1 A random sample of 15 days is taken from the large data set for Perth in June and July 1987.

The scatter diagram in Figure 1 displays the values of two of the variables for these 15 days.

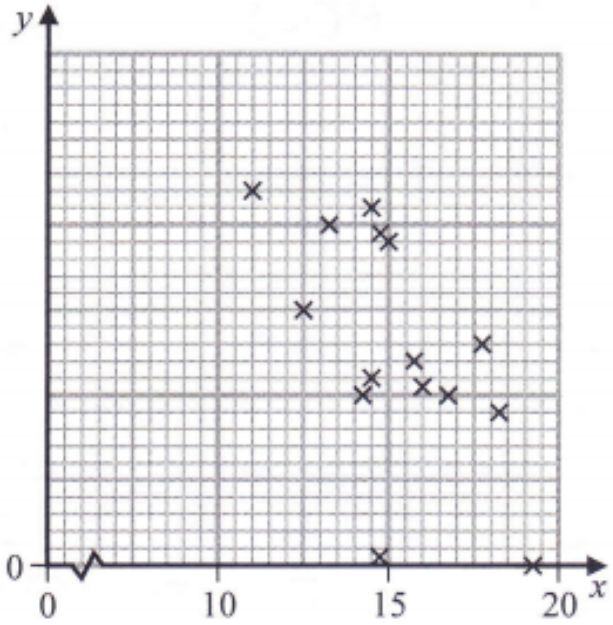


Figure 1

(a) Describe the correlation.

(1)

The variable on the x-axis is Daily Mean Temperature measured in °C.

(b) Using your knowledge of the large data set,

(i) suggest which variable is on the y-axis,

(ii) state the units that are used in the large data set for this variable.

(2)

Stav believes that there is a correlation between Daily Total Sunshine and Daily Maximum Relative Humidity at Heathrow.

He calculates the product moment correlation coefficient between these two variables for a random sample of 30 days and obtains $r = -0.377$

(c) Carry out a suitable test to investigate Stav's belief at a 5% level of significance. State clearly

- your hypotheses
- your critical value

(3)

On a random day at Heathrow the Daily Maximum Relative Humidity was 97%

(d) Comment on the number of hours of sunshine you would expect on that day, giving a reason for your answer.

(1)

a) Negative correlation

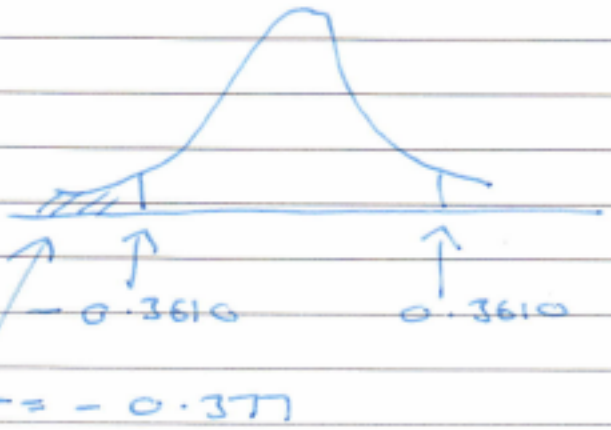
b) (i) Rainfall
(ii) mm

c) $r = -0.377$

PMCC tables $n = 30$ 2.5% (2 tail)
PMCC $= 0.3610$

$H_0: \rho = 0$ $H_1: \rho \neq 0$
two tail test

Critical value
$= -0.3610$ in lower tail



As $r = -0.377$ is in the lower tail reject H_0, accept H_1

There is a correlation between the Daily Total Sunshine and Daily Maximum Relative Humidity at Heathrow.

