Revision of Science Maths 3

Data can be either *qualitative* or *quantitative*.
The type of data is important in determining the method used to describe it.

- *Qualitative* data can be described by tabular and graphical methods
- *Quantitative* data can be described by tabular, graphical and numerical methods.

Figure:

Revision of Science Maths 3

Data can be either *qualitative* or *quantitative*.
The type of data is important in determining the method used to describe it.

- *Qualitative* data can be described by tabular and graphical methods
- *Quantitative* data can be described by tabular, graphical and numerical methods.

Revision of Science Maths 3

Qualitative data is divided into categories and the number of cases falling into each category is counted.
The number of cases in each category is called the **frequency.**
Example: In a sample of $n = 50$ students the exam grades they receive for a particular module. There are

- 20 students with A grades
- 15 with B grades
- 10 with C grades
- 5 with D grades.

Revision of Science Maths 3

Table: The frequency distribution table of the grades.

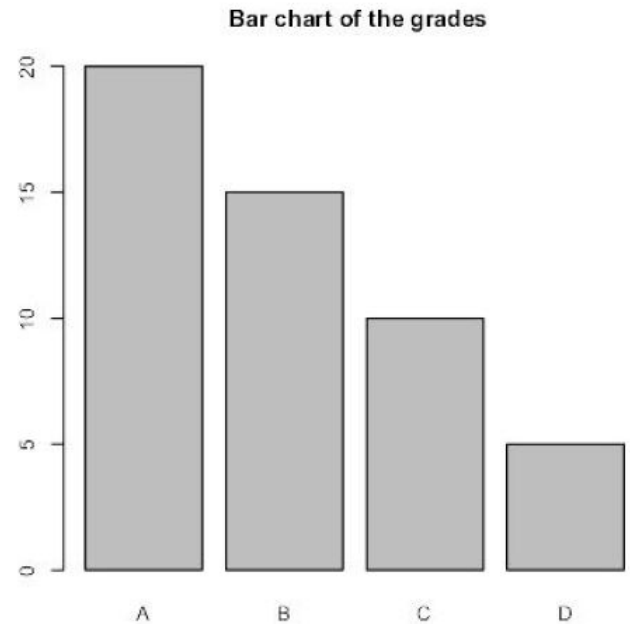
Grade	Frequency	Relative Frequency
A	20	0.40
B	15	0.30
C	10	0.20
D	5	0.10
Totals	50	1

The **relative frequency** is the proportion, or percentage, of items in each class. For a data set with n observations, the relative frequency of each class is:

$$RelativeFrequency = \frac{Frequency}{n}$$

Revision of Science Maths 3

A **bar chart** represents the frequency of the grades.
On the horizontal axis we specify the labels for the four classes: A, B, C and D. Draw a bar of fixed width above each class label. The length of the bar is equal the frequency.



Revision of Science Maths 3

- In quality control applications, bar charts are used to identify the most important causes of problems.
- When bars are arranged in descending order of height from left to right, with the most frequently occurring cause appearing first, the bar chart is called a **Pareto diagram**.

Revision of Science Maths 3

Quantitative data can also be summarized by a frequency distribution like the qualitative.
However, this requires more work since our data is numeric i.e. instead of just having categories like grades like (A, B, C, D) we may have all numbers between 1 and 100.
How do we group this data?

Grouping quantitative data

Use the following steps:

- Calculate the range of the data i.e. the largest value - smallest value.
- Divide the range into a number of intervals called class intervals - between 5 and 20.
  The width of each interval is the range divided by the number of class intervals.
- Count the number of cases falling into each interval.

Revision of Science Maths 3

Example . Consider the hours of personal computer usage during one week for a sample of 20 individuals. This is called raw data.

Table: The number of hours of personal computer usage.

