Statistical Thinking in Biology Research

Statistical Models

Terry Neeman

Australian National University

16th August 2022

A few key ideas for interpreting experimental data

- In an experiment, we compare responses under different conditions.
- The measured response is a mixture of "signal" and "noise".
- Signal: effect of treatments/conditions on response.
- Noise: measurement error, biological, environmental variation
- Statistical models turn data into information.

The goal of statistical modelling is to partition data into "signal" and "noise" or variation.

What is a Statistical Model?

- An informative summary of data
- A representation capturing important patterns
- A description of a data generating process
- A mathematical model that includes measures of uncertainty

One can fit a model to **explain** outcomes.

One can fit a model to **predict** outcomes.

A Statistical Model of an Experiment

Explanatory model

- Statistical model: a conceptualisation of experiment
  - **a measured response** - outcome variable
  - **experimental factors** - how do they influence outcome?
  - nuisance (design) factors - how do they influence outcome?
  - Other variation (unexplained)

Statistical Models: a principled way to learn from data

- data = signal + noise = mean response + variation
- mean response = f(experimental factors)
- variation = g(nuisance factors) + unexplained variation

Experimental design: a principled way to set up experiments to efficiently separate signal and noise.

Understanding statistical models: examples from Lecture 1

Example 1: Shigella vaccine challenge experiment

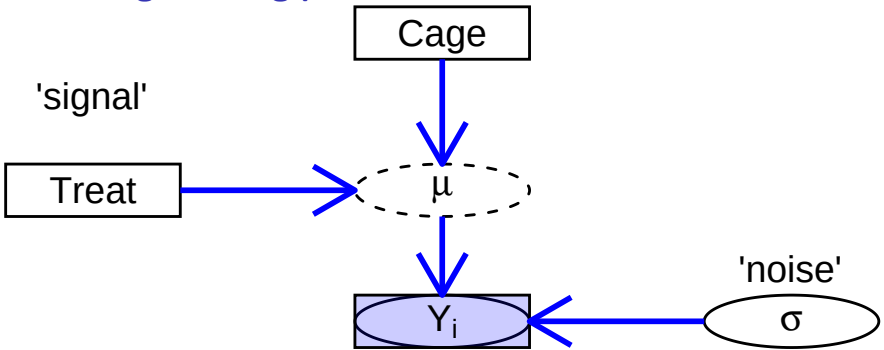
- **Outcome:** 7-day average symptom score post-challenge
- **Experimental factor:** vaccine received (saline/ low dose / high dose)
- All mice challenged with Shigella bacteria at Day 14

This time, 6 mice per cage (two per treatment), and 6 cages total.

- Potential *factors* influencing score: cage (6) and treatment (3)
- Can estimate *cage effects* (score differences between cages)
- Can estimate *treatment effects* **within** each cage

Example 1: Shigella vaccine challenge experiment

Proposed data generating process



Example 1: Shigella vaccine challenge experiment

Separating "signal" and "noise": Analysis of Variance

```
model_vaccine<- lm(Score~Treatment+factor(cageID), data=vaccine)
anova(model_vaccine)
```

```
## Analysis of Variance Table
##
## Response: Score
##                 Df  Sum Sq Mean Sq F value   Pr(>F)
## Treatment        2  5.6932  2.8466  4.3531 0.022587 *
## factor(cageID)   5 18.0115  3.6023  5.5087 0.001183 **
## Residuals       28 18.3100  0.6539
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Example 2: Drought resistance in GM tomato plants

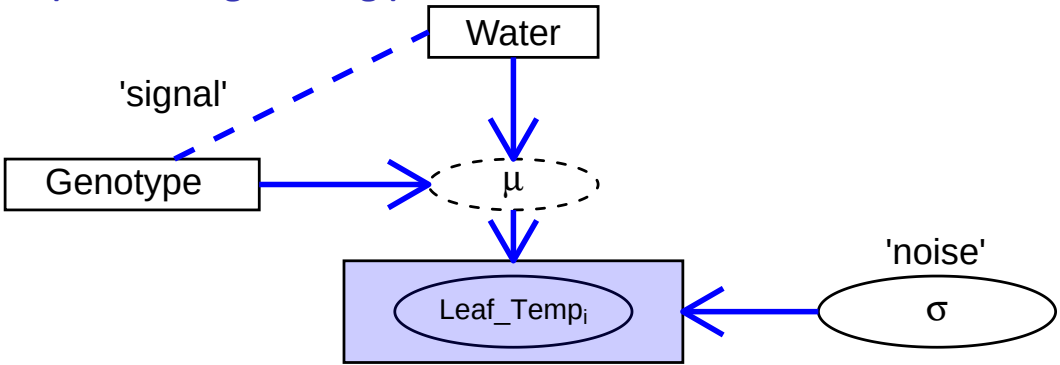
- **Outcome:** leaf temperature at 7 days post-treatment
- **Experimental factors:**
  - Genotype (WT/mutant)
  - Water (normal, drought)

Research question: Does mutation confer drought-resistance?

Equivalent statistical question: Do treatment effects differ by genotype? i.e. is there a treatment by genotype interaction?

Example 2: Drought resistance in GM tomato plants

Proposed data generating process



Example 2: Drought resistance in GM tomato plants

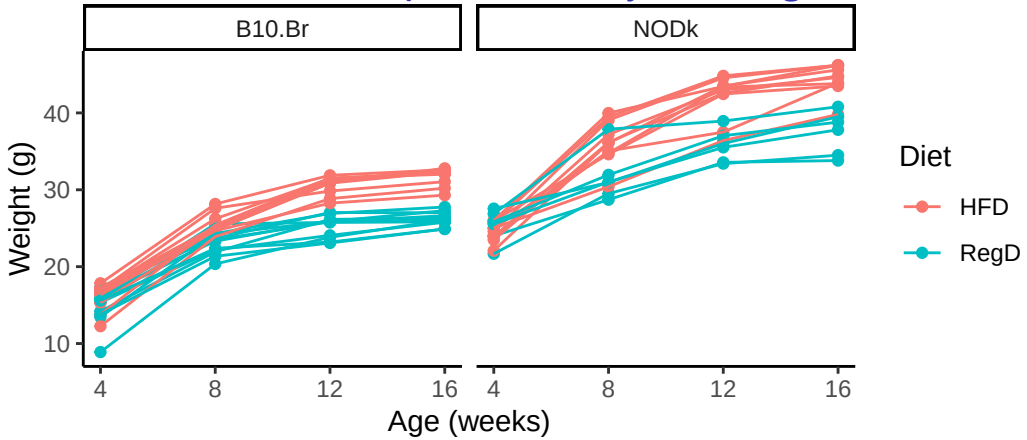
Separating "signal" and "noise": Analysis of Variance

