

Data and variables

DATA: the answers to questions or measurements from the experiment

VARIABLE = measurement which varies between subjects
e.g. height or gender

One variable per column

	A	B	C	D
	Subject ID	Gender	Year of study	Height
1	1	Male	1	170
2	2	Female	2	160
3	3	Female	3	165
4	4	Male	PG	175
5	5	Female	3	168

One row per subject

Data types

Data Variables

```
graph TD; A[Data Variables] --> B[Scale]; A --> C[Categorical: appear as categories, Tick boxes on questionnaires];
```

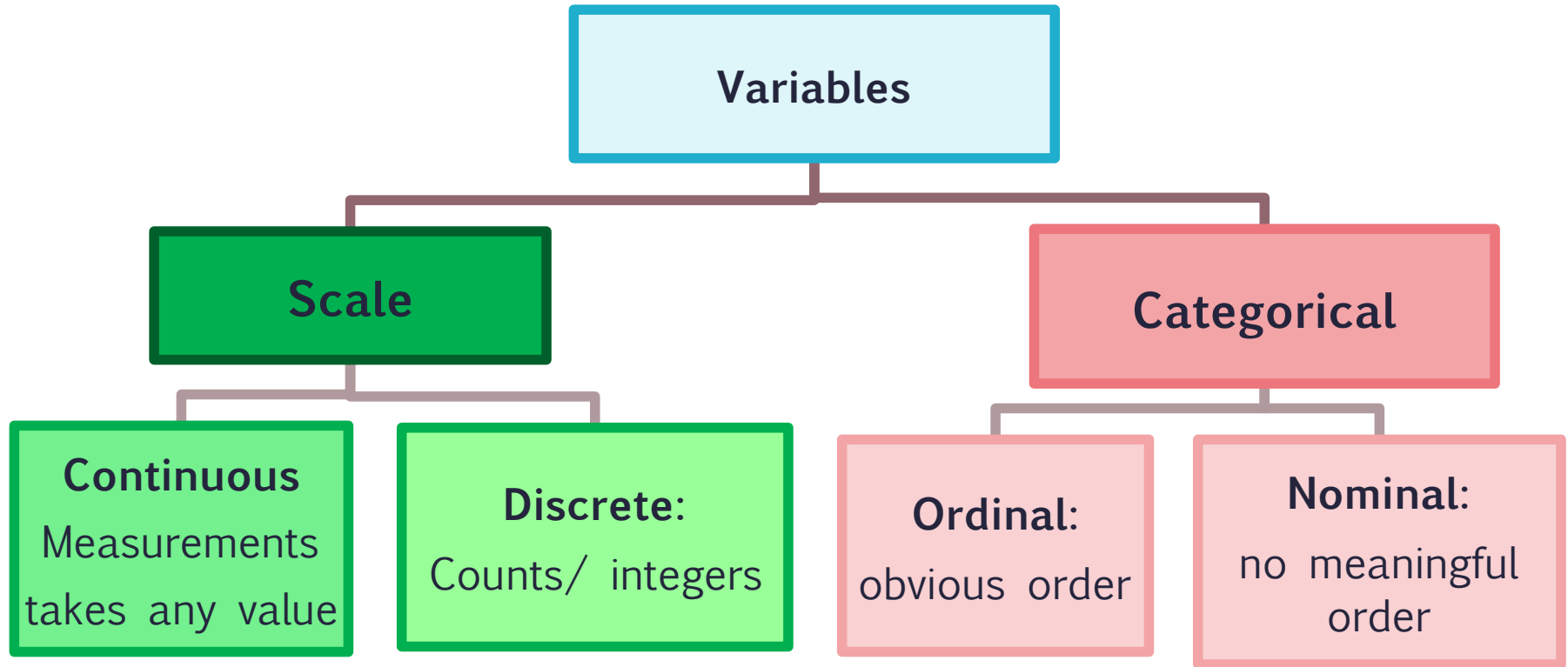
Scale

Measurements/ Numerical/
count data

Categorical:

appear as categories
Tick boxes on questionnaires

Data types



What data types relate to following questions?

➤ Q1: What is your favourite subject?

Maths	English	Science	Art	French
-------	---------	---------	-----	--------

▶ Q2: Gender:

Male	Female
------	--------

▶ Q3: I consider myself to be good at mathematics:

Strongly Disagree	Disagree	Not Sure	Agree	Strongly Agree
-------------------	----------	----------	-------	----------------

➤ Q4: Score in a recent mock GCSE maths exam:

Score between 0% and 100%

What data types relate to following questions?

➤ Q1: What is your favourite subject?

Nominal

Maths

English

Science

Art

French

➤ Q2: Gender:

Male

Female

Binary/ Nominal

➤ Q3: I consider myself to be good at mathematics:

