A Variables and Measures

Every **variable** needs a **measure** (or **standard**) by which its value can be determined.

- ❏ The standard could be a **simple measurement unit** e.g. the variable *height of a person* could be measured in centimetres.
- ❏ The standard could be **a scale** e.g. the variable *feelings about the school uniform* could be measured on a scale from 1 to 5 (*1 - unhappy, 2 - somewhat unhappy, 3 - neutral, 4 - reasonably happy, 5 - happy*)
- ❏ The standard may be **a rate** e.g. the variable *muscular strength* could be measured by number of push-ups per minute.

Depending on the standard of measurement used, we can classify variables into **category variables** and **numerical variables**.

A **category variable** (also called **qualitative** variable) is a property that can be organised in distinct subgroups. These subgroups must be defined without overlap so that every value belongs to exactly one subgroup.

A **numerical variable** (also called **quantitative** variable) is a property for which the values result from measuring with a measuring device or from counting. Therefore numerical variables are either **measurement variables** or **whole-number variables**.

Before you decide what type of variable you are dealing with, you need to look at the standard of measurement used. For instance, if *age of a person* is one of the variables of interest, it could be *measured* in years and months, or the number of birthdays celebrated could be *counted*, or the person could select an age-group (*category*) they belong to.

Example : Decide how to measure each variable, and then classify the variable as category, measurement or whole-number variable.

a) gender of patient
b) weight of patient
c) number of days in hospital
d) age-group

Answers :

a) (select male/female) category variable
b) (in kilograms) measurement variable
c) (count the days) whole-number variable
d) (class intervals) category variable

1 Below are variables for a survey about drivers and their cars.

i) Describe a way to measure each variable.
ii) Classify the variable as *category*, *measurement* or *whole-number* variable.

a) manufacturer of the car i) .....
ii) .....

b) number of doors i) .....
ii) .....

c) age of the car i) .....
ii) .....

d) fuel consumption i) .....
ii) .....

e) time taken to drive to work i) .....
ii) .....

2 A manufacturer of hand held vacuum cleaners called 'dust busters' is asking recent customers some questions in a telephone interview. There are 3 questions resulting in 3 variables. Describe the variable and suggest a possible standard by which to measure it.

Questions :
Q1 : *How long ago did you purchase your dust buster?*
Q2 : *How satisfied are you with your dust buster?*
Q3 : *How frequent do you use your dust buster?*

Q1 Variable / Standard .....
.....

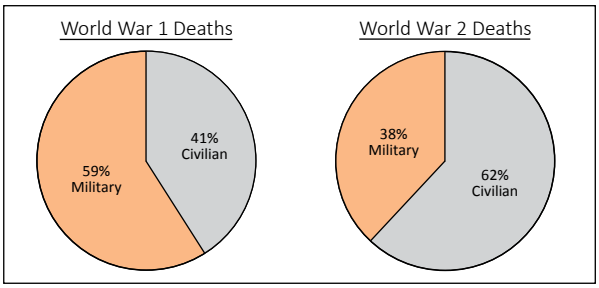
Q2 Variable / Standard .....
.....

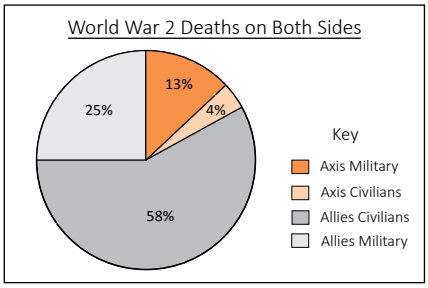
Q3 Variable / Standard .....
.....

7 Appropriate Visualisations 2 - continued

B War Statistics

1 Estimates of the number of deaths in World War 1 and World War 2 vary a lot. These graphs present data that is based on information that can be found in Wikipedia. In World War 2 the 'allies' were Britain, Russia, America and their supporters. The 'axis' were Germany, Japan, Italy and their supporters.





a) Use these graphs to compare deaths between the two world wars and to compare deaths between the ***Allies*** and the ***Axis*** during World War 2.

.....

.....

.....

.....

.....

