Chapter 2

Sets

Before embarking upon a detailed discussion of statistics, it is useful to introduce some notation to help in describing data. This section introduces elementary set theory notation and Venn diagrams so that these can be used in later discussions.

The notion of a set is a collection of object or elements. This collection can also be referred to simply as data, and the individual elements in the data can themselves be referred to as data (in the singular sense), or as an event or element. We usually denote a set with a capitol letter, for example Ω. The element of a set is denoted by either ω or x_i, where the latter explicitly references the i^{th} element of the set. When writing a set, the elements are contained within curly braces '{' and '}'. For example we can write a set A that contains the elements 1 and 0 as

$$A = \{1, 0\}.$$

The order of elements is irrelevant in a set, so we could write A in an equivalent form as

$$A = \{0, 1\}.$$

If we want to express the information that a given element is or is not a part of a set, we use the symbols $\in$ and $\notin$, respectively. For example we may write

$$0 \in A, \text{ and } 1 \in A,$$
$$\text{but } 2 \notin A,$$

to express that both 0 and 1 are elements of A, but 2 is not an element of this set. The set $A = \{1, 0\}$ is called the binary set.

Other useful sets include:

$\mathbb{R}$	the set of all real numbers
$\mathbb{R}^{\pm}$	the set of all positive/negative real numbers
$\mathbb{C}$	the set of all complex numbers
$\mathbb{Z}$	the set of all integer numbers (positive, negative and zero)
$\mathbb{N}^{\pm}$	the set of all positive/negative integers ($0 \notin N^{\pm}$)
$\mathbb{Q}$	the set of all rational numbers ($p/q \in \mathbb{Z}$, where $p, q \in \mathbb{Z}$, and $q \neq 0$)
$\emptyset$	an empty set (one that contains no elements)

One can define a set in two ways (i) using a list, or (ii) using a rule. An example of defining a set using a list is the binary set A introduced above:

$$A = \{1, 0\}.$$

We can consider using a rule to define more complicated sets, for example

$$B = \{x | x \in \mathbb{R}^+, \text{ and } x > 2\},$$

where B is the set of real positive numbers greater than 2.

One other set that can be useful is the Universal Set, which contains all elements of interest, and is often denoted by U.

2.1 Relationships between sets

Now that elementary properties of sets have been introduced, it is useful to examine relationships between different sets.

2.1.1 Equivalence

Two sets are said to be equal or equivalent if the contain the same elements as each other and nothing more. For example if we consider $A = \{0, 1, 2, 3\}$ and $B = \{3, 2, 0, 1\}$, we can see that all of the elements of A all occur in B and similarly all of the elements in B occur in A. Thus it is clear that in this case $A = B$.

2.1.2 Subset

Given a large set of data A, it can be useful to identify groups of elements from the set with certain properties. If the selection of identified elements form the set B, then B is called a subset of A and we may write

$$B \subset A,$$

where the symbol $\subset$ denotes a proper subset.

For example if we consider the set $\mathbb{R}$, then this completely contains the set $\mathbb{R}^+$. We can consider $\mathbb{R}^+$ a subset of $\mathbb{R}$. This can be written as

$$\mathbb{R}^+ \subset \mathbb{R}.$$

This notation can be extended slightly to include the possibility that the two sets may be equivalent by replacing $\supset$ with $\subseteq$. Using this notation it follows that on could write $\mathbb{R}^+ \subseteq \mathbb{R}^+$ instead of $\mathbb{R}^+ = \mathbb{R}^+$ as the set $\mathbb{R}^+$ contains all of the elements necessary to define itself, however this does in general introduce the ambiguity that the two sets are not necessarily equal. The symbol $\subseteq$ is the set notation analog of the numerical symbol $\leq$.

2.1.3 Superset

We can consider a related problem to that discussed above, where this time round, the set A is a subset of B that is B contains all elements of A. In this case we call B the superset of A and may write this as

$$B \supset A,$$

where the symbol $\supset$ denotes superset.

