Sources of Variation

Each time we take a sample set of measurements there is always some differences.

Natural Variation

Individual-to-individual variation or real variation is everywhere and is in everything, because all individuals are different they will give differences in their measurements.

The range of this variation is normally what we are trying to find using statistics.

Occasion-to-occasion variation happens because measurements on the same individual may change over time. For example, an individual's blood pressure changes during the day.

Measurement variation happens when repeated measurements on the same individual are different because the measurement device has its own variation. A slightly different placement of a ruler might give different results with different people using it.

Induced variation happens when the same quantity is affected by a change in other factors. For example, the difference in growth of two tomato plants from the same packet of seeds planted in the same garden could be due to differences in the growing conditions at slightly different places (soil fertility or exposure to sun or wind).

Sample variation

This happens because each sample is of different individuals. If a mean is calculated from a sample, a second sample will have a different mean, because the individuals in the same are different for the reasons given above.

This variation increases if the samples are taken at different times (due to occasion-to-occasion variation) or with different devices (measurement variation) or from different places (induced variation). So when comparing groups we try to reduce this variation as much as possible, by using the same devices at the same time and using groups that have as little change in other factors as possible.

Ideally measurements should always be spread across time and place to reduce the variation, but usually this does not happen for practical reasons.

Answering about Sources of Variation

When looking for ways in which a sample set of data might have sources of variation, you should consider the following:

- When the measurements were taken
- How the measurements were taken
- How the measurement was recorded
- Are they relying on unreliable methods, such as personal memory?

Bias

This is when our way of selecting the sample is likely to affect the result in a particular direction.

Either the way of selecting might be bad (asking who speaks Māori, but on a marae), the questions might be badly worded ("Are you an idiot who supports the So-and-so Party?") or just ignore some of the population (trying to measure the number of people who are unemployed, but only asking people who have a permanent address).

Examples of Issues

There is often nothing we can do about these issues, especially if we are given the data already collected, but we need to recognise that they exist. In other cases there may be better ways to collect the data, but those methods might be too expensive, too difficult or annoy those being sampled.

Time

If we want to see how much chocolate people eat, then if we measure around Easter or Christmas we are very likely to get much higher than usual.

It is better to see how much chocolate is sold each year, as that averages out.

How far SJC students run each week will be affected by the season, because more players train in winter (but others might run more in the nice weather of summer).

We should ask the same questions throughout the year, to get the full range.

How

If we ask a person how much they have in the bank they are likely to give a wrong number – they may exaggerate to look good, reduce it to seem humbler, or even just be wrong.

Ideally we would ask to see their bank balance.

If we measure the number of trucks along a road, we will always struggle to get two people to define exactly what a “truck” is.

Firm and agreed definitions should always be decided at the start.

When measuring obesity we usually use the BMI formula, because it is simple. However it is not a reliable measurement for different ages, different races and different levels of physical activity.

If we want to measure something we should measure it directly, not use some other measurement because it is easier.

Recording

Some measurements are rounded. For example weights are usually in whole kgs.

Ideally we would measure ourselves using accurate scales

Many results are collected for ranges. For example the data might have the incomes for people from 20 to 29. But the amount a 29-year-old earns is usually much more than they earned as a 21-year-old.

Ideally we would collect by exact ages.

Other

If you ask how many times in the last year a person has been to the movies, they are unlikely to remember exactly.

It is better to ask for shorter periods, so that memory is less of an issue.

If you ask a people to say how many friends they have, they will all have a different understanding of the word “friend”.

It is better to have a fixed measurable value than a vague term.

Privacy and Cultural Issues

While we would ideally collect data directly, by seeing how much a person has in the bank or measuring their exact weight we often cannot do this as often people do not want to give out personal details. When this happens we need to work with this and accept that the amount of variation in the answer is likely to be higher because we are not measuring directly.

Even if the data has already been gathered, such as the tax authorities knowing how much you earn, it is considered an invasion of privacy to release that data in a way that allows people to be identified.

Some data is considered as a *taonga* and needs to be respected as such. Data on the number of Māori spoken in an area, for example, might be best collected by the *iwi* involved, and kept by them.