

WORKING WITH DATA

Lecturer: Mark Gillan

DATA CATEGORIES

- Data can take a variety of forms, from numbers stored in databases to voice recordings and images. Each of these contain information, so they are all data
- Most data falls into two main categories:

○ Qualitative

○ Quantitative



DATA CATEGORIES IN MORE DETAIL

Qualitative – Descriptive, normally words rather than numbers

- Difficult to analyse
- Limited visualisation options
- Words often captured against different categories or labels enabling quantification of occurrences in a category
- Summary statistics cannot be calculated
- The data can be subjective, containing opinions rather than facts
- Used to answer questions such as:
 - What?
 - Why?

Quantitative – Measures of values or counts and expressed as numbers

- Discrete –
Whole numbers, obtained by counting
 - Continuous –
All possible values, obtained by measuring
-
- Obtained by counting
 - Easy to analyse
 - Can be visually displayed using many different chart types
 - Summary statistics can be calculated
 - Used to answer questions such as:
 - How many?
 - How often?



SCALES OF MEASUREMENT

NOMINAL

- Named variables
- No explicit order
- Examples include:
 - Names
 - Colours
 - Gender

INTERVAL

- Named variables
- Ordered variables
- Differences can be measured
- No true zero
- Examples include:
 - Celsius temperatures
 - Time

ORDINAL

- Named variables
- Ordered variables
- Differences cannot be measured
- Examples include:
 - Low/medium/high
 - Likert-scale
 - Age bands

RATIO

- Named variables
- Ordered variables
- Differences can be measured
- Has a true zero starting point
- Examples include:
 - Height
 - Weight



STRUCTURE OF DATA

STRUCTURED

- Ordered in rows and columns
- Defined by a data model
- Can be easily queried

SEMI-STRUCTURED

- Self-describing structure
- Allows for flexibility
- Examples include HTML, JSON and XML

UNSTRUCTURED

- No pre-defined structure
- 80-90% of data is in this format
- Difficult to analyse

Tidy Data:

Datasets arranged such that each variable is a column and each observation a row.

