Bivariate Data

Level 1

As 91036

3 Credits

This material has come from the Nayland school website

<http://maths.nayland.school.nz/Year_11/Level_1_home.htm>

and is rated PG (pretty good) by your maths teacher

. The Statistical Question .

A research question needs to be carefully planned.

What am I interested in, and what is a good research question?

Given some data, what is a good research question?

Start with “I wonder if there is a relationship between…”

Example “I wonder if there is a relationship between the number of texts sent and time talking on a cell phone for students in year 11 at Reporoa College”

Notice: 1) the underlying population is specified

 (year 11 students from Reporoa College)

 2) there are two variables – number of texts & time talking

To answer this problem collect data by taking a sample (small group of the population)

If there are not suitable variables then change your statistical problem

. Write a research question .

Given this data what research questions are possible?

Pam’s Pork Pies

|  |  |  |  |
| --- | --- | --- | --- |
| Pie | Pie Weightg | Pie VolumemL | Pie Diametercm |
| A | 377 | 157  | 5.1 |
| B | 392 | 165  | 5.3  |
| C | 338 | 134  | 4.8  |
| D | 333 | 132  | 5  |
| E | 418 | 172  | 5.9  |
| F | 402 | 167  | 5.3  |
| G | 319 | 118  | 5.2  |
| H | 324 | 128  | 4.9  |

What are the three research questions possible?

“I wonder if there is a relationship between ……….…… and ……………… from the 2009 Census at school Database”

. Variables .

Defining the variables used:

Variables need to be specified, with units and causality considered.

Eg Student height is measured in…

Student arm span is measured in…

There is probably an 'association' between the student height and arm span.

In this case the variables are probably neither ‘explanatory’ nor 'dependant'

Sometimes one variable probably causes the other to change

Eg: Fat content (g) in take away food, and the energy content (calories)

. Making a Plan .

“I wonder if there is a relationship between the number of texts sent and time talking on a cell phone for students in year 11 at Reporoa College”

What data would be useful to help answer this problem?

Is there a website to find this data or do we collect our own?

Websites: Census at School, or the Gamma maths data sets from [www.mathematics.co.nz](http://www.mathematics.co.nz)

Collect your own? Design a suitable survey, with quantitative questions (NOT descriptive answers)

Good question : How many texts did you send in the last 7 days?

DON’T USE: Do you use your cell phone: little, a bit, sometimes, often, lots,

. Cleaning the data .

“I wonder if there is a relationship between the number of texts sent and time talking on a cell phone for students in year 11 at Reporoa College”

|  |  |
| --- | --- |
|  | In last 7 days, Number of: |
| Gender | Texts sent  | Texts Received | Calls made | Calls Relieved |
| B | 14 | 10 | 3 | 6 |
| B | 241 | 183 | 24 | 48 |
| B | 45 | 37 | 30min | 45min |
| B | 27 | 100000 | 12 | 8 |
| G | 211 | 317 | 16 | 10 |
| G | -17 | 37 | 3 | 9 |
| G | 316 | 426 | n/a | 0 |
| G | 89 | 95 | 16 | 2.4 |

We ‘clean’ the data by removing individual data values that are errors

These could be typing errors, silly entries, incorrect units eg grams & Kg

Check ‘0’ entries carefully. Is it a zero or n/a entries

Sometimes more data will be needed to get a sample set of 30 data values

Only some of the available data will be needed!

. Collecting the data .

“I wonder if there is a relationship between the number of texts sent and time talking on a cell phone for students in year 11 at Reporoa College”

Describe how you will collect your data.

. Scatterplots (A) .

When investigating bivariate data making a scatterplot is the first step.

What types of scatterplot associations are possible?

How do we make a scatterplot using Excel or iNZight?

If the scatterplot indicates a linear model is appropriate then we can proceed.

One variable could be influencing (or be an influence in causing) the other variable. This is called the 'explanatory' or 'independent' or 'predictor' variable (should be plotted on the 'x' axis)

The other variable is the 'dependent' or 'response' variable
(plotted on 'y' axis)

If neither variable 'causes' the other then either
variable can be plotted on the 'x' axis.
There may be an ‘association’

. Making a Scatterplots .

When plotting graphs in Excel this is what to do:

1) Have the data explanatory (x) on left and dependent variable (y) on the right

2) Highlight both data sets (click and drag)

3) 'Insert' --> Scatterplot (under chart tab)

4) Click on the graph then the 'chart tools' --> Design --> Add axis labels

5) Label the axis with units.