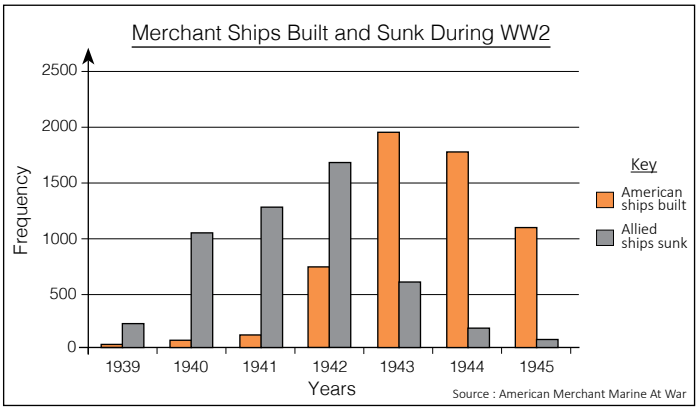
.....

b) What other statistical information would be useful when making these comparisons?

.....

.....

2 This graph shows the numbers of Allied (mainly British and American) merchant ships that were sunk each year during the Second World War (1939-1945). It also shows the number of merchant ships that were built in America over the same period of time.



a) How did the number sunk and the number built in 1941 compare with the number sunk and the number built in 1944?

.....

.....

.....

.....

.....

b) The fact that American ships could get to Britain was a key factor in the ability of Britain to continue fighting the war. Discuss how the graph shows how the number of merchant ships built in America and the number of sinkings changed as the war progressed.

.....

.....

.....

.....

.....

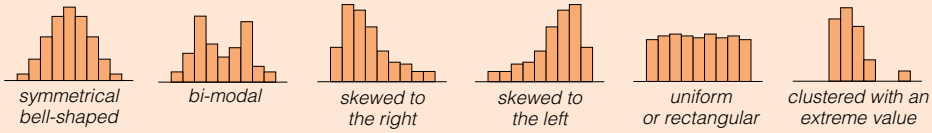
23 Data Display 3

A Describing a Distribution

The shape of a histogram reveals the distribution of the data.

Interesting details of the distribution could be the peak score (mode), gaps, extreme values, or clusters.

Vocabulary to describe the shape of distributions :



1 The histogram shows weights of checked in luggage on an Air New Zealand flight.

a) Name and describe the shape of the distribution. .....

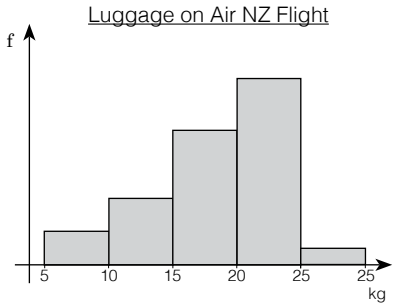
.....

b) Can you think of a reason for the data to have this shape? .....

.....

.....

.....



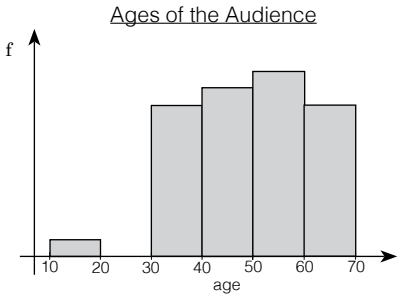
2 Ages of the audience of Shakespeare's play *'Hamlet'* are displayed in a histogram. Describe the shape and interesting features of the distribution.

.....

.....

.....

.....



3 Students of Year 11 at Bay View College drew the four graphs i) to iv) below. The four graphs have no titles, no labels, no scales. In the grey boxes are the titles, together with descriptions of the graph.

a) For each graph select the correct title - A, B, C or D.

b) Place labels and scales on the horizontal axes.

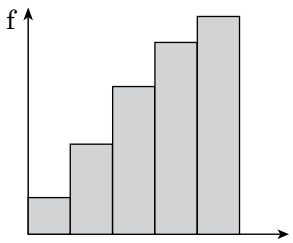
A Travel Time from Home to School
Times are clustered between 0 and 30 minutes with an extreme time of 40-50 minutes.

B Number of Students per Class
A triangular distribution, with class sizes ranging from minimum 5 students to maximum 30 students.

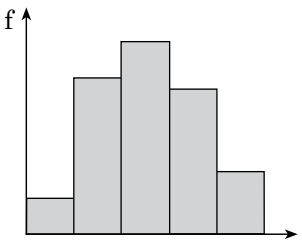
C Time Students Spend in the Library at Lunchtime
A bimodal distribution with most common times 0-5 mins and 15-20 minutes.

