Formula Sheet

Core – Data analysis

standardised score	$z = \dfrac{x - \bar{x}}{s_x}$
lower and upper fence in a boxplot	lower $Q_1 - 1.5 \times IQR$ upper $Q_3 + 1.5 \times IQR$
least squares line of best fit	$y = a + bx$, where $b = r\dfrac{s_y}{s_x}$ and $a = \bar{y} - b\bar{x}$
residual value	residual value = actual value – predicted value
seasonal index	seasonal index = $\dfrac{\text{actual figure}}{\text{deseasonalised figure}}$

Core – Recursion and financial modelling

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

Module 1 – Matrices

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

Module 4 – Graphs and relations

gradient (slope) of a straight line	$m = \dfrac{y_2 - y_1}{x_2 - x_1}$
equation of a straight line	$y = mx + c$

Significant figures vs. Decimal places

> 273.245 = 273.25
> (to 2 decimal places)
>
> 273.245 = 270
> (to 2 significant figures)

• 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)

• Decimal places

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

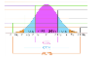
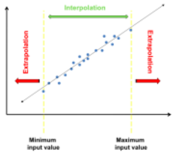
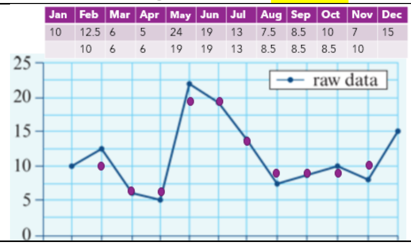
422 . 347

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"**

Topics	Data Types		Display/Analyse Tools		Report/Explain/Interpret/Describe		
Univariate Data	Categorical variables	Nominal data Ordinal data	Bar chart A3, Pie Chart, Frequency Table A1 P7, Segmented bar chart A4		Mode A1/ Modal Value A2 Frequency types, Frequency % = $\frac{Count}{Total\ Count} \times 100\%$		
	Numerical variables	Discrete data	Boxplots A6 P8, Frequency Tables	Stem plot A14/A13, dot plot	**Shape →** **Centre →** **Spread →** Outliers → P7	Symmetric A5 Mean $\bar{x}$ Standard Deviation S P19-21  68-95-99.7% rule A16 Z-score=Z=$\frac{x-\bar{x}}{S}$ A16 $x = \bar{x} + Z * S$	Skewed A6 Median M or Q_2 IQR, Range A7 Lower Fence $= Q_1 - 1.5 * IQR$ A8 Upper Fence $= Q_3 + 1.5 * IQR$ A6 5-figure summary: Min, Q_1, Q_2, Q_3, Max IQR=$Q_3 - Q_1$, Range=Max−Min
		Continuous data		Stem plot A14/A13, histogram A9/ A10, log A11 loghistogram A12			
Bivariate Data	Two categorical variables		Segmented bar chart A4, two-way frequency table, parallel bar chart A3		Mode/ Modal Value Frequency types		
	One categorical, one numerical variable		Back-to-back stem plots A14, parallel dot plots, parallel box plots A15P8		**Shape →** **Centre →** **Spread →P11**	Symmetric A5 Mean $\bar{x}$ Standard Deviation S	Skewed A6 Median M or Q_2 IQR, Range
	Two numerical variables  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 B1 Explanatory variable —explains/predicts→ Response variable **residual = actual data value y - predicted B3 value y** residual = $y - \hat{y}$ Nil pattern residual plot P13 P14 = Linear relation Curved/ patterned residual plot ≠ linear relation		**Strength →** **Direction →** **Form →P12P14** B2 P14 **The regression line y=a+bx** B4 P13 **slope** $b=\frac{r s_y}{s_x}$ B5 P14 **intercept** $a = \bar{y} - b\bar{x}$	Strong/Moderate/Weak (Check r value) B1 Positive / Negative Linear / Non-linear Reporting P12 P14 on Coefficient of Determination r^2 B1 Almost [r^2 in %] of [RV y] can be explained / predicted by [EV x].		
Time Series D1(Plot) D7(Fitted Line) D8(Prediction) P16	**Features P16**	**Moving smoothing P17**			**Seasonal Index S.I. D6**	**Deseasonalising D6**	
	Trend Cycles Seasonality Structure change Outliers	Moving Mean: 3/5 moving mean D2 18.1, 24.8, 26.4: $\frac{18.1+24.8+26.4}{3}=23.1$	Moving Mean: 2/4 Moving mean D3 18.1, 24.8, 26.4: $\frac{18.1+24.8}{2}=21.45$, $\frac{24.8+26.4}{2}=25.6$, $\frac{21.45+25.6}{2}=23.525$	Moving Median D4/D5 	S.I.=$\frac{Value\ for\ Season}{Yearly\ Average}$ Yearly Average=$\frac{Sum\ of\ Season\ Values}{No.of\ season\ per\ year}$	Deseasonalised Figure = $\frac{Actual\ Figure}{S.I.} = Actual\ Figure * \frac{1}{S.I.}$ Actual figure= Deseasonalised Figure * S.I.	

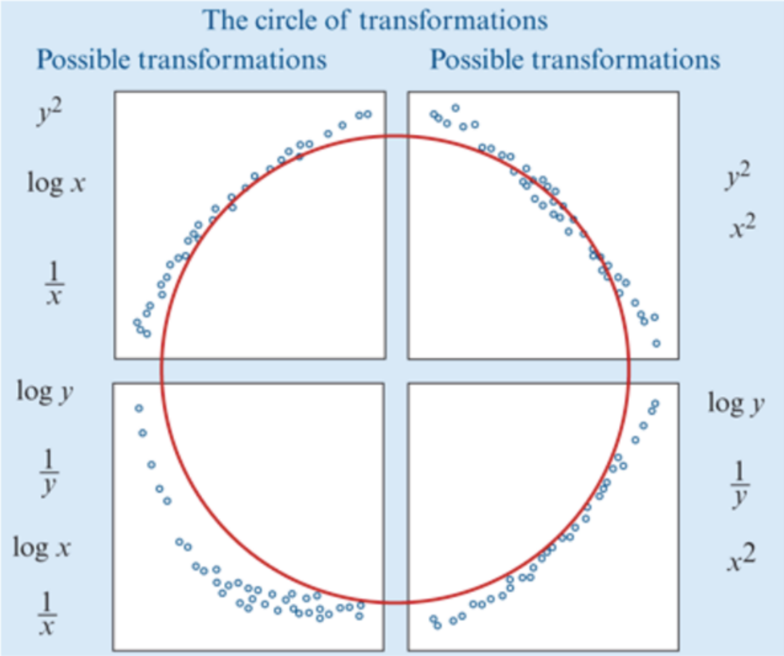
$Stretching\ transformation: Squared\ \&\ recipricol\ transformation$

