Sets and counting

Sharon Goldwater
Institute for Language, Cognition and Computation
School of Informatics, University of Edinburgh

DRAFT Version 0.95: 3 Sep 2015. **Do not redistribute without permission.**

1 Sets

1.1 Set notation, equivalence, cardinality

A SET is simply a collection of items, or ELEMENTS. We notate a set using braces, so

SET
ELEMENTS

$$S = \{Alice, Bob\} \tag{1}$$

means that set S contains the two elements *Alice* and *Bob*. (I'll use capital letters here for variables referring to sets.) When considering equivalence of sets, the order of elements is irrelevant, and repeated elements are ignored, so the following sets are equivalent:

$$\{Alice, Bob\} = \{Bob,\ Alice\} = \{Alice,\ Bob,\ Alice\} \tag{2}$$

The CARDINALITY of a set S, which we write as $|S|$, is just the number of elements in S. So, for S as defined in (1), $|S| = 2$.

CARDINALITY

1.2 Element notation, empty set, infinite sets

We write

$$\begin{aligned} Alice &\in S & (3)\\ Carla &\notin S & (4) \end{aligned}$$

to mean, respectively, "*Alice* is an element of S" (or "*Alice* is in S") and "*Carla* is not an element of S".

The EMPTY SET, which is the set containing no elements, can be written either $\{\}$ or $\emptyset$. It is *not* written $\{\emptyset\}$! That is a different set, which contains one element: the empty set.

EMPTY SET

We need not explicitly enumerate all of the elements in a set. Here are some examples of sets defined using properties:

$$\{x \mid x \text{ is an integer}\} \tag{5}$$

$$\{x \mid x \text{ is a three-letter word appearing in the Oxford English Dictionary}\} \tag{6}$$

$$\{x \mid x \text{ is a real number between 3 and 5 (inclusive)}\} \tag{7}$$

The $|$ symbol is read as "such that" (some people use a : symbol instead), so (5) would be read as "the set containing all elements x such that x is an integer", i.e., the set of all integers. This particular set comes up often, so we also use the shorthand notation $\mathbb{Z}$ to refer to it. Similarly, $\mathbb{R}$ is shorthand for the set of all real numbers, so we could also write (7) as $\{x \mid x \in \mathbb{R} \text{ and } 3 \leq x \leq 5\}$.

Hopefully you will have noticed that not all of these sets contain a finite number of elements. Both $\mathbb{Z}$ and $\mathbb{R}$ are infinite, and so is the set in (7). [1]

[1] Although both $\mathbb{Z}$ and $\mathbb{R}$ are infinite, they are infinite in different ways. If you're interested to know more, look up the definitions of *countably infinite* and *uncountably infinite* sets.

1.3 Subsets, intersection, union, set difference

SUBSET

For sets A and B, we say that A is a SUBSET of B iff[2] every element of A is also an element of B. We write this as

$$A \subseteq B \tag{8}$$

and we can also write $A \not\subseteq B$ to mean "A is not a subset of B".

Sometimes you might also see the following notation:

$$A \supseteq B \tag{9}$$

SUPERSET

which is read as "A is a SUPERSET of B" and is equivalent to $B \subseteq A$.

PROPER SUBSET

Notice that according to the definition of subset, every set is a subset of itself. If A is a subset of B and also does not equal B (i.e., there is at least one element in B that is not in A), then we say A is a PROPER SUBSET of B, written as

$$A \subset B \tag{10}$$

and similarly, $A \not\subset B$ means "A is not a proper subset of B".

INTERSECTION

DISJOINT

The INTERSECTION of sets A and B, written $A \cap B$, is the set of all elements that are members of both A and B. If $A \cap B = \emptyset$ (i.e., they have no elements in common), then we say A and B are DISJOINT.

UNION

The UNION of sets A and B, written $A \cup B$, is the set of all elements that are members of *either* A or B.

SET DIFFERENCE

The SET DIFFERENCE $A - B$ is the set of elements that are members of A but not of B.

