Section 3.1.4 – Describing Distributions

VCAA "Dot Points"

Investigating data distributions, including:

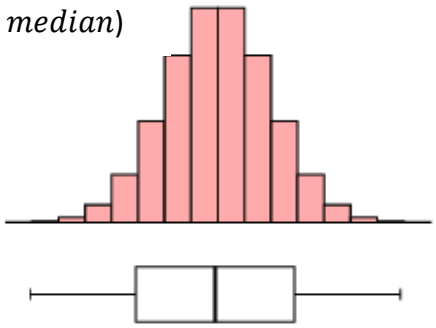
- *review of representation, display and description of the distributions of numerical variables: dot plots, stem plots, histograms*

Describing Distributions

In further maths students are often required to describe the distribution of a sample. The main three classifications are as follows:

Symmetric distribution

$(\bar{x} \approx median)$



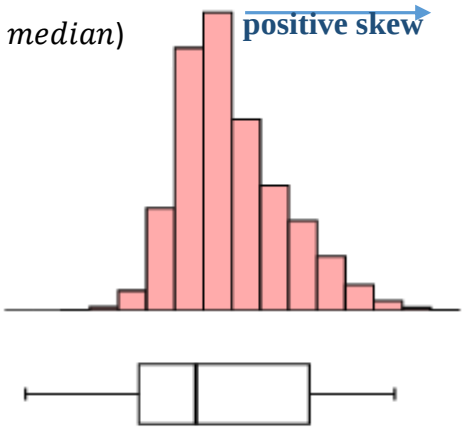
A symmetric distribution is one where the left and right hand sides of the distribution are roughly equally balanced around the mean.

For symmetric distributions, the mean is approximately equal to the median.

For a symmetric distribution, the left and right tails are equally balanced, meaning that they have about the same length.

Positively skewed

$(\bar{x} > median)$



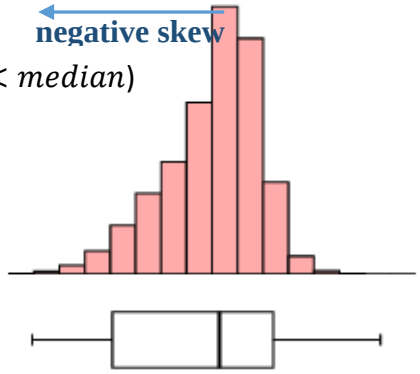
For a positively skewed distribution, the **mean is typically greater than the median.**

Also notice that the tail of the distribution on the right hand (positive) side is longer than on the left hand side.

From the box and whisker diagram we can also see that the median is closer to the first quartile than the third quartile. The fact that the right hand side tail of the distribution is longer than the left can also be seen.

Negatively skewed

$(\bar{x} < median)$

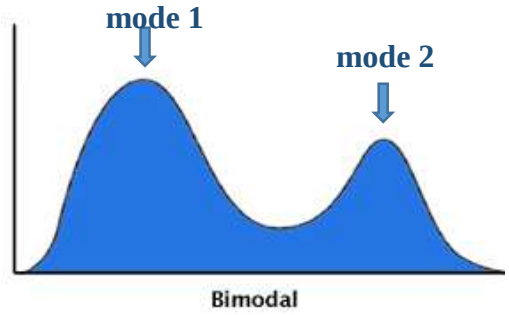


For a negatively skewed distribution, the **mean is typically less than the median**. Also notice that the tail of the distribution on the left hand (negative) side is longer than on the right hand side.

From the box and whisker diagram we can also see that the median is closer to the third quartile than the first quartile. The fact that the left hand side tail of the distribution is longer than the right can also be seen.

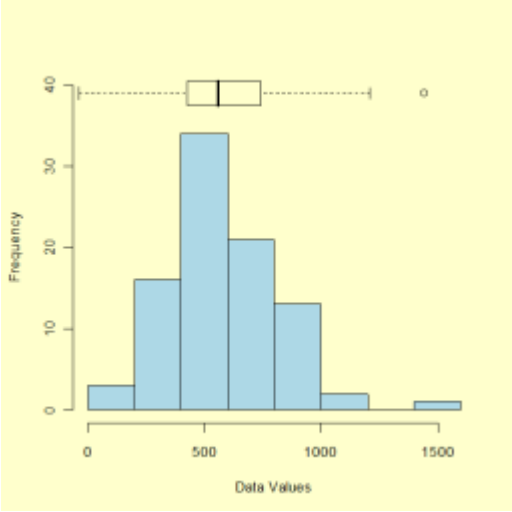
Other possible classifications include:

Bimodal distribution



A bimodal distribution, as the name suggests has two clear modes within the distribution.