$Compressing\ transformation:\ Logarithmic\ transformation$

$y^2 = a + bx$

$y = a + b\log x$

$y = a + \frac{b}{x}$



$y^2 = a + bx$

$y = a + bx^2$

$logy = a + bx$

$\frac{1}{y} = a + bx$

$y = a + b\log x$

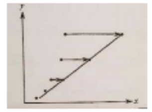
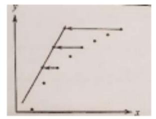
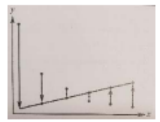
$y = a + \frac{b}{x}$

$logy = a + bx$

$\frac{1}{y} = a + bx$

$y = a + bx^2$

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	

C1(need laptop speed) /C2(old)

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

Write $2^3 = 8$ in logarithmic form.

- https://www.youtube.com/watch?v=zzu2POfYv0Y

Logarithm

Solution: $\log_2 8 = 3$

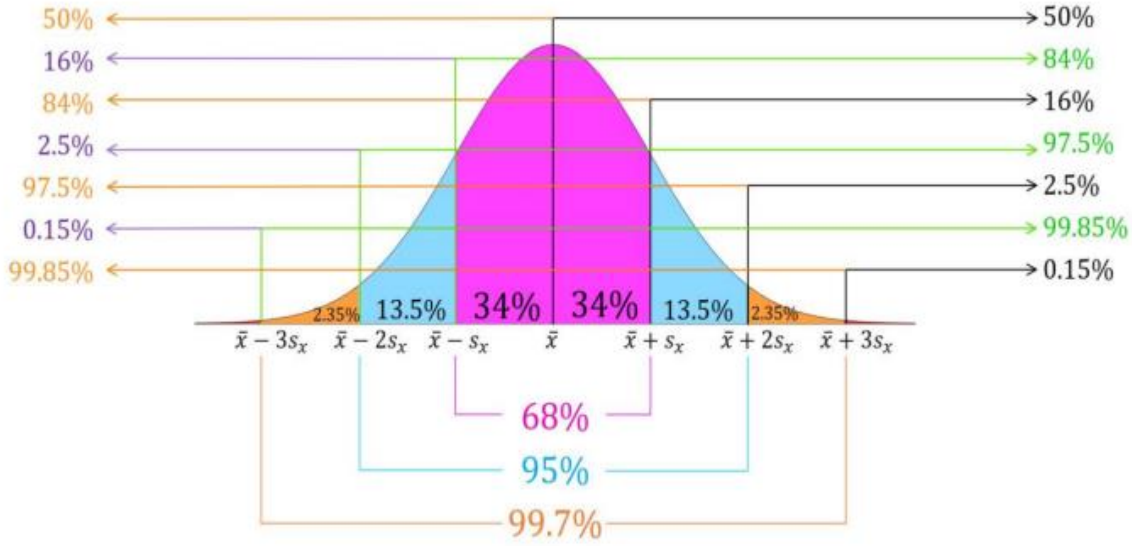
We read this as: "the log base 2 of 8 is equal to 3".

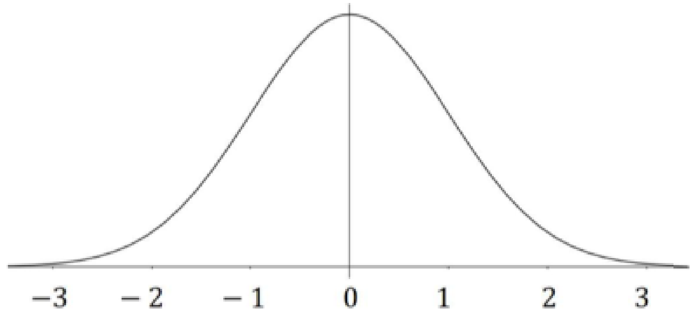
Convert to log form: $100 = 10^2$ $\log_{10} 100 = 2$

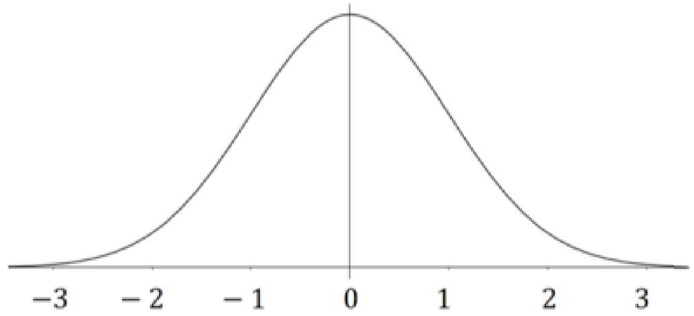
Convert to exponential form:

$\log_2 8 = 3$ $2^3 = 8$

The Normal Distribution







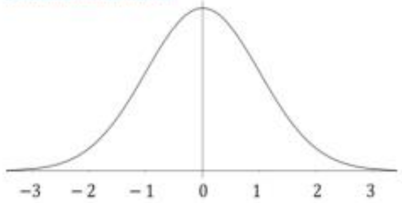
Converting to a Standard Score

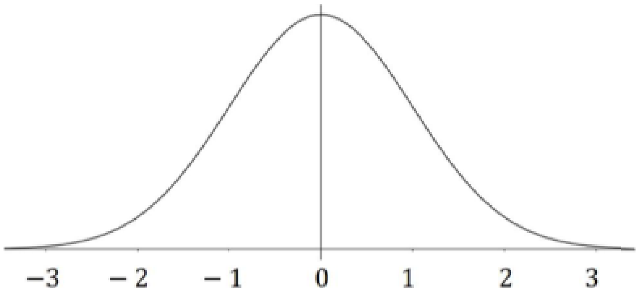
$$z = \frac{x - \bar{x}}{s}$$

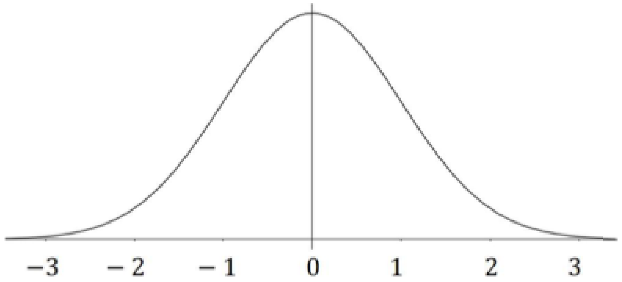
Converting to an Actual Score

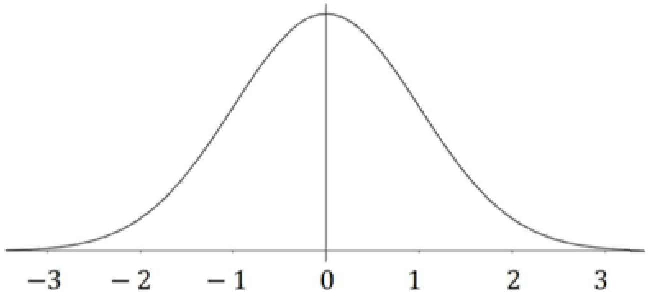
$$x = \bar{x} + z \times s$$

Standardised Bell Curve



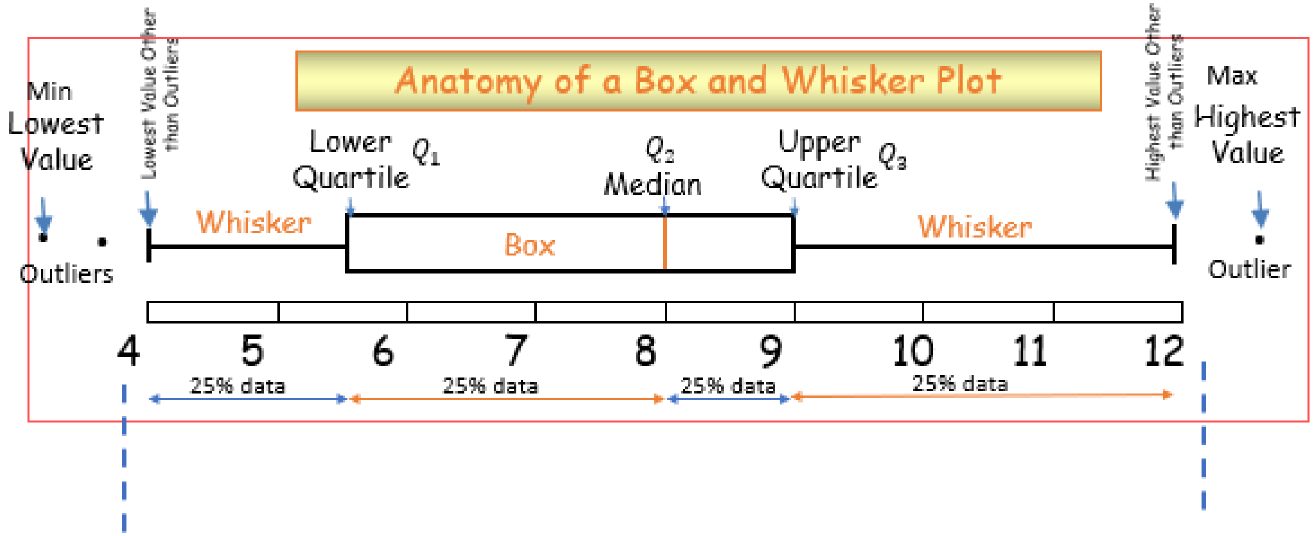






Box and Whisker Plots

Box plots are graphical representations of 5 number summary.



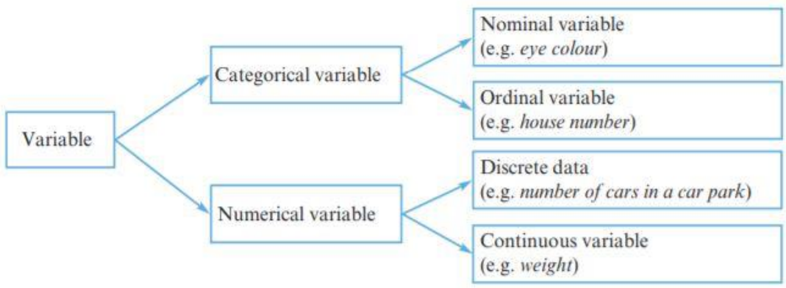
Lower Fence
$= Q_1 - 1.5 \times IQR$

50% of data is contained in the box.
25% of data is found in each whisker.

Upper Fence
$= Q_3 + 1.5 \times IQR$

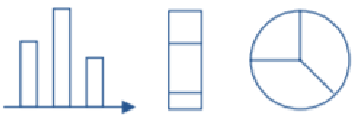
Strong positive association: r between 0.75 and 0.99
Moderate positive association: r between 0.5 and 0.74
Weak positive association: r between 0.25 and 0.49
No association: r between -0.24 and $+0.24$
Weak negative association: r between -0.25 and -0.49
Moderate negative association: r between -0.5 and -0.74
Strong negative association: r between -0.75 and -0.99

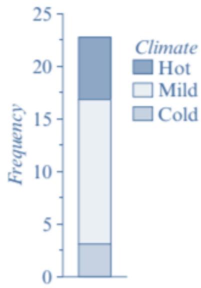
Core: Data Analysis



- ## Displaying categorical data

→ Bar charts, segmented bar charts, pie charts, dot plots.





Reports:

→ Modal Category:

From the majority of **[frequency type]** climate types in the **23 days, [modal value] 14 of the days** were found to be a **mild climate**. Of the remaining **[frequency type]** climate types, **[value X]** 3 of days were **[category X]** cold in climate and **[value Y]** 6 of the days were **[category Y]** hot in climate.

