# MAS187/AEF258

University of Newcastle upon Tyne

2005-6

# Contents

# Chapter 7

# Discrete Probability Models

## 7.1 Introduction

The link between probability and statistics arises because, in order to see, for example, how strong the evidence is in some data, we need to consider the probabilities concerned with how we came to observe the data. In this chapter, we describe some standard probability models which are often used used with data from various sources such as market research. However, before we describe these in detail, we need to establish some ground rules for counting.

## 7.2 Permutations and Combinations

### 7.2.1 Numbers of sequences

Imagine that your cash point card has just been stolen. What is the probability of the thief guessing your 4 digit PIN in one go? To answer this question, we need to know how many different 4 digit PINs there are. We are also assuming that the thief chooses in such a way that all possibilities are equally likely. With this assumption the probability of a correct guess (in one go) is

$$\begin{aligned} P(\text{Guess correctly}) &= \frac{\text{number of correct PINs}}{\text{number of possible 4 digit PINs}} \\ &= \frac{1}{\text{number of possible 4 digit PINs}}. \end{aligned}$$

There is, of course, only one correct PIN. The number of possible 4 digit PINs is calculated as follows. There are 10 choices for the first digit, another 10 choices for the second digit, and so on. Therefore the number of possible choices is

$$10 \times 10 \times 10 \times 10 = 10,000.$$

So the probability of a correct guess is

$$\begin{aligned} P(\text{Guess correctly}) &= \tfrac{1}{10\times10\times10\times10} \\ &= \tfrac{1}{10{,}000} \\ &= 0.00001. \end{aligned}$$

### 7.2.2 Permutations

The calculation of the card-thief's correct guess of a PIN changes if the thief knows that your PIN uses 4 different digits. Now the number of possible PINs is smaller. To find this number we need to work out how many ways there are to arrange 4 digits out of a lsit of 10.

In more general terms, we need to know how many different ways there are of arranging $r$ objects from a list of $n$ objects. The best way of thinking about this to consider the choice of each item as a different experiment. The first experiment has $n$ possible outcomes. The second experiment only has $n-1$ possible outcomes, as one object has already been selected. The third experiment has $n-2$ outcomes and so on until the $r$th experiment, which has $n-r+1$ possible outcomes. Therefore the number of possible selections is

$$n\times(n-1)\times(n-2)\times\cdots\times(n-r+1) = \frac{n\times(n-1)\times(n-2)\times\cdots\times3\times2\times1}{(n-r)\times(n-r-1)\times\cdots\times3\times2\times1} = \frac{n!}{(n-r)!}.$$

Here

$$n! = n(n-1)(n-2)(n-3)\times\cdots\times3\times2\times1$$

and is called $n$ *factorial* - it can be found on many calculators. The formula

$$\frac{n!}{(n-r)!}$$

is a commonly encountered expression in counting calculations (combinatorics) and has its own notation. The number of ordered ways of selecting $r$ objects from $n$ is denoted ${}^n\mathrm{P}_r$, where

$${}^n\mathrm{P}_r = \frac{n!}{(n-r)!}.$$

We refer to ${}^n\mathrm{P}_r$ as the number of *permutations* of $r$ out of $n$ objects.

If we are interested solely in the number of ways of arranging $n$ objects, then this is clearly just

$${}^n\mathrm{P}_n = n!$$

Returning to the example in which the thief is trying to guess your 4-digit PIN, if the thief knows that the PIN contains no repreated digits then the number of possible PINS is

$${}^{10}\mathrm{P}_4 = 5040$$

so, assuming that each is equally likely to be guessed, the probability of a correct guess is

$$P(\text{Guess correctly}) = \frac{1}{5040} = 0.0001984.$$

This illustrates how important it is to keep secret all information about your PIN. These probability calculations show that even knowing whether or not you repeat digits in your PIN is informative for a thief – it reduces the number of possible PINs by a factor of around 2.

We now consider a more complicated problem. In a tutorial group there are 40 students. What is the probability that at least two students share a birthday?

First, let's make some simplifying assumptions. We will assume that there are 365 days in a year and that each day is equally likely to be a birthday.

