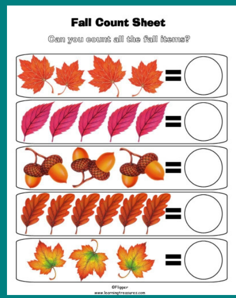
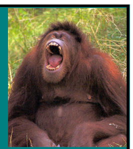
Ecological Data

Some methods for dealing with it

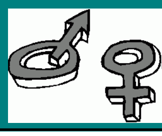
What's special about ecological data?



Types of Data



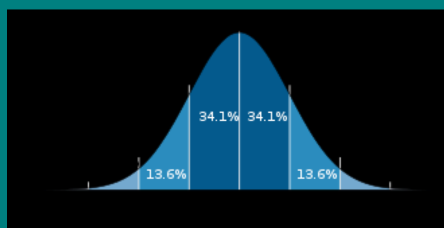
- Categorical data
- Count data
- Continuous data
- Binomial data
- Percentage data
- Time series
- Survival times
- Spatial data



What's special about ecological data?

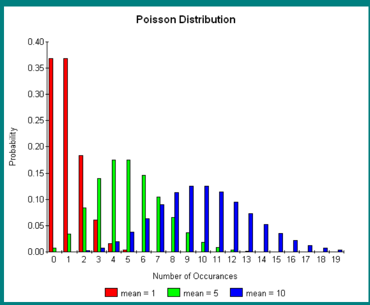
- Often counts rather than continuous data
- Often lots of zeros
- Not 'properly' designed and replicated
  - Random effects / spatial effects
  - Many possible explanatory variables - multiple regression
- Non-linear responses and thresholds
- Often multivariate – and correlated
  - Lots of explanatory (or dependent variables)

Counts



Why is this not a good model for count data?

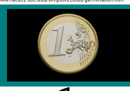
Poisson Distribution

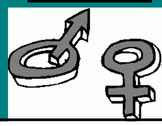


Why don't they look very normal?

What's Bernoulli Data?

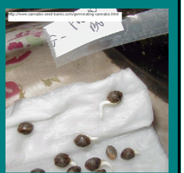
- Each data point is one of two possibilities:
  - Germinated / ungerminated
  - Viable / unviable
  - Dead / alive
  - Male / female
  - Heads / tails
  - Black / white
  - Asleep / awake
  - etc





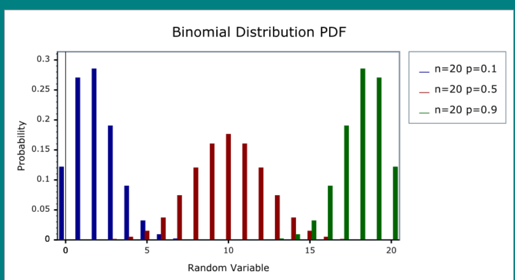
What's Binomial Data?

- Very similar – how many out of how many…
  - 23 out of 50 seeds germinated
  - 12 out of 123 died
  - 78 out of 79 passed the exam
  - 2 out of 3 could do a hand stand
  - 2 out 876 said that they loved statistics



What's different between Binomial Data and Count Data?

Binomial Distribution



Why don't they look very normal?

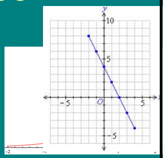
Generalised Linear Model (GLM)

- Model the variability or error with a Poisson or Binomial distribution
- Model the
  - mean number (Poisson)
  - proportion (Binomial)

as a function of the explanatory variables

(A standard linear model is a GLM with normal or Gaussian errors)

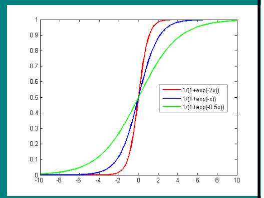
Link Function - Poisson



- Can be a 'standard' linear model
  - Expected count = ax+b
- What's the problem with that?
- Can solve that with a 'log' link function
  - Expected count = exp(ax+b) = e^{ax+b}
- This is the R default
- But can create it's own problems
  - Huge predictions
- Using transformed x variable can help

Link Function - Binomial

- Proportion: not just positive, between 0 and 1
- Default is logistic
- Can be probit or cauchit
- These are more extreme
  - Cauchit → 0 and 1 slowly
  - Probit → 0 and 1 fast!

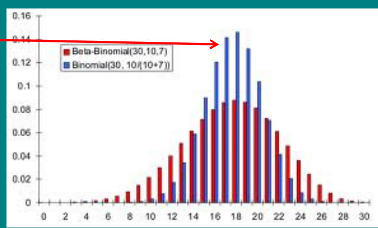


- But doesn't make a lot of difference unless you really care about the extremes

Break!