→ Equal Categories:

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

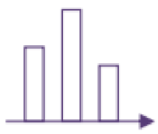
- ## Displaying numerical data

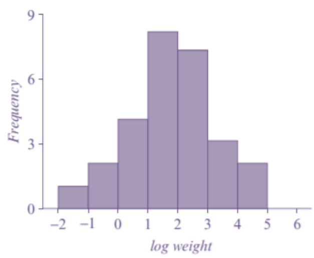
→ Discrete numerical data: stemplots, bar charts, dot-plots

→ Continuous numerical data: histograms, log charts (requires grouping data into intervals)

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

2 | 0 rep 2.0





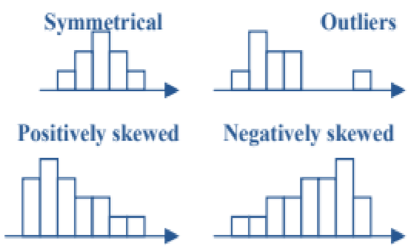
Report:

→ The shape of the distribution is **[symmetric/positively skewed/negatively skewed]**

→ The distribution has a **[range/IQR/standard dev.]** of **[value]**

→ The distribution has a **[mean/median/mode]** of **[value]**

→ The distribution **[has/has no]** outliers.



- ## Measures of centre

→ The median: **no. of values +1 ÷ 2**

→ The mean: **sum of all values ÷ no. of values**

→ The mode: **most repeated value / class interval**

Note:

→ Standard deviation and mean:

Do not use if data has outliers or is skewed

→ Median:

Use in any case

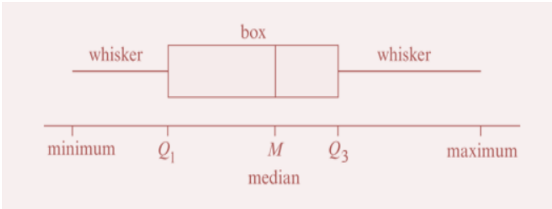
- ## Measures of spread

→ IQR: $\mathbf{Q_3 - Q_1}$

→ The range: **max value – min value**

→ The standard deviation: gives the average variation around the mean

- ## The five number summary and boxplots



→ How to make sure if a value is an outlier or not: calculate lower and upper fence and see if it lies outside either fence.

- Lower fence: $\mathbf{Q_1 - IQR \times 1.5}$

- Upper fence: $\mathbf{Q_3 + IQR \times 1.5}$

Report:

→ One boxplot:

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

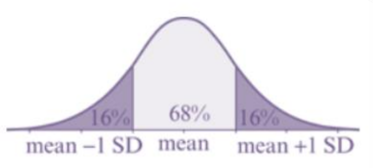
→ Comparing two boxplots:

The distributions at **[variable name]** are **[positively/negatively/symmetrically]** 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)**.

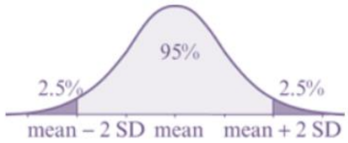
- ## Mean vs. Median to describe measures of centre

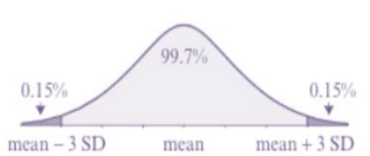
→ Choose either mean or median if: data is symmetric and has no outliers.
→ Choose only median if: data is skewed and there are outliers.

The 68- 95- 99.7 %rule



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

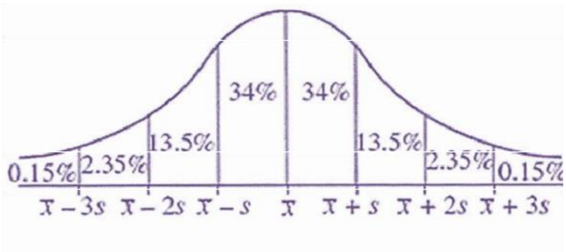




Standard scores

- Standard scores: z = actual value – mean value ÷ s.d
- Actual score: x= mean value + standard score x s.d
- Standard scores can be both positive and negative:

- Positive: actual score lies above the mean
- Negative: actual score lies below the mean
- Zero: actual score is equal to the mean



- Worked example:

Subject	Mark	Mean	Standard Deviation
Psychology	75	65	10
Statistics	70	60	5

If we assume that the *marks* are *normally distributed*, then *standardisation* and the 68–95–99.7% *rule* give us a way of resolving this issue.
Let us standardise the marks.

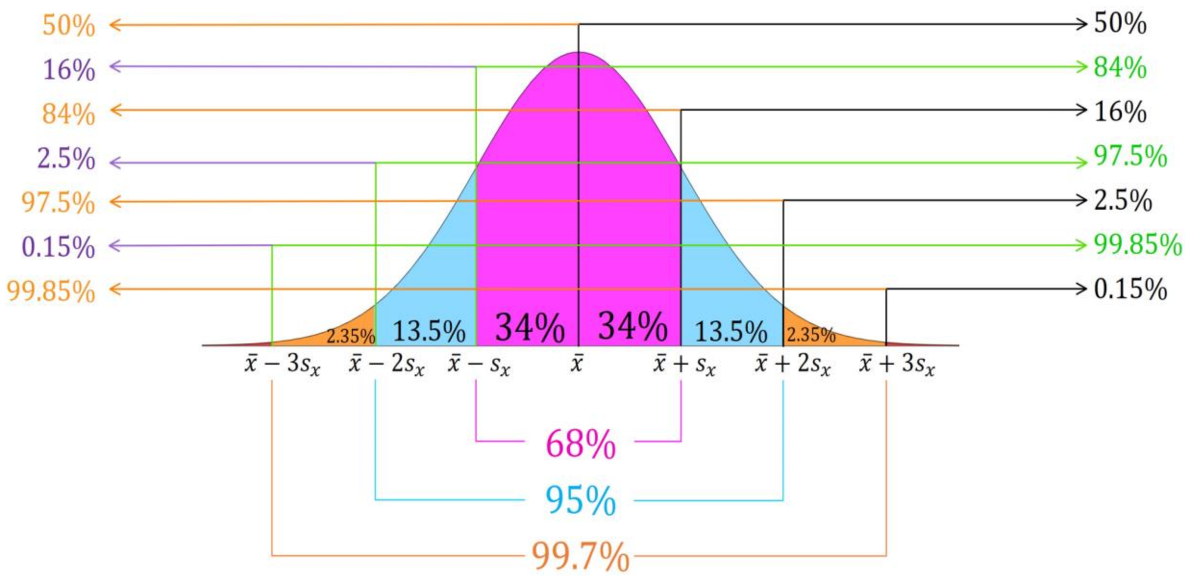
Psychology: standardised mark $z = \dfrac{75-65}{10} = 1$

Statistics: standardised mark $z = \dfrac{70-60}{5} = 2$

What do we see? The student obtained a higher score for Psychology than for Statistics. However, relative to her classmates she did better in Statistics.