D Heights of Year 11 Boys
Almost symmetrical bell shaped distribution centred at 170-180 cm, ranging from 150 cm to 200 cm.

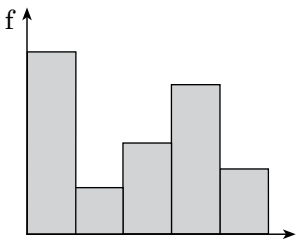
Graph i) ; Title



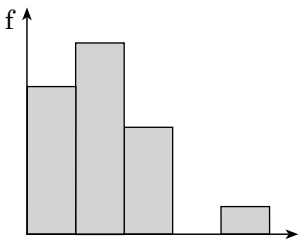
Graph ii) ; Title



Graph iii) ; Title

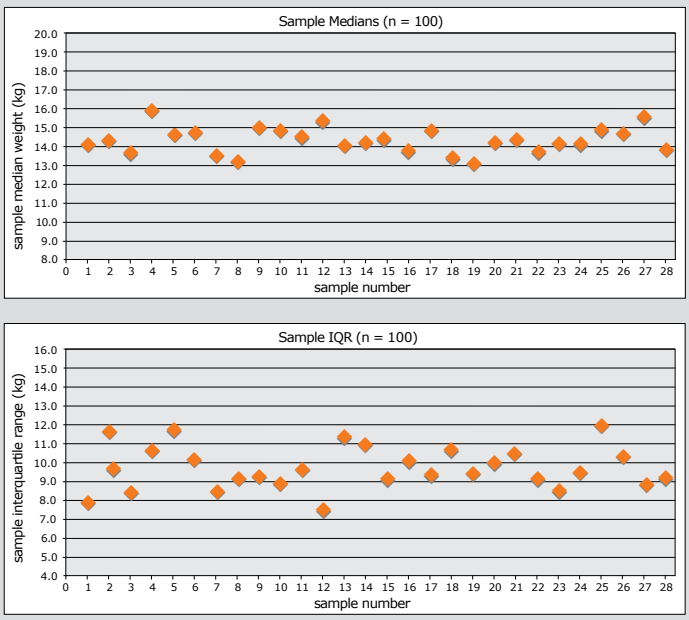


Graph iv) ; Title



A Same Population, Different Samples - continued

3 Students have been told that a sample of size 100 is more reliable, so they each take another sample, this time of size 100 and calculate the sample median and IQR. Another set of graphs is constructed on a spreadsheet.



a) Compare these graphs with the two on page 26.
Write about differences and similarities.

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

b) Explain what is meant by the statement : *"A sample of size 100 gives more reliable results than a sample of size 30"*.

.....

.....

.....

.....

.....

4 The graphs indicate a lot more variation in the interquartile ranges than in the medians. What could be the reason for that?

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

Conclusion : There is a fair amount of variation in sample statistics, even if the sample size is 100. If a new random sample was taken the box-and-whisker plot would look different, the median would be in a slightly different position and the length of the box would be longer or shorter.

This picture gives you an idea of the distortion :

53 Looking at Your Scatter Plot 1

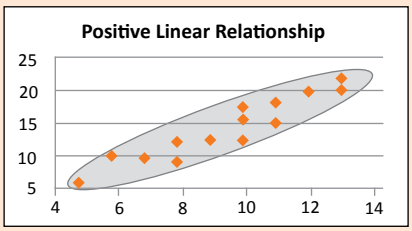
A The Nature of the Relationship

Now is the time to **analyse** your data. You need to describe *patterns* that you can see in your data.

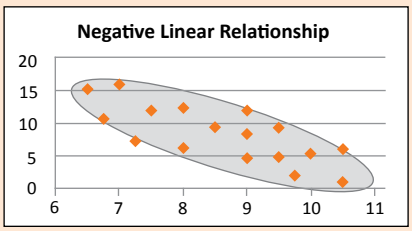
Start by drawing an oval (or band) that encloses most of the scattered points and comment on the **general trend** (or **nature of the relationship**) by studying this oval.

a) If you see a *cigar shape*, then there will be a **linear relationship**.
This can be :

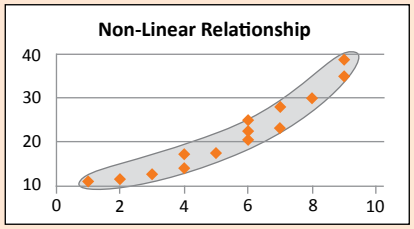
i) a **positive linear relationship**; as one variable increases, the other tends to increase.