12	14	19	18	15	15	18	17	20	27
22	23	22	21	33	28	14	18	16	13

Grouping data

Steps:

- Calculate the **range** $range = Max - min$= 33-12=21
- Chose the **number of intervals** $\rightarrow$ 5
  (since sample size small n=20).
  Calculate the **width** of each interval

$$Width = \frac{Range}{NumberIntervals} = \frac{21}{5} = 4.2$$

  For convenience we will use a class width of 5.
- Count the number of cases falling into each interval.

Revision of Science Maths 3

Note: If we take the first class of 10 - 15, note that 10 is the lower limit and 14 is the upper limit of this class. We consider the intervals to be left-closed and right open, i.e. the 10 - 15 interval contains 10 but **does not** contain the value 15.

Revision of Science Maths 3

Table: The frequency distribution table for the number of hours.

Interval	**Freq.**	**Relative Freq**	**Cumulative Relative Freq**
10-15	4	0.20	0.20
15-20	8	0.40	0.60
20-25	6	0.25	0.85
25-30	2	0.10	0.95
30-35	1	0.05	1
Total	20	1	

Revision of Science Maths 3

The **relative frequency** expresses the frequency counts as a fraction of the total number of cases.
The **cumulative relative frequency** expresses the relative frequencies on a cumulative basis. It shows the fraction of items with a value less than the upper limit of each class.

- For example, 85% of the sampled people spend less than 25 hours weekly on a PC.

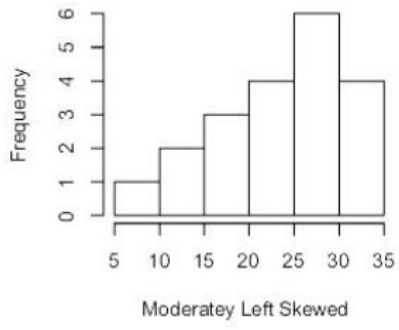
Histograms

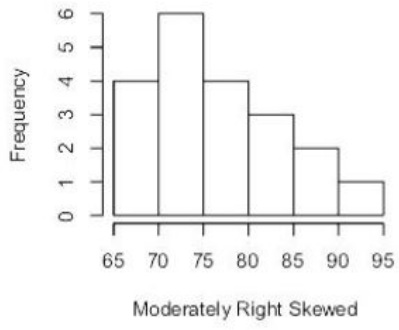
Histograms are a common graphical representation of the quantitative data. It can represent either the frequency or the relative frequency information.

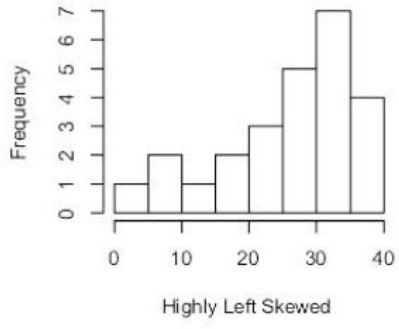
- The variable of interest is placed on the horizontal axis and the vertical axis represents the frequency scale.
- The heights of the rectangles drawn on the histogram are equal to or proportional to the frequencies.

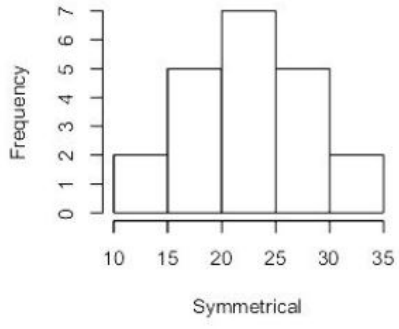
Revision of Science Maths 3

Histograms









Revision of Science Maths 3

This histogram is positively skewed since its right tail is longer than the left tail. This is caused by large values such as 33 which are are very different from the majority of cases.