- Her mark of 70 in Statistics is equivalent to a z-score of 2. This means that her mark was two standard deviations above the mean, placing her in the top 2.5% of students.
- Her mark of 75 for Psychology is equivalent to a z-score of 1. This means that her mark was only one standard deviation above the mean, placing her in the top 16% of students. This is a good performance, but not as good as for statistics.

The Normal Distribution



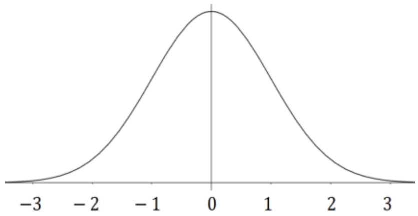
Converting to a Standard Score

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

Converting to an Actual Score

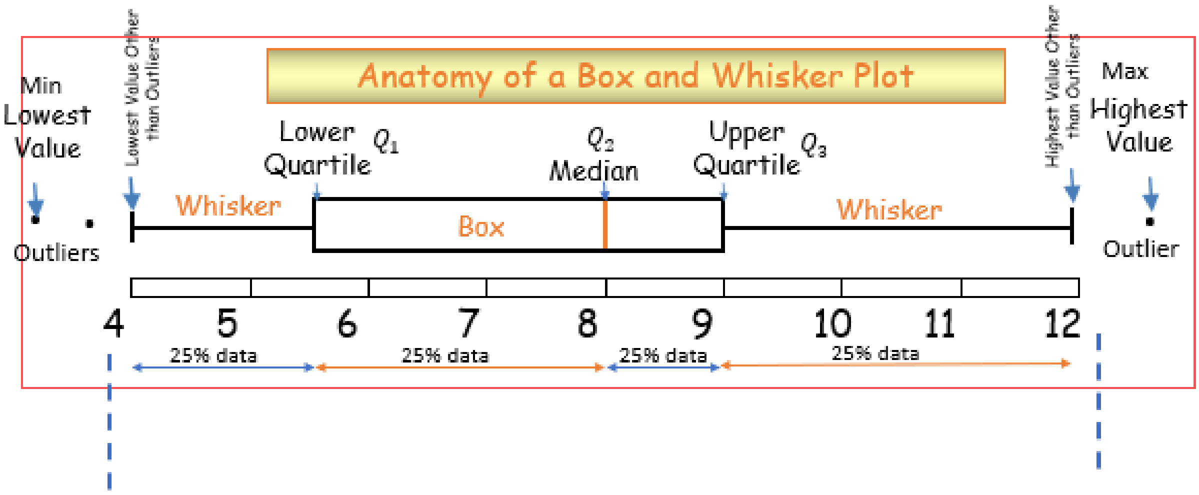
$x = \bar{x} + z \times s$

Standardised Bell Curve



Box and Whisker Plots

Box plots are graphical representations of 5 number summary.



Lower Fence
$= Q_1 - 1.5 \times IQR$

50% of data is contained in the box.
25% of data is found in each whisker.

Upper Fence
$= Q_3 + 1.5 \times IQR$

Response and explanatory variable

→ <u>Explanatory variable:</u> independent variable (x)
→ <u>Response variable:</u> dependent variable (y)

Association between two categorical variables

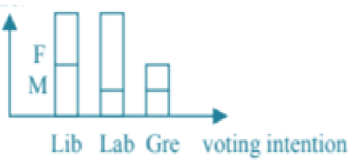
→ <u>Displayed through:</u> segmented bar chart, two-way table, parallel bar charts

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

<u>Report:</u>

→ <u>Worked example: Is there an association between interest in sports and age group?</u>

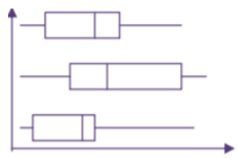
Yes, the percentage of males with a high level of interest in sport steadily decreases with age group from 56.5 % for the 'under 18 years' age group, to 35.0% for the '36-50 years' age group.



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

Association between numerical and categorical variables

→ <u>Displayed through:</u> back- to-back stem plots;
for **more than 2** EV categories: parallel dot-plots, parallel boxplots



<u>Report:</u>

→ Similar distributions

The shape of distribution A is **[symmetric/positively/negatively skewed/bi-modal]**. Distributiion A has a **[range/IQR/standard deviation]** of **[value]**, similarly Distribution B has a **[range/IQR/ standard dev.]** of **[value]**. Distribution A has a **[mean/median/mode]** of **[value]**, similarly Distribution B has a **[mean/median/mode]** of **[value]**. Distribution A and B **[have/have no]** outliers. Distributions have no association (because they are similar in almost everything, hence they can't be an association as there is no variation in results)

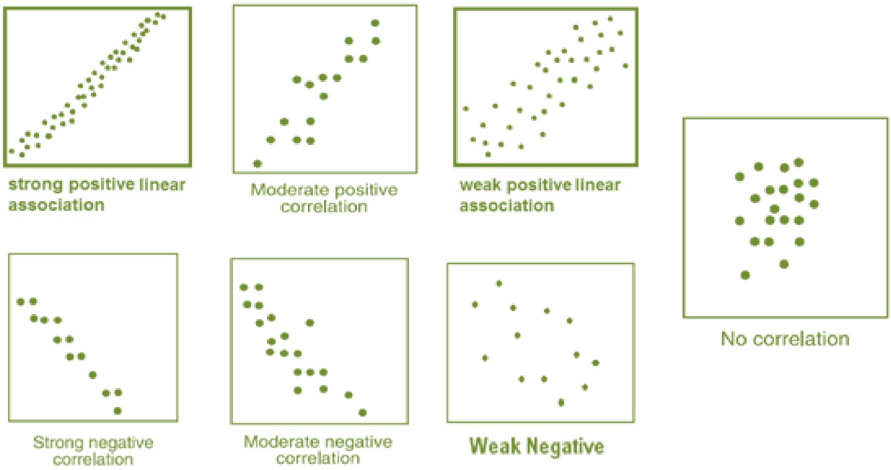
→ Different disributions

The shape of distribution A is **[symmetric/positively/negatively skewed/ bi-modal]** whereas the shape of distribution B is **[symmetric/positively/negatively skewed/ bi-modal]**. Distribution A has a **[range/ IQR/standard dev.]** of **[value]** whereas Distribution B has a **[range/ IQR/standard dev.]** of **[value]**. Distribution A has a **[mean/median/mode]** of **[value]**. Distribution A **[has/has no]** outliers while Distribution B **[has/has no]** outliers.They both have an associaton.