Call the event we are interested in $Match$. We will first calculate the probability of $No\,Match$, the probability that *no two* people have the same birthday, and calculate the probability we want using $P(Match) = 1 - P(No\,Match)$. The number of ways 40 birthdays could occur is $365^{40}$ since there are 365 choices for the first person and 365 choices for the second person and so on. The number of ways we can have 40 *distinct* birthdays is the same as the number of permutations of 40 objects from 365 objects, that is, ${}^{365}\mathrm{P}_{40}$. So, the probability of all birthdays being distinct is

$$P(No\,Match) = \frac{{}^{365}\mathrm{P}_{40}}{365^{40}} = \frac{365!}{325!365^{40}} \simeq 0.1$$

and so

$$P(Match) = 1 - P(No\,Match) \simeq 0.9.$$

That is, there is a probability of 0.9 that we have a match. In fact, the fact that birthdays are *not* distributed uniformly over the year makes the probability of a match even higher! This is a somewhat counter-intuitive result, and the reason is that people think more intuitively about the probability that someone has the same birthday as *themselves*. This is an entirely different problem.

An alternative way of thinking about the probability of $No\,Match$ is to consider the sequences of choices as individuals are picked from the tutorial group:

$$\begin{aligned} P(No\,Match) &= P(\text{Pick person } 1) \times P(\text{Pick person } 2 \text{ with different birthday to person } 1) \\ &\qquad \times P(\text{Pick person } 3 \text{ with different birthday to persons } 1 \text{ and } 2) \times \ldots \\ &= \frac{365}{365} \times \frac{364}{365} \times \ldots \times \frac{326}{365} \\ &\simeq 0.1. \end{aligned}$$

### 7.2.3 Combinations

We now have a way of counting permutations, but often when selecting objects, all that matters is *which* objects were selected, not the order in which they were selected. Suppose that we have a collection of $n$ objects and that we wish to make $r$ selections from this list of objects, where the order does not matter. An unordered selection such as this is referred to as a *combination*. How many ways can this be done? Notice that this is equivalent to asking how many different ways are there of choosing $r$ objects from $n$ objects.

For example, a company has 20 retail outlets. It is decided to try a sales promotion at 4 of these outlets. How many selections of 4 can be chosen? It may be important to know this when we come to look at the results of the trial.

This calculation is very similar to that of permutations except that the ordering of objects no longer matters. For example, if we select two objects from three objects $A$, $B$ and $C$, there are ${}^{3}\mathrm{P}_2 = 6$ ways of doing this:

$$A,\ B \qquad A,\ C \qquad B,\ A \qquad B,\ C \qquad C,\ A \qquad C,\ B.$$

However, if we are not interested in the ordering, just in whether $A$, $B$ or $C$ are chosen then $A, B$ is the same as $B, A$ *etc.* and so the number of selections is just 3:

$$A,\ B \qquad A,\ C \qquad B,\ C.$$

The effect of ignoring the ordering reduces the number of permutations by a factor of ${}^{2}\mathrm{P}_2 = 2$. In general, the number of combinations of $r$ objects from $n$ objects is

$$\begin{aligned}\frac{\text{number of ordered samples of size } r}{\text{number of orderings of samples of size } r} &= \frac{{}^{n}\mathrm{P}_r}{{}^{r}\mathrm{P}_r}\\ &= \frac{{}^{n}\mathrm{P}_r}{r!}\\ &= \frac{n!}{r!(n-r)!}.\end{aligned}$$

Again, this is a very commonly found expression in combinatorics, so it has its own notation:

$${}^{n}\mathrm{C}_r = \frac{n!}{r!(n-r)!}.$$

There are other commonly used notations for this quantity: $\mathrm{C}_r^n$ and $\binom{n}{r}$. These numbers are known as the *binomial coefficients*.