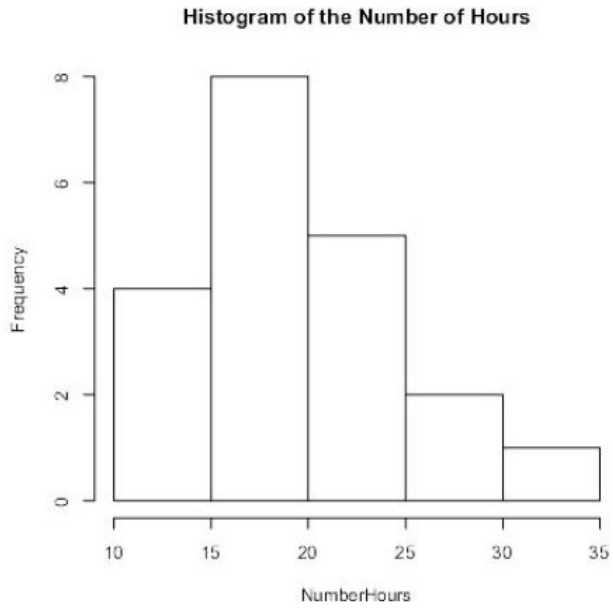


Figure: Histogram for the frequency of the number of computer usage hour.

Revision of Science Maths 3

Ogive

Ogives are the graphical representation of the cumulative relative frequency information.

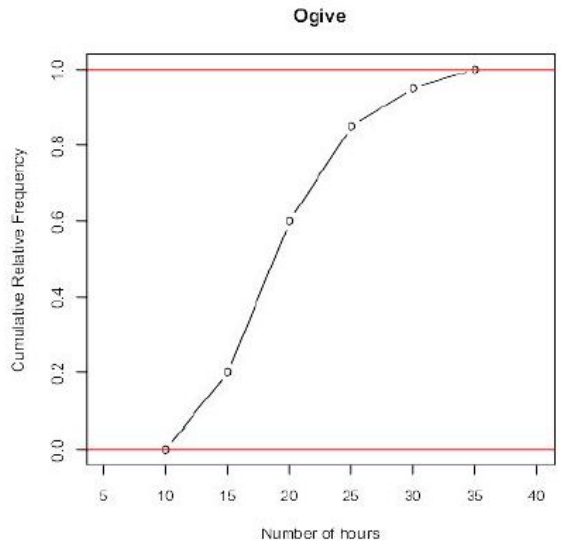


Figure: Ogive for the cumulative relative frequency of the number of computer usage hours.

Revision of Science Maths 3

Quantitative data can be described by tabular, graphical and **numerical methods**.

There are two main numerical descriptive measures:

1. **Measures of centrality** which measure the centre of the distribution.
2. **Measures of variability** which measure the dispersion of the data.

Revision of Science Maths 3

When we know the middle of our distribution and how spread out it is about the middle, we have two numbers which create a concise numerical summary of the data.

The measure of centrality and variability we use depends on what our data looks like i.e. the shape of the distribution.

Revision of Science Maths 3

Measures of centrality give one representative number for the location of the centre of the data. The most commonly used measures are:

- Mean
- Median
- Mode

Mean

The **mean** (average) = $\frac{\text{the sum of the observations}}{\text{the total number of observations}}$.

We use different symbols to distinguish between the

- mean of the sample $\overline{x}$
- the mean of the population μ

Sample Mean

The sample mean $\overline{x}$ (x bar) and it is calculated as:

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + ... + x_n}{n}$$

where n is the size of the sample and x_i is any element in the data set : $\{ x_1, x_2, ..., x_n \}$.

Population Mean

The population mean is denoted by the Greek letter μ (mu) and it is calculated as:

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + ... + x_N}{N}$$

where N is the size of the population.

Median

- The **median** is defined as the value of the number in the middle position when the data are arranged in numerical order.
- It splits the distribution into two halves.
- If n = odd, the median is the middle number, i.e. the number in the position $\frac{n+1}{2}$.
- If n = even, the median is the average of the two middle numbers, i.e. the average of numbers in position $\frac{n}{2}$ and $\frac{n}{2}+1$.

$\{2,5,6,7,9\} \rightarrow$ Median=6
$\{2,5,6,7,9,13\} \rightarrow$ Median=$\frac{6+7}{2}$=6.5

Mode

The mode of a set of data is defined as the number which occurs most frequently.

