L3 Data Essentials

- Module 4
- Statistics for Decision-Making
- 3-day Live event learning

Overview of 3-day learning

DAY 3

Module 4, revisit online content, knowledge check, task activities, and review.

Statistics and Data Modelling

SCHEDULE: DAY 3

AM

- Probability and experiments with dice
- Random variables and distributions
- Activity: Introduction to Normal Distribution

PM

- Hypotheses testing
- Activity: A/B testing
- Final Module Activity 4.2

Day 3 learning

PROBABILITY

Activity

Come up with the best definition you can of **probability**.

Extra challenge

Produce a definition that is not circular. That is, your definition of probability must not itself contain the word ‘probability’ or any close synonym such as ‘chance’, ‘likelihood’, etc. This is difficult!

PROBABILITY

Probability is given on a scale of 0 to 1.

- 0 : impossible
- 0.5 : 50% chance
- 1 : certain

They can be worked out in two ways:

- The number of ways an event could happen out of the size of the sample space (Bayesian Probability).

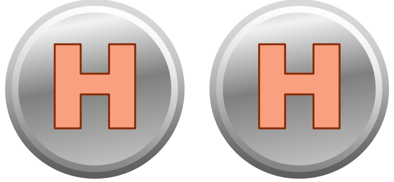
OR

- The number of times an event happens divided by the number of observations (Frequentist Probability).

COIN FLIPS

Question 1

You flip two coins. What is the probability of obtaining HH?

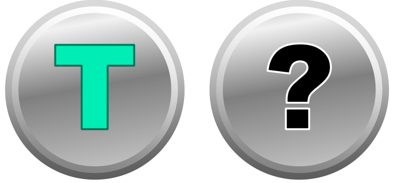


A: 3/4 B: 1/2

C: 1/4 D: 1/8

Question 2

You flip a coin. It lands on tails. You flip one more coin. What is the probability that the second coin lands on tails?



A: 3/4 B: 1/2

C: 1/4 D: 1/8

COIN FLIPS

Question 3

You flip the same coin twice. What is the probability that it lands first on heads, second on tails?

H T

A: 3/4   B: 1/2
C: 1/4   D: 1/8

Question 4

You flip the same coin twice. What is the probability that it lands exactly once on heads and once on tails?

H × 1

T × 1

A: 3/4   B: 1/2
C: 1/4   D: 1/8

RANDOM VARIABLES

A random variable can be informally defined as a variable whose value we cannot perfectly predict.

Examples of random variables:

- The outcome of a coin toss
- The age of a person randomly selected from the population
- A person’s reaction time
- The number of decays of a sample of radioactive material in one minute