Distributions with outliers



Distributions with outliers display gaps between the main body and data values in the tails.

Box plots display outliers with a dot.

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

$= Q_3 + 1.5 \times IQR$

Example.1

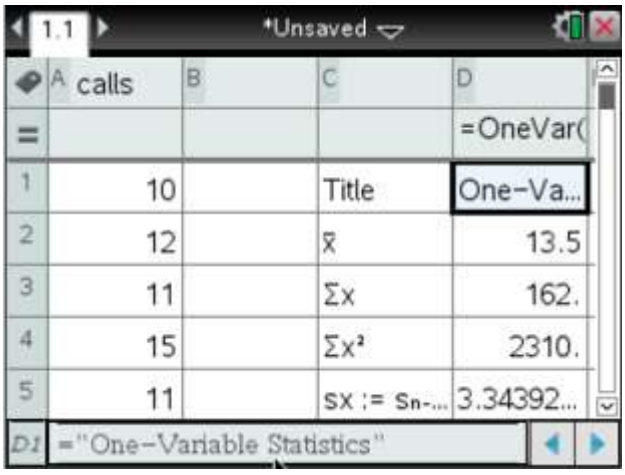
A sample of 12 students were surveyed and asked how many calls they received on their mobile phone on the previous day. The data is as follows:

10 12 11 15 11 14 13 17 12 22 14 11

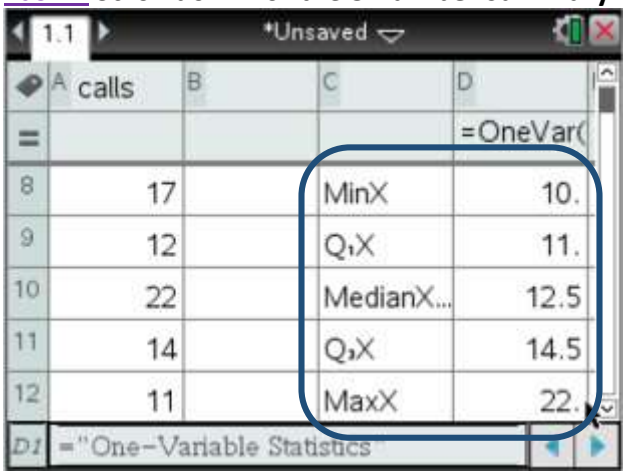
Use your Ti-nspire CAS calculator to:

1. Generate a univariate statistical analysis of the data
2. Create a 5 number summary
3. Construct a box plot
4. Describe the distribution

Task.1 Enter data into your TI-nspire CX CAS calculator
Home - add a Lists and Spreadsheet
Menu - option: 4:1:1 (Statistics: Stat Calculations: One-Variable Statistics)

