Where we are in the course

Weeks 1–4 (Data collection and summaries)

- How to *collect* data
- How to *summarise* data
  - Tabular
  - Graphical
  - Numerical (location and spread)

Where we are in the course

Weeks 5–7 (Probability)

- Introduction to probability
  - Interpretations of probability
  - Laws of probability
- Conditional probability and probability trees
- *EMV* and decision trees

Where we are in the course

Weeks 8–11 (Probability models)

- Models for discrete data
  - The Binomial distribution
  - The Poisson distribution
- Models for continuous data
  - The Normal distribution
  - The Uniform distribution
  - The exponential distribution

Don't forget – Assignment 1 will be available to download this week!

Lecture 8

DISCRETE PROBABILITY MODELS

Introduction

In this lecture we begin the final part of the course for Semester 1: **Probability models**.

Semester 2 will be devoted to the study of **Statistics**. The link between Probability and Statistics arises because in order to see, for example, how strong the evidence is in some data, we often need to consider the probabilities concerned with how we came to observe this data.

In this lecture, we describe some standard **probability models** which are often 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**".

Permutations and Combinations

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

$$P(\textbf{Guess correctly}) = \frac{\text{number of correct PINs}}{\text{number of possible 4 digit PINs}}$$

$$= \frac{1}{\text{number of possible 4 digit PINs}}$$

Obviously there is only **one** correct PIN.

Suppose your PIN consisted of only 2 digits. How many ways could they be arranged?

	0	1	2	3	4	5	6	7	8	9
0	0,0	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9
1	1,0	1,1	1,2	1,3	1,4	1,5	1,6	1,7	1,8	1,9
2	2,0	2,1	2,2	2,3	2,4	2,5	2,6	2,7	2,8	2,9
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	
9	9,0	9,1	9,2	9,3	9,4	9,5	9,6	9,7	9,8	9,9

So

$$\text{No. of possible } \mathbf{2} \text{ digit PINS} \quad = \quad 10 \times 10 = 100$$

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

$$\begin{aligned} P(\textbf{Guess correctly}) &= \frac{1}{10 \times 10 \times 10 \times 10} \\ &= \frac{1}{10,000} \\ &= 0.0001. \end{aligned}$$

Permutations

What if the thief knew that your PIN used **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 list of 10**.

More generally, 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
- The **rth** experiment has $n-(r-1)=n-r+1$ possible outcomes

Therefore the number of possible selections is

$$n \times (n-1) \times (n-2) \times \cdots \times (n-r+1)$$

$$= \frac{n \times (n-1) \times \cdots \times (n-r+1) \times (n-r) \times \cdots \times 3 \times 2 \times 1}{(n-r) \times (n-r-1) \times \cdots \times 3 \times 2 \times 1}$$

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

Here

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

and is called n **factorial**.

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.

This is often seen as $n\mathrm{P}r$ on calculators.

Back to the credit card thief

Thus, if the thief knows that the PIN contains no repeated digits then the number of possible PINs is

$$^{10}\mathrm{P}_4 = 5040 \ (=10 \times 9 \times 8 \times 7)$$

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

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

This illustrates how important it is to keep secret all information about your PIN!!

Combinations

We now have a way of counting permutations.

However, sometimes all that matters is **which** objects were selected, not **the order** in which they were selected.

Suppose 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**.

Example

A company has **20** retail outlets. The company decides to try a sales promotion at **4** of these outlets.

How many selections of **4** can be chosen?

This calculation is very similar to that for 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:

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

In general, the number of **combinations** of r objects from n objects is

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

Again, this is a very commonly found expression in combinatorics, so it has its own notation (usually the $n\mathrm{C}r$ button on a calculator).

We sometimes read this as "n choose r" or "Choose r objects from n".

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

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

The National Lottery

- There are 49 numbered balls
- 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
- 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**.

Let's consider the probability of winning the jackpot.

How many ways can 6 balls be chosen out of 49?

One option is $\{1, 2, 3, 4, 5, 6\}$; another is $\{1, 2, 3, 4, 5, 7\}$ ...

... in fact, there are

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

different ways 6 balls can be selected out of a possible 49.

Now out of these 13,983,816 different combinations, how many combinations match the drawn balls correctly?

Only one! There is only one set of six numbers that wins the jackpot! So the probability of winning the jackpot is just one in 13,983,816, i.e.

$$P(\text{match exactly 6 correct numbers}) = \frac{1}{13,983,816},$$

or just over a **one in fourteen million chance**!

The other probabilities used last in last week's lecture to calculate the Expected Monetary Value for the lottery can be found using similar arguments.

Probability distributions

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

In the die–rolling example, we used the interpretation of probability to obtain the probability distribution for X, the outcome of a roll on the die.

