T.M.J Tutoring
Perseverance Drives Success

Data analysis

G. Tang

Types of data

- Categorical
  - ordinal (ranking / order), eg. coffee sizes, movie ranking
  - nominal (no order / random), eg. postcodes, gender
- Numerical
  - discrete (counted), eg. the number of siblings
  - continuous (measured), eg. height, weight

Note: don't be purely swayed by just seeing numbers / words → read the description of the data itself (usually written within brackets)

Characteristics to describe data

Categorical:

- Briefly describe context of question
- Mention if there is a clear modal category
- Always include values to verify the point you've made
- Evaluate the major contrasts within the data
- No need to mention all categories unless needed.

3. Spread

(A) Range

R = largest data value − smallest data value.

(B) Interquartile range (IQR)

$IQR = Q_3 - Q_1$

★ the IQR is generally a more reliable measure of spread as it is not affected by outliers

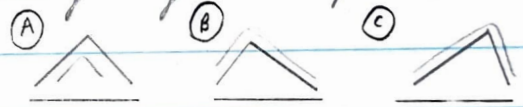
Numerical:

1. Shape

(A) Symmetry → mirrored image if folded.

(B) Positively skewed → peak towards left

(C) Negatively skewed → peak towards right:



2. Centre

(A) Mean → average

$\bar{x} = \frac{\text{add all values together}}{\text{total number of values}}$

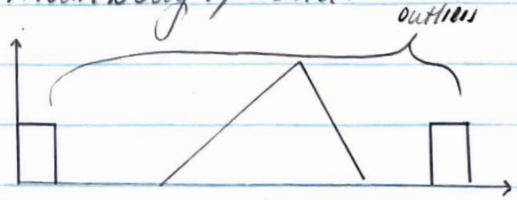
(B) Median → middle

$\frac{n+1}{2}$ → used to find the location of the middle number; NOT the median

★ The median is generally a more reliable measure of centre as it is not affected by outliers

4. Outliers

Any data values that stand out from the main body of data.



Displaying and describing the distributions of data.

Key:
- categorical
- numerical.

The frequency table

★ Percentage frequency = $\frac{\text{count}}{\text{total}} \times 100\%$.

Eg.

Lunch	Number	%
Sandwich	7	23·3
Salad	10	33·3
Pie	13	43·3
Total	30	99·9

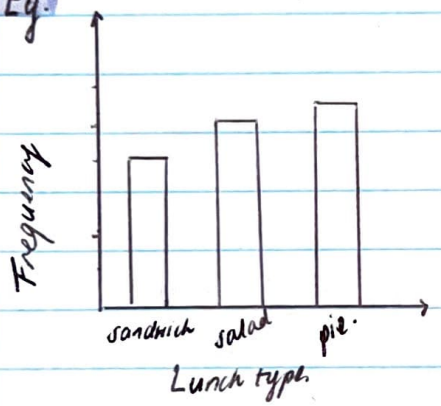
★ Always include **both** number and percentage.

Bar charts

★ x-axis: variable, y-axis: frequency

★ make sure there are gaps between categories

Eg.



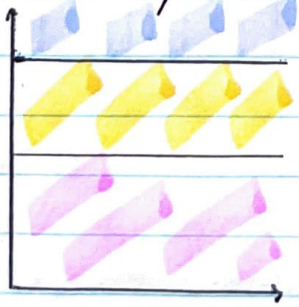
Segmented bar chart

★ multiple categories stacked upon each other (~4-5)

★ must include key / legend

★ must add up to 100%.

Eg.



Key:
- pie
- salad
- sandwich

G. Tang

The grouped frequency table

★ Groupings should be chosen according to the following principles:
- all are intervals
- intervals should not overlap
- no gaps between intervals

★ Used to take larger range of values, or when variables are continuous

Eg.

