STA2020 ANOVA Notes

Ané Cloete

2024-12-17

Table of contents

Preface

This book is not an exhaustive guide for designing an experiment or conducting ANOVA's, it has been tailored specifically for the learning outcomes and methods covered in STA2020. If you are interested in a more general text on experimental design, please see:

Instructions:

- Examples
- R code
- Tips
- Footnotes

Statistical Modelling

What is a Model?

A **statistical model** is a mathematical representation of how data is generated. It describes the relationship between observed data and underlying factors (parameters) while accounting for random variation. Suppose that we are interested in estimating the age of a tree from its stem diameter. To do this we need to know by how much the stem diameter increases per year. We could describe this relationship or process as follows:

$$D = \alpha + \beta \times Age$$

describing a linear increase of diameter with age. Once we have a good idea of how fast diameter increases with age ( ) we can predict diameter from age. The (mathematical) model above is a very simple representation of this process with only two parameters, the intercept and the growth rate.

With the chosen parameter values, diameter increases linearly with age. Of course, this model is not realistic except for special situations but it gives us powerful insights. In reality we don't know β, but usually need to estimate it from data. Also, not every tree grows equally fast, because of environmental and individual differences between trees. We can accept that the above is a simple model for the average behaviour of a tree, but to capture variability between trees (because of variability between environmental conditions from tree to tree, variability between individual trees, measurement error), we add an error term.

$$D = \alpha + \beta \times Age_i + e_i$$

The response that we observe is then described by an average behaviour, but the actual observed value will vary around this average. To summarise, the statistical model has a stochastic component which captures variability in the response that cannot be explained by the deterministic part of the model. Another distinguishing feature of statistical modelling is that we obtain estimates of the parameter values from the data, e.g. by fitting a line to the observations, i.e. we learn from data.

More generally

Statistical models are not perfect predictors of the data, rather they attempt to describe the "central tendency" of the observations. To get to the actual observed value some deviation from the central tendency needs to added (i.e. error). Such models typically have the following the form:

$$\text{Observed Response} = \text{Model Predicted Response} + \text{Error}$$

Mathematically this can be stated as:

$$Y = \hat{Y} + e$$

A simple example of a statistical model you may have encountered is the **mean** as a predictor. Suppose you measure the number of customers entering two stores over 20 days. The observed counts for each store fluctuate daily, but you may want to summarize the data using the average number of customers.

For each store i, a basic statistical model for these observations would be:

$$Y_{ij} = \mu_i + e_{ij}$$

where:

- Y_{ij} is the number of customers observed on day j at store 1,