Probability and Probability Models

Dr. Lee Fawcett

Newcastle University

Introduction

What is Probability?

We're all familiar with the concept of different events being **likely** or **unlikely**.

What is Probability?

We're all familiar with the concept of different events being **likely** or **unlikely**.

- It's currently raining in England

What is Probability?

We're all familiar with the concept of different events being **likely** or **unlikely**.

- It's currently raining in England $\implies$ LIKELY

What is Probability?

We're all familiar with the concept of different events being **likely** or **unlikely**.

- It's currently raining in England $\Longrightarrow$ LIKELY
- I will win Euromillions this weekend

What is Probability?

We're all familiar with the concept of different events being **likely** or **unlikely**.

- It's currently raining in England $\implies$ LIKELY
- I will win Euromillions this weekend $\implies$ UNLIKELY

What is Probability?

We're all familiar with the concept of different events being **likely** or **unlikely**.

- It's currently raining in England $\implies$ LIKELY
- I will win Euromillions this weekend $\implies$ UNLIKELY
- Tomorrow is Friday

What is Probability?

We're all familiar with the concept of different events being **likely** or **unlikely**.

- It's currently raining in England $\Longrightarrow$ LIKELY
- I will win Euromillions this weekend $\Longrightarrow$ UNLIKELY
- Tomorrow is Friday $\Longrightarrow$ CERTAIN

What is Probability?

We're all familiar with the concept of different events being **likely** or **unlikely**.

- It's currently raining in England $\Longrightarrow$ LIKELY
- I will win Euromillions this weekend $\Longrightarrow$ UNLIKELY
- Tomorrow is Friday $\Longrightarrow$ CERTAIN

We're probably also used to quantifying likelihood using percentages:

What is Probability?

We're all familiar with the concept of different events being **likely** or **unlikely**.

- It's currently raining in England $\Longrightarrow$ LIKELY
- I will win Euromillions this weekend $\Longrightarrow$ UNLIKELY
- Tomorrow is Friday $\Longrightarrow$ CERTAIN

We're probably also used to quantifying likelihood using percentages:

What is Probability?

We're all familiar with the concept of different events being **likely** or **unlikely**.

- It's currently raining in England $\Longrightarrow$ LIKELY
- I will win Euromillions this weekend $\Longrightarrow$ UNLIKELY
- Tomorrow is Friday $\Longrightarrow$ CERTAIN

We're probably also used to quantifying likelihood using percentages:

- "England have a 20% chance of winning the World Cup"

What is Probability?

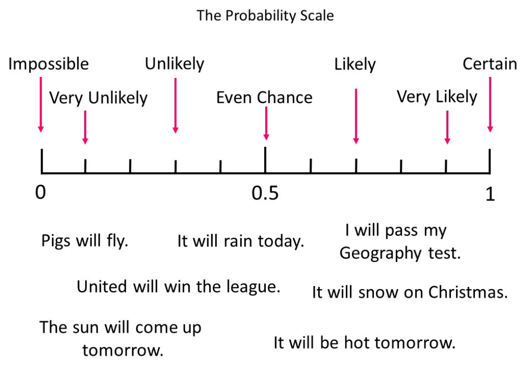
We're all familiar with the concept of different events being **likely** or **unlikely**.

- It's currently raining in England $\Longrightarrow$ LIKELY
- I will win Euromillions this weekend $\Longrightarrow$ UNLIKELY
- Tomorrow is Friday $\Longrightarrow$ CERTAIN

We're probably also used to quantifying likelihood using percentages:

- "England have a 20% chance of winning the World Cup"
- "I'm 99% sure I turned the oven off before I left for the airport..."

What is Probability?



The Probability Scale

Pigs will fly.

It will rain today.

I will pass my Geography test.

United will win the league.

It will snow on Christmas.

The sun will come up tomorrow.

It will be hot tomorrow.

Defining Probability

We can set-up a mathematically rigorous framework to define probability; we can then use this to calculate the probability of whether something we're interested in will happen.

First we need a few definitions:

- **Experiment**: Something that generates data
- **Outcome**: A possible result from the experiment
- **Event**: A collection of one or more outcomes

Classical Probability

Probability is normally introduced through scenarios involving playing cards, dice, coloured counters etc (where the probability of different outcomes are known, and are often equal).

This is known as **classical probability**.

Classical Probability

Probability is normally introduced through scenarios involving playing cards, dice, coloured counters etc (where the probability of different outcomes are known, and are often equal).

This is known as **classical probability**.

We can calculate probabilities in classical scenarios exactly:

$$P(A) = \frac{\text{The number of outcomes that satisfy } A}{\text{The number of possible outcomes}}$$

where $P(A)$ means “the probability that event A occurs”.

Classical Probability

Probability is normally introduced through scenarios involving playing cards, dice, coloured counters etc (where the probability of different outcomes are known, and are often equal).

This is known as **classical probability**.

We can calculate probabilities in classical scenarios exactly:

$$P(A) = \frac{\text{The number of outcomes that satisfy } A}{\text{The number of possible outcomes}}$$

where $P(A)$ means “the probability that event A occurs”.

