5 Week Modular Course in Statistics & Probability
Strand 1

Project ©

Maths

Tionscadal Mata

Development Team

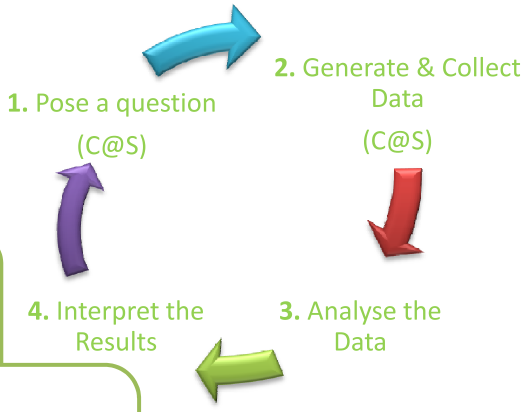
Module 1

Statistics

(a) Primary sources:
(i) Observational studies (JCHL, LCOL)
(ii) Designed experiments (JC)
(b) Secondary sources

Sampling: (i) Random (JC)
(ii) Stratified (LCHL)
(iii) Cluster (LCHL)
(iv) Quota (LCHL)

(a) Reliability of Data (JCHL)
(b) Summarise Data (Spreadsheets)
(c) Types of Data JC



Types of data:
Categorical/Numerical(JC)
(a) Univariate Categorical (JC)
Pie Charts (JC)
Bar Charts (JC)
Line Plots (JC)
Univariate Numeric
Histograms (JC)
Stem and Leaf(JC)
Back to Back (JCHL)
Line plots (JC)
(b) Bivariate (LC)
Bivariate Numeric
Scatter plots (LCOL)
Correlation (LCOL)

(a) Central Tendency
Mean (JCHL)
Median (JC)
Mode (JC)
(b) Spread
Range (JCOL)
Interquartile (JCHL)
Standard Deviation (Calculator)
(c) Histograms
Symmetry (LCOL)
Skewness (LCOL)
(d) Line of best fit (LCHL)
Correlation Coefficient
Meaning of (LCOL)
Calculate (LCHL)

Misuses and Misconceptions

Census at School (C@S)

Statistical Reasoning With an Aim to Becoming a Statistically Aware Consumer

Students learn about:

- The use of statistics to gather information from a selection of the population with the intention of making generalisations about the whole population
- They consider situations where statistics are misused and learn to evaluate the reliability and quality of data and data sources.

NCCA
National Council for Curriculum and Assessment
An Chomhairle Náisiúnta Curaclaim agus Measúnachta

(Syllabus)

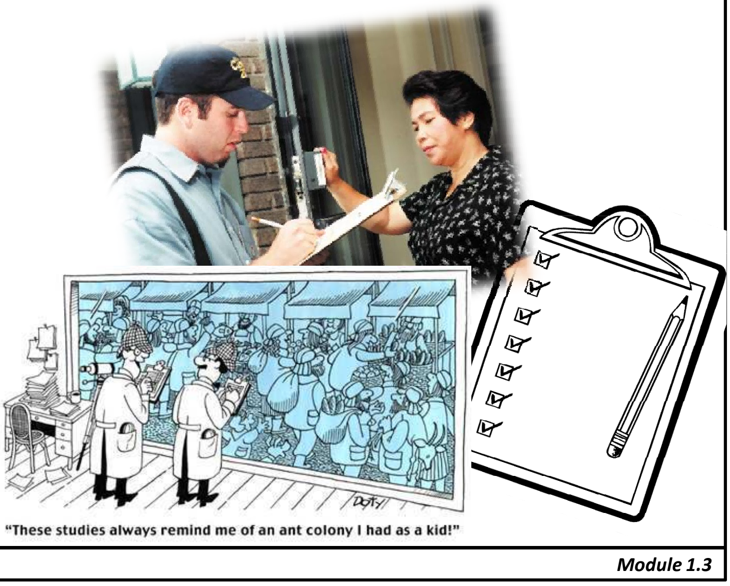
Producing Data

Primary Data

- Students collect the data themselves
- Observational studies:
  the researcher collects information but does not influence events e.g. surveys, epidemiological studies
- Experimental Studies:
  the researcher deliberately influences events and investigates the effects of the intervention, e.g. clinical trials, laboratory studies

Secondary Data

- Data collected by someone other than the user, i.e. the data already exists in books, journals, the internet etc.