- talk about how the increase in IQR, median (increases/decreases) and shape/skew (becomes more positively skewed as age increases, for example) all support the association between both variables.

• Association between two numerical variables

→ <u>Displayed through:</u> scatterplots



Report:
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.

→ <u>Pearson's correlation coefficient (**r**) :</u> helps determine association

- <u>It can only be used assuming that :</u>
  1. There are **no outliers** in the data
  2. The variables are **numeric**
  3. The association is **linear**

  ...Otherwise it could give misleading information!!

→ <u>The coefficient of determination ($\mathbf{r^2}$):</u>

Report:
[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.

Strong positive association: r between 0.75 and 0.99
Moderate positive association: r between 0.5 and 0.74
Weak positive association: r between 0.25 and 0.49
No association: r between −0.24 and +0.24
Weak negative association: r between −0.25 and −0.49
Moderate negative association: r between −0.5 and −0.74
Strong negative association: r between −0.75 and −0.99

<u>Remember:</u> When square-rooting r^2 to gain r value, identify whether the relationship is negative or positive and accordingly, r will take on a (–) or a (+)

<u>Note:</u> Even if variables swap, **r value will always remain the same**

- Least squares regression line

→ Minimises the sum of the squares of the residuals

→ **The assumptions for the least squares line is the same as for the correlation coefficient**

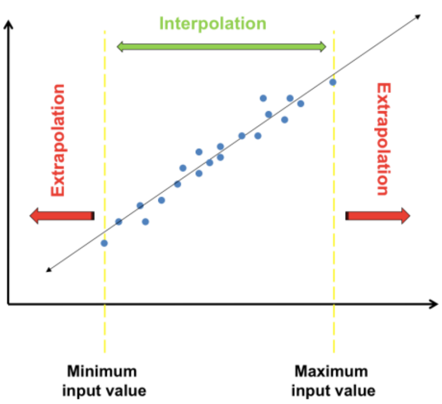
→ Equation of line: **a+bx**

- the slope (b) = $\boxed{b = r\frac{s_y}{s_x}}$
- the intercept (a) = $\boxed{a = \bar{y} - b\bar{x}.}$

→ r : correlation coefficient

→ s_x and s_y : standard deviations of x and y

→ $\bar{x}$ and $\bar{y}$: the mean values of x and y



→ Interpolation: predicting **within** the range of data

→ Extrapolation: predicting **outside** the range of data

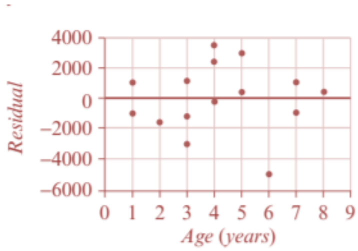
> Report:
>
> → Slope (b):
>
> On average, **[response variable] [increases/decreases]** by **[b units]** for every one unit increase in **[explanantory variable]**
>
> → y- intecept (a):
>
> When **[explanatory variable]** is 0, **[response variable]** is predicted to be **[a units]**

- Residuals: distance between the individual data points and the regression line

→ Residual value: **actual value – predicted value**

→ Residuals can be positive, negative or zero:

- Data points above regression line: positive residual
- Data points below residual line: negative residual
- Data points on the line: zero residual

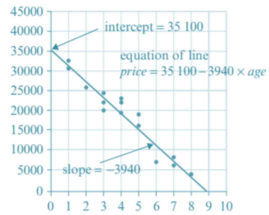


→ Residual plots: plot of the residual value for each data value

→ Random scatters indicate a linear relationship

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

- *A complete regression analysis*



<u>Report:</u>

→ Strength:

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.

→ 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 **[explanantory variable]**

→ y- intecept (a):

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

→ The coefficient of determination:

The coefficient of determination indicates that **[r^2 x 100]** of the variation in **[response variable]** is explained by **[explanatory variable]**

→ Residual plot:

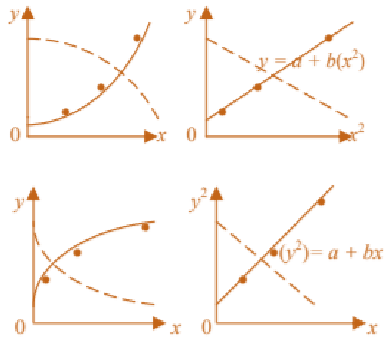
The residual plot shows a **[random scatter/ curved patter]** indicating there is a **[linear/non-linear]** relationship between **[response variable]** and **[explanatory variable]**

- *The square transformation:* **y = a+ b (x^2)**

→ <u>x^2 transformation:</u> spreads out the high x-values relative to lower x values

→ <u>y^2 transformation:</u> stretches out y-values

→ <u>x- axis</u>: x^2 values/ x values
<u>y-axis</u>: y values/ y^2 values

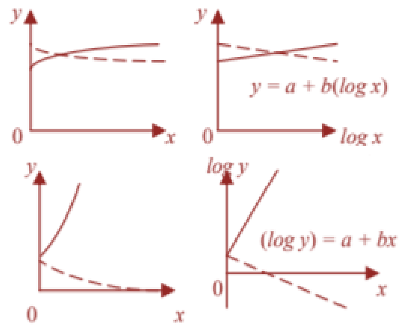


- *The log transformation:* **y = a + b (log x)**

→ <u>Log x transformation:</u> compresses the higher x values relative to lower x values

→ <u>Log y transformation:</u> compresses the higher y values relative to lower y values

