Representations of Data Cheat Sheet

Stats/Mech Year 1

Outliers

An outlier is commonly any value which fits into one of the following:

- Greater than $Q_3 + k(Q_3 - Q_1)$
- Less than $Q_1 - k(Q_3 - Q_1)$

The value of k will be given in the exam

Some questions have other ways of identifying the outliers. In the exam, you will be told which method to use.

Example 1: Some data is collected. $Q_1 = 46$ and $Q_3 = 68$. A value greater than $Q_3 + k(Q_3 - Q_1)$ or less than $Q_1 - k(Q_3 - Q_1)$ is defined as an outlier. Work out if a)7, b)88 and c)105 are outliers. The value of k is 1.5.

$$68 + 1.5(68 - 46) = 101$$
$$46 - 1.5(68 - 46) = 13$$

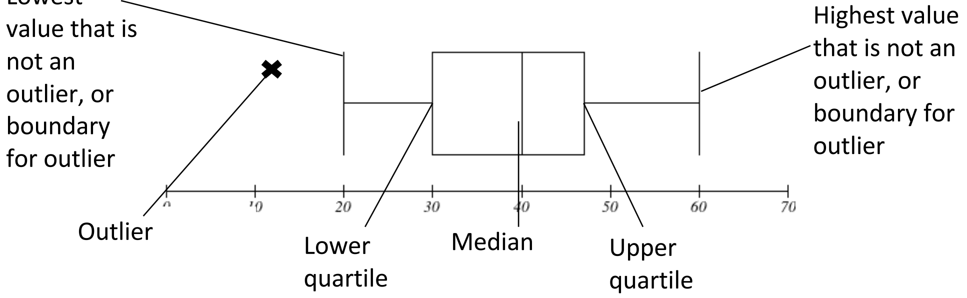
7<13 and 105>101 so 7 and 105 are outliers, 88 is not an outlier.

In some cases, the outliers are legitimate values which will still be correct. Some outliers are clearly an error and they are called anomalies. They can be due to experimental or recording error, or data values not relevant to the study. The process of removing anomalies from a data set is called data cleaning.

Boxplots

A boxplot shows the quartiles, maximum and minimum values and any outliers

Two sets of data can be compared using boxplots.


Lowest value that is not an outlier, or boundary for outlier
Highest value that is not an outlier, or boundary for outlier
0
10
20
30
40
50
60
70
Outlier
Lower quartile
Median
Upper quartile

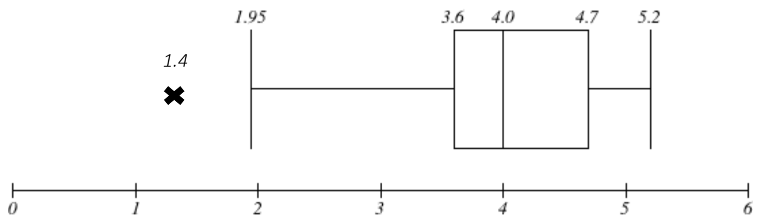
Example 2: The blood glucose level of 30 males is recorded. The results, in mmol/litre, are summarised below:

Lower quartile: 3.6
Upper quartile: 4.7
Median: 4.0
Lowest value: 1.4
Highest value: 5.2

An outlier is an observation that falls either 1.5x interquartile range above the upper quartile or 1.5x interquartile range below the lower quartile.

a. Given that there is only one outlier, draw a boxplot for this data:

1. Calculate the value of outlier:
   $3.6 - 1.5 \times 1.1 = 1.95$
   $4.7 + 1.5 \times 1.1 = 6.35$
   1.4 < 1.95, therefore the outlier is 1.4.
2. Draw the boxplot and label the axis.
   The end of the whisker is plotted at the outlier boundary since the actual figure is not known.



Blood glucose level (mmol/litre)

Cumulative Frequency

You can use a cumulative frequency diagram to help find estimates for the median, quartiles and percentiles in a grouped frequency table.

Histograms

Group continuous data can be presented using histograms. Histograms show the rough location and general shape of the data, and how spread out the data is.

The area of the bar is proportional to the frequency of each class.