Strongly
Disagree

Disagree

Not Sure

Agree

Strongly
Agree

Ordinal

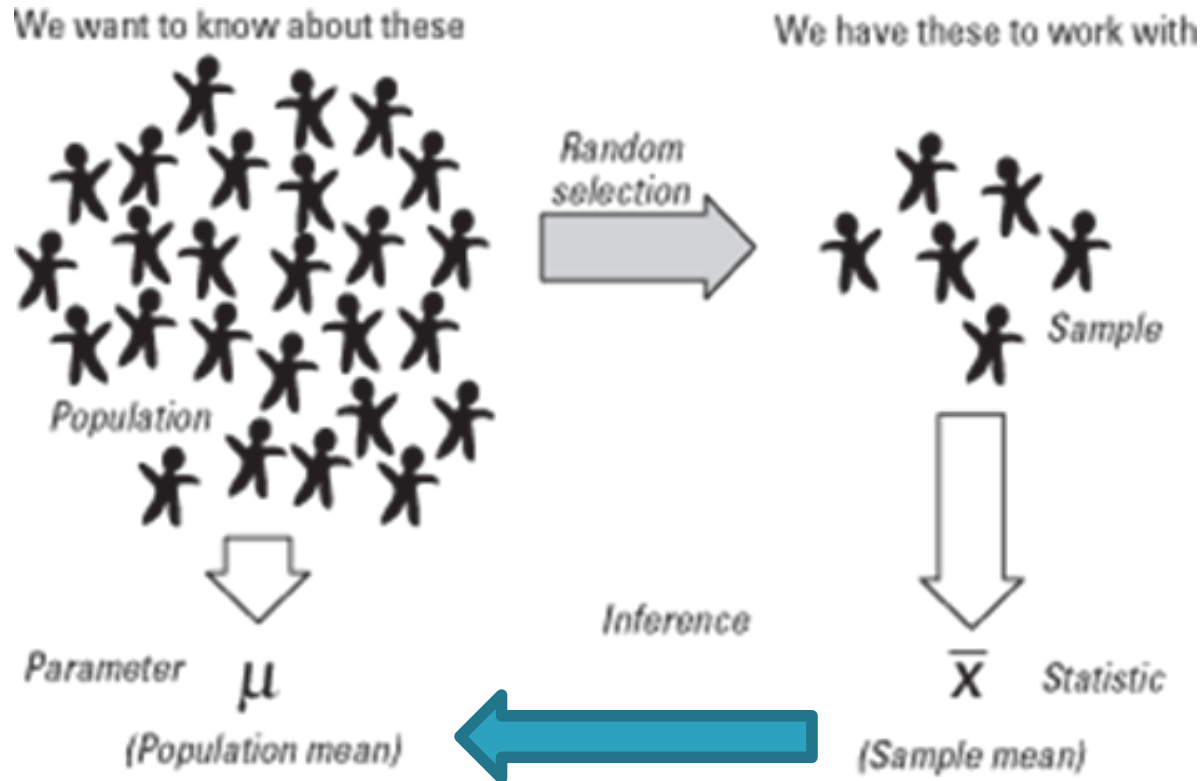
➤ Q4: Score in a recent mock GCSE maths exam:

Score between 0% and 100%

Scale

Populations and samples

- ▶ Taking a sample from a population



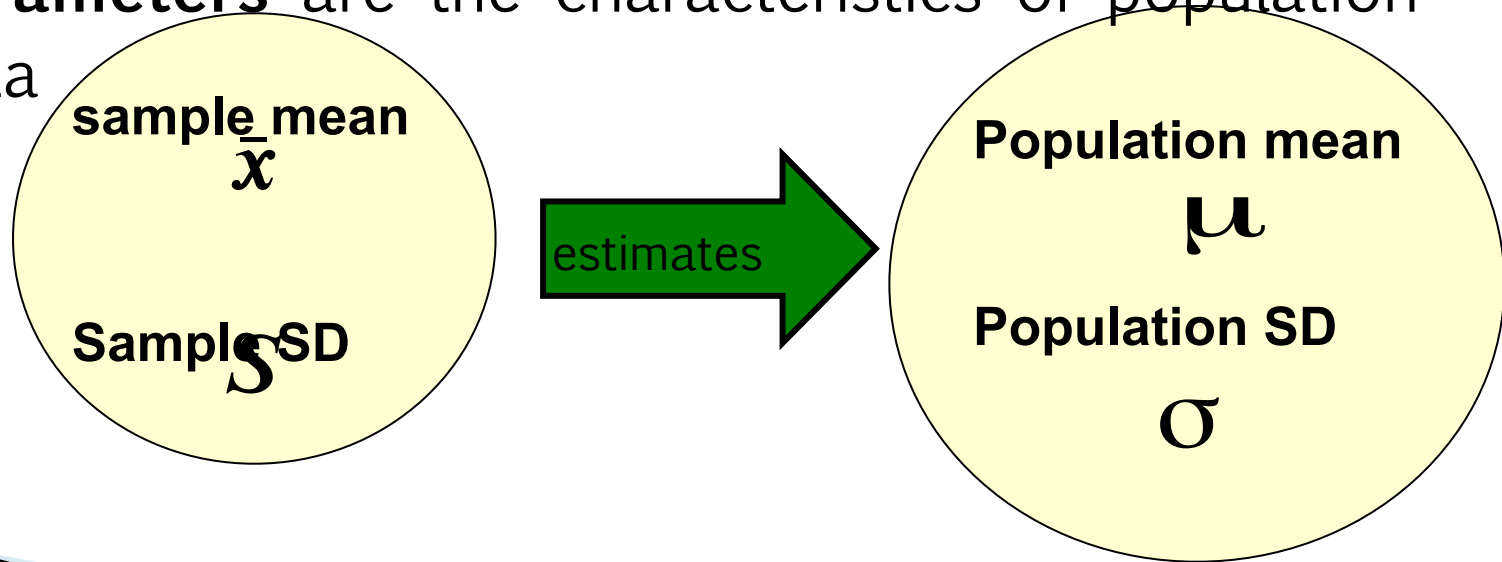
Sample data 'represents' the whole population

Point estimation

Sample data is used to estimate parameters of a population

Statistics are calculated using sample data.