DISTRIBUTION

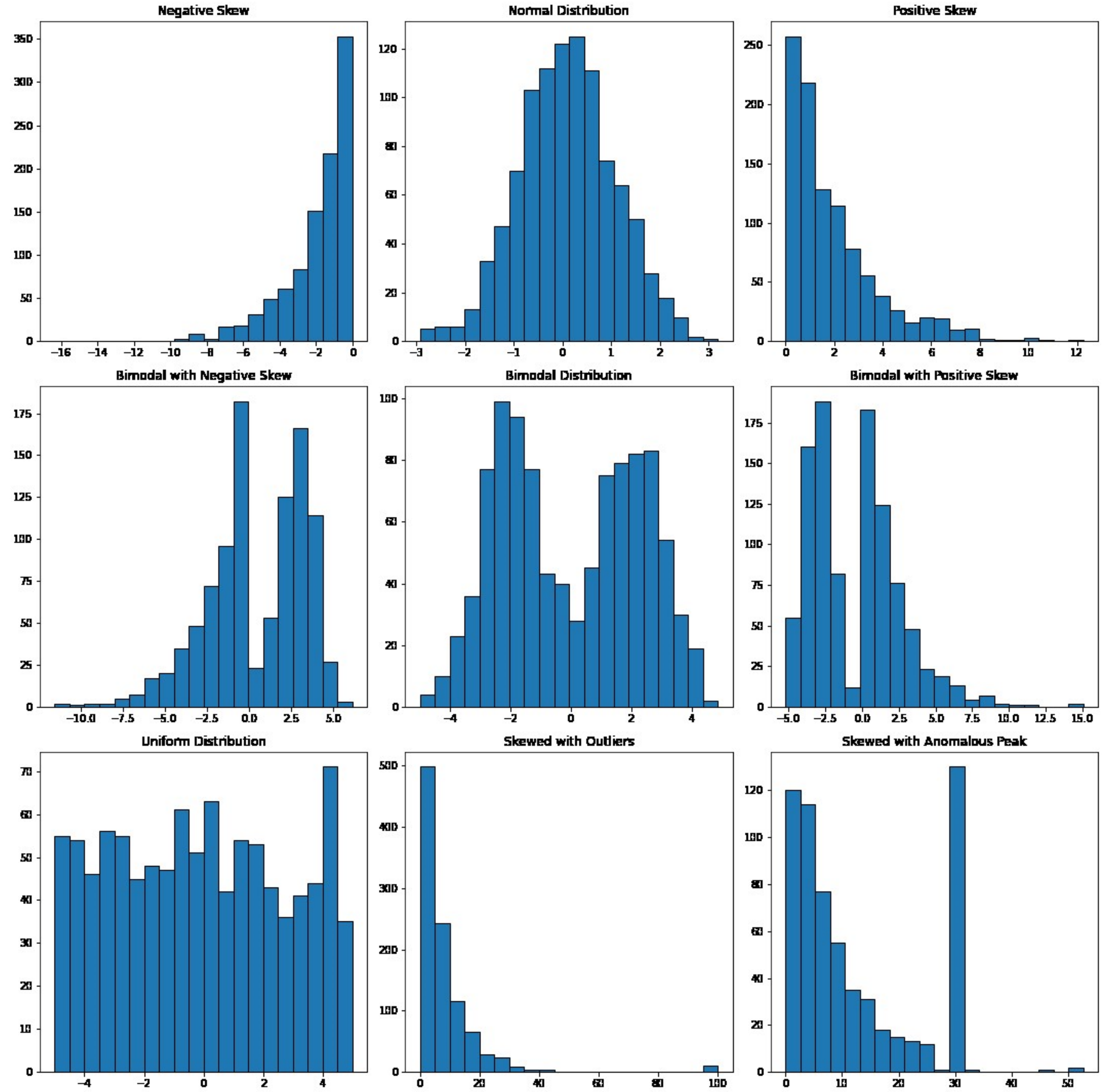
A **distribution** is a function that describes how frequently each possible value (or class of values) occurs.

We can approximate this by looking at what shape the data makes. This may give us insight into the behaviour of the variable. You might also hear the phrase the **structure** of the data.

We can see the distribution (or structure) of the data using a **histogram** to display:

- raw or scaled data.
- probability of each value in the dataset.

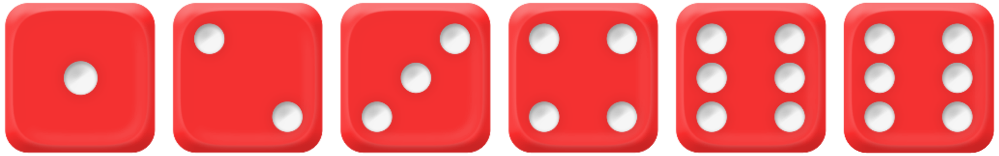
DISTRIBUTIONS: RAW DATA USING HISTOGRAMS



DISTRIBUTION

When we make an individual observation of a random variable, we can think of the value we observe as being drawn from a **probability distribution** of all the possible values.

Example: a biased die has the following faces.

