Data analysis

standardised score	$z = \dfrac{x - \bar{x}}{s_x}$
lower and upper fence in a boxplot	lower $\quad Q1 - 1.5 \times IQR \qquad$ upper $\quad Q3 + 1.5 \times IQR$
least squares line of best fit	$y = a + bx, \qquad$ where $\qquad b = r\dfrac{s_y}{s_x} \qquad$ and $\qquad a = \bar{y} - b\bar{x}$
residual value	residual value = actual value – predicted value
seasonal index	seasonal index $= \dfrac{\text{actual figure}}{\text{deseasonalised figure}}$

Recursion and financial modelling

first-order linear recurrence relation	$u_0 = a, \qquad u_{n+1} = Ru_n + d$
effective rate of interest for a compound interest loan or investment	$r_{\text{effective}} = \left[\left(1 + \dfrac{r}{100n}\right)^n - 1\right] \times 100\%$

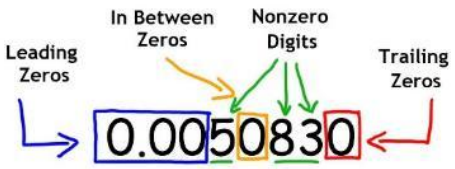
Matrices

determinant of a 2 × 2 matrix	$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}, \qquad \det A = \begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc$
inverse of a 2 × 2 matrix	$A^{-1} = \dfrac{1}{\det A}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}, \qquad$ where $\qquad \det A \neq 0$
recurrence relation	$S_0 =$ initial state, $\qquad S_{n+1} = T S_n + B$
Leslie matrix recurrence relation	$S_0 =$ initial state, $\qquad S_{n+1} = L S_n$

Networks and decision mathematics

Euler's formula	$v + f = e + 2$

Significant Figures



→ **All non- zero values are significant**

4 . 2 (2 sig figs)

→ **All zeros in between are significant**

40002 (5 sig figs)

Or, in the case of decimal values: **4 . 0002** (5 sig figs)

→ **Decimal values**

1. All **final zeros** after the decimal point are significant

4 . 200 (4 sig figs)

2. All **leading zeros** after a decimal point are **NOT** significant

0 . 000**422** (3 sig figs)

→ **Terminal zeros don't count UNLESS there is a decimal point at the end**

420 (2 sig figs)

420 . (3 sig figs)

420 . 0 (4 sig figs)

Rounding Decimals

→ Involves rounding values after the decimal point to however many decimal places

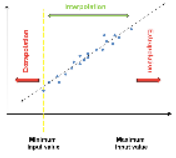
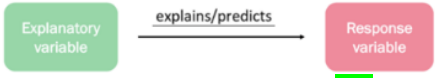
422 . 347

Round to 2 decimal places : 422 . **35**

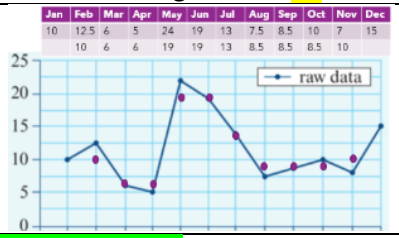
422 . 344

Round to 2 decimal places : 422 . **34**

→ **Money must always be rounded to 2 decimal places or "to the nearest cent"**

<table>
<tr><th>Topics</th><th colspan="2">Data Types 1A</th><th colspan="2">Display/Analyse Tools</th><th colspan="3">Report/Explain/Interpret/Describe</th></tr>
<tr><td rowspan="3">Univariate Data</td><td rowspan="2">Categorical variables1B</td><td>Nominal data
Ordinal data</td><td colspan="2">Bar chart, Pie Chart, Frequency Table1a, Segmented bar chart</td><td colspan="3">Mode/ Modal Value
Frequency types, Frequency % = $\frac{Count}{Total\ Count} \times 100\%$</td></tr>
<tr><td rowspan="2">Numerical variables1C</td><td>Discrete data</td><td rowspan="2">Boxplots 1c 1.2, Grouped Frequency Tables</td><td>Stem plot 1b, dot plot 1b 1.2</td><td rowspan="2">Shape → Centre → Spread → Outliers → 1F</td><td rowspan="2">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 $x = \bar{x} + Z * S$</td><td rowspan="2">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</td></tr>
<tr><td>Continuous data</td><td>Stem plot 1b, histogram 1b 1.1, log 1.2 log-histogram1.2</td></tr>
<tr><td rowspan="3">Bivariate Data</td><td>Two categorical variables2A</td><td colspan="2">Segmented bar chart, two-way frequency table2a, parallel bar chart 2.1</td><td colspan="3">Mode/ Modal Value
Frequency types</td></tr>
<tr><td>One categorical, one numerical variable 2B</td><td colspan="2">Back-to-back stem plots, parallel dot plots, parallel box plots 2b 2.2</td><td>Shape → Centre → Spread →</td><td>Symmetric
1.1 Mean $\bar{x}$
Standard Deviation S</td><td>Skewed
1.1 Median M or Q_2
IQR, Range</td></tr>
<tr><td>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.</td><td colspan="2">Scatterplot 2c 3.2
 

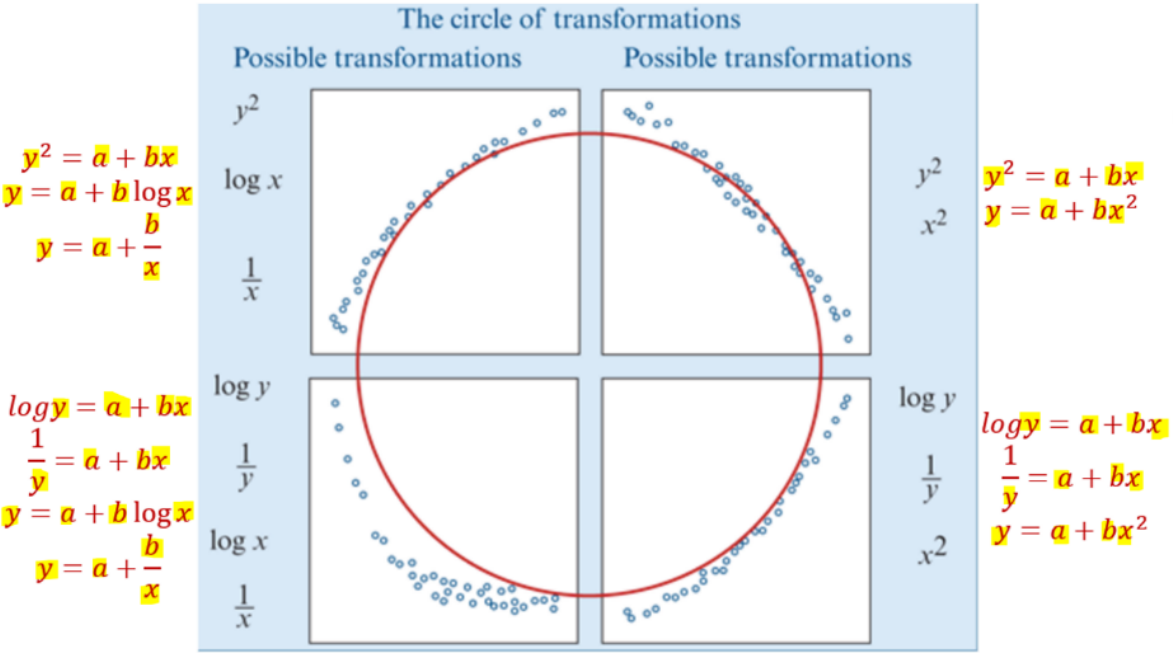
residual = actual data value y – predicted 3.3 value y
residual = $y - \hat{y}$
Nil pattern residual 3.2 plot 2f = Linear relation
Curved/ patterned residual plot ≠ linear relation</td><td>Strength →
Direction →
Form →
3A 3B 3C 2e
LSRL y=a+bx
3.1
2e slope $b=\frac{rs_y}{s_x}$
2e y-intercept
$a = \bar{y} - b\bar{x}$</td><td colspan="2">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].</td></tr>
<tr><td>Time Series
4A 4E: LSRL</td><td>Features4A
Trend
Cycles
Seasonality
Structure change
Outliers
  </td><td colspan="2">Moving smoothing 4.1
Moving Mean 4B 4.1
3/5 moving mean
 
2/4 Moving mean
 
Moving Median 4C
 
</td><td colspan="2">Seasonal Index S.I. 4D
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 ↓</td><td>Deseasonalising 4D
Deseasonalised Figure = $\frac{Actual\ Figure}{S.I.}$
= $Actual\ Figure * \frac{1}{S.I.}$
Actual figure= Deseasonalised Figure * S.I.</td></tr>
</table>

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

3D: Data Transformation & 3E 3.4

Stretching transformation: Squared & reciprocal transformation
Compressing transformation: Logarithmic transformation

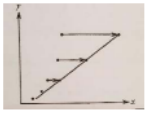
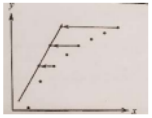
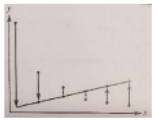


- *Best transformation: strongest r/r^2 value*

Types of transformations:

- *Log: compresses the data*
- *Square: stretches the data*
- *Reciprocal: compresses values greater than 1, stretches values less than 1.*

The Effect of Each Transformation:

Type of Transformation:	Description of Effect:	One Word Description:	Graph of Transformation:
Squared Transformations (x^2 and y^2)	Spreads out the high x-values relative to the lower x-values and vice versa.	Stretching transformation ➢ x^2 stretches high x-values ➢ y^2 stretches high y-values	
Log Transformation (Log_x and Log_y)	Compresses the higher x-values relative to the lower x-values and vice versa	Compressing Transformation	
Reciprocal Transformations	Compresses larger y-values relative to smaller y-values and vice versa	Stretching and Compressing Transformation	

1D Log Scales & Graphs

Log (Base 10) Scale

Logarithms

A logarithm, or log, is a power or exponent or index of a number. That is the log of a^b is b.
For example the logs of $2^3, 5^4$, and 10^6 are 2, 3, and 6 respectively.

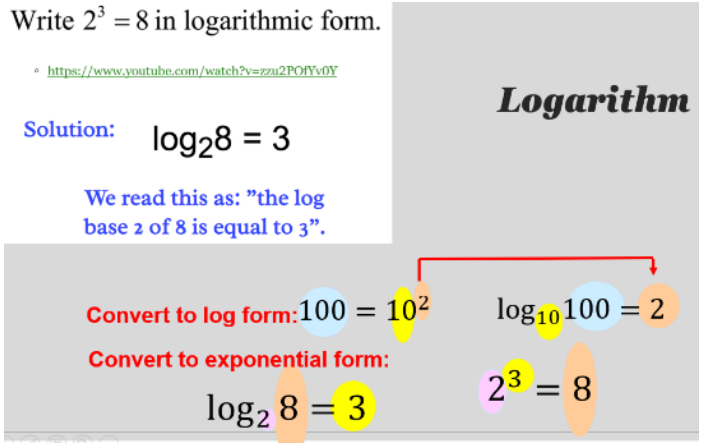
Log (Base 10) Scale

The log (base 10) scale is based of exponentials of base 10, i.e. $10, 10^2, 10^3, 10^4$.
Using the log (base 10) scale allows data ranging over several order of magnitude to be displayed.

