*Chapter “Introduction”*

Bullet points from the chapter are shown below in italics. Comments are made that relate the chapter contents to agri-food research.

*Turning experiences into* ***data is not straightforward****, and data is inevitably limited in its capacity to describe the world.*

Where does data come from in Agri-food research?

1. **Experiments e.g. in a greenhouse or lab**

Growth trials, like dope plants in a greenhouse. We take measurements at end, *destroying* the plants (chemical analysis and biomass).

Growing a sunflower “model” with different fertiliser levels. Here we might take measurements each week, and you get a time series (like height, leaf count and bud size).

Straight-forward to set-up controls, like equalising temperature, sunlight and initial conditions (like soil and moisture).

1. **Field trials**

Like an experiment, but *more noise in results.* The noise can be both *between* trials (say Launceston and Hobart arms of a trial) and within a particular field, relative to treatments.

Between trials:

No control over things like relative temperature, rain, soil conditions etc.

Within trials:

Imperfect control over giving each treatment the same conditions in terms of moisture, drainage, weeds, aspect to sun etc.

1. **“Observational trials” – for example, going to an orchard managed by a farmer.**

Usually over two or more seasons. Seasons are very different – the “season effect” is mostly noise and giant catchall term.

Little control over the set-up or “randomisation” of treatments to plants. For example, you cannot prune-and-train a cherry tree one way in season one and then a completely different way in season two.

You cannot control for initial conditions or randomise treatments – the set-up is just the way it is. For example, if the row that gets more nitrogen just happens to have at the start of the trial healthier vines, better soil or more sun, then anything you conclude about nitrogen is very tentative

*Statistical science has a long and successful history, but is now changing in the light of increased availability of data.*

Has much changed in Agrifood research in recent decades?

**Experiments.**

1. On one hand, not much has changed. **Experiments in labs and greenhouses are still roughly the same – expensive to run and small in scale.** The data and sample sizes look similar to the time of R A Fisher (one hundred years ago).

This includes microbiology set-ups which are handled in lab settings and very laborious. Many things cannot be automated.

1. On the other hand, there are many more labs doing similar experiments – it is an open question how to combine possibly related research results. Options include so-called meta-analysis projects. This is beyond the scope of our course.

**But technology has increased our capacity to take raw measurements.**

1. There are now vast measurements from remote technology e.g. satellites taking pictures of biomass across different days.
2. Soil and moistures sensors send back data to a base *every single second*. How can this be processed and made-sense-of?
3. We can now produce a very sophisticated taxonomy of living organisms with advanced DNA analysis. For example, you might find 50,000 different things living in a teaspoon of soil.
4. Chemical analysis is a capable of isolating hundreds of different things from very small samples.

In many cases you end up with lots more data than you have *information*. Careful choices need to be made about taking averages (e.g. averaging soil moisture per day, from per minute readings?) and reducing dimensions (e.g. lumping 50,000 organisms in fungi, bacteria and other?). Taking averages, and lumping/splitting choices, are fundamental to our course (and our use of R).

Modern practical problems include gear malfunctions and missing values. For example: cows wear monitors that record every breath etc. but they destroy them; batteries fail in soil sensors; and clouds prevent pictures being taken around rainy periods (exactly when you might want them). We learn to account for missing values in our course (initially with Rs “NA” placeholder in matrices).

*Skill in statistical methods plays an important part of being a data scientist.*

A good agri-food researcher designs, collects and analyses data. You also need to read other peoples’ research with a critical eye. The basic methods of statistical science are needed at *each* stage.

For example, you need to try to gauge the potential noise in your measurements, in order to plan for an adequate sample size (“n”). This is called “power analysis”, in which a higher n helps because you “average out” the noise and detect and interesting “treatment effect” or whatever. Power analysis is done very well in biostatistics, but very poorly, or not all, in agri-food research. We will be looking at power in our course.

The use of graphics and basic statistical summaries (e.g. means and standard deviations of “treatment A” contrasted with those from “treatment B”) are the most important steps in understanding data. From there more complex models can be built (but they are typically only the “icing on the cake” and are the least convincing part of agri-food research.)

*Teaching statistics is changing from a focus on mathematical methods to one based on an entire problem-solving cycle.*

This makes statistical classes more interesting and more practical.

*The PPDAC cycle provides a convenient framework: Problem – Plan – Data – Analysis – Conclusion and communication.*

Diagram

Description automatically generated

*Data literacy is a key skill for the modern world.*

Indeed – all the analysis (I would say over-analysis in many cases because there is so much gross uncertainty involved) of COVID 19 data is case in point. Here is David Spiegelhalter said (Feb 22 - yesterday) about the scrapping of virus rules in the UK: “I can see the figures are looking encouraging, but the consequences are very difficult to predict. **It may be fine, it may not be,”** he said – adding that surveillance of the virus should continue through the Office for National Statistics’ infection survey.

In general, data literacy is needed to criticise the justifications used by politicians (and Vice Chancellors) about public or institutional policy – our leaders all say they are “data-driven” and their decisions are “evidence based”; but in what sense and how can an individual challenge things?