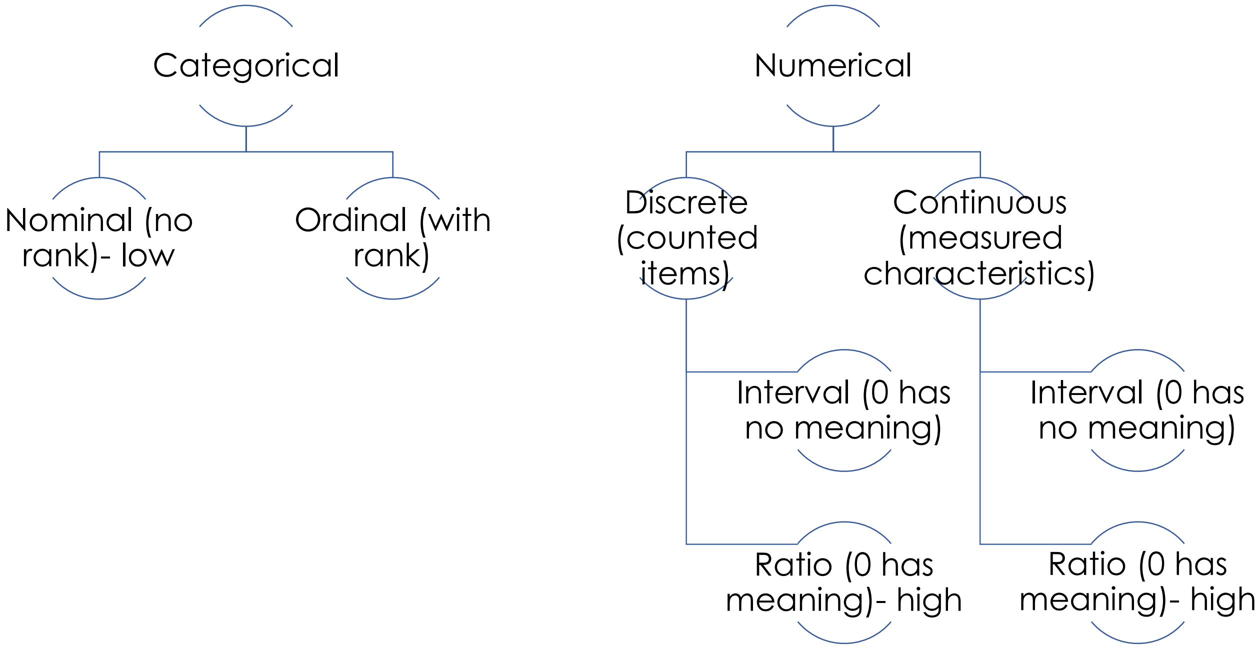
BUSS1020 Notes

Data Types



Categorical

Nominal (no rank)- low

Ordinal (with rank)

Numerical

Discrete (counted items)

Interval (0 has no meaning)

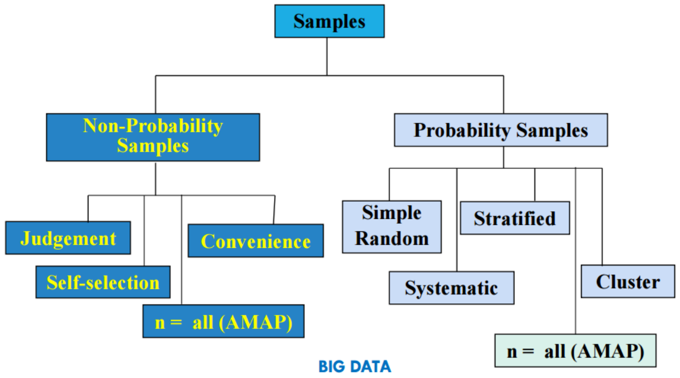
Ratio (0 has meaning)- high

Continuous (measured characteristics)

Interval (0 has no meaning)

Ratio (0 has meaning)- high

Sampling Data



- Non-probability sampling:
  - **Convenience sample:** selection easy, inexpensive, quick (e.g. 'snowball' sampling)
  - **Judgement sample:** 'experts' select most appropriate sample
  - **Self-selected sample:** individuals choose to participate
  - **Quota sample:** use pre-set quotas of groups chosen
- Probability sampling:
  - **Simple Random**: Every individual or item in the frame has equal chance of being selected.
  - **Systematic**: Divide your sample into **n** groups (equal size) and pick the **kth** person from each group. E.g. every 3rd person in each group here
  - **Stratified**: Divide data into important characteristics and select your sample. E.g. pick 10 people from each BUSS1020 tutorial class.
  - **Cluster:** Population is divided into several "**clusters**", each representative of the population. E.g. pick 3 BUSS1020 tutorials of all the tutorials
- Sampling Errors:
  - **Selection bias**: Exists if some groups are excluded from the frame and have no chance (or little chance) of being selected.
  - **Non-response error or bias**: 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**: Due to weaknesses in question design, respondent error and interviewer's effects on the respondent.