→ <u>x- axis:</u> log x values/ x values
→ <u>y- axis:</u> y values/ log y values

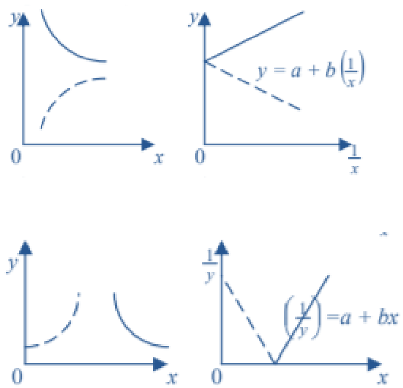


- *The reciprocal transformation:* **y= a+ b (1/x)**

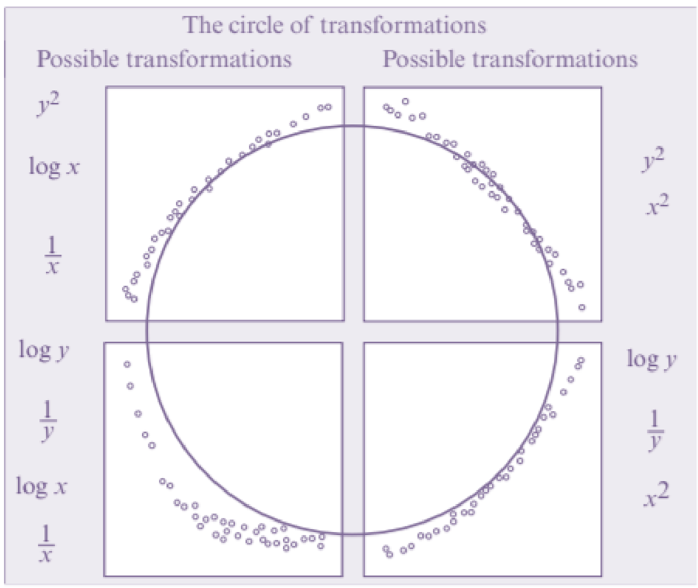
→ <u>1/x transformation:</u> compresses larger x values relative to lower x values

→ <u>1/y transformation:</u> compresses larger y values relative to lower y values

→ <u>x- axis:</u> 1/x values/ x values
→ <u>y- axis:</u> y values/ 1/y values



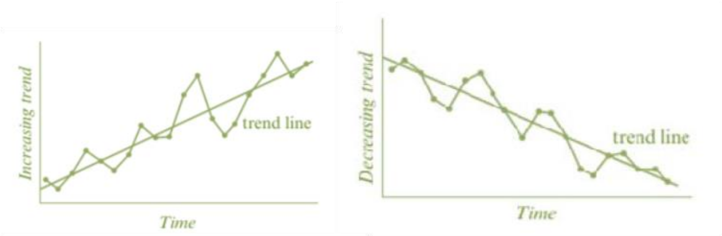
- *The circle of transformations*



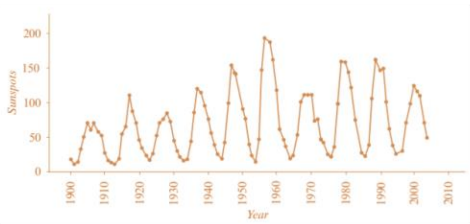
> <u>Note</u>: If the transformed data has a **high r^2, and if its residual plot is scattered,** then it is a very appropriate transformation to use

Features of a time series plot:

- <u>Trend:</u> increase/decrease in values



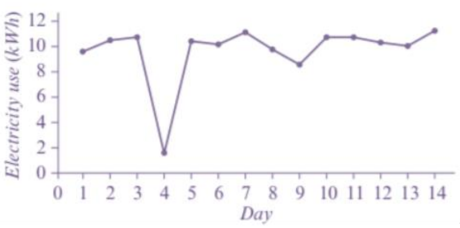
- <u>Cycles:</u> periodic movements in a time series (periods over a year)
  - Doesn't follow a set seasonal pattern, the cycles change from time to time



- <u>Seasonality:</u> present when there is a periodic movement in a time series that has a calendar-related period (years, months, weeks)
  - Follows a set seasonal pattern

- <u>Structural change:</u> present when there is a sudden change in the pattern of the graph and it occurs for a time frame (is not sudden like an outlier).
  - It takes a while for the data to return to its original structure.

- <u>Outliers:</u> out-standing values that occur suddenly (unlike structural change) and after which the plot is able to return to its normal structure.



- <u>Irregular (random) fluctuations:</u> includes all variations in a time series

- **Moving mean smoothing**

→ Three moving mean: $\text{smoothed } y_2 = \frac{y_1 + y_2 + y_3}{3}$ (in this case, for y_2)

→ Five moving mean: $\text{smoothed } y_3 = \frac{y_1 + y_2 + y_3 + y_4 + y_5}{5}$ (in this case, for y_3)

- Removes irregular fluctuations better than 3-mean smoothing

→ Two-mean smoothing with centring: (*in this case, **centred** at Tuesday*)

Day	*Temperature*	*Two-moving means*	*Two-moving mean with centring*
Monday	18.1		
		$\frac{(18.1 + 24.8)}{2} = 21.45$	
Tuesday	24.8		$\frac{(21.45 + 25.6)}{2} = 23.525$
		$\frac{(24.8 + 26.4)}{2} = 25.60$	
Wednesday	26.4		

- This process is similar for the four-moving mean smoothing

24.8 26.4 13.9 12.7 14.2

$$\text{Mean 1} = \frac{(24.8 + 26.4 + 13.9 + 12.7)}{4} = 19.45$$

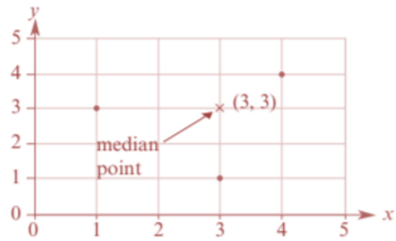
$$\text{Mean 2} = \frac{(26.4 + 13.9 + 12.7 + 14.2)}{4} = 16.8$$

$$\text{Centred mean} = \frac{(\text{mean 1} + \text{mean 2})}{2} = \frac{(19.45 + 16.8)}{2} = 18.125$$

- **Moving medians:** can be better than moving means if **there are outliers** in the data

→ Three-moving median: the y-value that is in between the other two y values becomes the median point.