Value	3	4	5	6	7	8	9
Frequency	2	3	2	2	1	1	2

Mode = 4

Comparing Measures of Centrality

Usually use the mean or the median to describe the centre of our distribution. How do we decide between them?

Data: 1.8 2.7 3.5 4.6 5.4

Mean = 3.6 Median = 3.5

Change the number **5.4** to **54**.

Data: 1.8 2.7 3.5 4.6 54

Mean=13.22 Median=3.5

Revision of Science Maths 3

The mean is sensitive to extreme values whereas the median is unaffected by extreme values.

Histogram is symmetric $\rightarrow$ mean = median = mode.

Histogram is symmetric $\rightarrow$ use the mean as measure of centrality.

Histogram is not symmetric $\rightarrow$ use the median as measure of centrality.

Revision of Science Maths 3

Measures of variability give one representative number for the dispersion of the data around the centre. The most commonly used measures are:

- Range
- Variance and standard deviation
- Interquartile Range

Range

Range = maximum point - minimum point

Since it is calculated using only 2 values, it tells us nothing about the data in between these two values and therefore, conveys the least information.

Variance

The variance is a measure of dispersion around the mean $\rightarrow$ can only be used with the mean.
Variance takes into account all the data values.

If the data are tightly clustered about the mean $\rightarrow$ the variance is small.

If the data are widely scattered around the mean $\rightarrow$ the variance is large.

Population Variance σ^2

For a population of size N and with mean μ, the population variance is:

$$\text{Population Variance} = \sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

The population standard deviation is the square root of the variance = σ.

$$\text{Population Standard Deviation} = \sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$$

Sample Variance s^2

For a sample of size n and with mean $\overline{x}$, the sample variance is:

$$\text{Sample Variance} = s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$$

The sample standard deviation is the square root of the variance = s.

$$\text{Sample Standard Deviation} = s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}}$$

Example

Calculate the sample variance and the sample standard deviation for the data: 3, 2, 4, 7, 4.

Sample size n=5

Sample mean $\overline{x} = \frac{3+2+4+7+4}{5} = \frac{20}{5}$=4

Sample variance $s^2 = \frac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{n-1}$= $\frac{(3-4)^2+(2-4)^2+(4-4)^2+(7-4)^2+(4-4)^2}{5-1}$= $\frac{14}{4}$=3.5

Sample standard deviation $s = \sqrt{s^2} = \sqrt{3.5}$= 1.870

Percentiles

A **percentile** is the point below which a certain percent of the observations lie.

The 50^{th} percentile is the point below which half the observations lie.

Important percentiles:

- $Q_1 = 25^{th}$ percentile or lower quartile
- $Q_2 = 50^{th}$ percentile or median
- $Q_3 = 75^{th}$ percentile or upper quartile

Interquartile Range

IQR = $Q_3 - Q_1$ is a measure of variability that is commonly used for skewed data.
IQR = width of an interval that contains the middle 50% of the sample,
IQR is smaller than the range and its value is less affected by outliers