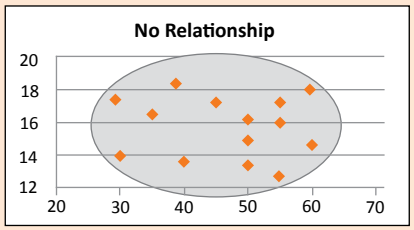
ii) a **negative linear relationship**; as one variable increases, the other tends to decrease.



b) If there seems to be a *curved band*, then the relationship is **non-linear**.

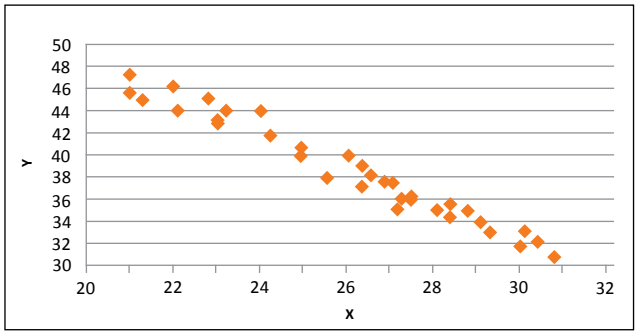


c) If the oval is *almost circular*, then there is **no relationship**.



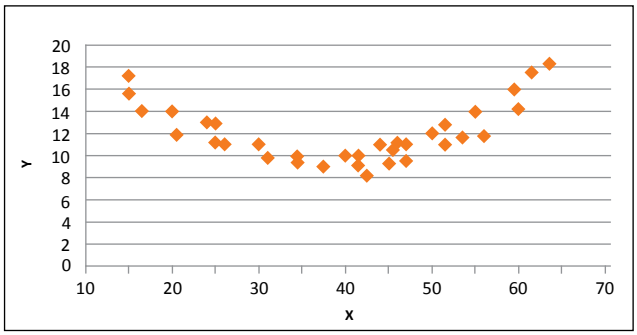
1 Describe the nature of the following relationships :

a)



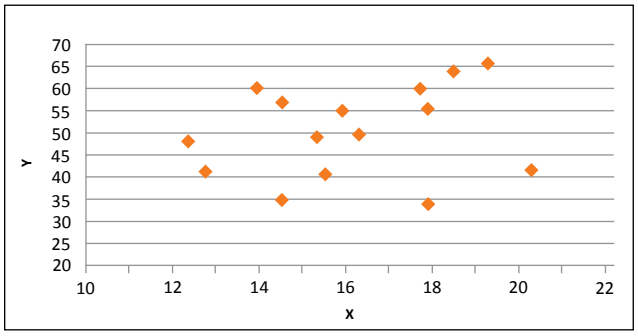
.....

b)



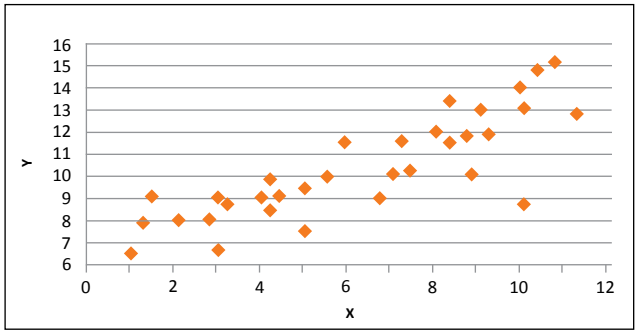
.....

c)



.....

d)

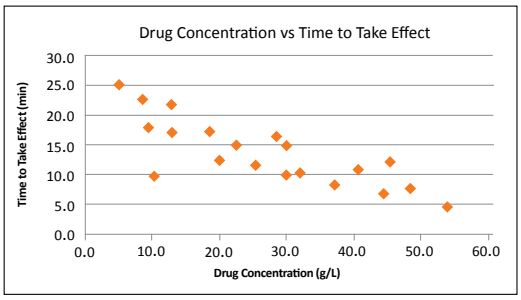


.....