For example:

$$P(\text{A playing card is a 9}) = \frac{4}{52} = \frac{1}{13} = 7.69\%$$

Estimating Probabilities

There are several ways of estimating the probabilities of real world events.

Here we will consider the **frequency** definition of probability.

Estimating Probabilities

There are several ways of estimating the probabilities of real world events.

Here we will consider the **frequency** definition of probability.

> *"If we repeated the experiment lots and lots of times, what proportion of those times did Event A happen?"*

Estimating Probabilities

There are several ways of estimating the probabilities of real world events.

Here we will consider the **frequency** definition of probability.

> *"If we repeated the experiment lots and lots of times, what proportion of those times did Event A happen?"*

$$P(\text{Event } A) = \frac{\text{Number of times event } A \text{ occurs}}{\text{Number of times the experiment was run}}$$

Random Variables

Random variables are outcomes of interest that will vary randomly between runs of an experiment. For example:

- The temperature outside this building each day
- The number of unique visitors to a website every hour
- A company's closing stock price etc.

Random Variables

Random variables are outcomes of interest that will vary randomly between runs of an experiment. For example:

- The temperature outside this building each day
- The number of unique visitors to a website every hour
- A company's closing stock price etc.

We are typically interested in the probability a random variable takes a particular value (or range of values).

Random Variables

Suppose we are interested in the probability of a fair, six-sided die taking different values.

We can easily write down the **probability mass function** (PMF) for X : Outcome of roll of die:

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

Random Variables

Suppose we are interested in the probability of a fair, six-sided die taking different values.

We can easily write down the **probability mass function** (PMF) for X : Outcome of roll of die:

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

```
> N = 100000
> game = sample(1:6, N, replace=TRUE)
> table(game)
game
    1     2     3     4     5     6
16473 16935 16677 16570 16504 16841
> table(game)/N
game
      1       2       3       4       5       6
0.16473 0.16935 0.16677 0.16570 0.16504 0.16841
```

Random Variables

x	1	2	3	4	5	6
(Probability Masss Function) $P(X = x)$	1/6	1/6	1/6	1/6	1/6	1/6
(Cumulative Distribution Function) $P(X \leq x)$	1/6	1/3	1/2	2/3	5/6	1

- From the table, we see that:
  - $P(X = 1) = P(X = 2) = \cdots = P(X = 6) = 1/6$

Random Variables

x	1	2	3	4	5	6
(Probability Masss Function) $P(X = x)$	1/6	1/6	1/6	1/6	1/6	1/6
(Cumulative Distribution Function) $P(X \leq x)$	1/6	1/3	1/2	2/3	5/6	1

- From the table, we see that:
  - $P(X = 1) = P(X = 2) = \cdots = P(X = 6) = 1/6$
  - $P(X = 2$ or $X = 5) = (1/6) + (1/6) = 2/6 = 1/3$

Newcastle University

Random Variables

x	1	2	3	4	5	6
(Probability Masss Function) $P(X = x)$	1/6	1/6	1/6	1/6	1/6	1/6
(Cumulative Distribution Function) $P(X \leq x)$	1/6	1/3	1/2	2/3	5/6	1

- From the table, we see that:
  - $P(X = 1) = P(X = 2) = \cdots = P(X = 6) = 1/6$
  - $P(X = 2$ or $X = 5) = (1/6) + (1/6) = 2/6 = 1/3$
  - $P(X \leq 4) = 4/6 = 2/3$

Random Variables

x	1	2	3	4	5	6
(Probability Masss Function) $P(X = x)$	1/6	1/6	1/6	1/6	1/6	1/6
(Cumulative Distribution Function) $P(X \leq x)$	1/6	1/3	1/2	2/3	5/6	1

- From the table, we see that:
  - $P(X = 1) = P(X = 2) = \cdots = P(X = 6) = 1/6$
  - $P(X = 2$ or $X = 5) = (1/6) + (1/6) = 2/6 = 1/3$
  - $P(X \leq 4) = 4/6 = 2/3$
- Suppose we roll the die twice:
  - $P(\text{Two sixes}) = (1/6) \times (1/6) = 1/36$

Random Variables

x	1	2	3	4	5	6
(Probability Masss Function) $P(X = x)$	1/6	1/6	1/6	1/6	1/6	1/6
(Cumulative Distribution Function) $P(X \leq x)$	1/6	1/3	1/2	2/3	5/6	1

- From the table, we see that:
  - $P(X = 1) = P(X = 2) = \cdots = P(X = 6) = 1/6$
  - $P(X = 2 \text{ or } X = 5) = (1/6) + (1/6) = 2/6 = 1/3$
  - $P(X \leq 4) = 4/6 = 2/3$
- Suppose we roll the die twice:
  - $P(\text{Two sixes}) = (1/6) \times (1/6) = 1/36$
- Also:
  - **Valid PMF**: All probabilities sum to 1

Newcastle University

Random Variables

x	1	2	3	4	5	6
(Probability Masss Function) $P(X = x)$	1/6	1/6	1/6	1/6	1/6	1/6
(Cumulative Distribution Function) $P(X \leq x)$	1/6	1/3	1/2	2/3	5/6	1

- From the table, we see that:
  - $P(X = 1) = P(X = 2) = \cdots = P(X = 6) = 1/6$
  - $P(X = 2$ or $X = 5) = (1/6) + (1/6) = 2/6 = 1/3$
  - $P(X \leq 4) = 4/6 = 2/3$
- Suppose we roll the die twice:
  - $P(\text{Two sixes}) = (1/6) \times (1/6) = 1/36$
- Also:
  - **Valid PMF**: All probabilities sum to 1
  - **Complement**: $P(X = 2) + P(X \neq 2) = 1/6 + 5/6 = 1$

Random Variables

What about the probability mass functions for more ‘interesting’ random variables?