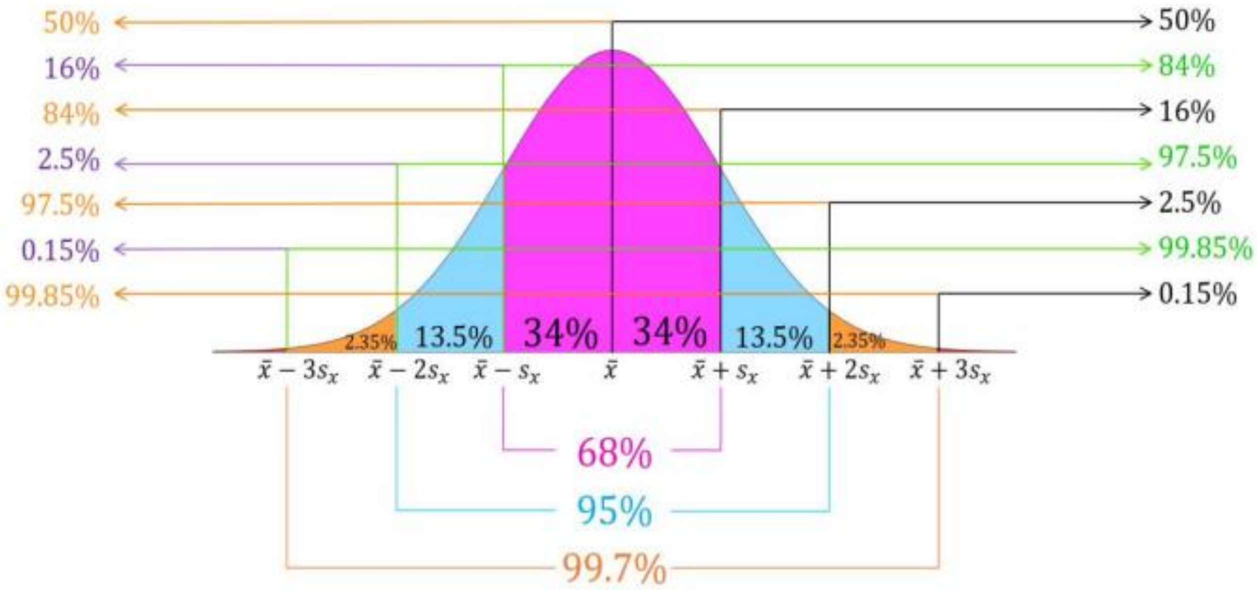
Converting Between Forms using the Log (Base 10) Scale

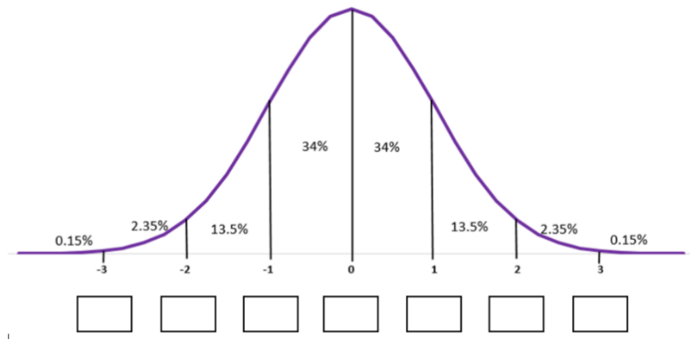
log value $= \log_{10}$(data value) data value $= 10^{\text{log value}}$

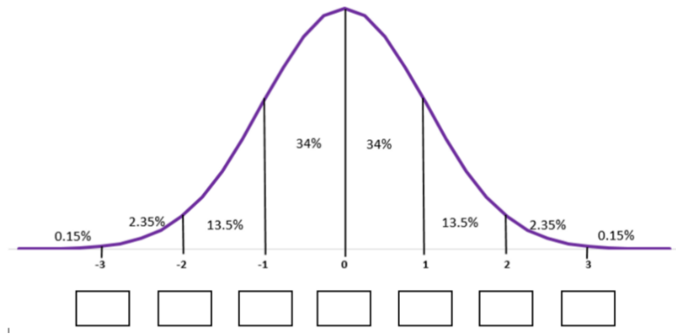
Data Value	0.001	0.01	0.1	10^n	1	10	100	1000
Log Form	$\log_{10} 0.001$	$\log_{10} 0.01$	$\log_{10} 0.1$	$\log_{10} 10^n$	$\log_{10} 1$	$\log_{10} 10$	$\log_{10} 100$	$\log_{10} 1000$
Log Value	-3	-2	-1	n	0	1	2	3
Exponent Form	10^{-3}	10^{-2}	10^{-1}	10^n	10^0	10^1	10^2	10^3



1H The Normal Distribution







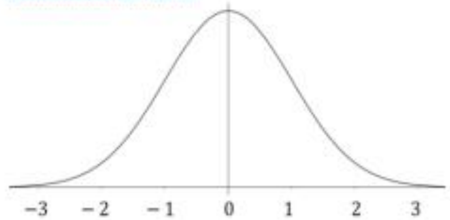
Converting to a Standard Score

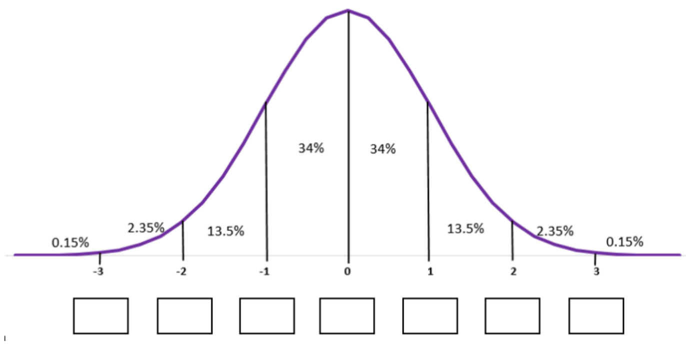
$$z = \frac{x - \bar{x}}{s}$$

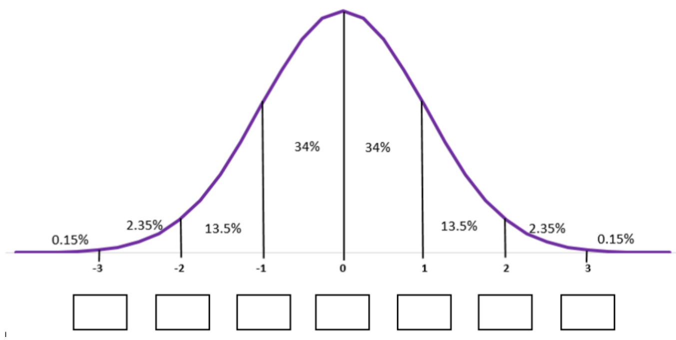
Converting to an Actual Score

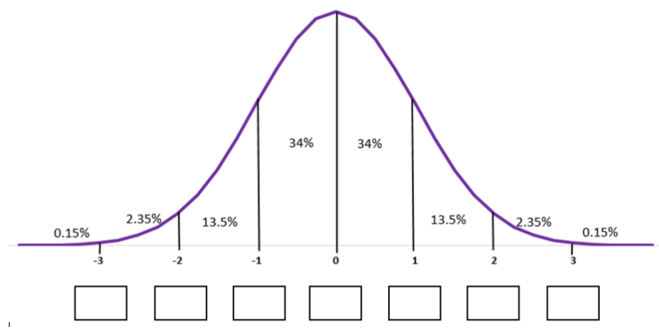
$$x = \bar{x} + z \times s$$

Standardised Bell Curve



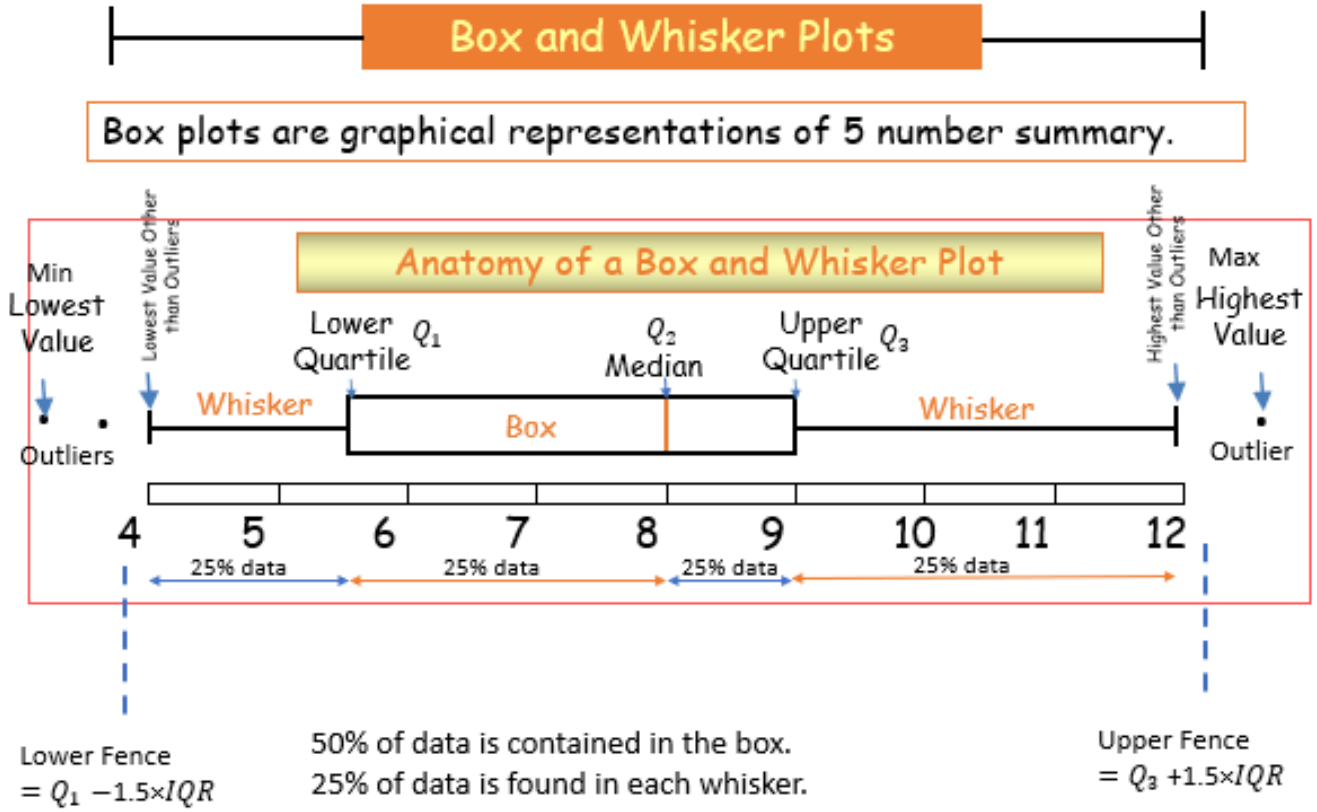




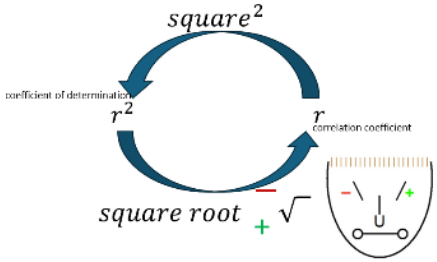


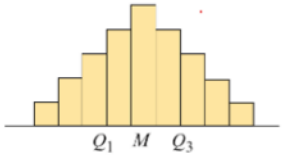
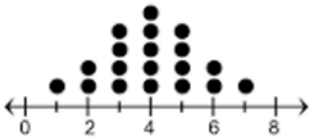
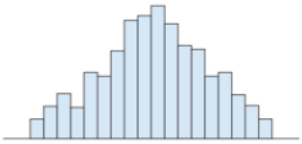
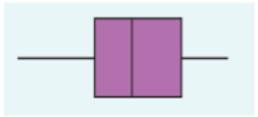
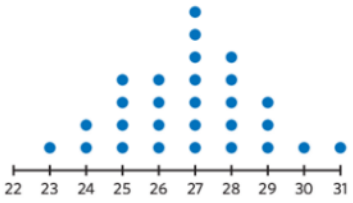
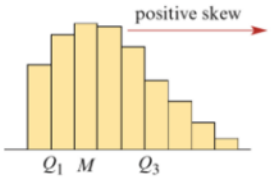
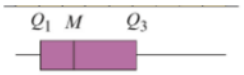
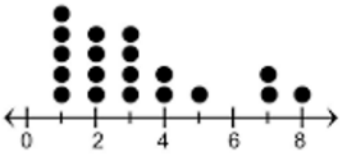
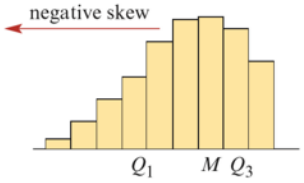
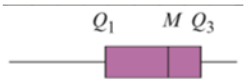
1E: The Box Plot **1.2**

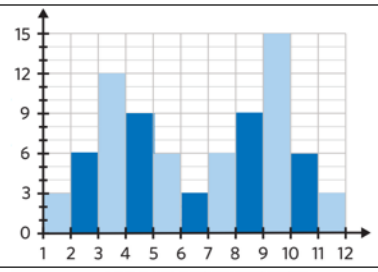
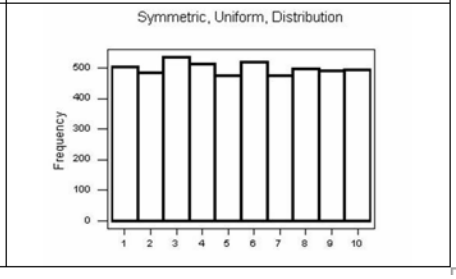
2D: Correlation 3.1& Causation



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



Shape	Histogram	Boxplot	Stem plot	Dot plot
Perfectly Symmetrical			3 | 7 4 | 36 5 | 017 6 | 1120 7 | 356789 8 | 0158 9 | 43 10 | 0	
Approximately Symmetrical			6 | 9 7 | 0 2 7 | 6 6 7 8 8 | 0 0 1 2 3 3 4 8 | 5 5 5 6 9 9 | 1 2 9 | 8 10 | 3	
Positive Skew			3 | 58 4 | 013345569 5 | 2446 6 | 17 7 | 8 | 2 9 | 5 10 | 0	
Negative Skew			3 | 0 4 | 5 5 | 6 | 2 7 | 17 8 | 2445 9 | 012235569 10 | 000	