6) Right click on the data values and 'add trendline, 'linear' & Show equation (tick box)

|  |  |
| --- | --- |
| Explanatory variable | Dependent Variable |
| **Carbohydrate total (g)** | **Energy (kJ)** |
| **28.7** | **1070** |
| **29.2** | **1280** |
| **37.6** | **1435** |
| **47.7** | **2086** |
| **21.8** | **1164** |
| **41** | **2414** |
| **etc** | **etc** |



**Data – Using iNZight**

How to plot graphs using iNZight

|  |  |
| --- | --- |
| First up we need to start iNZight by clicking on the shortcut in the Level 3 maths folder |  |
| After it has had some time to think it will open up a window that looks like this.To start off with we need to open the iNZight window by clicking on the ‘Run iNZight’ button (circled) |  |
| This will bring up the main iNZight window that looks like this. We then want to import the data, by clicking on ‘Data IN/OUT’ and ‘Import Data’. Browse for the right file in the Level 1 or Level 3 folder and follow the prompts, and once imported it should look like this. |  |
| Often it is useful to get an overview of the whole dataset. To do this go to ‘Advanced’ and ‘Scatter Plot Matrix’. This will give the window shown on the right where I have selected all of the variables apart from the name of the car. |  |
| Pressing plot will give us a graph of all of the possible combinations of variables like the one show on the right.From this we can work out what there is a relationship between, and what variables might be appropriate to use.  |  |
| Once we have determined the variables we would like to investigate you drag the title of each variable onto ‘Variable 1’ and ‘Variable 2’. Your independent variable goes onto variable 1 and your dependent variable goes onto variable 2.*Note: you can make it easier to see all the variables by clicking on the ‘View Variables’ button.*  |  |
| You also need to add a trend curve to the plot (we normally use a linear trend). You do this by clicking add to plot, and selecting add trend curve. After this you will need to press on ‘OK’ to get the options below to open. Tick the box next to linear as show to the right, and then you can close the window. |  |
| The final thing you need to do is get the summary, you get this by pressing the ‘Get Summary’ button.*Note: you can make it easier to copy the graph by clicking on the ‘New’ button under the plot which will open the graph in a new window which can be right clicked on to copy. Bitmaps normally work best when copying into a word document, as metafiles sometimes will lose some of the plotted points.* | Weight = 217.97 \* Engine\_size + 812.32Correlation = 0.8451 |

**Note: The axis titles given on iNZight are not sufficient, so once inserted into your document you will need to put a text box over the top of them to include a proper title and axis labels (including units)**

**. Describing Scatterplots .**

What do I look for in scatter plots?

**Trend**

 Do you see: a **linear** trend… **or** a **non-linear** trend?

 Do you see: a **positive** association… **or** a **negative** association?

as one variable gets bigger, the other gets smaller

as one variable gets bigger, so does the other

**Scatter**

Do you see:

a **strong** relationship… **or** a **weak** relationship?

lots of scatter

little scatter

Do you see: **constant** scatter… **or non-constant** scatter?

the scatter looks like a “fan” or “funnel”

roughly the same amount of scatter as you look across the plot

**. Scatterplots (cont) .**

When looking at a scatter graph we can ask two questions:

Is there a relationship between the two data variables?

How close does the data fit this relationship?

Eg. “I notice that there is a positive relationship between ........... and ............

This means that as .......... increases ........... also increases.

The relationship is strong as there is little scatter in the data points”

**. Even More Describing Scatterplots .**

**What do these scatterplots tell us?**

**1) 2) 3) 4)**

**5) 6)**

**. More Describing Scatterplots .**

**What do I look for in scatter plots?**

**Anything unusual**

Do you see any **outliers**?

Unusually far from the trend

Do you see any **groupings**?

**What do I see in these scatter plots?**

45

40

35

20

19

18

17

16

15

14

Latitude (°S)

Mean January Air Temperatures
for 30 New Zealand Locations

Temperature (°C)

There appears to be a linear trend.

There appears to be moderate constant

scatter about the trend line.

Negative Association.

No outliers or groupings visible.

Internet Users (%)

0

10

20

30

40

GDP per capita (thousands of dollars)

0

10

20

30

40

50

60

70

80

% of population who are Internet Users vs
GDP per capita for 202 Countries

There appears to be a non-linear trend.

There appears to be non-constant.

scatter about the trend line.

Positive Association.

One possible outlier (Large GDP, low % Internet Users).

Year

1990

1980

1970

1960

1950

1940

1930

30

28

26

24

22

20

Age

Average Age New Zealanders are First Married

Two non-linear trends (Male and Female).

Very little scatter about the trend lines

Negative association until about 1970, then a positive association.

Gap in the data collection (Second World War).

. Outliers (A) .

Outliers: How do I handle them?

 Integral part of data set?

 Mis-recorded or wrongly included?