In the die–rolling example, we used the **classical** interpretation of probability to obtain the probability distribution for X, the outcome of a roll on the die.

Consider the following **frequentist** example.

Let X be the number of cars observed in half–hour periods passing the junction of two roads. In a five hour period, the following observations on X were made:

$$2 \quad 3 \quad 2 \quad 5 \quad 5 \quad 3 \quad 4 \quad 5 \quad 6 \quad 7$$

Obtain the probability distribution of X.

2 3 2 5 5 3 4 5 6 7

We can calculate the following probabilities:

$$P(X = 0) = \frac{0}{10} = 0$$

$$P(X = 1) = \frac{0}{10} = 0$$

$$P(X = 2) = \frac{2}{10} = 0.2$$

$$P(X = 3) = \frac{2}{10} = 0.2$$

2 3 2 5 5 3 4 5 6 7

$$P(X = 4) = \frac{1}{10} = 0.1$$

$$P(X = 5) = \frac{3}{10} = 0.3$$

$$P(X = 6) = \frac{1}{10} = 0.1$$

$$P(X = 7) = \frac{1}{10} = 0.1$$

$$P(X > 7) = \frac{0}{7} = 0$$

Thus would give:

x	$P(X = x)$
< 2	**0**
2	**0.2**
3	**0.2**
4	**0.1**
5	**0.3**
6	**0.1**
7	**0.1**
> 7	**0**
sum	**1**

Does this make sense?

The Binomial Distribution

In many surveys and experiments data is collected in the form of **counts**. For example,

- the number of people in a survey who bought a CD
- the number of people who said they would vote Labour
- the number of defective items in a sample

All these variables have common features:

1. Each person/item has only **two** possible (exclusive) responses (Yes/No, Defective/Not defective etc)
   – this is referred to as a **trial** which results in a **success** or **failure**
2. The survey/experiment takes the form of a **random sample**
   – the responses are independent

Further, suppose that the true probability of a success in the population is p, and we are interested in the random variable X, the total number of successes out of n trials.

Example

Suppose we are interested in the number of sixes we get from 4 rolls of a dice.

Each roll of the dice is a trial which gives a "six" (success, or s) or "not a six" (failure, or f).

The probability of a success is $p = P(\text{six}) = 1/6$.

We have $n = 4$ independent trials (rolls of the dice).

Let X be the number of sixes obtained. We can now obtain the full probability distribution of X.

Example

For example, suppose we want to work out the probability of obtaining four sixes (four "successes" – i.e. *ssss* – or $P(X = 4)$).

Since the rolls of the die can be considered independent, we get:

$$\begin{aligned} P(ssss) &= P(s) \times P(s) \times P(s) \times P(s) \\ &= \frac{1}{6} \times \frac{1}{6} \times \frac{1}{6} \times \frac{1}{6} \\ &= \left(\frac{1}{6}\right)^4 \end{aligned}$$

Example

That one's easy... what about the probability that we get three sixes – i.e. $P(X = 3)$?

This one's a bit more tricky, because that means we need three s's and one f – i.e. three sixes and one "not six" – but the "not six" could appear on the first roll, or the second roll, or the third, or the fourth!

For example, for $P(X = 3)$, we *could* have:

$$\begin{aligned} P(fsss) &= \frac{5}{6} \times \frac{1}{6} \times \frac{1}{6} \times \frac{1}{6} \\ &= \left(\frac{1}{6}\right)^3 \times \frac{5}{6}. \end{aligned}$$

Example

Or we could have:

$$\begin{aligned} P(sfss) &= \frac{1}{6} \times \frac{5}{6} \times \frac{1}{6} \times \frac{1}{6} \\ &= \left(\frac{1}{6}\right)^3 \times \frac{5}{6} \end{aligned}$$

or *maybe*:

$$\begin{aligned} P(ssfs) &= \frac{1}{6} \times \frac{1}{6} \times \frac{5}{6} \times \frac{1}{6} \\ &= \left(\frac{1}{6}\right)^3 \times \frac{5}{6} \end{aligned}$$

Example

or *even*:

$$\begin{aligned} P(sssf) &= \frac{1}{6} \times \frac{1}{6} \times \frac{1}{6} \times \frac{5}{6} \\ &= \left(\frac{1}{6}\right)^3 \times \frac{5}{6}. \end{aligned}$$

Can you see that we therefore get:

$$P(X = 3) = 4 \times \left(\frac{1}{6}\right)^3 \times \frac{5}{6}.$$

Example

Thinking about it, there are actually **sixteen** possible outcomes for the four rolls of the die:

	Outcome	Probability
1	$ssss$	$\left(\frac{1}{6}\right)^4$
2	$fsss$	$\left(\frac{1}{6}\right)^3 \times \frac{5}{6}$
3	$sfss$	$\left(\frac{1}{6}\right)^3 \times \frac{5}{6}$
4	$ssfs$	$\left(\frac{1}{6}\right)^3 \times \frac{5}{6}$
5	$sssf$	$\left(\frac{1}{6}\right)^3 \times \frac{5}{6}$
6	$ssff$	$\left(\frac{1}{6}\right)^2 \times \left(\frac{5}{6}\right)^2$
7	$ffss$	$\left(\frac{1}{6}\right)^2 \times \left(\frac{5}{6}\right)^2$
8	$sfsf$	$\left(\frac{1}{6}\right)^2 \times \left(\frac{5}{6}\right)^2$
9	$fsfs$	$\left(\frac{1}{6}\right)^2 \times \left(\frac{5}{6}\right)^2$
10	$sffs$	$\left(\frac{1}{6}\right)^2 \times \left(\frac{5}{6}\right)^2$
11	$fssf$	$\left(\frac{1}{6}\right)^2 \times \left(\frac{5}{6}\right)^2$
12	$sfff$	$\frac{1}{6} \times \left(\frac{5}{6}\right)^3$
13	$fsff$	$\frac{1}{6} \times \left(\frac{5}{6}\right)^3$
14	$ffsf$	$\frac{1}{6} \times \left(\frac{5}{6}\right)^3$
15	$fffs$	$\frac{1}{6} \times \left(\frac{5}{6}\right)^3$
16	$ffff$	$\left(\frac{5}{6}\right)^4$

Example

So we get:

$$
\begin{aligned}
P(X=4) &= \left(\frac{1}{6}\right)^4 = 0.0008 \\
P(X=3) &= 4 \times \left(\frac{1}{6}\right)^3 \times \frac{5}{6} = 0.0153 \\
P(X=2) &= 6 \times \left(\frac{1}{6}\right)^2 \times \left(\frac{5}{6}\right)^2 = 0.1158 \\
P(X=1) &= 4 \times \frac{1}{6} \times \left(\frac{5}{6}\right)^3 = 0.3858 \qquad \text{and} \\
P(X=0) &= \left(\frac{5}{6}\right)^4 = 0.4823
\end{aligned}
$$

Example

So the full **probability distribution** for X is:

x	0	1	2	3	4
$P(X = x)$	0.4823	0.3858	0.1158	0.0153	0.0008

Now that was a bit long–winded... and that was just for four rolls of the die!

We would like a more concise way of working these probabilities out without having to list all the possible outcomes as we did above.

Luckily, there's a formula that does just that!

Example

If we have count data where each (independent) trial results in one of two possible outcomes ("success" or "failure"), then

$$P(X = r) = {}^{n}\mathrm{C}_{r}\, p^{r}(1-p)^{n-r}, \quad r = 0, 1, \ldots, n,$$

where p is the probability of "success" and n is the number of trials.

These probabilities describe how likely we are to get r out of n successes from independent trials, each with success probability p.

This distribution is known as the **binomial distribution** with index n and probability p.

We write this as $X \sim Bin(n, p)$. Here, n and p are known as the **parameters** of the Binomial distribution.

Example

In the die example, we know $n = 4$ and $p = P(\text{six}) = 1/6$.

Each roll of the dice is a trial which gives a "six" (success) or "not a six" (failure).

If X is the number of sixes obtained then

$$\begin{aligned} X &\sim Bin(n, p) \qquad \text{i.e.} \\ X &\sim Bin(4, 1/6), \end{aligned}$$

and so...

Example

$$
\begin{aligned}
P(X=0) &= {}^{4}\mathrm{C}_{0}\left(\frac{1}{6}\right)^{0}\left(1-\frac{1}{6}\right)^{4} \\
&= 1 \times 1 \times 0.4823 \\
&= 0.4823
\end{aligned}
$$

$$
\begin{aligned}
P(X=1) &= {}^{4}\mathrm{C}_{1}\left(\frac{1}{6}\right)^{1}\left(1-\frac{1}{6}\right)^{3} \\
&= 4 \times 0.1667 \times 0.5787 \\
&= 0.3858
\end{aligned}
$$

Example

$$
\begin{aligned}
P(X=2) &= {}^{4}\mathrm{C}_{2}\left(\frac{1}{6}\right)^{2}\left(1-\frac{1}{6}\right)^{2} \\
&= 6 \times 0.0278 \times 0.6944 \\
&= 0.1158
\end{aligned}
$$

$$
\begin{aligned}
P(X=3) &= {}^{4}\mathrm{C}_{3}\left(\frac{1}{6}\right)^{3}\left(1-\frac{1}{6}\right)^{1} \\
&= 4 \times 0.0046 \times 0.8333 \\
&= 0.0153
\end{aligned}
$$

Example

And finally,

$$\begin{aligned} P(X=4) &= {}^{4}\mathrm{C}_{4}\left(\frac{1}{6}\right)^{4}\left(1-\frac{1}{6}\right)^{0} \\ &= 1 \times 0.0008 \times 1 \\ &= 0.0008 \end{aligned}$$

This probability distribution shows that most of the time we would get either 0 or 1 successes and, for example, 4 successes would be quite rare.

