3 — DISCRETE DISTRIBUTIONS

It is always helpful when solving a problem to be able to relate it to a problem whose solution is already known and understood. In probability theory, many problems turn out to be special cases of standard examples. The most common standard examples are the well-known distributions.

Discrete Distributions

In simple terms, a *distribution* is an indexed set of probabilities whose sum is 1.

For the moment, discussion will be restricted to cases where there is a single discrete random variable X whose value r runs from zero upwards and serves as the index. It is possible to think of r running from 0 to ∞ where in most cases the indexed probability is zero.

A distribution may be expressed by a table or a function or graphically. Consider the distribution associated with a fair die.

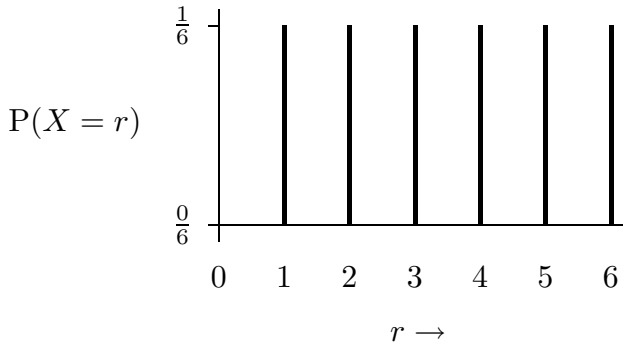
A tabular representation of the distribution is:

	$r \rightarrow$						
X	0	1	2	3	4	5	6
$\mathrm{P}(X = r)$	$\frac{0}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

A functional representation of the distribution is:

$$\mathrm{P}(X = r) = \begin{cases} \dfrac{1}{6}, & \text{if } r \in \mathbb{N} \wedge 1 \leqslant r \leqslant 6 \\ 0, & \text{otherwise} \end{cases}$$

A graphical representation of the distribution is:



Of the three representations, only the function makes it pedantically clear that unless $1 \leqslant r \leqslant 6$ the probability is zero.

The Uniform Distribution

When all the non-zero probabilities are the same and are indexed by a contiguous sequence of values of r, the distribution is said to be a *Uniform distribution*.

The behaviour of a fair die is an example of a Uniform distribution. All six non-zero probabilities are the same and the index r for these probabilities has the contiguous values 1, 2, 3, 4, 5 and 6.

One can imagine a fair die with a different number of faces. Consider a fair tetrahedral die whose faces happen to be numbered 5, 6, 7 and 8. The four probabilities are all $\frac{1}{4}$.

There is actually a family of distributions and the description:

$$\text{Uniform}(m, n)$$

is used to refer to the general case; m and n are the start and stop values of r and are called the *parameters* of the distribution.

A random variable whose value represents the outcome of throwing an ordinary fair die is said to be 'distributed Uniform(1,6)'. If the value represents the outcome of throwing the curious tetrahedral die the random variable is distributed Uniform(5,8).

In the general case there are $n - m + 1$ values for the index and, given the equiprobable nature of the Uniform distribution, the probabilities are all $1/(n - m + 1)$. The functional representation of the general case is:

$$\mathrm{P}(X = r) = \begin{cases} \dfrac{1}{n - m + 1}, & \text{if } r \in \mathbb{N} \wedge m \leqslant r \leqslant n \\ 0, & \text{otherwise} \end{cases}$$

It is good practice always to check that the probabilities in a distribution sum to 1. With the general Uniform distribution the check is straightforward:

$$\sum_{r=m}^{n} \frac{1}{n - m + 1} = \frac{n - m + 1}{n - m + 1} = 1$$

The Triangular Distribution

Many standard distributions will be discussed but one which has already been noted but not until now given a name is the *Triangular distribution*. The functional representation of an example of this distribution was given on page 1.7 as:

$$\mathrm{P}(X = r) = \begin{cases} \dfrac{r}{21}, & \text{if } r \in \mathbb{N} \wedge 1 \leqslant r \leqslant 6 \\ 0, & \text{otherwise} \end{cases}$$

