

## Chapter 2 – Text and Tests 5

### Statistics 1

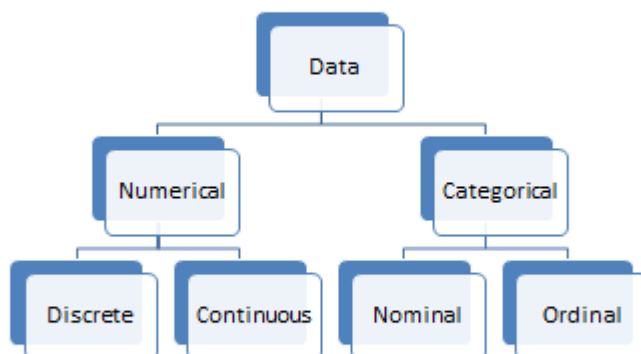
- (i) Collecting Data
- (ii) Presenting Data
- (iii) Analysing Data

**Descriptive statistics** – Graphs

**Summary statistics** – Averages to represent all data

**Inferential statistics** - Using a sample to draw conclusions about whole group

### Types of Data



**Primary Data – Collected yourself**

- Questionnaire
- Experiment
- Investigation
- Observations

**Secondary Data – Collected from existing statistics**

- Internet
- Data bases
- Newspapers, magazines
- Journals

Univariate Data - One item of data, age

Bivariate Data – 2 items of data, hrs of exercise and heart beat.

**Section 2.2 - COLLECTING DATA**

**1. Surveys**

- Post
- Personal
- Phone
- Observation

Advantages and Disadvantages of each

**2. Questionnaire**

- Clear
- Short
- Start simple
- Boxes and choice of responses
- No leading questions
- No personal questions

**Bias = Unfair  
Pilot Survey**

**3. Control group** Eg. Medicines  
Inactive substance = the control group

**4. Designed Experiments**  
Carry out a trial and record results

Explanatory variable = controlled  
Response variable = the effect observed

### **Section 2.3 Populations and sampling**

**Population** = Everyone in a particular group  
eg, all 17 yr olds in Ireland

**Census** = Everyone is surveyed

**Sample Survey** = a selection of the population to represent the whole population  
NB Size of the sample for reliability

**Bias** in a sample may arise from

- not representative
- not spread across the population
- poor response rate
- dishonest answers

## Sampling Methods

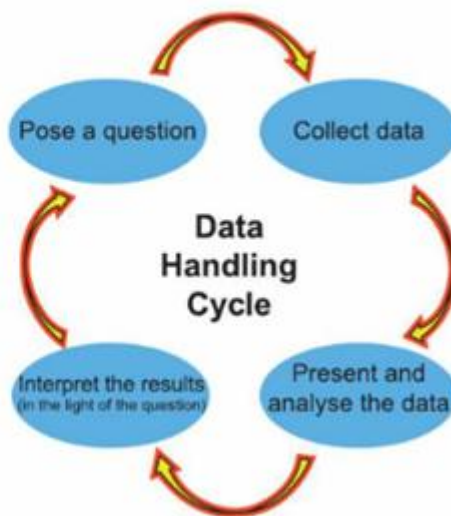
1. Simple random sampling
  - numbers picked from a hat
  - generate random numbers with calculator.  
(small population)
2. Stratified Sampling
  - population split into groups, (male/female)
  - representatives from each group of sample.
3. Systematic Sampling
  - select a random number from 1 – 10, say 7.
  - now select the 7<sup>th</sup>, 17<sup>th</sup>, 27<sup>th</sup> etc item/person
  
4. Quota Sample
  - Opinion polls
  - Interviewer told to survey n people from each group
5. Cluster Sampling
  - population split into groups/clusters (counties in Ire)
  - Number of clusters randomly selected and every item in the cluster is examined.
6. Convenience Sampling
  - Quick and easy
  - can lead to high levels of Bias.

- Ethical Issues
  - Informed consent
  - Confidentiality
- Reliability of data
  - Large enough
  - Random Sample (all an equal chance of selection)
  - High response rate

### Steps in a Statistical Investigation

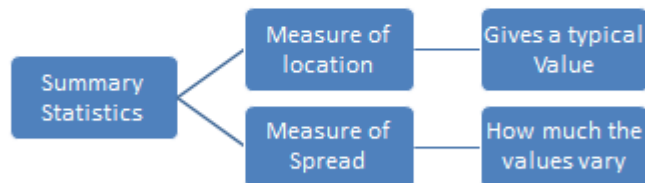
All statistical investigations begin with a question.  
Here are the steps in a statistical investigation:

- Pose a question.
- Collect data.
- Present the data.
- Analyse the data.
- Interpret the data.



## Section 2.4 - Measure of Location

### Lots of Data



### Measure of Location:

1. Mode
2. Mean
3. Median

1. Mode = most common value

2. Mean from a list:  $\bar{x} = \frac{\text{Sum of values}}{\text{total number of values}}$

$$\bar{x} = \frac{\sum fx}{\sum f}$$

Mean from a Table (frequency distribution) =

$$\frac{(\text{Top } x \text{ Bottom}) + (T \times B) + \dots \dots \dots}{\text{Bottom added}}$$

For grouped frequency use mid-intervals

### 3. Median

Write out Data in order, the median = the middle one  
If there are two numbers, the median = half way between them.

If there are  $n$  numbers in a list, the median is the  $\frac{1}{2}(n + 1)^{\text{th}}$  number.  
Eg. 11 numbers in a list—then  $\frac{1}{2}(11 + 1) = 6$   
the median is the 6<sup>th</sup> number.  
10 numbers in a list—then  $\frac{1}{2}(10 + 1) = 5.5$   
the median is between the 5<sup>th</sup> and 6<sup>th</sup> numbers.

The table below, which compares the advantages and disadvantages of each type of average, should help you make the correct decision.

Average	Advantages	Disadvantages
<b>Mode</b>	<ul style="list-style-type: none"><li>› Easy to find</li><li>› Not influenced by extreme values</li></ul>	<ul style="list-style-type: none"><li>› May not exist</li><li>› Not very useful for further analysis</li></ul>
<b>Median</b>	<ul style="list-style-type: none"><li>› Unaffected by extremes</li><li>› Easy to calculate if data is ordered</li></ul>	<ul style="list-style-type: none"><li>› Not always a given data value</li><li>› Not very useful for further analysis</li></ul>
<b>Mean</b>	<ul style="list-style-type: none"><li>› Uses all the data</li><li>› Easy to calculate</li><li>› Very useful for further analysis</li></ul>	<ul style="list-style-type: none"><li>› Distorted by extreme results</li><li>› Mean is not always a given data value</li></ul>

**Deciding which average to use:**

Mode: Used for categorical Data, shoe size, colour.  
Mode give a typical value for categorical data.

Mean  $\bar{x}$ : used to find typical value when data is closely grouped.  
Not useful when data is spread widely or  
if outlier(s) exist.

Median : used when data is well spread out or contains an  
outlier.

Outlier = very high or very low value not typical of the others