Bimodal – 2 peaks (not necessarily even)	
Uniform – no clear peaks, data evenly distributed	

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

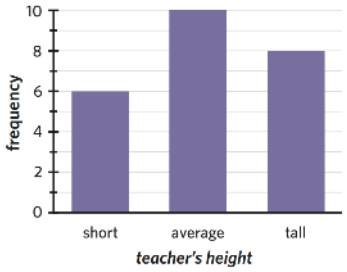
Mode: most frequently occurring value or category

Frequency Table:

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

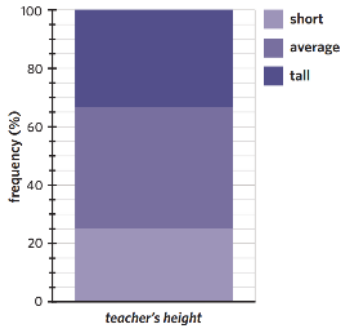
Bar chart:

- Must have gaps between bars



Segmented bar chart:

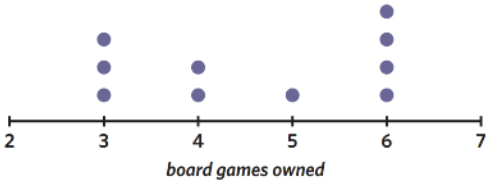
- Can be count or percentage frequency
- Must have a key



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)

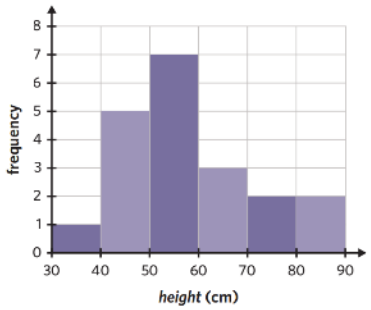
No intervals	With class intervals of 5
Key: 1 | 2 = 1.2	**Key:** 1 | 2 = 12
1 | 3 3 4 6 8	0 | 1 1 2 3 4
2 | 0 4 9	0 | 5 6 6 8 8 9
3 | 1 1 1 4 5 8	1 | 2 3 3
4 | 2	1 | 6 7 7 8 9 9

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



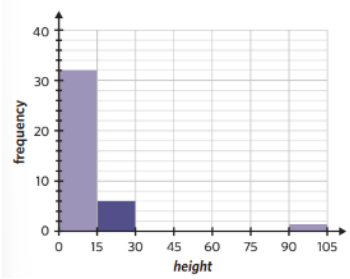
1D: Log scales and graphs

- Log scales are used to compress data that has a large range, making it more even and able to be displayed on the same set of axes.
- The base is always 10
- When undoing the log scale do ten to the power of the scale (eg. $10^{2.2} = 158.5$)

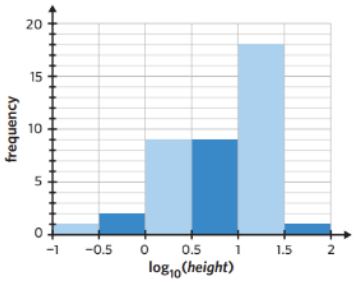
If... $\log_b(x) = y$ (argument, exponent, base)

then... $b^y = x$

From…



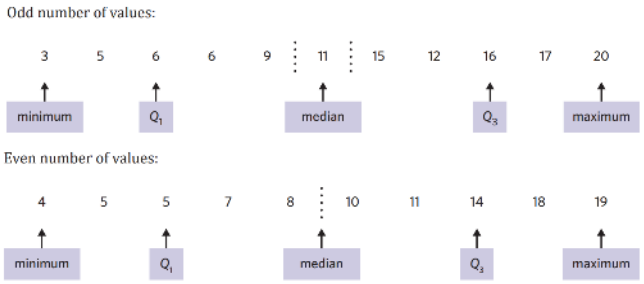
To…



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.

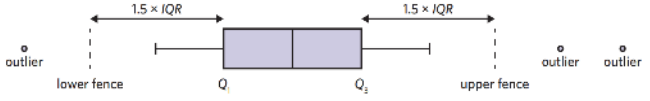
IQR = $Q_3 - Q_1$

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

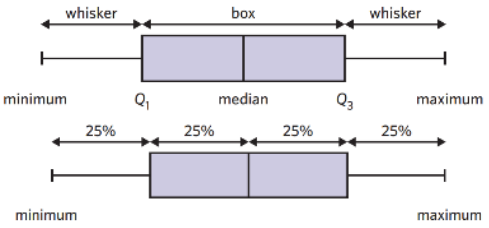
Fence: defines the boundary of what is an outlier. If a value is less than the lower fence or greater than the upper fence it is considered to be an outlier.

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

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

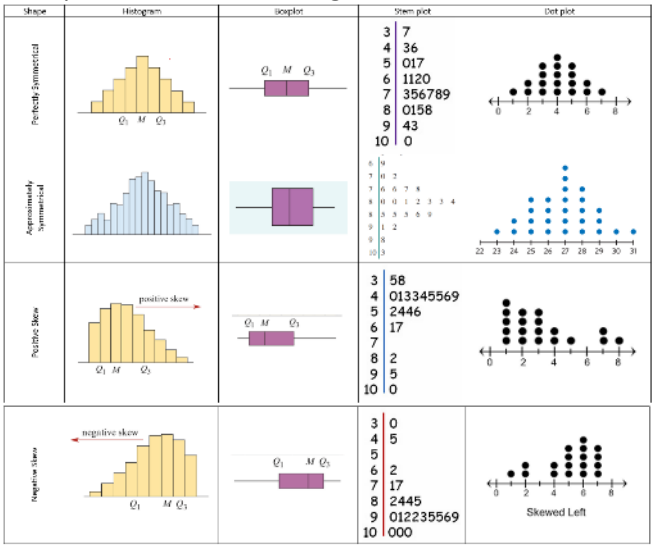


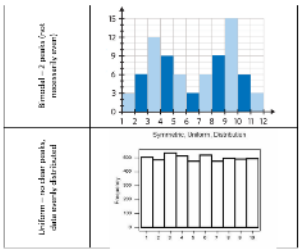
Boxplots:



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?





1G: Standard deviation

Population: the entire group is used to collect data.
Sample: smaller subset of the population (this is usually what is used).

Mean: measure of centre – the AVERAGE. $\bar{x}$

- Calculated by adding all the data values together and then dividing by the number of values.

$\bar{x} = \frac{\Sigma x}{n}$, where Σx is the 'sum of all values', and n is the number of values in the data set.

Standard deviation: measure of spread based on the average deviation of each data point compared to the mean. It can be calculated by hand but please use CAS.

$$s_x = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n - 1}}$$

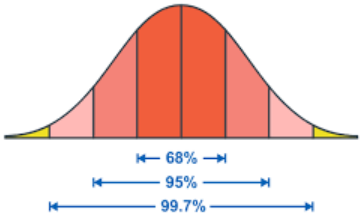
1H: The Normal Distribution

Normal Distribution: is a symmetrical (or approximately) numerical data set centred around the mean.

- Bell shaped
- Mean and median are equal

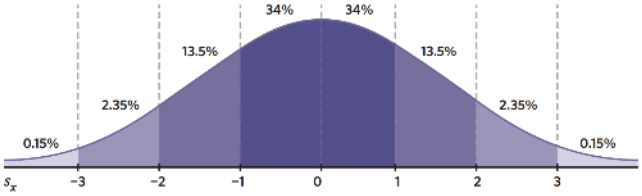
68-95-99.7% rule:

- 68% of the data lies within one standard deviation of the mean
- 95% of the data lies within two standard deviations of the mean
- 99.7% of the data lies within three standard deviations of the mean



The bell curve can be broken into each section:

- The mean lies in the centre (0)



1I: z-scores

Standardised score:

- Z-score
- Measure of the number of standard deviations between the mean and a data value
- Each data value is an 'actual score'
- Positive = above mean, negative = below mean, zero = equal to mean

$$z = \frac{x - \overline{x}}{s_x}$$

- z is the standardised score
- x is the actual score
- $\overline{x}$ is the mean
- s_x is the standard deviation

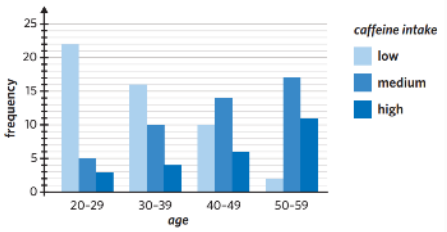
Actual score: $x = \overline{x} + (z \times s_x)$

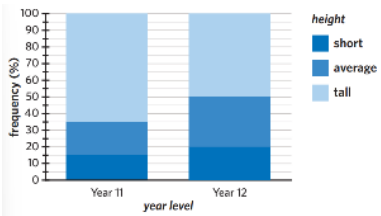
2A: association between 2 variables

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

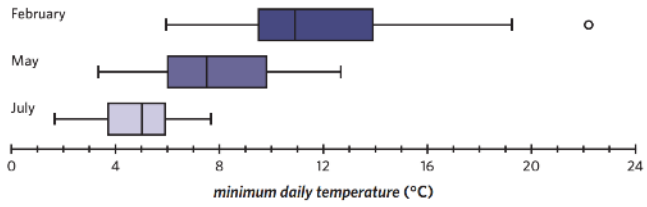
2B: association between numerical and categorical variables

- 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.

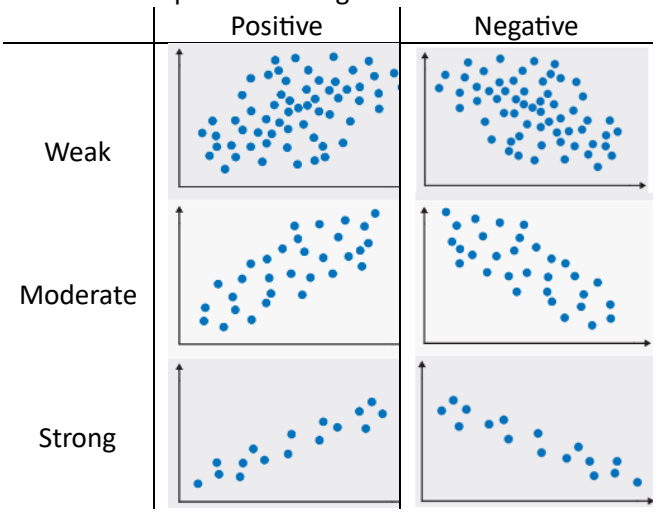
- 'EV explains the RV'

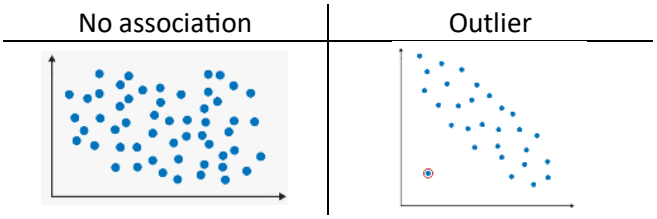
Scatterplots:

- EV = x axis, RV = y axis

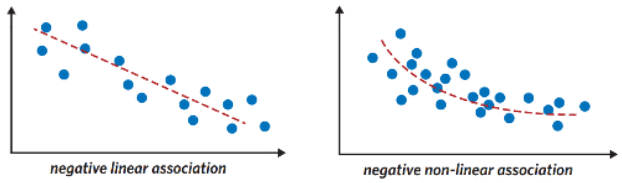
Describing relationship/analysing scatterplots:

- 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

Correlation and Causation: just because two variables have a high correlation, it doesn't mean that one causes the change in the other. Some explanations:

- **Common response:** a third variable that is the likely cause of correlation, acting on both variables. Eg. Number of people wearing sunscreen and feinting → the sunscreen isn't causing people to feint… the third variable would be temperature. This is common cause as temperature affects **both** variables.
- **Confounding variable:** external variable that can also produce a change to the RV. Eg. Plant height and water intake. Water intake does effect plant height (RV) but so does sun, soil quality, buys, season, temperature…
- **Coincidence:** two variables correlate but have no relation to each other. Pure chance. No logical explanation.

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:

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

Calculating the LSRL from summary statistics:

$$b = r \times \frac{s_y}{s_x}$$
$$a = \bar{y} - b\bar{x}$$

- a is the y-intercept
- b is the gradient
- r is Pearson's correlation coefficient
- $\bar{x}$ is the mean of the explanatory variable (x)
- $\bar{y}$ is the mean of the response variable (y)
- s_x is the standard deviation of the explanatory variable (x)
- s_y is the standard deviation of the response variable (y)

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.