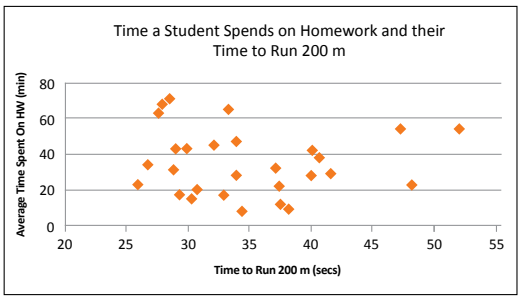
A Making Predictions - continued

For questions d) & e), if it is **sensible**, make and discuss a predition. If it is **not sensible** to make a predition, say why this is the case.

d) The time it takes for a drug at a particular concentration to take effect.

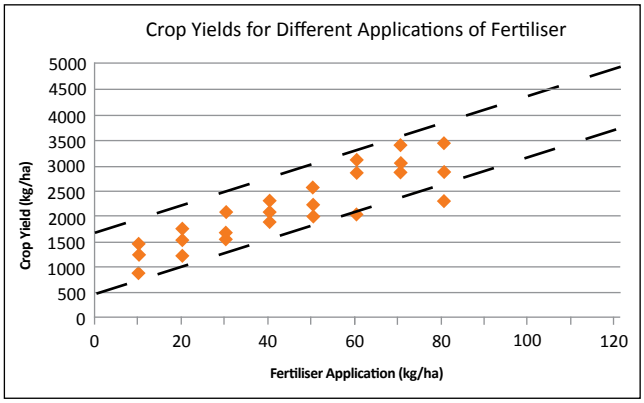


e) How long a student who can run 200 m in a particular time spends on homework.



B Going to Extremes

1 This is the graph for the fertilizer experiment (see page 23). Both the horizontal and vertical axes have been extended.



a) The yields for a wide range of application rates need to be predicted. Predict the yield for an application rate of . . .

i) 45 kg/ha. .....

ii) 85 kg/ha. .....

iii) 110 kg/ha. .....

b) Compare these three predictions, discussing how accurate you think they might be. Give reasons for your answer.

A Plotting a Time Series Graph by Hand - continued

b)

Median Quarterly Earnings of Teenagers Aged 15-19, Working in Retail in NZ, in $ over Five Years					
Q2 2020	Q3 2020	Q4 2020	Q1 2021	Q2 2021	Q3 2021
$3590	$4440	$4480	$3850	$4000	$4810

c)

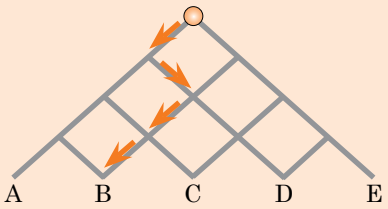
Yearly Greenhouse Gas Emissions from Argriculture in NZ, Kilotonnes of CO_2 equivalent					
1990	1991	1992	1993	1994	1995
33312	33557	33113	33506	34605	35176

Display the Experimental Data

A Tables and Graphs

For your investigation you must select and use at least **one display**.

Example : **Investigative question** : If a marble is released at the top of the grid what are the chances that the marble finishes at each of the points A, B, C, D, E?



Data : The data is a series of letters that the marble ended up at each time it was released. Sixty trials of this experiment produced the following data :

B B C A E	C D B C C	C D B D B
C E D B B	C C C A C	B D B C A
B C D B D	B C C D C	D A C B D
C D B D A	C D D B D	D C C B B

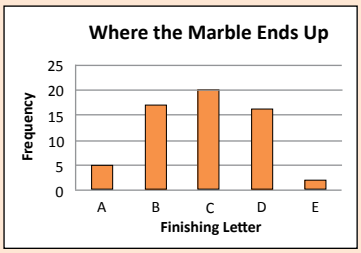
Display the data for this investigation.

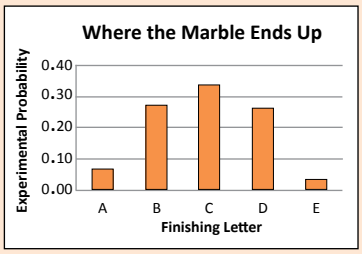
letter	tally	frequency	experimental probability
A	𝍸	5	0.08
B	𝍸 𝍸 𝍸 ||	17	0.28
C	𝍸 𝍸 𝍸 𝍸	20	0.33
D	𝍸 𝍸 𝍸 |	16	0.27
E	||	2	0.03

Possible Displays :

Display One is a tally table showing frequencies and the experimental probability distribution. (The experimental probability distribution is found by dividing each frequency by the total number of trials, in this case 60.)