Now we can see that the number of ways to select 4 retail outlets out of 20 is

$${}^{20}\mathrm{C}_4 = \frac{20!}{4!16!} = 4845.$$

An easy way to calculate binomial coefficients (at least small ones) is to use the fact that

$${}^{n}\mathrm{C}_r = \frac{n}{r} \times \frac{n-1}{r-1} \times \frac{n-2}{r-2} \times \cdots \times \frac{n-r+1}{1}.$$

For example,

$${}^{20}\mathrm{C}_4 = \frac{20}{4} \times \frac{19}{3} \times \frac{18}{2}\frac{17}{1}.$$

To see how combinations can be used to calculate probabilities, we will look at the UK National Lottery. In this lottery, there are 49 numbered balls, and six of these are selected at random. A seventh ball is also selected, but this is only relevant if you get exactly five numbers correct. The player selects six numbers before the draw is made, and after the draw, counts how many numbers are in common with those drawn. Players win a prize if they select at least three of the balls drawn. The order in which the balls are drawn in is irrelevant.

To begin with, let's calculate the probability that exactly 3 of the 6 numbers we select are drawn. First we need to count the number of possible draws (the number of different sets of 6 numbers), and then how many of those draws correspond to getting exactly three numbers correct. The number of possible draws is the number of ways of choosing 6 objects from 49. This is

$$^{49}\mathrm{C}_6 = 13,983,816.$$

The number of drawings corresponding to getting exactly three right is calculated as follows. Of the 49 balls from which the draw is made, 6 correspond to your selected numbers, and 43 correspond to other numbers. We want to know how many ways there are of choosing 3 of your selected numbers and 3 other numbers. This is the number of ways of choosing 3 from 6, multiplied by the number of ways of choosing 3 from 43. That is, there are

$$^{6}\mathrm{C}_3\ {}^{43}\mathrm{C}_3 = 246,820$$

ways of choosing exactly 3 of your selected numbers. So, the probability of matching exactly 3 numbers is

$$\frac{^{6}\mathrm{C}_3\ {}^{43}\mathrm{C}_3}{^{49}\mathrm{C}_6} = \frac{246,820}{13,983,816} \simeq 0.0177.$$

Similarly, we can calculate the probability of getting other prize-winning outcomes:

$$\begin{aligned}
P(\text{match exactly 6 correct numbers}) &= \frac{^{6}\mathrm{C}_6}{^{49}\mathrm{C}_6} = \frac{1}{13,983,816} \simeq 7 \times 10^{-8} \\
P(\text{match exactly 5 correct numbers plus bonus ball}) &= \frac{^{6}\mathrm{C}_5\ {}^{1}\mathrm{C}_1}{^{49}\mathrm{C}_6} = \frac{6}{13,983,816} \simeq 4 \times 10^{-7} \\
P(\text{match exactly 5 correct numbers}) &= \frac{^{6}\mathrm{C}_5\ {}^{43}\mathrm{C}_1}{^{49}\mathrm{C}_6} = \frac{258}{13,983,816} \simeq 2 \times 10^{-5} \\
P(\text{match exactly 4 correct numbers}) &= \frac{^{6}\mathrm{C}_4\ {}^{43}\mathrm{C}_2}{^{49}\mathrm{C}_6} = \frac{13545}{13,983,816} \simeq 1 \times 10^{-4}.
\end{aligned}$$

These outcomes are not very likely and so the prizes are chosen to reflect how likely you are to win. For example, in a recent lottery draw, the prizes were

| Number of balls matched | Prize |
|---|---|
| 6 | £2.4M |
| 5 plus bonus | £240K |
| 5 | £3K |
| 4 | £100 |
| 3 | £10 |
| $< 3$ | £0 |

This information allows us to calculate a fair price for such a bet. The expected monetary value of the bet is

$$
\begin{aligned}
\mathrm{EMV} = {} & P(\text{match 6 balls}) \times \mathrm{Prize}(\text{match 6 balls}) \\
& + P(\text{match 5 balls plus bonus}) \times \mathrm{Prize}(\text{match 5 balls plus bonus}) \\
& + \ldots \\
& + P(\text{match 3 balls}) \times \mathrm{Prize}(\text{match 3 balls}) \\
= {} & 2.4M \times \frac{1}{13,983,816} + 240K \times \frac{6}{13,983,816} + \ldots + 10 \times \frac{246,820}{13,983,816} \\
= {} & 0.6176.
\end{aligned}
$$

Therefore, a fair price for a ticket in this particular lottery is around $62p$. This difference between this and the standard £1 charge for a ticket goes to “good causes” and, of course, Camelot’s profits.

