Level 1 Multivariate Data

The requirements of this topic are that students plan and conduct investigations using the statistical enquiry cycle

- justifying the variables used
- identifying and communicating features in context (differences within and between distributions), using multiple displays
- making informal inferences about populations from sample data
- justifying findings, using displays and measures.

It is not sufficient to merely produce some statistics and say something vague about them. There must be a discussion about what the samples say about the population. This the "inference" stated above, which in general language is your conclusion or answer.

For the purposes of this year, students are expected to use the *statistical enquiry cycle* (Problem, Plan, Data, Analysis, Conclusion) in their investigation.

The understanding of, and correct use of, terminology is a vital part of this topic. The following words are absolutely key and you must be comfortable with them:

Population – All the members of some group. For example, all people aged 15 and over living in New Zealand on 1 January 2012. (Note how precisely the population is limited.)

Statistical measure or statistic – A number representing a property of a population or sample. Common examples are the population median and inter-quartile range.

Categorical – values which are not numerical and have no natural order, such as sex or colour.

Discrete – values of a variable that can only take on exact number values, usually whole numbers. Examples include number of children or percentage score in a test.

Continuous – values of a variable that can take any value inside a suitable range. That is there is always a potential value between any two other values. Note that for practical purposes we often round continuous distributions. Examples include weight and amount of rain fallen.

Sample – a selection of some members of the population. A **survey** is a measure of one or more variables from a sample (whereas a census is the same process applied to the entire population).

Feature – a property of the sample, such as clustering or skew, which makes it different from "normal".

Data

Your investigation will use **Data**.

> A collection of facts, numbers or information; the individual values of which are often the results of an experiment or observations.

In this unit the data will be **Multivariate**.

> Each item in the population will have measurements for different *variables*.

Missing Data Points

Often data sets are incomplete.

You must not attempt to even out sample sizes by leaving out data.

One of the reasons why statisticians use measures such as mean and median is so that sample of different sample can be compared.

Missing data points are not a problem unless they introduce bias (for example, if you are testing the effectiveness of a drug, the death of a person is not just a random missing result, but very important).

Calculate the statistical measures normally and then to discuss any possible effects of missing data in the analysis.

Sampling "Error"

Sampling error arises because every sample gives different values, even if the sample is completely random and unbiased. So any statistic calculated from a sample will differ from the same statistic calculated from the next sample. As a result, no sample statistic can be assumed to be anything but an *estimate* for the actual population measure.

The sampling error can be reduced only by taking larger samples, but can never be removed.

For this reason even if there is a difference between two measures in our samples – say one median and another – does not automatically mean that the difference is also in the populations. It may be due only to random chance.

Other Errors

Our aim with sampling is to get a sample that is, as far as possible, **Representative**.

> One where all groups of different elements of the population appear in the sample in proportion to their distribution in the population.

When discussing the meaning of your results, you should not move into populations where the sample are not representative.

> A random sample of Year 11 students taken at one school will be representative of Year 11 students at that school, but not necessarily of other year groups, and almost certainly not of other schools.
>
> A sample taken only in Hamilton will not necessarily be representative of the whole country.

Sampling needs, most of all, to avoid **Bias**.

> – An influence that leads to results which are different from the true value in a particular direction, e.g. consistently too high or too low, but also too varied or regular.

A **biased sample** is one in which the system used to create the *sample* produces results that are not *representative* of the *population* not just randomly, but reliably wrong in a particular way.

You can assume that the samples you are given in the assessment are not heavily biased.

Features of Data

While your answers must focus on the statistical measures of your samples you can also discuss the non-measureable features of the data separately.

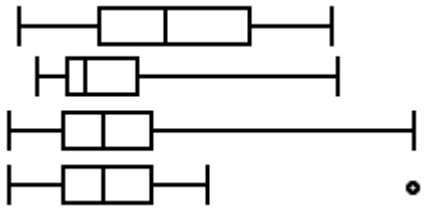
This should be done with care, as students are prone to considerably under-estimating the effects of natural variation. Just by sheer chance odd things are going to happen from time to time.

Symmetry and Skew

If the spread of values on one side of the median tend to be further away than on the other side, the data is skewed.

True skew requires that **both** the inner quartile box and whisker are longer on one side than the other. A long whisker by itself can be caused by a spread out series or extreme values without any actual skew – a tail – or by an outlier.

Small samples tend to lack good symmetry thanks to natural variation, even when the population is itself symmetrical. Students should therefore not read too much into a small lack of symmetry.



Normal spread

Skewed

Long whisker, which may be a tail of values or just an outlier

How to represent the above to show an outlier

Outlier

A variable's value significantly different from most other values in a sample.

An outlier may be genuine, indicating an individual of particular interest. Or an outlier may be the result of a mistake.

Students are far too prone to interpreting any extreme value as an outlier. Natural variation will always throw up extreme values.

You cannot have a group of outliers. By definition they must be lone values.

Students should, in general, do all their calculations including any extreme values, and then mention in their analysis the result of this (for example dragging the mean away from the true value).

Cluster

A distinct group of values in a distribution.

Ignore the presence of light clustering in small samples. Clusters will always occur as a result of natural variation, especially for the small sample sizes students deal with.

However very obvious clustering may represent a real features and is worth discussing. In particular a cluster at the ends may represent something important, as normally you expect the extreme values to be more spread out than the inner ones.

An extreme form of clustering is when there are two strongly separate groups visible in the data, each with their own centre and spread. Such values are said to be "bi-modal".

Measures of centre

These measure the typical value for a variable. Students should more or less stick to the *median* and the *inter-quartile range*, perhaps mentioning the *mean* if it adds something useful.