For example if we consider the set $\mathbb{R}^+$, then this completely contained in set $\mathbb{R}$. So we can consider $\mathbb{R}$ a superset of $\mathbb{R}^+$, and we may write

$$\mathbb{R} \supset \mathbb{R}^+.$$

2.1.4 Intersection

Given two sets A and B each containing a number of elements. If some of the elements contained in A also appear in B, then we can identify those common elements as the intersection of A and B. We may write

$$A \cap B = \{x | x \in A \text{ and } x \in B\},$$

where the symbol $\cap$ denotes intersection. If the intersection between sets A and B is an empty set we can write $A \cap B = \emptyset$.

2.1.5 Union

Given two sets A and B each containing a number of elements. We may write the union of the two sets, which is a combination of all elements from both of the original sets, as

$$A \cup B = \{x | x \in A \text{ or } x \in B\},$$

where the symbol $\cup$ denotes union. This definition of $A \cup B$ implicitly includes the subset of elements $A \cap B$.

2.1.6 Complement

The complement of some set A is the set of interesting elements not contained within A itself, and this is denoted by placing a bar above the set name. For example the complement of A is $\bar{A}$. It follows that $A \cup \bar{A} = U$ as together A and its complement contain all of the interesting elements. There is a similarly trivial solution to the intersection of a set and its complement: $A \cap \bar{A} = \emptyset$. Finally the complement of the complement of a set, is the original set, i.e. $\bar{\bar{A}} = A$.

2.1.7 Set difference

Given the set A and the set B, it can be useful to identify the elements that only exist in A. The set notation for the set of elements that exist in A, but not B is $A \setminus B$. For example if $A = \{1, 2, 3\}$, and $B = \{3, 4, 5\}$, then $A \setminus B = \{1, 2\}$. Similarly the elements of B that only exist in B are given by $B \setminus A = \{4, 5\}$.

2.2 Venn diagrams

It is often useful to pictorially represent the relationships between two or more sets. This is achieved through the use of Venn diagrams. If we consider the previous case of two sets of data represented in an idealized 2-dimensional space, then we can draw a Venn diagram to illustrate the data. In this space both A and B do not overlap then the sets are distinct, and $A \cap B = \emptyset$. This situation is illustrated in Figure 2.1. If A and B overlap non-trivially in this space so that there is a finite intersection $A \cap B \neq \emptyset$ we would use a Venn diagram like the one shown in Figure 2.2.

Elements of the sets A and B that are in the intersection region originate either in set A or B, and without additional information it is not possible to distinguish which set a given element belongs to. Those elements of A and B that lie outside of the intersection region can be unambiguously identified as being in set A and not set B, or visa versa.

If we now consider the same two sets, and consider $A \cup B$, then this encompasses everything of interest, that is to say the union of A and B is simply the universal set. This pictorial representation can be particularly useful when one wants to visualize event classification for complicated sets of data in a high number of dimensions. Such a problem can be symbolically considered as that of trying to classify subsets of events

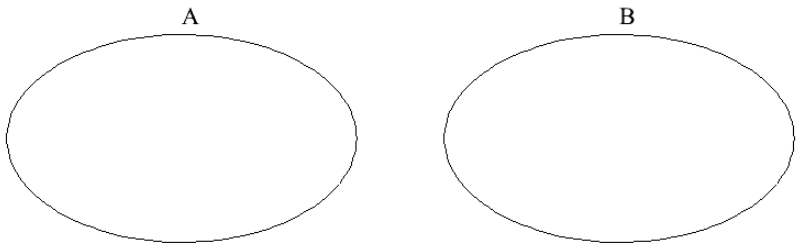

Figure 2.1: A Venn diagram illustrating two non-overlapping sets A and B.