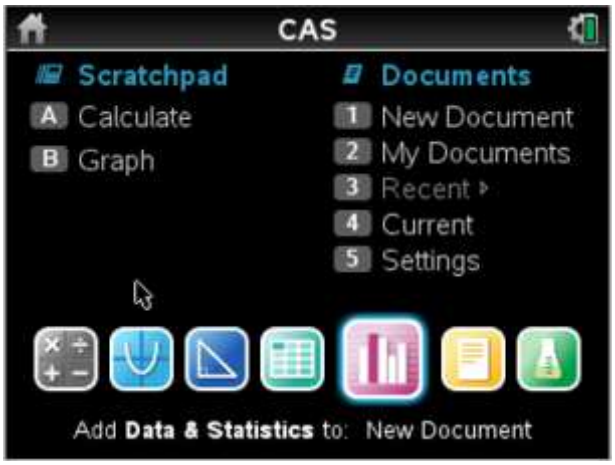


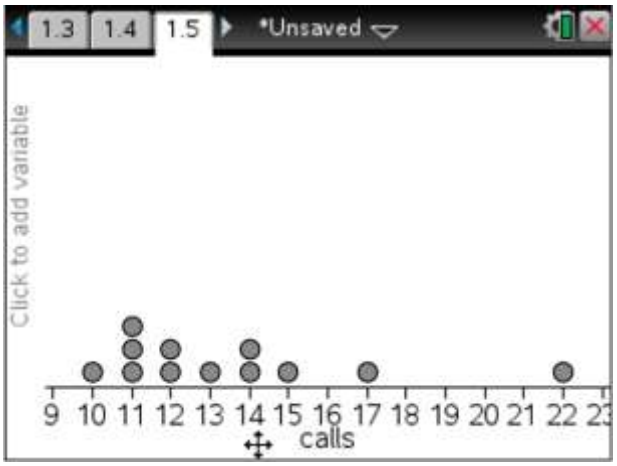
Task.2 Scroll down for the **5 number summary**

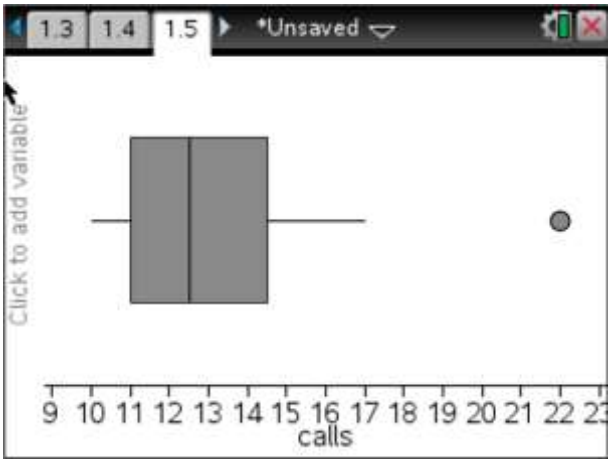


Task.3

Home - add a Data & Statistics Page
Construct a dot plot initially
Menu - option: 1:2 (Plot Type : Box Plot)







Task 4 Examine the box plot and compare $\bar{x}$ and the median (Q_2)

$\bar{x}$ = 13.5
Median = 12.5

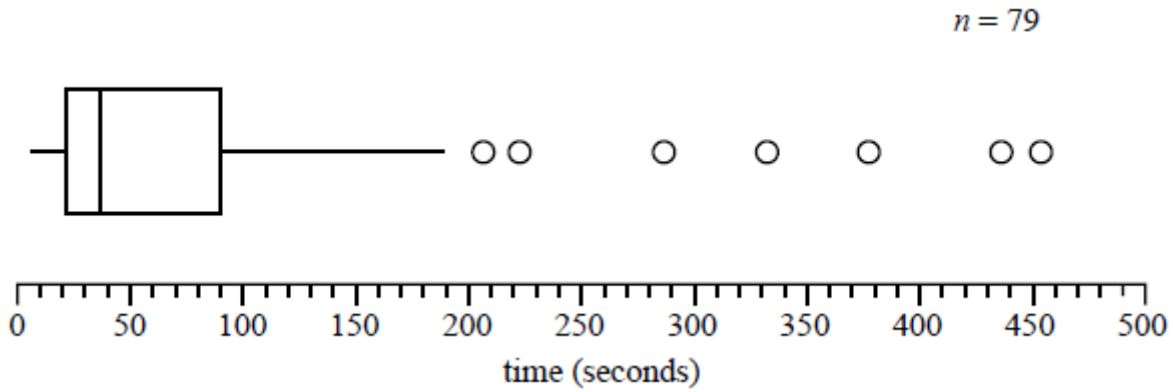
Upon observation and by showing that $\bar{x}$ > median, the distribution can be described as positively skewed with an outlier.

Checking the outlier via formula:
Lower fence = Q_1 – 1.5 x IQR = 11 – (1.5 x 3.5) = 5.75
Upper fence = Q_3 + 1.5 x IQR = 14.5 + (1.5 x 3.5) = 19.75

No values are less than 5.75
The value of 22 calls is greater than the upper fence of 19.75.
Therefore the value of 22 calls is considered an outlier.