Random Variables

What about the probability mass functions for more ‘interesting’ random variables?

Two of my personal favourites:

- X: Number of tropical storms in any given year for a location in the Gulf of Mexico

Random Variables

What about the probability mass functions for more ‘interesting’ random variables?

Two of my personal favourites:

- X: Number of tropical storms in any given year for a location in the Gulf of Mexico
- X: Number of rats developing renal tumours after expore to *Ochratoxin A*

Introduction

We will now break for a problem session covering this section.

Please pause the video here and go to **Problem Session 1** of **Lesson 4**.

See you in a bit!

Discrete Probability Models

Introduction

The table below shows the number of years, in the past 100 years, in which X tropical storms were observed for a location in the Gulf of Mexico.

x	0	1	2	3	4	5	6	7+
# Years	14	27	27	18	9	4	1	0

All things being equal, can we use this information to assess the likelihood of tropical storms at this location in future years?

Introduction

x	0	1	2	3	4	5	6	7+
# Years	14	27	27	18	9	4	1	0

We can do simple things using the probability we studied in the first part of this lecture. For example:

$$P(X = 3) = \frac{18}{100} = 0.18;$$

just under a one-in-five chance.

Introduction

x	0	1	2	3	4	5	6	7+
# Years	14	27	27	18	9	4	1	0

We can do simple things using the probability we studied in the first part of this lecture. For example:

$$P(X = 3) = \frac{18}{100} = 0.18;$$

just under a one-in-five chance. Similarly:

$$P(X = 6) = \frac{1}{100} = 0.01;$$

i.e. a once-in-a-century event. What about **compound events**?

Introduction

x	0	1	2	3	4	5	6	7+
# Years	14	27	27	18	9	4	1	0

We can do simple things using the probability we studied in the first part of this lecture. For example:

$$P(X = 3) = \frac{18}{100} = 0.18;$$

just under a one-in-five chance. Similarly:

$$P(X = 6) = \frac{1}{100} = 0.01;$$

i.e. a once-in-a-century event. What about **compound events**?

$$P(X \leq 3) = \frac{14 + 27 + 27 + 18}{100} = \frac{86}{100} = 0.86.$$

Introduction

x	0	1	2	3	4	5	6	7+
# Years	14	27	27	18	9	4	1	0

Using these data how likely is it that, next year, there will be more than 6 tropical storms at this location?

Introduction

x	0	1	2	3	4	5	6	7+
# Years	14	27	27	18	9	4	1	0

Using these data how likely is it that, next year, there will be more than 6 tropical storms at this location?

This simple approach to probability would give us

$$P(X > 6) = \frac{0}{100} = 0,$$

i.e. *impossible*! Does this make sense?

Introduction

We can use the information given to estimate our PMF and CDF for X:

x	0	1	2	3	4	5	6	7+
$\#$ Years	14	27	27	18	9	4	1	0

Introduction

We can use the information given to estimate our PMF and CDF for X:

x	0	1	2	3	4	5	6	7+
$\#$ Years	14	27	27	18	9	4	1	0
$P(X = x)$	0.14	0.27	0.27	0.18	0.09	0.04	0.01	0.00

Introduction

We can use the information given to estimate our PMF and CDF for X:

x	0	1	2	3	4	5	6	7+
$\#$ Years	14	27	27	18	9	4	1	0
$P(X = x)$	0.14	0.27	0.27	0.18	0.09	0.04	0.01	0.00
$P(X \leq x)$	0.14	0.41	0.68	0.86	0.95	0.99	1.00	1.00

Introduction

We can use the information given to estimate our PMF and CDF for X:

x	0	1	2	3	4	5	6	7+
$\#$ Years	14	27	27	18	9	4	1	0
$P(X = x)$	0.14	0.27	0.27	0.18	0.09	0.04	0.01	0.00
$P(X \leq x)$	0.14	0.41	0.68	0.86	0.95	0.99	1.00	1.00

- These probabilities are **frequency estimates** of the true probabilities

Introduction

We can use the information given to estimate our PMF and CDF for X:

x	0	1	2	3	4	5	6	7+
# Years	14	27	27	18	9	4	1	0
$P(X = x)$	0.14	0.27	0.27	0.18	0.09	0.04	0.01	0.00
$P(X \leq x)$	0.14	0.41	0.68	0.86	0.95	0.99	1.00	1.00