→ Five-moving median: similar to three-moving median but with 5 values



- **Seasonal indices**

→ **Seasonal indices always add up to one whole, so that the sum equals the number of seasons (for ex: seasons are months so seasonal indices add up to 12)**

→ The average seasonal index is always 1:
- If the seasonal index =1.2 = 120% = 20% above average
- If the seasonal index = 0.8 = 80% = 20% lower than average

→ Deseasonalising data : **actual figure / seasonal index**

Note: After smoothing or Deseasonalising, you **get rid of seasonality**, **cyclic nature etc,** hence smoothed data cannot be described as seasonal etc.

The only thing quality it contains is **trend**; increasing, decreasing or no trend at all.

- Removes seasonality from time series plot
- Revealed a clear underlying trend in the data

> Note: To obtain actual value, deseasonalised data needs to be reseasonalised

- Actual figure : deseasonalised figure x seasonal index
- Seasonal index : value for the one season / seasonal average (find the mean value of the season)
- Seasonal indices for several years' data: simply find the average of all the seasonal indices from all the years for each season.
- Correcting for seasonality: 100 / seasonal index
- Ex: 100/0.8= 125, the sales should be increased by 25%

⑩ Fitting a trend line and forecasting

- Fitting a trend line:
- Fit a least square regression line into the data/ graph (if given)
- Find the slope and interpret it

> **Report:**
>
> Over the period [period], the [response variable] [increased/decreased] at an average rate of [b] units per [one unit] in [explanatory variable]

- Forecasting: substitute value into regression line to find an approximate, possible, forecasted value.
- Forecasting with seasonality: (worked example)

Example 15 Forecasting (seasonality)

What sales do we predict for Mikki's shop in the winter of year 4? (Because many items have to be ordered well in advance, retailers often need to make such decisions.)

Solution

1 Substitute the appropriate value for the time period in the equation for the trend line. Since summer year 1 was designated as quarter '1', then winter year 4 is quarter '15'.

$$\text{Sales} = 838.0 + 32.1 \times \text{quarter}$$
$$= 838.0 + 32.1 \times 15$$
$$= 1319.5$$

Deseasonalised sales prediction for winter of year 4 = 1319.5

2 The value just calculated is the deseasonalised sales figure for the quarter in question. To obtain the *actual* predicted sales figure we need to reseasonalise this predicted value. To do this, we multiply this value by the seasonal index for winter, which is 1.30.

Seasonalised sales prediction for winter of year 4 $= 1319.5 \times 1.30$
≈ 1715

Sample standard deviation

Here's the formula again for sample standard deviation:

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

Here's how to calculate sample standard deviation:

Step 1: Calculate the mean of the data—this is   in the formula.

Step 2: Subtract the mean from each data point. These differences are called deviations. Data points below the mean will have negative deviations, and data points above the mean will have positive deviations.

Step 3: Square each deviation to make it positive.

Step 4: Add the squared deviations together.

Step 5: Divide the sum by one less than the number of data points in the sample. The result is called the variance.

Step 6: Take the square root of the variance to get the standard deviation.

Example: Sample standard deviation

A sample of   students was taken to see how many pencils they were carrying.

Calculate the sample standard deviation of their responses:
2, 2 , 5, 7

Step 1: Find the mean.

$$\bar{x} = \frac{2+2+5+7}{4} = \frac{16}{4} = 4$$

The sample mean is 4 pencils.

Step 2: Subtract the mean from each score.

Pencils: x	Deviation: x − $\bar{x}$
2	2−4= −2
2	2−4=−2
5	5−4=1
7	7−4=3

Step 3: Square each deviation.

Pencils: x	Deviation: x − $\bar{x}$	Squared Deviation: (x − $\bar{x}$)^2
2	2−4= −2	$(-2)^2=4$
2	2−4= −2	$(-2)^2=4$
5	5−4=1	$(1)^2=1$
7	7−4=3	$(3)^2=9$

Step 4: Add the squared deviations.

4+4+1+9=18

Step 5: Divide the sum by one less than the number of data points.

$$\frac{18}{4-1} = \frac{18}{3} = 6$$

Step 6: Take the square root of the result from Step 5.

$$\sqrt{6} \approx 2.45$$

The sample standard deviation is approximately 2.45.

Want to learn more about sample standard deviation? Check out this video.

Want to practice some problems like this? Check out this exercise on sample and population standard deviation.

Find the mean and the standard deviation for the values 9, 4, 5, 6

$$\bar{x} = \frac{(9+4+5+6)}{4} = 6$$ Find the mean.

x	$\bar{x}$	$x-\bar{x}$	$(x-\bar{x})^2$
9	6	3	9
4	6	-2	4
5	6	-1	1
6	6	0	0
		sum	14

Organize the next steps in a table.

$$\sigma = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}}$$ Find the standard deviation.

$$\sqrt{\frac{14\cdot 4}{4\cdot 4}} = \frac{\sqrt{56}}{4} = \frac{\sqrt{14}}{2} \approx 1.87$$

The mean is 6, and the standard deviation is about 1.87.

–How to calculate the correlation coefficient using the formula

Use the formula to calculate the correlation coefficient, ***r***, for the following data.

x	1	3	5	4	7
y	2	5	7	2	9

$\bar{x} = 4,\ s_x = 2.236$
$\bar{y} = 5,\ s_y = 3.082$

Give the answer correct to two decimal places.

$\bar{x} = 4 \quad s_x = 2.236$
$\bar{y} = 5 \quad s_y = 3.082 \quad n = 5$

x	$(x-\bar{x})$	y	$(y-\bar{y})$	$(x-\bar{x})\times(y-\bar{y})$
1	−3	2	−3	9
3	−1	5	0	0
5	1	7	2	2
4	0	2	−3	0
7	3	9	4	12
Sum	0		0	23

$$\therefore \sum(x-\bar{x})(y-\bar{y}) = 23$$

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{(n-1)s_x s_y}$$

$$\therefore r = \frac{23}{(5-1)\times 2.236\times 3.082}$$
$$= 0.834... = 0.83\ (2\ \text{d.p.})$$