As with all discrete distributions it satisfies the informal requirement of being a set of indexed probabilities whose sum is 1:

$$\sum_{r=0}^{6} \frac{r}{21} = \frac{1 + 2 + 3 + 4 + 5 + 6}{21} = 1$$

The Binomial Distribution

If the probability of a boy is p and the probability of a girl is q (where $p + q = 1$) it has already been shown that the four probabilities for two children sum to 1 as:

$$p^2 + pq + qp + q^2 = 1$$

The four terms are the probabilities of BB, BG, GB and GG respectively. Since $pq = qp$ the four terms can conveniently be reduced to three:

$$p^2 + 2pq + q^2 = 1$$

Using Binomial coefficients, this can be written as:

$$\binom{2}{0} p^2 q^0 + \binom{2}{1} p^1 q^1 + \binom{2}{2} p^0 q^2 = 1$$

The middle term is the probability of one boy and one girl without regard to order.

With four children the equivalent sum is:

$$p^4 + 4p^3 q + 6p^2 q^2 + 4pq^3 + q^4 = 1$$

As with the previous example, the term which represents all boys is first and the term which represents all girls is last. Reversing the order gives:

$$q^4 + 4pq^3 + 6p^2 q^2 + 4p^3 q + p^4 = 1$$

Using Binomial coefficients, this can be written as:

$$\binom{4}{0} p^0 q^4 + \binom{4}{1} p^1 q^3 + \binom{4}{2} p^2 q^2 + \binom{4}{3} p^3 q^1 + \binom{4}{4} p^4 q^0 = 1$$

The five terms are, respectively, the probabilities of having 0, 1, 2, 3 and 4 boys in a family of four children without regard to order.

Using the random variable X to refer to the number of boys:

$$\mathrm{P}(X = r) = \begin{cases} \binom{4}{r} p^r q^{4-r}, & \text{if } r \in \mathbb{N} \wedge 0 \leqslant r \leqslant 4 \\ 0, & \text{otherwise} \end{cases}$$

This is an indexed set of probabilities whose sum is 1 and so is a distribution. It is an example of the Binomial distribution as it applies to 4 children. The term $\binom{4}{r} p^r q^{4-r}$ begins with $\binom{4}{r}$ (the number of ways of there being r boys in 4 children) and this is multiplied by p^r (the probability of having r boys) and q^{4-r} (the probability that the remaining $4 - r$ children are girls).

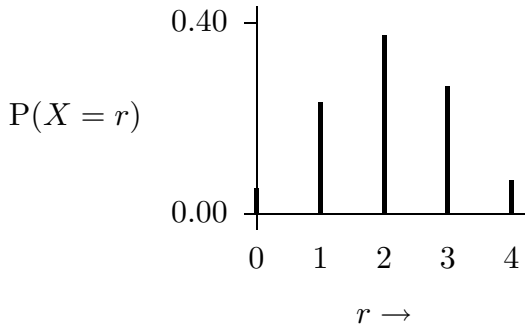
As a distribution it is not completely specified until a value is given for p (and hence q) as well as saying how many children there are.

Taking $p = 0.515$ and $q = 0.485$, and for once not using fractions, the probabilities may be tabulated thus:

	$r \rightarrow$				
X	0	1	2	3	4
$\mathrm{P}(X = r)$	0.055	0.235	0.374	0.265	0.071

It is easy to check that the five values sum to 1. Notice also that when there are four children of the same sex, the probability that they are all boys is noticeably greater than the probability that they are all girls.

A graphical representation of the distribution is:



As with the Uniform distribution, the Binomial distribution is a family of distributions, indeed a family of families. The description

$$\mathrm{Binomial}(n, p)$$

is used to refer to the general case; n and p are the parameters. In the example just considered, the random variable X is said to be distributed $\mathrm{Binomial}(4, 0.515)$.

The (general) Binomial distribution applies to many circumstances where there is a finite number of entities each of which may be one of two possibilities. A random variable whose value represents the number of heads which appear when 4 fair coins are tossed is distributed $\mathrm{Binomial}(4, \frac{1}{2})$. If you have 4 machines each of which has a 1% probability of failing in a given time interval, the appropriate distribution is $\mathrm{Binomial}(4, \frac{1}{100})$.