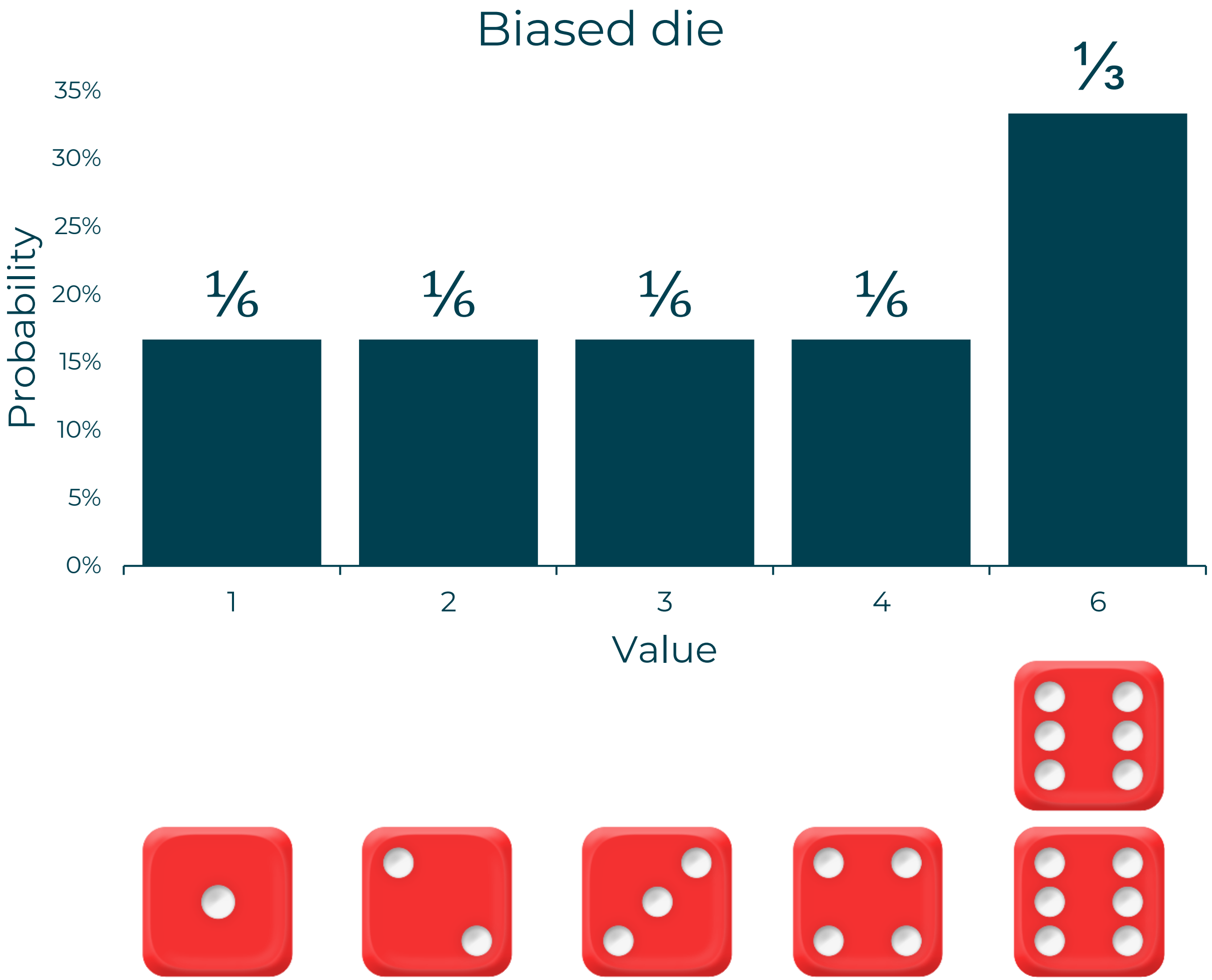


Question

What is the probability of rolling a 6?

