General Maths Unit 3

Core Topic: Data Analysis
Chapter 2 Part 1

Investigating Associations between two variables

Area of Study 1 – Data Analysis

- response and explanatory variables and their role in investigating associations between variables
- contingency (two-way) frequency tables, two-way frequency tables and their associated bar charts (including percentaged segmented bar charts) and their use in identifying and describing associations between two categorical variables
- back-to-back stemplots, parallel dotplots and boxplots and their use in identifying and describing associations between a numerical and a categorical variable

Bivariate Data – Classifying variables (Ex 2A)

- So far we have learned how to display, describe and compare the distributions of single variables. This is univariate data.
- A set of data involving two variables where one affects the other is called **bivariate data**

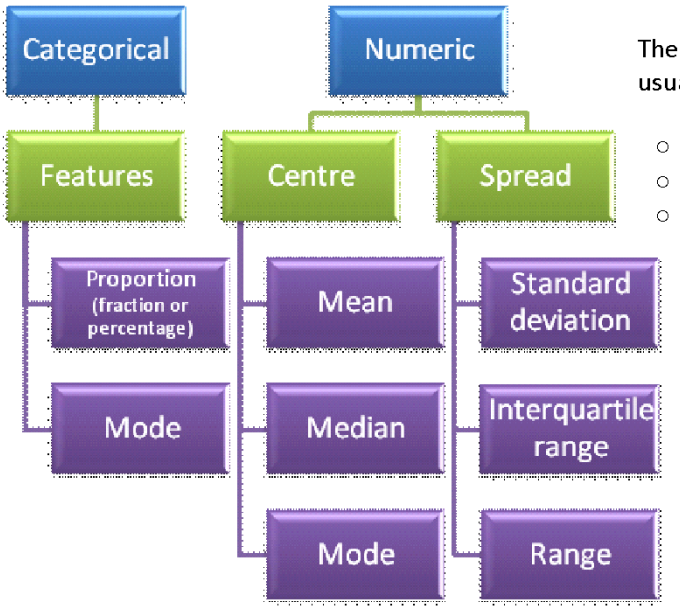
We look at the relationship between the 2 variables
Eg. Attendance at the zoo and weather conditions.

Types of Data

Remember there are 2 types of data:

1. Numerical – data values that are counted
2. Categorical – data values that are measured.

Analysing data



The following summary statistics are usually used together:

- Mean and standard deviation
- Median and IQR
- Mode and range

Statistic	Term number
Q1	$\frac{n+1}{4}$
Median	$\frac{n+1}{2}$
Q3	$\frac{3(n+1)}{4}$

Classifying the Variables (Ex 2A)

Response & Explanatory variables

- A set of data involving two variables where one affects the other is called **bivariate data**
- If the value of one variable "responds" to the values of another variable, we call it the **response variable**
- The **explanatory variable** influences the other variable.

When a relationship between 2 sets of variables is being examined, it is useful to know if one of the variables depends on the other.

Graphically
If one of the variables influences the other, then the variable being influenced is said to be the ***response (dependent) variable*** (y). The variable causing the influence is said to be the ***explanatory*** (***independent variabl***e) (x).

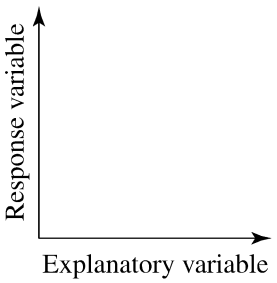
Eg 1. *The weight of a person compared to their height.*
The weight depends on their height therefore, it is being influenced and is the response variable. The height influences the weight, therefore it is the explanatory variable.

Eg 2. *The height of a company's employees and their annual salaries*.
One variable does not influence the other, therefore it is not appropriate to call one the response and the other the explanatory.

Eg 3. *The age of a company's employees and their annual salaries*.
There may be a reasonable relationship – as the age of a person increases so does their salary. Therefore the age causes the influence and is the explanatory variable. Their salary depends on or is influenced by their age, therefore it is the response variable.

Summary
When investigating the association between two variables the **explanatory variable (EV)** is the variable we expect to explain or predict the value of the **response variable (RV).**

We can not say the explanatory variable (EV) causes the response variable (RV). We say it influences / effects the dependent variable.