## 7.3 Probability Distributions

### 7.3.1 Introduction

In Chapter 1 we saw how surveys can be used to get information on population quantities. For example, we might want to know voting intentions within the UK just before a General Election. Why does this involve *random* variables? In most cases, it is not possible to measure the variables on every member of the population and so some sampling scheme is used. This means that there is uncertainty in our conclusions. For example, if the true proportion of Labour voters were 40%, in a survey of 1,000 voters, it would be possible to get 400 Labour voters, but it would also be possible to get 350 Labour voters or 430 Labour voters. The fact that we have only a sample of voters introduces uncertainty into our conclusions about voting intentions in the population as a whole. Sometimes experiments themselves have inherent variability, for example, the toss of a coin. If the coin were tossed 1000 times and heads occurred only 400 times, would it be fair to conclude that the coin was a biased coin? The subject of Statistics has been developed to understand such variability and, in particular, how to draw correct conclusions from data which are subject to experimental and sampling variability.

Before we can make inferences about populations, we need a language to describe the uncertainty we find when taking samples from populations. First, we represent a random variable by $X$ (capital X) and the probability that it takes a certain value $x$ (small x) as $P(X = x)$.

The *probability distribution* of a discrete random variable $X$ is the list of all possible values $X$ can take and the probabilities associated with them. For example, if the random variable $X$ is the outcome of a roll of a die then the probability distribution for $X$ is

| $x$ | $P(X=x)$ |
|---|---|
| 1 | $1/6$ |
| 2 | $1/6$ |
| 3 | $1/6$ |
| 4 | $1/6$ |
| 5 | $1/6$ |
| 6 | $1/6$ |
| sum | 1 |

Just as with sample data, it is useful to have some summary information about probability distributions. For example, what is the average value of the random variable? How much variation is there in this distribution?

### 7.3.2 Expectation and the population mean

The mean of a quantitative random variable is a weighted sum of its possible values, where each weight is the probability of the value occurring. This is known as the expected value of the random variable or the population mean of the random variable and is usually written as $\mathrm{E}(X)$ or $\mu$. Therefore, for a discrete random variable,

$$\mathrm{E}(X) = \mu = \sum x\, P(X = x).$$

Previously we have seen a similar calculation when determining the expected monetary value

$$EMV = \sum P(\text{Event}) \times \text{Monetary value of Event}.$$

The expected value is the average value which we would get in an infinitely long sequence of identical experiments.

For example, suppose that the population of interest is this class and that it contains $N$ students. Suppose that we are interested in the number of times that students have bought a particular product (e.g. a cinema ticket) in the last month. Clearly the population mean is just the average of this variable in the class:

$$\mu = \frac{1}{n} \sum_{i=1}^{n} x_i$$

where $x_i$ is the number of times student $i$ has bought the product. We can also write this as

$$\mu = \frac{1}{n} \sum_{j=0}^{\infty} j f_j = \sum_{j=0}^{\infty} j \frac{f_j}{n}$$

where $f_j$ is the frequency of $x = j$ in the population and $f_j/n$ is the relative frequency. If we choose a student at random from the class then the probability that we choose a student with $x = j$ is

$$P(X = j) = \frac{f_j}{n}$$

the relative frequency and so

$$\mu = \sum_{j=0}^{\infty} jP(X = j).$$

It is also clear that this is the average which we would get if we kept on sampling, with replacement, for a very long time.

For the die-rolling experiment, the average number of spots we would get if we repeated the experiment an "infinite" number of times is

$$E(X) = \sum x\,P(X = x) = 1 \times \frac{1}{6} + 2 \times \frac{1}{6} + \ldots + 6 \times \frac{1}{6} = 3.5.$$

This concept can be generalised to calculate the expected value of any function of $X$. For instance, in the lottery example discussed previously, the prize was determined by the number of matches. In the die-rolling experiment, we could consider a prize worth the square of the number showing: £1 for a 1, £4 for a 2, £9 for a 3, and so on. In this case the expected prize money is

$$\begin{aligned} E\left(X^2\right) &= \sum x^2\,P(X = x) \\ &= 1 \times \frac{1}{6} + 4 \times \frac{1}{6} + \ldots + 36 \times \frac{1}{6} \\ &= \frac{91}{6} \simeq \pounds 15.17. \end{aligned}$$

### 7.3.3 Population variance and standard deviation

In addition to having the population mean as a measure of location, it is also useful to know about the spread of the random variable about this value. The variance of a random variable is denoted $\text{Var}(X)$ or sometimes $\sigma^2$ and is determined by

$$\text{Var}(X) = \sigma^2 = E\left[(X - \mu)^2\right].$$

It is simply the average squared deviation from the mean. Note that this is the same sort of calculation as with sample variances. The larger the value for the variance, the larger the spread.

