Statistics

3 Parts

Collecting Data [Gathering Information]

Analysing Data [Average Spread]

Displaying/Presenting Data [Graphs, Charts, etc.]

Types of Data:

Types of Data: [Lots of Words to Learn, (Sorry)]

Data

- Primary Data
- Secondary Data
- Uni-Variate
- Bi-Variate

Data

- Numerical
  - Discrete
  - Continuous

- Categorical
  - Nominal
  - Ordinal (Ordered)

Sampling

- Random Sample
- Biased Sample
- Reliability of Sample
- Population/Sample

Collecting Data / (Surveys)

- Face-to-Face
- Telephone
- Postal Questionnaire
- Online Questionnaire
- Observation

See Definitions Sheet for definition of each term, and examples of each.
Statistics Definitions

<table>
<thead>
<tr>
<th>Type of Data</th>
<th>Definition</th>
<th>Sample Question/Example of Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Numerical Data</td>
<td>Data which is recorded as numbers</td>
<td>How many brothers/sisters do you have?</td>
</tr>
<tr>
<td>Discrete (Numerical) Data</td>
<td>Can only have a fixed number of values/answers</td>
<td>How many bedrooms are in your house?</td>
</tr>
<tr>
<td></td>
<td></td>
<td>What is your shoe size? (note: can’t be 11.345)</td>
</tr>
<tr>
<td>Continuous (Numerical)</td>
<td>Can have an infinite number of possible answers, is usually measured on a</td>
<td>What is your height?</td>
</tr>
<tr>
<td>Data</td>
<td>scale</td>
<td></td>
</tr>
<tr>
<td>Categorical Data</td>
<td>Data which is not recorded as numbers</td>
<td>How do you get to school?</td>
</tr>
<tr>
<td>Ordinal Data</td>
<td>Data which can be ordered in some way</td>
<td>Junior Cert Grades (A, B, C, D, etc)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Month of Birth</td>
</tr>
<tr>
<td>Nominal Data</td>
<td>Categorical data which can’t be ordered</td>
<td>What mobile phone network do you use?</td>
</tr>
<tr>
<td></td>
<td></td>
<td>What is your favourite film?</td>
</tr>
</tbody>
</table>

Other definitions:

Data can be **Primary/Secondary**
- **Primary data** is collected by or for the person who is going to use it.
- **Secondary data** is data which is taken from another source

Data can be **Uni-variate** or **Bi-variate**
- **Uni-variate** means that you’re just interested in one thing at a time, for example, the height of students in a school
- **Bi-variate** data is “linked”/ “paired” data – so you might be interested in the hours spent studying and the marks in an exam of students in the school, to see if there is a link between the two...

Samples:
- The population is the entire group that is being studied
- A sample is a group that is taken/selected from the population

A **Simple Random Sample** is a sample in which each person in the population has an equal chance of being selected

A **Biased Sample** is a sample which does not fairly represent the population. For example, if I was trying to find out what the most popular sport in Dublin was, and I decide to ask 1,000 people coming out of the All-Ireland Hurling Final, this might be a biased sample.

Miscellaneous:

A **Leading Question** is one which suggests a possible answer. For example: “Taxes are too high: Should they be reduced?”
DESIGNING QUESTIONNAIRES

NEEDS TO (BE):

- CLEAR / EASY TO UNDERSTAND
- USEFUL / RELEVANT
- ALLOW ALL POSSIBLE ANSWERS
- HAVE NO LEADING QUESTIONS
- ASK ONLY ONE QUESTION AT A TIME.

FREQUENCY TABLES

Often when we collect data we are interested in how often something occurs.

E.g. the number of different types of vehicle which pass by a particular set of traffic lights.

We make a frequency table using a "tally" to keep track of counts.

E.g. TALLY

<table>
<thead>
<tr>
<th>Hit</th>
<th>Hit</th>
<th>Hit</th>
<th>Hit</th>
<th>Hit</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>7</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

= 13

As we mark off each "fifth" mark, we have a group of 5.

Why?

