the event 'A or B occurs', e.g. A = {2, 4, 6} and B ={5, 6} then $A \cup B$ = {2, 4, 5, 6}.



$A \cap B$, the INTERSECTION A and B contains all elements inside both A and B, $A \cap B$ is the event 'A and B both occur' e.g. A = {2, 4, 6} and B = {5, 6} then $A \cap B$ = {6}.



Two sets A and B are **DISJOINT** (or **MUTUALLY EXCLUSIVE**) if they have no elements in common, i.e. if their intersection is the empty set $\varnitothing$, i.e. if $A \cap B = \varnitothing$.



Exercise: Draw a diagram for a sample space S with two events A and B and shade each of the following events in a different colour:

$A \cap B$, $A \cap \overline{B}$, $\overline{A} \cap B$, $\overline{A} \cap \overline{B}$

1.2 Probability defined to be the limit of relative frequency

Let S be the sample space for an experiment and let A be an event in S. Suppose the experiment is 'independently and identically repeated' n times. Let f_A count the frequency of times event A occurs out of the n repetitions.

The PROBABILITY of event A occurring (in a single repetition of the experiment) is defined to be the limiting value of the sample proportion $\hat{p}_A$ of times event A occurs, i.e.

$$p(A) = \lim_{n\to\infty} \hat{p}_A = \lim_{n\to\infty} \frac{f_A}{n}$$

e.g. toss a fair coin n times and count f_H the frequency of times event H 'heads' occurs, then

$$p(H) = \lim_{n\to\infty} \hat{p}_H = \lim_{n\to\infty} \frac{f_H}{n} = \frac{1}{2}$$

1.3 Axioms of probability

Let S be a sample space and A and B be two events in S. Probability satisfies the following axioms:

A1 $0<p(A)<1$ probability lies between 0 and 1

A2 $p(S)=1$ S is certain to occur

A3 $p(A \cup B)=p(A)+p(B)$ provided A and B are disjoint events

This is the Addition Rule for disjoint events

1.4 Rules derived from the axioms of probability

Many rules can be derived from the three basic axioms of probability, e.g.

Rule 1 $p(\varnothing)=0$

Rule 2 $p(\overline{A}) = 1 - p(A)$

Rule 3 $p(A \cup B) = p(A) + p(B) - p(A \cap B)$

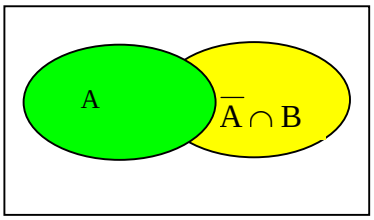
$p(A \text{ or } B) = p(A) + p(B) - p(A \text{ and } B)$

This is the General Addition Rule for any two events

Proof 3 $p(A \cup B) = p(A) + p(\overline{A} \cap B)$ since A and $\overline{A} \cap B$ are disjoint

$p(B) = p(A \cap B) + p(\overline{A} \cap B)$ since $A \cap B$ and $\overline{A} \cap B$ are disjoint

$$p(A \cup B) = p(A) + p(B) - p(A \cap B)$$



1.5 Calculating probabilities from simple events

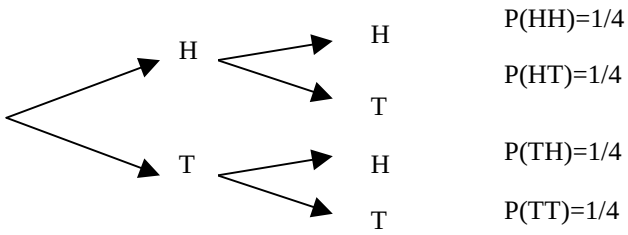
Rule 4: Let $A = (a_1, a_2, ..., a_{n(A)})$ comprising n(A) simple events be an event in sample space S

$$p(A) = \sum_{i=1}^{n(A)} p(a_i)$$

Examples:

i) Toss a fair dice and record the score S = {1, 2, 3, 4, 5, 6}
Let A = event 'getting an odd score' = {1, 3, 5}
$p(A) = p(1) + p(3) + p(5) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{3}{6} = \frac{1}{2}$

ii) Toss a fair coin twice and count the number of heads S = {0, 1, 2}
Let A = event 'getting at least one head' = {1, 2}
$p(A) = p(1) + p(2) = \frac{1}{2} + \frac{1}{4} = \frac{3}{4}$



