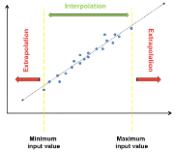
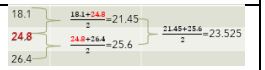
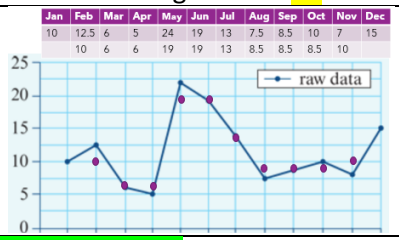
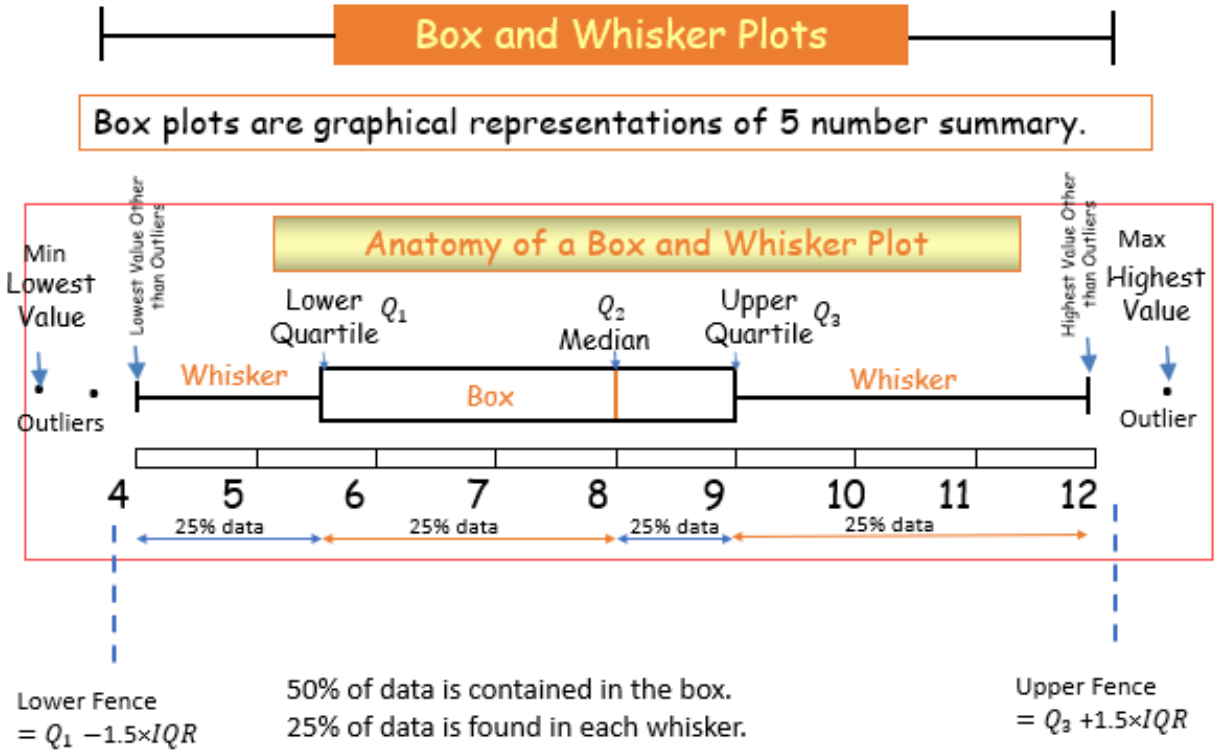
Topics	Data Types 1A		Display/Analyse Tools		Report/Explain/Interpret/Describe		
Univariate Data Chapter 2	Categorical variables 1B	Nominal data Ordinal data	Bar chart, Pie Chart, Frequency Table 1a, Segmented bar chart		Mode/ Modal Value Frequency types, Frequency % = $\frac{Count}{Total\ Count} \times 100\%$		
	Numerical variables 1C	Discrete data	Boxplots 1c 1.2, Grouped Frequency Tables	Stem plot 1b, dot plot 1b 1.2	Shape → Centre → Spread → Outliers → 1F	Symmetric Mean $\bar{x}$ 1.1 Standard Deviation S 1G 1.1  68-95-99.7% rule 1H 1I Z-score=Z=$\frac{x-\bar{x}}{S}$ 1I $x = \bar{x} + Z * S$	Skewed 1.1 Median M or Q_2 IQR, Range $\begin{cases} \text{Lower Fence} = Q_1 - 1.5 * IQR \\ \text{Upper Fence} = Q_3 + 1.5 * IQR \end{cases}$ 1.1 5-figure summary: Min, Q_1, Q_2, Q_3, Max IQR=$Q_3 - Q_1$, Range=Max−Min
		Continuous data		Stem plot 1b, histogram 1b 1.1, log 1.2 log-histogram 1.2			
Bivariate Data Chapter 2	Two categorical variables 2A		Segmented bar chart, two-way frequency table 2a, parallel bar chart 2.1		Mode/ Modal Value Frequency types		
	One categorical, one numerical variable 2B		Back-to-back stem plots, parallel dot plots, parallel box plots 2b 2.2		Shape → Centre → Spread →	Symmetric 1.1 Mean $\bar{x}$ Standard Deviation S	Skewed 1.1 Median M or Q_2 IQR, Range
Chapter 7	Two numerical variables 2C  Interpolation Extrapolation The assumptions for fitting a least squares line 1. the data is numerical 2. the association is linear 3. there are no clear outliers.		Scatterplot 2c 3.2 Explanatory variable → explains/predicts → Response variable residual = actual data value y − predicted 3.3 value y residual = $y - \hat{y}$ Nil pattern residual 3.2 plot 2f = Linear relation Curved/ patterned residual plot ≠ linear relation		Strength → Direction → Form → 3A 3B 3C 2e LSRL y=a+bx 3.1 2e slope $b=\frac{rs_y}{s_x}$ 2e y-intercept $a = \bar{y} - b\bar{x}$	Strong/Moderate/Weak (Check r value) 2D 3.1 Positive / Negative Linear / Non-linear Reporting 2d on Coefficient of Determination r^2 3.1 Almost [r^2 in %] of [RV y] can be explained / predicted by [EV x].	