- These probabilities are **frequency estimates** of the true probabilities
- What are the true probabilities? Difficult to know, as we can't assume equally likely outcomes

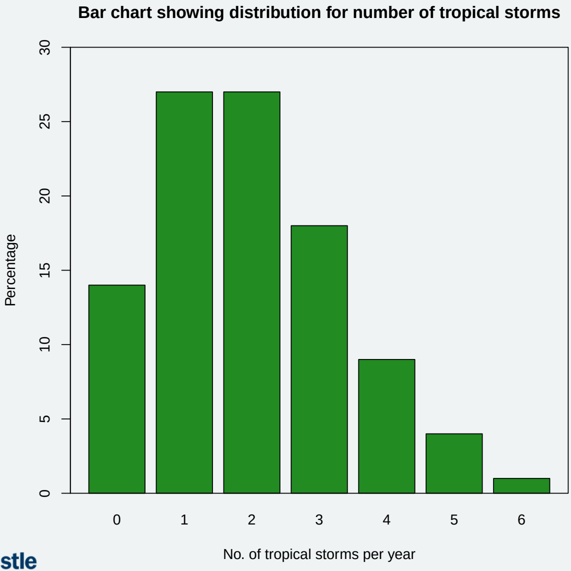
Introduction

We can use the information given to estimate our PMF and CDF for X:

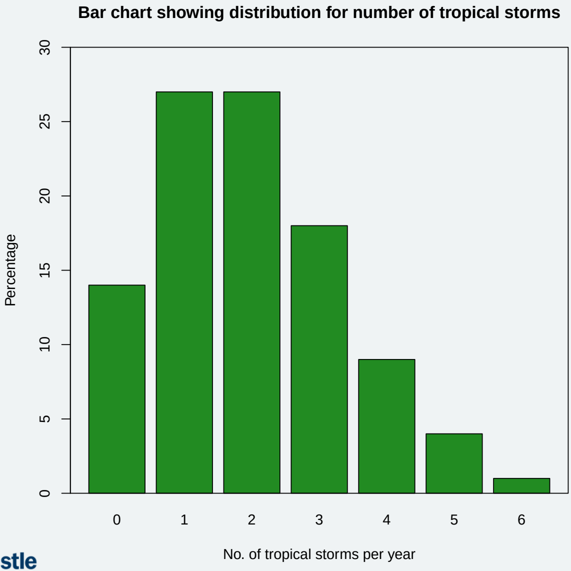
x	0	1	2	3	4	5	6	7+
# Years	14	27	27	18	9	4	1	0
$P(X = x)$	0.14	0.27	0.27	0.18	0.09	0.04	0.01	0.00
$P(X \leq x)$	0.14	0.41	0.68	0.86	0.95	0.99	1.00	1.00

- These probabilities are **frequency estimates** of the true probabilities
- What are the true probabilities? Difficult to know, as we can't assume equally likely outcomes
- We might consider the PMF/ as an approximation to the 'true' underlying probability distribution, governed by a **probability model**

Introduction



Introduction



The Poisson distribution

Suppose we are interested in X, the number of events occurring in a given time interval.

Suppose further that the following statements hold:

- Events occur **independently**, at some **constant rate** (let's call this λ)

The Poisson distribution

Suppose we are interested in X, the number of events occurring in a given time interval.

Suppose further that the following statements hold:

- Events occur **independently**, at some **constant rate** (let's call this λ)
- There is no **natural/known upper limit** to the number of events that might occur in any given time interval

The Poisson distribution

What does all of this mean? Think about our tropical storms example.

The Poisson distribution

What does all of this mean? Think about our tropical storms example.

Here, we are interested in X, the number of tropical storms occurring in any given year.

The Poisson distribution

What does all of this mean? Think about our tropical storms example.

Here, we are interested in X, the number of tropical storms occurring in any given year.

- **Independence**: The number of storms occurring in one year does not influence what happens in the next year

The Poisson distribution

What does all of this mean? Think about our tropical storms example.

Here, we are interested in X, the number of tropical storms occurring in any given year.

- **Independence**: The number of storms occurring in one year does not influence what happens in the next year
- **constant rate**: We are assuming there is some underlying average ‘annual storm rate’ that our data are hovering around

The Poisson distribution

What does all of this mean? Think about our tropical storms example.

Here, we are interested in X, the number of tropical storms occurring in any given year.

- **Independence**: The number of storms occurring in one year does not influence what happens in the next year
- **constant rate**: We are assuming there is some underlying average ‘annual storm rate’ that our data are hovering around
- **Known upper limit?** Can we say what the maximum number of tropical storms will be in any given year?

The Poisson distribution

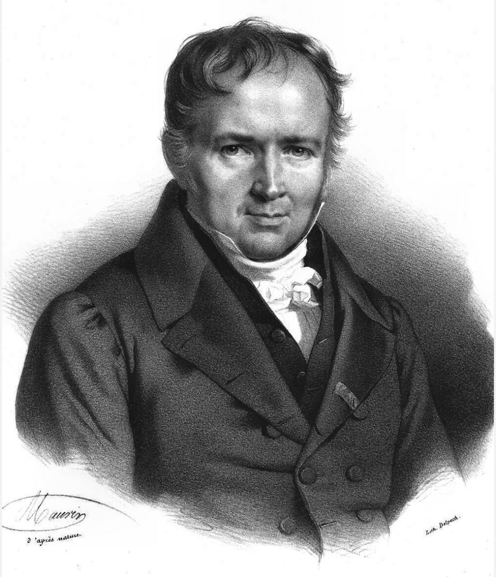
What does all of this mean? Think about our tropical storms example.