Parameters are the characteristics of population data



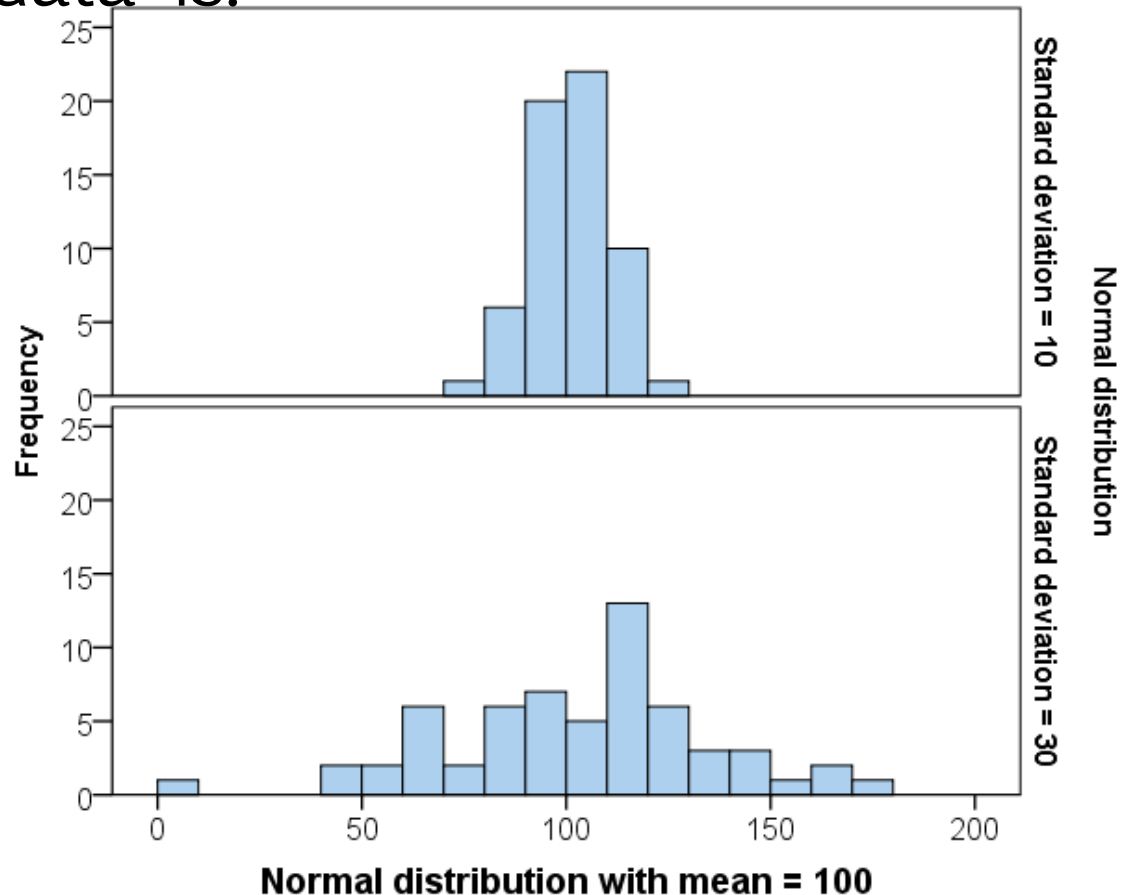
How can exam score data be summarized?

Exam marks for 60 students (marked out of 65)

48	37	1	33	26	22	15	22	40	30	12	36
21	20	29	13	44	52	28	39	16	48	56	27
47	12	35	24	10	36	18	34	9	25	31	42
31	27	64	25	58	17	26	38	28	43	33	5
25	55	7	32	39	23	49	43	11	37	22	54

Interpretation of standard deviation

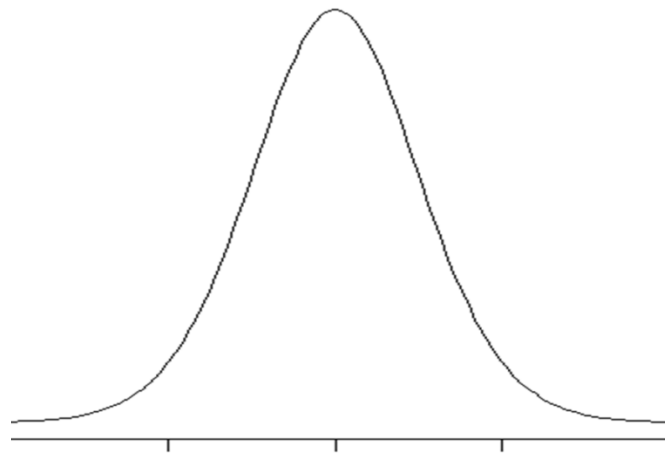
- ▶ The larger the standard deviation, the more spread out the data is.