Example

Let's see how close these "theoretical" probabilities are to some "observed" values obtained by actually rolling a dice four times and counting the number of sixes we get. *Actually, we'll not roll a dice, but will use Minitab instead!*

No. of sixes	Binomial probability	Observed probability
0	0.4823	
1	0.3858	
2	0.1158	
3	0.0153	
4	0.0008	
Sum		

Another example

A salesperson has a 50% chance of making a sale on a customer visit and she arranges 6 visits in a day.

What are the probabilities of her making 0,1,2,3,4,5 and 6 sales?

Let X denote the **number of sales**. Assuming the visits result in sales independently, $X \sim Bin(6, 0.5)$ and

No. of sales	**Probability**	**Cumulative Probability**
r	$P(X = r)$	$P(X \leq r)$
0	0.015625	0.015625
1	0.093750	0.109375
2	0.234375	0.343750
3	0.312500	0.656250
4	0.234375	0.890625
5	0.093750	0.984375
6	0.015625	1.000000
sum	1.000000	

The formula for binomial probabilities enables us to calculate values for $P(X = r)$. From these, it is straightforward to calculate cumulative probabilities such as the probability of making no more than 2 sales:

$$\begin{aligned} P(X \leq 2) &= P(X = 0) + P(X = 1) + P(X = 2) \\ &= 0.015625 + 0.09375 + 0.234375 = 0.34375. \end{aligned}$$

These cumulative probabilities are also useful in calculating probabilities such as that of making more than 1 sale:

$$P(X > 1) = 1 - P(X \leq 1) = 1 - 0.109375 = 0.890625.$$

If X is a random variable with a binomial $Bin(n, p)$ distribution then its **mean** and **variance** are

$$E(X) = n \times p, \qquad Var(X) = n \times p \times (1 - p).$$

For example, if $X \sim Bin(6, 0.5)$ then

$$E(X) = n \times p = 6 \times 0.5 = 3$$

and

$$Var(X) = n \times p \times (1 - p) = 3 \times 0.5 \times 0.5 = 1.5$$

Also

$$SD(X) = \sqrt{Var(X)} = \sqrt{1.5} = 1.225.$$

The Poisson Distribution

The **Poisson distribution** is another very important discrete probability distribution.

1. It is often used to model **count data**
2. Unlike the binomial distribution, there is **no known fixed upper limit** of counts
3. The **rate** of occurrence, λ, is the parameter here

If these conditions are reasonable, then we say

$$X \sim Po(\lambda)$$

If X is a random variable with a Poisson distribution with parameter λ, then the probability it takes different values is

$$P(X = r) = \frac{\lambda^r e^{-\lambda}}{r!}, \quad r = 0, 1, 2, \ldots \ .$$

The parameter λ has a very simple interpretation as the rate at which events occur. The distribution has mean and variance

$$E(X) = \lambda, \qquad Var(X) = \lambda.$$

Example

Consider

X: **number of calls made in a 1 minute interval to an ISP**

If the ISP knows that on average 5 calls will be made in this 1 minute interval, then

$$X \sim Po(5)$$

Then for example,

$$P(X = 4) = \frac{5^4 e^{-5}}{4!} = 0.1755.$$

We can use the formula for Poisson probabilities to calculate the probability of all possible outcomes:

	Probability	**Cumulative Probability**
r	$P(X = r)$	$P(X \leq r)$
0	0.0067	0.0067
1	0.0337	0.0404
2	0.0843	0.1247
3	0.1403	0.2650
4	0.1755	0.4405
5	0.1755	0.6160
6	0.1462	0.7622
7	0.1044	0.8666
8	0.0653	0.9319
⋮	⋮	⋮
sum	1.000000	

Therefore the probability of receiving between 2 and 8 calls is

$$P(2 \leq X \leq 8) = P(X \leq 8) - P(X \leq 1) = 0.9319 - 0.0404 = 0.8915$$

Using such a model we can also account for "extreme" situations.

For example, suppose that, for this ISP, we observed the following number of calls per minute over a five minute period:

$$6,\ 3,\ 5,\ 4,\ 6.$$

Using simple frequentist reasoning, we would have

$$P(7 \text{ calls made}) \quad = \quad \frac{0}{5} = 0,$$

i.e. we will *never* observe seven calls in any one minute period!

However, using the Poisson model, we have

$$P(X = 7) \quad = \quad 0.1044,$$

which is obviously more realistic.