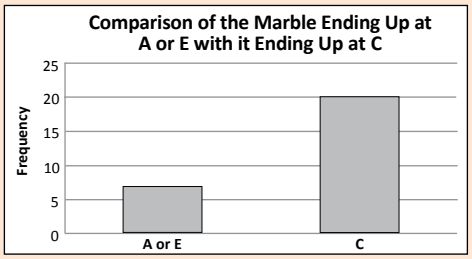
Display Two could be a frequency graph or a graph of the experimental probabilities.





Alternative Display : While a frequency graph using all the outcomes is likely to be best for noticing patterns, other graphs could also be appropriate. These must relate to the investigative question.

For example if the question was *'How does the chance of finishing at A or E compare with the chance of finishing at C?'*, the frequencies for finishing at A or E and the frequency of finishing at C could be graphed.



Display the data for the following investigations (remember the experimental probability distribution **must** be on one of your displays). You can either draw the graphs yourself or paste a print-out of a computer generated one in the box.

1 **Investigative Question :** What is the probability of getting each of the possible number of complete levels on a card house before it collapses?

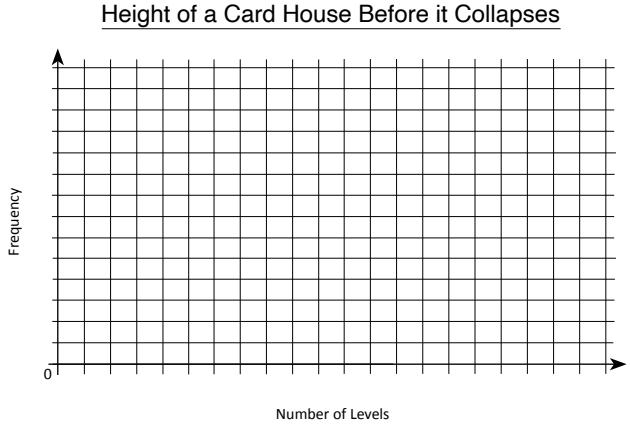
Data : The number of complete levels for 40 trials :

1 0 1 2 3 3 0 2 4 4 4 4 3 2 0 4 4 4 5 3 2 0 5 6 4 3 5 0 4 3 0 4 3 1 3 0 0 4 3 2

Display One :

Nº levels	tally	frequency	experimental probability
0			
1			
2			
3			
4			
5			
6			

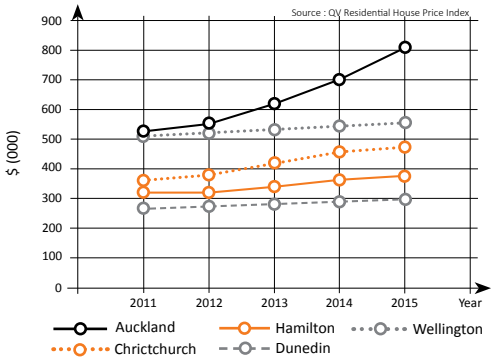
Display Two :



A Auckland House Prices

1 These graphs relate to housing in Auckland.
A claim relating to each graph has been made.
Use information from the graph to discuss each claim.

Average House Price (April)



a) *'The graph of average house prices shows that it is getting more expensive to buy a house in Auckland than in other cities.'*

.....

.....

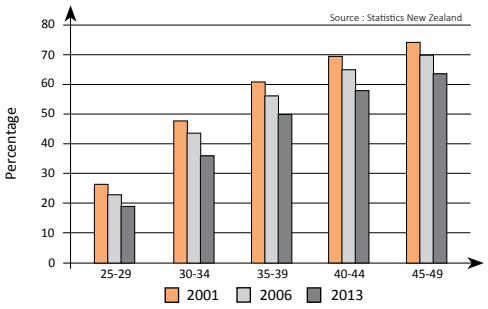
.....

.....

.....

.....

Auckland Home Ownership Rates for Different Age Groups



b) *'The graph of home ownership rates shows that the percentage of Auckland house owners is dropping only for younger people.'*

.....

.....