Ex 2A
Investigating Associations Between Categorical Variables (Ex 2b)

DISPLAYING BIVARIATE DATA

Bivariate data can be displayed in a number of ways:

- **Back to back stemplots:**
  Displays bivariate data involving a **numerical variable** and a **categorical variable** with **2 categories.**

- **Parallel boxplots**
  Displays a relationship between a **numerical variable** and a **categorical variable.** However, the categorical variable can have 2 or more categories.

- **2 way frequency tables:**
  A table that examines the relationship between **2 categorical relationships.** A 2 way frequency table can then be drawn as a segmented bar chart.

- **Scatterplots:**
  Graphical display of **2 numerical variables**.

Exam Questions Relating to Displaying Bi-variate data:

2003

The relationship between **resting pulse rate** (in beats per minute) and **fitness level** (below average, average, above average) is best displayed using

A. a histogram.
B. a scatterplot.
C. a time series plot.
D. parallel boxplots.
E. back-to-back stemplots.

2007

The relationship between the variables

size of car (1 = small, 2 = medium, 3 = large)

and

salary level (1 = low, 2 = medium, 3 = high)

is best displayed using

A. a scatterplot.

B. a histogram.

C. parallel boxplots.

D. a back-to-back stemplot.

E. a percentaged segmented bar chart.

2003

The level of water usage of 250 houses was rated in a survey as low, medium or high and the size of the houses as small, standard or large. The results of the survey are displayed in the table below.

Level of water usage	**Size of house**		
	small	standard	large
low	15	14	9
medium	22	71	11
high	15	47	46

The variables, **level of water usage** and **size of house**, as recorded in this survey, are

A. both numerical variables.

B. both categorical variables.

C. neither numerical nor categorical variables.

D. numerical and categorical variables respectively.

E. categorical and numerical variables respectively.

2011

Question 4

The variables

region (city, urban, rural)

population density (number of people per square kilometre)

A. are both categorical.
B. are both numerical.
C. are categorical and numerical respectively.
D. are numerical and categorical respectively.
E. are neither categorical nor numerical.

Investigating Associations between Categorical Variables

- **We use two way frequency tables and segmented bar charts to investigate associations between categorical variables**

Two way (contingency) frequency tables

The 2 way frequency table is an excellent tool for examining the relationship **between 2 categorical relationships**.

The example below shows how a person's attitude to gun control is associated with their gender.

Table 3.1

		Sex	
	Attitude	*Male*	*Female*
RV	For	32	30
	Against	26	12
	Total	58	42

EV

- In this example, the respondent's gender is referred to as the **explanatory variable**, and the attitude to gun control is the **response variable**.
- When constructing two way tables the **RV defines the rows** and the **EV the columns**.

From the table above, we can see that more males than females favoured gun control. However, this doesn't tell us very much at all because there were more males in the sample.

To solve this problem we turn our table entries into percentages.

Table 3.2

	Sex (%)	
Attitude	*Male*	*Female*
For	55.2	71.4
Against	44.8	28.6
Total	100.0	100.0

****We calculate column percentages as the EV is shown in the columns*

The Big Question – Association or No Association????

Identifying and describing associations between two categorical variables from a two-way table – often used in exam questions

- If there is no association between attitude to gun control and sex, approximately equal percents of males and females would be "for" gun control.
- If there is an association between attitude to gun control and sex, there would be a notable difference between the the females and males "for" and "against" gun control.
- Usually you will be asked to use one of the variables in your answer i.e. just the fors or the against
- When drawing conclusions and describing associations actual values must be used!!

Example

Example 3 **Identifying and describing associations between two categorical variables from a two-way table**

A survey was conducted with 100 people.

As part of this survey, people were asked whether or not they supported banning mobile phones in cinemas. The results are summarised in the table.

Ban mobile phones	*Sex*	
	Male	*Female*
Yes	87.9%	65.8%
No	12.1%	34.2%
Total	100.0%	100.0%

Is there is an association between support for banning mobile phones in cinemas and the sex of the respondent? Write a brief response quoting appropriate percentages.

Example

Example 4 Identifying and describing associations between two categorical variables from a two-way table (no association)

In the same survey people were asked whether or not they supported Sunday racing. The results are summarised in the table.

	Sex	
Sunday racing	*Male*	*Female*
For	55.6%	54.5%
Against	44.4%	45.5%
Total	100.0%	100.0%

Is there an association between support for Sunday racing and the sex of the respondent? Write a brief response quoting appropriate percentages.

Example (2021 NH)

Use the following information to answer Questions 8 and 9.

