# BUSS1020: Quantitative Business Analysis

## Week 1

Statistics – methods that collect, describe, and transform data into useful insights for decision makers.

Statistical Analysis Framework (DCOVA):

- **D**efine the problem/objective and data required.
- **C**ollect the required data.
- **O**rganise the data.
- **V**isualise the data.
- **A**nalyse the data.

Branches of Statistics:

- Descriptive = collecting, summarizing, presenting and organizing data.
- Predictive = using a model and data to make forecasts.
- Inferential = using data from a small group to draw conclusions about a larger group.
    - Estimation.
    - Hypothesis testing.

| Term | Definition |
|---|---|
| Variable | Characteristics of an item/individual that can change. |
| Data | Collection of observed outcomes of one or more variables. |
| Population | Consists of all items and individuals that you are interested in. |
| Sample | Part of the population. |
| Parameter | Numerical measure that describes a relevant characteristic of a population. |
| Statistic | Numerical measure that describes a characteristic of a sample. |

### Types of Variables

- **Categorical** – variables that have values that can only be placed into categories.
- **Numerical** – variables that have values that represent actual number quantities.
    - Discrete = counting process (integers).
    - Continuous = measuring process (can have decimals).

### Levels of Data Measurement

- **Nominal** data – no ranking of data e.g. employment classification.
- **Ordinal** data – ranking of data e.g. grade A is better than grade B.
- **Interval** data – numerical data but with no scale (no true zero) e.g. 0 degrees does not mean that there is no temperature.
- **Ratio** data – numerical data with scale (has a true zero) e.g. $0 means you have no money.

### Sampling

- Sampling is less time-consuming and less costly than selecting every item in the population.

**Non-probability sampling**

- Items are chosen without regard to their probability of occurrence.
- Convenience sample – selected based on being easy, inexpensive and quick to sample.
- Judgment sample – perceived experts/most appropriate items are selected, by convenience.
- Self-selection – individuals choose to participate.
- Quota sample – pre-set quotas of groups are chosen, by convenience.

**Probability sampling**

- Items chosen randomly which have regard to their probability of occurrence.
- Simple random (SRS) – every individual/item in the frame has equal chance of being selected.
- Systematic – divide your sample into n groups and pick the $k^{th}$ person from each group.

- Stratified – divide data into important characteristics and select your sample e.g. pick 10 people from Y7,8,9,10,11,12.
- Cluster – population is divided into several clusters, each representative of the population.

Comparing sampling methods

- SRS and Systematic sampling – simple, cheap, effective against bias BUT may not give the best representation of the population's underlying characteristics.
- Stratified sampling – ensures representation of individuals across entire population, effective against bias, most efficient BUT costly.
- Cluster sampling – cost effective BUT less efficient (need large samples to provide significant insight).

## Sources of Data

- Primary sources – analyst collects the data.
- Secondary sources – analyst is not the data collector.
- Data from a designed experiment:
  - Consumer testing of different versions of a product.
  - Quality testing.
  - Market testing.
- Survey data.
- Data from observational studies.
- Automated and streaming data.

## Data Format

- Structured vs unstructured (messy, not stored easily, stored in several different locations etc.).
- Unstructured data needs; data linking, preparation, cleaning etc. before organization or analysis.
- Cleaning data = removing errors, flagging outliers, fill-in/delete missing data.

## Types of Survey Errors

- **Coverage error/selection bias:** exists if some groups are excluded from the frame and have no/little chance of being selected.
- **Non-response error:** people who choose not to respond may be different from those who do respond.
- **Sampling error:** variation from sample to sample; always exists.
- **Measurement error:** due to weakness in question design, respondent error and interviewer's effects on the respondent.

# Week 2

*Data is organized and visualized so as to reveal and communicate the information, especially the main features and patterns, that are hidden within it.*

## Categorical Data

**Organising One Variable Categorical Data**

- Summary table: indicates the frequency, amount, percentage or proportion of items in each of a set of categories. Allows you to see;
  - The relative frequency of each category.
  - The differences between the categories.

A Survey of 1000 Bank Customers:

| Banking Preference? | Percent |
|---|---|
| ATM | 16% |
| Telephone | 2% |
| Drive-through service at branch | 17% |
| In person at branch | 41% |
| Internet | 24% |

**Visualising One Variable Categorical Data**

- Bar chart: one bar for each category; length represents amount/frequency/percentage of values in that category.
  - Each bar is separate.
  - Order does not matter.
- Pie chart: a shaded circle with one slice for each category. Slice size represents percentage.
- Pareto chart: for categorical, nominal scale data.
  - Vertical bar chart, categories shown in descending order of frequency.

The Pareto Chart