Here, we are interested in X, the number of tropical storms occurring in any given year.

- **Independence**: The number of storms occurring in one year does not influence what happens in the next year
- **constant rate**: We are assuming there is some underlying average ‘annual storm rate’ that our data are hovering around
- **Known upper limit?** Can we say what the maximum number of tropical storms will be in any given year?

Assuming ‘Yes’, ‘Yes’ and ‘No’ gives rise to the **Poisson distribution** as a suitable probability model for X.

The Poisson distribution


Maurin
d'après nature.
Lith. Delpech.

The Poisson distribution: Notation and formulae

We say $X \sim Po(\lambda)$, with

$$P(X = x) = \frac{e^{-\lambda} \cdot \lambda^x}{x!},$$

where $\lambda > 0$, $x = 0, 1, \ldots$ and $x! = x \cdot (x-1) \cdot (x-2) \ldots$.

The Poisson distribution: Notation and formulae

We say $X \sim Po(\lambda)$, with

$$P(X = x) = \frac{e^{-\lambda} \cdot \lambda^x}{x!},$$

where $\lambda > 0$, $x = 0, 1, \ldots$ and $x! = x \cdot (x-1) \cdot (x-2) \ldots$.

We also have

- Exepcted value $= E(X) = \lambda$
- Variability $= Var(X) = \lambda$

The Poisson distribution: Notation and formulae

For example, in our tropical storms dataset we have

$$\lambda = E(X) \approx \bar{x} = \frac{1 \cdot 27 + 2 \cdot 27 + \cdots + 6 \cdot 1}{100} = 1.97;$$

in other words, we assume a constant rate of around 1.97 tropical storms per year. Then

The Poisson distribution: Notation and formulae

For example, in our tropical storms dataset we have

$$\lambda = E(X) \approx \bar{x} = \frac{1 \cdot 27 + 2 \cdot 27 + \cdots + 6 \cdot 1}{100} = 1.97;$$

in other words, we assume a constant rate of around 1.97 tropical storms per year. Then

$$P(X = 3) = \frac{e^{-1.97} 1.97^3}{3!} = 0.178$$ *(Compared to 0.18 earlier)*

The Poisson distribution: Notation and formulae

For example, in our tropical storms dataset we have

$$\lambda = E(X) \approx \bar{x} = \frac{1 \cdot 27 + 2 \cdot 27 + \cdots + 6 \cdot 1}{100} = 1.97;$$

in other words, we assume a constant rate of around 1.97 tropical storms per year. Then

$$P(X = 3) = \frac{e^{-1.97} 1.97^3}{3!} = 0.178 \quad \textit{(Compared to 0.18 earlier)}$$

Similarly,

$$P(X = 6) = \frac{e^{-1.97} 1.97^6}{6!} = 0.011 \quad \textit{(Compared to 0.01 earlier)}$$

The Poisson distribution: Notation and formulae

For the full probability distribution, we have:

x	0	1	2	3	4	5	6	7+
# Years	14	27	27	18	9	4	1	0
Empirical	0.14	0.27	0.27	0.18	0.09	0.04	0.01	0.00

The Poisson distribution: Notation and formulae

For the full probability distribution, we have:

x	0	1	2	3	4	5	6	7+
# Years	14	27	27	18	9	4	1	0
Empirical	0.14	0.27	0.27	0.18	0.09	0.04	0.01	0.00
Poisson	0.139	0.275	0.271	0.178	0.088	0.034	0.011	0.004

The Poisson distribution: Notation and formulae

For the full probability distribution, we have:

x	0	1	2	3	4	5	6	7+
# Years	14	27	27	18	9	4	1	0
Empirical	0.14	0.27	0.27	0.18	0.09	0.04	0.01	0.00
Poisson	0.139	0.275	0.271	0.178	0.088	0.034	0.011	0.004

Notice:

- How close the Poisson probabilities are to the **empirical** probabilities

The Poisson distribution: Notation and formulae

For the full probability distribution, we have:

x	0	1	2	3	4	5	6	7+
# Years	14	27	27	18	9	4	1	0
Empirical	0.14	0.27	0.27	0.18	0.09	0.04	0.01	0.00
Poisson	0.139	0.275	0.271	0.178	0.088	0.034	0.011	0.004

Notice:

- How close the Poisson probabilities are to the **empirical** probabilities
- That although $P(X > 6)$ is very small, at least we don't rule this out now!

The Binomial distribution

In a study to investigate mammalian toxicity to *Ochratoxin A* (OTA), five male rats were exposed to OTA. Subsequently, the number of rats developing renal tumours was recorded.