In general, where a random variable X is distributed $\mathrm{Binomial}(n, p)$, the probability $\mathrm{P}(X = r)$ is:

$$\mathrm{P}(X = r) = \begin{cases} \binom{n}{r} p^r q^{n-r}, & \text{if } r \in \mathbb{N} \wedge 0 \leqslant r \leqslant n \\ 0, & \text{otherwise} \end{cases}$$

The sum of these $n+1$ probabilities is:

$$\binom{n}{0}p^0q^n + \binom{n}{1}p^1q^{n-1} + \binom{n}{2}p^2q^{n-2} + \cdots + \binom{n}{r}p^rq^{n-r} + \cdots + \binom{n}{n}p^nq^0$$

It is immediately clear from the Binomial theorem that the sum is 1 since the expression can be rewritten:

$$\sum_{r=0}^{n}\binom{n}{r}p^rq^{n-r} = (q+p)^n = 1$$

Note that $(q+p)^n$ is shown (in preference to $(p+q)^n$) since, reading from left to right, the terms in its expansion are normally written with ascending powers of p (compare with the expansion of $(x+y)^n$ on page 2.12). The key point is that the general case satisfies the informal requirement of having a set of indexed probabilities whose sum is 1.

A Point to Ponder

In the context of children, the particular term $\binom{n}{r}p^rq^{n-r}$ is the probability of there being r boys and $n-r$ girls in a family of n children. The coefficient $\binom{n}{r}$ is the number of ways in which n children may divide as n boys and $n-r$ girls and this coefficient multiplies the probability of one such case.

With 4 children the probability of there being 1 boy (and 3 girls) is $\binom{4}{1}p^1q^3 = 4pq^3$. This is really the sum:

$$pqqq + qpqq + qqpq + qqqp = 4p^1q^3$$

The multiplication theorem holds for each term because the boy-girl events are independent and the addition rule holds overall because the B+3G events are mutually exclusive. Since each of the four separate B+3G events has the same probability, the addition amounts to multiplying the probability of one of them by 4.

The previous paragraph is worth pondering. Are the boy-girl events really independent? If a couple have three girls would you really put the same odds on the next child being a boy as you would if they were expecting their first baby? Is the value of p in $qqqp$ really the same as the p in $pqqq$? Demographic experts generally agree that it is.

The Trinomial Distribution

The Binomial distribution applies when considering entities which have two states, boy-girl, heads-tails, working-broken and so on.

There are circumstances when three states are appropriate. For example a bicycle has three principal states: Parked, Ridden or Pushed and traffic lights can be Red, Green or Changing. There was a brief period when ternary computers were thought worth exploring: voltages would have been positive, zero or negative.

There are many three-state examples in genetics. It would be fanciful to imagine that children came in three sexes but, if both parents have blood group AB, then each offspring will necessarily have blood group, AA, AB or BB and there are known probabilities for each.

One can reasonably ask the probability of such parents with four children having two children AA, one AB and one BB. Problems involving entities which have three states lead naturally to a discussion of the *Trinomial distribution*.

Before proceeding, consider a summary of the case of a family of four children and the two-state boy-girl analysis:

- There are two salient probabilities p and q; these are the probabilities of a child being a boy or a girl respectively. Necessarily $p + q = 1$.
- If order is taken into account, there are $2^4 = 16$ ways of having four children being GGGG, GGGB, ..., BBBB.
- If order is taken not into account, there are $4 + 1 = 5$ ways of having four children which can be listed as 0 boys, 1 boy, ..., 4 boys.
- The probability of there being r boys is:

$$\binom{4}{r} p^r q^{4-r} = \frac{4!}{r!\,(4-r)!}\, p^r q^{4-r}$$

Suppose the three blood groups are labelled a, b and c and are regarded as three sexes:

- There are three salient probabilities; call these p_a, p_b and p_c and note that necessarily $p_a + p_b + p_c = 1$.
- If order is taken into account, there are $3^4 = 81$ ways of having four children because each may be one of three possibilities.
- If order is taken not into account, the number of ways of having four children turns out to be 15 and these ways are most easily presented in a triangular array:

$$\begin{array}{ccccc} & & aaaa & & \\ & aaab & & aaac & \\ aabb & & aabc & & aacc \\ abbb & & abbc & abcc & accc \\ bbbb & bbbc & bbcc & bccc & cccc \end{array}$$

- There is no direct parallel to 'the probability of there being r boys' because of the complication introduced by having three possibilities...