Rule 5: Let A be an event in sample space S comprising n(S) **EQUALLY LIKELY** simple events

$$p(A) = \frac{n(A)}{n(S)}$$

$$\text{i.e.} \quad p(A) = \frac{\text{number of simple events in A}}{\text{number of simple events in S}}$$

***Proof*:**

$$p(A) = \sum_{i=1}^{n(A)} p(a_i) = \sum_{i=1}^{n(A)} \frac{1}{n(S)} = \frac{n(A)}{n(S)}$$

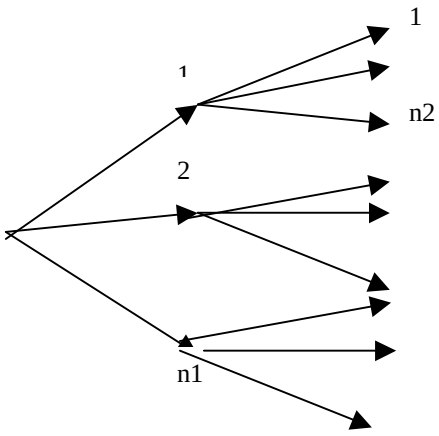
In example i), since the dice is fair all n(S) = 6 outcomes from the sample space S={1, 2, 3, 4, 5, 6} are equally likely, so for A={1, 3, 5},

$$p(A) = \frac{n(A)}{n(S)} = \frac{3}{6}$$

Multiplication rule for finding n(S)

If an experiment comprises two procedures P1 and P2 each with n1 and n2 simple events respectively, then

n(S) = n1*n2



n(S) = n1 * n2

Example: Toss two fair dice (one red and one blue) and record the score on each dice. Sample space S is given by S = {11, 12, 13, 14, 15, 16, 21, 22, …….., 61, 62, 63, 64, 65, 66}

In tabular form the sample space S is given by

	blue					
red	1	2	3	4	5	6
1	11	12	13	14	15	16
2	21	22	23	24	25	26
3	31	32	33	34	35	36
4	41	42	43	44	45	46
5	51	52	53	54	55	56
6	61	62	63	64	65	66

n(S) = n1*n2 = 6*6 = 36

Since both dice are fair each of the 36 outcomes are **equally likely,**

hence $$p(A) = \frac{n(A)}{n(S)}$$

What is the probability of each of the following events?

i) A = two sixes = (66)

$$p(A) = \frac{n(A)}{n(S)} = \frac{1}{36}$$

ii) B = the same score on each dice = (11, 22, 33, 44, 55, 66)

$$p(B) = \frac{n(A)}{n(S)} = \frac{6}{36}$$

iii) C = the total score from the two dice is 10 = (64, 55, 46)

$$p(C) = \frac{n(A)}{n(S)} = \frac{3}{36}$$

1.6 Estimating probabilities from large sample proportions

Example 1

The following data are taken from Newell, D. J. (1964) JRSS, A, 127, 1-33
and Hand, D.J. 'Small data sets' p97.

The duration of pregnancy was recorded for 1669 women, giving the following:

Number of weeks	Frequency	Proportion
10-15	1	0.0006
15-20	2	0.0012
20-25	8	0.0048
25-30	17	0.0102
30-35	84	0.0503
35-40	683	0.4092
40-45	859	0.5147
45-50	14	0.0084
50-55	0	0
55-60	1	0.0006

Estimated probability of an event = sample proportion = $\frac{\text{frequency}}{\text{total}} = \frac{\text{f}}{\text{n}}$

i) Estimate the probability a pregnancy lasts from 40 up to (but not including) 45 weeks.

Let A = event 'pregnancy lasts from 40 up to 45 weeks'
p(A) = 0.5147 i.e. approximately 51.47%

ii) Estimate the probability that a pregnancy lasts 50 or more weeks.

Let A = event 'pregnancy lasts 50 or more weeks'
p(A) = 0.0006 i.e. approximately 0.06%

iii) Estimate the probability that a pregnancy lasts less than 25 weeks.

Let A = event 'pregnancy lasts less than 25 weeks'
p(A) = 0.0066 i.e. approximately 0.66%

iv) Estimate the probability that a pregnancy lasts from 35 up to (but not including) 45 weeks.

