Statistical Thinking in Biology Research

An introduction

Terry Neeman

Australian National University

30th July 2020

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

Some history of statistical methods in biology - 20th century

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

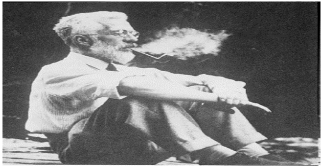


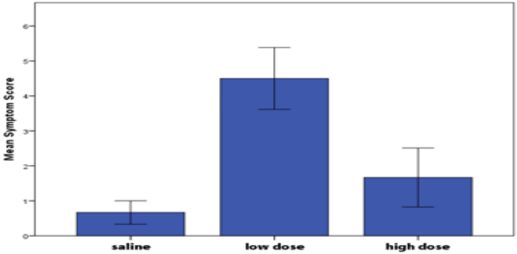
Figure 1: R.A. Fisher 1890 - 1962

Some false narratives ("cautionary tales")

"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

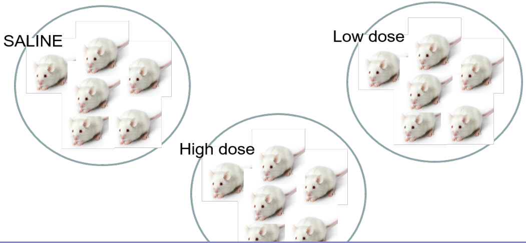


Was there a cage effect or a vaccine effect?

The observed difference in symptom scores could be due to:

- animal cage
- vaccine treatment

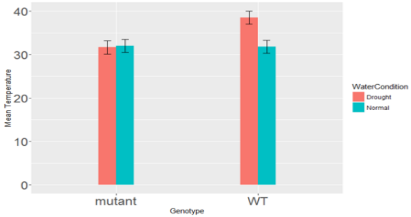
These two factors are **CONFOUNDED**. It is impossible to separate out these two effects.



“$P > 0.05$ means ‘same’; $P < 0.05$ means ‘different’ ”

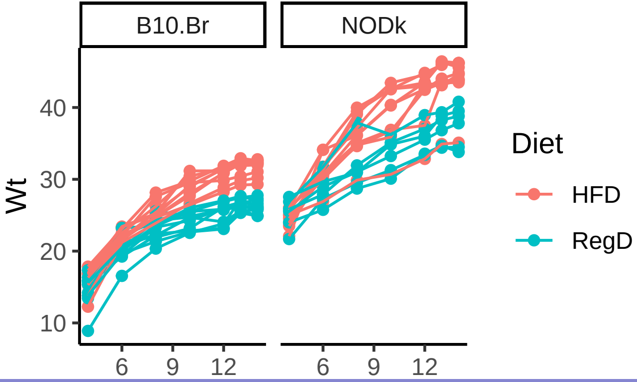
Experimental set-up: Are temperature mechanisms modified in a genetically-engineered tomato plant?

- 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"

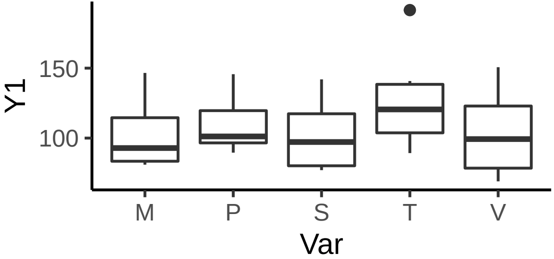
Research questions: Are mice susceptible to obesity when exposed to a high fat diet? Are NODk mice **MORE** susceptible than mice without mutation?



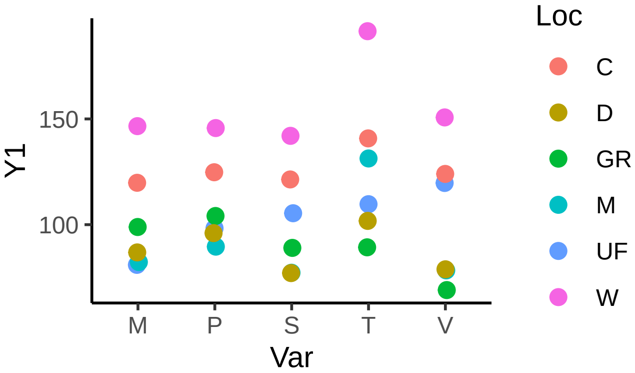
"More than 2 groups? Use 1-way ANOVA"

Research question: Which barley variety has the biggest yield?

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

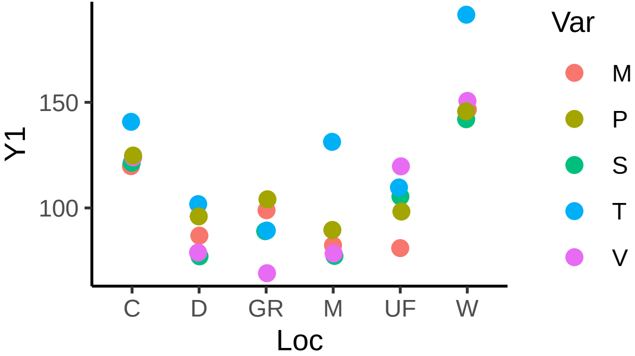


Location contributes to the variation in yield



Yield is highest in Locations C and W. Yield is lowest in Locations D and GR.

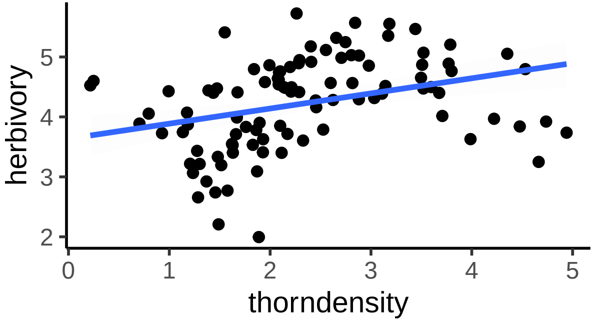
Varieties should be compared *within* location



Notice that: Variety M is near the bottom in most locations
Variety T is near the top in most locations

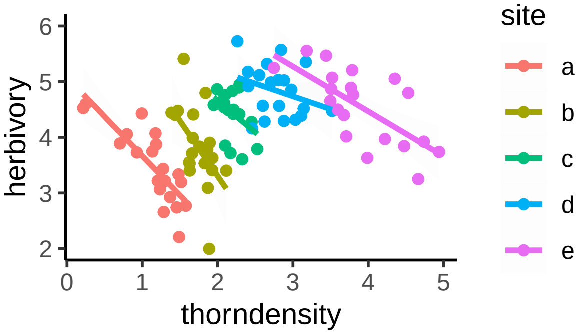
"When I see a scatterplot, I fit a linear regression"

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 from simple contrasts cannot tell us if the contrasts are different.
- 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
  - useing data to build information about complex systems
  - using statistics to allow the data to speak