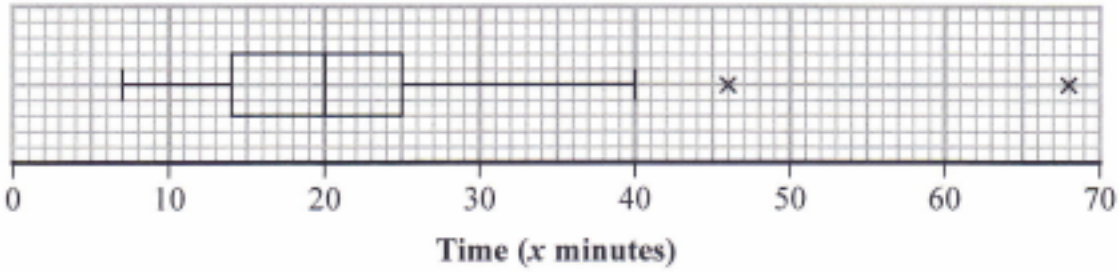
d) As humidity is high (97%) and there is negative correlation ($r < 0$), we would expect the number of hours of sunshine at Heathrow to be lower than the average.

2.

Each member of a group of 27 people was timed when completing a puzzle.

The time taken, x minutes, for each member of the group was recorded.

These times are summarised in the following box and whisker plot.



Time (x minutes)

(a) Find the range of the times.

(1)

(b) Find the interquartile range of the times.

(1)

For these 27 people $\sum x = 607.5$ and $\sum x^2 = 17623.25$

(c) calculate the mean time taken to complete the puzzle,

(1)

(d) calculate the standard deviation of the times taken to complete the puzzle.

(2)

Taruni defines an outlier as a value more than 3 standard deviations above the mean.

(e) State how many outliers Taruni would say there are in these data, giving a reason for your answer.

(1)

Adam and Beth also completed the puzzle in a minutes and b minutes respectively, where $a > b$.
When their times are included with the data of the other 27 people

- the median time increases
- the mean time does not change

(f) Suggest a possible value for a and a possible value for b, explaining how your values satisfy the above conditions.

(3)

(g) Without carrying out any further calculations, explain why the standard deviation of all 29 times will be lower than your answer to part (d).

(1)

a) range = 68 − 7 = 61

b) IQR = UQ − LQ = 25 − 14 = 11

c) $\text{mean} = \frac{\Sigma x}{n} = \frac{607.5}{27} = 22.5$

d) $\sigma = \sqrt{\frac{\Sigma x^2}{n} - \bar{x}^2} = \sqrt{\frac{17623.25}{27} - 22.5^2}$

$\sigma = 12.1$

e) $3 \times 12.1 = 36.3 = 3$ sd

upper $22.5 + 36.3 = 58.8$
lower $22.5 - 36.3 = -13.8$

Only one outlier (68) above 58.8 minutes

f) 7 . . . . 20 22 23 68

added

↑ 14th median

↑ 15th (new median)

current total time = $22.5 \times 27 = 607.5$
add 2 people
total time = $22.5 \times 29 = 652.5$
(we've added 45)
say, 22 + 23
New median will be 15th (22)

- Median increased from 20 to 22
- Mean the same (22.5)

g) Both 22 and 23 are less than 1 s.d. (12.1 units) from the mean, so s.d. of all 29 values will be smaller.

5.

