Table of Contents

INTRODUCTION TO BUSINESS ANALYTICS

Key Definitions

- **Statistics-** A branch of mathematics concerned with the collection and variation of data (collection, analysis, interpretation & presentation)
- **Variables-** Characteristics or attributes that can be expected to differ from one individual to another EG: Gender
- **Entity-** label
- **Data-** The observed values of Variables
- **Population-** consists of all the members of a group about which you want to draw a conclusion
  Two factors need to be specified when defining a population:
  1. The **entity** (e.g. People or motor vehicles)
  2. The **boundary** (e.g. Registered to vote in NZ or registered in Victoria for road use)
- **Sample-** A sample is the proportion of the population
- **Census-** data collected on the whole population (rare)
- **Parameter-** Is a numerical measure that describes a characteristic of a population (Greek letters)
- **Statistic-** a numerical measure that describes a characteristic of a sample (Roman/English letters)
- **Descriptive Statistics-** Focus on collecting, summarizing & presenting a set of data to draw conclusions about a population (graph)
- **Inferential Statistics-** uses sample data to draw conclusions about a population
- **Observational-** no attempts made to control EG: Survey
- **Random sampling** is the best way to collect data.
- Data collected is not bias or ambiguous
- **Primary Data-** collected first hand
- **Secondary Data-** already available (someone else got it)
- **Time Series Data-** collected over time
- **Cross-Sectional Data-** collected at one fixed point in time
- **Error-** error made within probability

Types of Variables

- **Categorical-** worded answers EG: male or female, day of the week
  - **Nominal-** distinct groups, no ranking EG: favourite food, political party, type of fuel used (WEAK)
  - **Ordinal-** distinct groups, ranked EG: S, M, L- clothes, satisfaction- very satisfied, satisfied (STRONG)
- **Numerical**- numerical responses EG height, weight, times seen
  - **Discrete**- whole numerical responses that arise from a counting process EG: 1,2,3
  - **Continuous-** any numerical responses by measuring process. EG: height, weight, time, length
  - **Interval-** fixed term measurement, no true zero, intervals are equal EG exam score, Celsius, shoe size (WEAK)
  - **Ratio Scale-** meaningful value, zero must be included EG: length, weight, age, salary (STRONG)

DATA VISUALISATION

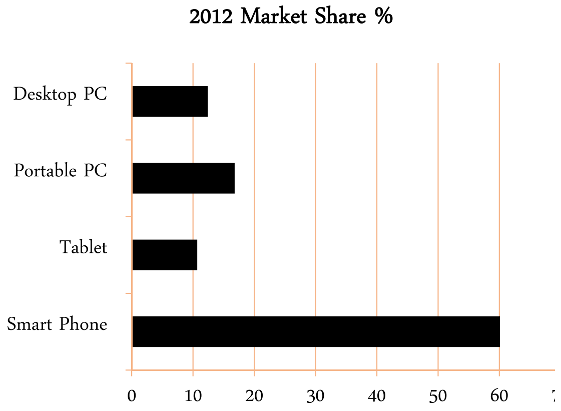
Tables and Charts for Categorical Data

Summary Table

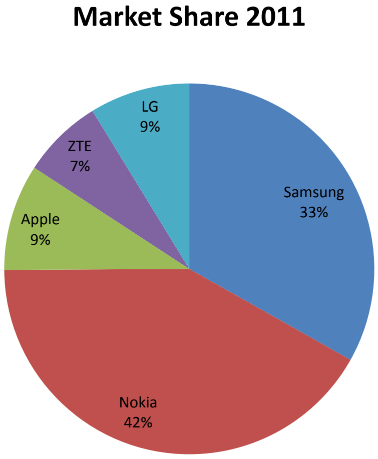
Gives the frequency, proportion or percentage of the data in each category

Type of device	2012 Shipments (in millions)	2012 Market Share
Smart Phone	722.4	60.1%
Tablet	128.3	10.7%
Portable PC	202	16.8%
Desktop PC	148.4	12.4%
Total	**1201.1**	**100%**

