

Business and Economic Statistics

Chapter 1 - Data

Statistics: involves collecting, analysing and interpreting data

Population: refers to the entire pool from which a statistical sample is drawn

Sample: a small portion of data that is drawn from the studied population -> this allows a manageable sample where conclusions can be drawn about the larger population

Types of Data:

A **variable** is a characteristic of a population or of a sample from a population eg. Weight, Gender, Age

- We look at values or **observations** of a variable
- A **data set** contains observations on variables

Variables may be:

- **Discrete** (whole numbers or categories that are binary eg. Exam score, Football Score) or **continuous** (any value that can be taken from a continuum eg. Time remaining in a football game, temperature)
- **Quantitative** (numerical) or **qualitative** (categorical: nominal (no-order), ordinal)
 - Exams scores and time are quantitative variables whereas, gender is a nominal qualitative variable
- **Ordinal** qualitative data is when the values have an intrinsic order
Eg. Course evaluations: poor, good, very good
Eg. Age: infant, youth, teen, adult, senior
- **Identifier** variables are categorical variables whose only purpose is to assign a unique identifier code to each individual in the data set eg. Student ID no.

In order to apply statistical analyses directly to qualitative data, we must convert it somehow to quantitative data.

Types of Observations:

The type of observation made by the statistician can also be used to classify data:

- **Time series data** consists of measurements of the same variable at different points in time. Eg. Someone's HR in hospital over 12 hours in regular 1-hour intervals, Total Revenue of Woolworths every year for 10 years.
- **Cross sectional data** consists of measurements of several variables at a single point in time. Eg. Age, gender and marital status of a sample of UNSW staff in a particular year.

Chapter 2 – Graphical Descriptive Statistics

- **Descriptive statistics** involves organising, summarizing and presenting data in an informative way.
- There are many tools and techniques including graphical and numerical methods.
- The type of data you have partly dictates which techniques are sensible to use.
- The best data displays observe a fundamental principle of graphing known as the **area principle**, which says that the area occupied by a part of the graph should correspond to the magnitude of the value it represents. Otherwise, it can distort the reader's understanding.

Displaying Qualitative Data

1) Frequency Distribution Tables

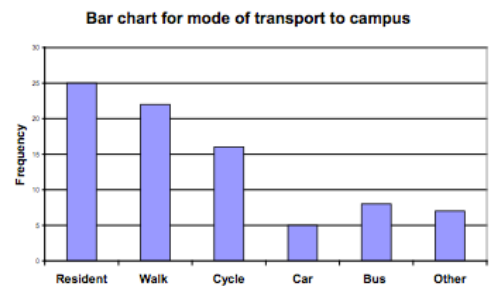
- **Summarises categorical data** using counts.
- Categories must be **mutually exclusive** (every observation belongs to only one category) and **exhaustive** (all possibilities are recorded even if no-one is in that category)

Example: UNSW is interested in how many students get to campus, for long term planning.

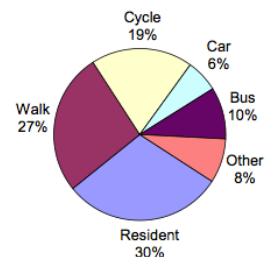
Mode of transport to Campus	Frequency – how often the event occurs	Relative Frequency – lists the categories and the proportion in which they occur of the total
Resident	25	$25/83 \times 100\%$
Walk	22	$22/83 \times 100\%$
Cycle	16	$16/83 \times 100\%$
Car	5	$5/83 \times 100\%$
Bus	8	$8/83 \times 100\%$
Other	7	$7/83 \times 100\%$
Total	83	100

2) Bar Charts

- Graphical representation displaying the distribution of a categorical variable, showing the counts for each category next to each other for easy comparison.
- LEAVE SPACES between categories
- Must label axis
- Alternatively, the counts could be replaced with percentages – relative frequency bar chart



Pie chart of mode of transport to campus



3) Pie Charts

- Graphical representation of data presenting the relative frequency (proportion of the total)
- Typically used for qualitative data.

Displaying Quantitative Data

1) Histograms

- Obvious categories for the values of a quantitative variable may not exist
- But we **can create** categories or classes by defining the lower and upper class limits. These categories must be **mutually exclusive** and **exhaustive**.
- These categories are known as **bins**.
- Bins don't need to be of equal width and may be open-ended at the top or bottom eg. If age is the x-axis the last bin may be '70+'
- There should be no gaps between bars unless it is because there is no data in one bin
- An alternative is to report the percentage of cases in each bin, creating a **relative frequency histogram**.