Scale data

If have scale data
analyse it we often
assume it follows a

Normal distribution



Normal

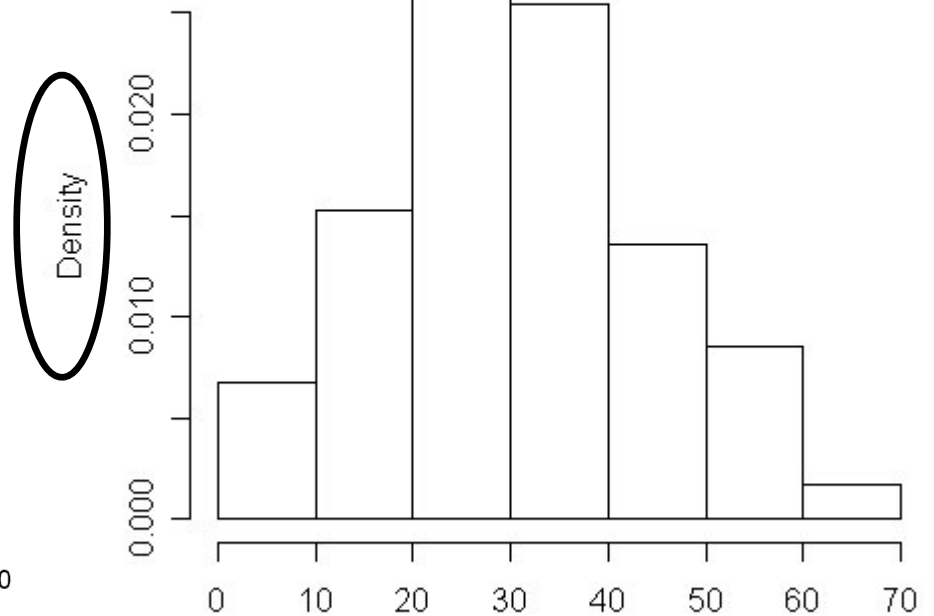
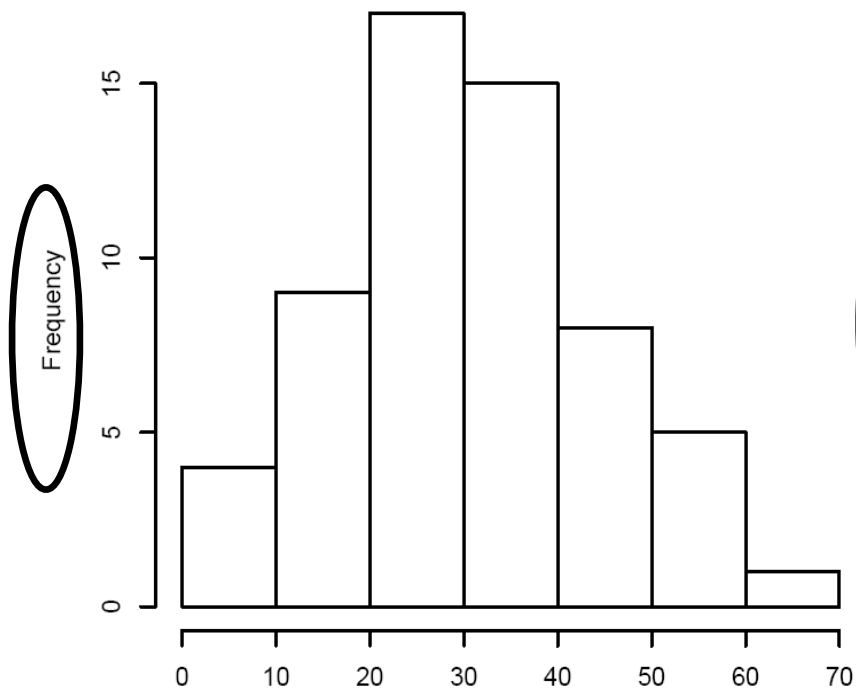
Discussion

- ▶ How could you explain to a someone what we mean by data being assumed to follow a Normal Distribution?

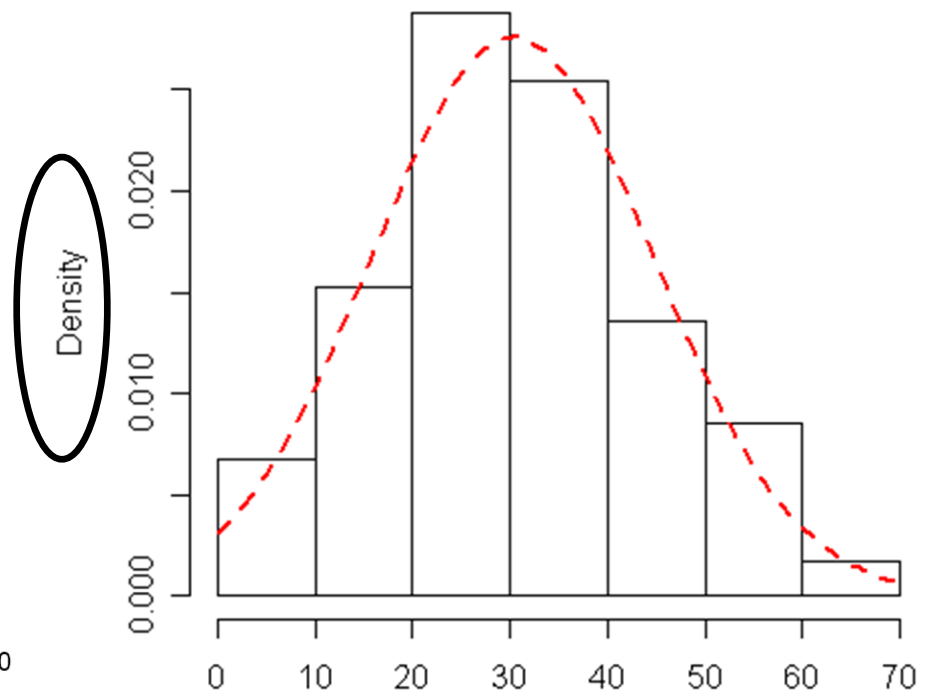
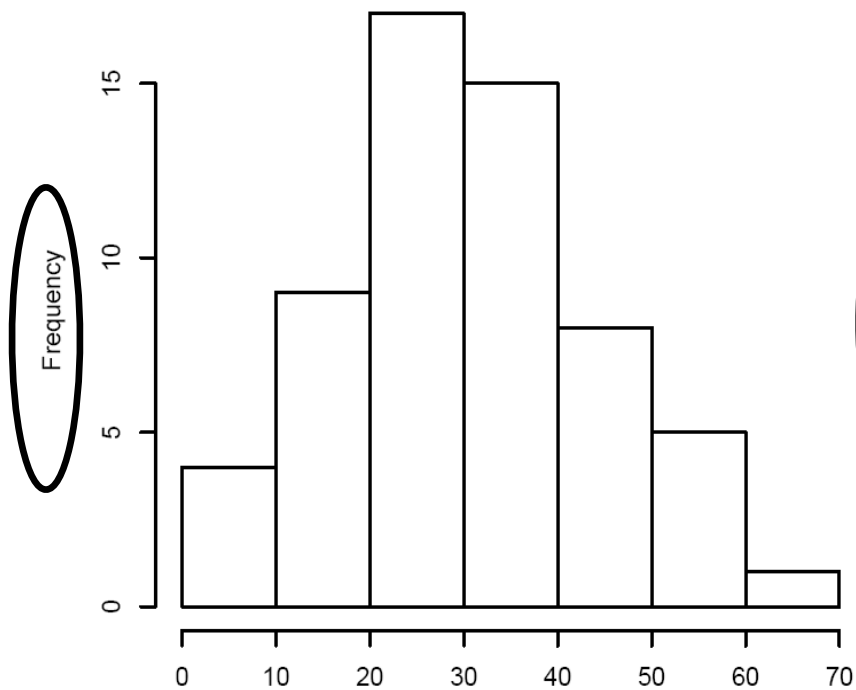
Group Frequency Table