Bar Charts



Pie Chart



Stem and Leaf Plots

Helpful to order large amounts of data

Stem unit: $	leaf unit: 10 cents
4	8 3 99 3
5	4 6 8 5
6	1 4 6 8 9

Tables and Charts for Numerical Data

Frequency Distributions

- Allow you to condense a set of data.
- Summary table for numerical data
- Select an appropriate number of classes and suitable **class width**
- Example: Class width = 49 / 10 = 4.9
- Construct the frequency distribution table by first establishing clearly defined **class boundaries** (upper and lower values used to define classes for numerical data)
- The center of each class is called the **class mid-point**

Relative Frequency Distributions and Percentage Distributions

- Instead of the frequency, knowing the percentage of each of the data may be more useful
- A **relative frequency distribution** is obtained by dividing the frequency in each class by the total number of values. (EG: 3/52)
- From this a **percentage distribution** can be obtained by multiplying each relative frequency by 100%. (EG: 3/52x100)

Cumulative Distributions

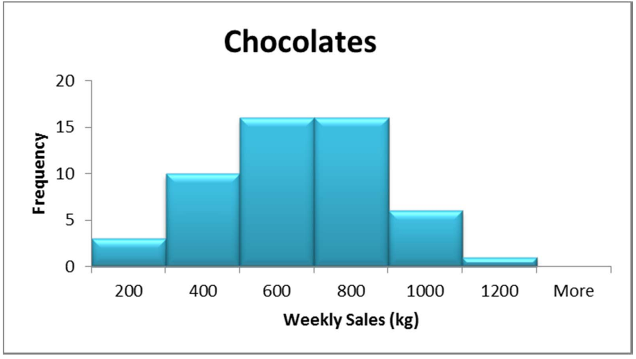
A **cumulative percentage distribution** gives the percentage of values that are less than a certain value. Percentage smallest to largest, just add as you go down.

Weekly Sales	**Count**	**Percentage**	**Cum. Percentage**
0 kg < 200 kg	3	5.8%	5.8%
200 kg to < 400 kg	10	19.2%	25%
400 kg < 600 kg	16	30.8%	55.8%
600 kg < 800 kg	16	30.8%	86.6%
800 kg < 1000 kg	6	11.5%	98.1%
1000 kg < 1200 kg	1	1.9%	100%
Total		**100%**	

Histogram

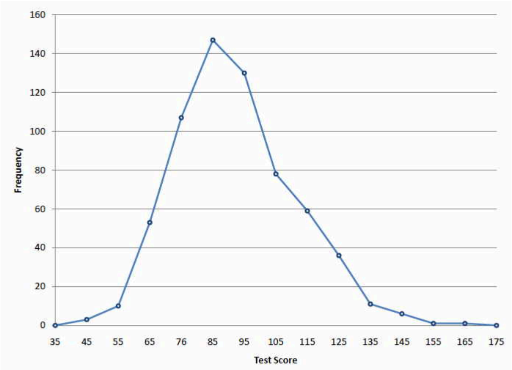
A grouped frequency, relative frequency or percentage distribution can be graphically represented by a **histogram.**

- **Ogive-** place dot on midpoint of class on histogram & connect lines

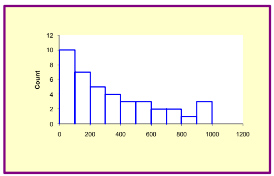


Polygons

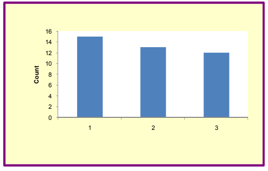
- When comparing two or more sets of data we can construct polygons on the same set of axes.
- Percentage Polygon- plotting % for each class above the midpoint & join lines



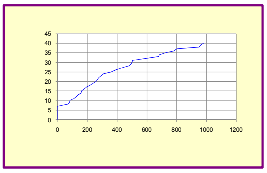
Exercise – 2 Type of graph



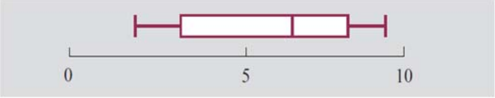
Histogram – Continuous Numerical data. Good for overview of distribution of data



Column Chart - Discrete/Categorical data. Good for overview of distribution of data



Line Chart – Time series data. Good for identifying trends/patterns over time



Box plot – Numerical data. Good for a quick overview of key features of data.