The *blood pressure* (normal, high) and the *age group* (60–69 years, 70–79 years, 80+ years) of a group of people were recorded. The results are shown in the two-way frequency table below.

Blood pressure	Age group		
	60–69 years	70–79 years	80+ years
normal	42	23	11
high	38	37	29

Question 8

In this group of people, the percentage of people who have high blood pressure and are in the 70–79 years age group is closest to

A. 21%
B. 23%
C. 35%
D. 37%
E. 38%

Question 9

In this group of people, the percentage of people in the 60–69 years age group who have normal blood pressure is closest to

A. 23%
B. 38%
C. 42%
D. 48%
E. 53%

Ex 2B Question 1

1 The following data were obtained when a sample of 20 Year 12 students were asked if they intended to go to university (*university*). The *gender* of the student was also recorded.

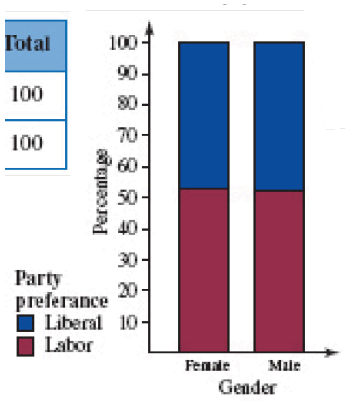
Student No.	*Gender*	*Intends to go to university*	Student No.	*Gender*	*Intends to go to university*
1	F	Yes	11	F	Yes
2	M	Yes	13	M	Yes
3	F	No	13	F	No
4	F	Yes	14	F	Yes
5	M	No	15	M	No
6	M	Yes	16	M	Yes
7	F	Yes	17	F	Yes
8	M	No	18	M	No
9	F	No	19	F	No
10	F	Yes	20	F	Yes

a Identify which is the explanatory and which is the response variable.

b Create a two-way frequency table from the data, with the values of the explanatory variable labelling the columns.

Segmented Bar Charts

- The graphical method to represent **two categorical variables**
- Each column in the bar chart matches a column in the table
- The vertical axis shows a percentage
- The column is divided up into the proportion of each group
- A key is vitally important



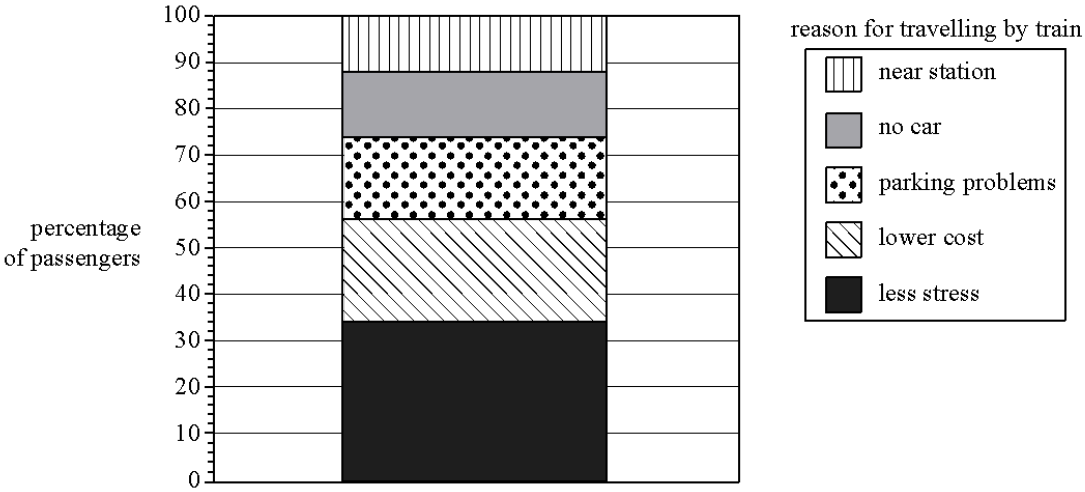
-

Exam Question:

2010

Question 4

The passengers on a train were asked why they travelled by train. Each reason, along with the percentage of passengers who gave that reason, is displayed in the segmented bar chart below.



The percentage of passengers who gave the reason 'no car' is closest to

A. 14%
B. 18%
C. 26%
D. 74%
E. 88%

2012

Question 7

The table below shows the percentage of students in two age groups (15–19 years and 20–24 years) who regularly use the internet at one or more of three locations.