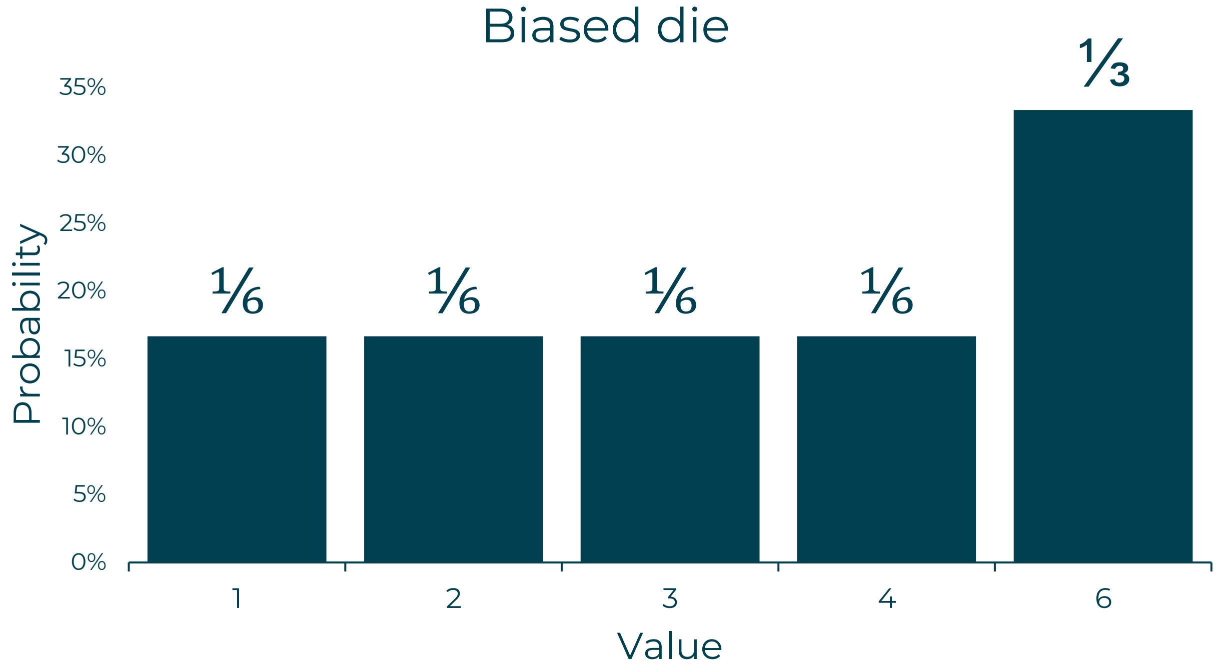
DISTRIBUTION

The probability distribution for the biased die:



DISTRIBUTION

The probability distribution for the biased die is a **discrete** probability distribution: the number of possible outcomes is **finite**.

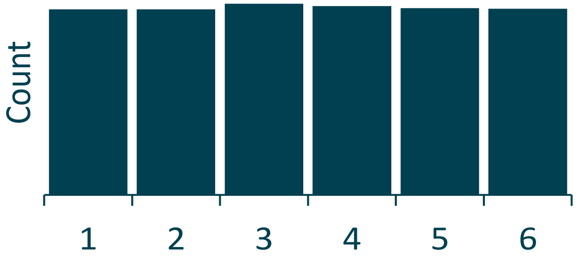


The roll of a die is a **discrete random variable**.

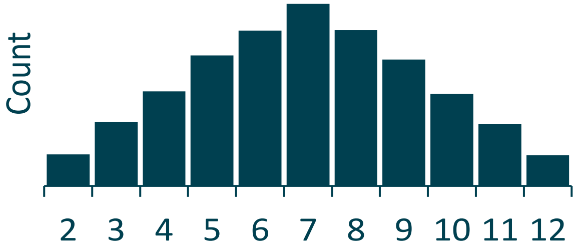
EXPERIMENTS WITH DICE: FINDINGS

Findings:

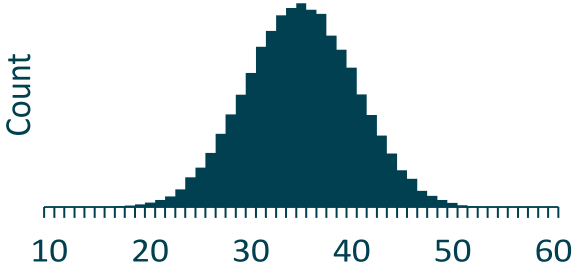
- One six-sided die



- Two six-sided dice

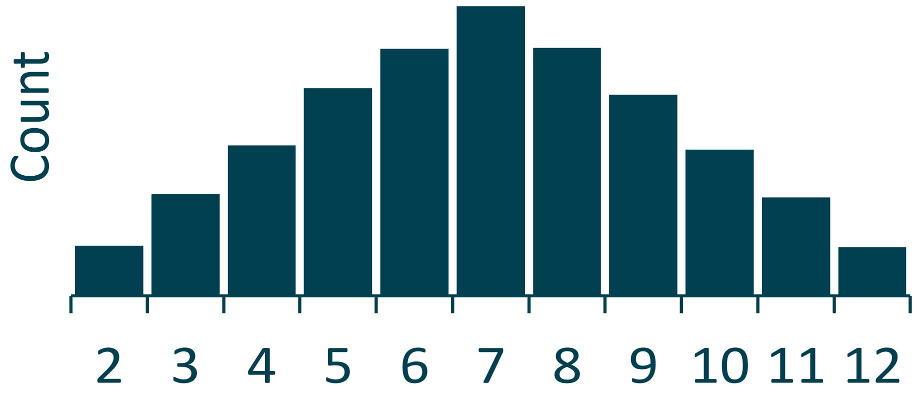


- Ten six-sided dice



ACTIVITY: ESTIMATING A PROBABILITY DISTRIBUTION

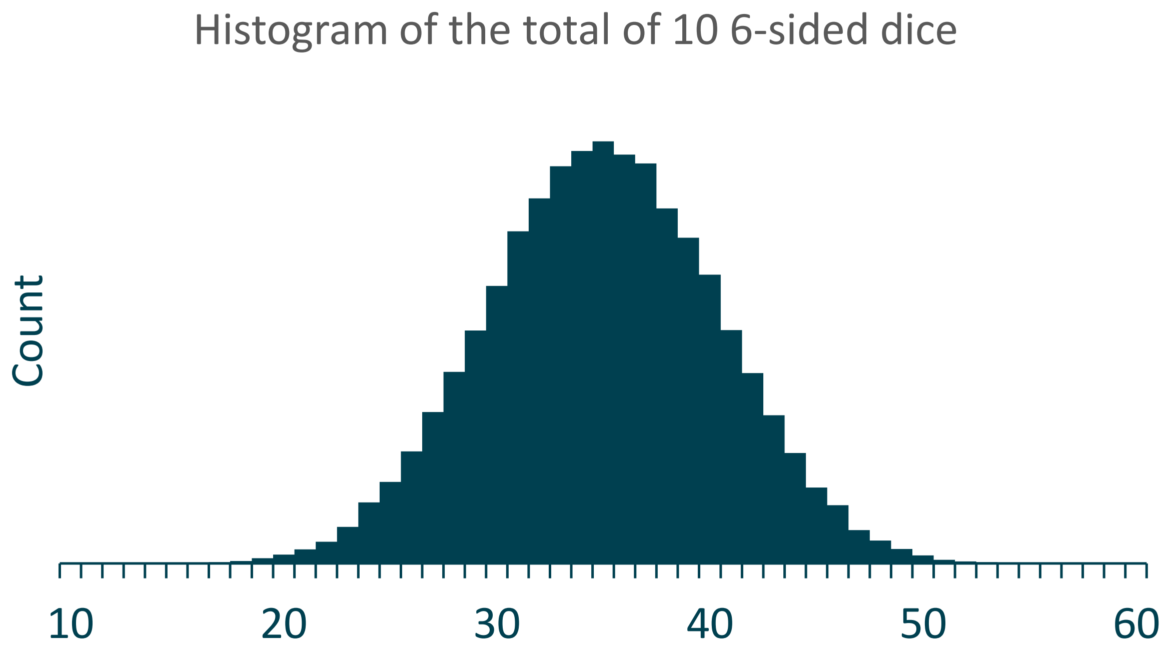
The histogram for two six-sided dice has the following symmetric shape:



Estimate and visualise, using Excel, the probability distribution of the random event associated to the outcome of launching two six-sided dice.