Average hours worked	Number	Percentage
30.0 - 34.9	1	4.3
35.0 - 39.9	6	26.1
40.0 - 44.9	8	34.8
45.0 - 49.9	5	21.7
50.0 - 54.9	3	13.0
Total	23	99.9

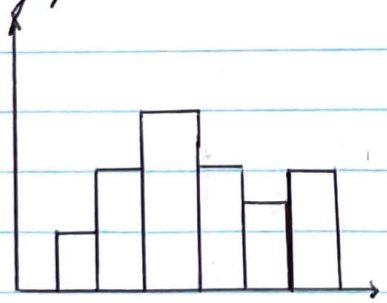
Histogram

★ Graphical display of intervals - each bar

★ x-axis: intervals, y-axis: frequency

★ no gaps

Eg.



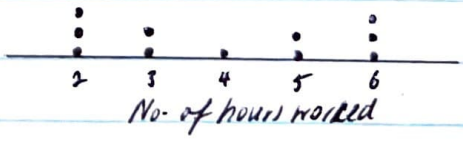
Dot plot

★ simplest display of discrete data

★ best for small data sets

★ dots must be placed evenly

Eg.



Stem and leaf plot

- ★ Works well for both discrete and continuous data
- ★ best for data sets up to 50
- ★ must include key
- ★ evenly spaced values required

Eg. Marks

Key: 1|5 = 15 marks

0	2
1	5 9 9 9
2	0 4 5 5 6 7 7 8
3	0 3 5 5 8 9

Using logarithmic scales to display data

★ this scale allows us to represent both small and large populations in a more compact way:

$10^0, 10^1, 10^2, 10^3, 10^4$ etc... ⇒ we always use the base of 10

↳ the logarithms are the numbers that are displayed as powers (the small numbers raised to base of 10)

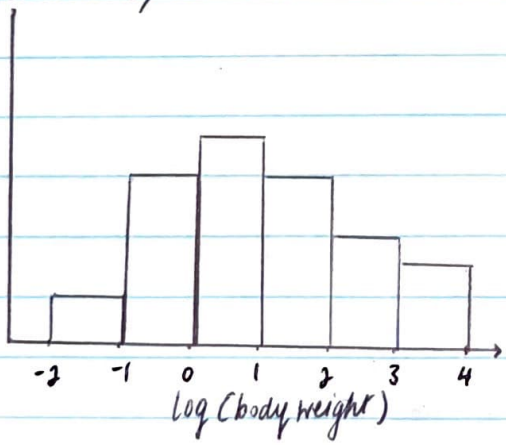
$\log_{10} x = b$, so $10^b = x$

Eg. $\log_{10}(100) = 2$, so $10^2 = 100$

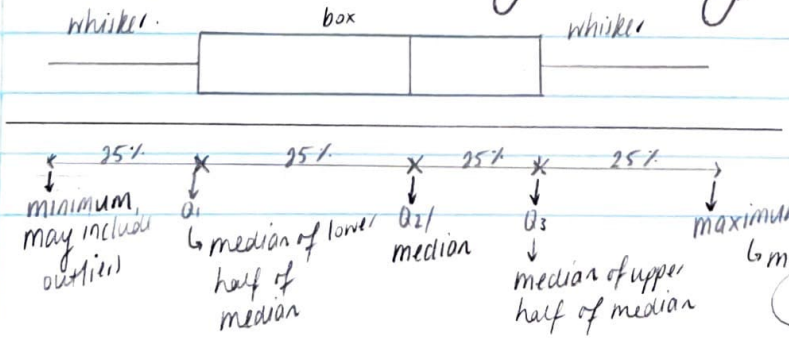
Log histograms

★ x-axis represents the powers to the base of 10. So to find the 'actual' value, put your powers to the base of 10 → $10^{x \leftarrow x\text{-axis value}}$.

Eg.



The boxplot and the five-figure summary



5-figure summary: Min, Q_1, median, Q_3, Maximum.