- at home
- at an educational institution
- at work

Location of internet use	Age group	
	15–19 years	20–24 years
at home	95%	95%
at an educational institution	85%	18%
at work	38%	74%

For the students surveyed, which one of the following statements, by itself, supports the contention that the location of internet use is associated with the age group of the internet user?

A. 85% of students aged 15–19 years used the internet at an educational institution.

B. 95% of students aged 15–19 years used the internet at home, but only 38% of 15–19 year olds used it at work.

C. 95% of students aged 15–19 years used the internet at home and 18% of 20–24 year olds used the internet at an educational institution.

D. The percentage of students who used the internet at an educational institution decreased from 85% for those aged 15–19 years to 18% for those aged 20–24 years.

E. The percentage of students who used the internet at home was 95% for those aged 15–19 years and 95% for those aged 20–24 years.

Edrolo link: https://edrolo.com.au/s/2678922/

Ex 2B

Investigating the association between a numerical and a categorical variable (Ex 2c)

Back to back stem plots

- If two variables are related or linked in some way, we say they are associated.
- Associations between the two variables can be identified by comparing the way in which the distribution of the numerical variable changes between categories in terms of **shape, centre and spread.**

A back to back stemplot is used to display bivariate data **involving a numerical variable and a categorical variable with 2 categories**.

Key: 1|2 = 12

Leaf Boys	Stem	Leaf Girls
3 3 3 2	1	2
5 5 4 4 4	1	5
6	1	6 6 6 7 7
	1	9 9 9

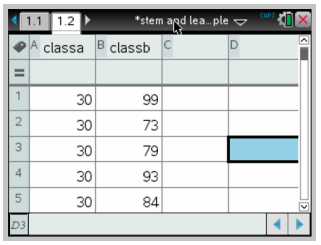
Summary Statistics are:

The information that gives a quick and simple description of the data.

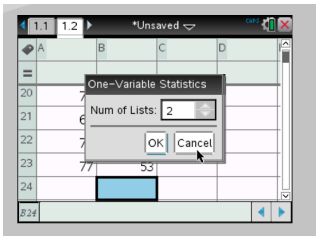
Can include mean, median, mode, minimum value, maximum value, range, standard deviation, etc.

Using CAS for Summary Statistics of two categories

1. On a lists and spreadsheets page enter the data into column A and B, naming each group as shown

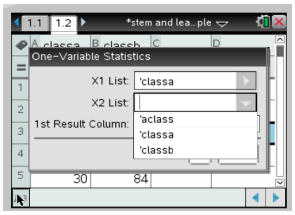


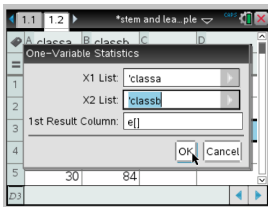
2. Then press Menu
   - 4. Statistics
   - 1. Stat Calculations
   - 1. One-Variable statistics
3. Change the number of lists to 2, as shown



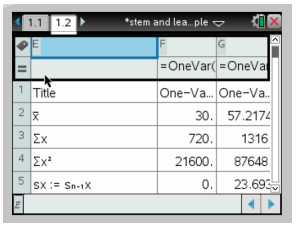
4. Press OK

To calculate two sets of summary statistics, complete the table as shown. Remember to press tab to move between each section, and to scroll up!!!





5. Press OK
6. Results!!!



Example 2

Using a back-to-back stem plot to identify and describe associations

The back-to-back stem plot below displays the distribution of life expectancy (in years) for 13 countries in 2010 and 1970.

Key: 6|7= 67 years

1970									2010								
							8	3									
								4									
						9	4	2	5	8							
9	9	9	8	7	7	1	0	6	9								
							4	7	1	2	4	4	6	8	9	9	
								8	0	0	0						

Do the back-to-back stem plots support the contention that life expectancy is increasing over time? Write a brief explanation based on your comparisons of the two medians.

2006

The back-to-back **ordered** stemplot below shows the distribution of maximum temperatures (in !Celsius) of two towns, Beachside and Flattown, over 21 days in January.

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

Question 1

The variables

temperature (!Celsius)

and

town (Beachside or Flattown)

are

A. both categorical variables.

B. both numerical variables.

C. categorical and numerical variables respectively.

D. numerical and categorical variables respectively.

E. neither categorical nor numerical variables.

2009

The following information relates to Questions 1 to 3.