Let A = event 'pregnancy lasts from 35 up to 45 weeks'
p(A) = 0.9239 i.e. approximately 92.39%

1.7 Calculating probabilities using set theory: joint and marginal probabilities

Example 2

The following data are taken from Goodman, L.A. (1981) JASA, 76, 320-334 and Hand, D.J. 'Small data sets' p146

The hair and eye colour were recorded for a sample of 22,361 children in Aberdeen, Scotland giving the following frequencies:

	HAIR	
EYES	light	dark
blue	2579	399
not blue	13947	5436

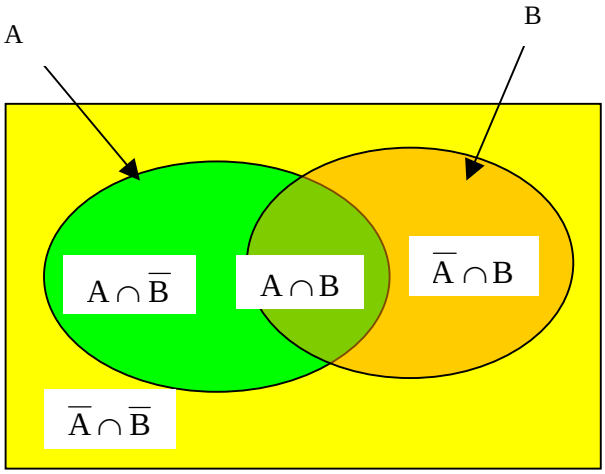
Sample proportion $= \frac{\text{frequency}}{\text{total}} = \frac{f}{n}$ where n = 22,361

	HAIR		
EYES	not dark,D	dark,D	Total
blue, B	0.1153	0.0178	0.1331
not blue, B	0.6237	0.2431	0.8668
Total	0.7390	0.2609	1

For a child selected at random from the sample,
let B = event 'child has blue eyes'
let D = event 'child has dark hair'

What are the following probabilities?
p(B) = 0.1331 i.e. 13.31%
p(D) = 0.2609 i.e. 26.09%
p(B∩D) = 0.0178 i.e. 1.78%

Find the following probabilities:
p(B)
P(D)
$p(B \cap \overline{D})$
p(B∪D)

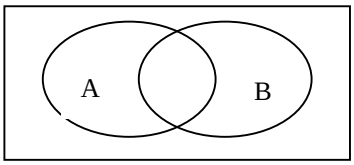


Draw the events B and D on a diagram and identify probabilities for the following events:
$B \cap D$, $B \cap \overline{D}$, $\overline{B} \cap D$, $\overline{B} \cap \overline{D}$.

1.8 Conditional probability

Given that the event B has occurred, then the probability that A occurs is called the **conditional probability** of A given B, i.e. $p(A|B)$.

$$p(A|B) = \frac{p(A \cap B)}{p(B)}$$



This gives the general **Multiplication Rule for any two events,** i.e**.** $p(A \cap B) = p(A|B) * p(B)$**.**

Remember event B occurs if any element of B occurs.

Example 3

Toss a fair dice and record the score, S=(1,2,3,4,5,6).

Let A = event that the score on the dice is odd, A=(1,3,5).

Let B = event that the score on the dice is high, B=(4,5,6).

What is the probability of event A occurring?

$$p(A) = \frac{n(A)}{n(S)} = \frac{3}{6} = \frac{1}{2}$$

Suppose I toss the fair dice, look at the result and tell you that the score is high (i.e. event B has occurred), what is now the probability of event A, an odd score?

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

Example 4: Reconsider example 2.

	HAIR		
EYES	not dark,D	Dark,D	Total
blue, B	0.1153	0.0178	0.1331
not blue, B	0.6237	0.2431	0.8668
Total	0.7390	0.2609	1

Suppose a child is picked at random from the sample.

What is the probability that the child has blue eyes, given that the child has dark hair?

$$p(B|D) = \frac{p(B \cap D)}{p(D)} = \frac{0.0178}{0.2609} = 0.0682 \quad \text{i.e. } 6.82\%$$

What is the probability that the child has blue eyes, given that the child does NOT have dark hair?

$$p(B|\overline{D}) = \frac{p(B \cap \overline{D})}{p(\overline{D})} = \frac{0.1153}{0.7390} = 0.156 \quad \text{i.e. } 15.6\%$$

Find the probability that a child has dark hair given that the child has blue eyes.
Find the probability that a child has dark hair given that the child does NOT have blue eyes.

