Quantitative Business Analysis

General information on Statistics

Analytics: discovery and communication of meaningful patterns in data. It involves statistics, computer programming and operations research.

Statistics: branch of maths that makes numbers (**data**) informative.

- Descriptive statistics are the methods that help collect, summarise, present and analyse data.
- Inferential statistics are methods that use the data collected from a small group to draw conclusions about a larger group.
- 4 important uses of statistics
  - Visualise and summarize data (e.g. tables, charts. Descriptive method)
  - Reach conclusions about a large group based on data collected from small group (Inferential).
  - Make reliable predictions based on statistical models (Inferential)
  - Improve business processes.

Statistical Process

1. Define: set clearly defined goals for the investigation and formulate the research question/hypothesis.
2. Collect: decide what data is appropriate and how to collect them. Collect the data.
3. Organise: display, describe and summarise the data. E.g. tables. Check for any usual features.
4. Extract information: Choose and apply appropriate statistical methods to extract useful information.
5. Conclusion: Interpret information, draw conclusions and communicate results to others.

- Or DCOVA for Define, collect, organise, visualise and analyse.

Sources, sampling and variables overview

After defining goal and variables, data must be collected. Data can be a **primary or secondary source.**

- Primary source: the analyst is the data collector.
  - Organisations and individuals that **collect and publish data** use their data as a primary source (e.g. ABS) and let others use their data as a secondary source.
  - E.g. data from a **political survey**, **experiment** and **directly observed data**.
- Secondary source: data for analysis has been collected by someone else.
  - Company database, internet information, textbook, data made on stock markets.

- Population: all the items or **individuals (elements)** which you want to reach conclusions.
- Sample: portion of a population selected for analysis. The process of selecting the sample is **sampling.**
  - If sample is to be informative of total population, sampling must be done carefully, impartially and objectively.
  - **Random sampling**: selecting individuals in a totally random fashion to prevent bias. Can be used to draw conclusions about larger population.
- Characteristics of an individual/element is a **variable** and this affects results. Variables can be:
  - Categorical (qualitative): variables can be placed into categories like yes/no or true/false.
    - Nominal scale: classifies data into categories without ranking. E.g. type of email provider.
    - Ordinal scale: classifies values into distinct categories with ranking. E.g. rating service.
  - Numerical (quantitative): variables have values that represent actual quantities.
    - Discrete: values that arise from **counting process** (integer values). E.g. No. of TV channels.
    - Continuous: arise from a **measuring process** and can be assigned any value within a given interval. Can have decimal point values E.g. room temperature/height of child.
    - Numerical variables are measured on an interval or ratio scale.

- Ratio: ordered scale in which the difference between the measurements involves a true 0 point.
- Interval: ordered scale in which the difference between the measurements is a meaningful quantity but does NOT involve a true 0 point.

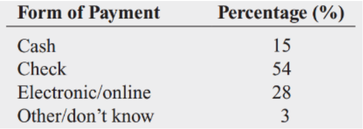
Organising and visualising categorical data (Tables and Charts)

Data may be organised in different structures depending on whether:

- You are summarising variables or looking at relationship between variables.
- The variables are categorical or numerical.
- Ask yourself "Is there enough data points to merit a graph?" If not, use a table. E.g. a graph is unnecessary when people present a few numbers in a bar chart.

Summary Table

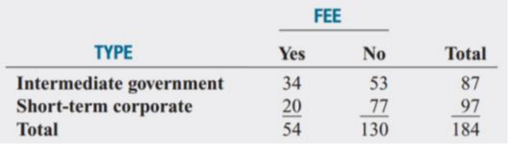
- Presents tallies of frequencies or percentages for each category.
- Helps to see the differences among categories.



Form of Payment	Percentage (%)
Cash	15
Check	54
Electronic/online	28
Other/don't know	3

Contingency Table

- Allows you to study patterns between the responses of 2 or more categorical variables.
- Cross tabulated form that tallies the responses of the categorical variables together.



	FEE		
TYPE	Yes	No	Total
Intermediate government	34	53	87
Short-term corporate	20	77	97
Total	54	130	184

Creating charts for **visualizing data** enhances the discovery of patterns and relationships. It also depends on type of variable and number of variable.

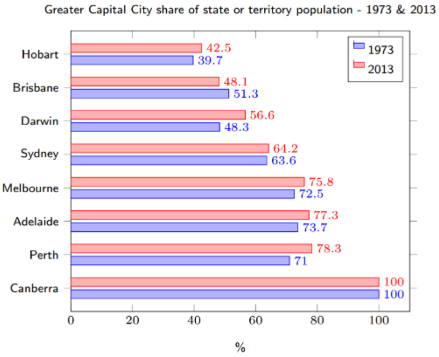
	Numerical	Categorical
One variable	Histogram, box-plot	Bar chart
Two variables	Scatter plot, time series plot, percentage polygon	Side-by-side bar chart.

Bar Chart

- Compares different categories by using **individual** bars to represent tallies.
- Unlike histogram, bar chart separates bars between each category.
- Can be horizontal or vertical.

Side by Side Bar Chart

- Two or more bars to represent tallies.
- Shows joint responses from 2 categorical variables.
- 3D bar charts are confusing - hard to determine exact values.


Greater Capital City share of state or territory population - 1973 & 2013

Pie Chart

- Uses parts of a circle to represent tallies. Lets you visualise the portion of the entire pie in each category.
- Very few instances where pie chart is more informative than a simple bar chart.

Organising Numerical Data (Graphs)

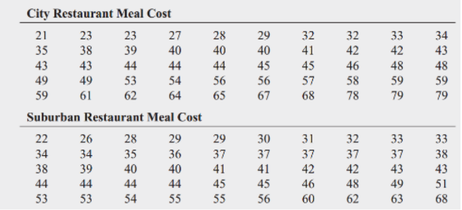
You organize numerical data by creating ordered arrays or distributions. Method you choose depends on amount of data that you have and what you see to discover about your variables.