3C: Performing a regression analysis:

Coefficient of determination (r^2): calculated by squaring the r value. It is turned into a percentage (×100) then interpreted. Use the statement by inputting the variable names and percentages:

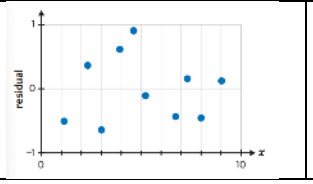
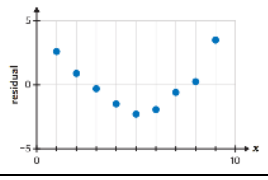
- **r^2 %** of the variation in the **RV** can be explained by the variation in the **EV**. The remaining **%** can be explained by other factors.

Residuals: residuals are the vertical distances between the data point and the LSRL.

$$residual = actual\ data\ value - predicted\ data\ value$$

- Actual value: found in the question/table of data
- Predicted value: must use the LSRL to predict the RV from the EV
- Positive residual = data point above LSRL, negative residual = data point below LSRL, zero residual = data point on the LSRL.

Residual plots:

Random scattering of points in the residual plot, supports the assumption of linearity	Clear curved pattern in the residual plot, does not support the assumption of linearity
	

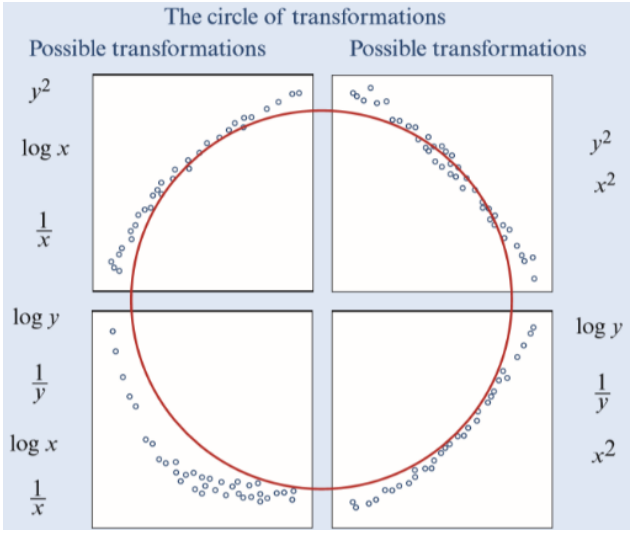
3D: Data Transformations:

You shouldn't perform a linear regression analysis for data that is nonlinear. Therefore, nonlinear data is **transformed.**

- Transformation linearise data so that regression analysis can be performed accurately.
- Match the nonlinear scatterplot with one in the diagram to help you determine the best transformation.
- **Best transformation:** strongest r/r^2 value

Types of transformations:

- Log: compresses the data
- Square: stretches the data
- Reciprocal: compresses values greater than 1, stretches values less than 1.



3E: Data transformations – applications

LSRL: once you have transformed your data you must create a new LSRL equation and include the transformation in the rules.
Eg. From

$$y = -16.14 + 9.39x$$

To

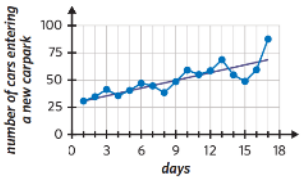
$$y = -0.73 + 1.05x^2$$

Making predictions: the limits of extrapolation are still present. When calculating, use solve as this will undo the transformation for you.

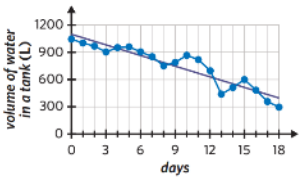
4A: Time series data and their graphs

Trends: general upwards (increasing) or downwards (decreasing) movement over time. Trend lines can be fitted directly to trends. There can be multiple trend lines.

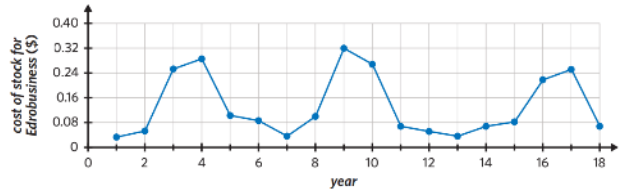
Increasing trend



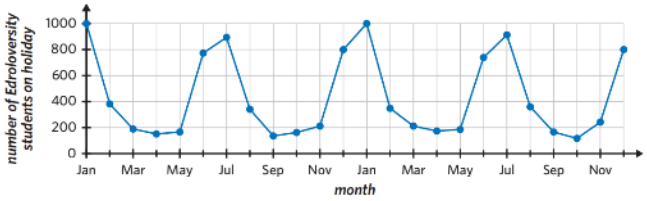
Decreasing trend



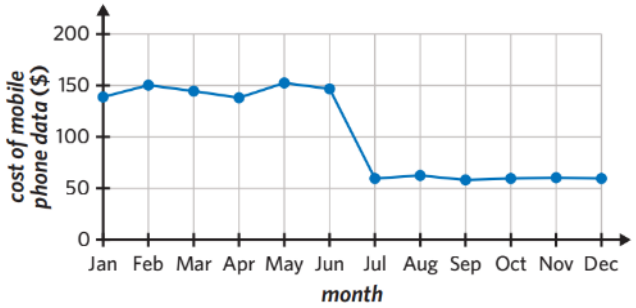
Cycles/cyclical variation: periodic movements over a period greater than 1 year. Peaks of cycles occur at approximately the same intervals, cycles can have a period which changes slightly between peaks.



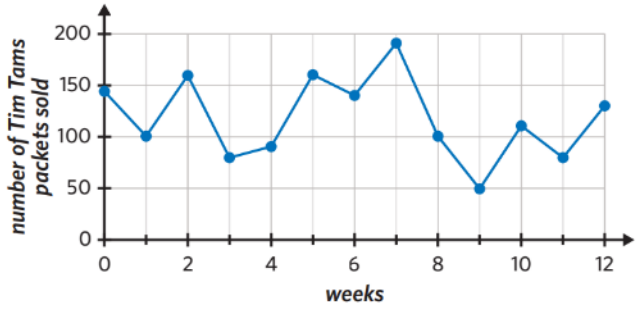
Seasonality: cyclical variation within a calendar-related period (week, month, quarter). A seasonal time series plot has regular peaks and troughs that occur at the same time each period and the length of the period must be a year or less.



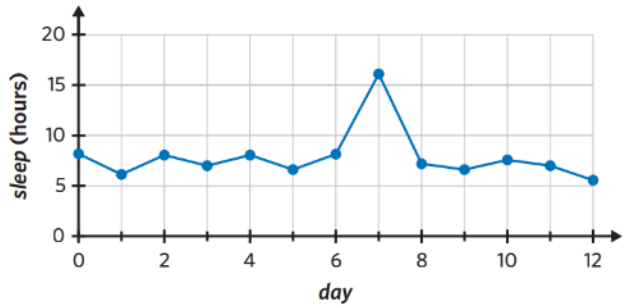
Structural Change: When an established pattern is suddenly altered. The graph then continues on the same level post structural change.



Irregular fluctuations: random variations that cannot be explained by trend, seasonality, cycles or structural change.



Outliers: stands out from the general body of data. It then returns to follow the original pattern/trend



4B: Smoothing – moving means

Smoothing: evens out fluctuations to help identify any underlying trends

- Only smooth the RV
- The larger the mean smooth, the more effective (5 more effective than 3)

3 mean: use 3 values and find the mean

5 mean: use 5 values and find the mean

- Always centred around the value you are trying to smooth

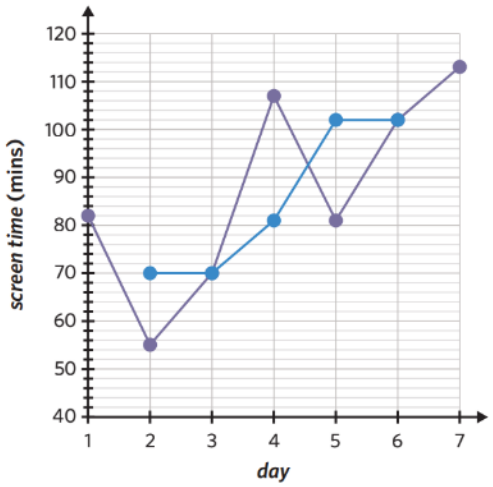
day	*temp.* (°C)	*calculation*	three-mean smoothed *temperature* (°C)
Mon	24	–	–
Tue	27	$\frac{24+27+21}{3}$	24
Wed	21	$\frac{27+21+18}{3}$	22
Thu	18	$\frac{21+18+15}{3}$	18
Fri	15	$\frac{18+15+15}{3}$	16
Sat	15	$\frac{15+15+12}{3}$	14
Sun	12	–	–

Smoothing with centring: an additional step when smoothing with an even number of points. Finding the mean of two non-centred means.

day	*temp.* (°C)	*before centring*	*after centring*
Mon	24		–
		$\frac{24+27}{2} = 25.5$	
Tue	27		$\frac{25.5+24}{2} = 24.75$
		$\frac{27+21}{2} = 24$	
Wed	21		$\frac{24+19.5}{2} = 21.75$
		$\frac{21+18}{2} = 19.5$	
Thu	18		–

4C Smoothing – moving medians

- Smoothed directly on the graph
- Median smoothing only uses and odd number of points
- Smooth the RV



4D: Seasonal adjustments:

- Seasonal fluctuations exist.
- Seasonal indices (SI) are used to de-seasonalise the data to minimise the effects of seasonality. This allows trends to be more easily observed.

Rules

1. $Seasonal\ index\ (SI) = \frac{value\ for\ season}{seasonal\ average}$
2. $Seasonal\ average\ (SA) = \frac{sum\ of\ all\ seasons}{number\ of\ season}$
3. $Deseasonalised\ figure\ (DS) = \frac{value\ for\ season}{seasonal\ index}$
4. Reseasonalising data:

$$value\ for\ season = deseasonalised\ figure \times seasonal\ index$$

How to interpret a seasonal index:

$$(seasonal\ index - 1) \times 100 = \_\_\_\_\_\%$$

- A negative %: (season) is below the seasonal average by ___%
- A positive %: (season) is above the seasonal average by ___%

Correcting for seasonality:

$$(\frac{1}{seasonal\ index} - 1) \times 100 = \_\_\_\_\_\%$$

- A negative %: To correct (season) for seasonality, (unit) need to be decreased by ____%
- A positive %: To correct (season) for seasonality, (unit) need to be increased by ____%

Notes:

- The sum of the seasonal indices is equal to the number of seasons (if you are working with months of the year there are 12 seasons and therefore the seasonal indices will sum to 12)
- If there were no fluctuations, the seasonal average is 1

4E: Time series data and LSRL modelling:

Trend lines: can be fitted to time series plots if there appears to be an increasing or decreasing trend.

- The LSRL is used
- If seasonality is present, data needs to be deseasonalised first before fitting the LSRL to the deseasonalised values

Forecasting: making a prediction for the future

- You need to re-seasonalise the value if the prediction was made from a deseasonalised LSRL

Which Graph? Bivariate data

Types of Variables			
Response	Explanatory	Which Graph?	Statistics
Categorical	Categorical	Segmented bar chart, parallel bar chart	Mode Percentages
Numerical	Categorical	Parallel box plot, parallel dot plot	5 number summary – use the median to make comparisons. Mean- remember it is influenced by outliers and skew. Standard deviation. Range and IQR. Percentages – use your understanding of boxplots and 25% per section to make comparisons. Shape. Outliers and the upper and lower fence.
Numerical	Categorical (two categories only)	Back-to-back stem plot, parallel dot or box plot	
Numerical	Numerical	Scatterplot	A full regression analysis involves several processes, which include: • Constructing a scatterplot • Calculating the correlation coefficient to indicate the strength of the relationship • Determining the equation of the regression line • Interpreting the coefficients, the intercept (a) and the slope (b) of the least squares line $y = a + bx$ • Using the coefficient of determination to indicate the predictive power of the association • Using the regression line to make predictions • Calculate residuals and use a residual plot to confirm the assumption of linearity • Write a report on your findings

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].**

2h. The appropriateness of fitting a transformed regression analysis

Transformations are **[appropriate/ not appropriate]** as the original data produces a **[non-linear/ linear]** scatterplot, confirmed with a clear **[curved pattern/ random scatter]** in the residual plot.
[Log [EV]/Reciprocal [EV]/[EV]Squared/ Log[RV]/Reciprocal[RV]/[RV]Squared] is the most appropriate transformation, given it has the strongest coefficient of determination r^2 of **[r^2 x 100]** %, compared with **[r^2 x 100]** % and **[r^2 x 100]** %.

2i. A residual analysis on the number

The residual value is **[positive/ negative], [above/ below]** the LSRL or the LSRL is an **[under prediction/ over prediction]**.

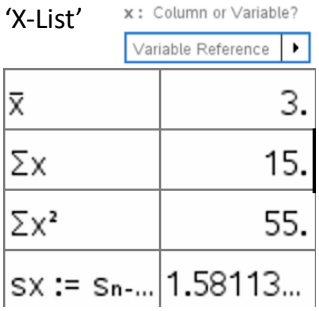
2j. Confounding variables

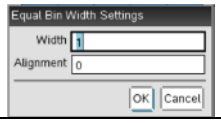
Since the correlation coefficient **[r value]** indicates a **[strong/moderate/weak]** association. Therefore, factors (confounding variables) other than the **[explanatory variable]** must be considered. Some of these factors include **[list potential reasons here]**.

