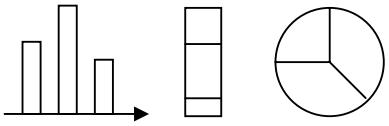
Further mathematics Summary sheets (Core only)

Data analysis

Univariate data (i.e. data of a single variable)
Categorical data: e.g. nationality, language, transportation, profession, sport, blood type.
Discrete numerical data: e.g. petrol price/litre, number of students in a class, road toll/year.
Continuous numerical data: monthly rain fall, daily maximum temperature, height of a tree.

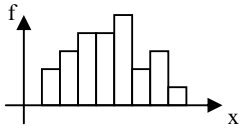
Bar-charts, segmented bar-charts, pie-charts are suitable for displaying categorical data.



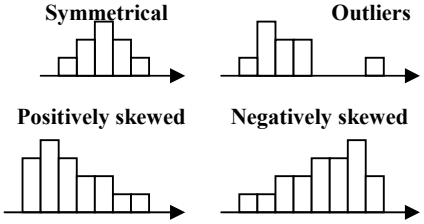
Stemplots are suitable for discrete numerical data.

2 | 0 0 3 7 8	2 | 0 0 3
3 | 2 3 3 6 9 9 9	2 | 7 8
4 | 1 1 1 7 7 8	3 | 2 3 3
	3 | 6 9 9 9
2 | 0 rep 2.0	4 | 1 1 1
	4 | 7 7 8

Frequency histograms are suitable for both discrete and continuous numerical data when they are grouped into class intervals.



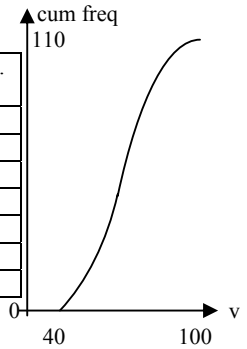
Terms used to describe a set of data:



Modal class is the interval with the highest frequency. Two or more modal classes are possible in the same data set. Mode is a value with the highest frequency in a data set

Cumulative frequency distribution and curve (or **ogive**):

Speed v (kmh^{-1})	Cumul. Freq.
<40	0
<50	6
<60	19
<70	40
<80	75
<90	100
<100	110



Summary statistics: **mean** $\bar{x}$ and **standard deviation** s for describing the centre and spread of numerical data, e.g.

x	f	fx	fx^2
2	3	6	12
5	5	25	125
7	4	28	196
8	2	16	128
	$n=\Sigma f=14$	$\Sigma fx=75$	$\Sigma fx^2=461$

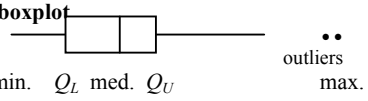
$\bar{x} = \frac{\Sigma fx}{n} = \frac{75}{14} = 5.36$, $s = \sqrt{\frac{1}{n-1}\left(\Sigma fx^2 - n\bar{x}^2\right)}$

$= \sqrt{\frac{1}{13}\left(461 - 14 \times 5.36^2\right)} = 2.13$

$\bar{x}$ and s are good measures of the centre and the spread provided the data set has no extreme outliers.

Median and **interquartile range** are more appropriate measures of the centre and spread if the data set contains extreme **outliers**. Arrange the data in ascending order. Median is the middle number (or the average of the middle two) of the arranged data. The middle of the half before the median is called the **lower quartile** Q_L, the middle of the other half after the median is the **upper quartile** Q_U. Interquartile range $IQR = Q_U - Q_L$.

These statistics can be presented in the form of a **boxplot**



For '**bell shaped**' data sets:
68% of the data lies within $x = \bar{x} \pm 1s$,
95% within $x = \bar{x} \pm 2s$,
99.7% within $x = \bar{x} \pm 3s$.
A data point is an outlier if it is less than $Q_L - 1.5 \times IQR$ or greater than $Q_U + 1.5 \times IQR$.

Bell shaped data sets are modelled by the normal distribution. Values (x) from different data sets are converted to standard z scores for comparison. $z = \frac{x - \bar{x}}{s}$.

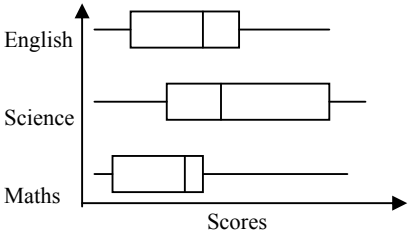
Bivariate data (i.e. data for two variables)
Dependent (response) variable on the y-axis and independent (explanatory) variable on the x-axis for graphing.

Back-to-back stemplot: to display relationship between numerical variable (e.g. age) and two-valued categorical variable (e.g. gender), e.g.

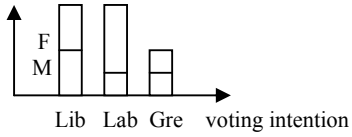
3 0	3	1 1 4 5
5 5 2	4	3 6 6 8 9
7 4 2 1	5	2 2 7

0 | 3 | 1 rep 30 yrs old F, 31 yrs old M.