"These studies always remind me of an ant colony I had as a kid!"

An Example of an Observational Study

Step 1: **Pose the question**
How accurate are students at estimating, to within 5 seconds, the number of seconds in a minute?

Step 2: **Collect the data**
Working in pairs (students A and B) and using a stop watch, students estimate the number of seconds in a minute. Student A signals when he/she is starting the stop watch and student B says "stop" when he/she thinks a minute has elapsed. Student B records the number of seconds estimated for a minute. The students then switch roles so that this time student A estimates and B operates the stop clock.

Step 3: **Analyse the data**
Estimated times from all the groups are recorded and a stem and leaf plot is produced for the whole class.

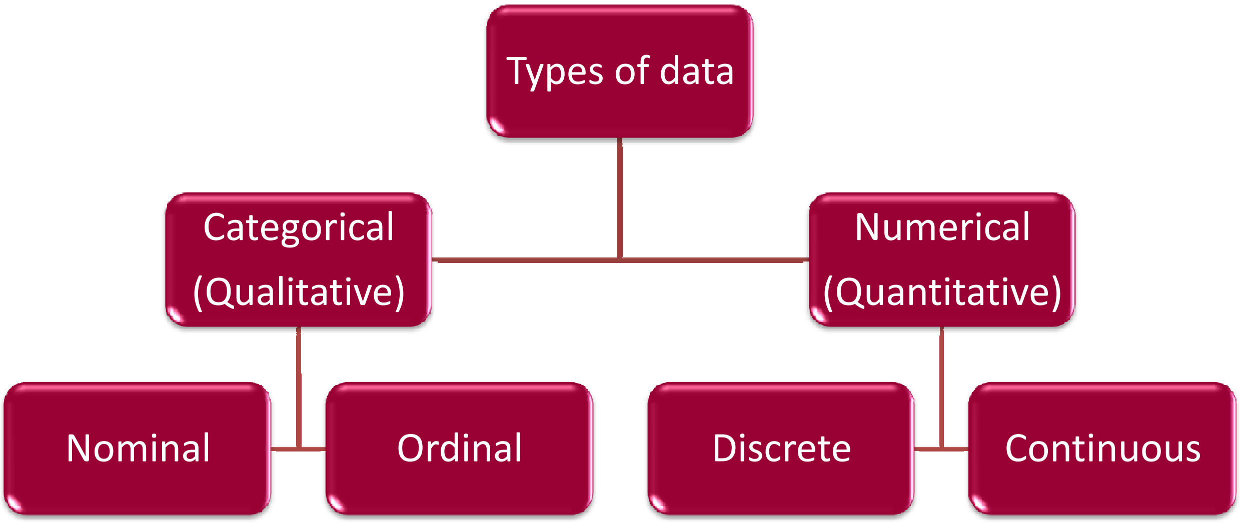
Step 4: **Interpret the result**
Are there any/many outliers? Find the mode and median from the stem plot and calculate the mean.
Are these values close to 60 seconds?
Answer the original question.

Extension Question:
Do students get better at estimating the number of seconds in a minute with practice?
What happens to the number of outliers which each successive trial?
What number are most of the data points clustered around etc.?

How Reliable is Secondary Data?

Who carried out the survey?
What was the population?
How was the sample selected?
How large was the sample?
What was the response rate?
How were the subjects contacted?
When was the survey conducted?
What were the exact questions asked?

Data Types


Types of data

Categorical (Qualitative)

Numerical (Quantitative)

Nominal

Ordinal

Discrete

Continuous

Univariate Data

(1) Involves a **single variable i.e.** we look at one item of data at a time from each subject e.g. height