	Frequency	Percent
0 but less than 10	4	6.7
10 but less than 20	9	15.0
20 but less than 30	17	28.3
30 but less than 40	15	25.0
40 but less than 50	9	15.0
50 but less than 60	5	8.3
60 or over	1	1.7
Total	60	100.0

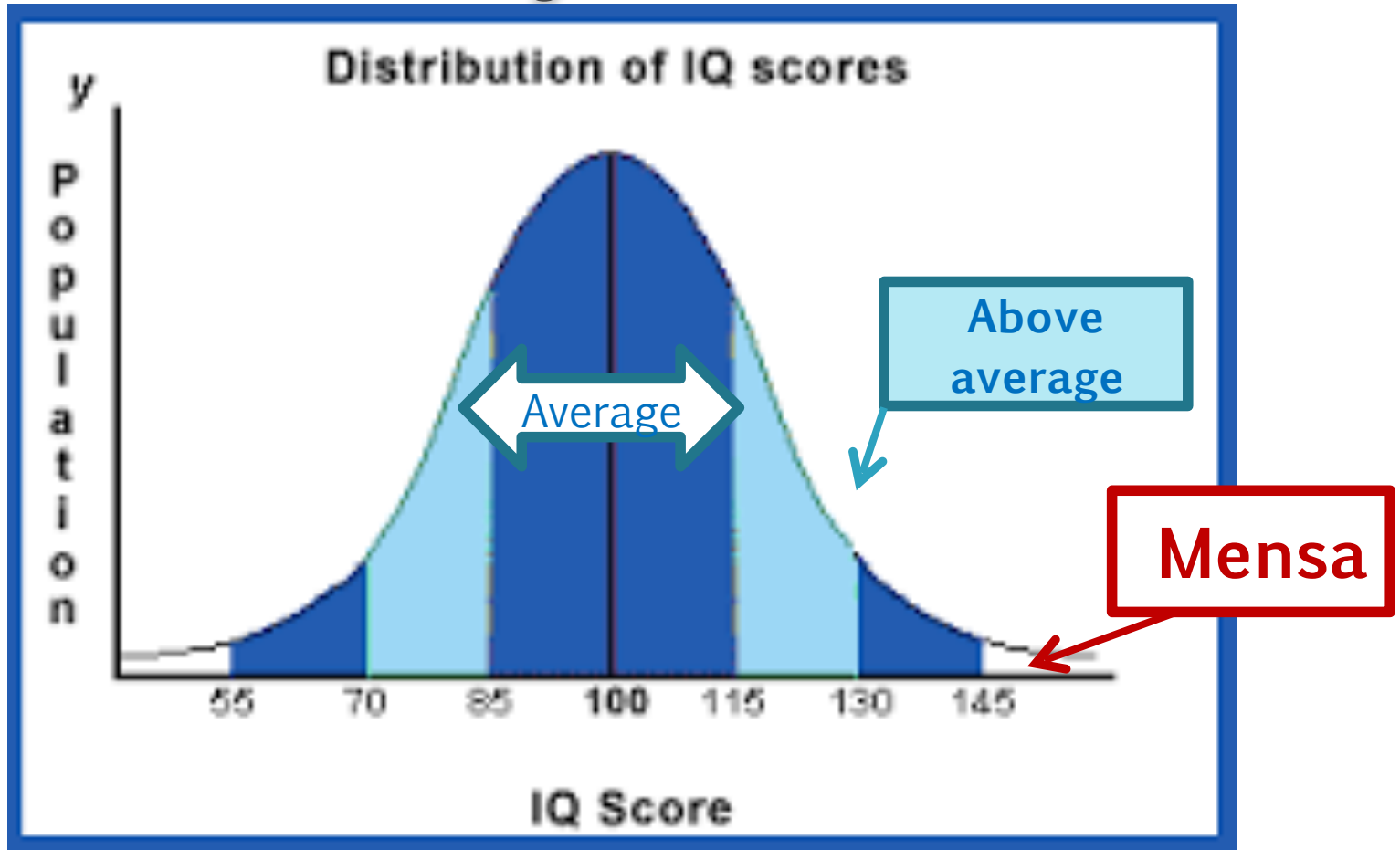
Histogram and Probability Distribution for Exam Marks Data