Time Series 4A 4E: LSRL	Features 4A	Moving smoothing 4.1			Seasonal Index S.I. 4D	Deseasonalising 4D
	Trend Cycles Seasonality Structure change Outliers 	Moving Mean 4B 4.1 3/5 moving mean  	2/4 Moving mean  	Moving Median 4C  	S.I.=$\frac{Value\ for\ Season}{Yearly\ Average}$ Yearly Average=$\frac{Sum\ of\ Season\ Values}{No.of\ season\ per\ year}$ **Correct** S.I. = $(\frac{1}{S.I.} - 1) \times 100$ + means ↑, − means ↓	Deseasonalised Figure = $\frac{Actual\ Figure}{S.I.}$ = $Actual\ Figure * \frac{1}{S.I.}$ Actual figure= Deseasonalised Figure * S.I.

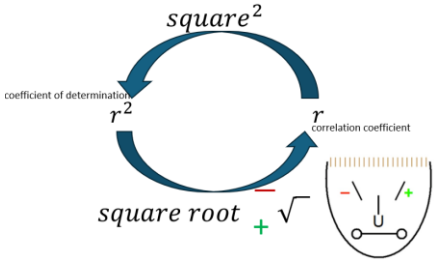
Note: Textbook Summary Notes Section #, Report Instruction Notes #. CAS Instruction Notes #

1E: The Box Plot **1.2**



2D: Correlation 3.1& Causation

$0.75 \leq r \leq 1$ **Strong, positive, linear association**
$0.5 \leq r < 0.75$ **Moderate, positive, linear association**
$0.25 \leq r < 0.5$ **Weak, positive, linear association**
$-0.25 < r < 0.25$ **No association**
$-0.5 < r \leq -0.25$ **Weak, negative, linear association**
$-0.75 < r \leq -0.5$ **Moderate, negative, linear association**
$-1 \leq r \leq -0.75$ **Strong, negative, linear association**



