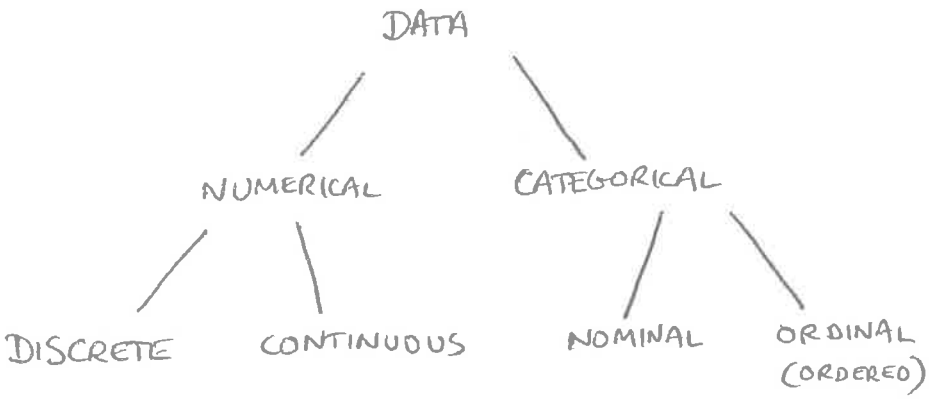
STATISTICS

3 PARTS
- COLLECTING DATA [GATHERING INFORMATION]
- ANALYSING DATA [AVERAGE, SPREAD]
- DISPLAYING / PRESENTING DATA [GRAPHS, CHARTS, ETC]

TYPES OF DATA : [LOTS OF WORDS TO LEARN, (SORRY)]



- PRIMARY DATA
- SECONDARY DATA
- UNI-VARIATE
- BI-VARIATE

SAMPLING

- RANDOM SAMPLE
- BIASED SAMPLE
- RELIABILITY OF SAMPLE
- POPULATION / SAMPLE

SEE DEFINITIONS SHEET FOR DEFINITION OF EACH TERM, AND EXAMPLES OF EACH

COLLECTING DATA /(SURVEYS)

- FACE-TO-FACE
- TELEPHONE
- POSTAL QUESTIONNAIRE
- ONLINE QUESTIONNAIRE
- OBSERVATION

KNOW PROS / CONS OF EACH :

- COST
- BIASED SAMPLE
- CLEAR / EASY TO UNDERSTAND

Statistics Definitions

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

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 ← QUESTIONS WHICH SUGGEST A POSSIBLE ANSWER
- ASK ONLY ONE QUESTION AT A TIME.

FREQUENCY TABLES

OFTEN WHEN WE COLLECT DATA WE ARE INTERESTED IN HOW OFTEN SOMETHING OCCURS.

eg THE NUMBER OF DIFFERENT TYPES OF VEHICLE WHICH PASS BY A PARTICULAR SET OF TRAFFIC LIGHTS.

WE MAKE A FREQUENCY TABLES USING A "TALLY" TO KEEP TRACK OF COUNTS.

eg TALLY : 𝍸 𝍸 III = 13

AS WE MARK OFF EACH "FIFTH" MARK, WE HAVE A GROUP OF 5

→ WHY? THIS IS EASIER TO COUNT

PRESENTING DATA

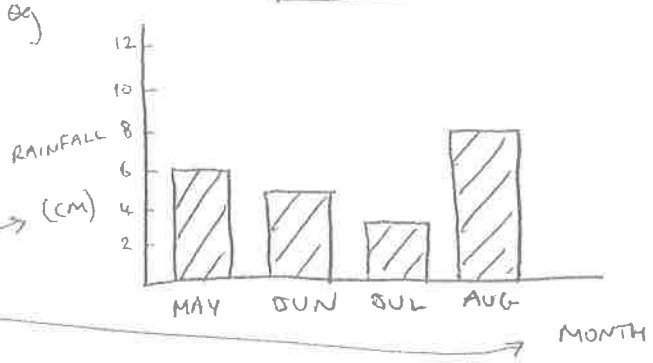
- BAR CHARTS / LINE PLOTS
- PIE CHARTS
- STEM & LEAF DIAGRAMS
- HISTOGRAMS [LIKE A BAR CHART]