Histogram and Probability Distribution for Exam Marks Data

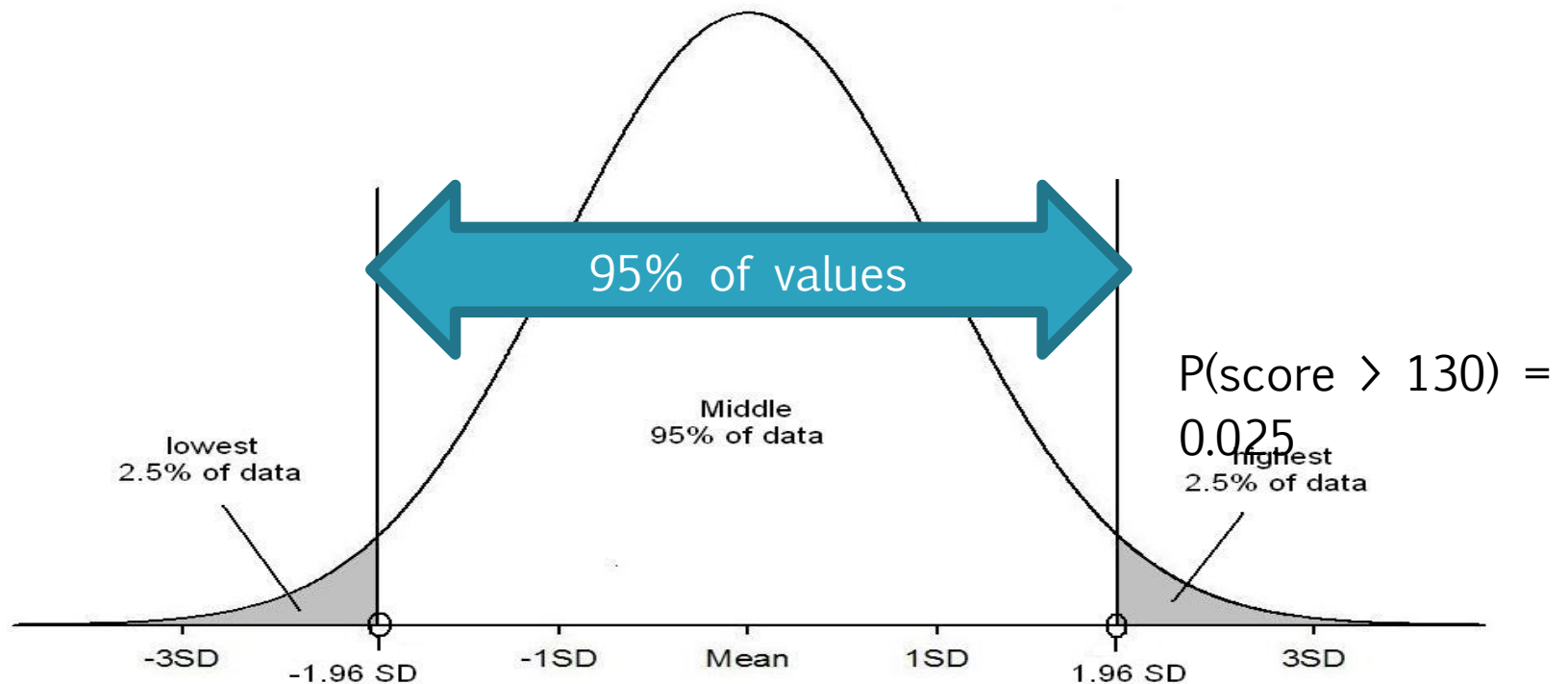


IQ is normally distributed



Mean = 100, SD = 15.3

95% 1.96 x SD's from the mean



$$P(\text{score} > 130) =$$

$$0.025$$

highest
2.5% of data

70

100

130

$$\text{mean} - (1.96 \times SD)$$

$$100 - (1.96 \times 15.3) = 70$$

$$\text{mean} + (1.96 \times SD)$$

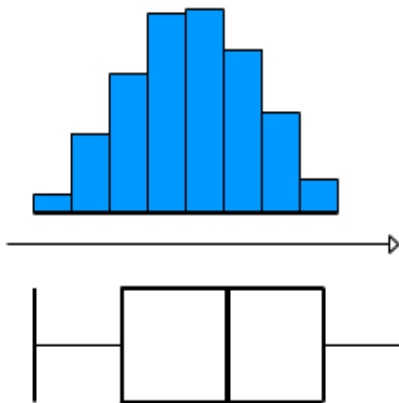
$$100 + (1.96 \times 15.3) = 130$$

95% of people have an IQ between 70 and 130

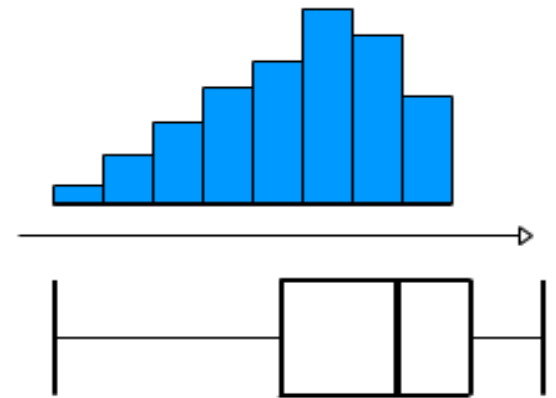
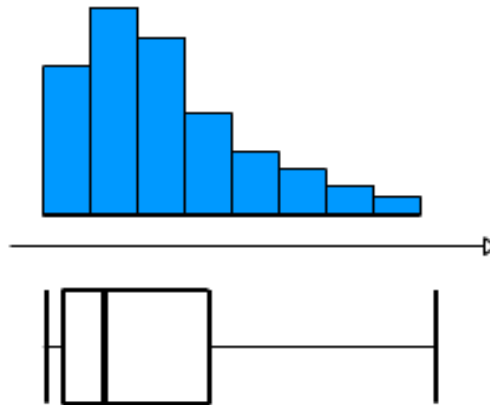
Assessing Normality