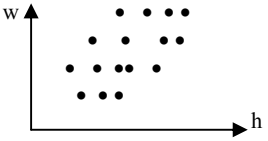
Parallel boxplots: to display relationship between a numerical variable (e.g. exam scores)and a two or more level categorical variable (e.g. subjects), e.g.



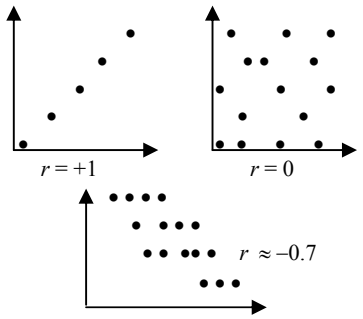
Segmented bar charts: to display the relationship between two categorical variables, e.g voting intention and gender.



Scatterplots: to display the association between two numerical variables, e.g.height and weight,



Describe a scatterplot in terms of **direction** (positive or negative association), **form** (linear or non-linear association) and **strength** (strong, moderate or weak association).
Correlation is a term used in statistics for association or connection.
Estimation of **Pearson's product-moment correlation coefficient** r from a scatterplot:
For the scatterplot above, $r \approx +0.5$



r is always a number between –1 and +1. It measures the strength of the linear association between two variables. $r = 0$ shows that there is no association and it suggests that the two variables are independent of each other. A positive (negative) value of r shows an increasing (decreasing) trend. $r = \pm 1$ shows a definite linear association.

Extreme outliers in the data set can result in a misleading value for r.

A strong correlation between two variables does not imply that a change in one variable causes a change in the other because there may be a third variable that influences both in the same way. A strong correlation only allows prediction of one variable from the other.

Coefficient of determination r^2: e.g. $r = 0.95$, $r^2 = 0.90$. The interpretation of $r^2 = 0.90$ is that 90% of the variation of one variable with the other can be explained by the relationship that exists between them.

Example. Refer to the scatterplot for weight and height above. $r = 0.5$, $r^2 = 0.25$, hence 25% of the variation of weight w with height h can be explained by the linear relationship between w and h.

Regression encompasses analysis of data in order to develop mathematical relationship (equation) between variables and exploration of the relationship to make predictions.
Methods to find regression line $y = a + bx$.

(1) **Fitting line by eye**: Draw line of best fit. Mark two convenient points on line as far apart as possible. Use the coordinates of the two marked points to calculate the slope of the line $b = \frac{y_2 - y_1}{x_2 - x_1}$, a is the y-intercept and can be found from the scatterplot, or by substituting the coordinates of one of the marked points in the equation.

(2) **The three-median line**: Divide the data into three groups (left L, middle M, right R) of equal number of points if possible, otherwise a single extra point goes to the middle or put one in each outer region for two extra points. For each group find the medians for the x and y values, they form the summary points (x_L, y_L), (x_M, y_M) and (x_R, y_R). Then calculate slope $b = \frac{y_R - y_L}{x_R - x_L}$ and y-intercept $a = \frac{1}{3}[(y_L + y_M + y_R) - b(x_L + x_M + x_R)]$

Draw the three-median line by placing a ruler in alignment with (x_L, y_L) and (x_R, y_R) and then sliding the ruler vertically one-third of the way towards (x_M, y_M).

(3) Estimation of line of best fit $y = a + bx$ from scatterplot by formulas $b = r\frac{s_y}{s_x}$ and $a = \bar{y} - b\bar{x}$.

(4) **The least squares line**: is the line for which the sum of the squares of the vertical deviations is a minimum. Use graphics calculator to find a and b in $y = a + bx$.

Slope b represents the rate of change in y as x increases, i.e. how much y changes when x increases by 1. Intercept a is the y value when $x = 0$.

Extrapolation: Using the regression line to make prediction beyond the data set.
Interpolation: Using the regression line to insert a value within the data set.

Residual analysis: For checking quality of fit.
Residual is defined as $y - \hat{y}$ where y is the observed (actual) value and $\hat{y}$ is the value predicted by a regression line. A regression line for which the sum of the squares of the residuals is smaller is a better fit.

Example Which of the following regression lines is a better fit for the data set (x, y).
Line 1: $y_1 = 2.6 + 1.3x$, line 2: $y_2 = 2.5 + 1.2x$

x	y	$\hat{y}_1$	resid	$(\text{resid})^2$
1	3.7	3.9	–0.2	0.04
2	5.2	5.2	0	0
3	6.8	6.5	+0.3	0.09
4	7.2	7.8	–0.6	0.36
				$\Sigma = 0.49$

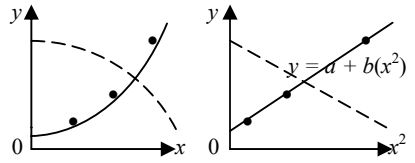
x	y	$\hat{y}_2$	resid	$(\text{resid})^2$
1	3.7	3.7	0	0
2	5.2	4.9	+0.3	0.09
3	6.8	6.1	+0.7	0.49
4	7.2	7.3	–0.1	0.01
				$\Sigma = 0.59$

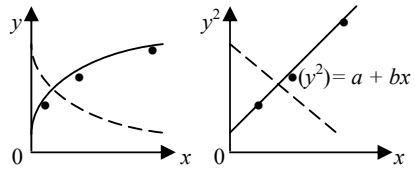
$\therefore$ $y_1 = 2.6 + 1.3x$ is a better fit.