Referring again back to the die-rolling experiment, if $X$ is the number of spots, we can calculate the variance (using $\mu = 3.5$):

| $x$ | $P(X = x)$ | $(x - \mu)^2$ | $(x - \mu)^2 P(X = x)$ |
|---|---|---|---|
| 1 | 1/6 | 6.25 | 1.0417 |
| 2 | 1/6 | 2.25 | 0.3750 |
| 3 | 1/6 | 0.25 | 0.0417 |
| 4 | 1/6 | 0.25 | 0.0417 |
| 5 | 1/6 | 2.25 | 0.3750 |
| 6 | 1/6 | 6.25 | 1.0417 |
| sum | 1 | | 2.9167 |

Hence

$$\text{Var}(X) = 2.9167.$$

As with sample variances, there is an alternative way of calculating population variances, using

$$\mathrm{Var}(X) = \mathrm{E}(X^2) - \mu^2.$$

Using this formula with the above example gives

| $x$ | $P(X = x)$ | $x^2$ | $x^2P(X = x)$ |
|---|---|---|---|
| 1 | 1/6 | 1 | 0.1667 |
| 2 | 1/6 | 4 | 0.6667 |
| 3 | 1/6 | 9 | 1.5000 |
| 4 | 1/6 | 16 | 2.6667 |
| 5 | 1/6 | 25 | 4.1667 |
| 6 | 1/6 | 36 | 6.0000 |
| sum | 1 | | 15.1667 |

and so

$$\mathrm{Var}(X) = \sum x^2P(X = x) - \mu^2 = 15.1667 - 3.5^2 = 2.9167.$$

The standard deviation of a random variable is

$$\mathrm{SD}(X) = \sqrt{\mathrm{Var}(X)}.$$

In this example, $\mathrm{SD}(X) = \sqrt{2.9167} = 1.7078$.

## 7.4 Exercises 7

1. Consider a lottery that is slightly different to the National Lottery in that there are 48 balls instead of 49. What is the probability of winning the jackpot in this lottery? (That is, you choose six balls and exactly these six are drawn).

2. A market survey has identified 10 desirable features for a new product. However, due to cost constraints, only four of these features can be included. If the features are selected randomly, what is the probability that your four favourites are chosen in your preferred ordering?

3. If you dial 7 digits at random on a (non-mobile) telephone in Newcastle, what is the probability you dial Dr. Farrow's office number (which has 7 digits)?

4. A sample of four mass-produced items is examined for quality control purposes. Each item can be either satisfactory (S) or unsatisfactory (U). Each item has a probability of 0.2 of being unsatisfactory and each item is independent of every other item

   (a) Consider the sequence of 4 items. In how many different sequences can we get

      i. no unsatisfactory items?
      ii. exactly 1 unsatisfactory item?
      iii. exactly 2 unsatisfactory items?
      iv. exactly 3 unsatisfactory items?
      v. four unsatisfactory items.

   (b) Find the probability of a particular sequence containing

      i. no unsatisfactory items.
      ii. exactly 1 unsatisfactory item.
      iii. exactly 2 unsatisfactory items.
      iv. exactly 3 unsatisfactory items.
      v. four unsatisfactory items.

   (c) Hence find the probability that we get

      i. no unsatisfactory items.
      ii. exactly 1 unsatisfactory item.
      iii. exactly 2 unsatisfactory items.
      iv. exactly 3 unsatisfactory items.
      v. four unsatisfactory items.

   (d) Find the mean number of unsatisfactory items.

   (e) Find the variance and standard deviation of the number of unsatisfactory items.