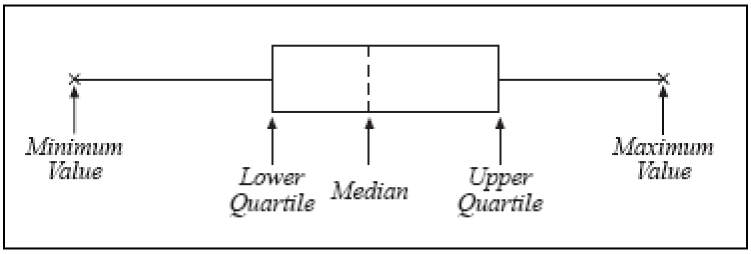
Section 3.1.9 – Parallel Box Plots

VCAA "Dot Points"

Investigating associations between two variables, including:

- *back-to-back stem plots, parallel dot plots and boxplots and their use in identifying and describing associations between a numerical and a categorical variable*

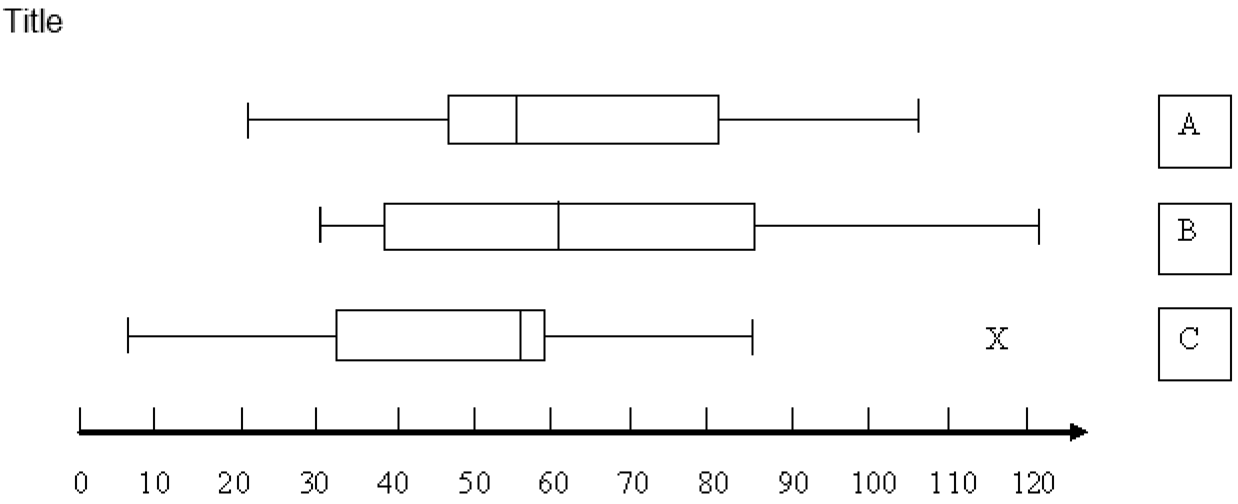
Box Plot



Parallel box plots (also known as box and whisker plots) are used to **compare** the numerical results taken from **two or more groups**. They are called parallel because they are placed one above the other using only one number line or common scale

A box plot is constructed using the **5 number summary** (X_{min}, Q_1, Median, Q_3 & X_{max}), as shown in the diagram above.

These are often useful in comparing features of distributions.



NB: When constructing parallel box plots be sure to:

- Draw a suitable axis and scale
- Label each box plot
- Provide a title

Comparing Box Plots

When comparing parallel box plots be sure to examine the following features:

- Central tendencies - compare the median from each box plot and make comment upon the increasing or decreasing order.
  NB: Be sure to support your observation with actual median values.

- Variation/Spread - compare the IQR and/or range and comment upon increasing or decreasing spread.
  - If the data cover a wide range, then the spread is large.
  - If the data is clustered around a single value, the spread is smaller.

  NB: Be sure to support your observations with actual IQR and/or range values.

- Shape - classify the shape of the distribution as either symmetrical, positively skewed or negatively skewed.