Name of Variation	Symbol of Variation	Equation of Variation	CAS	
			Slope/Equation	Plot/Equation
Direct Variation	$y \propto x$	$y = kx$	**3.1**	$y = kx + c$
Inverse Variation	$y \propto \frac{1}{x}$	$y = \frac{k}{x}$	**3.1**	$y = \frac{k}{x} + c$
Logarithm Variation	$y \propto \log x$	$y = k \log x$	**3.1**	$y = k \log x + c$
Variation involving powers	$y \propto x^2$	$y = kx^2$	**3.1**	$y = kx^2 + c$
Other variations	$y \propto \sqrt{x}$ $y \propto x^3$ $y \propto \frac{1}{x^2}$	$y = k\sqrt{x}$ $y = kx^3$ $y = \frac{k}{x^2}$	**3.1**	

Chapter 9

1A: Types of data

Categorical: characteristics/qualities

- Nominal: grouped according to characteristics
- Ordinal: can be grouped and ordered

Numerical: numbers/quantities

- Discrete: whole numbers, can be counted
- Continuous: is measured

1B: displaying categorical data

Count frequency: number of times the category appears in the data

Percentage frequency: $\frac{count\ frequency}{total\ count} \times 100$

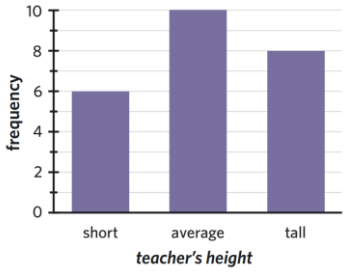
Mode: most frequently occurring value or category

Frequency Table:

teacher's height	frequency	
	number	%
short	6	25.0
average	10	41.7
tall	8	33.3
total	24	100.0

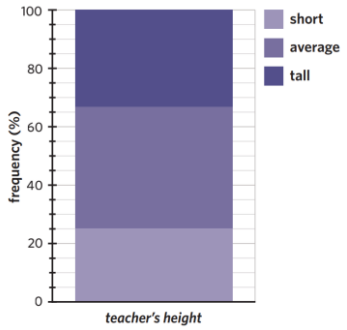
Bar chart:

- Must have gaps between bars



Segmented bar chart:

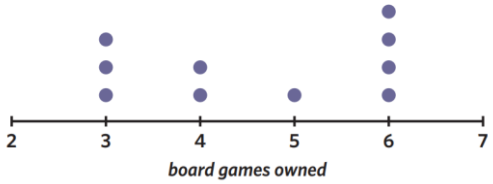
- Can be count or percentage frequency
- Must have a key



1C: Displaying Numerical data

Dot plot

- Discrete data
- Small data sets



Stem and leaf plot

- Needs a key
- Can have class intervals (splitting the stem in two if it is really large)

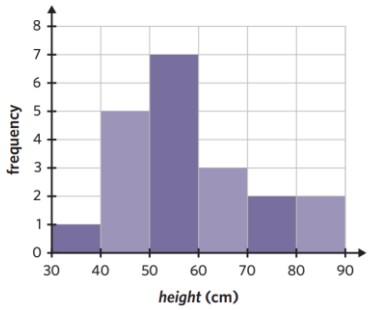
No intervals	With class intervals of 5
Key: 1 | 2 = 1.2	**Key:** 1 | 2 = 12
1 | 3 3 4 6 8	0 | 1 1 2 3 4
2 | 0 4 9	0 | 5 6 6 8 8 9
3 | 1 1 1 4 5 8	1 | 2 3 3
4 | 2	1 | 6 7 7 8 9 9