Independent events

Events A and B are INDEPENDENT $\Leftrightarrow$ $p(A|B) = p(A)$

$\Leftrightarrow$ $p(A \cap B) = p(A) * p(B)$

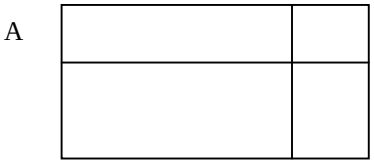
$\Leftrightarrow$ $p(B|A) = p(B)$

This gives the **Multiplication Rule for Independent Events, i.e.** $p(A \cap B) = p(A) * p(B)$

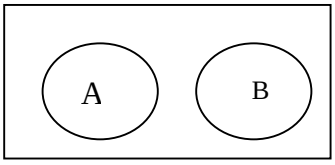
Disjoint events

Events A and B are disjoint if $A \cap B = \varnothing$
Hence $p(A \cap B) = 0 \neq p(A)*p(B)$, therefore A and B are not independent.

INDEPENDENT



DISJOINT



1.9 Bayes Theorem

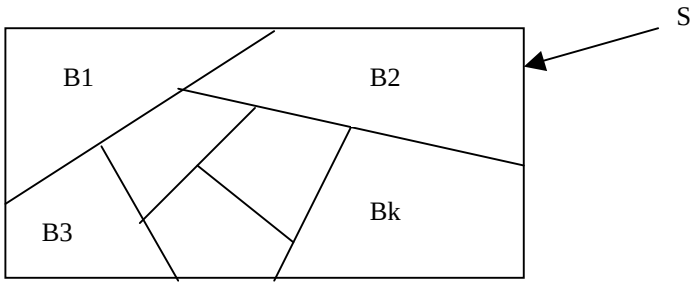
PARTITION of the sample space S

The events $B_1, B_2, ..., B_k$ form a **PARTITION** of the sample space S $\Leftrightarrow$

i) $B_i \cap B_j = \varnothing$ for all $i \neq j$ [**EXCLUSIVE**]

ii) $\bigcup_{i=1}^{k} B_i = S$ i.e. $\bigcup_{i=1}^{k} B_i = B_1 \cup B_2 \cup ..... \cup B_k = S$ [**EXHAUSTIVE**]

iii) $p(B_i) > 0$ for all i



PARTITION = MUTUALLY EXCLUSIVE and **EXHAUSTIVE**

Theorem of TOTAL PROBABILITY

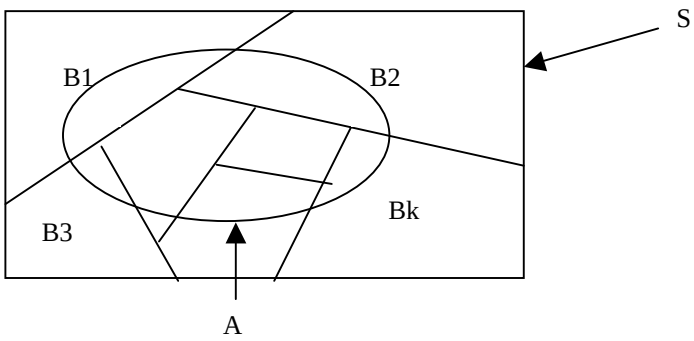
Let $B_1, B_2, ..., B_k$ form a partition of the sample space S, then

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

Proof

$$p(A) = \sum_{i=1}^{k} p(A \cap B_i) = \sum_{i=1}^{k} p(A \mid B_i)\, p(B_i)$$

since $A = (A \cap B_1) \cup (A \cap B_2) \cup ...(A \cap B_k)$ is the union of disjoint sets



***Bayes Theorem**:* Let $B_1, B_2, ..., B_k$ form a partition of the sample space S, then

$$p(B_j \mid A) = \frac{p(A \mid B_j)\, p(B_j)}{\sum_{i=1}^{k} p(A \mid B_j)\, p(B_j)}$$

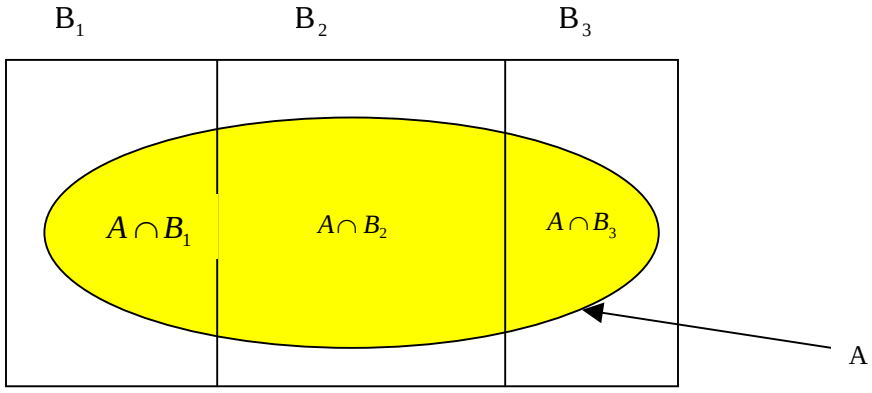
Proof

Conditional probability

$$p(B_j \mid A) = \frac{p(A \cap B_j)}{p(A)} = \frac{p(A \mid B_j)\, p(B_j)}{\sum_{i=1}^{k} p(A \mid B_j)\, p(B_j)}$$

Theorem of Total Probability

e.g.