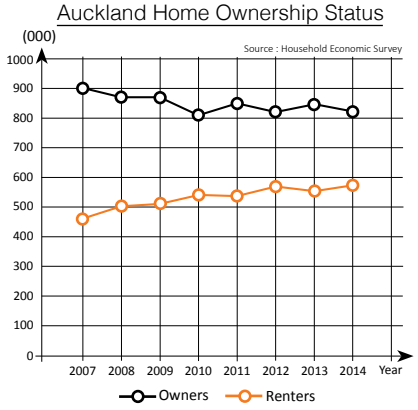
.....

.....

.....

.....

Auckland Home Ownership Status



c) *'The home ownership status graph shows that there has been no change in the number of people owning and the number of people renting in recent years.'*

.....

.....

.....

.....

.....

.....

2 An Auckland real estate agent records the selling price of the last 15 houses that she has sold. Here are her results : $645 000, $748 000, $548 000, $867 000, $603 000, $987 000, $645 000, $445 000, $710 000, $1 243 000, $534 000, $605 000, $625 000, $706 000, $810 000

a) Calculate the mean and median selling price for these houses.

.....

.....

b) Which measure gives the better *'average'* for this data, mean or median? Give a statistical reason for your answer.

.....

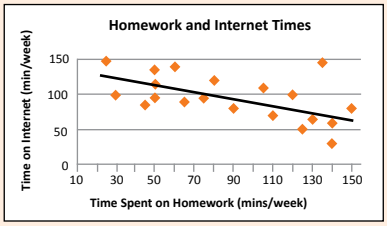
.....

.....

Pages 63-64 - Practice Investigation - continued

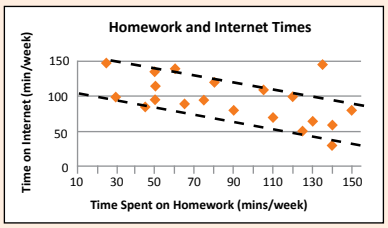
Answers for pages 63-64 are examples only :

A5 Conclusion :



Students who spend more time on homework **tend** to spend less free time on the internet. This can be seen by the **negative gradient** of the trend line. You would expect this because students who spend a lot of free time on the internet would have less time available for other things, for example homework. For student spending 90 minutes on homework per week, we would predict that they would spend roughly between 70 and 120 minutes per week on the internet. As the data is very scattered around the trend line, this prediction is a very rough estimate.

The fact that the data is quite spread from the trend line is not surprising because, for example, you would expect some students to make relatively more time for both homework and the internet.



Page 66-67 - Define the Problem

A1 a) I'm going to study the monthly eftpos spend in $NZD Millions in New Zealand over five years and determine if there are any patterns or trends.
Variables of interest: Time measured in months, Monthly Eftpos Spend in $NZD Millions in New Zealand.

b) I'm going to study the median quarterly earnings of teenagers aged 15 to 19 in New Zealand over five years. I wonder, are there any patterns or trends?
Variables of interest: Time measured in quarters, median quarterly earnings, in $, of teenagers aged 15 to 19 in New Zealand.

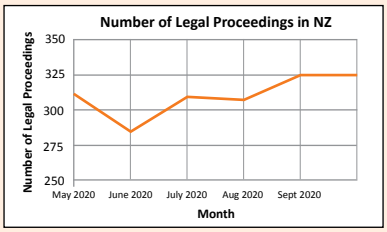
c) Can any patterns or trends in yearly greenhouse gas emissions from agriculture in NZ be identified to help with a better understanding of this issue?
Variables of interest: Time measured in years, gas emissions due to agriculture measured in the equivalent of Kilotonnes of CO2 in NZ.

d) I'm going to study the number of legal proceedings per month in NZ for motor vehicle thefts in New Zealand over six years and determine if there are any patterns or trends.
Variables of interest: Time measured in months, number of legal proceedings for motor vehicle thefts in NZ.

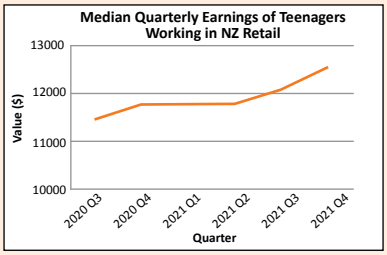
e) Can any patterns or trends in quarterly broadband usage in NZ be identified to help with a better understanding of broadband consumers?
Variables of interest: Time measured in quarters, broadband usage measured in GB in NZ.