Grouped frequency tables

height (cm)	frequency	
	number	%
30–<40	1	5
40–<50	5	25
50–<60	7	35
60–<70	3	15
70–<80	2	10
80–<90	2	10
total	20	100

Histogram:

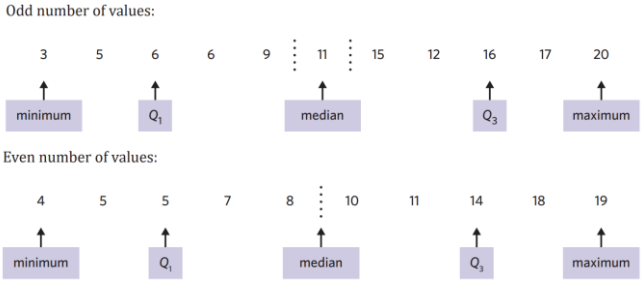
- Continuous data
- Intervals – no gaps between bars
- No gaps between bars
- X-axis markers are always a whole number



1E: the five-number summary and boxplots

5 number summary:

- Minimum: smallest value in data set
- Q1: median of the lower half
- Median: middle value in an ordered data set
- Q3: median of the upper half
- Maximum: largest value in data set



Spread: refers to how variable the data set it

Range = maximum – minimum

Interquartile range: measure of spread of the middle 50% of a data set. Accurate measure of spread when outliers are present.

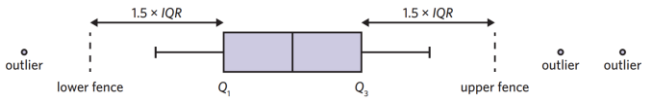
IQR = $Q_3 - Q_1$

Outliers: values which fall outside of what is 'normal'. Outliers are still the minimum and maximum value!

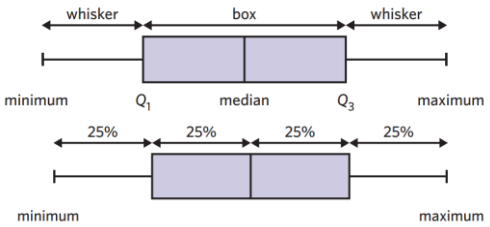
Fence: defines the boundary of what is an outlier. If a value is less than the lower fence or greater than the upper fence it is considered to be an outlier.

$$lower\ fence = Q_1 - (1.5 \times IQR)$$

$$upper\ fence = Q_3 + (1.5 \times IQR)$$

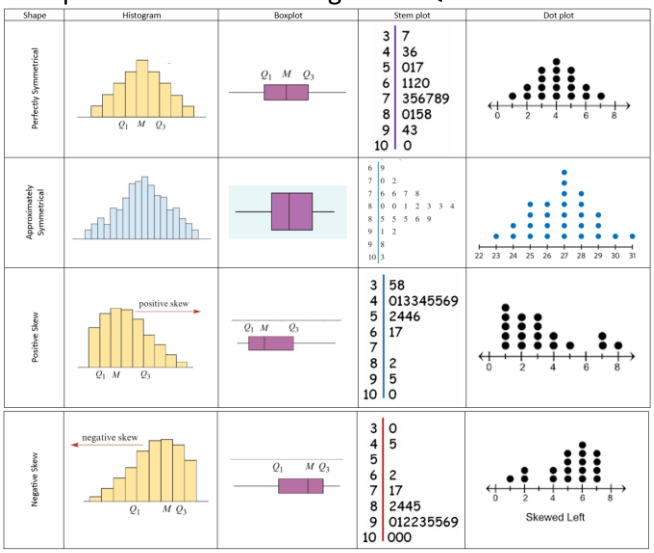


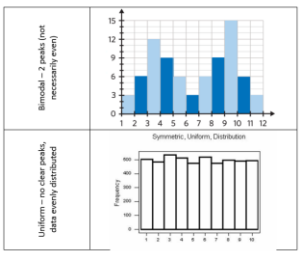
Boxplots:



1F: describing numerical data

- Shape: is the data symmetrical, skewed or have any outliers?
- Centre: What is the median value?
- Spread: What is the range and IQR?



