# Achievement Statement 2.5

## Statistics

Numerical facts about populations are called parameters.

The parameters of most interest are:
- the population mean $\mu$
- the population standard deviation $\sigma$
- the population proportion $\pi$

To find the values of these parameters we must carry out a census (a survey of the whole population).

Usually this is impractical, so we make a sample (a sub-gr of the population).
Numerical facts about samples are called statistic.
The statistics of most interest are:
- the sample mean $\bar{x}$
- the sample s.d. $s$.
- the sample proportion $p$.

We use our sample statistics to make estimates of the population parameters.

## Sampling

A statistical survey involves gathering information from a population.
The target population is the population of ~~interest~~

eg
- all pine trees in a forest
- all sheep on a farm
- all retired people living in New Zealand

It is usually not possible to survey the whole population, so we take a sample.

A sampling frame is a list of the target population. Frequently used sampling frames are:

- a school roll
- the telephone book
- the electoral roll

## Selecting a Sample

A good sample must be representative of the population. It should include each strata in approx. the correct proportions.

It should not include outliers. The inclusion of outliers may lead to an overestimate or an underestimate of the population parameters.

The sample should be sufficiently large to get reliable estimates of the population parameters.
Statistics calculated from a small sample are unlikely to be reliable.
If we were to take another sample, we could get quite different results.

### Commonly used sampling methods are:

(assume we are selecting a sample of n=50 from a population of 700 students on the school roll)

#### Cluster sampling

The population can often be divided into groups or clusters.
eg. Our sampling of 50students could simply be found by randomly choosing 2 classes.

This method is quick, cheap, and easy to do, but is unlikely to be representative of the population.

eg. Both classes could be from Yr 9

Random sampling

We require an alphabetic listing of the school roll with each student assigned a number from 1 to 700.

- We set our calculator to generate random integers between 1 and 700.
- Match each number generated with a student from the list.
- Ignore repeats
- Continue until a sample of 50 is chosen.

The advantage of this method is that no bias exists. Every student has an equal chance of being chosen.

This method does not guarantee that the sample will be representative of the population.
eg. Yr 11 students could be over-represented in our sam

## Stratified Sampling

Sometimes the population can be divided into layers or strata.

e.g.

1) For the population "all trees in a forest" the strata may be
   - trees on a exposed hillside
   - trees in a shady gully
   - trees on a fertile plain

2) For the population "students at OBHS" the strata are the 5 year groups in the school.

   If our sample is to be representative of the population we would randomly select from each strata, in ~~population~~ proportion to the size of the strata.

   eg. If 24% of the school roll are year 11 students, then our sample of 50 students should include $0.24 \times 50 = 12$ randomly chosen from year 11.

To use stratified sampling the boundaries of each strata must be clearly defined.
Stratified sampling is more expensive, more complicated and more time consuming than other sampling methods

[redacted]

## Systematic Sampling

There is a system, or pattern to the way the sample is chosen.

eg. To select our sample of n=50 students from 700 students we could take every 14th person from the number list.

This can guarantee a good spread of data throughout the population. Clusters of data in the sample will not occur.

## Sample Statistics

### Measure of average

The mean $\bar{x}$ is the arithmetic average

It can be affected by outliers so that $\bar{x}$ is an overestimate or an underestimate of the population mean $\mu$.

It does however give an overall picture of the average data value

$$\bar{x} = \frac{\Sigma x}{n}$$

eg. For 5, 8, 8, 11, 16

$$\bar{x} = \frac{\Sigma x}{n}$$

$$= \frac{5+8+8+11+16}{5}$$

$$= \frac{48}{5}$$

$$= 9.6$$

The median is the middle data value.
It is not affected by outliers, but does not give an overall picture.

eg 1) 48 48 49 52 99

Median = 49

2) 48 48 49 52 54 58

Median 50.5

## Measure of spread

The standard deviation 's' is the most commonly used measure of spread. On a calculator, to find 's' we use

$x\sigma n$ function

It is a measure of the average deviation of the data values from the mean
If the standard deviation is large (more than 20% of mean) this tells us that there is a large variation in the data values.
If we were to take another sample we could get quite a different result, whereas, if the standard deviation (s.d.) is small a second should give similar results.

Standard deviation is affected by outliers, but gives a good overall picture of spread as all data values contribute.

Find the mean and s.d. of the data
23, 28, 32, 39, 44, 47, 57, 60
mean $\bar{x}$ = 41.25
s.d. $s$ = 12.45

The interquartile range is the spread of the middle 50% of the data

Interquartile range = U.Q - L.Q.

It ignores values at the extremes so is not affected by outliers, but does not give an overall picture as only half the data contributes to its value.