Pages 68-69 - Analyse the Data

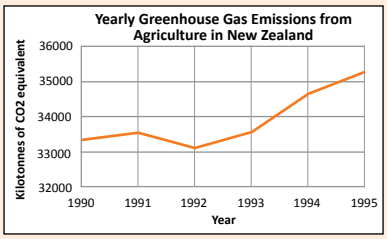
A1 a)



b)

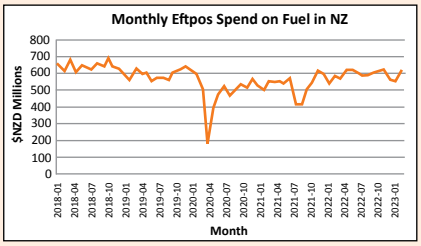


c)



Page 70 - Time Series Graph Using Technology

A1 a)



Pages 71-73 - Time Series Analysis

A1 a) Values of hot pie sales have tended to increase over the years. For example winter values have gone up steadily from about $13 000 to about $29 000.

b) The values of hot pie sales are high in winter (colder weather) and low in summer (hot weather).
The autumn of Year 3 appears to be unusual in that the value of sales is lower than what would be expected.

B1 a) The long term trend for both Retail and Fuel Sales is increasing. The Retail data is increasing at a steadier rate than the Fuel Sales which is more erratic.

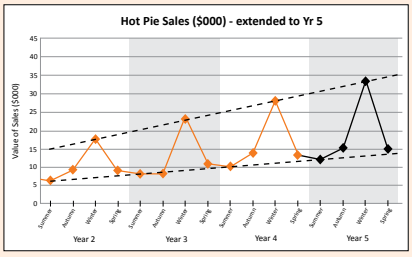
b) Both datasets show a regular seasonal peak in the fourth quarter, with this being more regular in the Retail Sales data. Both sets of data have significant troughs in the third quarter of 2020.

B2 a) Both the male and female sets of data show a general decreasing trend for unemployment. However this trend is not consistent, with the female data being more irregular than the male data.

b) There is no obvious seasonal pattern to either the male or female unemployment data. There is a significant peak in the unemployment data for both datasets in the third quarter of 2020, the female data shows a larger relative peak than the male data.

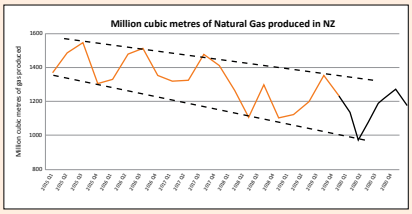
Pages 74-75 - Time Series Forecasting 1

A1 a)



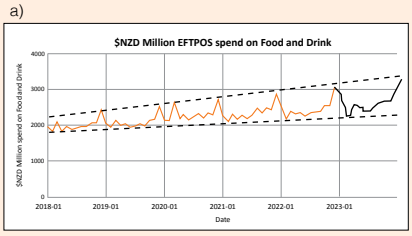
b) About $34 000. I would be relatively confident in this prediction because the trend is constant and the seasonal effects are regular. Also the prediction is for a time that is not very far into the future.

A2 a)



b) I forecast that the Natural Gas production for Q4 2020 might be around 1300 Million cubic metres. I am not very confident in my forecast as the pattern for this time series is quite erratic.

A3 a)



b) I forecast that the eftpos spend on food and drink in NZ will be roughly $3300 Million in December 2023. I am fairly confident in this forecast as the pattern for December is quite consistent in being a significant peak in the data, the long term trend is also a steady increase.

Page 76 - Time Series Forecasting 2

A1 a) Doug's earnings tended to increase over the five years. For example in the spring of Year 1 he earnt about $17 000 and from then spring earnings increased steadily to about $30 000 in the spring of the fifth year.

b) There are seasonal highs in spring (when things start growing) and lows in winter (when there is little growth). There are also secondary highs in autumn (pre-winter clean-ups). Doug's earnings in autumn of the second year appear to be a little lower than what would be expected.

c) (see graph for this answer on page 76)

d) $34 000. You can be reasonably confident in this prediction because of the consistency in the trend and seasonal patterns and also because the prediction is not too far into the future.

Pages 78 - Form a Conclusion

A1 a) Overall the monthly eftpos spend on fuel has gradually decreased over the ten years to 2023, there are no seasonal patterns in the monthly data however there are two significant troughs in the data in 2020 and 2021. As a further investigation, it could be interesting to look at how fuel prices have changed over that time.