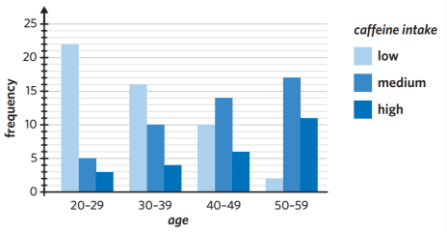


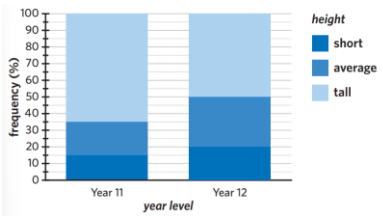
2A: association between 2 variables

Two-way frequency table:

- Columns = EV, Rows = RV
- Percentage frequency is used for greater accuracy when making comparisons if sample sizes are different

Grouped and segmented bar charts:





Describing the association between two variables:

- Whether or not an association between the two variables exists
- Appropriate percentages to support findings

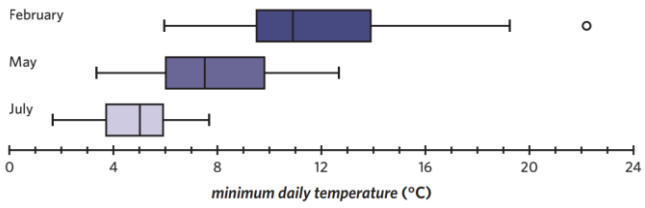
2B: association between numerical and categorical variables

- Back to back stem plot

Key: 3 | 7 = 37 points

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

- Parallel boxplot



- Making comparisons: refer to 1F and compare shape, centre and spread of the two categories

2C: association between two numerical variables

Response variable: RV, may be explained or predicted by changes in the explanatory variable.

Explanatory variable: EV, used to explain or predict the changes observed in the response variable.

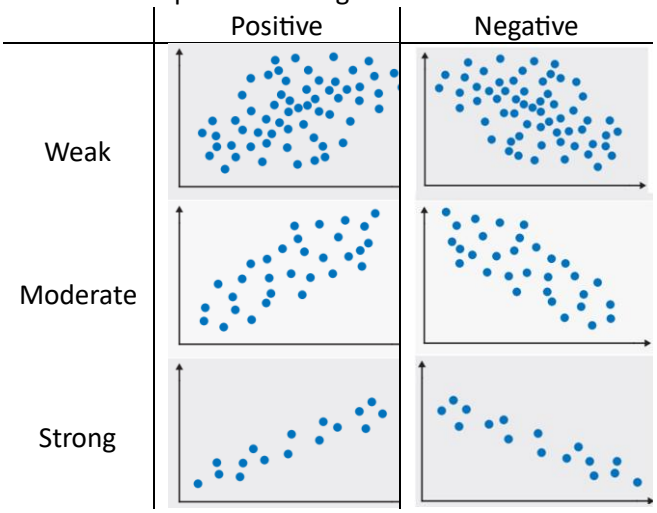
- 'EV explains the RV'

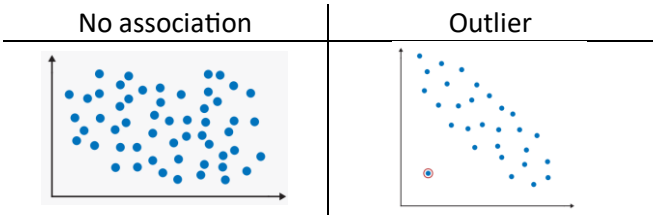
Scatterplots:

- EV = x axis, RV = y axis

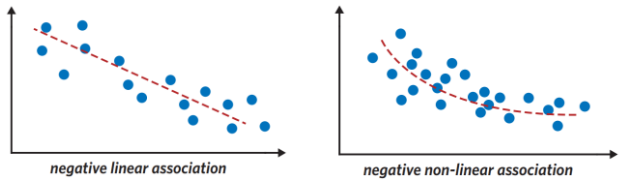
Describing relationship/analysing scatterplots:

- Strength: how close the data points are together
- Direction: positive or negative





- Form: linear (straight) or non-linear (curved)



2D: Correlation and causation

Pearson's correlation coefficient (r): numerical value that determines strength and direction between two numerical variables, assuming:

- Data is linear
- Data is numeric
- No outliers present

$0.75 \leq r \leq 1$	Strong, positive, linear association
$0.5 \leq r < 0.75$	Moderate, positive, linear association
$0.25 \leq r < 0.5$	Weak, positive, linear association
$-0.25 < r < 0.25$	No association
$-0.5 < r \leq -0.25$	Weak, negative, linear association
$-0.75 < r \leq -0.5$	Moderate, negative, linear association
$-1 \leq r \leq -0.75$	Strong, negative, linear association

3A: fitting a least squares regression line

Least squares regression line (LSRL): is the line which creates the minimum sum of the squares of residuals. There are assumptions:

- Data is numerical
- The relationship between variables is linear
- There are no clear outliers present

The line is used to show the general trend in the data and is given by the equation:

$$y = a + bx$$

Intercept Slope

Determining LSRL from a graph: Find the intercept (a) and the slope (b).

- Intercept: read directly from the graph when the EV is 0
- Slope: choose two points on the line that you can clearly ready the coordinates. Use the rule:

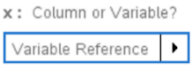
$$b = \frac{rise}{run} = \frac{y_2 - y_1}{x_2 - x_1}$$

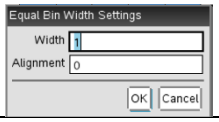
Drawing the LSRL on a graph: Sub in the first value on the x-axis and the last value on the x-axis into the equation. Plot the two points, join the line using a ruler.

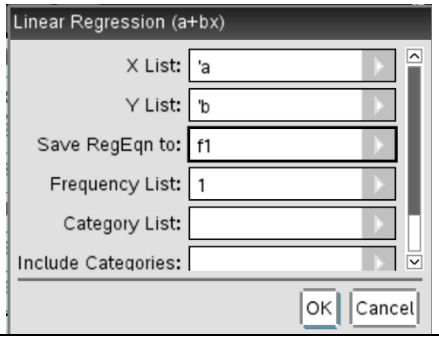
3B: Interpreting LSRL: use the following statements, fill in EV and RV and values of a and b.
y-intercept: when the **EV** is 0, the **RV** is **a**.
Slope: for every one-unit increase in the **EV**, the **RV increases/decreases** by **b**. (If b is positive, increases, if b is negative, decreases)

Making predictions: the LSRL can be used to predict the value of the RV from the EV.
Interpolation: predicting within the range of data.
Extrapolation: predicting outside the range of data. Less reliable.
How to predict: sub the EV value that you are predicting for into the LSRL equation to predict the RV.

CAS reference sheet: Data

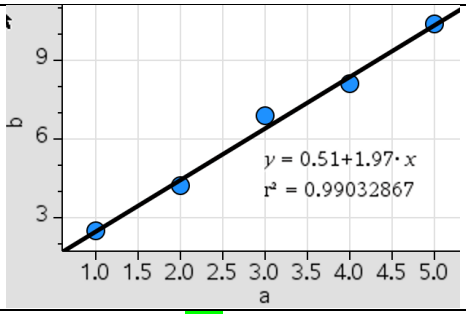
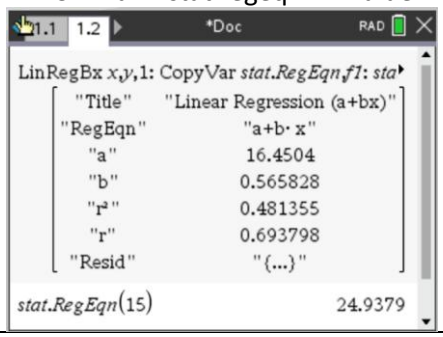
<table>
<tr>
<td>1.1 Univariate Data Statistics: to find the 5 number summary, mean and standard deviation</td>
<td>
1. Ctrl n > 4, Enter data in ‘list and spreadsheets’ (ensure a title)
2. Menu > 4 > 1 > 1 OR ctrl i > 1 > menu > 6 > 1 > 1 in a new page
3. Select the title you gave the data in the ‘X-List’