Example:

In the USA during 1989 there were 30,232 suicides, of which 24,102 were male and 6,130 were female. Among the male suicides (i.e. given that the suicide was a male) the proportions of different methods used were as follows:

MALES

METHOD	Poison,P	Hanging,H	Guns,G	Other,O
PROPORTION	0.133	0.154	0.651	0.062

Among the female suicides (i.e. given that the suicide was a female) the proportions of different methods used were as follows:

FEMALES

METHOD	Poison,P	Hanging,H	Guns,G	Other,O
PROPORTION	0.364	0.126	0.408	0.102

Select one case at random.

Q1 What is the probability the person committing suicide was female?

$$p(\mathrm{F}) = \frac{n(\mathrm{F})}{n(\mathrm{S})} = \frac{6130}{30232} = 0.203$$ i.e. 20.3 %

Q2 What is the probability the person committing suicide was male?

p(M) = 1-p(F) = 1-0.203 = 0.797 i.e. 79.7 %

Q3 What is the probability poison was used in the suicide?

Since (M,F) form a partition of the sample space,

use the Theorem of Total Probability:

$p(A) = P(A|M)\,p(M) + P(A|F)\,p(F) = 0.133*0.797+0.364*0.203=0.106+0.074 = 0.18$, i.e. 18%

Q4 What is the probability the person committing suicide was female, **GIVEN** that poison was used in the suicide?

Using Bayes Theorem:

$$p(F \mid A) = \frac{p(F \cap A)}{p(A)} = \frac{p(A|F)\,p(F)}{p(A)} = \frac{0.074}{0.18} = 0.41$$ i.e. 41.1 %

EXERCISE 1 PROBABILITY

Q1 EXPERIMENT: Toss a fair coin 3 times and record the sequence of heads or tails on each toss (i.e. the sequence 1st toss, 2nd toss, 3rd toss).

a) Write out the sample space of possible outcomes of the experiment.
(Hint: try using a tree diagram)

b) Find the probability of each of the following events:
i) three heads i.e. (HHH)
ii) the sequence (HTH)
iii) exactly one heads
iv) one or more heads

Q2 EXPERIMENT: Toss a fair dice 2 times and record the sequence of scores.

a) Write out the sample space of possible outcomes of the experiment.
(Hint: try using a two-way table)

b) Find the probability of each of the following events:
i) two sixes i.e. (66)
ii) exactly one six
iii) one or two sixes
iv) a total score of exactly 7
v) a total score of more than 7

Q3 EXPERIMENT: Toss a fair dice 3 times and record the sequence of scores.

a) Write out the sample space of possible outcomes of the experiment.
(Hint: try using a tree diagram)

b) Find the probability of each of the following events:
i) three sixes i.e. (666)
ii) two sixes followed by a 1
iii) two sixes followed by any score not a six
iv) exactly two sixes
v) a total score of 4 or less

Q4 The USA Bureau of Labour Statistics reported in *Employment and Earnings* the age distribution of employed people aged 16 and over (in thousands):

Age (yrs)	Number
16-20	5899
20-25	11748
25-35	28429
35-45	23597
45-55	15216
55-65	10163
65-	2737
Total	97789

[From: Weiss (1995), 4.3, p188]

Calculate the corresponding sample proportions for each age group.

If an employed person is picked at random, what is the probability of each of the following events?

a) the person is aged from 16 up to 20
b) the person is aged from 25 up to 45
c) the person is aged over 65

Q5 A study conducted by the USA Census Bureau on the method used to travel to work gave the following frequencies (in thousands):

	AREA	
METHOD	Urban, $\overline{R}$	Rural, R
Car, C	45000	15000
Public transport, $\overline{C}$	6500	500

[From: Weiss (1995), 4.16, p189]

Calculate the corresponding sample proportions for each group.

