BUSS1020 Theory and Formulas Notes

Chapter 1: Introduction

Branches of analytics

1. Descriptive: collecting, summarising, presenting, analysing
   - e.g. survey, tables, sample mean
2. Inferential: Data from small group to draw conclusions about larger groups
   - e.g. estimation, hypothesis testing
3. Predictive: Model and data to make forecast and outcomes
   - e.g. statistical model

Types of variables

- Categorical (qualitative): yes/no - defined categories
- Numerical (quantitative): represent actual quantities
  - Discrete: counting item (number of kids) - 5
  - Continuous: Measuring characteristic (financial return) - 5.398

Levels of data measurement

- Nominal
  - Labels are used to distinguish different categories (e.g. employment classification - teacher, doctor...)
- Ordinal
  - Labels to classify AND indicate rank with underlying scale but levels not comparable (e.g. computer tutorial was helpful or not etc)
- Interval
  - Numerical data and different between values have consistent meaning (location of 0 is matter of convenience e.g. celsius temp)
- Ratio
  - Same as interval AND 0 has true meaning - significance, represents absence of the phenomenon measured (e.g. measurement, price)

Chapter 2: Organising and visualising data

DCOVA

Define variables

Collecting data

- Primary Sources: Analyst collects data from political survey, from experiment
- Secondary: Analysing census data, consultant analysing company database
  - Distributed by organisation/individual
  - Designed experiment
  - Survey
  - Observational study

Organising data

Categorical data

- 1 categorical table = summary table / 2+ = contingency table
- Pivot table (visual): Automatically sort, count total or give the average of the data stored in one table or spreadsheet.

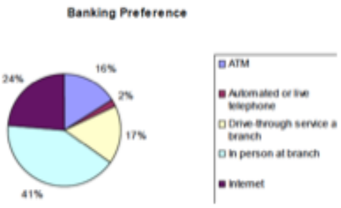
Numerical data

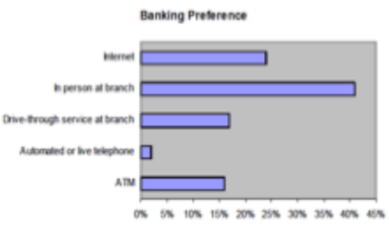
- Ordered array
  - Sequence rank from smallest to largest –range and outliers
- Frequency distribution
  - Summary table – numerically ordered classes to see characteristics
    - Classes: boundaries and 5-15 classes (determined by number of values)
    - Class interval: range/class groupings (ascending, range, select class number, interval, boundaries, midpoints)
    - Relative frequency (% adds up to 1)
- Cumulative distribution
  - Frequency and percentage adding up to n

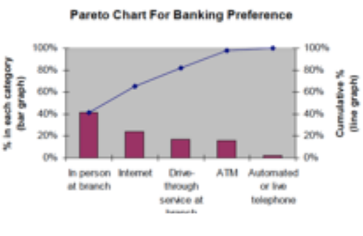
Visualising data

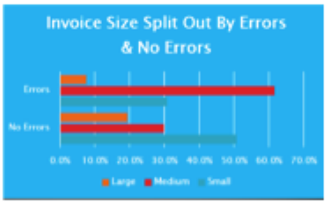
Categorical data

- 1 variable
  - Bar chart: amount, frequency, percentage of value
  - Pie: broken to slices rep categories
  - Pareto chart: vertical bar in descending order & cumulative polygon (separate vital few from trivial many)
  - Side-by-side: data from contingency table



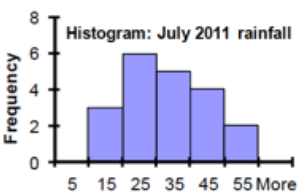


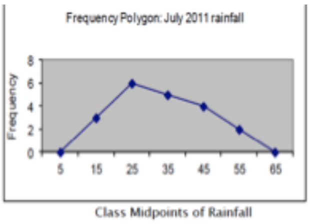


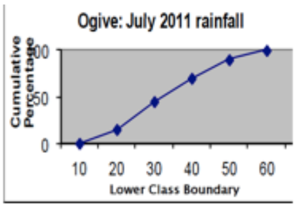


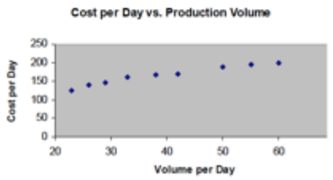
Numerical data

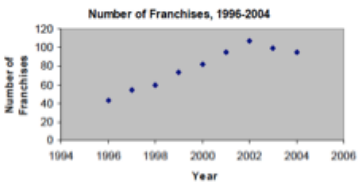
- 1 variable
    - Histogram: organise data into groups (bins)
        - frequency distribution
        - No gaps - continuos data
        - X-axis: class boundaries / Y: Frequency/relative frequency, percentage
    - Polygon: midpoint of each class represents data in that class, connecting by midpoints
        - At class midpoints (between 2 values)
    - Ogive: X-axis - variable / Y: cumulative percentage (cumulative percentage polygon)
        - At class points (not between)
- 2+ variables
    - Scatter Plot
        - Paired observations - one on each axis
        - Examine relationships
    - Time-series plot:
        - Study patterns in values
        - X: Time period / Y: numeric variable
    - Multidimensional data - to discover patterns, summary, contingency tables

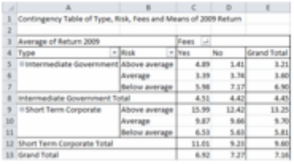












	A	B	C	D	E
1	Contingency Table of Type, Risk, Fees and Means of 2009 Return				
2					
3	Average of Return 2009		Fees		
4	Type	Risk	Yes	No	Grand Total
5	Intermediate Government	Above average	4.89	1.41	3.21
6		Average	3.39	3.76	3.60
7		Below average	5.98	7.17	6.90
8	Intermediate Government Total		4.51	4.42	4.45
9	Short Term Corporate	Above average	15.99	12.42	13.25
10		Average	9.87	9.66	9.70
11		Below average	6.53	5.63	5.81
12	Short Term Corporate Total		11.01	9.23	9.60
13	Grand Total		6.92	7.27	7.16

Principles of graphs

- Not distorted, char junk, must begin at 0, label, title, simple, source data, covey the message

Chapter 3: Numerical descriptive measures

Central Tendency

Arithmetic Mean

Median

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

- Middle number (not affected by outliers)

$$\frac{n+1}{2}$$

Mode

- Value occurs most often (not affected by outliers)

Geometric mean

- Rate of change of variable over time

$$\bar{X}_G = (X_1 \times X_2 \times \cdots \times X_n)^{1/n}$$

Geometric mean rate of return

- Measures status of investment over time

$$\bar{R}_G = [(1+R_1) \times (1+R_2) \times \cdots \times (1+R_n)]^{1/n} - 1$$

E.g.

$$X_1 = \$100{,}000 \quad X_2 = \$50{,}000 \quad X_3 = \$100{,}000$$

$$\underbrace{X_1 \quad X_2}_{\text{50\% decrease}} \quad \underbrace{X_2 \quad X_3}_{\text{100\% increase}}$$

$$\bar{R}_G = [(1+R_1)\times(1+R_2)\times\cdots\times(1+R_n)]^{1/n} - 1$$
$$= [(1+(-.5))\times(1+(1))]^{1/2} - 1$$
$$= [(.50)\times(2)]^{1/2} - 1 = 1^{1/2} - 1 = 0\%$$

Variation

Range

- Sensitive to outliers
- $X_{largest}$ - $X_{smallest}$