2k. Reason for smoothing Time Series Plot

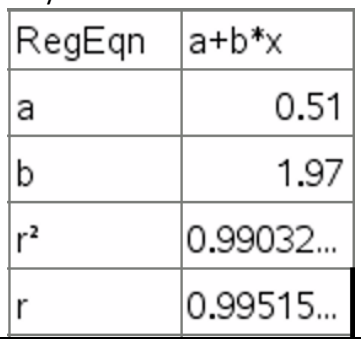
Reduce Random Fluctuations; Highlight the Underlying Trend; Identify Seasonal Patterns; Prepare for Forecasting; Outlier Detection.

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’


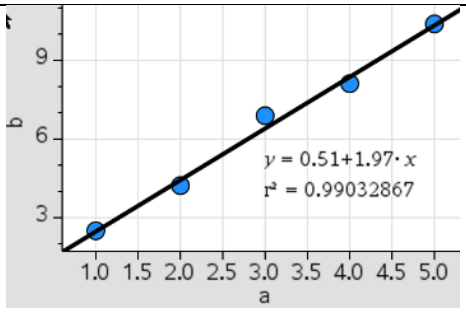
</td>
</tr>
<tr>
<td>1.2 Graphing Univariate Data: dot plot, boxplots, histograms</td>
<td>
1. Ctrl n > 4, Enter data in ‘list and spreadsheets’ (ensure a title)
2. Ctrl doc OR ctrl i OR home menu > 5 to open a new page and select ‘data and statistics’
3. Click to add data to the x-axis
4. Menu > 1: 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’

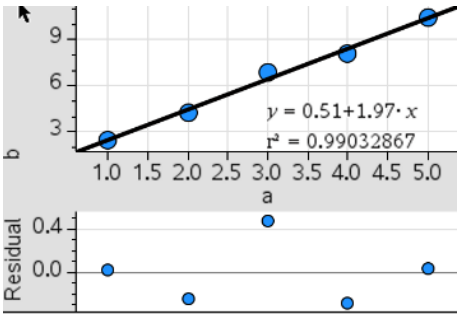
</td>
</tr>
<tr>
<td>3.2 Scatterplots</td>
<td>
1. Ctrl n > 4, Enter data in ‘list and spreadsheets’ (ensure you give both columns a title)
2. Ctrl doc OR ctrl i OR home menu > 5 to open a new page and select ‘data and statistics’
3. Click to add the EV to the x-axis and the RV to the y-axis

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



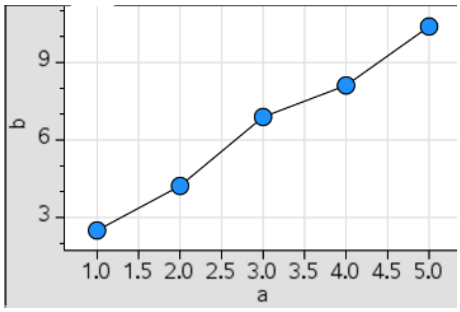
Residual Plot:

4. Menu 4 > 7 > 2



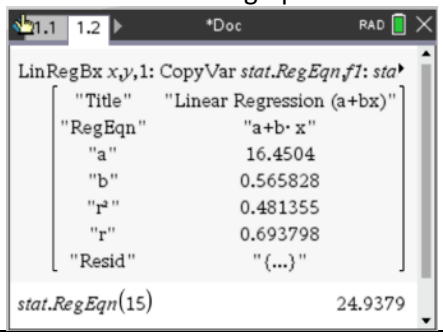
Timeseries line:

5. Menu > 2 > 1



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 ↵



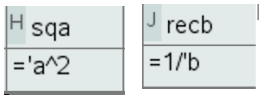
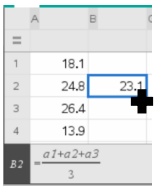
3.4 Transformations

1. Enter data in 'list and spreadsheets' (ensure you give both columns a title)
2. In a new column title the transformation you are completing by using:

Square transformation	sq(variable name)
Log transformation	log(variable name)
Reciprocal transformation	rec(variable name)

3. In the second row of this column (grey shaded) you are going to put the transformation in by:

Square transformation	=(variable name)^2
Log transformation	=$\log_{10}$(variable name)
Reciprocal transformation	=1/(variable name)

	 4. From here you can get the 'Bivariate Data statistics' and create a 'scatterplot' – remember that you only use one transformation at a time. So if you transform x you use the original y when calculating statistics and graphing (vice versa).
4.1 Mean Smoothing	1. Enter values in: time for column A, RV data for column B. 2. If 2 or 3 smoothing start in column C cell 2. If 4 or 5 smoothing start in column C cell 3. If 6 or 7 smoothing start in column C cell 4. If 8 or 9 smoothing start in column C cell 5….. 3. Write a. 2 mean =(b1+2b2+b3)/4 b. 3 mean =(b1+b2+b3)/3 c. 4 mean =(b1+2b2+2b3+2b4+b5)/8 d. 5 mean =(b1+b2+b3+b4+b5)/5 e. 6 mean =(b1+2b2+2b3+2b4+2b5+b6)/12 f. 7 mean =(b1+b2+b3+b4+b5+b6+b7)/7 or g. 3 median =median(b1:b3) h. 5 median =median(b1:b5) i. 7 median =median(b1:b7) j. 9 median =median(b1:b9)  4. Menu > 3 > 3, drag and it will fill the values. 5. Comparing time series plots: Menu > 2 > 8
4.2 Seasonal indices	1. Define seasonal data ctrl [ ], displaying multiple data by using curly brackets ctrl ( ), comma in between. $y1:=\{3051,8430,12340,4302\}$ define sales $y1 \quad \{3051.,8430.,12340.,4302.\}$ $\text{mean}(y1) \quad 7030.75$ yearly average $\frac{y1}{\text{mean}(y1)} \quad \{0.433951,1.19902,1.75515,0.611884\}$ $\text{round}\left(\frac{y1}{\text{mean}(y1)},2\right) \quad \{0.43,1.2,1.76,0.61\}$ seasonal indices $\frac{\frac{w1}{\text{mean}(w1)}+\frac{w2}{\text{mean}(w2)}}{2}$ 0.548733,1.2271,1.28012,1.61988,1.20117 AVRG S.I for 2 weeks $s:=\{0.7,1.35,0.8,0.75,1.2,1.7,0.5\}$ $\{0.7,1.35,0.8,0.75,1.2,1.7,0.5\}$ seasonal index $sp:=\{7500,14000,6000,8000,13000,17000,4$ $0.,14000.,6000.,8000.,13000.,17000.,4000.\}$ actual value $\frac{sp}{s}$ 10714.3,10370.4,7500.,10666.7,10833.3,10 deseasonalised fig.
4.3 Solve	2. Calculator Page, Menu > 3 > 1 3. There must be an equals sign (=), must choose a variable (x) and use the same variable at the end and do 'comma variable' (,x) $\text{solve}(10=x+2,x) \quad x=8.$
4.4 Back to top of List	Ctrl > 7

1A Data Types

numerical
- quantities
- numbers

categorical
- characteristics
- qualities

discrete
- whole numbers
- can be counted
- # students in each class
- # cars in a car park
- # siblings

continuous
- decimal numbers
- can be measured
- height
- weight
- temperature
- amount of rain

nominal
- group according to characteristics
- hair/eye colour
- favourite sport
- fare animal

ordinal
- can be grouped and ordered
- maccas fries Size (S-m-L)
- months of the year

1B – displaying and describing categorical data

- count frequency: # of times a category appears in the data

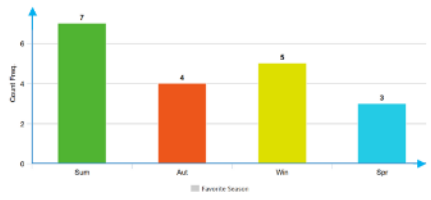
- percentage frequency $= \frac{\text{count frequency}}{\text{total count}} \times 100$

- mode: most frequently-occurring value

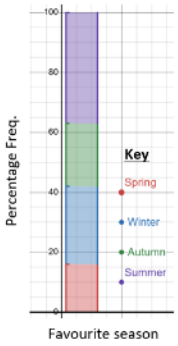
Examples: What is your favourite season?

	summer	autumn	winter	spring
tally	𝍸 ||	||||	𝍸	|||
count	7	4	5	3
%	$\frac{7}{19}$ x 100 = 37%	$\frac{4}{19}$ x 100 = 21%	$\frac{5}{19}$ x 100 = 26%	$\frac{3}{19}$ x 100 = 16%

a. create a bar chart using count freq.
***bars do not touch!! ***



b. create segmented bar chart using % freq.



c. describe the distribution:

1) what is the data type and total number of data?

2) what is the mode/s?

3) compare percentages.

1) Favourite season is categorical nominal data; there were 19 pieces of data.

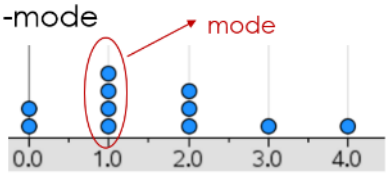
2) The most popular season/mode is summer.

3) 37% of the class preferred summer, next was winter at 26%, then autumn at 21%, and the least favourite was spring at 16%.

1C Displaying Numerical Data

Dot Plot:

-simple
-small data sets
-mode



siblings

Stem Plots:

-numerical order
-leading digits in stem
-trailing digit in leaf
-key!!
-class intervals: helps keep leaves small → a.k.a "stem intervals"

Eg 1

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

Key = 0 | 2 = 2
or 0 | 2 = 0.2
0 | 2 = 2%

Eg 2: class intervals

15	1 1 4
15	5 6 7 8 9
16	0 1 2 3 4
16	5 6 7 8 9
17	0 2
17	8 8

**class intervals of 5*
First interval 0–4
second interval 5–9

Key = 15 | 4 = 154

Grouped Frequency Tables

-intervals
-intervals don't overlap
-lower bound is inclusive
-turns into histogram

Eg: Plant height data:

32.0 40.2 40.5 45.1 47.0 49.1 50.1 53.7 54.2 55.3 56.9 57.2 58.2 67.2 68.9 69.0 72.3 77.6 82.1 88.5

grouped frequency table:

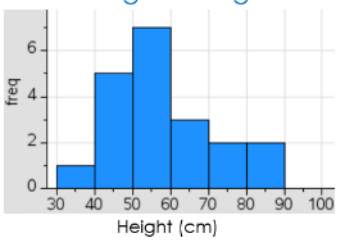
height (cm)	# freq	% freq
30 – < 40	1	$\frac{1}{20}$ x 100 = 5%
40 – < 50	5	25%
50 – < 60	7	35%
60 – < 70	3	15%
70 – < 80	2	10%
80 – < 90	2	10%
	20	100%

lower bound
• x axis on histogram.

Histogram:

-no spaces between bars
-displays numerical continuous data

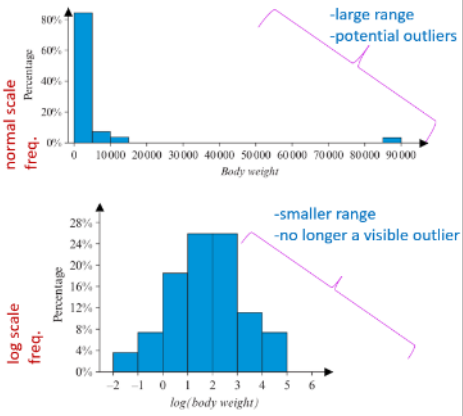
Eg: Plant height Histogram



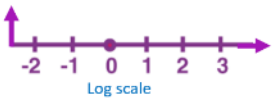
1D Logs → we only use $\log_{10}$ base 10!

purpose: using a log scale compresses the data, making a more even spread of data.