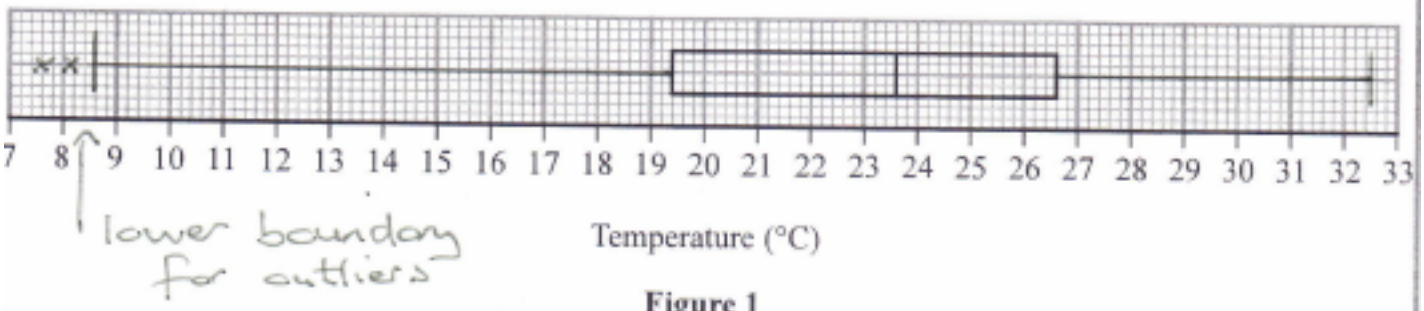


Figure 1

The partially completed box plot in Figure 1 shows the distribution of daily mean air temperatures using the data from the large data set for Beijing in 2015

An outlier is defined as a value
more than 1.5 × IQR below Q_1 or
more than 1.5 × IQR above Q_3

2 Outliers

The three lowest air temperatures in the data set are 7.6 °C, 8.1 °C and 9.1 °C
The highest air temperature in the data set is 32.5 °C upper value

(a) Complete the box plot in Figure 1 showing clearly any outliers.

(4)

(b) Using your knowledge of the large data set, suggest from which month the two outliers are likely to have come.

(1)

Using the data from the large data set, Simon produced the following summary statistics for the daily mean air temperature, x °C, for Beijing in 2015

$$n = 184 \qquad \sum x = 4153.6 \qquad S_{xx} = 4952.906$$

(c) Show that, to 3 significant figures, the standard deviation is 5.19 °C

(1)

Simon decides to model the air temperatures with the random variable

$$T \sim N(22.6, 5.19^2)$$

(d) Using Simon's model, calculate the 10th to 90th interpercentile range.

(3)

Simon wants to model another variable from the large data set for Beijing using a normal distribution.

(e) State two variables from the large data set for Beijing that are **not** suitable to be modelled by a normal distribution. Give a reason for each answer.

(2)

b) October – northern hemisphere coldest of months as data from May to October.

a) From box plot

$IQR = 26.6 - 19.4 = 7.2$

$1.5 \times 7.2 = 10.8$

$LQ - 10.8 = 19.4 - 10.8 = 8.6$

$UQ + 10.8 = 26.6 + 10.8 = 37.4$

Outliers would be below 8.6 or above 37.4

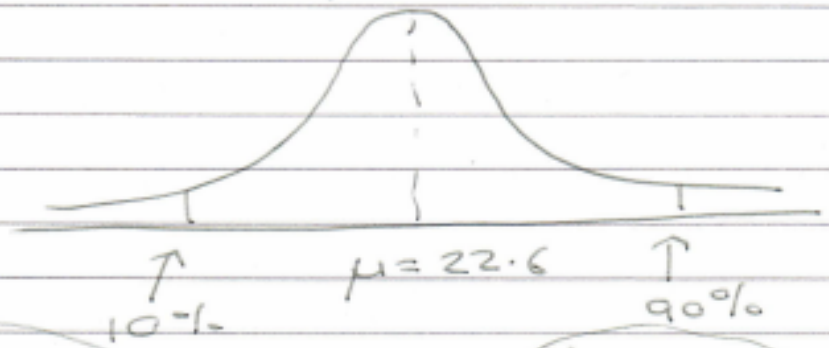
c) $\sigma = \sqrt{\frac{S_{xx}}{n}} = \sqrt{\frac{4952.906}{184}}$

$= 5.18825$

$= 5.19°C$ (3 sf)

d) $T \sim N(22.6, 5.19^2)$

μ $\quad$ σ



Inverse normal
Area = 0.1
$\sigma = 5.19$
$\mu = 22.6$

$= 15.9487$

Inverse normal
Area = 0.9
$\sigma = 5.19$
$\mu = 22.6$

$= 29.251253$

$29.251253 - 15.9487 = 13.3025$

Turn over for a spare grid if you need to redraw your box plot.