Charts can be used to **informally** assess whether data is:

Normally
distributed



Or...Skewed

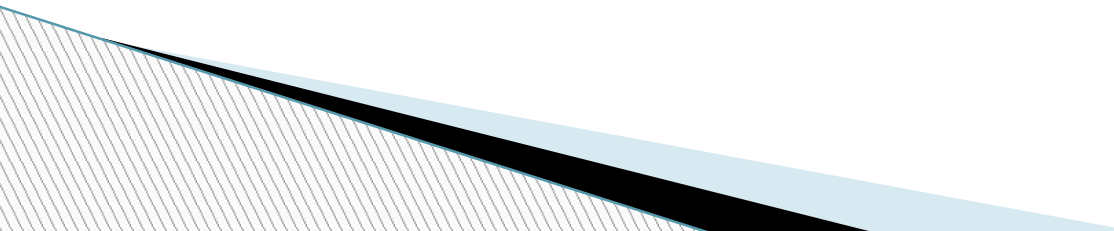


The mean and median are
very different for skewed
data.

Discussion

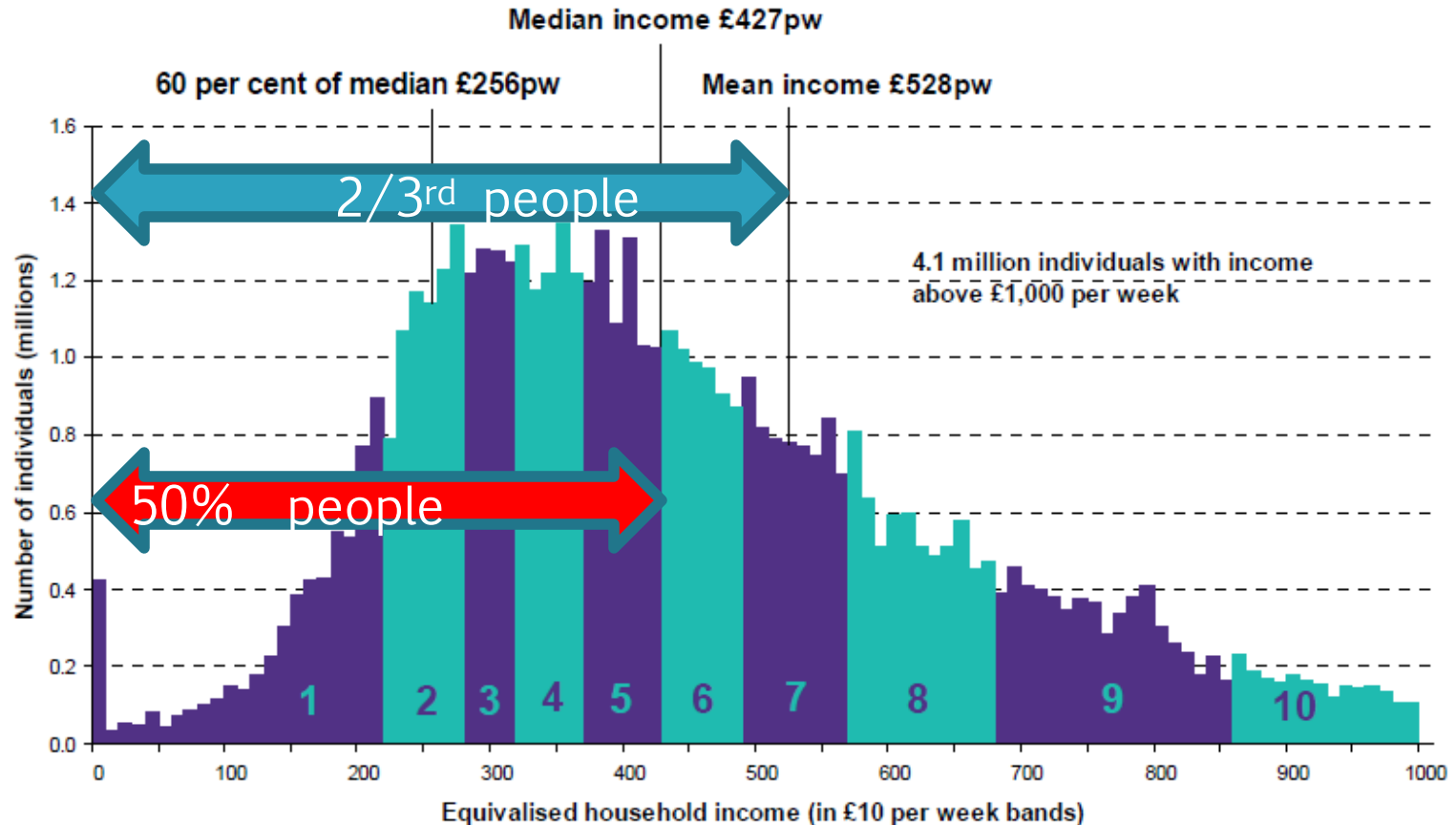
Is the following statement:

“2 out of 3 people earn less than the average income”

