Note Title 14/10/2008

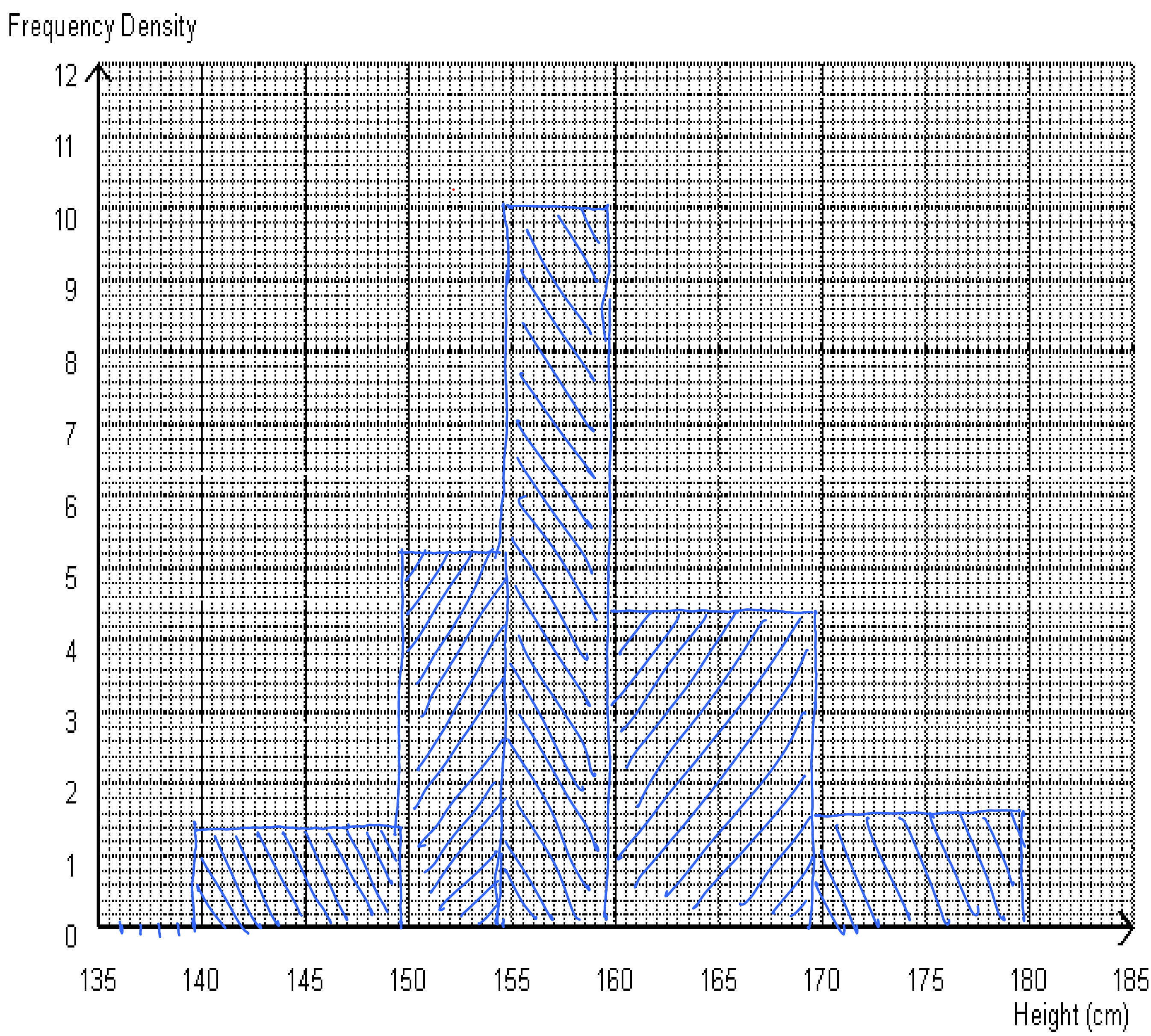
Histograms

Example: The table below shows the heights (in cm) of 150 6th Form girls. Draw a histogram to illustrate this.

Height	Frequency	Class Boundaries	Class Width	Frequency Density $= \frac{\text{Freq}}{\text{Class Width}}$
140 - 149	14	$139.5 \leq x < 149.5$	10	1.4
150 - 154	26	$149.5 \leq x < 154.5$	5	5.2
155 - 159	50	$154.5 \leq x < 159.5$	5	10
160 - 169	44	$159.5 \leq x < 169.5$	10	4.4
170 - 179	16	$169.5 \leq x < 179.5$	10	1.6

Note:

- Horizontal axis **need not** start at zero
- Vertical Axis **must** start at zero
- Label Axes!