<table>
<tr><td>MinX</td><td>1.</td></tr>
<tr><td>Q₁X</td><td>1.5</td></tr>
<tr><td>MedianX...</td><td>3.</td></tr>
<tr><td>Q₃X</td><td>4.5</td></tr>
<tr><td>MaxX</td><td>5.</td></tr>
</table>
<table>
<tr><td>$\bar{x}$</td><td>3.</td></tr>
<tr><td>Σx</td><td>15.</td></tr>
<tr><td>Σx^2</td><td>55.</td></tr>
<tr><td>sx := sn-...</td><td>1.58113...</td></tr>
</table>
</td>
</tr>
<tr>
<td>1.2 Graphing Univariate Data: dot plot, boxplots, histograms</td>
<td>
1. Ctrl n > 4, Enter data in ‘list and spreadsheets’ (ensure a title)
2. Ctrl doc OR ctrl i OR home menu > 5 to open a new page and select ‘data and statistics’
3. Click to add data to the x-axis
4. Menu > 1: choose the plot type you are after
Histogram settings:
5. Menu > 2 > 2 > 2 > 1
6. Width is where you put your interval size. Alignment is where the graph will start on the x-axis
7. Zoom: menu > 5 > 2
8. Frequency or Percent: menu > 2 > 2 > 1


</td>
</tr>
<tr>
<td>2.1 Graphing Bivariate Data: Grouped Bar Chart</td>
<td>
1. Create first Bar Chart
Menu > 2 > 9 > select second set of data
</td>
</tr>
<tr>
<td>2.2 Graphing Bivariate Data: parallel boxplots</td>
<td>
2. Create first boxplot
3. Menu > 2 > 5 > select second set of data
</td>
</tr>
<tr>
<td>3.1 Bivariate Data Statistics: a, b, r, r^2</td>
<td>
1. Ctrl n > 4, Enter data in ‘list and spreadsheets’(ensure you give both columns a title)
2. Menu > 4 > 1 > 4 (Spreadsheets) OR ctrl i > 1 > Menu > 6 > 1 > 4 (Calculator)
3. Select your EV in the ‘X-List’ and your RV in the ‘Y-List’


<table>
<tr><td>RegEqn</td><td>a+b*x</td></tr>
<tr><td>a</td><td>0.51</td></tr>
<tr><td>b</td><td>1.97</td></tr>
<tr><td>r^2</td><td>0.99032...</td></tr>
<tr><td>r</td><td>0.99515...</td></tr>
</table>
</td>
</tr>
<tr>
<td>3.2 Scatterplots</td>
<td>
1. Ctrl n > 4, Enter data in ‘list and spreadsheets’ (ensure you give both columns a title)
2. Ctrl doc OR ctrl i OR home menu > 5 to open a new page and select ‘data and statistics’
3. Click to add the EV to the x-axis and the RV to the y-axis

Scatterplot Least Squares Regression Line:
4. Menu 4 > 6 > 2 OR ctrl i > 1 > Menu > 6 > 1 > 4 (Calculator)
</td>
</tr>
</table>

	
3.3 Prediction by LSRL	1. Above 3.2 step1-3 to entre EV and RV. 2. ctrl i > 1 > Menu > 6 > 1 > 4 (Calculator) 3. var > stat.regeqn > x value ↲ 
4.2 Solve	1. Calculator Page 2. Menu > 3 > 1 3. You write inside the brackets, there must be an equals sign (=), you must choose a variable (x) and then use the same variable at the end and do 'comma variable' (,x) $\text{solve}(10=x+2,x) \quad x=8.$
4.3 Back to top of List	Ctrl > 7