Auckland Sydney air ticket price for 10 passengers of different

weights

|  |  |
| --- | --- |
| Weight (kg) | Price Paid ($) |
| 120 | 277 |
| 75 | 277 |
| 126 | 281 |
| 105 | 277 |
| 116 | 281 |
| 220 | 530 |
| 88 | 290 |
| 92 | 277 |
| 71 | 281 |
| 83 | 286 |
| 89 | 264 |

Including Outlier

 Further information needed

 strong positive

 Association (sensible?)



Excluding Outlier

No outlier = weak

 negative association

= A Lack of Resistance - So the single outlier has a significant influence on the relationship

. Outliers (B) .

An Outlier is a data point significantly far from the trend or group of other data.

Is the outlier an error? 🡪 remove it

Is the outlier an actual data value? 🡪 produce TWO relationships

 1) With outlier

 2) Without outlier: is there significant variation

Is it appropriate to remove outlier?

Do NOT remove outlier just to make model fit better!

Discuss what the outlier could be and if it is appropriate to remove or not.

Discuss Groups – is there a reason for groupings or is it just a result of the sampling process?

. Fitting a line of fit .

A line can be drawn to ‘fit’ the relationship

Using Excel – right click on a data point 🡪 Add Trend line 🡪 tick the ‘show equation’ box near the bottom

The equation line can then be used to predict ‘y’ and ‘x’ values

Eg

If Variable ‘A’ = 10, predict a value for ‘B’

 y = 1.026x - 0.3249

so B = 1.026×10 - 0.3249 = 9.9351 which should be rounded sensibly

 B = 9.9 and **remember units**

. Fitting a line of fit B .

Predicting ‘x’ from a given ‘y’ is involves solving an equation.

Eg

If Variable ‘B’ = 14, predict a value for ‘A’

 y = 1.026x - 0.3249

so 14 = 1.026×A - 0.3249

14.3249 = 1.026×A

13.96189 = A

Round sensibly A = 14

Check the if the solution matches the graph

. Fitting a line of fit C .



1) If a student had a Maths score of 14, predict their English Score.

2) If a student had a Maths score of 6, predict their English Score.

3) If a student had a English score of 8, predict their Maths Score.

4) If a student had a English score of 16, predict their Maths Score.

5) How accurate are these predictions?

. Population Inference A .

We use the sample statistics and graphs to make statements about the population.

We need to be aware that the sample will vary from the population

If the sample shows that there is a relationship between the two variables AND the sample size is reasonable (at least 30) then we can say there probably is a relationship in the population as well.

Remember the relationship can be…

Linear or Non Linear, Positive or Negative,

Strong or Weak, with Constant Scatter or Non-Constant Scatter.

**To ensure that a sample represents a population we need a sample size of at least 30**

. Calculating Sample Statistics .

Calculating ‘Sample Statistics’ may be useful as well.

Use Excel to calculate the mean, median, mode, maximum, and minimum for the data.

You will have to calculate the sample statistics for EACH of the variables.

 =Average(click and drag over the data cells)

 =Median(click and drag over the data cells)

 =Mode(click and drag over the data cells)

 =Maximum(click and drag over the data cells)

 =Minimum(click and drag over the data cells)

=Range(click and drag over the data cells)

=Quartile(click and drag over the data cells,2)

=Quartile(click and drag over the data cells,4)

Interquartile Range

Mean shown as $\overbar{x}$

Mode as mod

Median as med

Lower quartile Q1

Upper quartile Q3

Maximum value max

Minimum value min

* Select 2 from the menu to get STAT.
* Delete any existing data by pressing F6 F4 F1 {(more) DEL\_A YES] and repeat if required
* Move the cursor to column one and enter the data values.

**12, 15, 21, 14, 13, 12, 11, 24, 19, 17, 18, 17, 21, 12,**

* Select F2 F6 (CALC, SET) and check that 1Var Xlist: list1 and 1Var Freq: 1
* Then press EXE F1
* Scrolling down what summary statistics have you been given?

**Using your calculator to find summary statistics.**

Interquartile range = Q3 – Q1

Range = maximum - minimum



. Conclusion .

In your conclusion you MUST refer back to your research question

To answer the research question use…

“I notice that…. “

Talk about :

The relationship (State the obvious “as height of students increases, arm span also increases) and Strong/Weak/Linear/Non-Linear/Constant Scatter /Non-Constant

The trend line (slope and equation) what it means and how well it represents the data (for all or part of the data range)

How the trend line can be used to make predictions and how appropriate these predictions are.

Any outliers or groups and what may have caused this.

The summary statistics referring to central tendency and spread

How well (or not) the sample represents the population.

Were there any bias, limitations, errors, improvements that were possible?