- No gap between bars!
- Horizontal axis has a <u>scale</u>, not labels

Box and Whisker Plots

Example: The reaction times (in seconds) of 33 female students were recorded in a stem and leaf diagram as follows:

Reaction Times — Key 1|3 means 0.13s

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

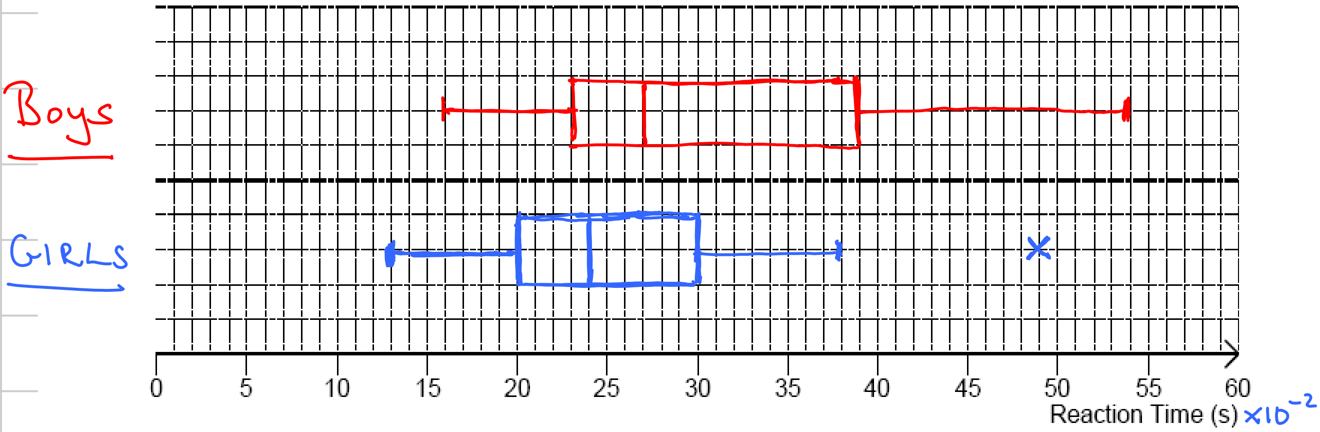
The results were analysed , and the median and quartiles were found to be:

$Q_1 = 0.20$s $\quad Q_2 = 0.24$s $\quad Q_3 = 0.30$s

The reaction times of a set of boys were also analysed, with results as follows:

Least value = 0.16s $\quad Q_1 = 0.23$s $\quad Q_2 = 0.27$s $\quad Q_3 = 0.39$s $\quad$ Greatest = 0.54s

Draw a box plot illustrating these data. Outliers should be taken as values which are 1.5 × IQR above Q_3 or below Q_1. Comment on the graphs.



Girls : $1.5 \times IQR = 1.5 \times 0.10 = 0.15$s

$Q_3 + 0.15 = 0.30 + 0.15 = 0.45$s so 0.49s is an outlier

$Q_1 - 0.15 = 0.20 - 0.15 = 0.05$s so no outliers at the bottom end

Boys : $1.5 \times IQR = 1.5 \times 0.16 = 0.24$

$Q_3 + 0.24 = 0.63$ so no outliers at top end

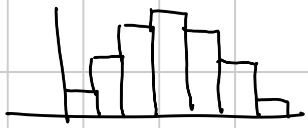
$Q_1 - 0.24 = -0.01$ so no outliers at bottom end

Notes on Box and Whisker Plots

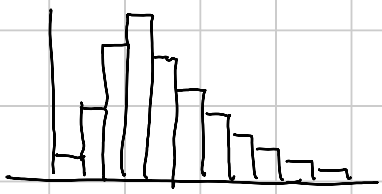
- Often called box plots for short – 'box plot' still includes the whiskers.
- There is no standard way of deciding on outliers – any method required will be specified in the question. The method used in this example is the most common.

Skewness

A distribution can be SYMMETRIC :–



or POSITIVELY SKEW



or NEGATIVELY SKEW

(reflect the diagrams above!)

p67 Ex 4A Q 12, 15(c). By next Tuesday

Another property of a skew distribution is that

POSITIVE SKEW	$\Longleftrightarrow$	mode < median < mean
NEGATIVE SKEW	$\Longleftrightarrow$	mode > median > mean

So to identify skewness we can use this, or

Use

$$\boxed{\begin{array}{lcl} \text{POSITIVE SKEW} & \iff & Q_3 - Q_2 > Q_2 - Q_1 \\ \text{NEGATIVE SKEW} & \iff & Q_3 - Q_2 < Q_2 - Q_1 \end{array}}$$

or just look at a box plot or a histogram.

Which Measures to Use?

The mode and range are seldom used; the range only depends on the two most extreme values.

The mean and standard deviation have the advantage of taking into account every data value, and should be used UNLESS

- there are OUTLIERS which distort the distribution

or

- the distribution is noticeably SKEW,

in which case the median and interquartile range should be used.

$$\boxed{\begin{array}{llll} \text{p 67} & \text{Ex 4A} & \text{Q 13} \\ \text{p 111} & \text{Review Ex 1} & \text{Q 50} \end{array}}$$