Previous studies (e.g. Herman, D. and Mantle, P., 2021) estimate that 35% of male rates exposed to OTA will go on to develop renal tumours.

The Binomial distribution

If we are interested in X: The number of rats that will develop tumours, we might assume a **Binomial distribution**:

- We have a fixed number of ‘trials’ (call this n, with $n = 5$ in this example), meaning there is a fixed upper limit to the value X can take

The Binomial distribution

If we are interested in X: The number of rats that will develop tumours, we might assume a **Binomial distribution**:

- We have a fixed number of 'trials' (call this n, with $n = 5$ in this example), meaning there is a fixed upper limit to the value X can take
- Each trial has two possible outcomes: 'success' or 'failure'

The Binomial distribution

If we are interested in X: The number of rats that will develop tumours, we might assume a **Binomial distribution**:

- We have a fixed number of ‘trials’ (call this n, with $n = 5$ in this example), meaning there is a fixed upper limit to the value X can take
- Each trial has two possible outcomes: ‘success’ or ‘failure’

Further, if we assume:

- There is a constant probability of success (call this p, with $p = 0.35$ in this case)

The Binomial distribution

If we are interested in X: The number of rats that will develop tumours, we might assume a **Binomial distribution**:

- We have a fixed number of ‘trials’ (call this n, with $n = 5$ in this example), meaning there is a fixed upper limit to the value X can take
- Each trial has two possible outcomes: ‘success’ or ‘failure’

Further, if we assume:

- There is a constant probability of success (call this p, with $p = 0.35$ in this case)
- The outcome of each trial is independent of every other trial

The Binomial distribution

If we are interested in X: The number of rats that will develop tumours, we might assume a **Binomial distribution**:

- We have a fixed number of 'trials' (call this n, with $n = 5$ in this example), meaning there is a fixed upper limit to the value X can take
- Each trial has two possible outcomes: 'success' or 'failure'

Further, if we assume:

- There is a constant probability of success (call this p, with $p = 0.35$ in this case)
- The outcome of each trial is independent of every other trial

then we might say $X \sim Bin(n = 5, p = 0.35)$.

The Binomial distribution: Notation and formulae

We say $X \sim Bin(n, p)$, with

$$P(X = x) = \left(\begin{array}{c} n \\ x \end{array} \right) p^x (1 - p)^{n-x},$$

where $n = 1, 2, 3, ...,$ $0 \leq p \leq 1$ and $x = 0, 1, \ldots$.

The Binomial distribution: Notation and formulae

We say $X \sim Bin(n, p)$, with

$$P(X = x) = \left(\begin{array}{c} n \\ x \end{array} \right) p^x (1 - p)^{n-x},$$

where $n = 1, 2, 3, ...,$ $0 \leq p \leq 1$ and $x = 0, 1, \ldots$.

We also have

- $E(X) = np$
- $Var(X) = np(1 - p)$

The Binomial distribution: Notation and formulae

We say $X \sim Bin(n, p)$, with

$$P(X = x) = \left(\begin{array}{c} n \\ x \end{array} \right) p^x (1 - p)^{n-x},$$

where $n = 1, 2, 3, ...,\ 0 \leq p \leq 1$ and $x = 0, 1, \ldots$.

We also have

- $E(X) = np$
- $Var(X) = np(1 - p)$

This gives the following PMF for $X \sim Bin(5, 0.35)$:

x	0	1	2	3	4	5
$P(X = x)$	0.116	0.312	0.336	0.181	0.049	0.005

Discrete Probability Models

We will now break for a problem session covering this section.

Please pause the video here and go to **Problem Session 2** of **Lesson 4**.

See you in a bit!

The Normal distribution

Studying Normal populations

It is often the case that collected data have a distribution with the characteristic shape of the **Normal distribution**.

Studying Normal populations

It is often the case that collected data have a distribution with the characteristic shape of the **Normal distribution**.

To explain this phenomenon, let's consider an example.

Haematocrit (Packed Cell Volume)

The **haematocrit** (Ht or HCT) or **packed cell volume** (PCV) is a measurement of the proportion of blood volume that is occupied by red blood cells.

It is normally 45 ± 7 $(38\% - 52\%)$ for males and 42 ± 5 $(37\% - 47\%)$ for females.

Haematocrit (Packed Cell Volume)

In a recent investigation, a haematocrit measurement was taken from 126 female students at Newcastle University:

42.0	42.0	40.0	45.0	42.0	43.0	36.0	43.0	40.0	42.0	35.0
46.0	45.0	38.0	40.0	46.0	45.0	43.0	42.0	41.0	44.0	41.0
49.0	42.0	35.0	46.0	44.0	37.0	43.0	42.0	41.0	36.0	41.0
44.0	38.0	40.5	42.0	38.0	35.0	42.0	42.0	40.0	45.0	40.0
55.0	45.0	44.0	44.0	43.0	42.0	42.0	43.0	42.0	45.0	43.0
44.0	41.0	46.0	49.0	44.0	44.0	32.5	42.0	44.0	49.0	41.5
48.0	42.0	39.0	45.0	41.0	42.0	43.0	46.0	40.0	43.0	42.0
44.0	40.0	44.0	41.0	44.0	43.0	49.0	40.5	39.0	48.0	40.0
41.0	41.0	41.0	45.0	36.0	39.0	36.0	38.0	45.0	46.0	41.0
40.0	43.0	34.5	42.0	42.0	39.0	41.5	46.0	42.0	44.0	46.0
44.0	40.0	38.5	40.0	40.0	44.0	39.0	40.0	39.0	43.0	42.0
36.0	46.0	44.5	48.0	45.5						