- A. True
 - B. False
 - C. Don't know
- 

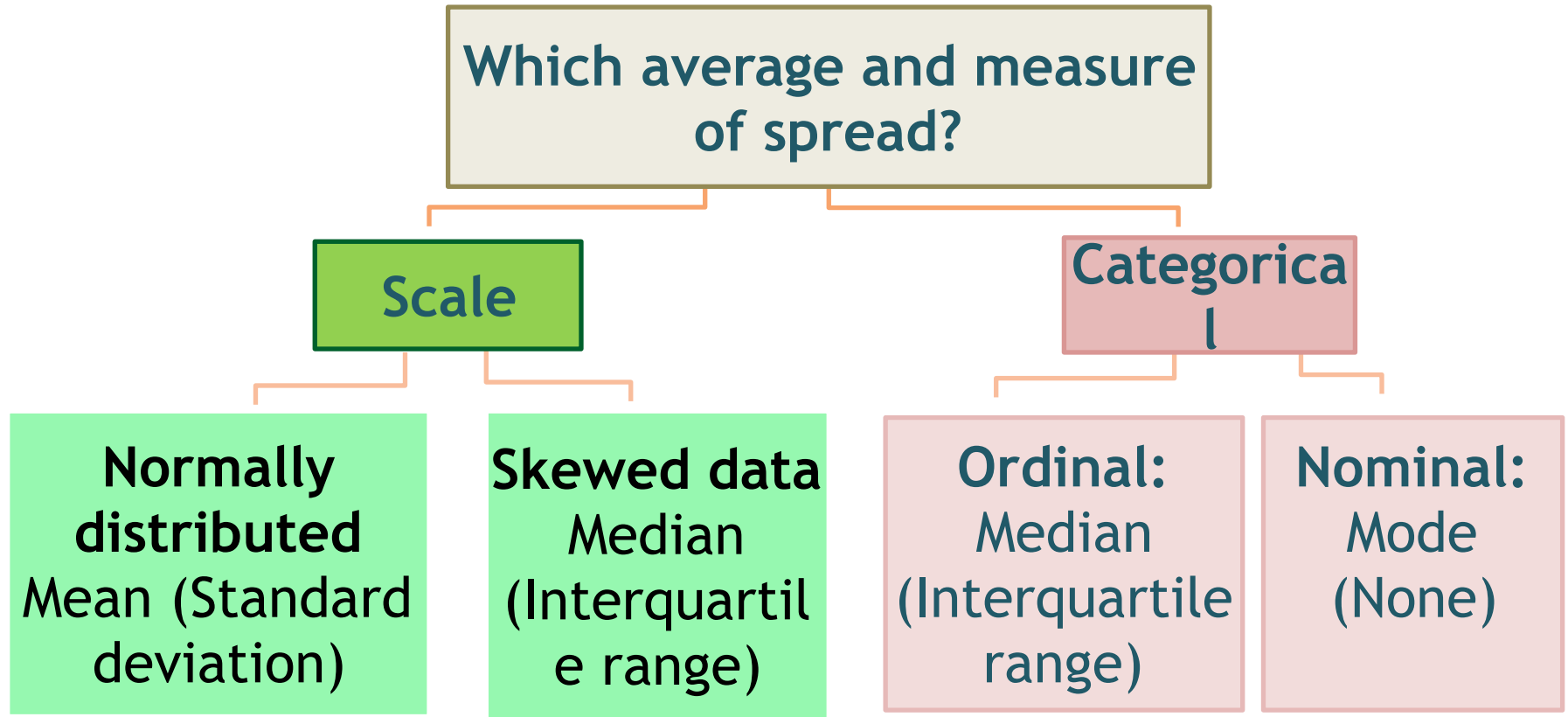
Sometimes the median makes more sense!

Chart 1.2 (BHC): Income distribution for the whole population, 2011/12



Source: Households Below Average Income: An analysis of the income distribution 1994/95 – 2011/12, Department for Work and Pensions

Choosing summary statistics



Which graph? Exercise

1 st variable	Only 1 variable	Scale	Categorical
Scale	Histogram	Scatter plot	Box-plot/ Confidence interval plot
Categorical	Pie/ Bar	Box-plot/ Confidence interval plot	Stacked/ multiple bar chart

Which graph would you use when investigating:

- 1) Whether daily temperature and ice cream sales were related?
- 2) Comparison of mean reaction time for a group having alcohol and a group drinking water

Exercise: Ticket cost comparison

Summary statistics for cost of Titanic ticket by survival

Cost of ticket	Survived?	
	Died	Survived
Mean	23.4	49.4
Median	10.5	26
Standard Deviation	34.2	68.7
Interquartile range	18.2	46.6
Minimum	0	0
Maximum	263	512.33

- a) Is there a big difference in average ticket price by group?
- b) Which group has data which is more spread out?
- c) Is the data skewed?
- d) Is the mean or median a better summary measure?

Which graph? **Solution**

1 st variable	Only 1 variable	Scale	Categorical
Scale	Histogram	Scatter plot	Box-plot/ Confidence interval plot
Categorical	Pie/ Bar	Box-plot/ Confidence interval plot	Stacked/ multiple bar chart

Which graph would you use when investigating:

1) Whether daily temperature and ice cream sales were related?

Scatter

2) Comparison of mean reaction time for a group having alcohol and a group drinking water **Boxplot or confidence interval plot**

Exercise: Ticket cost comparison Solution

a) Is there a big difference in average ticket price by group?

The mean and median are much bigger in those who survived.

b) Which group has data which is more spread out?

The standard deviation and interquartile range are much bigger for those who survived so that data is more spread out

c) Is the data skewed?

Yes. The medians are much smaller than the means and the plots show the data is positively skewed.

d) Is the mean or median a better summary measure?

The median as the data is skewed

