*Chapter Two notes* ***for Agri-food research***

Summary from page 72

*A variety of statistics can be used to summarize the empirical distribution of data-points, including measures of location and spread.*

Mean, median and mode – location of the histogram.

Skewness and standard deviation – spread of the histogram.

*Skewed data distributions are common, and some summary statistics are very sensitive to outlying values.*

For example, income. In class we have the bad apples data.

The mean is different to the median when a distribution is skewed (or asymmetric).

Robust summaries – like Spearman rank correlations (look up the book glossary!) can help reduce the influence of outliers in summary statistics.

Nothing beats plotting the data and examining a wide range of summary statistics alongside the plots.

*Data summaries always hide some detail, and care is required so that important information is not lost.*

Think about the dinosaur and “five groups” example in the book and correlation coefficients – the chart is more informative in this case.

*Single sets of numbers can be visualized in strip-charts, box-and-whisker plots and histograms.*

The histogram is fundamental and the best way (Ian says) and visualising an empirical distribution. But there are alternative ways to display and highlight particular attributes of datasets and their distributions (which affects how we **compare distributions**).

*Consider transformations to better reveal patterns, and use the eye to detect patterns, outliers, similarities and clusters.*

Yes – but we are not going to focus on transformations in our class (despite them being important in advanced model analysis).

*Look at pairs of numbers as scatter-plots, and time-series as line-graphs.*

Yes yes. That is why we are using R.

*When exploring data, a primary aim is to find factors that explain the overall variation.*

This means focusing on standard deviation and the “sampling error” of what we are interested the most. In our case, things like the **mean yield for different treatments is the key interest: so the question becomes “how accurately have we estimated the means” – this turns out to be a simple function of the standard deviation of the underlying data.**

*Graphics can be both interactive and animated. Infographics highlight interesting features and can guide the viewer through a story, but should be used with awareness of their purpose and their impact.*

Not so important in our course.