Haematocrit (Packed Cell Volume)

We can use techniques from Lecture 1 to summarise these data:

Haematocrit (Packed Cell Volume)

We can use techniques from Lecture 1 to summarise these data:

- We could summarise the measurements **numerically**

Haematocrit (Packed Cell Volume)

We can use techniques from Lecture 1 to summarise these data:

- We could summarise the measurements **numerically**
    - the **mean** haematocrit measurement is 42.15

Haematocrit (Packed Cell Volume)

We can use techniques from Lecture 1 to summarise these data:

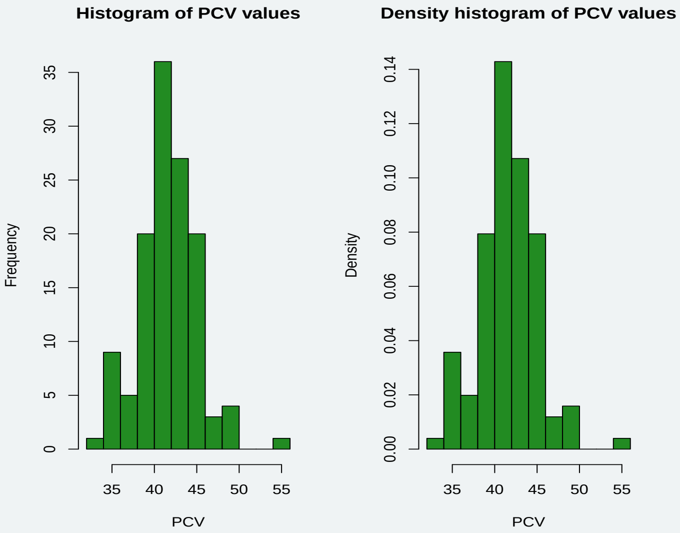
- We could summarise the measurements **numerically**
  - the **mean** haematocrit measurement is 42.15
  - the **standard deviation** is 3.44

Haematocrit (Packed Cell Volume)

We can use techniques from Lecture 1 to summarise these data:

- We could summarise the measurements **numerically**
  - the **mean** haematocrit measurement is 42.15
  - the **standard deviation** is 3.44
- We could summarise the measurements **graphically** – perhaps using a **histogram**

Haematocrit (Packed Cell Volume)



Haematocrit (Packed Cell Volume)

These data show the classic shape of the **Normal distribution**.

Haematocrit (Packed Cell Volume)

These data show the classic shape of the **Normal distribution**.

It is characterised by the symmetrical **bell-shaped curve**:

- values near the mean are more common
- values further away *tail off* in terms of their frequencies

Haematocrit (Packed Cell Volume)

These data show the classic shape of the **Normal distribution**.

It is characterised by the symmetrical **bell-shaped curve**:

- values near the mean are more common
- values further away *tail off* in terms of their frequencies

A ‘perfect’ Normal distribution curve is given by the function

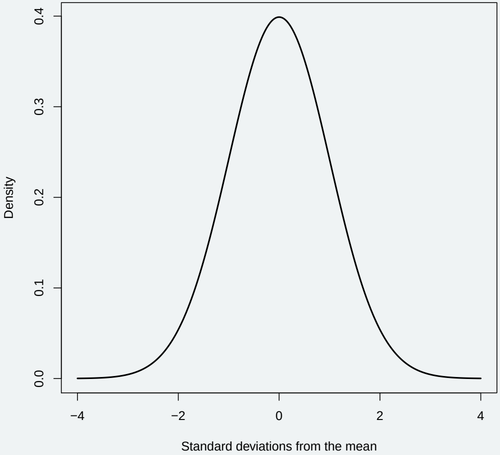
$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2 \right\}.$$

This is known as the **probability density function** (PDF) of the Normal distribution, where μ and σ represent the mean and standard deviation (respectively).

The Normal (Gaussian) distribution

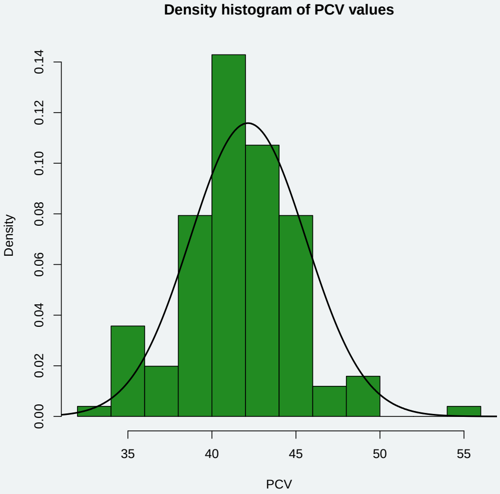
Haematocrit (Packed Cell Volume)