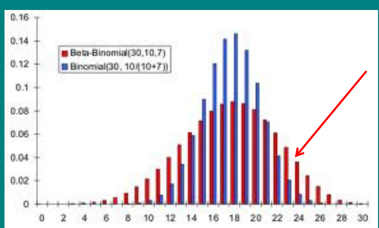
Over-dispersion

- 30 seeds in a Petri dish – 10/17 = 59% chance germinating - how many actually germinate?
  - If exactly 59% chance



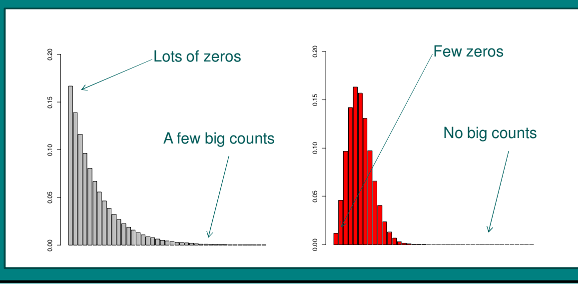
Over-dispersion

- Many Petri dishes – 30 seeds in each - how many germinate?
  - Still 59% on average, but why more variability?



Over-dispersion in Poisson Model

- Same mean (5) but different variability



Over-dispersion

- Doesn't affect means of fitted model
- But affects confidence in the mean, and thus tests of significance
- Must account for over-dispersion (if present)

(a bit like accounting for non-constant variance in linear models)

GLM in R

```
fm <- glm(y~x,family=binomial,data=myData)
fm <- glm(y~x,family=poisson)
fm <- glm(y~x*tr,family=binomial)
fm <- glm(y~(x1+x2+x3+tr)^2,family=poisson)
```

With poisson, y is just the counts

With binomial, y is the number of 'successes' and the number of 'failures' eg.

y <- cbind(n.germ,n.not.germ)

GLM in R

- Similar output to lm
- Can do an 'anova' on model to get significance of terms (actually an analysis of deviance)

```
Null deviance: 451.824  on 23  degrees of freedom
Residual deviance:  33.036  on 22  degrees of freedom
AIC: 103.07
```

- AIC general measure of model fit – can be used to compare between models (lower is better!)

GLM in R – check for overdispersion

```
Null deviance: 451.824  on 23  degrees of freedom
Residual deviance:  33.036  on 22  degrees of freedom
AIC: 103.07
```

- Residual deviance much greater than residual degrees of freedom → overdispersion

```
Null deviance: 451.824  on 23  degrees of freedom
Residual deviance:  330.36  on 22  degrees of freedom
AIC: 103.07
```

GLM in R – accounting for overdispersion

```
fm <- glm(y~x,family=quasipoisson)
fm <- glm(y~x*tr,family=quasibinomial)
```

It will tell you the estimated dispersion parameter

If there is over-dispersion, this will be much bigger than 1.

GLM in R – changing link function

```
fm <- glm(y~x,family=poisson(link='identity'))
fm <- glm(y~x*tr,family=binomial(link='probit'))
```

Other GLMs for Counts

- Beta-binomial and negative-binomial models alternative ways of dealing with over-dispersion
- Zero-inflated Poisson models (ZIPs) good when you have lots of zeros
  – Model the probability of zero separately

More to counting than you expected!

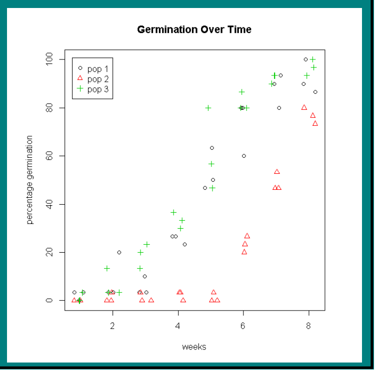
- All covered briefly in lab
- But in more detail in relevant R books
- R Book Ch 13-17
- Lots of practice!

Poisson GLM ‘regression’ example

- Effect of time since fire on abundance

Binomial GLM ‘ANCOVA’ example

- Seeds from three populations
- Germination over time for each population



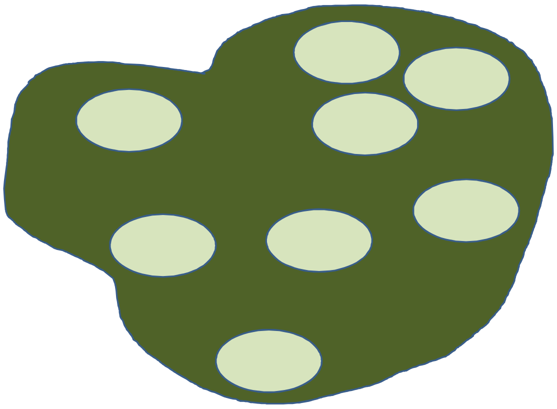
Break!