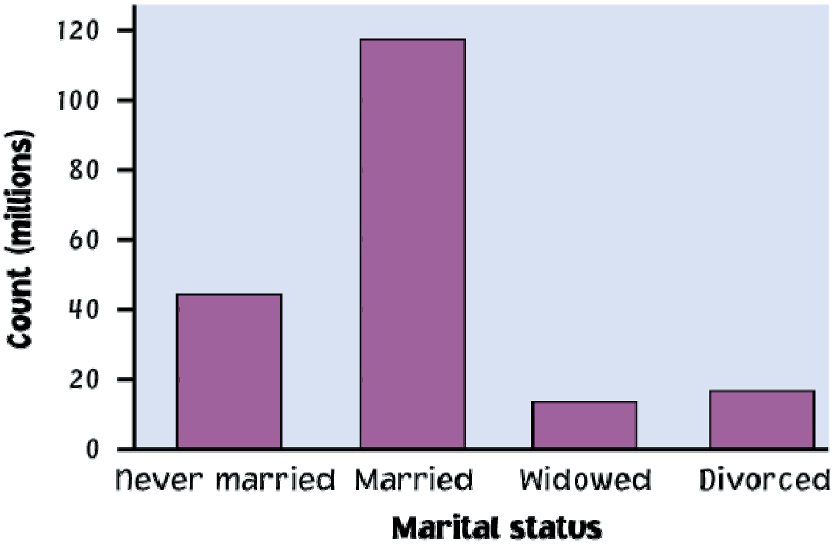
(2) Not dealing with relationships between variables

(3) The major purpose of Univariate analysis is to describe

Sample question:

How many of the students in the class are female?

A Bar Chart



Note: A bar chart describes categorical data, and has gaps, whereas a histogram describes continuous data and hence has no gaps.

Categorical

Type	Nominal	Ordinal
Description	Can be identified by particular names or categories and cannot be organised according to any natural order	Data which looks like numbers but are really just labels, they can be identified by categories which can be ordered in some way
Examples	**Gender:** Male or female **Hair Colour:** black, blonde etc., **Favourite Sport:** Soccer, rugby	ISBN Numbers, Visa card no., **Watching TV**: Never, Rarely, Sometimes, A lot
Suitable Graphical Representation	Bar chart, Line plot, Pie chart	Bar chart, Line plot, Pie chart

Numerical

Type	Discrete	Continuous
Description	Data can only have a finite number of values	Data can assume an infinite number of values between any two given values. Students height may be 1.4325 m
Examples	No. of peas in a pod, age in years (as opposed to age)	Height, arm span, foot length
Suitable Graphical Representation	Bar chart, Pie chart, Line plot, Stem plot	Histogram

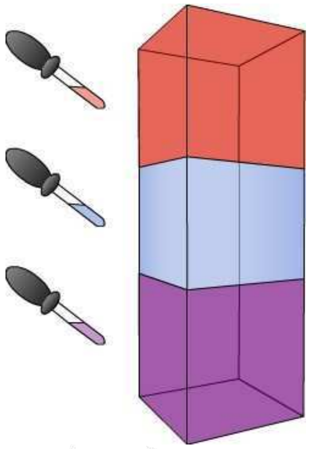
Let's Define Some Terms

A subset of the population.

- **Population:** The entire group of subjects about which information is required
- **Sample:** Any subset of a population e.g. a representative subset of students from the school
- **Variable:** We measure its value for each person and it varies from person to person e.g. the height of an individual or their favourite sport
- **Parameter:** Some value we are interested in calculating for the population
- **Statistic:** Some value we are interested in calculating for the sample

Types of Sampling

- Simple Random Sampling
- Stratified Random Sampling
- Cluster Sampling
- Quota Sampling



Simple Random Sample

One way of collecting data is to use a Sample. Whenever you need to take a sample the sample will need to be a Random Sample which is Representative of the population.

Example:
A new business with 100 employees wants to know whether staff would like to have childcare facilities on site.

An estimate could be made by asking a sample of 20 employees if they would use the childcare facilities and multiplying the number who say yes by 5.

If we do this, we have to decide which people to ask.

Biased Samples

When we are taking samples it is very important to avoid Bias.

Suppose we want to estimate how many students watch the X-Factor in a school with 1000 students.

Suppose we take a random sample of 50 and ask if they watch the X-Factor

. . . and all in the sample happen to be girls.

(Very unlikely but possible)

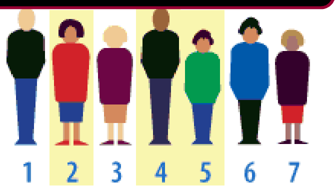
If girls are more, or less, likely to watch the X-Factor than boys we would have a biased sample.

Our results could be Unreliable

So, we need to avoid bias.

