Topic: Exploratory Data Analysis (EDA)

Data Types

School of Mathematics and Applied Statistics

UNIVERSITY
OF WOLLONGONG
AUSTRALIA

Making sense of data

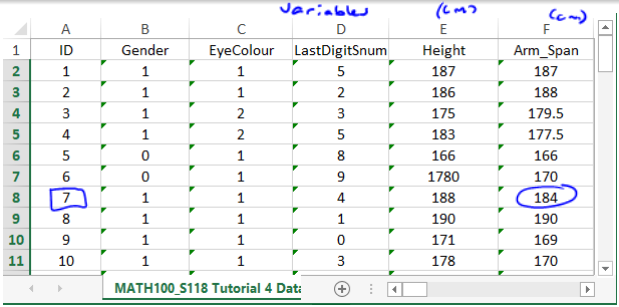
We need to know

- Context
- Units of measurement

How we analyse our data depends upon

- The structure of the data
- The type of measurement of our variables
- The questions being asked

Data set example



	A	B	C	D	E	F
1	ID	Gender	EyeColour	LastDigitSnum	Height	Arm_Span
2	1	1	1	5	187	187
3	2	1	1	2	186	188
4	3	1	2	3	175	179.5
5	4	1	2	5	183	177.5
6	5	0	1	8	166	166
7	6	0	1	9	1780	170
8	7	1	1	4	188	184
9	8	1	1	1	190	190
10	9	1	1	0	171	169
11	10	1	1	3	178	170

Basic Terms

- A **data set** is a collection of observations on one or more variables.
- An **observational unit** is the entity providing the information eg. student (row)
- A **variable** is a characteristic under study that assumes different values for different units eg Height (column)
- The item response for a unit is called an **observation** (cell)

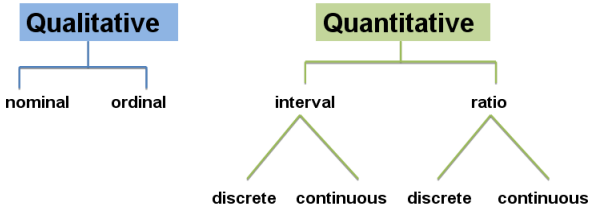
Types of Data

Qualitative Variable

A variable that can be classified into two or more categories is called a qualitative variable.

Quantitative Variable

A variable that can be measured numerically is called a quantitative variable.



Qualitative Data

Qualitative data is classified as **nominal** or **ordinal**.

Nominal (or Categorical) variable

A qualitative variable that can be classified into two or more categories which have no order.

eg. Own a bicycle: Y, N. Binary ⇒ 2

eg. Brand of mobile phone

eg. Country of birth

Ordinal variable

A qualitative variable that can be classified into two or more categories which have some order.

eg. Age-group 18-25, 26-40, 41-60

eg. ATAR band

eg. Qn with Likert scale: Strongly agree - Strongly disagree

Quantitative Data

Quantitative data can be classified on an **interval** or **ratio** scale.

Interval scale

An interval scale is one in which the same difference between two values means the same thing everywhere on the scale, but ratios of differences are not meaningful.

An interval scale may have a zero, but it is not a true zero with respect to the property it is measuring.

eg. Temperature (degrees Celsius)

°F.

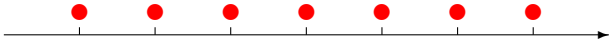
Ratio Scale

On a ratio scale something with twice the value has twice the property.

eg. Height: a child that is 1.4m tall is twice as tall as a child that is 70cm tall.

Discrete or Continuous

- If the possible values are separate points on the number line, a measurement is said to be **discrete**, *e.g.* no. of emails.



- The possible values of a **continuous** measurement form 1 or more intervals on the number line, *e.g.* length, weight.

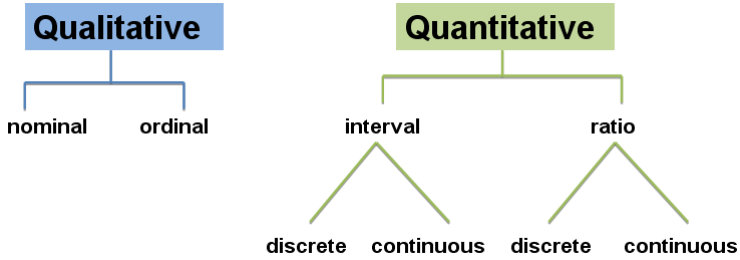
Note that length, weight are on a continuous scale even though it can only take positive values.

Activity

Classify the following variables

- **Handedness** Left / right. nominal ⇒ binary 2 categories.
- **Time** in mins spent on FaceBook on a particular day. quant. ratio / cont. ⇒ ratio
- Number of **texts** sent on a particular day. quant. ⇒ ratio. ⇒ d. ratio.
- **Coffee** cup size ordered: S/M/L. qual. ⇒ ordinal. 3 categories.
- **Coffee** type ordered: Capp/Latte/Flat white / ... etc qual. ⇒ nominal.

Types of Data - summary



Once we know the type of data we can choose the best way

- to summarise the data; and
- represent the data in tables and graphs.

i.e. How can we summarise and display the data effectively?