10^{th} to 90^{th} interpercentile range = 13.3°C

6-

Charlie is studying the time it takes members of his company to travel to the office. He stands by the door to the office from 0840 to 0850 one morning and asks workers, as they arrive, how long their journey was.

(a) State the sampling method Charlie used.

(1)

(b) State and briefly describe an alternative method of non-random sampling Charlie could have used to obtain a sample of 40 workers.

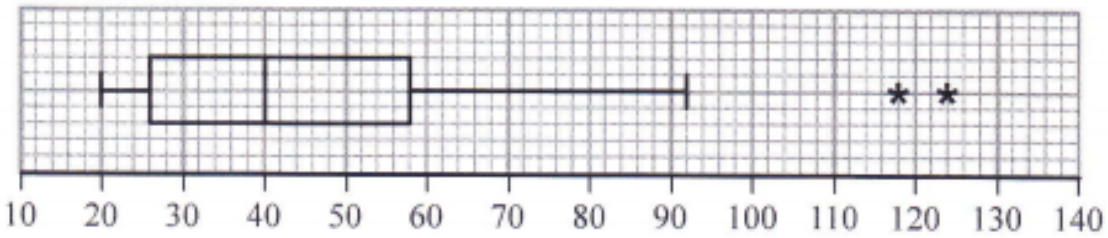
(2)

Taruni decided to ask every member of the company the time, x minutes, it takes them to travel to the office.

(c) State the data selection process Taruni used.

(1)

Taruni's results are summarised by the box plot and summary statistics below.



Journey time (minutes)

$n = 95 \quad \sum x = 4133 \quad \sum x^2 = 202\,294$

(d) Write down the interquartile range for these data.

(1)

(e) Calculate the mean and the standard deviation for these data.

(3)

(f) State, giving a reason, whether you would recommend using the mean and standard deviation or the median and interquartile range to describe these data.

(2)

Rana and David both work for the company and have both moved house since Taruni collected her data.

Rana's journey to work has changed from 75 minutes to 35 minutes and David's journey to work has changed from 60 minutes to 33 minutes.

Taruni drew her box plot again and only had to change two values.

(g) Explain which two values Taruni must have changed and whether each of these values has increased or decreased.

(3)

a) Convenience (or opportunity) sampling

b) Use quota sampling.

Instead of just picking the first 40, instead take 4 people every 10 minutes until 40 picked.

c) This is a Census

d) IQR = 58 − 26
= 32 minutes

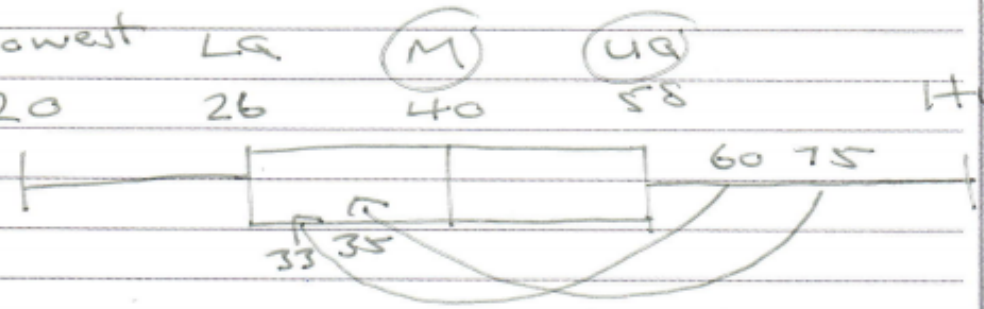
e) Mean = $\mu = \frac{\Sigma x}{n} = \frac{4133}{95}$
$= 43.505526$

Standard deviation = $\sigma = \sqrt{\frac{\Sigma x^2}{n} - \mu^2} = \sqrt{\frac{202294}{95} - (43.5\ldots)^2}$

$\sigma = 15.385514$

f) There are 2 outliers for the data, so these will skew the mean and range.
Better to use Median and IQR.

g)



The lowest and highest values will be unchanged, as will be the LQ (see diagram above)
From diagram, Median and upper quartile will change.