The back-to-back ordered stem plot below shows the female and male smoking rates, expressed as a percentage, in 18 countries.

smoking rates (%)

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

Question 1

For these 18 countries, the lowest female smoking rate is

A. 5%

B. 7%

C. 9%

D. 15%

E. 19%

Question 2

For these 18 countries, the interquartile range (IQR) of the female smoking rates is

A. 4

B. 6

C. 19

D. 22

E. 23

Question 3

For these 18 countries, the smoking rates for females are generally

A. lower and less variable than the smoking rates for males.

B. lower and more variable than the smoking rates for males.

C. higher and less variable than the smoking rates for males.

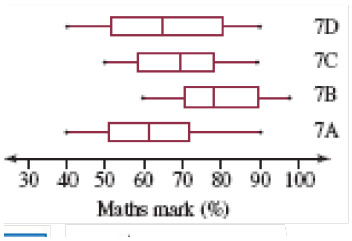
D. higher and more variable than the smoking rates for males.

E. about the same as the smoking rates for males.

Parallel Box Plots

Parallel boxplots also displays a **relationship between a numerical variable and a categorical variable**. However, the categorical variable can have **2 or more** categories.

A Parallel boxplot is obtained by constructing individual boxplots for each distribution using a common scale.

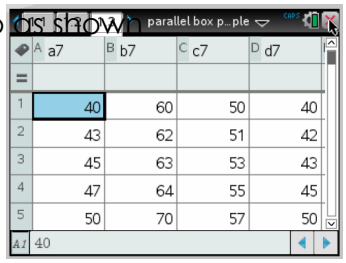


Eg. Four Year 7 classes complete the same end of topic maths test. The marks expressed as percentages are given below. Display the data using a parallel boxplot and use this to describe any similarities or differences in the distributions of the marks between the four classes.

7A	40	43	45	47	50	52	53	54	57	60	69	63	63	68	70	75	80	85	89	90
7B	60	62	63	64	70	73	74	76	77	77	78	82	85	87	89	90	92	95	97	97
7C	50	51	53	55	57	60	63	65	67	69	70	72	73	74	76	80	82	82	85	89
7D	40	42	43	45	50	53	55	59	60	61	69	73	74	75	80	81	82	83	84	90

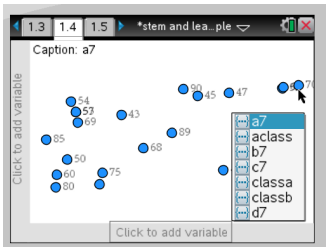
Using CAS

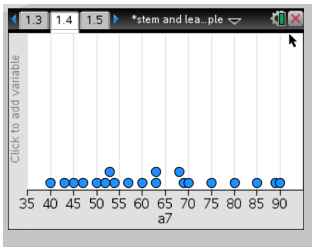
1. On a lists and spreadsheets page enter the data into column A and B, naming each group as shown



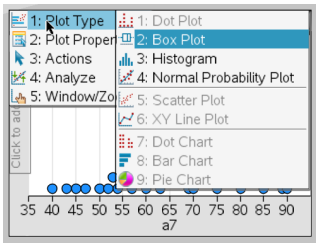
2. Go to Home and open graphs and statistics

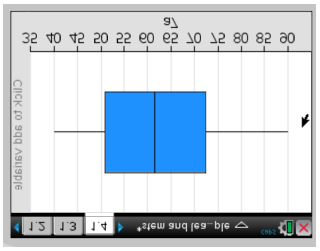
3. Go to the horizontal axis – press on click to add variable and then go to a7





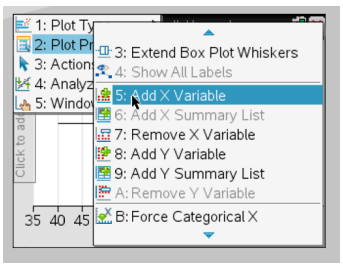
3. Go to Menu – 1. Plot Type – 2. Box Plot