- Quartiles - look for comparisons between parallel box plots relative to quartiles.
  Use statements such as "the top 50% of scores for class A were better than the best class B score"

- Unusual features - such as gaps in data and/or outliers

Calculating outliers

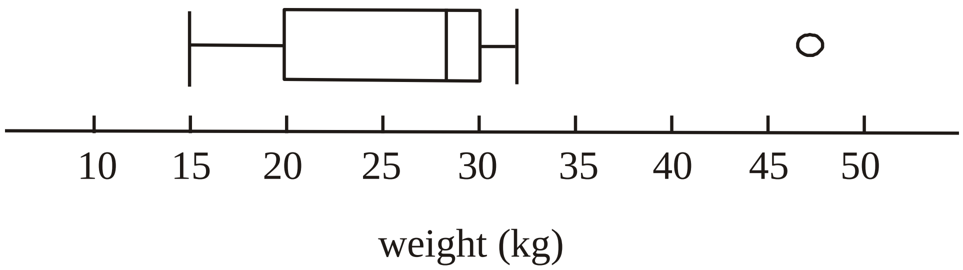
An outlier is an observation that lies an abnormal distance from other values in a random sample from a population. It can be considered as an outlier for being too large or too small in value.

To test if a value is an outlier, the value must be compared against an **upper** and **lower** boundary or **fence**.

Lower fence
$= Q_1 - 1.5 \times IQR$

Upper fence
$= Q_3 + 1.5 \times IQR$

The following information relates to examples 1 and 2.



Example.1

The data represented by the box plot shown above is best described as

A. negatively skewed
B. bell shaped
C. symmetric
D. increasing
E. positively skewed.

A

The data represented by the box plot is negatively skewed since the data values are grouped to the right hand side of the distribution and gradually tail off to the left.

∴ Option A

Example.2

The data value 47, has been correctly identified as a possible outlier because it is

A. less than 50
B. greater than 32
C. greater than 35
D. greater than 40
E. greater than 45

E

A piece of data is classified as a possible outlier if it is:

$> Q_3 + 1.5 \times IQR$, or

$< Q_1 - 1.5 \times IQR$

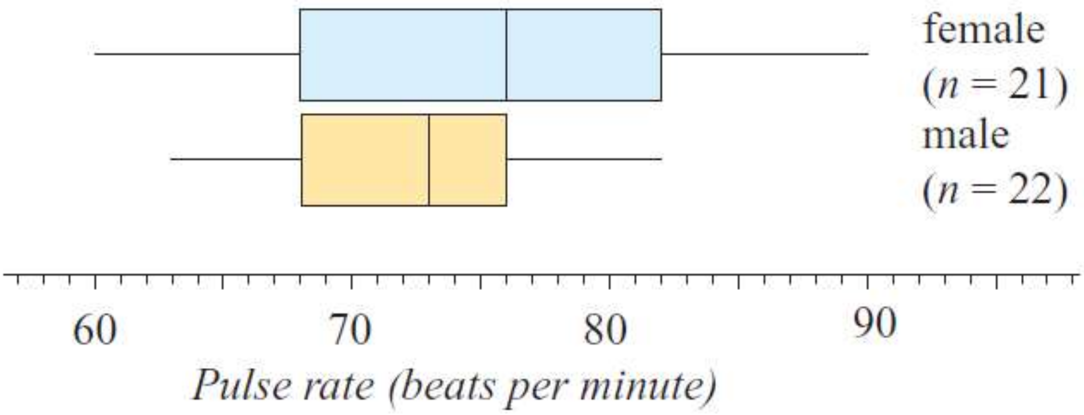
IQR = $Q_3 - Q_1$
= 30 – 20
= 10

Lower fence
= Q_1 – 1.5xIQR
= 20 – (1.5 x 10)
= 20 – 15
= 5

Upper fence
= Q_3 + 1.5xIQR
= 30 + (1.5 x 10)
= 30 + 15
= 45

Since 47 > 45 it is a possible outlier.
∴ Option E.

The below parallel box plots show the distribution of pulse rate of 21 adult females and 22 adult males.