1.4 Cartesian product

CARTESIAN PRODUCT

Suppose we have two (possibly equal) sets A and B. The CARTESIAN PRODUCT of A and B, written $A \times B$, is defined as:

$$A \times B = \{(x,y) \mid x \in A, y \in B\} \tag{11}$$

ORDERED PAIRS

That is, $A \times B$ the set of all possible ORDERED PAIRS (x,y) where x is an element of A and y is an element of B. As implied by the name, the ordering matters in an ordered pair (unlike in a set), so for example

$$\{a,b\} \times \{a,b\} = \{(a,a),(a,b),(b,a),(b,b)\} \tag{12}$$

not just {(a, a), (a, b), (b, b)}. Similarly, $\{a,b\} \times \{c,d\}$ includes (a,c) but not (c,a).

1.5 Powerset

POWERSET

The POWERSET of a set S is the set of all possible subsets of S, including the empty set and S itself. For example, if $A = \{1,2,3\}$, then the powerset of A is $\{\{\},\{1\},\{2\},\{3\},\{1,2\},\{1,3\},\{2,3\},\{1,2,3\}\}$. Notice that whereas A has 3 elements, the powerset has 8, or 2^3 elements. In fact, for any set S with n elements, the powerset of S has 2^n elements (and we'll explain why in the Counting section below!). For this reason, the powerset of S is often written as 2^S.

1.6 Excercises

[2] *iff* means "if and only if"

Exercise 1.1
Say whether each statement is true or false.

a) $\{a,b,c,d\} = \{c,b,d,a\}$

b) If we define $A = \{c,a,b,c,\}$, then $|A| = 3$.

c) $|\mathbb{Z}|$ is a finite number.

d) $3.2 \in \mathbb{Z}$

e) 3.2 is in $\mathbb{R}$

f) $\{b\} \notin \{c,b,d,a\}$

g) $\emptyset = \{\}$

h) $\mathbb{Z} \subseteq \mathbb{R}$

i) $\{a,b,c,d\} \subset \{a,b,c,d\}$

j) The set $\{\{a,b\}\}$ contains 2 elements.

Exercise 1.2
What set is specified by each of the following expressions?

a) $\{a,b\} \cup \{a,c,d\}$

b) $\{a,b\} \cap \{a,c,d\}$

c) $\{a,b\} \cap \{c,d\}$

d) $\{a,b\} \cup \emptyset$

e) $\{a,b\} \times \{Alice, Bob, Carla\}$

f) 2^A, where $A = \{red, green\}$.

Exercise 1.3
Describe each of the following sets in words.

a) $\{x \mid \frac{x}{2} \in \mathbb{Z}\}$

b) $\{x \mid 2x \in \mathbb{Z}\}$

c) $\{x \mid 2x \in \mathbb{R}\}$

d) $\mathbb{Z} \cap \mathbb{R}$

e) $\mathbb{Z} - S$, where S is $\{x \mid \frac{x}{2} \in \mathbb{Z}\}$

f) $\mathbb{Z} - S$, where S is $\{x \mid 2x \in \mathbb{Z}\}$

Exercise 1.4
For each statement, either explain why it is true, or give a counterexample to prove it is false.

a) $\{a,b,c,d\} \not\subseteq \{a,b\} \cup \{c,b\} \cup \{d\}$

b) $\{x \mid \frac{x}{2} \in \mathbb{Z}\} \subseteq \mathbb{Z}$

c) $\{x \mid 2x \in \mathbb{Z}\} \subseteq \mathbb{Z}$

d) For any two sets A and B, if $A \not\subseteq B$, then $B \subseteq A$.

e) For any two sets A and B, $A \times B = B \times A$.

f) For any two sets A and B, if $A = B$, then $A \times B = B \times A$.

2 Counting

Of course we all know how to count things that we can enumerate one by one. But when writing algorithms or dealing with probabilities, we often need to be able to count things without explicitly enumerating them. For example, we might want to know many times will a line of code be executed (to estimate how fast a program will run), or how many different 5-word sequences are possible given a particular vocabulary (so that we know how much computer memory is required to store them). Counting is also the basis of one view of probability theory.