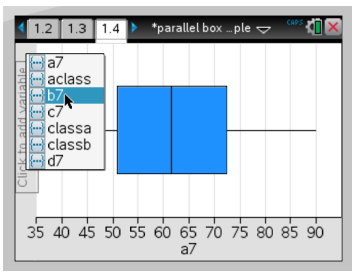




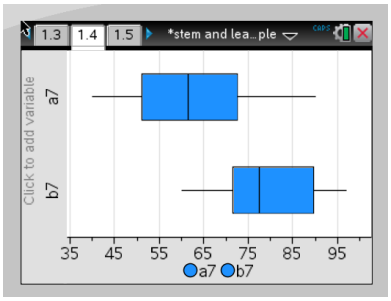
For Parallel Box Plots

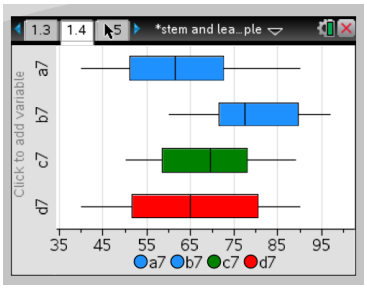
Go to Menu – 2. Plot Properties – 5. Add X Variable – Choose b7





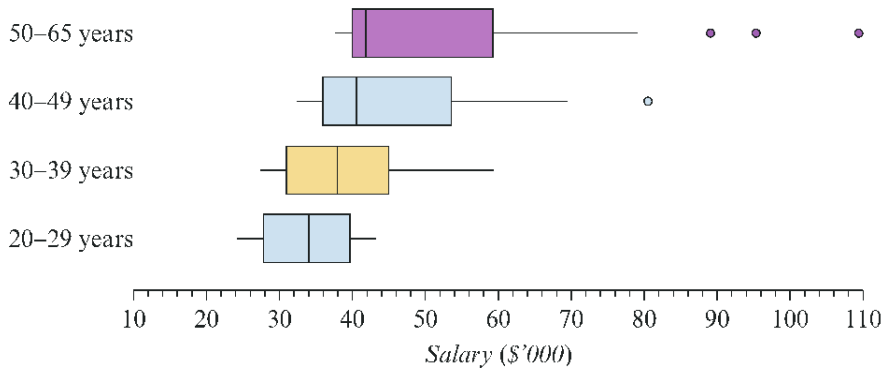
Repeat this process, choosing all 4 graphs





Analysing parallel box plots

The parallel box plots below compare the salary distribution for four different age groups: 20–29 years, 30–39 years, 40–49 years and 50–65 years. Here the numerical variable *salary* is the *response* variable and the categorical variable *age group* is the *explanatory* variable.



If there is *no association* between salary and age group, the salary distribution will be similar for all age groups. Equivalently, when working with parallel box plots, each *box plot* will be the *same*. This is not the case here.

There are several ways of identifying and systematically describing the association between salary and age group from this display. This can be done by:

Comparing Medians

Report

The parallel box plots show that median salaries and age group are associated because median salaries increase with age group. For example, the median salary increased from $34 000 for 20–29 year-olds to $42 000 for 50–65 year-olds.

Comparing IQRs and/or ranges

Report

From the parallel box plots we can see that the spread of salaries is associated with age group. For example, the IQR increased from around $12 000 for 20–29-year-olds to around $20 000 for 50–65-year-olds.

Comparing Shapes

Report

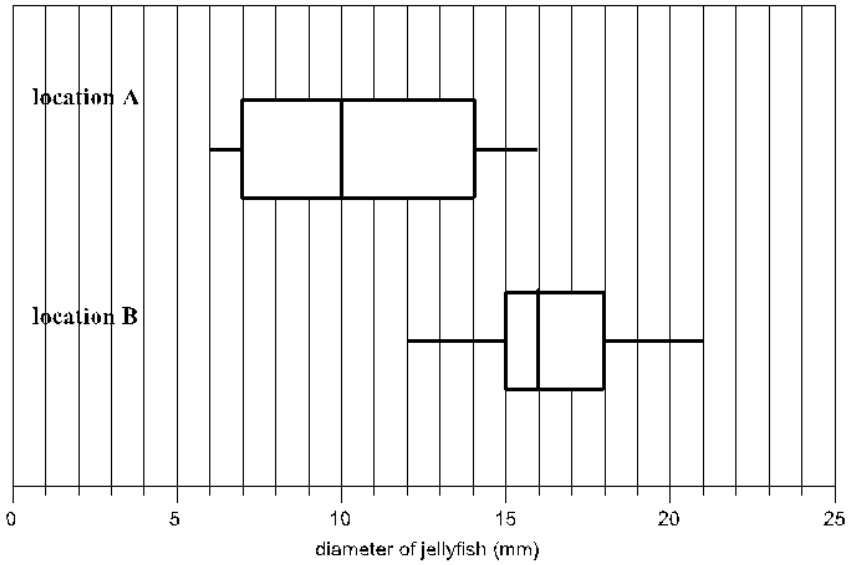
From the parallel box plots we can see that the shape of the distribution of salaries is associated with age group because of the distribution, which is symmetric for 20–29-year-olds, and becomes progressively more positively skewed as age increases. Outliers also begin to appear.

