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

Summary from page 94

*Inductive inference requires working from our data, through study sample and study population, to a target population.*

Sample versus population example.

**Sample:** the measurements taken by your soil moisture probe, in a ten metre grid, in three paddocks, on particular days of a particular year.

**Study population:** those paddocks on *all* days and *all* spots within the paddocks (your grid was patchy and you didn’t take the measurements every day).

**Target population:** anopen question – other paddocks in the region across multiple years? Other soil types (all your paddocks are sandy loam but have you learnt something about soil moisture for clay soils? – maybe yes or maybe not).

Induction versus deduction example.

**Deductive inference**: working out pieces of “knowledge” based on assumptions (which may or may not be true) and simple rules about how to make conclusions from sets of assumptions.

**Inductive inference:** forming “new knowledge” based on empirical observation. The process of learning about general principles from specific examples.

The main difference between inductive and deductive inference is that induction is inherently ***risky*.**  A deductive conclusion is always “correct” because it merely assumes things to be true. Induction *relies* on things being true.

Checking and reasoning about our assumptions in statistics is really important – induction is more “useful” than mere deduction.

**Reliability:** something works again and again, perhaps with a known error rate.

**Valid:** something that correctly gets to a conclusion, given the assumptions. Assumptions may be wrong but a deductive conclusion can still be correct.

*Problems and biases can crop up at each stage of this path.*

Think of the iterative cycle of using data to find things out about the world.

Map

Description automatically generated with low confidence

The path contains a variety of instances in which you would want to go from a sample, to a study population and then generalise to a target population. Sometimes you may have re-adjust your ambitions and downscale the claimed generalisations. Or, if you are lucky, it might be that unforeseen generalisations can be made!

There are many potential **biases** in Agrifood research. The main one is having to use a particular set-up on a farm/orchard/vineyard in which you simply cannot “randomly select” things for different treatments. The bias here is that there may be something systematically different about the individuals being measured that coincides with the “treatments” you are interested in. For example, if all the cherry trees trained to be tall were chosen for that training system because they were the most vigorous after the initial planting (it may appear that the training that the ”tall training system” beats other systems just because the trees are healthier!).

Just be honest – like this from a recent cherry trial we did at TIA (accepted for publication):

The sampling scheme for the data was defined by the pre-existing orchard layout and its established management practices. Thus, the sample is essentially observational, rather than “experimentally designed”. All observations come from a four-hectare “Trial Block” within the orchard, which contains rows of trees at various age, cultivar, training system and rootstock combinations. Five different training systems were chosen that had trees consistent in age, cultivar and rootstock. A diagram of the Trial Block and sampling scheme is in Figure 1. The basic sampling unit is a tree (where there are multiple measurements taken from a tree, the values are averaged). Six trees were randomly selected from each of the five training systems, with the selection of trees being different for both seasons. This means *n* = 30 for each season. The random tree selection was done with two constraints. First, the Trial Block was rendered in half by a path, so for each season, three trees on either side of the path were selected (this step was arbitrary and the side-of-path structure was not incorporated into any models). Secondly, the random selection was done such that each tree was at least two trees away from any other tree in the sample. In summary, within a training system and season, we treated the six trees as being randomly selected from the same “population”. For the ANOVA analysis, *n* = 30 for the one-way ANOVA linear models and *n* = 60 for the two-way ANOVA linear models (which incorporate a system-by-season interaction). There were no missing values.

The best way to deal with potential biases that you cannot avoid is to openly acknowledge the limitations of your research. Often this means that the “target population” of your work is not as broad or wide-ranging as what you might hope.

*The best way to proceed from sample to study population is to have drawn a random sample.*

This can help spread the risk of bias. But “observational” studies (like following the crops from different trees trained in a particular way) are still useful. An “experimental” sample, e.g. in which trees are randomly assigned a training system, may be better in theory but impossible in practice.

*A population can be thought of as a group of individuals, but also as providing the probability distribution for a random observation drawn from that population.*

Ultimately, we want to plot the distribution of our samples and have this **represent** a wider population. Summary statistics, such as means, standard deviations and quantiles (percentiles) help summarize these distributions

The idea is to go from summaries of samples to generalisations about wider populations.

The risky “inductive assumption” is always something akin to “my histogram is indicative of the wider population” so I have learnt “such and such about the wider world”.

You can reasons for (and against) why such assumptions are good, but you can never know for sure.

*Populations can be summarized using parameters that mirror the summary statistics of sample data.*

See above.

*Often data does not arise as a sample from a literal population. When we have all the data there is, then we can imagine it drawn from a metaphorical population of events that could have occurred, but didn’t.*

This is usually not the case in Agri-food research – our samples are usually from small-scale experiments (labs and greenhouses) or from limited field studies.

In theory we might have “all the data”. For example, assume that you measured the fertiliser levels and yields for every truffle farm in Tasmania. This would more-or-less be a summary of the “entire [local] truffle industry”. But why would you do this? Because you want to know about the future or reason about counterfactual setups e.g. what will fertiliser levels do next year for truffles; or what would happen if you altered the fertiliser levels in the low yielding plots etc.