Example.2

The box plot below shows the distribution of the time, in seconds, that 79 customers spent moving along a particular aisle in a large supermarket.



The shape of the distribution is best described as

A. symmetric.
B. negatively skewed.
C. negatively skewed with outliers.
D. positively skewed.
E. positively skewed with outliers.

This distribution is positively skewed as the variation in the top 50% of data is much greater than that of the bottom 50% of the data.

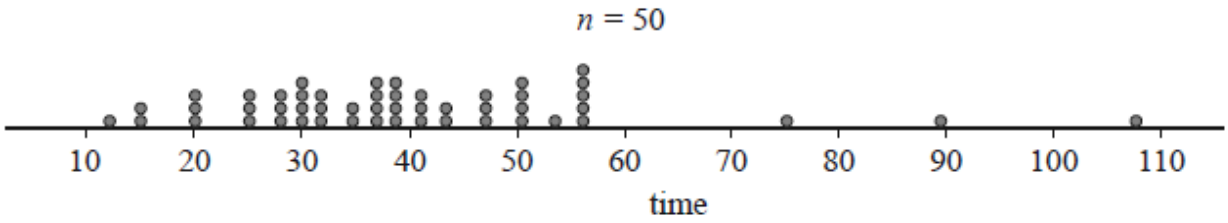
Clearly there are also several outliers.

E

$\therefore$ Option E

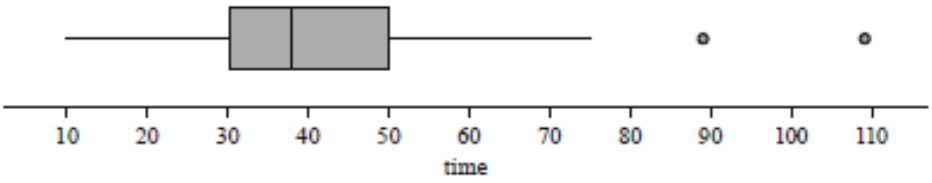
Example.3

The dot plot below shows the distribution of the time, in minutes, that 50 people spent waiting to get help from a call centre.

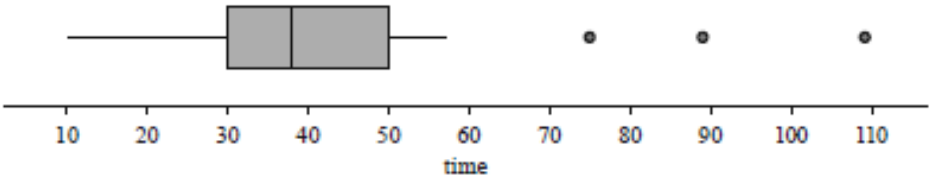


Which one of the following boxplots best represents the data?

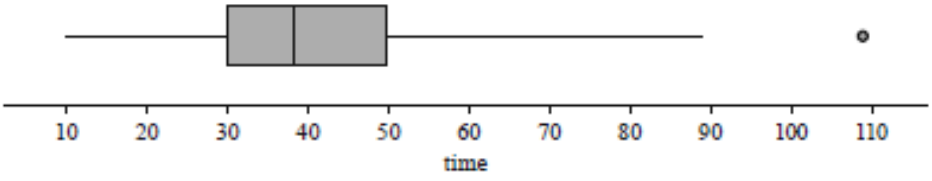
A.



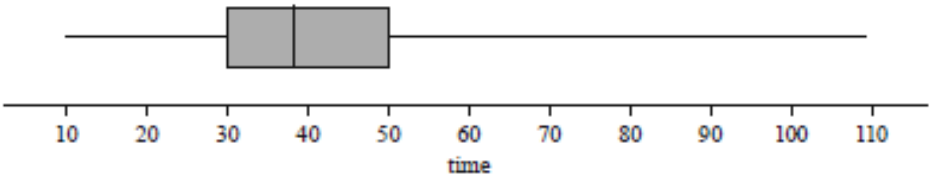
B.