The plot below shows the PDF of the ‘perfect’ Normal distribution:



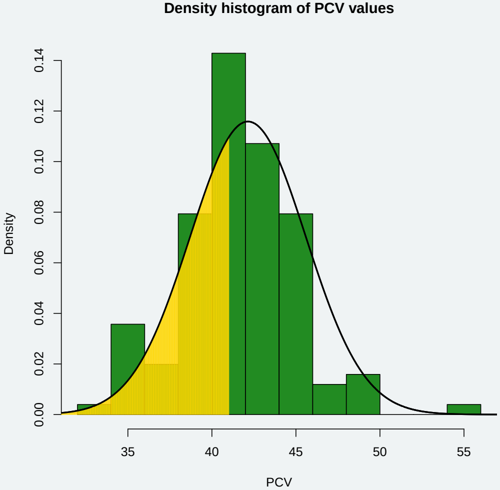
Haematocrit (Packed Cell Volume)

The Normal distribution curve for our haematocrit data, using $\mu = 42.15$ and $\sigma = 3.44$, looks like:



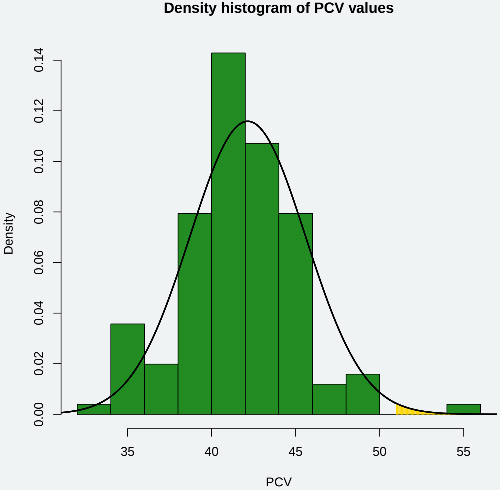
Haematocrit (Packed Cell Volume)

Areas under the curve represent probabilities; for example, for $P(X < 41)$:

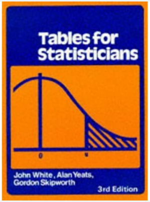


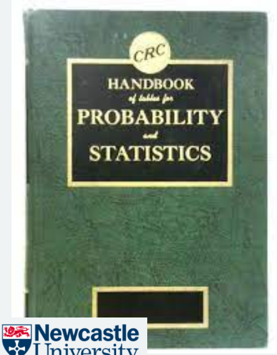
Haematocrit (Packed Cell Volume)

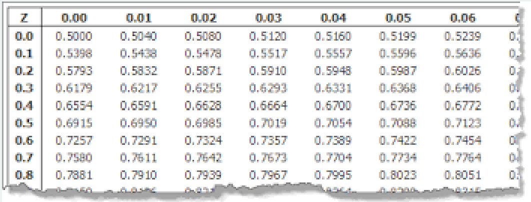
Areas under the curve represent probabilities; for example, for $P(X > 51)$:



Haematocrit (Packed Cell Volume)







Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051

Normal Q-Q plots

In the Haematocrit example, the sample size was sufficiently large to compare a histogram with the fitted PDF of a Normal distribution.

Normal Q-Q plots

In the Haematocrit example, the sample size was sufficiently large to compare a histogram with the fitted PDF of a Normal distribution.

With smaller samples, assessing Normality in such a way can often be challenging.

Normal Q-Q plots

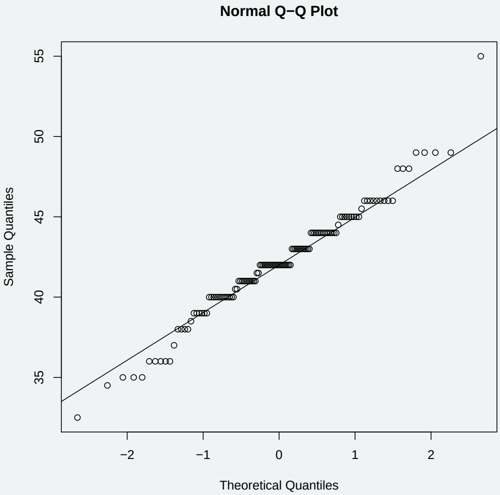
In the Haematocrit example, the sample size was sufficiently large to compare a histogram with the fitted PDF of a Normal distribution.

With smaller samples, assessing Normality in such a way can often be challenging.

Normal quantile plots (often referred to as ‘Q-Q plots’) are more often used to assess Normality.

Such plots provide the main way of checking the assumption of Normality underpinning much of what we will cover in tomorrow’s sessions.

Normal Q-Q plots



The Normal distribution

We will now break for a problem session covering this section.

Please pause the video here and go to **Problem Session 3** of **Lesson 4**.

See you in a bit!

Conclusions

1. **Basic ideas of probability** – probability scale, language/notation, classical and frequency calculations
2. **Probability models**
    - For discrete random variables: The Poisson and binomial distributions
    - For Continuous random variables: The Normal distribution
3. Checking **Normality**