Boxplots

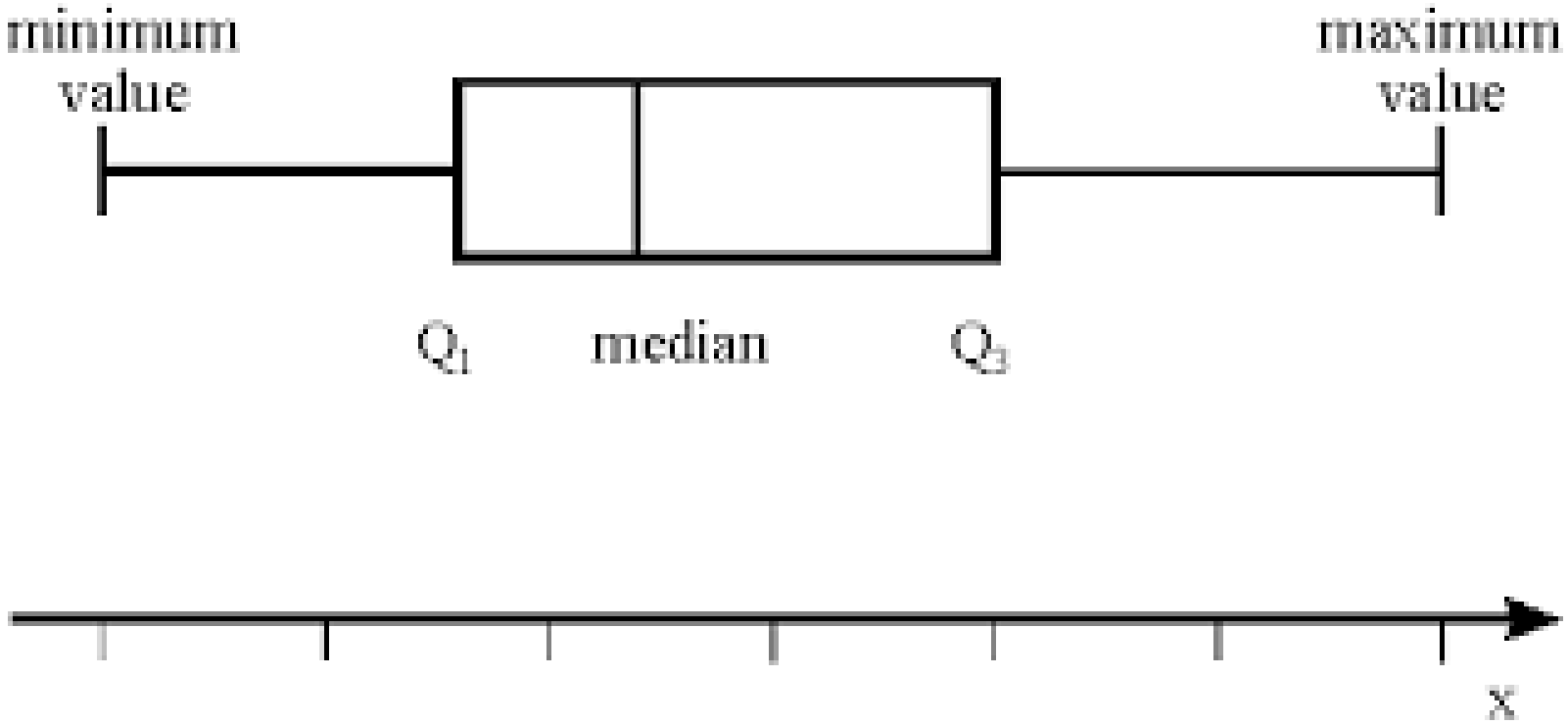
- Boxplots are uniform in their use of the box: the bottom and top of the box are always the first and third quartiles, and the band inside the box is the median.
- But the ends of the whiskers can represent several possible alternative values (see next slide. We will use the **Tukey Boxplot** variant).
- **Outliers:** Any data not included between the whiskers should be plotted as an outlier with a dot, small circle, or star, but occasionally this is not done.

Revision of Science Maths 3

Boxplots: Different ways of computing "Whiskers"