Some different designs

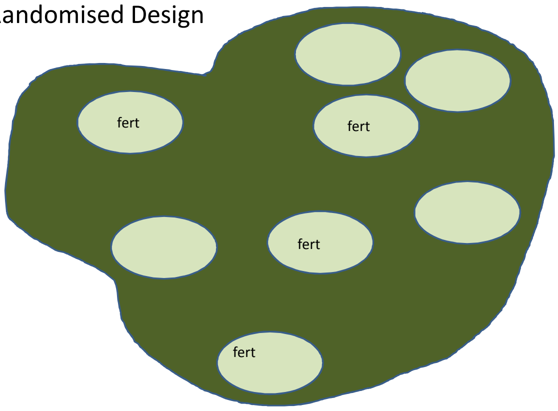
And how to analyse them

Effects of different treatments on species diversity

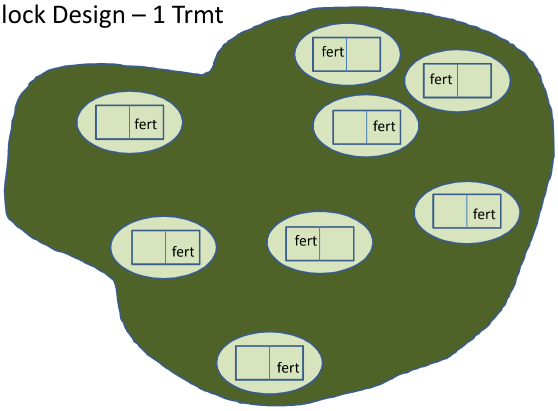
- Grazing (fencing)
- Fertilisation



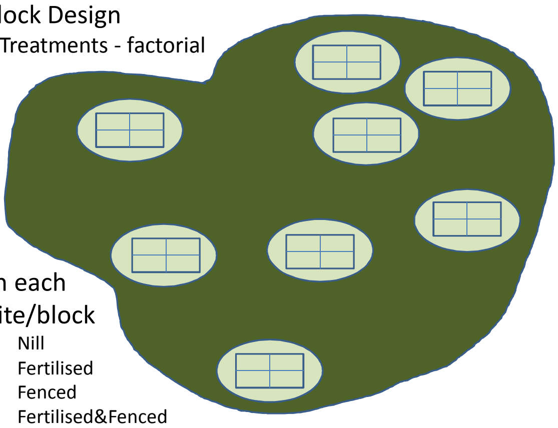
Randomised Design



Block Design – 1 Trmt



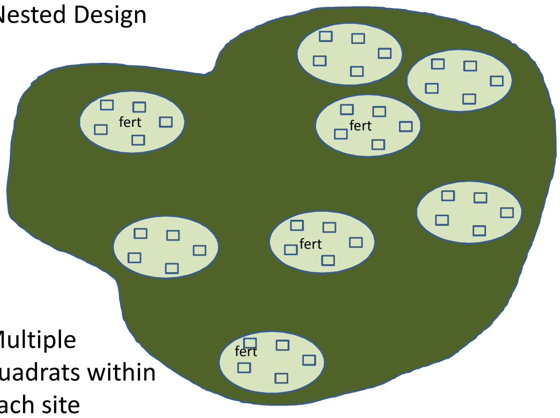
Block Design
2 Treatments - factorial



In each site/block

- Nill
- Fertilised
- Fenced
- Fertilised&Fenced

Nested Design



Multiple quadrats within each site

Nested design

- So 8x5=40 measurements for each treatment
- Looks like lots of replicates
- But they aren't independent
- So to treat them as independent replicates is 'pseudoreplication'
- Which is bad!

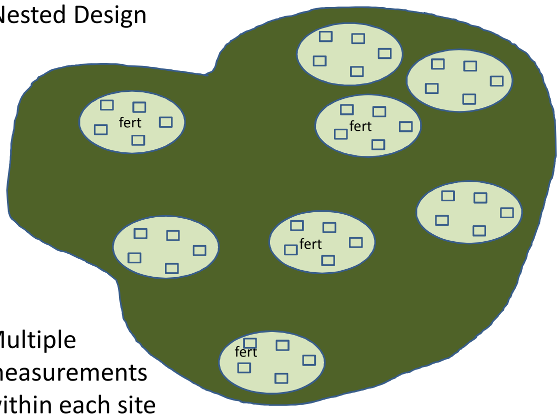
Nested design

- You have repeated measures (in space) on the same unit (…site, animal, tree etc)
- Could also be repeated measures on the same unit over time
- Or different depths in the soil (microbial diversity)
- 'Repeated measures' or 'hierarchical design'
- But treating them as independent replicates is always bad!