BAR CHARTS

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

eg

MONTHLY RAINFALL

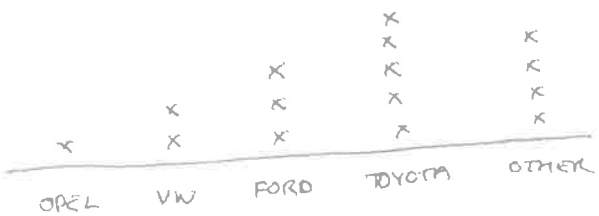


LINE PLOT

- BASICALLY THE SAME BUT LINE> INSTEAD OF BARS
- USES SYMBOL (EG "X") TO MARK EACH ITEM OF DATA

eg

CAR BRANDS IN CAR PARK



PIE CHARTS

- SHOW HOW DATA IS DIVIDED / SHARED
- THE SIZE OF THE ANGLE REPRESENTS THE SIZE OF THE SHARE
- REMEMBER: FULL CIRCLE = 360°

FAVOURITE COLOUR

STEM-AND-LEAF DIAGRAMS

- SEPARATE EACH VALUE INTO 2 PARTS < STEM / LEAF

  eg 27 BECOMES 2|7

- ALL LEAVES WHICH HAVE THE SAME STEM ARE PLOTTED TOGETHER
  - LIKE A BAR CHART ON ITS SIDE
  - ARRANGE LEAVES IN ORDER OF SIZE.

- MUST HAVE A "KEY"

- EVERY PERSON / ITEM IS REPRESENTED BY A LEAF

eg MARKS IN AN EXAM

Stem	Leaf
0	7, 9
1	3, 4
2	9
3	5
4	3, 4, 4, 7, (9)
5	2, 2, 3, 5, 7, 7, 8
6	1, 2, 5, 8, 9
7	3, 4, 5, 9
8	4, 7
9	1

Key: 3|5 = 35

- EACH ITEM OF DATA HAS ITS OWN LEAF.

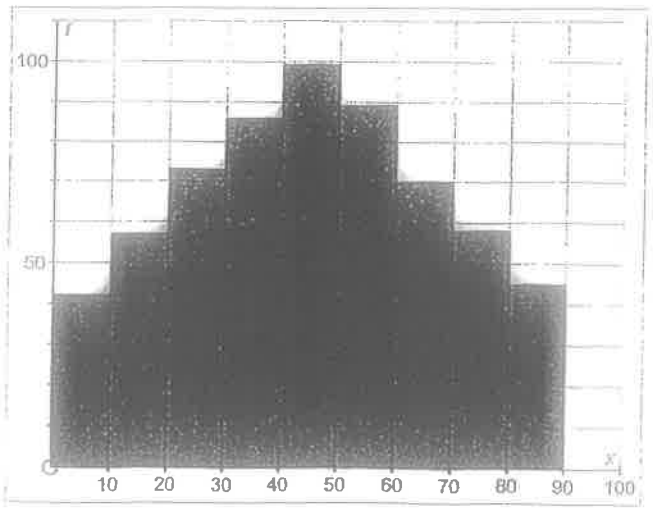
NUMBERS GO IN ORDER OF SIZE

THIS 9 DOESN'T MEAN 9. IT MEANS 49 BECAUSE ITS STEM IS 4

DON'T FORGET THE KEY.

HISTOGRAM [LIKE A BAR CHART - JOINED UP - NO GAPS]

- USED FOR "CONTINUOUS" DATA. eg HEIGHT



- NOTE: BARS JOINED TOGETHER - NO GAPS

ANALYSING DATA

2 THINGS
- ① CENTRAL TENDENCY (AVERAGE)
- ② SPREAD / VARIABILITY

① AVERAGES — (A) MEAN / (B) MEDIAN / (C) 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

eg 3, 4, 4, 5 | 7, 8, 9, 11

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

③ MODE — MOST OFTEN

- MOST COMMON VALUE IN THE SET OF DATA.
- USED FOR CATEGORICAL DATA.
- eg 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 → THIS IS OUR FAVOURITE. WE WANT TO 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.

eg A SMALL COMPANY EMPLOYS 5 PEOPLE, WITH THE FOLLOWING SALARIES : ("THE BOSS"!!)

€20,000 €30,000 €31,000 €32,000 €200,000

MIDDLE NUMBER

MEDIAN = **€31,000** → WE USE THIS AS A GOOD AVERAGE

$$\text{MEAN} = \frac{20\,000 + 30\,000 + 31\,000 + 32\,000 + 200\,000}{5} = €62,600$$

WAAAY TOO HIGH DOESN'T REPRESENT "AVERAGE" SALARY VERY WELL.

② 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 1 MEASURE OF SPREAD: THE **RANGE**

[BE CAREFUL, THIS IS NOT THE SAME RANGE WE'RE TALKING ABOUT IN FUNCTIONS, ie DOMAIN + RANGE]

- **RANGE = BIGGEST − SMALLEST**

[ie DIFFERENCE BETWEEN TOP + BOTTOM]