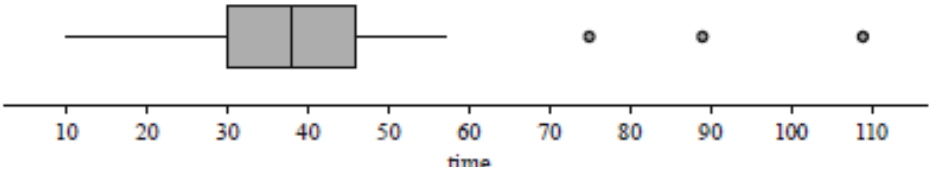
C.



D.



E.



A

Waiting time is an outlier if it is greater than ($Q_1 + 5.1 \times IQR = 50 + 5.1 \times 20 =$) 80 minutes.
$\therefore$ Option A

Example.4

The stem plot below displays the average number of decayed teeth in 12-year-old children from 31 countries.

key: 0|2 = 0.2

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

Based on this stem plot, the distribution of the average number of decayed teeth for these countries is best described as

A. negatively skewed with a median of 15 decayed teeth and a range of 45
B. positively skewed with a median of 15 decayed teeth and a range of 45
C. approximately symmetric with a median of 1.5 decayed teeth and a range of 4.5
D. negatively skewed with a median of 1.5 decayed teeth and a range of 4.5
E. positively skewed with a median of 1.5 decayed teeth and a range of 4.5

E

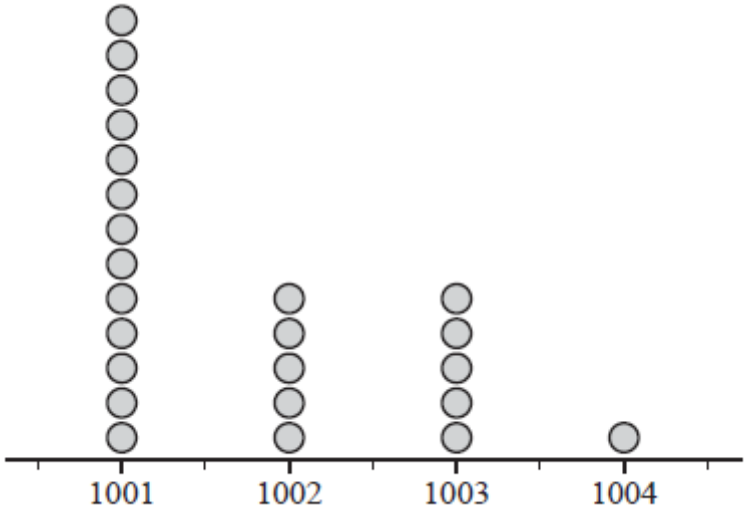
This distribution is clearly positively skewed, as there is a tail towards the larger numbers.

The median is 1.5 and the range 4.5

$\therefore$ Option E

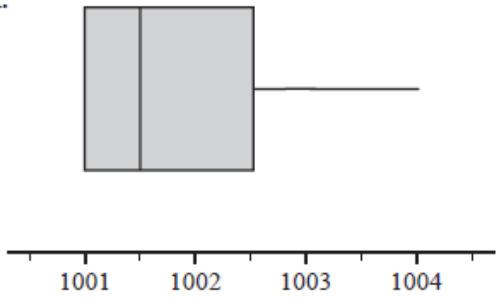
Example 5

A dot plot for a set of data is shown below.

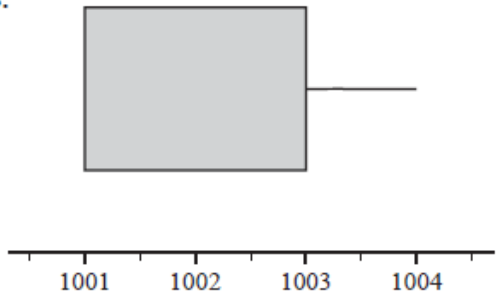


Which one of the following boxplots would best represent the dot plot above?

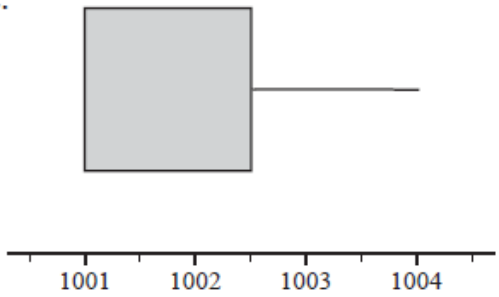
A.