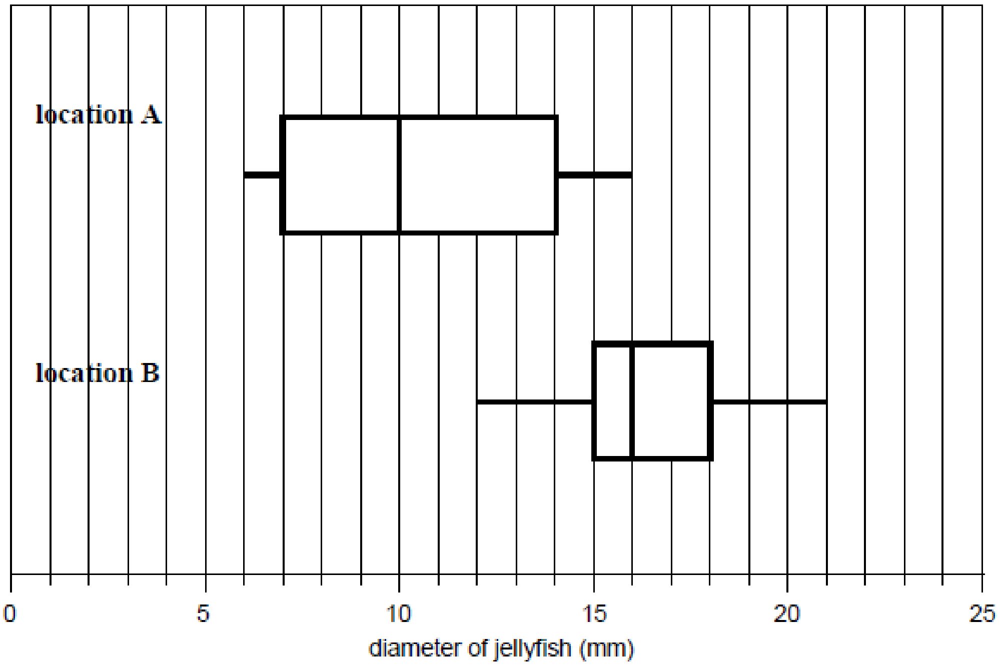
Example.3

Compare the two distributions in terms of central tendencies, spread and shape. Give appropriate supporting values at a level of accuracy that can be read from the plot.

- The female median pulse rate of 76 bpm was higher than the male median pulse rate of 73 bpm.
- The female range (90 - 60 = 30) and IQR (82 – 68 = 14) were both greater than the male range (82 – 63 = 19) and IQR (76 – 68 = 8). Therefore the females had a much larger spread.
- Upon inspection it would appear that both the female and male data had a slightly negative skewness.
- The top 25% of females had a higher pulse rate than any males recorded.

The following information relates to Examples 4 and 5.

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.



Example.4

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%

Q_3 for boxplot A = 14
∴ 25% of the data remains in the final quartile.