Consider a rewrite of the expression for the probability of there being r boys in the Binomial case:

$$\frac{4!}{r!\,(4-r)!}\, p^r q^{4-r} = \frac{4!}{r_a!\,r_b!}\, {p_a}^{r_a} {p_b}^{r_b}$$

The boy-girl probabilities p and q have been replaced by p_a and p_b and the boy-girl numbers r and $4 - r$ have been replaced by r_a and r_b. Clearly $p_a + p_b = 1$ and $r_a + r_b = 4$.

This latter expression generalises to the Trinomial case:

$$\frac{4!}{r_a!\,r_b!\,r_c!}\,{p_a}^{r_a}{p_b}^{r_b}{p_c}^{r_c} \quad \text{where } p_a+p_b+p_c=1 \text{ and } r_a+r_b+r_c=4$$

Here, within the total of four children, r_a, r_b and r_c are the numbers with blood groups AA, AB and BB respectively. The only possibilities for r_a, r_b and r_c are 3 permutations of (0,0,4), 6 permutations of (0,1,3), 3 permutations of (0,2,2) and 3 permutations of (1,1,2) making a total of 15 possibilities.

These 15 possibilities can be plugged into the expression to give the 15 probabilities:

$$\begin{array}{ccccccccc}
 & & & & p_a^4 & & & & \\
 & & & 4p_a^3p_b & & 4p_a^3p_c & & & \\
 & & 6p_a^2p_b^2 & & 12p_a^2p_bp_c & & 6p_a^2p_c^2 & & \\
 & 4p_ap_b^3 & & 12p_ap_b^2p_c & & 12p_ap_bp_c^2 & & 4p_ap_c^3 & \\
p_b^4 & & 4p_b^3p_c & & 6p_b^2p_c^2 & & 4p_bp_c^3 & & p_c^4
\end{array}$$

By way of illustration, take the middle term in the middle row. This is the probability that two of the children have blood group AA, one is blood group AB and one is blood group BB. Thus $r_a=2$, $r_b=1$ and $r_c=1$. So:

$$\frac{4!}{r_a!\,r_b!\,r_c!}\,{p_a}^{r_a}{p_b}^{r_b}{p_c}^{r_c}=\frac{4!}{2!\,1!\,1!}\,{p_a}^{r_a}{p_b}^{r_b}{p_c}^{r_c}=12{p_a}^{r_a}{p_b}^{r_b}{p_c}^{r_c}$$

Note that the sum of the coefficients is 81, accounting for the 81 possibilities if order is important. [Equivalently, the sum of the coefficients in $q^4+4pq^3+6p^2q^2+4p^3q+p^4$ is 16, accounting for the 16 possibilities in the binomial case if order is important.]

The sum of the 81 probabilities turns out to be:

$$(p_a+p_b+p_c)^4=1 \qquad \text{given that } p_a+p_b+p_c=1$$

It is not difficult to expand this fourth power by hand and verify that the 15 terms which result correspond to those in the triangle.

It is an essential requirement of any distribution that the overall total probability is 1 and the Trinomial distribution satisfies this. A difference from the Uniform, Triangular and Binomial distributions is that the constituent probabilities of the Trinomial distribution are not indexed in a linear way.

There is nothing special about 4 as the number of children and the general expression for the Trinomial distribution is:

$$\frac{n!}{r_a!\,r_b!\,r_c!}\,{p_a}^{r_a}{p_b}^{r_b}{p_c}^{r_c} \quad \text{where } p_a+p_b+p_c=1 \text{ and } r_a+r_b+r_c=n$$

Given n children, this is the probability that r_a are of blood group AA, r_b are of blood group AB and r_c are of blood group BB.

