BUSS1020 Notes

Week 1

Introduction to statistics

➢ How businesses collect and use data

- **D**efine the problem or objective and the data required (design)
- **C**ollect the required data in an appropriate manner
- **O**rganise the data → prepare it for analysis
- **V**isualise the data
- **A**nalyse the data
- **DCOVA**

➢ Some ways they use statistics and analytics

- Statistics → methods that collect, describe, transform data into useful insights for decision makers.
  - Descriptive statistics → collecting, summarising, presenting & organising data
  - Inferential statistics → using data collected from a small group to draw conclusions about a larger group
  - Predictive statistics → using a model and data to make forecasts of outcomes.
- Used to:
  - Summarise business data
  - Draw conclusions from business data
  - Make forecasts about business activities
  - Improve business processes.

➢ The basic vocabulary of statistics and of analytics

- **Variables** are characteristics of an item or individual → also known as attributes
- **Data** are the observed values or outcomes of one or more variables.
- **Operational definition** → variables should have universally accepted meanings that are clear to all associated with an analysis.
- **Population** consists of all the items or individuals about which you want to draw a conclusion. → large group
- **Sample** is the portion of a population selected for analysis → small group
- **Parameter** is a numerical measure that describes a relevant characteristic of a population.
- **Statistic** is a numerical measure that describes a characteristic of a sample.

➢ The types of data used in business

- Types of variables:
  - **Categorical** (qualitative) variables have values that can only be placed into categories, e.g yes or no
  - **Numerical** (quantitative) variables have values that represent actual number quantities → eg counted or measured.
    - **Discrete** variables arise from a counting process

        - **Continuous** variables arise from a measuring process: can be assigned any value within a given interval.
- Levels of data measurement:
    - **Nominal** (lowest level) → categorical variables
        - Classify or categorize eg occupation. → labels used to distinguish different categories that have no order
    - **Ordinal** → categorical variables
        - Labels are used to classify AND to indicate rank or order. Eg how helpful was this: 1,2,3,4 etc.
    - **Interval** → numerical variables
        - Data are numerical and differences between values have a consistent meaning. Eg Celsius temperature; calendar time. → but 0 doesn't imply nothing.
    - **Ratio** (highest level) → numerical variables
        - Same properties as interval data, with the addition that zero has a true meaning and represents the absence of the phenomenon being measured. Eg measurements such as height, weight & volume.

➢ The sources of data used in business

- **Primary sources** → analyst collects the data
    - Data from a survey; data collected from an experiment; directly observed data
- **Secondary sources** → analyst not the data collector
    - Analysing census data; consultant analysing company database; examining data published on internet; analysing data collected by stock markets
- Sources of data:
    - Data distributed by organisatons or individuals: financial data; market data
    - Data from a designed experiment: results from tests of different product versions; quality testing; market testing
    - Survey data: political polls; internet polls
    - Data from observational studies: measuring time it takes customers to be served; measuring volume of traffic through intersection
    - Automated and streaming data: website visit data; GPS data

➢ The principles of effective data sampling

- Why sample?
    - Collecting information from a sample is less time-consuming & less costly than selecting every item in the population (census).
- Sampling frame → list of items in popular that CAN be sampled.
    - Inaccurate or biased results can result if parts of population are excluded.
- Types of samples:
    - **Non-probability samples** → items chosen without knowing their probability of selection. → cannot be used for statistical inferences
        - **Judgement:** perceived experts or most appropriate items are selected
        - **Convenience:** selected based on being easy, inexpensive, quick
        - **Self selection**: individuals choose to participate
    - **Probability samples** → select items based on known probability
        - **Simple random**

  - Every individual or item in the frame has equal chance of being selected. → easiest sample to select
  - Disadvantages: results are often subject to more variation; when the frame is very large, it is time consuming & expensive;
  - Number every item in the frame from 1 to N. The chance any item will be selected on the first selection is 1/N → if it is sampling without replacement, the second time it is 1/(N-1).
- **Systematic**
  - Divide population into groups, & from first group pick random number and take sample from this number in every group.
    - Advantages: simplicity that allows a degree of system or process into the random selection; Assures population will be evenly sampled
    - Disadvantages: prone to selection bias when there is a pattern in the frame; requires large sample size.
    - Split N items in the frame into n groups of k items:
      - K = N/n
- **Stratified**
  - Divide frame into strata according to an important characteristic. → eg dividing into males & females to get right proportion of each in survey.
  - Simple random within each strata, and combine the results.
  - Ensures representation of individuals across the entire population, possibly in the right proportions → effective against bias. → more efficient
  - Requires you can determine the variable on which to base the stratification → thus expensive to implement.
- **Cluster**
  - Population divided into clusters, each representative of the population. → each cluster is supposed to be representative of the entire population → random sample within 1 or more clusters.
  - Cost effective, but less efficient. → requires large sample size to product precise results.

- Types of survey errors:
  - **Coverage error** → if some groups are excluded from the frame & have no chance of being selected. → results in selection bias
  - **Non-response error** → people who choose not to respond may be different from those who do respond
  - **Sampling error** → variation from sample to sample will always exist
  - **Measurement error** → weaknesses in question design, respondent error and interviewer's effect on respondent