∴ Option C

C

Example.5

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

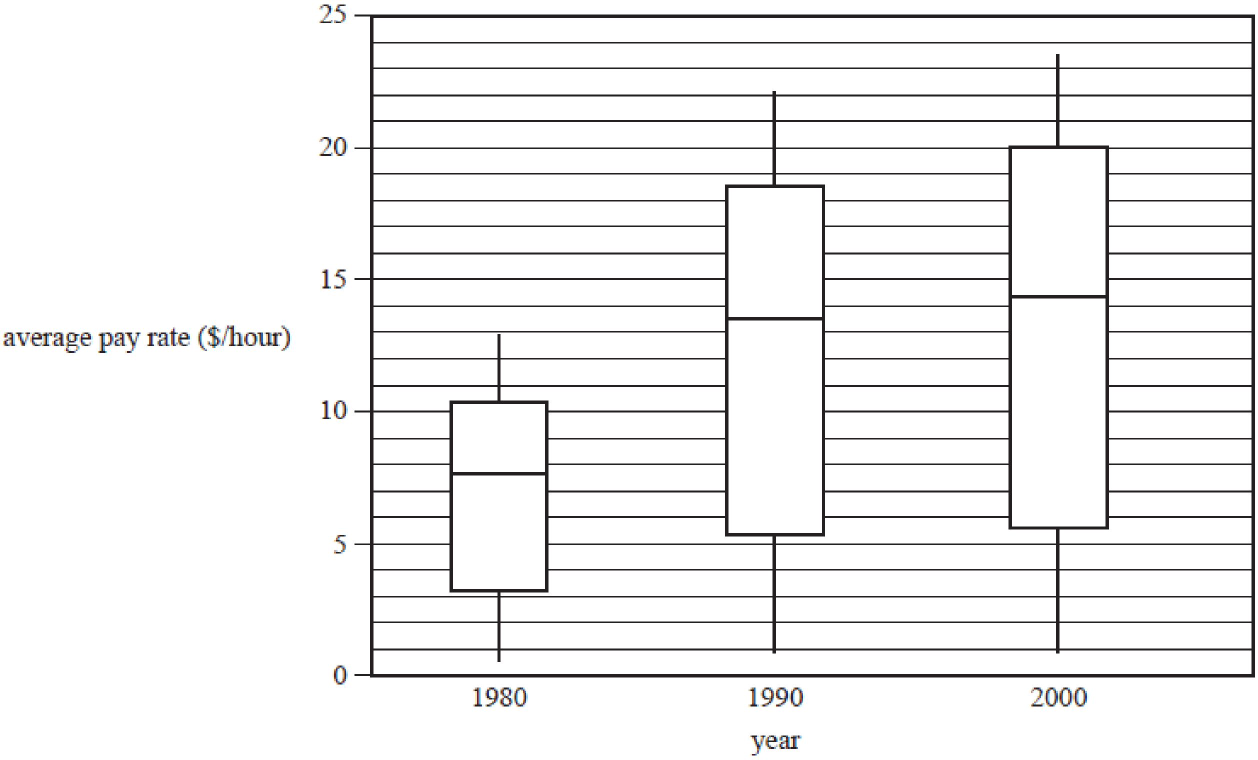
A. similar to the diameters of the jellyfish taken from location B.
B. less than the diameters of the jellyfish taken from location B and less variable.
C. less than the diameters of the jellyfish taken from location B and more variable.
D. greater than the diameters of the jellyfish taken from location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

C

Jellyfish from location A have a lower median and a larger range.
∴ Option C

Example.6

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 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 countries in 1980

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

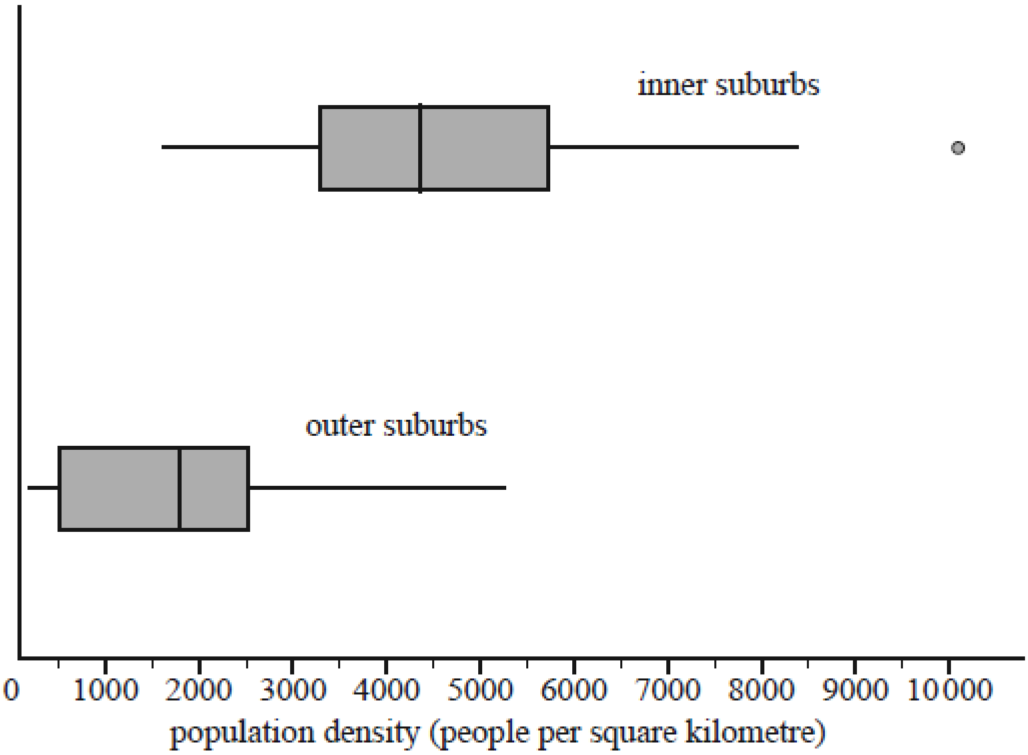
E. In 2000, over 75% of the countries had an average pay rate greater than the median average pay rate in 1980

E

Statement E is not true!
∴ Option E

Example 7

The parallel boxplots below summarise the distribution of population density, in people per square kilometre, for 27 inner suburbs and 23 outer suburbs of a large city.



