Statistical Thinking in Biology Research

An introduction

Terry Neeman

Australian National University

9th August 2022

A few key ideas

- Statistics in biology is the study of biological variation
- Understanding biological variation informs experimental design
- Understanding biological variation informs data analysis

Statistical thinking is an essential component of scientific thinking

Statistical methods in biology - 20th century

- Agricultural experiments in Rothamsted Station, UK
- Stochastic processes in genetics
- Clinical trials

Figure 1: R.A. Fisher 1890 - 1962

The ideas from these intellectual movements gave us foundations for how we think about and interpret data as scientific evidence.

Some false narratives ("cautionary tales")

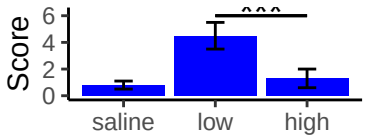
When teaching statistics, these ideas can get distilled, degraded into a simplistic and false narrative.

“Statistical analysis is all about getting a p-value”

Vaccine challenge experiment

- 6 mice per vaccine group (saline/ low dose / high dose)
- All mice challenged with Shigella bacteria at Day 14
- Outcome: 7-day average symptom score post-challenge

Statistical analysis: one-way ANOVA, p=0.04 post hoc Bonferroni adjusted

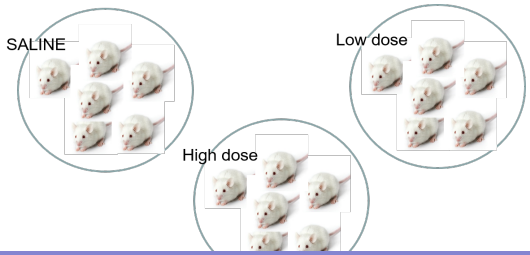


Was there a cage effect or a vaccine effect?

The observed difference in symptom scores could be due to:

- animal cage
- vaccine treatment

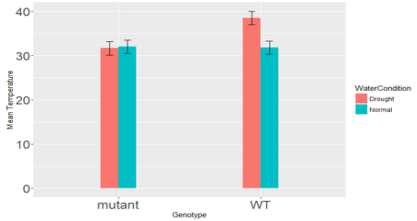
These two factors are **CONFOUNDED**. It is impossible to separate out these two effects.



"P > 0.05 means 'same'; P < 0.05 means 'different'"

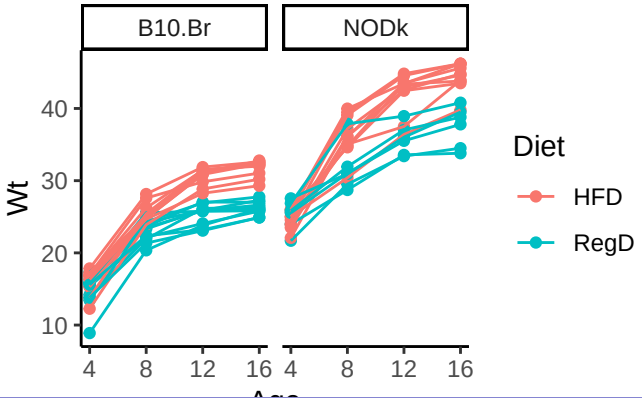
Experimental set-up: Are temperature mechanisms modified in a genetically-engineered tomato plant?

- Genotypes: WT or mutant
- watering conditions: normal or drought
- Outcome: leaf temperature at 7 days post-treatment



"When in doubt, use lots of t-tests"

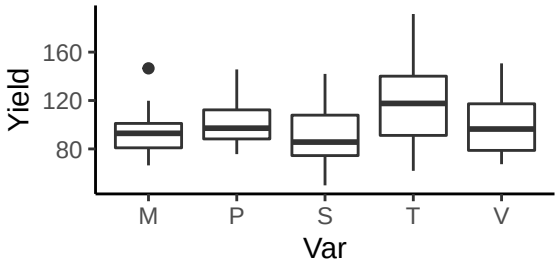
Research questions: Are mice susceptible to obesity when exposed to a high fat diet? Are NODk mice **MORE** susceptible than B10.Br mice?