The Multinomial Distribution

If entities have k states then the *Multinomial distribution* may apply. The expression that should be noted is:

$$\frac{n!}{r_1!\,r_2!\ldots r_k!}\,{p_1}^{r_1}{p_2}^{r_2}\ldots{p_k}^{r_k} \quad \text{where } p_1+p_2+\cdots+p_k=1 \text{ and } r_1+r_2+\cdots+r_k=n$$

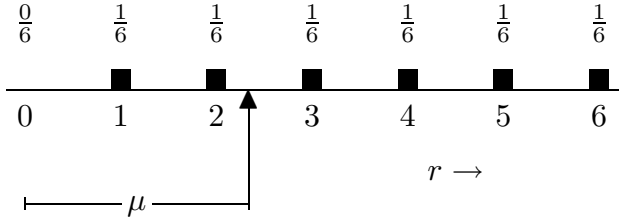
Given n entities, this is the probability that r_1 are in state 1, r_2 are in state 2 and so on up to r_k being in state k.

Expectation or Mean

If you repeatedly throw a fair die you would intuitively expect the long-term average of the values shown to be $3\frac{1}{2}$. On this occasion, intuition provides the right answer but a more formal approach is merited.

The terms *expectation* (usually denoted by the letter E) and *mean* (usually denoted by μ) are used to describe the long-term average. The mean may be calculated by thinking of weights and moments to determine a centre of gravity.

Including the contrived zero, the values which can result from throwing a die are 0, 1, 2, 3, 4, 5 and 6. Imagine marking these values off at unit intervals along a light beam and at each of the seven positions placing a weight whose mass is proportional to the associated probability:



The figure shows such an arrangement with little squares representing the weights. The leftmost weight has mass zero and so is not shown. A pivot has been placed at distance μ along the beam and it is at once clear that its position would not leave the beam in balance.

To achieve balance, consider the net clockwise moment about the pivot. The required value of μ has to be such that the net moment is zero. Accordingly, μ must satisfy:

$$(0-\mu).\frac{0}{6}+(1-\mu).\frac{1}{6}+(2-\mu).\frac{1}{6}+(3-\mu).\frac{1}{6}+(4-\mu).\frac{1}{6}+(5-\mu).\frac{1}{6}+(6-\mu).\frac{1}{6}=0$$

Consider the last term on the left, $(6-\mu).\frac{1}{6}$. The value $6-\mu$ is the distance of the rightmost weight from the pivot and this is multiplied by the mass of the weight (equal to the probability). The same consideration applies to the other terms but notice that if a weight is to the left of the pivot, the distance (as $2-\mu$ for example) is negative, correctly implying that the moment is anti-clockwise.

Rearrange the equation:

$$\mu\left(\frac{0}{6}+\frac{1}{6}+\frac{1}{6}+\frac{1}{6}+\frac{1}{6}+\frac{1}{6}+\frac{1}{6}\right)=0.\frac{0}{6}+1.\frac{1}{6}+2.\frac{1}{6}+3.\frac{1}{6}+4.\frac{1}{6}+5.\frac{1}{6}+6.\frac{1}{6}$$

The item in brackets is the sum of the probabilities and this, as always, is 1. Accordingly:

$$\mu = 0.\frac{0}{6} + 1.\frac{1}{6} + 2.\frac{1}{6} + 3.\frac{1}{6} + 4.\frac{1}{6} + 5.\frac{1}{6} + 6.\frac{1}{6} = \frac{1+2+3+4+5+6}{6} = \frac{21}{6} = \frac{7}{2}$$

The value $\frac{7}{2}$ or $3\frac{1}{2}$ comes as no surprise as the long-term average outcome of throwing an ordinary fair die.

The expression for μ can be rewritten:

$$\mu = 0.\mathrm{P}(X=0) + 1.\mathrm{P}(X=1) + 2.\mathrm{P}(X=2) + \cdots + 6.\mathrm{P}(X=6) \quad \text{or} \quad \mu = \sum_{r=0}^{6} r.\mathrm{P}(X=r)$$

The analysis applies to any distribution which is an indexed set of probabilities whose sum is 1. The general formula for the expectation or mean of a single random variable is written as:

$$\mathrm{E}(X) = \mu = \sum_{r} r.\mathrm{P}(X=r) \tag{3.1}$$

The item $\mathrm{E}(X)$ is pronounced 'the expectation of X'. The sum over r is left open-ended but this is taken to refer to the range which is appropriate.