THE NORMAL DISTRIBUTION

The histogram for ten six-sided dice has a special shape: it is very close to the **normal distribution**.



ACTIVITY: INTRODUCTION TO THE NORMAL DISTRIBUTION

Open file

‘Introduction to the Normal Distribution v{#.#.#}.xlsx’

THE NORMAL DISTRIBUTION: RECAP

Question 1

The probability distribution of a random variable tends to approximate the normal distribution when the variable is produced by many random phenomena acting in what ways?

a _ _ _ _ _ _ _ ly and i _ _ _ _ _ _ _ _ _ _ ly.

Question 2

What properties of the normal distribution are represented by μ and σ?

Question 3

Approximately what percentage of the area of the normal distribution lies within one standard deviation of the mean?

Question 4

What is the total area under the normal distribution curve?

Activity: Understanding statistic types

In your online learning content for Module 4, you were introduced to different types of statistics:

- **Hypotheses Testing**
- **P-value**

Task

1. The class is split in break-out sessions to review and analyse the different statistics, and report back to the class a brief description of the same supported by notes.

2. Discuss as a class what is the use of each statistics and provide examples of practical use of the same, particularly and ideally within your working organisations.

STATISTICAL INFERENCE: STEPS TO TAKE WHEN TESTING HYPOTHESES

- **State the null and alternative hypotheses** – how you think the population behaves: **H_0** and **H_1** with the value(s) of any parameters and a descriptive sentence to help you interpret the results later. Identify if the test is one or two tailed.

- Decide the **significance level**, α.

- **State that you are assuming the null hypothesis is true** and state any logical deductions from this (e.g., the assumed distribution of the population).

- **Use a sample to calculate a test statistic or p-value** (based on the assumption that the null hypothesis is true).

- Check **if** the **test statistic is extreme or** if **the p-value is low** – if either is the case **you can reject the assumption that H_0 is true**.
- The result from the sample is significant and conclude (in context) that there might be enough evidence to suggest the alternative is true.

SIGNIFICANCE LEVEL

The **significance level**, α.

How unlikely does something have to be before we start to wonder if we have been making bad assumptions?

We can’t set it at 0% or we would **always** stick with our initial assumption and there would be no point doing the test!

But it also represents the chance of switching to an alternative hypothesis incorrectly – it’s the chance of a **false positive**.

The level should be decided **before** calculations on the sample – to avoid scientific bias.

T-TEST: AN EXAMPLE OF A/B TESTING

Basic idea: Are the means of two samples similar or different? If they are similar, they could come from the same population.

Examples:

- Is the average sentiment value from before and after a marketing campaign the same, or have we made a positive impression?

  Note: This is an example of **1 tailed**: specifically, have we made a positive impact (rather than just, is it different, which would be **2 tailed**)?

- Is the average investment performance different if we change strategy?

A/B testing can be used to compare before vs. after, or distinct groups in the same time frame.

T-TEST EXAMPLE SET UP

Group A: all the investments follow the standard strategy.

Group B: all the investments use the new strategy.

H_0: The population means of group A and B are the same, i.e., it makes no difference which strategy is used.

H_1: The population means of group A and B are different, i.e., it does make a difference which strategy is used.

This is **2 tailed**.