Random Sample

Random does NOT mean that we can just pick anyone for the sample.



Assign Numbers, Auto-Generate Random Selections

To get a Random Sample of 20 people we could give each person a number from 1 to 100 and then select 20 numbers using a random method.

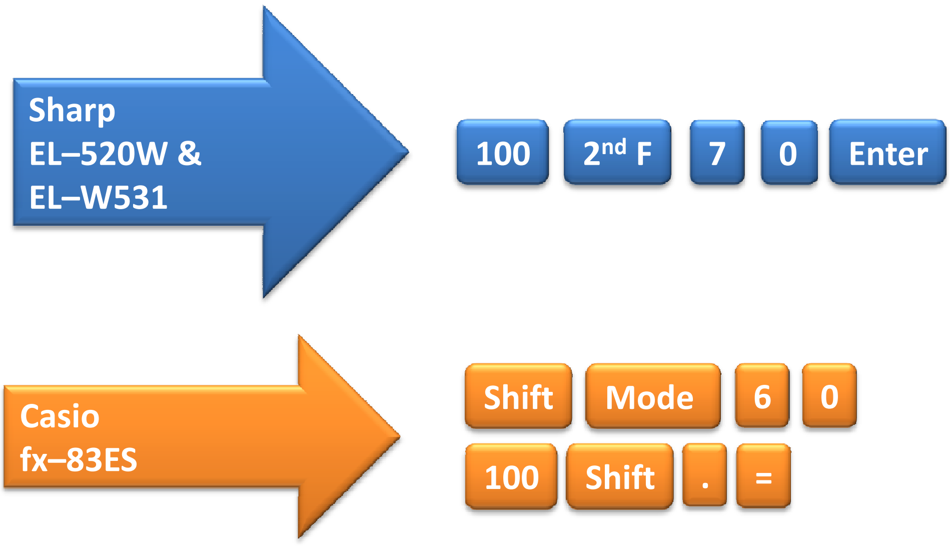
One Random method is to write the 100 numbers on separate slips, put them in a bag, shake them, and take 20 of them out without looking.

A better random method for large samples is to use the Random Number Generator found on your calculator.

Generating Random Numbers using a Calculator

The button might say “RANDOM” (SHARP).
Other makes may have a button “Ran” or “Ran#” or “RanInt”.
Whichever you have, selecting and pressing ENTER repeatedly gives random numbers.

Generate a Random Number between 0 and 99.



Sharp EL–520W & EL–W531: 100, 2nd F, 7, 0, Enter

Casio fx–83ES: Shift, Mode, 6, 0, 100, Shift, ., =

N.B. Calculator should be in LINE IO mode. Shift mode 2

Stratified Random Sample

Example:

Suppose there are 500 girls and 500 boys.

Decide with the person beside you how you could avoid gender bias in taking a sample of 50.

Answer:

Take 2 random samples, one of 25 boys and one of 25 girls, and then combine them.

However, we are unlikely to have exactly equal numbers of boys and girls.

Can you see what to do if the school has 560 boys and 440 girls and we need a sample of 50 ?

Answer:

We sample in proportion to the numbers in the categories.

Boys : $\frac{560}{1000}\times 50 = 28$

Girls : We find the number in the final category by subtracting from the total sample size: $50 - 28 = 22$

Problem

How many of each of the 3 types of computer component should be taken in a sample of 100 categorised by type of component?

Type	A	B	C	Total
Number	300	260	40	600

Solution :

The total number of components $= 600$

Component A: $\frac{300}{600}\times 100 = 50$

Component B: $\frac{260}{600}\times 100 = 43.3$ [Round to 43]

Component C: $100 - 50 - 43 = 7$

We need 50, 43 and 7 respectively.

Cluster & Quota Sampling

Cluster

- Splitting the population into similar parts or clusters can make sampling more practical.
- Then we could simply select one or a few clusters at random and perform a **census** within each of them.