1a. Univariate Categorical Data: Frequency Table

The **[types of categories]** of **[total frequency]** **[frequency type]** were classified as **[list of categories].**

Modal Category

The majority of **[frequency type], [modal percentage],** were found to be **[modal category].**
Of the remaining **[frequency types], [percentage X]** were found to be **[category X],** while **[percentage Y]** were found to be **[category Y],** and while **[etc.].**

Equal Categories

The **[frequency types]** all had roughly the same percentages where **[category X]** had **[percentage X], [category Y]** had **[percentage Y], [etc.].**

1b. Univariate Numerical Data: Histogram, Dot Plot, Stem Plot

The shape of the distribution is **[symmetric/positively skewed/negatively skewed]**
The distribution has a **[standard dev./range/IQR]** of **[value]**
The distribution has a **[mean/median/mode]** of **[value]**
The distribution **[has #/has no]** outliers.

Refer to 1F: Describing numerical data

1c. Univariate Numerical Data: Box Plot

The distribution is **[positively skewed/negatively skewed]** with **[outliers/no outliers]**. The distribution is centred at **[value],** the median value. The spread of the distribution, is measured by the IQR, is **[value]** and, as measured by the range **[value].** If outliers present: There are **[value]** many outliers: **[list of outliers]**

2a. Bivariate Data (Both Categorical): Two-way Frequency Table

Worked example: Is there an association between interest in sports and age group?
Yes, the percentage of males with a high level of interest in sport steadily decreases with age group from **56.5 %** for the 'under 18 years' age group, to **35.0%** for the '36-50 years' age group.

Interest in sport	*Age group (%)* *Under 18 years*	*19–25 years*	*26–35 years*	*36–50 years*
High	56.5	50.2	40.7	35.0
Medium	30.1	34.4	36.8	45.8
Low	13.4	13.4	22.5	20.3
Total	100.0	100.0	100.0	100.0

2b. Bivariate Data (one categorical, One Numerical): Comparing two boxplots:

The distributions at **[variable name]** are **[symmetric/positively skewed/negatively skewed]** for both **[boxplot variables]**. There **[are/are no]** outliers. The median **[variable name]** is higher for **[boxplot 1], (M= value),** than **[boxplot 2], (M= value).** The IQR is also greater for **[boxplot 1], (IQR= value),** than **[boxplot 2], (IQR= value).** The range of **[variable name]** is also greater for **[boxplot 1], (R= value),** than **[boxplot 2], (R= value).** Yes, there is an association between **[variable name1] [variable name 2]** as their median change from **(M= value) to (M= value).**

2c. Bivariate Data (Both Numerical): Scatter Plot

There is a **[strong/moderate/weak], [positive/negative], [linear/non-linear]** relationship between **[response variable y]** and **[explanatory variable x]**. There **[are/are no]** clear outliers.

2d. The coefficient of determination (r^2):

The coefficient of determination indicates that **[r^2 x 100]** % of the variation in **[response variable]** is explained by the variation in **[explanatory variable]** and **[remaining %]** is explained by other factors and not explained by **[explanatory variable]**.

2e. Least squares line:

The equation of the regression line is: **[response variable]** = **[a]** + **[b]** x **[explanatory variable]**

Slope (b):

On average, **[response variable]** **[increases/decreases]** by **[b units]** for every one **[unit]** increase in **[explanatory variable].**

y- intercept (a):

When **[explanatory variable]** is 0, **[response variable]** is predicted to be **[a units].**

2f. Residual Plot

The residual plot shows a **[random scatter/ curved pattern]** indicating there is a **[linear/non- linear]** relationship between **[response variable]** and **[explanatory variable].**

2g. Prediction Reliability

The prediction using **[explanatory variable]** of **[value of EV]** is **[reliable/ unreliable]** as it is **[within/ outside]** the date range and is therefore **[interpolation/extrapolation].**