2.1 Counting by multiplying

Many things we might want to count can be viewed as different possible outcomes of a multi-step procedure where each step determines part of the final outcome. If the number of choices at step i doesn't depend on the outcome of the previous step $i-1$, then we can count the total number of outcomes of steps $1 \dots N$ by simply multiplying together the number of outcomes at each step. In the simplest case, when the number of choices at each step is the same (say, n), then the total number of possible outcomes is

$$n^N \tag{13}$$

Example 2.1.1. How many different 3-character sequences can be constructed using only the characters a and b?

Solution: Each sequence can be thought of as requiring 3 steps: choose the first character, then the second character, then the third character. Each of the three steps has two possible outcomes, so the total number of outcomes is $2 \cdot 2 \cdot 2$, or 2^3.

If it isn't immediately clear why this rule works, here is a slightly more detailed explanation. At the first step, there are two possible outcomes: a and b. For each of those two outcomes, there are a further two outcomes at the second step (so, a becomes either aa or ab, b becomes either ba or bb). That gives us $2 \cdot 2 = 4$ outcomes after two steps, and again in the next step there are two further outcomes for each of those four, yielding $4 \cdot 2 = 8$ altogether.

Example 2.1.2. Suppose we have a set S of size N. How many different subsets of S are there?

Solution: I already told you in the section on powersets that the answer is 2^N. But now I will explain why. Formulate the problem as follows: to create a subset S' of S, consider each of the N elements of S in turn. For each element, specify one of two choices: either the element is in S', or it is not. The number of distinct subsets is equal to the number of distinct sequences of choices. Two choices for each of N elements yields 2^N possible subsets.

Depending on the problem, we may not always have the same number of possibilities at each step. But, if the number of possible outcomes at step i (call it n_i) does not depend on the outcome of the previous step, we can still just multiply the possibilities at each step to get the total number of possibilities:

$$n_1 \cdot n_2 \cdot \ldots \cdot n_N \tag{14}$$

Example 2.1.3. Let's say we have a system with ID numbers containing exactly two letters (a-z) followed by three numbers (0-9). How many different ID numbers are there?

Solution: Here, we have 26 choices for each of the first two steps, and 10 choices for each of the following three steps, for a total of $(26^2)(10^3)$ possible ID numbers.

Warning: There are cases that may at first look similar to those given here, but where the simple multiplication rule in (14) doesn't apply.

Example 2.1.4. Suppose we want a password containing exactly 3 digits (0-9) and 8 letters (a-z), but in any order. How many possible passwords are there?

Solution: This is not a straightforward question to answer using the multiplication rule. The first three characters each have 36 possibilities, but once we have chosen those three, the number of possibilities for the next character could be either 26 (if the first three characters are digits) or 36 (otherwise). Things get even more complicated for the fifth character, where there are only 26 choices if *any* three of the preceding characters are digits. And so forth. There are ways to count outcomes in cases like this, but the most important point here is just to consider whether it's ok to use simple multiplication or not.

2.2 Permutations and factorials

A special case of using the multiplication rule in (14) comes up when we want to compute the number of PERMUTATIONS (orderings) of a set.

PERMUTATIONS

Example 2.2.1. How many different permutations are there of the numbers in the set $\{0,1,2\}$?

Solution: Each permutation is just an ordered sequence of the three numbers in the set. We can solve the problem by noting that for each permutation, we start by choosing one of the three elements to be the first number in the sequence. Once we have done that, there are two elements left to choose from, and once we have chosen again, there is only one element left. So, the total number of permutations is 3*2*1.

From this example it should be clear that in general, the number of permutations of a set of n elements is

$$n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot 2 \cdot 1 = n! \tag{15}$$

where $n!$ is read as "n factorial".

Warning: Formula (15) is correct only if we are considering permutations of a *set*, i.e. there are no repeated elements. If instead we are considering permutations of an ordered sequence with repetitions, like $(1,2,5,1,3)$, then some of the permutations are actually the same as others and should not be counted separately. In this example, it doesn't matter which 1 is in which position, but the above formula assumes each element is distinct so it will count each sequence twice, once with each 1 in each position. The correct number of distinct permutations in this case is 5!/2, and can be found more generally by computing the number of permutations as if each one is distinct, and then dividing by the number of identical sequences (which can be found using similar techniques to those described here). I won't go into more detail except to say that counting permutations and combinations can become quite tricky in some situations! Hopefully you won't have to deal with them much but if you are thinking about permutations, you should at least make sure you know whether you need to worry about repeated elements or not.