Significance level: 5% = 0.05. As this is 2 tailed, the significance level is split when comparing to the p-value later: 0.025

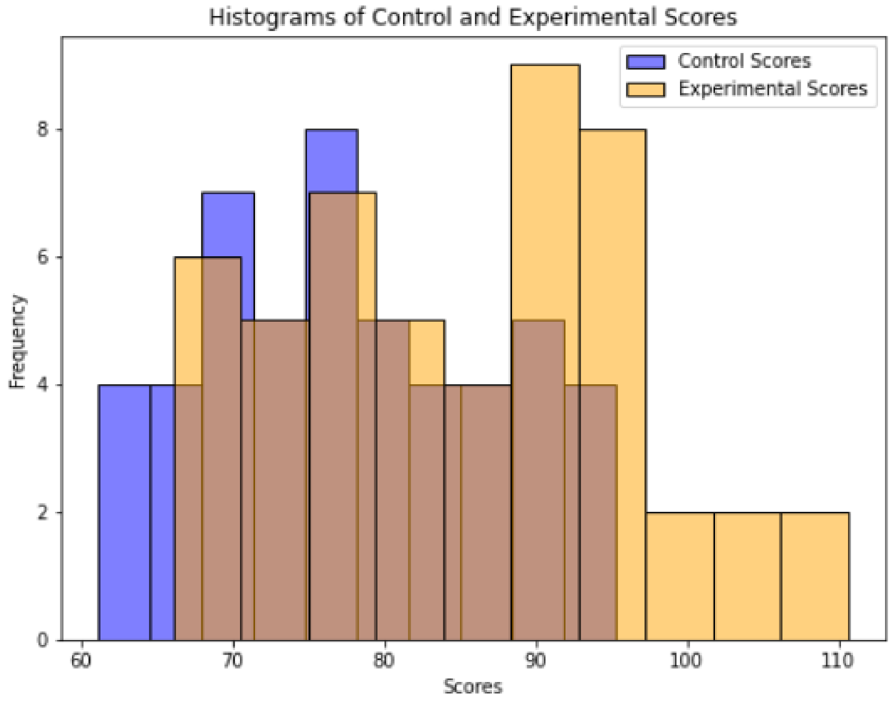
Assume the null hypothesis is true – this is done automatically by the tool or programming language you will use.

T-TEST VISUALISED AND IN CODE

Are these different enough or could they have come from the same population?

p-value: What's the chance the samples behave this way if the assumption is correct (that they come from the same population)?

We can use tools like the Data Analysis ToolPak in Excel to calculate the p-value.



ACTIVITY: COMPARE TWO INVESTMENT STRATEGIES

1. View and analyse the data contained in the file investment_performance.csv.
2. The scores are results from two investment strategies – the control group (using the current method) and the experimental group (using a new investment strategy).
3. Implement and A/B test to compare the means of the two control groups using the Data Analysis ToolPak in Excel.
4. In the context of the available data, decide if there is enough evidence to suggest that there is a difference between the two investment strategies in relation to their means.
5. Report and discuss the results in class.

Activity 4.2

Live Activity – How statistics, algorithms, and data modelling can support business decision–making.

Activity 4.2 task requirements

Tasks:

1. Discuss how algorithms could be used to support the following organisation types:
   - I. A training provider (any sector)
   - II. A hospital/GP surgery
   - III. An online retailer
   - IV. A train provider (e.g., Greater Anglia)

   In particular, provide four examples of how algorithms could be used as a solution to a business problem within those organisation types, including at least one example of automated algorithm.

2. Within the same context of four organisation types, provide examples of how statistics and statistical models could be used in the definition of relevant KPIs (e.g., total x, average x, change in x, trend direction of x, etc.).

Deliverables:

Tasks 1 and 2 will be delivered as a presentation (15/20 slides).

Q&A

Reflection, review, and 1:1 time

Close and final remarks

CONTINUE GATHERING EVIDENCE FOR ACTIVITY 4.3 BACK IN THE WORKPLACE

RE-VISIT MODULE 4 ONLINE LEARNING CONTENT TO CONSOLIDATE UNDERSTANDING

POWERING
POTENTIAL

THANK YOU

Hope you enjoyed this learning journey!