Queen Mary
University of London

MAS 108 **Probability I**

Notes 2 **Autumn 2005**

Sampling

I have four pens in my satchel; they are red, green, blue, and purple. I take out a pen and lay it on the desk; each pen has the same chance of being selected. In this case, $\mathcal{S} = \{R, G, B, P\}$, where R means 'red pen chosen' and so on. If A is the event 'red or green pen chosen', then

$$P(A) = \frac{|A|}{|\mathcal{S}|} = \frac{2}{4} = \frac{1}{2}.$$

More generally, if I have a set of N objects and choose one, with each one equally likely to be chosen, then each of the N outcomes has probability $1/N$, and an event consisting of m of the outcomes has probability m/N.

What if we choose more than one pen? We have to be more careful to specify the sample space.

First, we have to say whether we are

- *sampling with replacement*, or
- *sampling without replacement*.

Sampling with replacement means that we choose a pen, note its colour, put it back and shake the satchel, then choose a pen again (which may be the same pen as before or a different one), and so on until the required number of pens have been chosen. If we choose two pens with replacement, the sample space is

$$\begin{array}{cccc}
\{RR, & RG, & RB, & RP, \\
GR, & GG, & GB, & GP, \\
BR, & BG, & BB, & BP, \\
PR, & PG, & PB, & PP\}
\end{array}$$

The event 'at least one red pen' is $\{RR, RG, RB, RP, GR, BR, PR\}$, and has probability $7/16$.

In general, if we choose n items from a set Ω of size N, and the sampling is done with replacement, then the sample space $\mathcal{S}$ consists of all ordered n-tuples of the form $(\omega_1, \omega_2, \ldots, \omega_n)$, where ω_i denotes the object taken out on the i-th occasion.

Sampling without replacement means that we choose a pen but do not put it back, so that our final selection cannot include two pens of the same colour. In this case, the sample space for choosing two pens is

$$\begin{array}{cccc} \{ & RG, & RB, & RP, \\ GR, & & GB, & GP, \\ BR, & BG, & & BP, \\ PR, & PG, & PB & \} \end{array}$$

and the event 'at least one red pen' is $\{RG, RB, RP, GR, BR, PR\}$, with probability $6/12 = 1/2$.

Now there is another issue, depending on whether we care about the order in which the pens are chosen. We will only consider this in the case of sampling without replacement. Sometimes it doesn't really matter whether we choose the pens one at a time or simply take two pens out of the drawer; we are not always interested in which pen was chosen first. If we are not interested then the sample space is

$$\{\{R,G\},\{R,B\},\{R,P\},\{G,B\},\{G,P\},\{B,P\}\},$$

containing six elements. (Each element is written as a set since, in a set, we don't care which element is first, only which elements are actually present. So the sample space is a set of sets!) The event 'at least one red pen' is $\{\{R,G\},\{R,B\},\{R,P\}\}$, with probability $3/6 = 1/2$. We should not be surprised that this is the same as in the previous case.

If order is important, the sample space $\mathcal{S}$ still consists of ordered n-tuples of the form $(\omega_1, \omega_2, \ldots, \omega_n)$, but now all of the ω_i must be different. If order is not important then $\mathcal{S}$ consists of all subsets of Ω of size n.

There are formulae for the sample space size in these three cases. These involve the following expressions:

$$\begin{aligned} N! &= N(N-1)(N-2)\cdots 1 \\ {}^N P_n &= N(N-1)(N-2)\cdots(N-n+1) \\ {}^N C_n &= {}^N P_n / n! \end{aligned}$$

Note that $N!$ is the product of all the whole numbers from 1 to N; and

$${}^N P_n = \frac{N!}{(N-n)!},$$

so that

$$ {}^{N}C_n = \frac{N!}{n!(N-n)!}. $$

Theorem The number of selections of n objects from a set of N objects is given in the following table.

	with replacement	without replacement
ordered sample	N^n	${}^{N}\mathrm{P}_n$
unordered sample		${}^{N}\mathrm{C}_n$

In fact the number that goes in the empty box is ${}^{N+n-1}\mathrm{C}_n$, but this is much harder to prove than the others, and you are very unlikely to need it.

Here are the proofs of the other three cases. First, for sampling with replacement and ordered sample, there are N choices for the first object, and N choices for the second, and so on; we multiply the choices for different objects. (Think of the choices as being described by a branching tree.) The product of n factors each equal to N is N^n.

For sampling without replacement and ordered sample, there are still N choices for the first object, but now only $N-1$ choices for the second (since we do not replace the first), and $N-2$ for the third, and so on; there are $N-k+1$ choices for the kth object, since $k-1$ have previously been removed and $N-(k-1)$ remain. As before, we multiply. This product is the formula for ${}^{N}\mathrm{P}_n$.