This is easier to count.
PRESENTING DATA

- BAR CHARTS / LINE PLOTS
- PIE CHARTS
- STEM & LEAF DIAGRAMS
- HISTOGRAMS [LIKE A BAR CHART]

BAR CHARTS

- HEIGHT OF BAR - IMPORTANT
- WIDTH OF EACH BAR IS SAME
- ALWAYS INCLUDE LABELS & UNITS

MONTHLY RAINFALL

LINE PLOT

- BASICALLY THE SAME BUT LINES INSTEAD OF BARS
- USES SYMBOL (Eg "X") TO MARK EACH ITEM OF DATA

CAR BRANDS IN CAR PARK

- FAVOURITE COLOUR

- SHOW HOW DAM IS DIVIDED / SHARED
- THE SIZE OF THE ANGLE REPRESENTS THE SIZE OF THE SHARE
- REMEMBER: FULL CIRCLE = 360°
**STEM - AND - LEAF DIAGRAMS**

1. SEPARATE EACH VALUE INTO 2 PARTS - **STEM**
   - EQ: 27 BECOMES 2 | 7
2. ALL LEAVES WHICH HAVE THE SAME STEM ARE
   PLOTTED TOGETHER - **LIKE A BAR CHART ON ITS SIDE**
   - **ARRANGE LEAVES IN ORDER OF SIZE**
3. MUST HAVE A "KEY"
4. EVERY PERSON / ITEM IS REPRESENTED BY A LEAF

---

**MARKS IN AN EXAM**

- EACH ITEM OF DATA HAS ITS OWN LEAF.
- NUMBERS GO IN ORDER OF SIZE
- THIS 9 DOESN'T MEAN 9, IT MEANS 49
  - BECAUSE ITS STEM IS 4
- DON'T FORGET THE KEY:
  - Key: 3|5 = 35

---

**HISTOGRAM**

- LIKE A BAR CHART - JOINED UP - NO GAPS
- USED FOR "CONTINUOUS" DATA, E.G. HEIGHT

---

- NOTE: BARS JOINED TOGETHER - NO GAPS
ANALYSING DATA

1. CENTRAL TENDENCY (AVERAGE)

2. SPREAD / VARIABILITY

1. AVERAGES
   (A) MEAN / MEDIAN / MODE

   (A) MEAN = \frac{\text{SUM OF THE NUMBERS}}{\text{NUMBER OF NUMBERS}}

   MOST COMMONLY USED AVERAGE.

(B) MEDIAN = THE "MIDDLE" NUMBER
   (NUMBERS MUST BE LISTED IN ORDER)

   CAN BE HALF-WAY BETWEEN 2 MIDDLE NUMBERS

   \[ 3, 4, 4, 5, 7, 8, 9, 10 \]

   MEDIAN = 6 < HALF-WAY BETWEEN 5 AND 7.

   METHOD: CROSS OFF BIGGEST + SMALLEST
   KEEP DOING THIS UNTIL YOU'RE LEFT WITH EITHER
   1 OR 2 "MIDDLE" NUMBERS

   YOU CAN BE ASKED TO
   FIND MEDIAN OF A STEM
   AND LEAF DIAGRAM.

NOTE:
THE MEDIAN SPLITS
THE DATA IN 2, THERE
SHOULD BE SAME NUMBER
OF NUMBERS LESS THAN
THE MEDIAN AS THERE ARE
NUMBERS BIGGER THAN THE
MEDIAN.

(C) MODE

MOST OFTEN

MOST COMMON VALUE IN THE SET OF
USED FOR CATEGORICAL DATA.

E.G. CAN'T GET MEAN / MEDIAN OF
COLOUR OF CAR / EYES ETC.
Choosing which average to use:

- If there are extreme values, use the median.
- Otherwise, use the mean.
- Unless it's categorical data, when we have to use mode.

Why? Extreme values make the mean way too big or too small. The median is not affected by extreme values.

Eq: A small company employs 5 people, with the following salaries:

- €20,000
- €30,000
- €31,000
- €22,000
- €200,000

The boss: "The mean!"

Median = €31,000

Mean = \( \frac{20,000 + 30,000 + 31,000 + 22,000 + 200,000}{5} = \$62,000 \)

Way too high. Doesn't represent average salary very well.
**Spread** [or Variability]

- How "spread out" is our data?

- This can be important to us if we're doing a survey/statistical analysis. We can see how consistent/agreed the data is.

For your junior cycle, there is only one measure of spread: the **Range**

\[ \text{Range} = \text{biggest} - \text{smallest} \]

Be careful, this is not the same range we're talking about in functions, i.e., Domain + Range.

ie difference between top + bottom.