Exam Questions

2007

The following information relates to Questions 5 and 6.

Samples of jellyfish were selected from two different locations, A and B. The diameter (in mm) of each jellyfish was recorded and the resulting data is summarised in the boxplots shown below.



Question 5

The percentage of jellyfish taken from location A with a diameter greater than 14 mm is closest to

A. 2%
B. 5%
C. 25%
D. 50%
E. 75%

Question 6

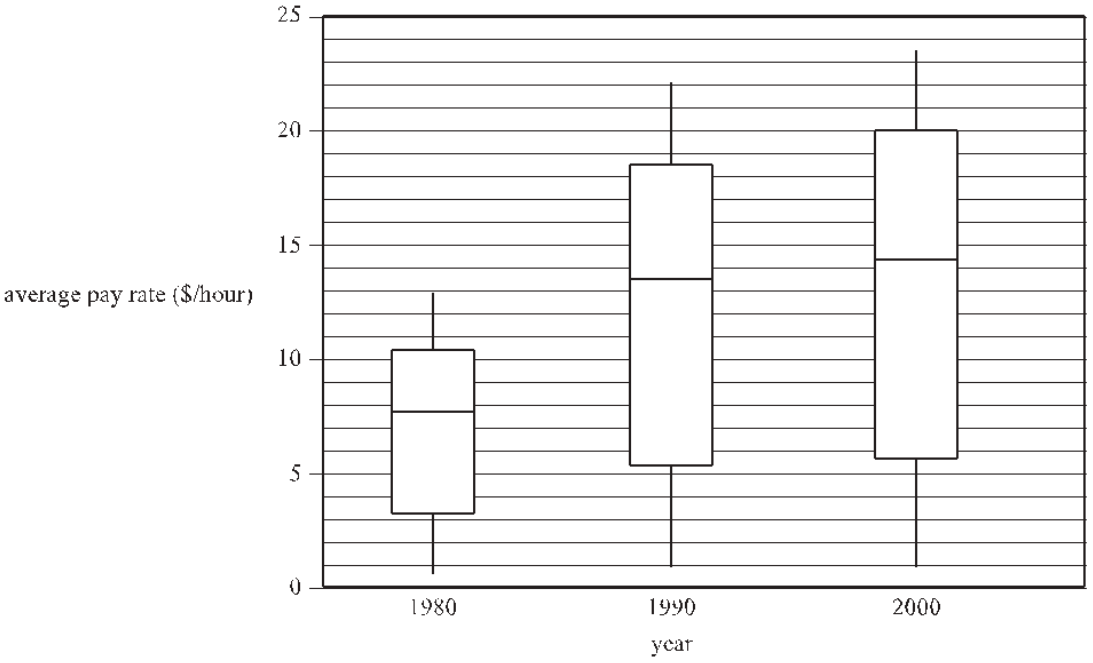
From the boxplots, it can be concluded that the diameters of the jellyfish taken from location A are generally

A. similar to the diameters of the jellyfish taken from location B.
B. less than the diameters of the jellyfish taken from location B and less variable.
C. less than the diameters of the jellyfish taken from location B and more variable.
D. greater than the diameters of the jellyfish taken from location B and less variable.
E. greater than the diameters of the jellyfish taken from location B and more variable.

2011

Question 5

The boxplots below display the distribution of average pay rates, in dollars per hour, earned by workers in 35 countries for the years 1980, 1990 and 2000.



Based on data obtained from: www.data.un.org

Based on the information contained in the boxplots, which one of the following statements is **not** true?

A. In 1980, over 50% of the countries had an average pay rate less than $8.00 per hour.

B. In 1990, over 75% of the countries had an average pay rate greater than $5.00 per hour.

C. In 1990, the average pay rate in the top 50% of the countries was higher than the average pay rate for any of the countries in 1980.

D. In 1990, over 50% of the countries had an average pay rate less than the median average pay rate in 2000.

E. In 2000, over 75% of the countries had an average pay rate greater than the median average pay rate in 1980.

Edrolo link: https://edrolo.com.au/s/2678923/

Ex 2C