Quota

- Non probability sampling method
- Example: Opinion Polls
  - 1000 – (2000) in all 43 constituencies
  - Split by gender, age, rural, urban, etc.
  - Not truly random, not equal chance of being selected as interviewer has been told what to get

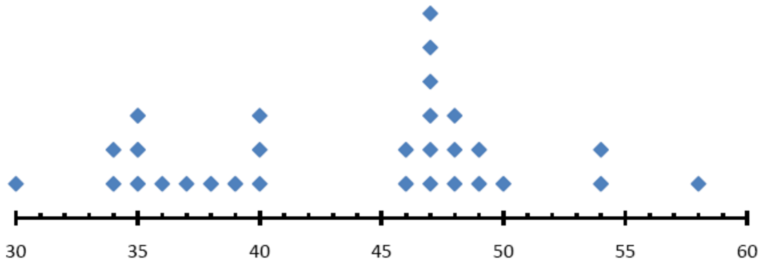
Representing Data Graphically – Line Plots (Univariate)

Example: Suppose thirty people live in an apartment building. These are their ages:

58, 30, 37, 36, 34, 49, 35, 40, 47, 47, 39, 54, 47, 48, 54,
50, 35, 40, 38, 47, 48, 34, 40, 46, 49, 47, 35, 48, 47, 46

Represent this data using a line plot.

Solution:



Note:

Clusters are isolated groups of points, such as the ages of 46 through 50.

Gaps are large spaces between points, such as 41 and 45.

Outliers are items of data which lie far away from the overall pattern of the rest of the data, such as the data for ages 30 and 58.

Representing Data Graphically, Stem and Leaf Plots

The ages of the 30 members of an aerobics class are:

19	22	31	17	8	12	23	47	53	47	19	46	38	59	47
52	21	58	54	26	32	47	55	62	64	36	37	43	15	51

These are presented as a stem and leaf diagram by using the first digit as **stem**, and the second digit as the **leaf**:

19 is written 1|9
22 is written 2|2
31 is written 3|1

17 is then 1|9 7

and 8 is 0|8

0	8
1	9 7
2	2
3	1
4	
5	
6	

So the complete diagram is:

0	8
1	9 7 2 9 5
2	2 3 1 6
3	1 8 2 6 7
4	7 7 6 7 7 3
5	3 9 2 8 4 5 1
6	2 4

A Stem and Leaf plot is like a histogram but it shows the individual values

Now put the leaves in order:

0	8
1	2 5 7 9 9
2	1 2 3 6
3	1 2 6 7 8
4	3 6 7 7 7 7
5	1 2 3 4 5 8 9
6	2 4

15th value

16th value

mode

Key : 6|2 means 62 years

You should include a key with a stem and leaf plot

From this you can pick out the mode and identify the median.

Mode = 47

The median is the $\frac{30+1}{2} = 15\frac{1}{2}$th value $= 40.5$

The average of the 15th and 16th values is $\frac{38+43}{2} = 40.5$

Back to Back Stem Plot Example

Sample Paper, OL Q9

The students in a Leaving Certificate class decided to investigate their heights. They measured the height of each student, in centimetres. The heights of the boys and the girls in the class are given below:

Boys						Girls					
173	180	174	175	178	176	167	161	160	157	164	172
180	171	170	187	176	166	168	149	161	167	167	171

(a) Construct a back-to-back stem and leaf plot of the above data.

(b) State one difference and one similarity between the two distributions.

Solution

(a)

Boys		Girls
	14	
	14	9
	15	
	15	7
	16	0 1 1 4
6	16	7 7 7 8
4 3 1 0	17	1 2
8 6 6 5	17	
0 0	18	
7	18	

Key for boys : 17|3 represents 173 cm

Key for girls : 16|1 represents 161 cm

(b) The boys are taller on average

The spread of the data is about the same for each set

Advantages of Stem Plots

- A Stem Plot displays each separate data value
- They give a quick clear picture of a distribution and make it easy to identify clusters of data from the lengths of the branches
- A Stem and Leaf Plot shows how wide a range of values the data cover, where the values are concentrated and whether the data has any symmetry
- As the values on the branches are ordered it is very easy to pick out the median, quartiles, maximum and minimum values and to identify any outliers
- They are easily created by hand
- Two data sets can be compared using a Back-to-Back Stem Plot
- Both discrete and continuous data sets can be displayed

Disadvantages of Stem Plots

- They cannot display categorical type data
- Small data sets with a large range can be difficult to display on the stem plot without rounding e.g. 432, 507, 534, 581, 609, 626, 671, 712, 719

  [These data points could be displayed with the hundreds digits as the stem and the tens as the leaves.]

Notes