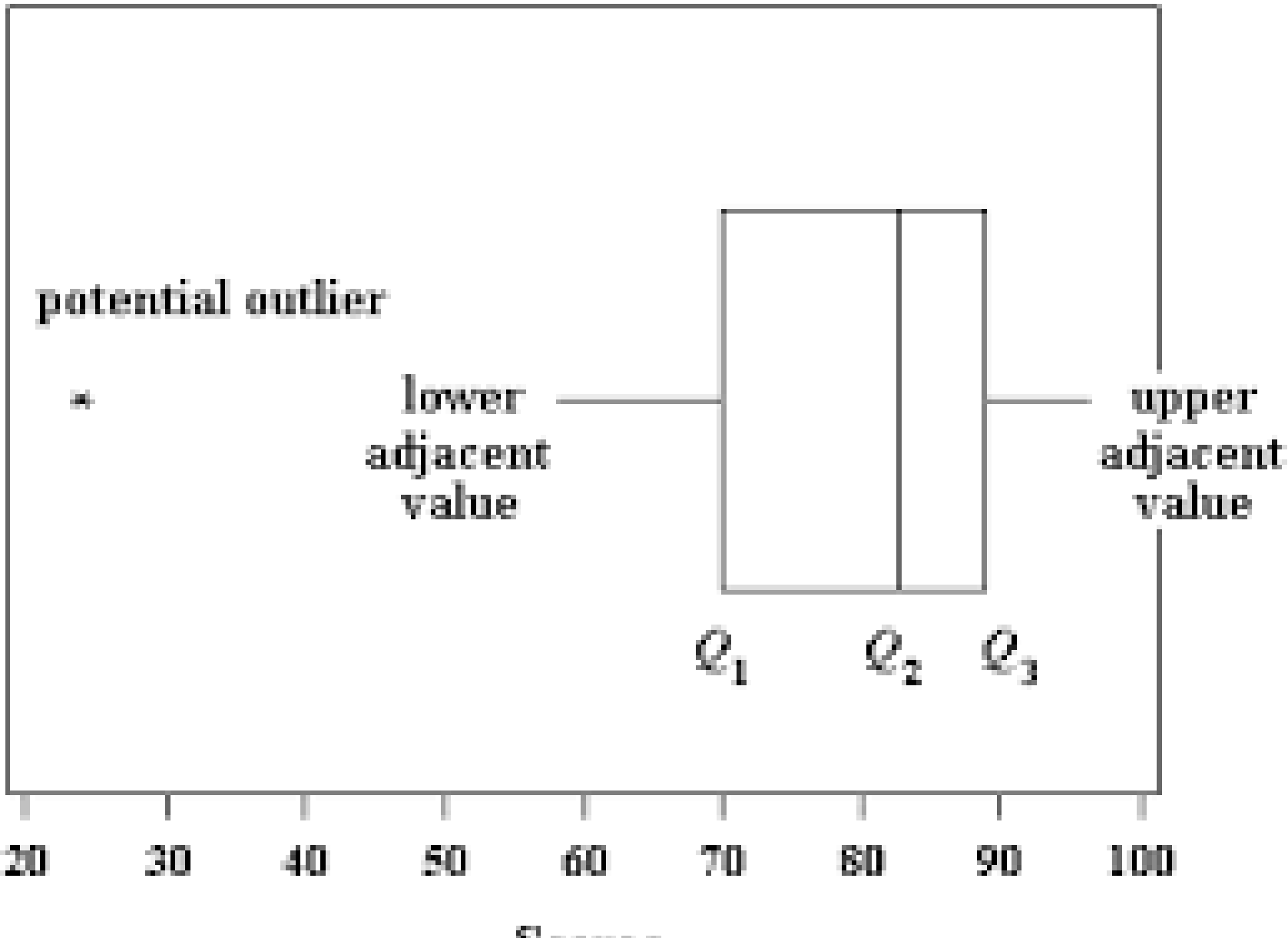
- the minimum and maximum of all of the data
- **Tukey boxplot** the lowest datum still within 1.5 IQR of the lower quartile, and the highest datum still within 1.5 IQR of the upper quartile.
- one standard deviation above and below the mean of the data
- the 9th percentile and the 91st percentile
- the 2nd percentile and the 98th percentile.