Check for accuracy before use.
Please inform mathlinE (mathline@itute.com) re typing or mathematical errors.

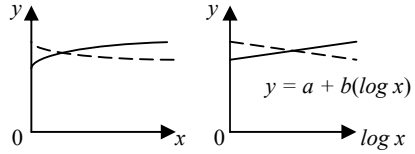
Transformation of some forms of non-linear data to linearity:

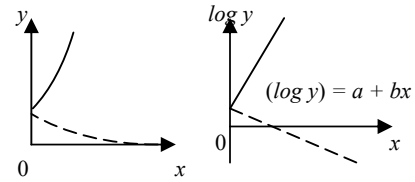
Square transformation



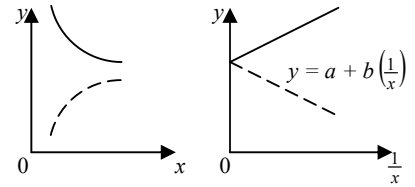


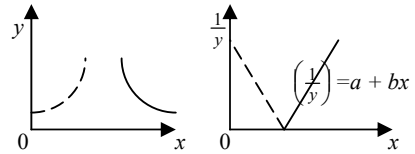
Log transformation



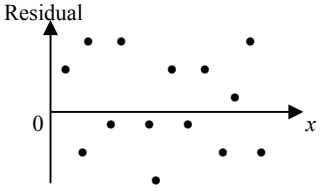


Reciprocal transformation





Residual plots: For displaying quality of fit.



A **random residual plot** suggests the regression line (representing certain relationship) is a good fit. If the residual plot shows a **pattern**, then another relationship (depending on the pattern) may result in a better fit.

Qualitative analysis of **time series**: (1) **Trend pattern**-increase or decrease in the data over time. (2) **Seasonal pattern**-variation of the data due to seasonal factors, e.g. the month of the year, the day of the week or the hour of the day. (3) **Cyclic pattern**-longer term (years) fluctuations in the data. (4) **Random pattern**-unpredictable short term variations of the data about a constant value.

How to distinguish seasonal from cyclic pattern: Seasonal pattern has a constant length in the cycle and constant magnitude in the variation. Cyclic pattern has a changing cycle length and magnitude from cycle to cycle.

Seasonal pattern exists in many time series. To study long term trends, remove the seasonal effects from the data, the process in doing so is called **deseasonalising** the data and the resulting figures are called **seasonally adjusted figures**. The estimates of seasonal effects are called **seasonal indices**.

Example

Year	Q1	Q2	Q3	Q4	Quart. Av.
1996	68	70	64	55	64.25
1997	65	67	64	55	62.75
1998	64	66	64	55	62.25
1999	61	65	59	53	59.50
2000	60	64	59	52	58.75

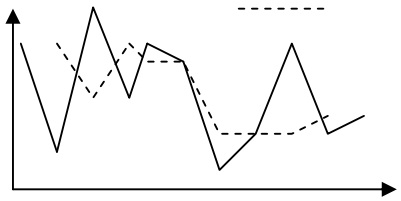
*Quarterly average=(Q1 + Q2 + Q3 + Q4) ÷ 4
Divide each quarterly entry by the quarterly av. for that year to obtain the following table.

Year	Q1	Q2	Q3	Q4
1996	1.0584	1.0895	0.9961	0.8560
1997	1.0359	1.0677	1.0199	0.8765
1998	1.0281	1.0602	1.0281	0.8835
1999	1.0252	1.0924	0.9916	0.8908
2000	1.0213	1.0894	1.0043	0.8851
	1.0338	*1.0798*	*1.0080*	*0.8784*

The entries in the last row of the above table are the seasonal indices for the quarters. Q1 seasonal index is obtained by taking the average of the 5 entries for the 5 years under Q1, etc. NB. Sum of seasonal indices = 4.
Deseason.figure = actual figure/seasonal index
The following table shows the deseasonalised figures and can be analysed for long term trends without the effects of seasonal factors.

Year	Q1	Q2	Q3	Q4
1996	65.78	64.83	63.49	62.61
1997	62.87	62.05	63.49	62.61
1998	61.91	61.12	63.49	61.48
1999	59.01	60.20	58.53	60.34
2000	58.04	59.27	58.53	59.20

Random variations or seasonal effects in time series can also be removed by smoothing procedures: e.g. *3- moving median smoothing*



e.g. *centred 4-moving average smoothing* (suitable for removing seasonal effects in patterns that repeat every 4 terms, e.g. quarterly data)

Year	Qtr	x	4-m. av	Centred
1996	1	68		
	2	70		
			64.25	
	3	64		63.88
			63.50	
	4	55		63.13
			62.75	
1997	1	65		62.75
			62.75	
	2	67		62.75
			62.75	
	3	64		
	4	55		

For daily seasonal data, use 7-moving average; for monthly data, use 12-moving average.
A regression line can be fitted for the deseasonalised or smoothed data and used to make predictions.