To calculate the height of each bar (frequency density):

Area of bar= $k \times$ frequency

When $k = 1$,

$$\text{Frequency density} = \frac{\text{frequency}}{\text{class width}}$$

Joining the middle of the top of each bar in histogram forms a frequency polygon.

Example 3: A random sample of 200 students was asked how long it took them to complete their homework. Their responses are summarised in a table:

Time, t(min)	$25 \leq t < 30$	$30 \leq t < 35$	$35 \leq t < 40$	$40 \leq t < 50$	$50 \leq t < 80$
Frequency	55	39	68	32	6

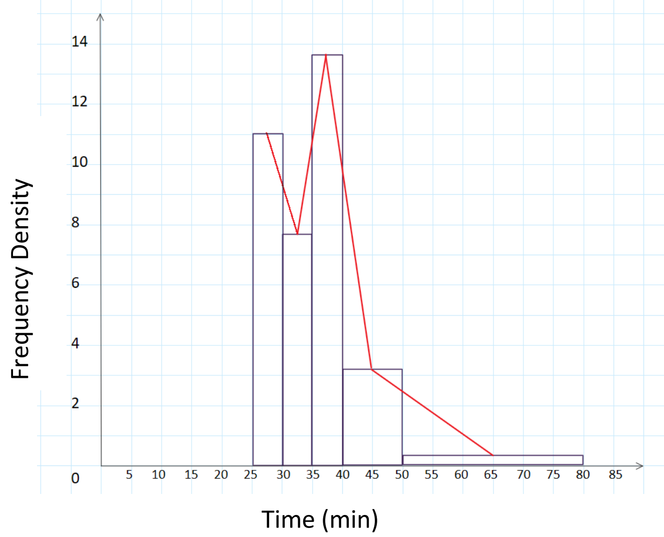
a. Draw a histogram and frequency polygon to present the data.

1. Find the class width and frequency density of each class.

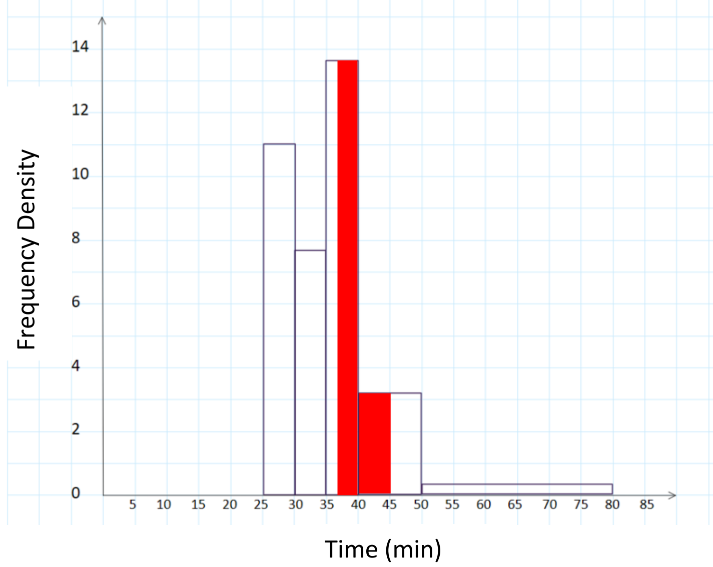
Time, t(min)	Frequency	Class width	Frequency density
$25 \leq t < 30$	55	5	11
$30 \leq t < 35$	39	5	7.8
$35 \leq t < 40$	68	5	13.6
$40 \leq t < 50$	32	10	3.2
$50 \leq t < 80$	6	30	0.2

$$\text{Frequency density} = \frac{\text{frequency}}{\text{class width}}$$

2. Draw the histogram using class width as the width of each bar and frequency density as the height.
3. To draw the frequency polygon, join the middle of the top of each bar of the histogram.



b. Estimate how many students took between 36 and 45 minutes to complete their homework.



The number of students is directly proportional to the area under graph between 36 and 45 minutes.

Area: $(40 - 36) \times 13.6 + (45 - 40) \times 3.2 = 70.4$ students

Comparing data

When comparing data, you can comment on

- A measure of location
- A measure of spread

You can use the mean and standard deviation or median and interquartile range (suitable for data sets with extreme values) Median should not be used with standard deviation and mean should not be used with interquartile range.