Glossary

The following technical terms have been introduced:

distribution	Triangular distribution	Multinomial distribution
Uniform distribution	Binomial distribution	expectation
parameter	Trinomial distribution	mean

Exercises — III

Work in fractions whenever possible.

1. If the probability of hitting a target is $\frac{2}{5}$ and five shots are fired, what is the probability that the target will be hit at least twice? What is the conditional probability that the target will be hit at least twice, assuming that at least one hit is scored?
2. A supermarket has 20 check-outs: 5 have A-type cash registers and 15 have B-type. The A-type has a probability a of breaking down during the first hour of trading and the B-type has a probability b. The supervisor arrives at the end of this time and learns that one register has broken down. Determine the probability that the broken register is (a) A-type and (b) B-type.
3. 12 dice are thrown. What is the probability that each face appears twice?
4. Given Pascal's theorem $\left(\text{expressed as } \binom{n+1}{r+1} = \binom{n}{r} + \binom{n}{r+1}\right)$ prove that $\displaystyle\sum_{r=0}^{n} \binom{n}{r} = 2^n$
5. Prove the Binomial Theorem (page 2.12).
   Hint: it may be helpful to assume that the expansion holds for $(x+y)^n$ and to consider the effect of multiplication by one more $(x+y)$.

6. Using (3.1), determine the expectation of the Triangular distribution:

$$\mathrm{P}(X = r) = \begin{cases} \dfrac{r}{21}, & \text{if } r \in \mathbb{N} \wedge 1 \leqslant r \leqslant 6 \\ 0, & \text{otherwise} \end{cases}$$

7. [From Part Ia of the Mathematical Tripos, 1973] A princess is equally likely to sleep on anything from six to a dozen mattresses of the softest down, and beneath the lowest of these on just half the nights of the year is placed a pea. Being a young lady of refined sensibility her sleep is invariably disturbed by the presence of a pea beneath a mere six mattresses; with seven however a pea may pass unnoticed in one case out of ten, with eight it may escape detection in two cases out of ten, and so on, so that with the full twelve mattresses she slumbers on notwithstanding the offending pea as often as six times in ten. One morning, on being wakened by Bayes, her maid, she announces delightedly that she has spent the most tranquil of nights.

   What is the expected number of mattresses upon which she slept?

8. The answers to the following questions may be expressed as decimals:
   (*a*) What is the probability of obtaining at least one six when six dice are thrown?
   (*b*) What is the probability of obtaining at least two sixes when 12 dice are thrown?
   (*c*) What is the probability of obtaining at least three sixes when 18 dice are thrown?

9. A report of a possible breakthrough in the treatment of Nerd's Syndrome described a preliminary trial of a new drug. The drug was administered to 10 sufferers all of whom were immediately cured of the affliction. Tragically, later trials showed that 10% of patients treated with this drug die from an unfortunate side-effect.

   Suppose that n patients take part in a trial and that p is the probability that a trial participant suffers the fatal side-effect. Let S be the probability that at least one of the n patients dies. [Thus S is the probability that the trial reveals the side-effect.]

   Now consider the following (where, again, probabilities may be expressed as decimals):
   (*a*) In the preliminary trial, $p = \frac{1}{10}$ and $n = 10$. What is the probability that none of the 10 patients dies (as was the case in the reported trial)?
   (*b*) Again taking $p = \frac{1}{10}$, what is the minimum value of n needed to be 90% sure that the trial reveals the side-effect? [Thus what is the minimum value of n needed to ensure that $S \geqslant \frac{90}{100}$?]
   (*c*) The value 90% is sometimes called the *confidence*. With $p = \frac{1}{10}$, what is the minimum value of n needed to ensure a confidence of 99%? [That is $S \geqslant \frac{99}{100}$?]
   (*d*) For arbitrary (but known) p and arbitrary (but known) confidence C, what is the minimum value of n needed to ensure that $S \geqslant C$?
   (*e*) In real life there is a well-known heart drug for which the probability of suffering a serious side-effect is 10^{-5}. The risk of using the drug is deemed acceptable because there is a very much greater probability that an untreated patient will die. How large a trial would be needed to be 99% confident that the trial reveals the side-effect?