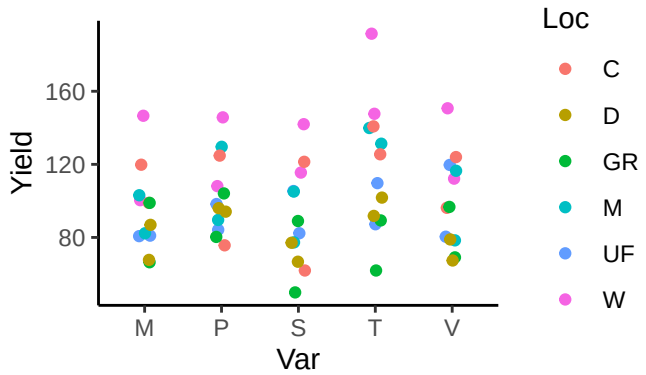
“More than 2 groups? Use 1-way ANOVA”

Research question: Which barley variety has the biggest yield?

- Five barley varieties, grown in 6 locations
- Two growing seasons
- Outcome: yield (tonnes/hectare)

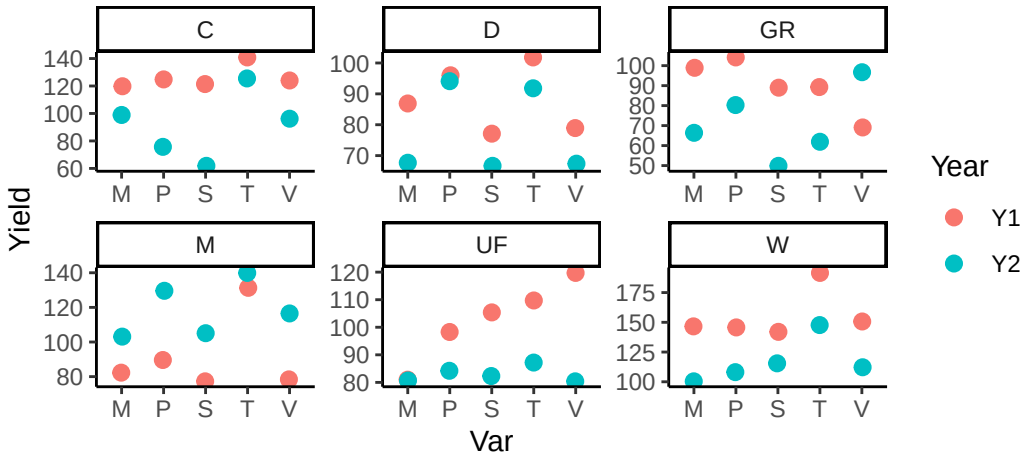


Location contributes to the variation in yield



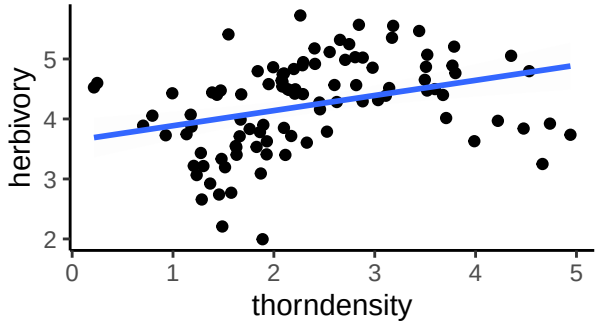
Yield is highest in Location W. Yield is lowest in Location GR.

Varieties should be compared *within* location and year



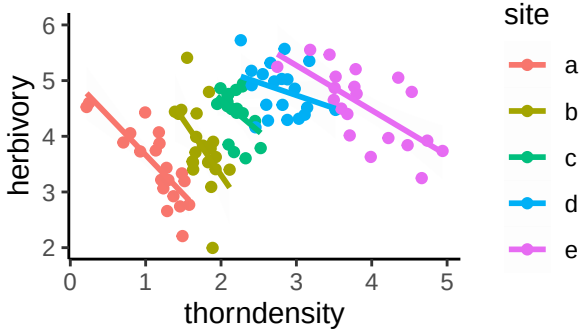
“When I see a scatterplot, I fit a linear regression”

Ecology researchers recorded **density of thorn-like plants** in multiple locations across five regions, and measured **per hectare consumption** of plant material by herbivores.



Herbivory vs Thorns, by Site

Ecology researchers recorded **density of thorn-like plants** in multiple locations across **five sites**, and measured **per hectare consumption** of plant material by herbivores.



Summary

- Message 1: Building a scientific case for a treatment effect is not just about the p-value. Must understand the context of experiment(s).
- Message 2: P-values are measures of evidence. In particular, insufficient evidence (p large) $\neq$ evidence of "no difference".
- Message 3: Interpreting experimental results needs more than t-tests.
- Message 4: We need to incorporate known sources of variation into statistical analyses.
- Message 5: What's more important than p-values and t-tests?
  - recognising patterns in data
  - understanding sources of variation
  - using data to build information about complex systems