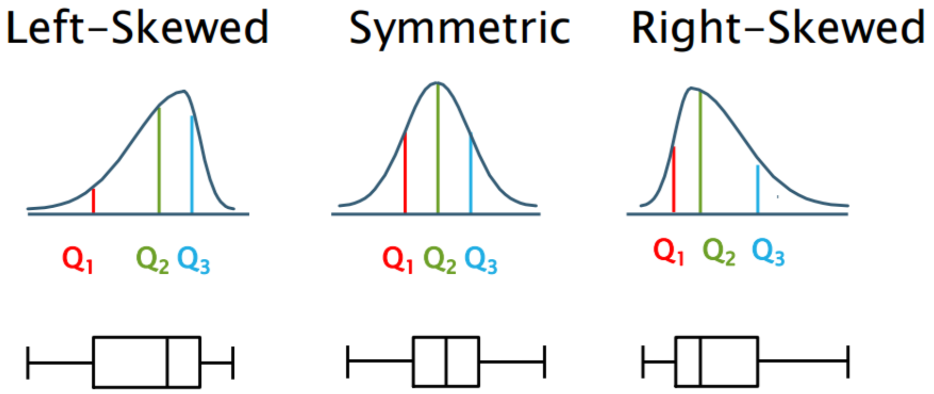
Organising and Visualising Data

Variable type	Organising	Visualising
Categorical (1 variable)	Summary Table (frequency and/or percentage)	Bar charts Pie charts Pareto charts
Categorical (2 variables)	Contingency Table	Side-by-side bar chart
Numerical (1 variable)	Ordered Array Frequency Distributions Cumulative Distributions	Histogram Polygon Ogive
Numerical (2 variables)	Same as above	Scatter plot Time series plot

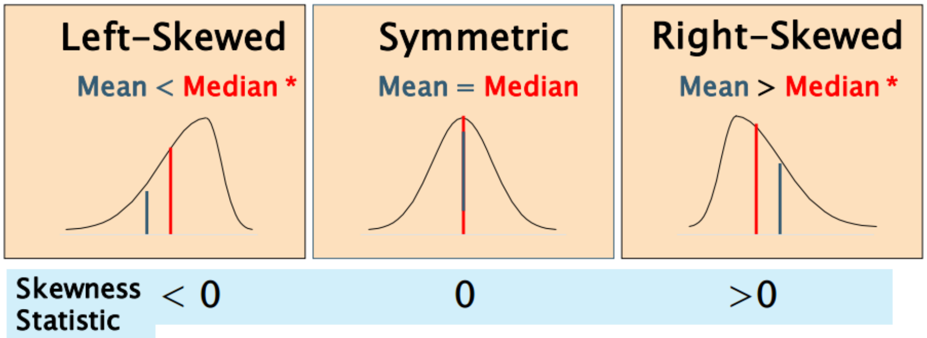
Numerical Descriptive Measures

- **Central tendency**: extent to which the data values group around a central value.
- **Variation**: amount of dispersion around the central value.
- **Shape**: pattern of distribution from lowest to highest value.

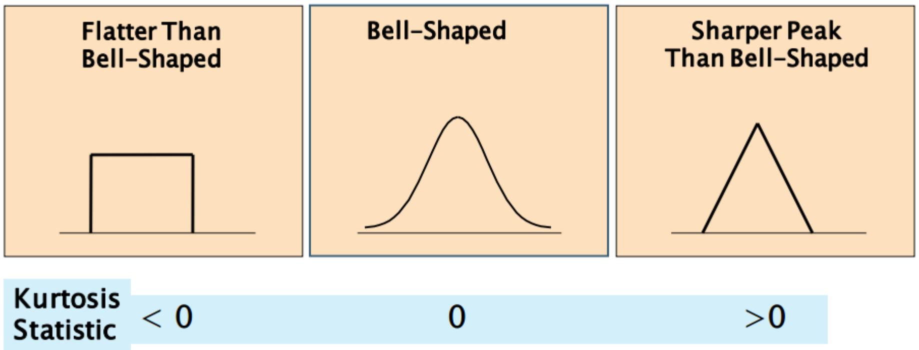
- Measures of Central tendency
  - **Mean**: the average value of the observation.
  - **Median**: middle value in the ordered array.
  - **Mode**: Most frequently observed value
  - **Geometric mean**: Rate of change of a variable, over time.
    - $$\bar{X}_G = (X_1 \times X_2 \times \cdots \times X_n)^{1/n}$$
    - Rate of return
    - $$\bar{R}_G = [(1+R_1)\times(1+R_2)\times\cdots\times(1+R_n)]^{1/n} - 1$$
- Common measures of variation
  - **Range:** difference between largest and smallest value
  - **Sample Variance:** avg. of squared deviations of values from mean
  - **Sample Standard Deviation:** square root of variance
  - **Interquartile Range:** measures spread in middle 50% of data
  - **Coefficient of Variation:** measures relative variation compared to the mean
  - **Z score:** calculate how many standard deviations a value is from the sample mean
  - **The five-number summary:**
    - Minimum
    - First quartile
    - Median
    - Third quartile
    - Maximum



- Distribution Shape
  - **Skewness**
    - This describes the amount of **asymmetry** in a distribution



- **Kurtosis**
  - Describes relative **concentration** of values in the center as **compared to the tails**



Measure	Population	Sample
Mean	μ	$\overline{X}$
Variance	σ^2	S^2
Standard Deviation	σ	S

- Empirical Rule
  - The empirical rule describes that in the bell shape distribution, **approximately**
  - **68%** of data is within **one** standard deviation from the mean;
  - **95%** of data is within **two** standard deviation from the mean;
  - **99.7%** of data is within **three** standard deviation from the mean;
- Chebyshev's Rule
  - At least **(1-1/k²) * 100%** of the values will fall within **k** standard deviations of the mean **(k>1)**