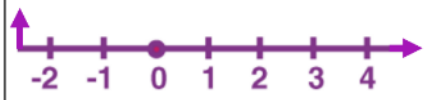
eg:



understanding the log scale:



to undo a log scale, you do $10^{\square}$ → power of the scale



10^{-2}	10^{-1}	10^{0}	10^{1}	10^{2}	10^{3}	10^{4}
0.01	0.1	1	10	100	1000	10000

↓ these numbers are what a log scale represents!

calculating a log value

eg: what is the log value of 12,456 and 0.82

a. $\log_{10}(12456) = 4.1$
b. $\log_{10}(0.82) = -0.086$

calculating the value from a log scale

eg: what is the value for a log of 4.7 and −3.2

a. $10^{4.7} = 50118.72$
b. $10^{-3.2} = 0.00063$

*$\log_{10}(x) = y \leftrightarrow 10^{y} = x$ *

Convert to log form: $100 = 10^{2}$ $\log_{10}100 = 2$

IE: 5 number summary + boxplots

5 number summary and statistics:
minimum, Q1, median, Q3, maximum

median = middle number (Q2)

↓ to locate position of median: $\frac{n+1}{2}$

range = maximum − minimum

Interquartile range (IQR) = Q3 − Q1

identifying outliers:

– outliers are extreme values, that fall outside of what is "reasonable"

– calculate fences to determine outliers

• outliers can be the min or max value *

upper fence = Q3 + 1.5 × IQR

lower fence = Q1 − 1.5 × IQR

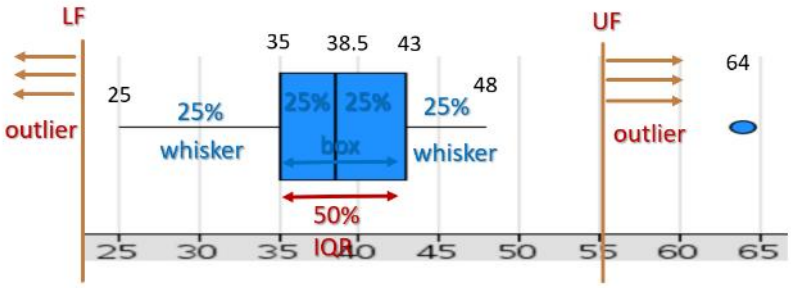
↓ a data value greater than the upper fence or lower than the lower fence is an outlier.

constructing a boxplot:

example: 25, 30, 31, 34, 35, 35, 35, 36, 38, 39, 40, 40, 42, 43, 44, 46, 48, 64

(5th, 9th, 10th, 5th)

Q1 Q3



min 25
Q1 35
med 38.5
Q3 43
max 64

UF = 43 + 1.5 × (43 − 35) = 55

LF = 35 − 1.5 × (43 − 35) = 23

$\frac{n+1}{2} = \frac{18+1}{2} = 9.5$ → 9th – 10th data point → $\frac{38+39}{2} = 38.5$

$\frac{9+1}{2} = 5^{th}$

IF Describing Numerical Data.

* when analysing histograms, dot plots, boxplots and stem plots we describe:

- centre : median value histograms – approx. interval

- spread : range = max – min IQR = Q3 – Q1 (if outliers present)

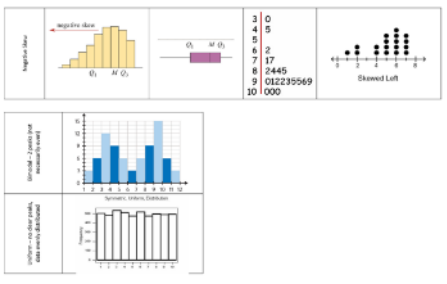
- shape : symmetrical
  approx. symmetrical
  positive skew
  negative skew
  outliers (mention)

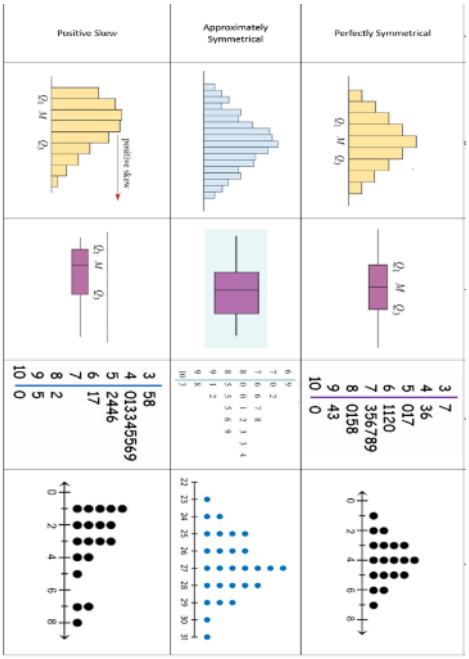
centre

mean or median??

- both measure centre
- mean is affected by skew/outliers,

- in symmetrical distributions with no outliers the mean or median can be used to describe centre
- in skewed distributions with or without outliers the median is used to describe centre.





1G : Introduction to standard deviation

mean: "average"

- measure of centre
- $\bar{x}$ → "x bar"
- mean = $\frac{\text{sum of all data points}}{\text{number of data points}}$ } $\bar{x} = \frac{\Sigma x}{n}$
- population: entire groups data
- sample: smaller subset of population

standard deviation:

- another measure of spread
- how far away each data point is from the mean
- Sx
- calculate using CAS

example:

1. calculate the mean for the following weights of 10 rugby players:

80, 95, 85, 91, 102, 93, 87, 78, 84, 90

$$\bar{x} = \frac{80 + 95 + 85 + 91 + 102 + 93 + 87 + 78 + 84 + 90}{10}$$

$$\bar{x} = 88.5$$

2. Homer was embarrassed about his weight loss results and would not tell anyone. 14 other men compiled their results, and the data is shown below. Given that the **mean weight lost** was **3.8 kg**, how much weight did Homer lose over the month? *(2 MARKS)*

Weight lost (kg)	5	5.2	3.8	3.7	2.5	2.9	4.8	5.6	4.7	3.8	3.6	2.4	4.1	4.2	??

$$\frac{5 + 5.2 + 3.8 + 3.7 + 2.5 + 2.9 + 4.8 + 5.6 + 4.7 + 3.8 + 3.6 + 2.4 + 4.1 + 4.2 + x}{15} = 3.8$$

solve $\left(\frac{56.3+x}{15} = 3.8, x\right)$

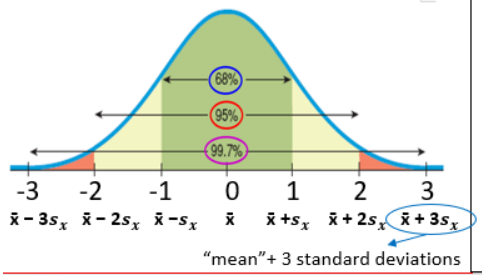
$$x = 0.7$$

1H: Normal Distribution

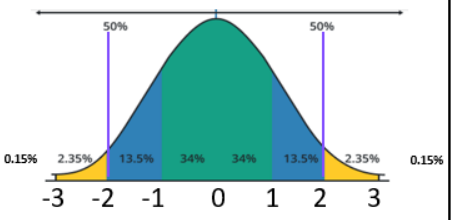
normal distribution = symmetrical (approx.) distributions, "bell-shaped"

- numerical data sets
- mean and median are both appropriate
- mean and standard deviation are used

68–95–99.7% rule

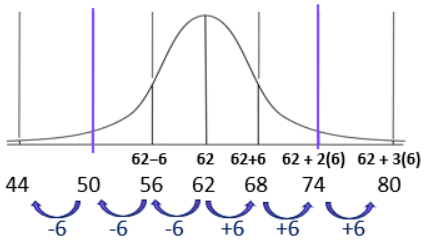


Section breakdown



example: New plants are planted in the TC gardens and are measured after a year. There are **50 plants**, the **mean is 62 cm** & a **standard deviation of 6 cm.**

a. construct a bell curve:



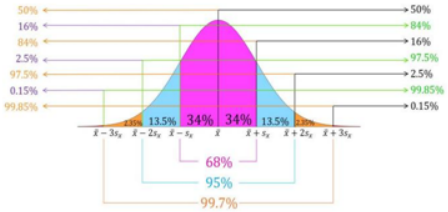
b. what % of plants are smaller than 68cm?

34 + 34 + 13.5 + 2.35 + 0.15 = **84%**

c. how many plants are between 50 and 74cm?

13.5 + 34 + 34 + 13.5 = **95%**

$\frac{95}{100}$ × 50 = **47.5 ≈ 48 plants**



11. z - scores

- "standardised score"
- provides a precise measure of the location of each data point in the normal distribution curve.

calculating z scores:

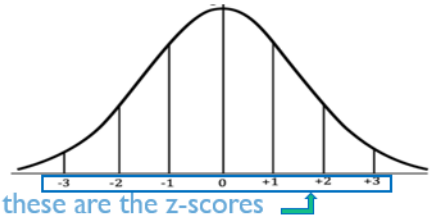
$z = \frac{x-\bar{x}}{s_x}$

if z is:
+ above mean
− below mean
0 same as mean

z = standardised score
x = actual data point
$\bar{x}$ = mean
s_x = standard dev.

calculating actual scores:

$x = \bar{x} + (z \times s_x)$



these are the z-scores

examples:

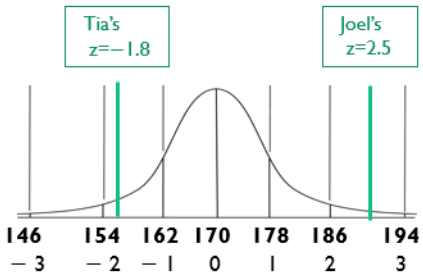
1. The average height of this class is **170cm**, the standard dev. is **8cm**.

a. Joel is 190cm tall, what is his standardised score?

$z = \frac{x-\bar{x}}{s_x} = \frac{190-170}{8} = \mathbf{2.5}$

b. Tia's z-score is −1.8, what is her actual height?

x = 170 + (−1.8 × 8)
= **155.6 cm**



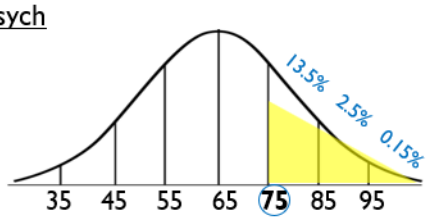
2. Which subject did Oliver do better in?

(determine the z-scores and interpret)

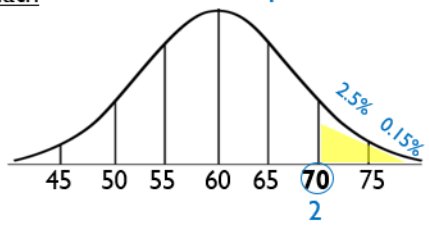
subject	mark	class mean	s_x
Psych	75	65	10
math	70	60	5

Oliver did better in maths (relative to the class) as the z-score is 2 compared to 1 in Psych.

Psych



Math



2A Associations between two categorical variables

bivariate data = **two** bits of data

two-way frequency tables:

- allow two sets of categorical data to be compared
- columns = EV
- rows = RV
- percentages are used when sample sizes of data are different

describing the association between two categorical variables:

- analyse for an association or pattern
- percentages to support findings

displaying:

grouped bar chart:

- displays 2 categorical variables
- frequency on vertical (y) axis
- EV on horizontal (x) axis
- RV represented by the columns (key)

percentage segmented bar charts:

- gap between columns
- each bars height is 100%
- frequency on vertical (y) axis
- EV on horizontal (x) axis
- RV represented by the columns (key)

example: A group of people were surveyed as to their chocolate preference (dark, milk, white), their age was also recorded

a. complete the two-way frequency tables:

number

Choc preference	20–29	30–39	40–49	50–59
dark	5	20	22	30
milk	i. 21	18	13	10
white	21	3	4	1
total	47	ii. 41	39	41

age

i. 47 − 21 − 5 = 21
ii. 20 + 18 + 3 = 41

percentage

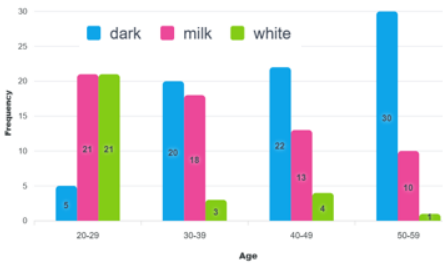
Choc preference	20–29	30–39	40–49	50–59
dark	11%	i. 49%	56%	73%
milk	45%	ii. 44%	33%	24%
white	45%	iii. 7%	10%	2%
total	101%	100%	99%	99%

age

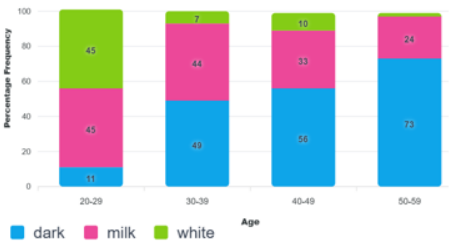
* when rounding to the nearest whole % it won't always add to 100%! *

i. 20 ÷ 41 × 100 = 49%
ii. 18 ÷ 41 × 100 = 44%
iii. 3 ÷ 41 × 100 = 7%

b. display as a grouped bar chart:



c. display as a % segmented bar chart:



d. determine if there is an association between chocolate preference and age:

Yes, there is an association. As age increases preference for dark chocolate increases (11%, 49%, 56%, 73%). Preference for milk (45%, 44%, 33%, 24%) and white (45%, 7%, 10%, 2%) decreases.