If an employed person is picked at random, what is the probability of each of the following events:

a) the person travels to work by car
b) the person lives in a rural area

c) the person travels to work by car and lives in a rural area
d) draw the events C and R on a diagram and identify the probabilities of each of the events $C \cap R$, $C \cap \overline{R}$, $\overline{C} \cap R$, $\overline{C} \cap \overline{R}$
e) find the probabilities of each of the following: C, R, $C \cup R$, $C \cup \overline{R}$, $\overline{C} \cup R$, $\overline{C} \cup \overline{R}$

EXERCISE 2

CONDITIONAL PROBABILITY and BAYES THEOREM

Q1 The following data were given in question 5 of Exercise 1.
A study conducted by the USA Census Bureau on the method used to travel to work gave the following proportions:

	AREA	
METHOD	Urban, $\overline{R}$	Rural, R
Car, C	0.6716	0.2239
Public transport, $\overline{C}$	0.0970	0.0075

[From: Weiss (1995), 4.16, p189]

If an employed person is picked at random, find the following:

a) the probability the person travels to work by car, **GIVEN** that the person lives in a rural area.
b) the probability the person travels to work by car, **GIVEN** that the person lives in an urban area.
c) $p(\overline{C} \mid R)$
d) $p(\overline{C} \mid \overline{R})$

Q2 The American Lung Association has reported that
i) 7% of the population has Lung disease
ii) of people with Lung disease, 90% are smokers
iii) of people without Lung disease, 25% are smokers

[From: Weiss (1995), Example 4.31, p240]

Suppose a person is picked at random from the population.
Let L = event the person has Lung disease

Let S = event the person is a Smoker.

a) Write each of i), ii) iii) as probabilities of events.
b) Find the probability the person does **NOT** have Lung disease, $p(\overline{L})$
c) Find $p(\overline{S} \mid L)$ and $p(\overline{S} \mid \overline{L})$
d) Find the probabilities of events $S \cap L$, $S \cap \overline{L}$, $\overline{S} \cap L$, $\overline{S} \cap \overline{L}$
and write these joint probabilities in a two-way table.
Now, find the marginal probabilities of the events S and $\overline{S}$.
e) Find the probability the person has Lung disease, **GIVEN** that the person is a Smoker.
f) Find the probability the person has Lung disease, **GIVEN** that the person is **NOT** a Smoker.

Q3 An investigation into the relationship between Snoring and Heart disease published by Norton and Dunn (1985) *British Medical Journal,* **291**, p630-632
found the following results:

i) 4.4% of people suffered from heart disease (H).
ii) Among the people with heart disease,
21.8% were non-snorers (N),
31.8% were occasional snorers (O), and
46.4% were regular snorers (R).
iii) Among the people **without** heart disease,
57.1% were non-snorers (N),
25.4% were occasional snorers (O), and
17.5% were regular snorers (R).

Suppose a person is picked at random from the investigation.

a) Write each of the results in i), ii) and iii) as probabilities of events.
b) Draw a Venn diagram with events N, O, R and H, $\overline{H}$ labelled.
c) Find probabilities for the following events
$N \cap H$, $O \cap H$, $R \cap H$, and $N \cap \overline{H}$, $O \cap \overline{H}$, $R \cap \overline{H}$
Fill in the probabilities on your Venn diagram.
d) Write the joint probabilities in a two-way table.
Now, find the marginal probabilities of the events N, O and R.
e) Hence find the following probabilities
that the person has heart disease **GIVEN** that they are non-snorers
that the person has heart disease **GIVEN** that they are occasional snorers
that the person has heart disease **GIVEN** that they are regular snorers
f) Display the results in e) using a tree diagram with first branches defined by heart disease (H) and **not** heart disease ($\overline{H}$).
g) Compare the results in e) graphically using a probability bar chart.

Q4 The US department of Defence has reported that among all US males on active military service,

i) 14.2% are officers (O) and the remaining 85.8% are enlisted (E).
ii) Among the officers, 88.2% were white, 5.3% were black and 6.5% were neither.
iii) Among the enlisted, 67.5% were white, 20.6% were black and 11.8% were neither.

[From: Weiss (1995), 12.37, p716]

Suppose a military serviceman is picked at random.

a) Write each of the results in i), ii) and iii) as probabilities of events.
b) Draw a Venn diagram with events labelled.
c) Find the joint probabilities and record them in a two way table.
d) Fill in the probabilities on your Venn diagram.
e) Hence find the following probabilities
that the person is an officer **GIVEN** that they are black
that the person is an officer **GIVEN** that they are white
that the person is an officer **GIVEN** that they are neither.