```
model_drought<- lm(Temperature~Genotype*WaterCondition, data=drought)
anova(model_drought)
```

```
## Analysis of Variance Table
##
## Response: Temperature
##                          Df  Sum Sq Mean Sq F value    Pr(>F)
## Genotype                  1  89.111  89.111  33.289 3.407e-06 ***
## WaterCondition            1  80.645  80.645  30.127 7.304e-06 ***
## Genotype:WaterCondition   1 101.531 101.531  37.929 1.195e-06 ***
## Residuals                28  74.953   2.677
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Example 3: Diet and obesity

Are NODk mice more susceptible to obesity with a high fat diet?



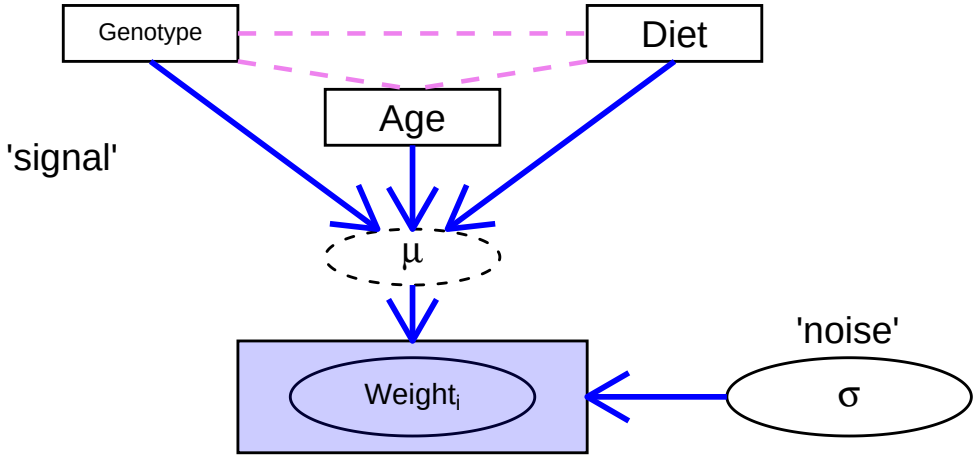
Example 3: Diet and obesity

Are NODk mice more susceptible to obesity with a high fat diet?

- **Outcome:** mouse weight (g)
- **Experimental factors:**
  - Genotype: WT or NODk
  - Diet: normal or high fat
  - Age: measured over time

Does diet impact *growth*? Does diet have stronger impact on growth in NODk mice?

Example 3: Diet and obesity



Example 3: Diet and obesity

```
model_mice<-lmer(Wt~Age*Diet*Strain + (1|MouseID), data=mice)
anova(model_mice)
```

```
## Type III Analysis of Variance Table with Satterthwaite's method
##                 Sum Sq Mean Sq NumDF  DenDF  F value    Pr(>F)
## Age             4169.0  4169.0     1 107.00 592.2165 < 2.2e-16 ***
## Diet               5.7     5.7     1 139.94   0.8027   0.37184
## Strain           377.1   377.1     1 139.94  53.5651 1.783e-11 ***
## Age:Diet         176.4   176.4     1 107.00  25.0595 2.200e-06 ***
## Age:Strain        34.1    34.1     1 107.00   4.8423   0.02992 *
## Diet:Strain        3.6     3.6     1 139.94   0.5116   0.47563
## Age:Diet:Strain    9.0     9.0     1 107.00   1.2785   0.26070
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 ' ' 0.1 ' ' 1
```

Summary

- Statistical models: conceptualisation of experiment
- Fitting a Statistical model to data: estimates of signal and noise
- Signal (pattern): treatment effects and interactions
- Noise: randomly distributed scatter around pattern.
- Fitting Statistical model: make inferences about treatment effects/interactions

In the next workshop, we'll fit models to data using R