STATISTICS — Part 1

Data Collection & Analysis

① STATISTICS — KEY TERMS

Statistics: collecting, presenting, analysing data

Descriptive: graphs to show the data

Summary: represent data as a whole (avg)

Inferential: small group → conclusions about big group

Population: all of group under consideration

Census: everyone

Sample: part of population

$\bar{x}$ = sample mean μ = population mean

② TYPES OF DATA

Categorical: car colour, brand

• **Nominal:** no order (film types)

• **Ordinal:** order (exam grades)

Numerical: uses numbers

• **Discrete:** single value (place in race)

• **Continuous:** range (age groups)

Univariate: 1 variable **Bivariate:** 2 paired

Primary: collected by you **Secondary:** from source

③ SAMPLING METHODS

Simple Random: everyone equal chance — NO bias

Systematic: every 10th person — biased

Stratified: split into groups, sample by % of population

Cluster: split into clusters, sample whole clusters

Quota: split into groups, fixed quota per group

Convenience: ease of access — biased

Bias: non-representative, non-response, dishonest answers

④ SURVEYS & QUESTIONNAIRES

Postal: cheap, large reach. BUT poor response, limited data

Personal: many questions, high response. BUT expensive, interviewer bias

Observation: systematic. BUT laborious, time-consuming

Good questionnaires:

• Brief, clear, simple first

• Multiple choice answers, clear instructions

• NO leading or embarrassing questions

⑤ MEAN, MODE, MEDIAN

Mean: (sum) / (number of values)

Frequency table: Mean = $\Sigma(fx) / \Sigma f$

Grouped data: use mid-interval value

Mode: most common (modal class for grouped)

Median: middle when in order — pos = (n+1)/2

Comparison:

• Mode/Median: not affected by extremes

• Mean: distorted by extremes, best for analysis

⑥ WORDED MEAN PROBLEMS

Sum = mean × number of values

Missing number when given mean:

1. Sum of known + x = mean × n
2. Solve for x

Missing frequency when given mean:

1. $\Sigma(fx)$ = mean × Σf
2. Solve for missing f

✓ **TIP:** Always rearrange Mean = Sum/n to find what you need

⑦ PRESENTING DATA

Bar Chart: Categorical / Discrete

Histogram: Continuous (no gaps)

Stem & Leaf: leaf = ONE digit; KEY essential

Double-sided S&L: compare 2 datasets

Line Plot: dots instead of bars

Scatter: plot points, draw line of best fit

★ *Frequency on VERTICAL axis ; always LABEL both axes*

⑧ SCATTER & LINE OF BEST FIT

Scatter diagram: plots x-y pairs to spot relationships

Line of best fit: straight line through middle of data

Calculator method:

Mode → Setup 2 → y = a + bx → option 4

Gives equation: y = a + bx, plus correlation r

✓ **TIP:** Use line of best fit to PREDICT y for a given x

⑨ CORRELATION r & CAUSALITY

r = correlation coefficient ($-1 \leq r \leq 1$)

• r = 1: strong positive (both up)

• r = −1: strong negative (one up, one down)

• r = 0: no linear relationship

⚠ **WATCH:** r is NOT the slope of the line of best fit!

Causality ≠ Correlation:

• Smoking → cancer (causal)

• Hot day ↔ ice cream sales (no causality)

⑩ POPULATIONS & SAMPLES

Population: group you need information on

Census: counting everyone in population

Sample size: too small → unreliable; too big → costly

Sources of bias:

• Sample not representative

• Failure to respond

• Dishonest answers

★ *Random sample = no bias ; Stratified = best for diverse population*