2.3 Exercises

For each question, first consider whether it is possible to solve straightforwardly using the simple multiplication methods above. If not, can you think of a clever way to break down the problem or look at it differently in order to solve it anyway? (Remember, the most important part is recognizing whether or not you can use the straightforward method; but you might also like to think about how to get the solution anyway.)

Exercise 2.1

Using the 26 lowercase letters of English, how many different 6-character strings (character sequences) are possible if letters can be reused? What if letters cannot be reused?

Exercise 2.2

Consider a language where all words consist of alternating consonant-vowel sequences, starting with a consonant. (Some real languages, like Japanese and Hawaiian, are almost this simple.) If there are 5 vowels $\{a,e,i,o,u\}$ and 15 consonants, how many 6-character words are possible that obey the consonant-vowel restriction?

Exercise 2.3

Many natural language processing and speech recognition systems use *n-gram language models* which, if implemented naively, require storing the probability of each possible sequence of n words in the language. Suppose our dictionary lists 15,000 words. If storing a single probability takes 1 byte (this is an underestimate), how much storage space would we need for this naive n-gram model if $n = 3$? $n = 4$? $n = 5$? (Remember, 1 gigabyte = 1 billion bytes.)

Exercise 2.4

Consider a language in which every three-character word has *exactly* one vowel. The number of vowels and consonants is the same as in question 2.2. How many different three-character words are possible in this language?

Exercise 2.5

In English, every three-character word has *at least* one vowel. Using the 26 lowercase letters of English, with $\{a,e,i,o,u\}$ as the vowels (we simplify by assuming y is always a consonant), how many 3-letter strings are there with at least one vowel?

3 Solutions to selected exercises

Solution 1.1

a) True.

b) True. Since duplicates are ignored in sets, A can also be written as $\{a,b,c\}$, so its cardinality is 3.

c) False. $\mathbb{Z}$ is the set of all integers, of which there are an infinite number.

d) False. $\mathbb{Z}$ is the set of all integers and doesn't contain non-integer values.

e) True.

f) True. The element b *is* a member of the set, but the element $\{b\}$ is not.

g) True. $\emptyset$ and $\{\}$ are different ways to write the empty set.

h) True.

i) False. But it would be correct to write $\{a,b,c,d\} \subseteq \{a,b,c,d\}$.

j) False. This set contains a single element, the set $\{a,b\}$. *That* set contains two elements.

Solution 1.3

a) The set of all even integers.

b) The set of all integers and integers plus 0.5 (that is, any number that is half of an integer).

c) The set of all real numbers. (Note that any number that is half a real number is also a real number).

d) The set of integers.

e) The set of odd numbers.

f) The empty set.

Solution 2.1

If letters can be reused, then there are 26 choices for each of the six characters, so 26^6 (or about 300 million) possible strings. If letters cannot be reused, then there are $26 \cdot 25 \cdot 24 \cdot 23 \cdot 22 \cdot 21$ (which we can also write as $\frac{26!}{20!}$, or about 165 million) possible strings.

Solution 2.3

For $n = 3$, we need $15,000^3$ or about 3.37×10^{12} (3.37 trillion) probabilities (or bytes), i.e., 3,370 gB of storage. For $n = 4$, we need $15,000^4$ or about 5.06×10^{16} bytes ($50,600,000$ gB), and for $n = 5$, we need $15,000^5$ or about 7.59×10^{20} bytes (759 billion gB). Hopefully you can see why we call this a naive method.

Solution 2.5

This problem cannot be solved directly by multiplying together the number of choices for each position, because the number of choices in the third position depends on what happened in the previous positions: if there was already a vowel, then there are 26 choices, otherwise only 5. However, there is a clever way to solve this problem: notice that the number of strings with at least one vowel is equal to the number of *all* strings (26^3) minus the number of strings with *no* vowels (21^3). So the answer is $26^3 - 21^3$, or 8315.