**Lecturecast Week 8: Summary Measures and Inference**

**Level of Measurement**

Here, we look at a hierarchy of 4 levels of measurement of dataset variables. They are from the highest, most informative to the lowest, least informative level

1. Ratio Scale: applies to numeric data, differences between values are meaningful
2. Interval scale: applies to numeric data, differences between values are meaningful
3. Ordinal data: applies to numeric data but the numbers are used for ranking, differences between values are not meaningful
4. Nominal Data: represents a simple classification e.g sex, marital status etc.

The type of statistical methodology applied to data investigation depends on the level of the data.

Higher levels 1 & 2 are quantitative or metric data. Lower levels 3 & 4 are qualitative, non-metric data.

Discrete and Continuous variables only apply to quantitative data. Discrete variables are whole numbers (integer) while continuous are all possible values within a range.

**Graphical Summaries**

Bar charts can be used to summarise qualitative and discrete quantitative data

Histograms on the other hand are for continuous quantitative data.

**Summary Measures**

To summarise numerical data (so we don’t have to look through every single observation), there are two means:

1. Location, the value taken by a typical observation
2. Dispersion, which gives some idea of how spread out the data is.

Sample mean and Sample standard deviation

The sample mean and sample standard deviation can only be used for quantitative (i.e. interval and ratio) data.

The sample median and interquartile range represent alternative measures of location and dispersion respectively that can be used both for quantitative and for ordinal data.

* The sample median M is defined to be that value below which 50% (i.e. one half) of the observations lie.

The first sample quartile Q1 is defined to be that value below which 25% (i.e. one quarter) of the observations lie.

The third sample quartile Q3 is defined to be that value below which 75% (i.e. three quarters) of the observations lie.

Our measure of dispersion is then the interquartile range IQR = Q3 – Q1. The “middle 50%” of the observations fall within the interquartile range.

**How do we decide which to use to measure location and dispersion?**

The sample mean and sample standard deviation OR The sample median and interquartile range.

We’ll need to consider the sensitivity of various measures to extreme observations.

**Inference**

Statistical inference takes two forms;

1. Using the data to estimate some underlying population value.
2. Using the data to investigate whether some assumption of interest regarding the underlying population is likely to be true – Hypothesis testing

Hypothesis Testing

There are 4 steps in hypothesis testing:

1. State the hypothesis
2. Set the Criteria for testing
3. Compute the Test Statistic
4. Make a Decision.