Ordered Array

- Arranges values of numerical variable in rank order, from smallest to largest.
- Helps get a better sense of range and any outliers.
- If data set contains a large number of values, ordered array is difficult.

TABLE 2.7B

Ordered Arrays of Cost per Person at 50 City Restaurants and 50 Suburban Restaurants



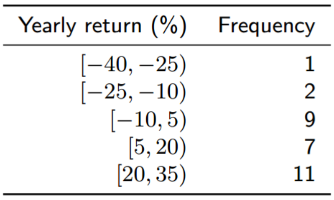
City Restaurant Meal Cost									
21	23	23	27	28	29	32	32	33	34
35	38	39	40	40	40	41	42	42	43
43	43	44	44	44	45	45	46	48	48
49	49	53	54	56	56	57	58	59	59
59	61	62	64	65	67	68	78	79	79

Suburban Restaurant Meal Cost									
22	26	28	29	29	30	31	32	33	33
34	34	35	36	37	37	37	37	37	38
38	39	40	40	41	41	42	42	43	43
44	44	44	44	45	45	46	48	49	51
53	53	54	55	55	56	60	62	63	68

Frequency Distribution

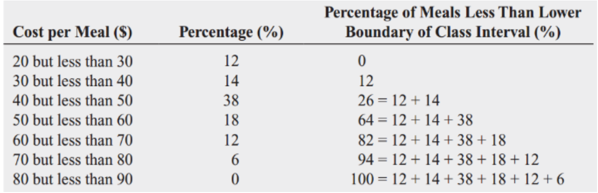
- Summary table where data is assigned to numerically ordered categories, called **classes**.

- Frequency distribution procedure:
  1. Sort the observations (i.e. in an ordered array)
  2. Determine number of **classes** (5-15 is ideal) (empirical rule: ≈ √n where n is number of values).
  3. Determine **class interval width** (range of values in each class). Formula is: $\text{interval width} \approx \dfrac{\text{maximum} - \text{minimum}}{\text{number of classes}}$
  4. Determine **class boundaries.** E.g. [-40,-25) → between -40 and -25 not including 25.
  5. Count the number of observations in each class.



Yearly return (%)	Frequency
$[-40, -25)$	1
$[-25, -10)$	2
$[-10, 5)$	9
$[5, 20)$	7
$[20, 35)$	11

- Useful for numerically continuous variables.
- Assesses where observations lie and whether the observations are clustered or spread out.
- Tips:
  - Different class boundaries may provide different pictures (especially in small sample sizes)
  - Shifts in data concentration may show up when different class boundaries are chosen.
  - As sample size increases, impact of alterations in class boundaries is reduced.
  - When comparing 2 or more groups with different sample sizes, you must use either a relative frequency or percentage distribution. $\text{Proportion} = \text{relative frequency} = \dfrac{\text{number of values in each class}}{\text{total number of values}}$
- **Cumulative percentage distribution**: presents information about percentage of values that are less than a specific amount. This includes all the classes that are below the specific amount.



Cost per Meal ($)	Percentage (%)	Percentage of Meals Less Than Lower Boundary of Class Interval (%)
20 but less than 30	12	0
30 but less than 40	14	12
40 but less than 50	38	26 = 12 + 14
50 but less than 60	18	64 = 12 + 14 + 38
60 but less than 70	12	82 = 12 + 14 + 38 + 18
70 but less than 80	6	94 = 12 + 14 + 38 + 18 + 12
80 but less than 90	0	100 = 12 + 14 + 38 + 18 + 12 + 6

There are multiple ways to visualise numerical data.

- Stem and leaf display: shows how data are distributed and where concentration of data exists. Organises data into groups (the stems) so that values within each group (leaves) branch out to left.

Histogram

- A bar chart for grouped numerical data where **vertical** bars have no spaces between one other and each bar represents the frequencies or percentages.
- Variable of interest (independent variable) on X axis and frequency (dependent variable) on the Y axis.
- When commenting on histograms, consider the following:
  - **Location** (where most of data are) and **spread** (variability of data).
  - **Symmetric** or **skewed** to the left or right.
  - Unimodal (1 peak) or bimodal/multimodal.
  - **Bell-shaped** (if symmetric and unimodal).
  - **Range** of the outlier and any **outliers.**

Figure: Histograms with different spread

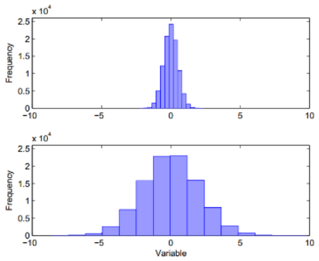
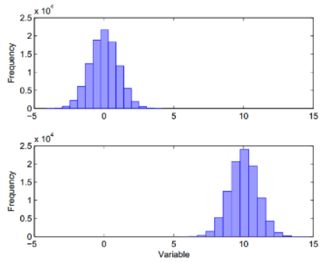


Figure: Histograms with different locations



- When looking at 2 or more variables, **percentage polygon** joins the midpoints of each class interval in a histogram (line graph). Easier to interpret than multiple histograms on the same graph as there would be overlaps between the bars.
  - **Cumulative percentage polygon (Ogive)**: graphs cumulative percentage distribution.

Scatter Plot

- Plots one numerical variable against another to show relationships between the variables.
- To make a scatter plot, we organise data into a collection of ordered pairs {(-3,1), (28,26), (18,15)…}.
- When commenting on scatter plot, focus on the nature of relationship:
  - Existent vs non-existent.
  - Strong vs weak