Which one of the following statements is **not** true?

A. More than 50% of the outer suburbs have population densities below 2000 people per square kilometre.

B. More than 75% of the inner suburbs have population densities below 6000 people per square kilometre.

C. Population densities are more variable in the outer suburbs than in the inner suburbs.

D. The median population density of the inner suburbs is approximately 4400 people per square kilometre.

E. Population densities are, on average, higher in the inner suburbs than in the outer suburbs.

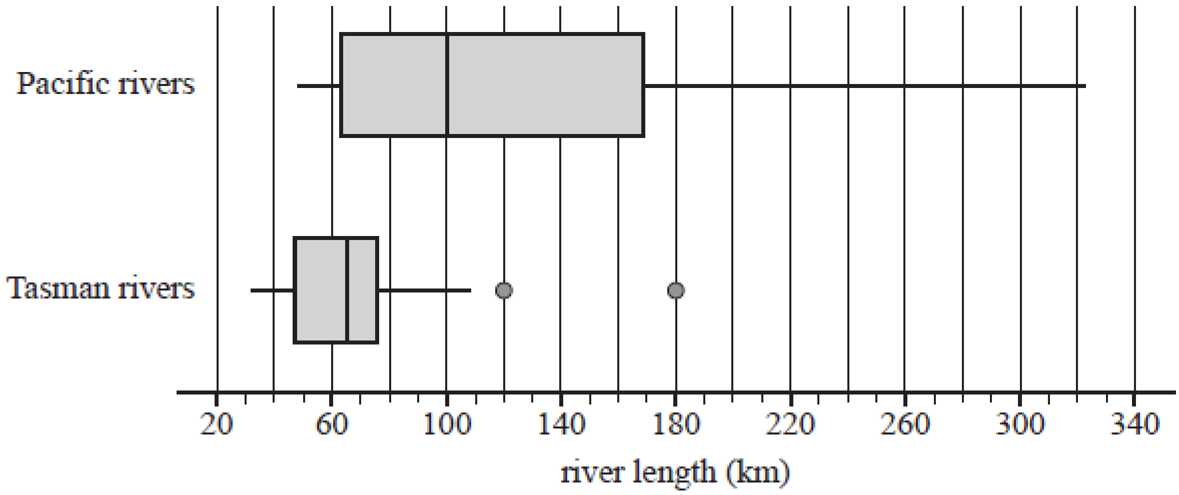
$\boxed{\text{C}}$

Statement C is not true! There is a much greater variation in the inner suburb, in both Range and IQR, than that of the outer suburbs.
∴ Option C

Use the following information to answer Questions 8 and 9.

In New Zealand, rivers flow into either the Pacific Ocean (the Pacific rivers) or the Tasman Sea (the Tasman rivers).

The boxplots below can be used to compare the distribution of the lengths of the Pacific rivers and the Tasman rivers.



Example 8

The five-number summary for the lengths of the Tasman rivers is closest to

A. 32, 48, 64, 76, 108
B. 32, 48, 64, 76, 180
C. 32, 48, 64, 76, 322
D. 48, 64, 97, 169, 180
E. 48, 64, 97, 169, 322

B

Whilst points 120 and 180 are recorded as "outliers", they still are part of the data that makes the 5 number summary.
∴ Option B

Example 9

Which one of the following statements is **not** true?

A. The lengths of two of the Tasman rivers are outliers.
B. The median length of the Pacific rivers is greater than the length of more than 75% of the Tasman rivers.
C. The Pacific rivers are more variable in length than the Tasman rivers.
D. More than half of the Pacific rivers are less than 100 km in length.
E. More than half of the Tasman rivers are greater than 60 km in length.

D

The median of the Pacific rivers is exactly 100 km long.
Clearly, statement D is incorrect.
∴ Option D