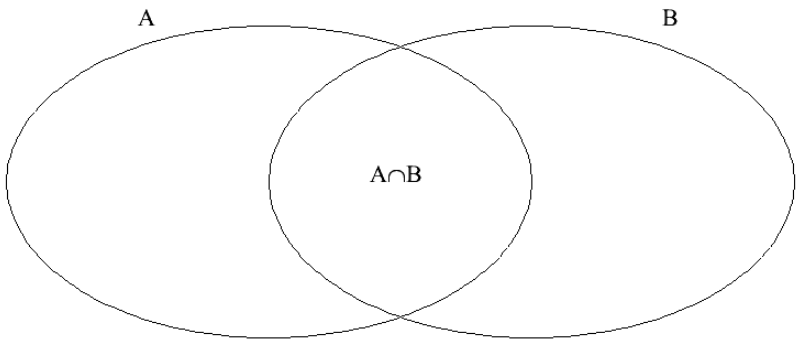

Figure 2.2: A Venn diagram illustrating two sets A and B, and their intersection $A \cap B$.

as coming from a specific target sample (or of a given type), with some overlap between the different target samples. When solving such a problem one wants to minimize any such overlap.

A third more complicated scenario can be discussed where we introduce a third set C. This case is shown in Figure 2.3. In this scenario there are seven distinct regions:

Region	Elements
(i)	Only existing in set A: $A \setminus B \setminus C = \{x \mid x \in A \text{ and } (x \notin B \text{ or } x \notin C)\}$
(ii)	Only existing in set B: $B \setminus A \setminus C = \{x \mid x \in B \text{ and } (x \notin A \text{ or } x \notin C)\}$
(iii)	Only existing in set C: $C \setminus A \setminus B = \{x \mid x \in C \text{ and } (x \notin A \text{ or } x \notin B)\}$
(iv)	$A \cap B$
(v)	$A \cap C$
(vi)	$B \cap C$
(vii)	$A \cap B \cap C$

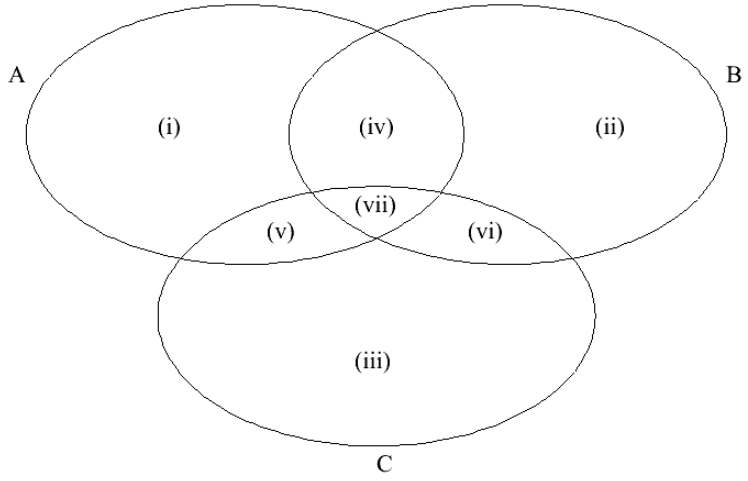

Figure 2.3: A Venn diagram illustrating three sets A, B, and C.

2.3 More useful rules

This section serves as a short summary of some of the rules already mentioned, and extends this to introduce more complicated notation that can be useful when manipulating sets.

Firstly there are the following common sense rules for a set A

$$\begin{aligned} A \cap A &= A, \\ A \cup A &= A. \end{aligned}$$

The $\cap$ and $\cup$ operators are commutative, so for two sets A and B it follows that

$$\begin{aligned} A \cap B &= B \cap A, \\ A \cup B &= B \cup A. \end{aligned}$$

Similarly these operators obey associativity which follows on naturally from their commutative nature

$$\begin{aligned} (A \cap B) \cap C &= A \cap (B \cap C), \\ (A \cup B) \cup C &= A \cup (B \cup C). \end{aligned}$$

These operators also follow a distributative law, so

$$\begin{aligned} A \cap (B \cup C) &= (A \cap B) \cup (A \cap C), \\ A \cup (B \cap C) &= (A \cup B) \cap (A \cup C). \end{aligned}$$

Two ways of defining an identity relation for a set A are

$$\begin{aligned} A &\cup \emptyset \\ A &\cap U. \end{aligned}$$

Familiarity with the notation introduced in this section is assumed when discussing data samples and techniques to analyse data in the following sections.