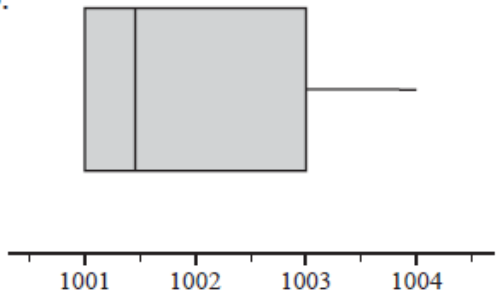
B.



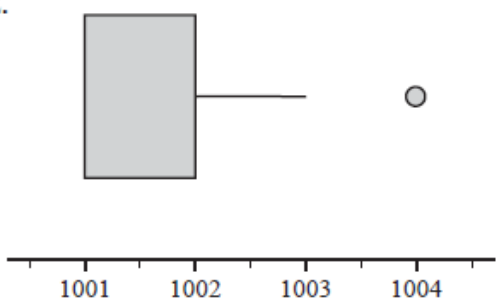
C.



D.



E.



From the dot plot it can be seen that the X_{min}, Q_1 and Medium are all the same value, namely 1001. Whilst Q_3 is equal to 1002.5

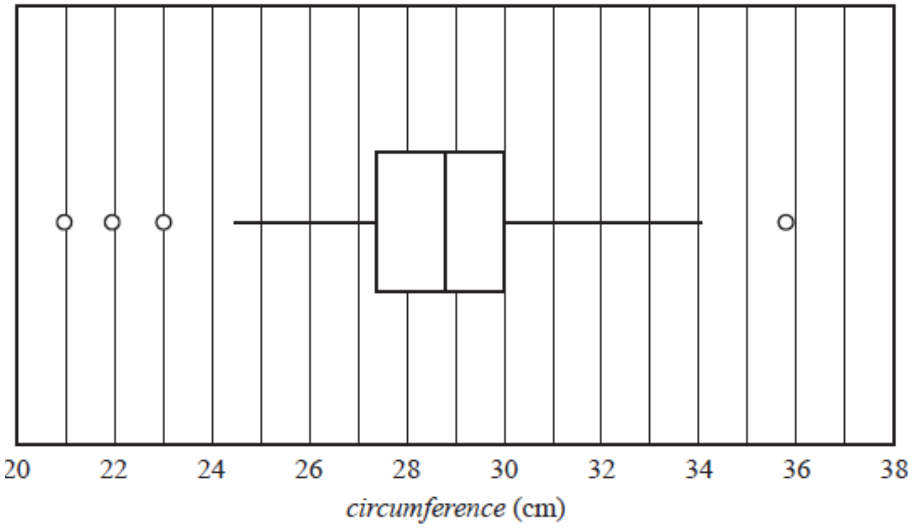
$\therefore$ Option C

C

Exam Styled Questions (current study design) – Multiple Choice

Use the following information to answer Questions 1 and 2.

The boxplot below shows the distribution of the forearm *circumference*, in centimetres, of 252 people.



Question 1
(2017 Exam 1 Section A – Qn 1)

The percentage of these 252 people with a forearm *circumference* of less than 30 cm is closest to

A. 15%
B. 25%
C. 50%
D. 75%
E. 100%

From the box plot it can be seen that Q_3 is equal to a circumference of 30 cm. This means that 75% of the data has a circumference < 30 cm.

∴ Option D

D

Question 2
(2017 Exam 1 Section A – Qn 2)

The five-number summary for the forearm *circumference* of these 252 people is closest to

A. 21, 27.4, 28.7, 30, 34
B. 21, 27.4, 28.7, 30, 35.9
C. 24.5, 27.4, 28.7, 30, 34
D. 24.5, 27.4, 28.7, 30, 35.9
E. 24.5, 27.4, 28.7, 30, 36

$X_{min} = 21$
$Q_1 = 27.4$
Med = 28.7
$Q_3 = 30$
$X_{max} = 35.9$

B

∴ Option B
(**NB:** Outliers do count towards the 5 number summary)