For sampling without replacement and unordered sample, think first of choosing an ordered sample, which we can do in ${}^{N}\mathrm{P}_n$ ways. But each unordered sample could be obtained by drawing it in $n!$ different orders. So we divide by $n!$, obtaining ${}^{N}\mathrm{P}_n/n! = {}^{N}\mathrm{C}_n$ choices.

In our example with the pens, the numbers in the three boxes are $4^2 = 16$, ${}^{4}\mathrm{P}_2 = 12$, and ${}^{4}\mathrm{C}_2 = 6$, in agreement with what we got when we wrote them all out.

Note that, if we use the phrase ‘sampling without replacement, ordered sample’, or any other combination, we are assuming that *all outcomes are equally likely*.

Example The names of the seven days of the week are placed in a hat. Three names are drawn out; these will be the days of the Probability I lectures. What is the probability that no lecture is scheduled at the weekend?

Here the sampling is without replacement, and we can take it to be either ordered or unordered; the answers will be the same. For ordered samples, the size of the sample space is ${}^{7}\mathrm{P}_3 = 7\cdot 6\cdot 5 = 210$. If A is the event ‘no lectures at weekends’, then A occurs precisely when all three days drawn are weekdays; so $|A| = {}^{5}\mathrm{P}_3 = 5\cdot 4\cdot 3 = 60$. Thus, $P(A) = 60/210 = 2/7$.

If we decided to use unordered samples instead, the answer would be ${}^{5}\mathrm{C}_3/{}^{7}\mathrm{C}_3$, which is once again $2/7$.

Example Ten coins are tossed: each is equally likely to come down heads or tails. What is the probability that we get exactly three heads?

This is equivalent to sampling from $\{H,T\}$ *with replacement*, so $|\mathcal{S}| = 2^{10} = 1024$.

Let A be the event 'exactly three heads'. Then $|A|$ is equal to the number of ways of choosing 3 things from 7, which is

$$^{10}\mathrm{C}_3 = \frac{10!}{3!\,7!} = \frac{10\times 9\times 8}{3\times 2\times 1} = 120.$$

If all outcomes are equally likely then $P(A) = 120/1024 = 15/128 \approx 0.117$.

Example I have 10 coins in my pocket; 3 are copper and 7 are silver. I take out 4 coins, one after another. Let

$$\begin{aligned} D &= \text{'2 silver followed by 2 copper'} \\ E &= \text{'all 4 are silver'} \\ F &= \text{'2 silver and 2 copper, in any order'.} \end{aligned}$$

This is sampling without replacement.

For event D, the order matters, so we consider ordered samples. Then $|\mathcal{S}| = {}^{10}\mathrm{P}_4 = 10\times 9\times 8\times 7$. For event D we must choose an ordered sample of 2 from the 7 silver coins followed by an ordered sample of 2 from the 3 copper coins, so $|D| = {}^{7}\mathrm{P}_2 \times {}^{3}\mathrm{P}_2 = 7\times 6\times 3\times 2$. Therefore $P(D) = |D|/|\mathcal{S}| = 1/20$.

For event E, we choose an ordered sample of 4 from the 7 silver coins,so $|E| = {}^{7}\mathrm{P}_4 = 7\times 6\times 5\times 4$ and $P(E) = |E|/|\mathcal{S}| = 1/6$.

Event F is like event D except that we have to choose which 2 of the 4 positions should have the silver coins. There are ${}^{4}\mathrm{C}_2$ ways of doing this, which is 6, so $P(F) = 6P(D) = 6/20 = 3/10$.

If we didn't want to know about event D then we could use unordered samples. Then $|\mathcal{S}| = {}^{10}\mathrm{C}_4$ and $|E| = {}^{7}\mathrm{C}_4$ so

$$P(E) = \frac{7!}{4!\,3!} \times \frac{6!\,4!}{10!} = \frac{1}{6}.$$

Also, $|F| = {}^{7}\mathrm{C}_2 \times {}^{3}\mathrm{C}_2$, because each choice of two silver coins can be combined with each choice of two copper coins. Thus

$$P(F) = \frac{7!}{2!\,5!} \times \frac{3!}{2!\,1!} \times \frac{4!\,6!}{10!} = 3/10.$$

The results for E and F are the same for ordered and unordered samples, as they should be.

Summary: In a sampling problem, you should first read the question carefully and decide whether the sampling is with or without replacement. If it is without replacement, decide whether the sample is ordered (e.g. does the question say anything about the first object drawn?). If so, then use the formula for ordered samples. If not, then you can use either ordered or unordered samples, whichever is convenient; they should give the same answer. (Usually it is easier to use unordered samples whenever you can.) If the sample is with replacement, or if it involves throwing a die or coin several times, then use the formula for sampling with replacement.