2B : Association between numerical and categorical variables

- compares distribution of 2 or more categorical variables
- if distributions differ an association between numerical and categorical may exist.

displaying:

- back-to-back stem plot
- parallel dot plots
- parallel boxplots

describing: use comparative language (eg: larger/smaller)

- shape: positive/negative skew, approx. symmetrical, outliers
- centre: median
- spread: range or IQR if outliers are present

examples:

1. back-to-back stem plot: number of points scored by two bball teams

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

median position $= \frac{14+1}{2} = 7.5$ (between 7th and 8th point)
croc: 44
zebras: $\frac{69+70}{2} = 69.5$

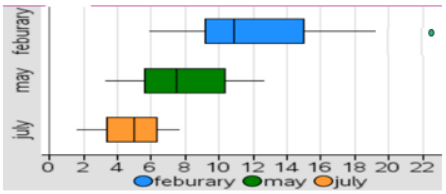
a. compare the results:

Team Zebra was the more successful team, their results are negatively skewed and have a median of 69.5 points, this is larger than the crocodiles median score of 44 and their data is positively skewed. The crocodiles' results are more varied with a range of 60, compared to 47 for the zebras.

2. parallel boxplots

The five-number summary for the distribution of minimum daily temperature for the months of February, May and July in 2017 is shown in the table. The associated boxplots are shown following the table.

month	minimum	Q1	median	Q3	maximum
February	5.9	9.5	10.9	13.9	22.2
May	3.3	6.0	7.5	9.8	12.7
July	1.6	3.7	5.0	5.9	7.7



Explain why the information given supports the contention that minimum daily temperature is associated with the month. Refer to the values of an appropriate statistic in your response. (2 MARKS)

Minimum daily temperature is associated with month as the median temperature decreases from Feb to May to July from 10.9 to 7.5 to 5.0. The range/variability is larger in Feb, 12.7°C, compared to 6.7 and 4.0 (May and July). Feb is positively skewed with an outlier, May is slightly positively skewed (approximately symmetrical and July is approximately symmetrical.

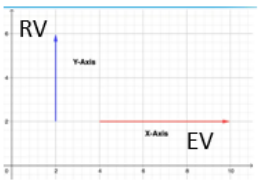
2C : association b/w 2 variables: scatterplots.

bivariate data: two variables are analysed to determine an association

EXPLANATORY VARIABLE (EV): can be used to predict or explain the changes observed in the RV.

RESPONSE VARIABLE (RV): may be explained or predicted by changes in the EV

" EV explains the RV "



eg:

" temp explains the time of day " ✗

OR

" time of day explains the temp " ✓

constructing scatterplots:

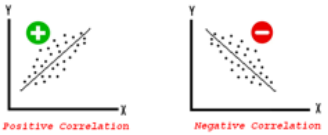
see CAS notes sheet

describing relationships b/w 2 variables / scatterplots:

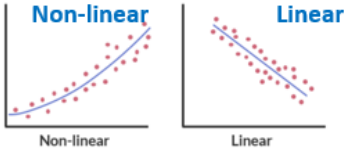
strength: weak, moderate or strong
- weak = wide band of points
- strong = narrow band of points

See hand out

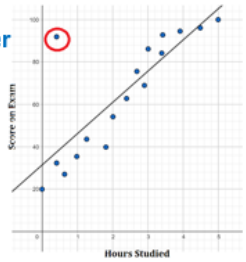
direction: positive or negative
- positive = RV ↑ as EV ↑
- negative = RV ↓ as EV ↑



form: linear or non-linear
- outliers: a point away from rest of graph



outlier



2D. Pearsons's correlation coefficient

- a number that describes the strength of an association

r

assumptions:

- data is linear
- numerical
- no outliers are present

see hand out!

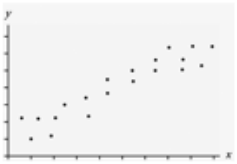
$0.75 \le r \le 1$ Strong, positive, linear association
$0.5 \le r < 0.75$ Moderate, positive, linear association
$0.25 \le r < 0.5$ Weak, positive, linear association
$-0.25 < r < 0.25$ No association
$-0.5 < r \le -0.25$ Weak, negative, linear association
$-0.75 < r \le -0.5$ Moderate, negative, linear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

correlation and causality

- just b/c 2 variables have a high correlation; it doesn't mean that one causes the change in the other

some explanations:

1. **common cause:** a third variable that explains the correlation
   eg # ppl wearing sunscreen and fainting
   the third variable could be: temperature
   → actually effecting both variables



2. **confounding variable:** another variable that impacts the others
   eg plant height and water intake: water intake effects plant growth
   but so do... sun, soil, bugs, oxygen, season, temperature
3. **coincidence:** two things correlate but have no relation to each other. Pure chance. No logical explanation

3A LSRL: least squares regression line

equation $y = a + bx$ (slope, RV, y-intercept, EV)

- a LSRL minimises the sum of the squared values of the residuals
- a "line of best fit"

residual: vertical distance b/w the straight line and any given point on the scatterplot

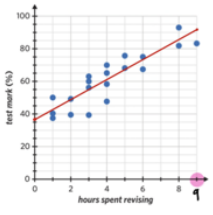
assumptions made when fitting LSRL:

- data is numerical
- the relationship b/w the variable is linear
- there are no clear outliers

CAS: follow CAS cheat sheet

ctrl N 4 enter data ctrl i 5
EV → x-intercept RV → y-intercept
Menu 4 6 2 (9, 90.9)

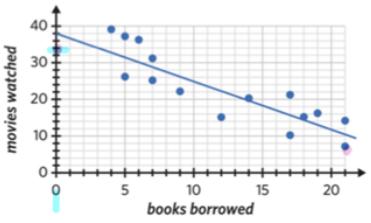
sketching LSRL by hand



(0,36)

test mark = 36 + 6.11 × hours revising
(where the line starts) (0, 36)
- to find a second point, sub in the last number on the x-axis, in this case 9, to the LSRL
36 + 6.11 × 9 = 90.9
∴ the second point is (9, 90.9)
→ join the two dots w a ruler

determining LSRL from a graph:



- need to determine the value of the y-intercept (a) and the slope (b)
- y-int: where the line cuts the y-axis when x = 0

when x = 0, y-int = 38 ∴ a = 38

- slope: gradient / steepness of the graph
→ choose two points on the line you can clearly read the coordinates
(0, 38) and (21, 10)

$$b = \frac{\text{rise}}{\text{run}} = \frac{y_2 - y_1}{x_2 - x_1} = \frac{10 - 38}{21 - 0} = \frac{-28}{21} = -1.33$$

∴ b = −1.33

LSRL: movies watched = 38 − 1.33 × books borrowed

Calculating LSRL from summary statistics:

$y = a + bx$ $\quad b = r * \frac{s_y}{s_x}$ $\quad a = \bar{y} - b * \bar{x}$

r = pearsons $\quad \bar{y}$ = mean of y
s_Y = standard dev. y $\quad \bar{x}$ = mean of x
s_x = standard dev. x

Example: determine LSRL from the following statistics:
r = 0.845 $\bar{x}$ = 11 $\bar{y}$ = 29.2 s_x = 6.06 s_Y = 16.8

$b = 0.845 \times \frac{16.8}{6.06} = 2.34$

$a = 29.2 - 2.34 \times 11 = 3.43$

$\therefore y = 3.43 + 2.34 \times x$

* You can be asked to find r, y bar, x bar, s_Y, and s_Y also-fill in the equation and use solve *

3B Interpreting + making predictions:

LSRL:

$y = a + bx$

RV y-int slope EV

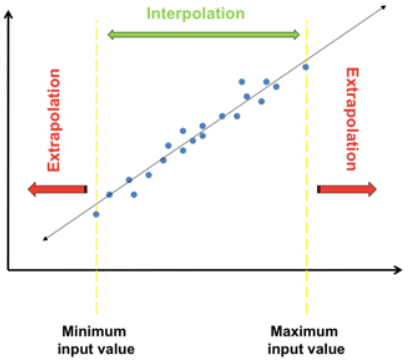
interpret statements:

y-intercept (a): when the *EV* EV is 0, the *RV* is *a*

slope (b): for every one-unit increase in the *EV*, the *RV* increases/decreases by *b*

if b is + choose increase

if b is − choose decrease



fill variable names and a and b values!

making predictions:

* 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

$y = a + b\,x$

predict for RV sub EV value here

3C: performing regression analysis

coefficient of determination (r^2) COD

- calculated by squaring r
- turn it into a percentage then interpret (×100)

interpret statement:

r^2 % of the variation in the *RV* can be explained by the variation in the *EV*

The remaining *(%)* can be explained by other factors.

(100 − r^2 %)

- *input variable names and percentages*

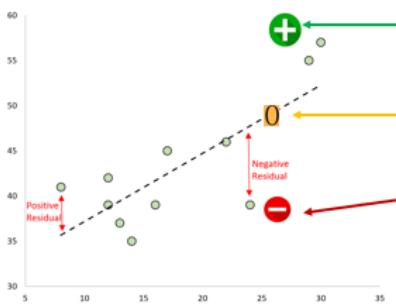
residuals

- residuals are the vertical distances between the data point and the LSRL

residual = actual − predicted
value value value

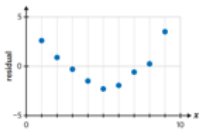
found in table of data/the question (RV value)

must use LSRL to predict the RV from the EV

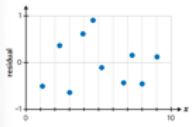


- data point above LSRL is positive residual
- data point on the LSRL is zero residual
- data point below LSRL is negative residual

residual plots



clearly curved pattern = non-linear association



random scattering = linear association

3D: Data Transformations

– you shouldn't perform linear regression analysis for data that is non-linear

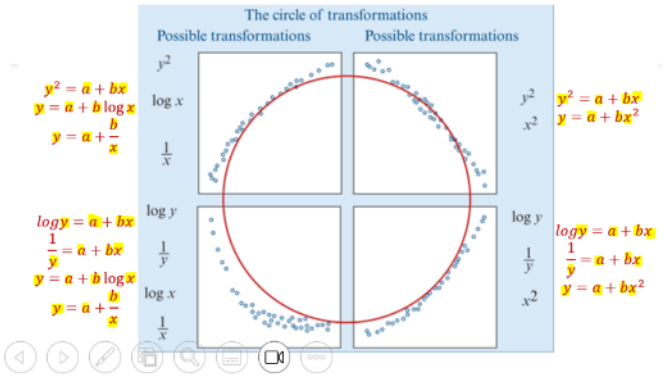
– ∴ nonlinear data is transformed

– transformations linearise data so that you can accurately perform regression analysis

– match the nonlinear scatterplot with one in the diagram to help you determine the best transformation

strongest r value (closest to +1 or −1)

closest r^2 % to 100%



types of transformations:

log: compresses the data $\log_{10}(x)$ or $\log_{10}(y)$

square: stretch the data x^2 or y^2

reciprocal: compresses values greater than 1, stretches values less than 1 $\frac{1}{x}$ or $\frac{1}{y}$

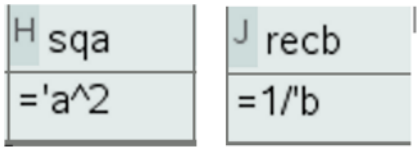
CAS see CAS reference sheet.

1. Enter data in 'list and spreadsheets' (ensure you give both columns a title)
2. In a new column title the transformation you are completing by using:

Square transformation	sq(variable name)
Log transformation	log(variable name)
Reciprocal transformation	rec(variable name)

3. In the second row of this column (grey shaded) you are going to put the transformation in by:

Square transformation	=(variable name)^2
Log transformation	=$\log_{10}$(variable name)
Reciprocal transformation	=1/(variable name)



4. From here you can get the 'Bivariate Data statistics' and create a 'scatterplot' – remember that you only use one transformation at a time. So if you transform x you use the original y when calculating statistics and graphing (vice versa).

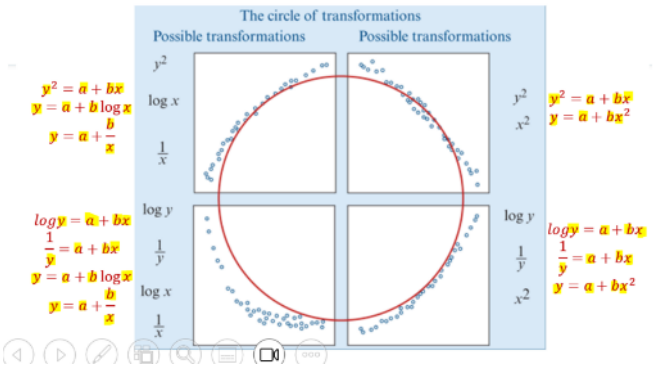
3E - data transformations - applications.

LSRL:

- once you have transformed your data you must create a new LSRL equation using the transformation!!

making predictions:

- the limits of extrapolation are still present.
- use solve - as this will do the transformation for you.