Structure of Boxplot, when no outliers are present

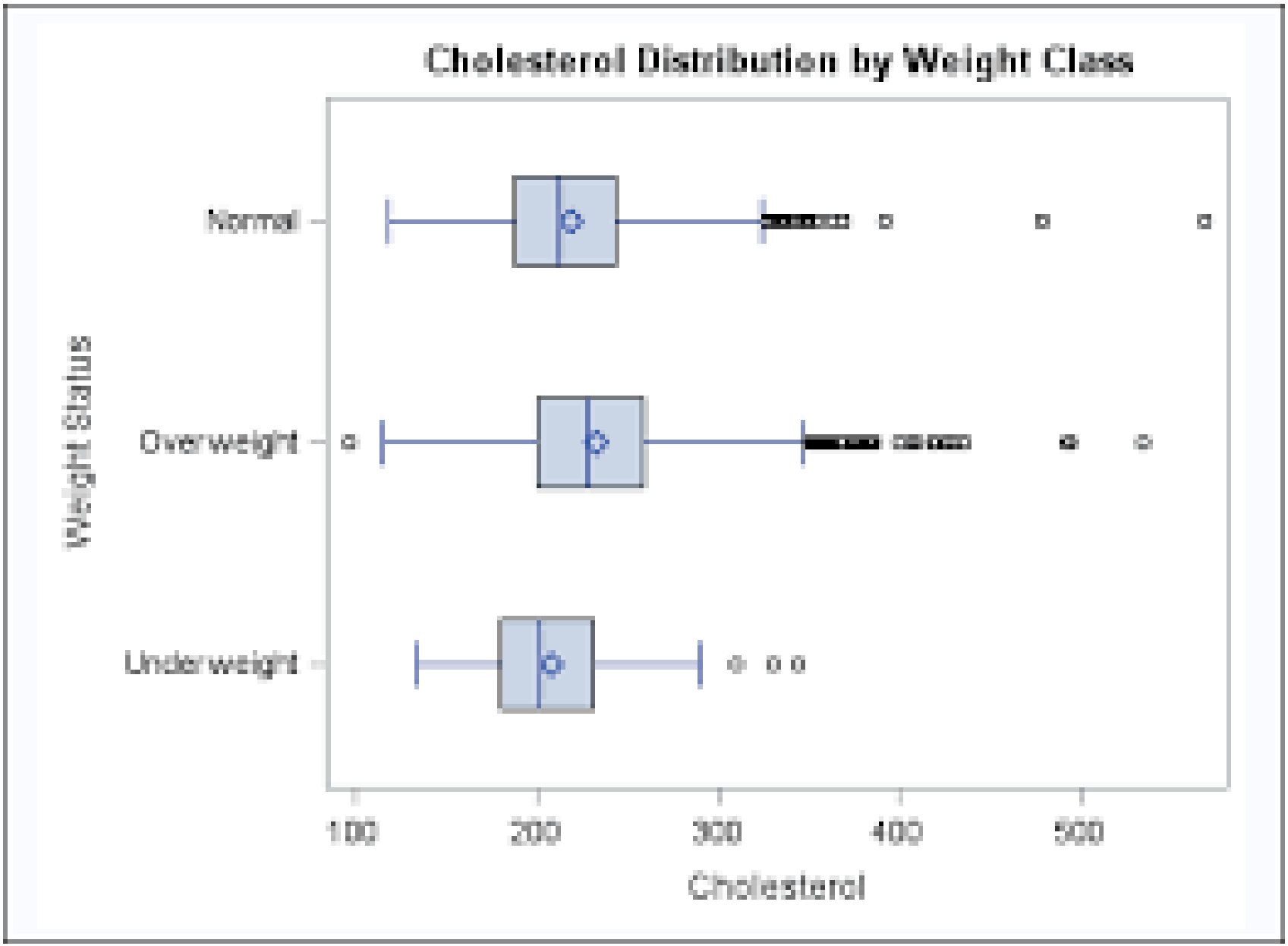


Revision of Science Maths 3

Boxplots can indicate outliers



Revision of Science Maths 3



Coefficient of variation

What happens if you have two sets of data with two different means and two different standard deviations?

How do we decide which set is more spread out?

Coefficient of variation

The coefficient of variation (cv) is used to compare the relative dispersion between two or more sets of data.

It is formed by dividing the standard deviation by the mean and is usually expressed as a percentage i.e. (multiplied by 100). We distinguish between the population and sample coefficient of variation.

$$CV_{population} = \frac{\sigma}{\mu} * 100$$

$$CV_{sample} = \frac{s}{\overline{x}} * 100$$