Collection, Classification & Organisation of Data

Prepared by Nikolas Agisilaou

Qualitative vs Quantitative

- **Qualitative** (Categorical): Describes attributes — **not** numbers.
  e.g. nationalities, genres of music, types of transport.
- **Quantitative** (Numerical): Can be measured or counted.
  e.g. lengths, speeds, populations, test scores.

Discrete vs Continuous

Both are types of **quantitative** data.

- **Discrete** → **Count** it. Takes separate values.
  e.g. number of apps, goals scored, pages in a book.
- **Continuous** → **Measure** it. Any value in a range.
  e.g. mass, time, capacity, temperature.

Gap test: Can you have a value between 3 and 4?
3.6 cats? No ⇒ discrete. 3.6 kg? Yes ⇒ continuous.

Primary vs Secondary Data

- **Primary:** You collect it yourself (surveys, experiments, observations).
  + Tailored, current, you control quality.
  – Time-consuming, costly.
- **Secondary:** Already collected by someone else (reports, databases, websites).
  + Fast, cheap, large datasets.
  – May not fit your needs, could be outdated.

Data Collection Methods

- **Survey / Questionnaire:** Structured questions to a group.
- **Observation:** Watching and recording without interfering.
- **Experiment:** Change one variable, measure the effect, control the rest.

Key experiment terms:
Independent = what you change
Dependent = what you measure
Control = what stays the same

Designing Good Surveys: Common Pitfalls

Avoid these mistakes:

- **Leading / biased language:**
  "Don't you love the new library?"
- **Double-barrelled questions:**
  "How much do you spend on snacks *and* transport?"
- **Overlapping response categories:**
  3–6, 6–10 ← where does 6 go?

Fix them like this:

- Use **neutral** wording:
  "How would you rate the library? (Poor / Satisfactory / Good / Excellent)"
- Ask **one thing per question**.
- Use **non-overlapping** ranges:
  Less than 3, 3 to less than 6, 6 to less than 10, 10 or more.

The Panda Math Tutor

Tally Charts & Frequency Tables

Group every 5 marks: 𝍸 = 5

1. List all categories or values.
2. Tally each data item one by one.
3. Count tallies → frequency.
4. **Check:** $\sum$ frequencies = total data items.

Example: 𝍸 | | | = 8 𝍸 𝍸 | = 11

Grouped Frequency Tables

Use when data has many values or is continuous.

1. Range = Max − Min.
2. Pick a class width (5, 10, etc.).
3. Write intervals: $10 \le x < 20$, etc.
4. Tally data into each class.
5. $\sum$ frequencies = total.

No overlaps! $10 \le x < 20$, $20 \le x < 30$ ✓
No gaps! 10–19, 21–30 × (20 is missing!)

Class Boundaries & Midpoints

For the interval $25 \le x < 35$:

- **Lower boundary** = 25
- **Upper boundary** = 35
- **Class width** = 35 − 25 = 10
- **Midpoint** = $\dfrac{25+35}{2} = 30$

Used for: estimating the mean, plotting frequency polygons.

Two-Way Tables

Display two categorical variables at once.

1. Fill cells you can find from row/column totals.
2. Unknown = Total − Known.
3. Work rows → columns → totals.
4. **Check:** Row totals = Column totals = Grand total.

	Fict.	**Non-f.**	**Tot.**
Yr 10	19	15	34
Yr 11	15	11	26
Tot.	34	26	60

Sampling Methods & Key Formulas

Method	**How it works**
Random	Every member has an equal chance
Systematic	Every nth item from an ordered list
Stratified	Proportional sample from each subgroup
Cluster	Select whole groups at random

- **Systematic interval:** $n = \dfrac{\text{Population}}{\text{Sample size}}$
- **Stratified sample:** $\dfrac{\text{Stratum size}}{\text{Population}} \times$ Sample size
- **Class midpoint:** $\dfrac{\text{Lower} + \text{Upper}}{2}$

Quick Exam Shortcuts

Keywords: "Classify the data" → qualitative/quantitative, discrete/continuous. "Design a questionnaire" → check bias & overlapping options. "Complete the table" → row/column totals + subtraction. "What percentage…" → $\frac{\text{part}}{\text{total}} \times 100\%$.
Always check: $\sum$ frequencies = total data items. Two-way table: row totals = column totals = grand total.

The Panda Math Tutor