Median

The middle value of a data set when placed in order.

The median is the most stable measure of centre as it is not influenced by the random appearance of extreme values.

Mean

Calculated by adding the values and then dividing this total by the number of values.

The mean can be heavily influenced by unusually large or unusually small values. Generally it should be stressed far less heavily than the median, especially for skewed distributions.

Measures of spread (measures of variability, measures of dispersion)

A number that conveys the degree to which values in a distribution differ from each other.

Interquartile range

The width of an interval that contains the middle 50% of the values in the distribution.

The difference between the *upper quartile* and *lower quartile*.

The interquartile range is a stable measure of spread in that it is not influenced by unusually large or unusually small values.

Range

The difference between the largest and smallest values in the distribution.

The range is strongly influenced by single values – in this case the largest and smallest. Therefore it should be used with caution. You can talk about the spread excluding extreme values if that helps.

Finding the Median and Quartiles

Sort the data into order. The median is the middle one.

If there are an even number of samples the median is halfway between the two middle values.

e.g. for: 3, 3, 4, 5, 6, 6 the median is 4.5, being halfway between the middle two values.

The lower quartile is the value a quarter the way along the sorted data. It is found by taking the middle value of all values below, but not including, the median. The upper quartile is the 75% value, found by taking the middle value of all values above the median. As with the median calculations, if this falls between two values the middle of them is taken.

e.g. 30, 31, 31, 31. 35, 36, 39, 40, 40, 41, 43, 44, 49

The lower quartile is 31, and the upper quartile is 42 (shown with arrows).

The process of finding median and quartiles should divide the data into four equal quarters.

Writing a Question

The structure of the assessment will require you to write a "question" that can be answered using the data you are given.

It is important to achieve at Year 11 that the question be at a sufficiently high level.

- The data must be **divided into two groups** to compare. Usually it is best to divide based on some category such as sex or education.
- The question should ask if a **difference between the groups** can be seen.

  You should avoid asking questions which can be answered with one word answers. Instead ask "Is there is a difference between Group A and Group B" or "Has there been a change from Time 1 to Time 2".
- It must use **measurable statistics**.

  It is no good asking which is "best" or if they are "different" unless "best" or "different" is well defined by some number value.
- The question must **talk about the population**, not the sample.

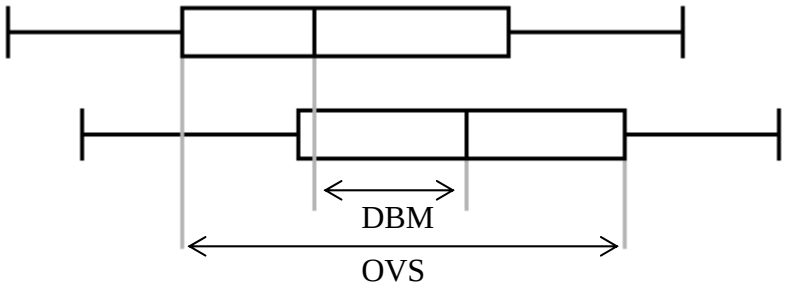
  The sample statistics are used to estimate the population, so we can say something meaningful about the real world.

The answer can be negative. You are not punished for finding that some measure is not increased by some activity or that different parts of the population have the same values. So do not spend excessive time looking for something exciting in the data to report on.

Do not expect to use all the data presented to you in the test. That generally means that you can cross out all the data columns except the one you used to group the data into two categories, and the measurement that you are investigating.

"Making the Call"

When deciding whether the difference between medians in the sample is likely to represent a real difference between medians in the populations, and not just be due to random variation ("sample error") we can use the DBM – difference between medians – compared to OVS – overall visible spread.



For sample size around 30, the DBM must be more than a third of OVS to make the call that the difference between medians is significant – not just due to sample variation.

For a sample size around 100, the DBM must be more than a fifth of OVS to make the call.

Notes on Conclusions

1) **Compare medians** as your starting point to talk about typical values.

   "**Make the call**" about whether the DBM compared to OVS allows you to say the difference between them is "significant" – that is we can be sure the higher median in the sample is also the higher in the populations – or we that we cannot tell whether the higher median is not just due to "sample error" (random variation in sample).

   You can then bring up the difference in the means if you think it will add anything (but bear in mind they can be distorted by extreme values or skew).

2) **Compare Inter-Quartile Ranges** to discuss the range of typical values, using words like "consistent", "inconsistent", "variability" and "spread of values".

   You can compare the differences between the quartiles themselves, not just the IQR, especially if there is a clear pattern.

   Compare the Ranges, but don't just look at the box-and-whisker graphs, but take into account the effect of any individual extreme values.

3) Discuss briefly the overall **features** of the data or graphs.

   a) whether the distributions are symmetrical, skewed or have a tail.

   b) any extreme values or outliers.

   c) whether the data is smooth or clumped.

   d) any other differences between the two graphs.

   Discuss whether the difference between the features is important in the real world.

6) Directly answer the question you posed at the start of the assessment.

   Relate your conclusion to the populations. The purpose of sampling is to make some conclusion (inference) about the real world. Where do you think the result might apply? Clearly state any assumptions you make when doing this.

   Include any negative conclusions, such as that you cannot see any significant differences between the samples (and as a result, between the populations).

   Include variability as well as the typical values when making a conclusion, if you can.

   Don't get too caught up with saying the result "proves" anything. Saying it is "consistent with" or "supports" is far better.

Try to always put **numbers** with all comments, including the size of differences – you have them so you should use them. Feel free to refer to actual data points.

Comments must be **statistical** in nature. If the comment isn't about the data, graphs or statistics, it probably isn't relevant. Keep any comments about why something is seen very short.