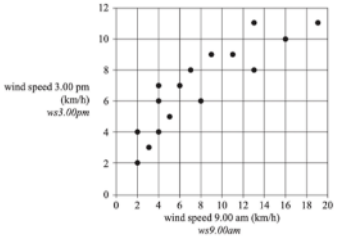
Example:

Apply the squared transformation to the variable ws3.00 pm and determine the equation of the least squares regression line that allows $(ws3.00\ pm)^2$ to be predicted from ws9.00 am. In the boxes provided, write the coefficients for this equation, correct to 3 significant figures.

$(ws3.00\ pm)^2 =$ 3.38 + 6.59 × ws9.00 am r^2=0.82

$(ws3.00\ pm) =$ 0.29 + 8.27 × log(ws9.00 am) r^2=0.84 ✓

$(ws3.00\ pm) =$ 10.2 + 16.7 / ws9.00 am r^2=0.74



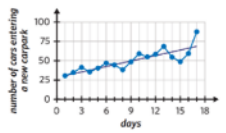
Wind speed (km/h) 9.00 am	3.00 pm
2	2
4	6
4	7
4	4
13	11
6	7
3	3
16	10
6	7
13	8
11	9
2	4
7	8
5	5
8	6
6	7
19	11
9	9

4A Time series data and their graphs

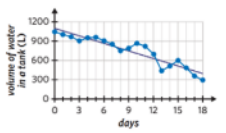
Characteristics of time series data

Trends: general upwards (increasing) or downwards (decreasing) movement over time. Trend lines can be fitted directly to trends. There can be multiple trend lines.

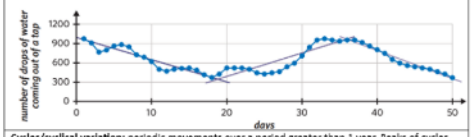
Increasing trend:



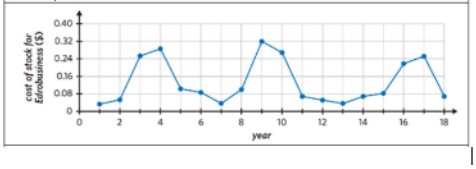
Decreasing trend:



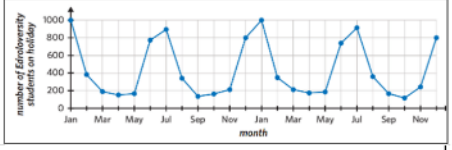
Sometimes, a time series may have multiple trends that change over time.



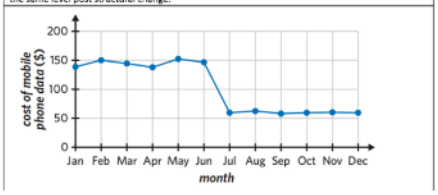
Cycles/cyclical variation: periodic movements over a period greater than 1 year. Peaks of cycles occur at approximately the same intervals. cycles can have a period which changes slightly between peaks.



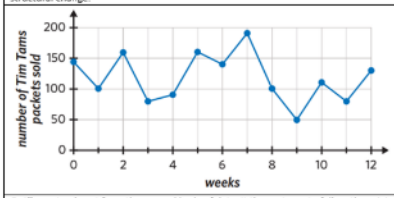
Seasonality: cyclical variation within a calendar-related period (week, month, quarter). A seasonal time series plot has regular peaks and troughs that occur at the same time each period and the length of the period must be a year or less.



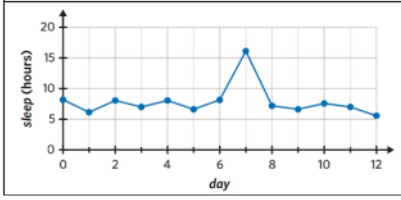
Structural Change: When an established pattern is suddenly altered. The graph then continues on the same level post structural change.



Irregular fluctuations: random variations that cannot be explained by trend, seasonality, cycles or structural change.



Outliers: stands out from the general body of data. It then returns to follow the original pattern/trend



4B: mean smoothing.

why smooth? to even out fluctuations and identify any underlying trends.

- ✓ only smooth the RV!
- ✓ larger means are more effective in smoothing (5 better than 3).

smoothing

- ✓ 3 mean = use 3 values, find the mean
- ✓ always centred around value trying to smooth.
- ✓ centering is an addition step when smoothing with an even number of points
- ✓ Centering rules

ODD 3

$$y_2 = \frac{y_1 + y_2 + y_3}{3}$$

EVEN 2

$$y_2 = = \frac{y_1 + y_2 + y_2 + y_3}{4}$$

ODD SMOOTHING (3,5,7…) example

day	M	T	W	T	F	S	S
temp	18.1	24.8	26.4	13.9	12.7	14.2	24.9

a. 3 mean smooth Tuesday:

$$\frac{M + (T) + W}{3} = \frac{18.1 + 24.8 + 26.4}{3} = 23.1$$

b. 5 mean smooth Thursday

$$\frac{T + W + (Th) + F + S}{5}$$

$$\frac{24.8 + 26.4 + 13.9 + 12.7 + 14.2}{5} = 18.3$$

c. 7 mean smooth Thursday

$$\frac{M + T + W + (Th) + F + S + S}{7}$$

EVEN smoothing 2,4,6… w/ CENTERING

day	M	T	W	T	F	S	S
temp	18.1	24.8	26.4	13.9	12.7	14.2	24.9

a. 2 mean smooth Tuesday

long way:

$$\frac{108.1 + 24.8}{2} = 21.45$$

$$\frac{24.8 + 26.4}{2} = 25.6$$

Centering

$$\frac{21.45 + 25.6}{2} = 23.5$$

short way:

$$\frac{M + T + T + W}{4} = \frac{18.1 + 24.8 + 24.8 + 26.4}{4} = 23.5$$

b. 4 mean smooth Thursday

$$\frac{T + W + W + Th + Th + F + F + S}{8} = \frac{24.8 + 26.4 + 26.4 + 13.9 + 13.9 + 12.7 + 12.7 + 14.2}{8} = 18.1$$

2 mean: like 3 mean, middle value used twice, divided by 4

4 mean: like 5 mean, middle values used twice, divided by 8.

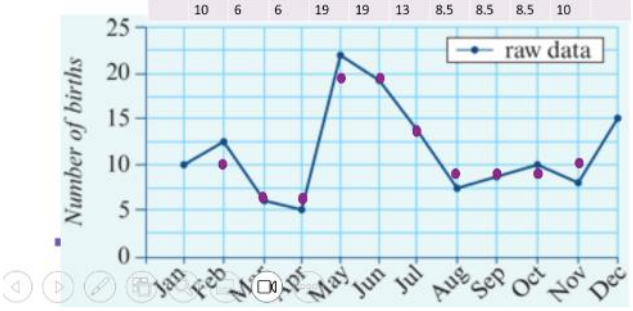
6 mean: like 7 mean middle values used twice, divided by 12.

4C: Smoothing – moving medians

Time Period	Raw Data	Three-point median	Five-point median
1	119		
2	97	97	
3	86	97	105
4	131	x	97
5	105	105	105
6	93	105	115
7	130	115	105
8	115	115	115
9	105	115	y
10	141	115	115
11	115	115	
12	107		

a three-median smoothed plot of the time series plot shown below.

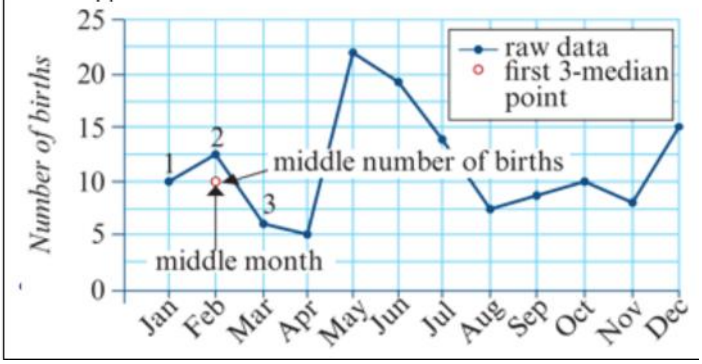
Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
10	12.5	6	5	24	19	13	7.5	8.5	10	7	15
	10	6	6	19	19	13	8.5	8.5	8.5	10	



a five-median smoothed plot of the time series plot shown opposite.

Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
10	12.5	6	5	24	19	13	7.5	8.5	10	7	15
		10	12.5	13	13	13	10	8.5	8.5		

Construct a three-median smoothed plot of the time series plot shown opposite.



4D Seasonal Indices:

example: Charlie owns an umbrella shop the following table shows his 2022 sales:

	summer	autumn	winter	spring
umbrella sales ($)	205	377	528	442

1. find the seasonal average for 2022

$$SA = \frac{205 + 377 + 528 + 442}{4} = 388$$

2. calculate the seasonal indices:

	summer	autumn	winter	spring
Seasonal indices	$\frac{205}{388} = 0.528$	$\frac{377}{388} = 0.972$	$\frac{528}{388} = 1.361$	$\frac{442}{388} = 1.139$

3. Interpret summer and winter SI's:

summer: $(0.528 - 1) \times 100 = -47.2\%$

Summer is 47.2% below the seasonal av.

winter: $(1.361 - 1) \times 100 = 36.1\%$

winter is 36.1% above the seasonal av.

4. summer and winter should be corrected for seasonality:

summer:

$(\frac{1}{0.528} - 1) \times 100 = 89.4\%$

To correct summer for seasonality, sales need to increase by 89.4%

winter:

$(\frac{1}{1.361} - 1) \times 100 = -26.5\%$

To correct winter for seasonality, sales need to decrease by 26.5%

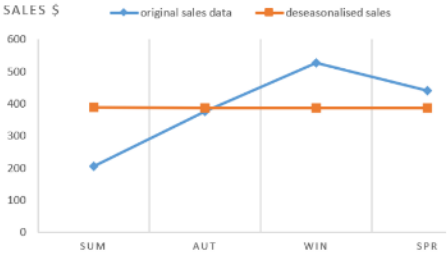
5. deseasonalise the seasons

	summer	autumn	winter	spring
Deseasonalise seasons	$\frac{205}{0.528} = 388.26$	$\frac{377}{0.972} = 387.86$	$\frac{528}{1.361} = 387.95$	$\frac{442}{1.139} = 388.06$

6. reseasonalise autumn:

RS = 387.86 × 0.972 = 376.99 ≈ 377

7. graph the deseasonalised values



4E: Trendlines and forecasting

- trend lines can be fitted to time series plots if there appears to be an increasing or decreasing trend.

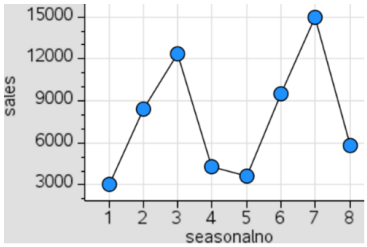
→ the LSRL is used

- if seasonality is present, data needs to be deseasonalised first before fitting the LSRL
- forecasting: making a prediction for the future

→ you need to re-seasonalise the value if the prediction was made from a deseasonalised LSRL.

example: the following sales are from a ski hire shop

	SUM	AUT	WIN	SPR	Season av.
2014	3051	8430	12340	4302	7030.75
2015	3651	9471	14960	5793	8468.50



seasonality present → need to deseasonalise

1) determine the average SI for each season:

	SUM	AUT	WIN	SPR
2014 SI	$\frac{3051}{7030.75} = 0.44$	$\frac{8430}{7030.75} = 1.2$	$\frac{12340}{7030.75} = 1.76$	$\frac{4302}{7030.75} = 0.61$
2015 SI	$\frac{3651}{8468.5} = 0.43$	$\frac{9471}{8468.5} = 1.12$	$\frac{14960}{8468.5} = 1.77$	$\frac{5793}{8468.5} = 0.68$
AVG SI	$\frac{0.44+0.43}{2} = 0.435$	$\frac{1.2+1.12}{2} = 1.16$	$\frac{1.76+1.77}{2} = 1.765$	$\frac{0.61+0.68}{2} = 0.645$

2) deseasonalise each value using the average seasonal index

	SUM	AUT	WIN	SPR
Season number	1	2	3	4
2014 DS	$\frac{3051}{0.435} = 6834.09$	$\frac{8430}{1.16} = 7267.24$	$\frac{12340}{1.765} = 6991.50$	$\frac{4302}{0.645} = 6669.77$
Season number	5	6	7	8
2015 DS	$\frac{3651}{0.43.5} = 8388.10$	$\frac{9471}{1.16} = 8164.66$	$\frac{14960}{1.765} = 8475.92$	$\frac{5793}{0.645} = 8981.40$

3) use deseasonalised values to determine LSRL

CAS

ds sales = 6362.35 + 304.97 × season number

4) Forecasting: predict actual sales for winter 2016

season number: 11

DS sales: 6362.35 + 304.97 × 11 = 9717.02

reseasonalise: 9717.02 × 1.765 = 17150.56