Random vs Fixed Effects

- Fixed effects – labels are meaningful, you wouldn't switch them around
  - Treatments, species, sex
- Random effects – labels are not meaningful, you could easily switch them around
  - Plot number, quadrat number, block number

Nested Design

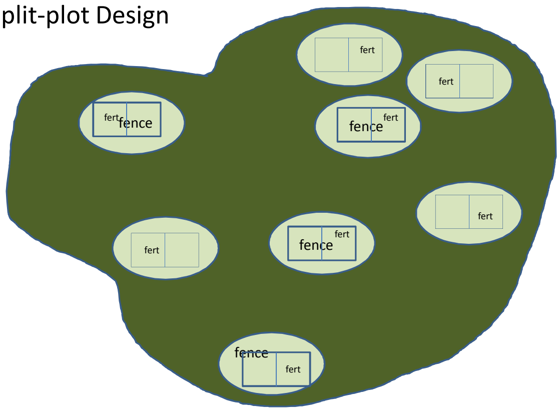


Multiple measurements within each site

Random vs Fixed Effects ??

- Site number?
- Fertilised / unfertilised ?
- Nested effect
  - Quadrat number?
  - Depth?
  - Time?

Split-plot Design



More complicated!

- Eight sites (blocks) – random
- Each divided into two plots – one fenced, the other not
- Each plot divided into three subplots – one high fertiliser, one low, one not
- Each subplot sampled at five times
- At each time – two random quadrats sampled
- quadrats/times/subplots/plots/blocks
- Treatments at the subplot and plot level

How to analyse

- Various options
- lm and t.tests will sometimes be ok
- 'Error' terms in 'aov' function ok in more cases
- 'lme' function from 'nlme' package is more robust
  - Eg. works when values are missing
- 'lmer' function from 'lme4' package is more robust still
  - Eg. lets you do glmers

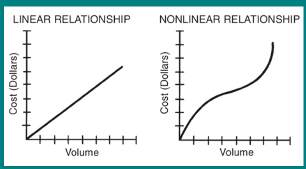
Break!

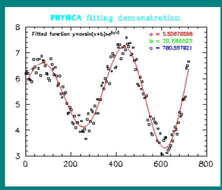
What have we done?

- Week 1 – General intro to R and basic classical stats tests
- Week 2 – General linear model (lm)
- Week 3 – checking assumptions of general linear model, and fixing problems
- Week 4: extensions of lm
  - Generalised linear models (GLMs)
  - Linear mixed effects models

What's next?

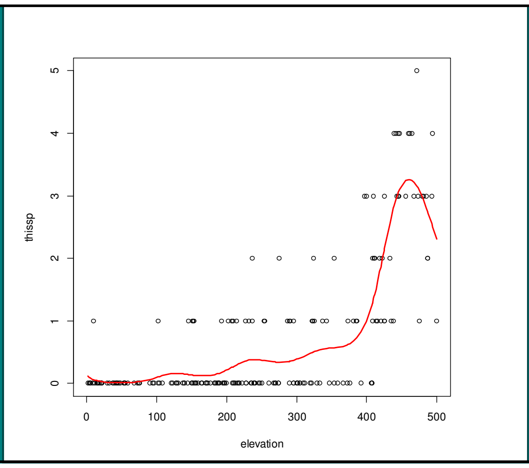
Non-linear responses





Non-linear responses

- GLM accounts for certain types of simple non-linearity – logistic, log, exp, inverse...
- Transformations and polynomials can help deal with some types of non-linearity (see Day 3)
- Generalised Additive Models allow very flexible modelling of non-linearity (look up online and see lab example and R Book Ch 18)



Non-linear responses

- Non-linear regression using R nls for non-linear response curves where you know what the form should be
  - Michaelis-Menten
  - Photosynthesis curves
  - Bounded growth curves
  - Etc etc
- See help(nls) in R, R Book Ch 20

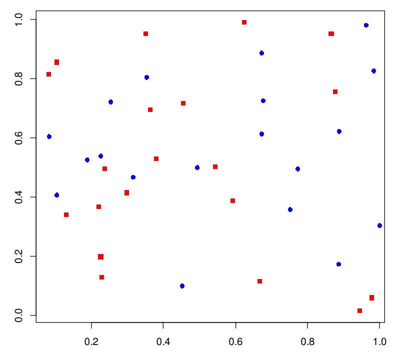
Mixed effects

- Random effects – mixed effects models (Day 5, R Book Ch 19)
- GLMMERs

Survival Analysis

- How long things survived (when they died)
- GLM with Gamma distribution errors
- Other methods if not all have died yet or you don't know about some (censoring of data)
- See R Book Ch 25

Spatial Analysis



Spatial Analysis

- See R Book Ch 24 to get started

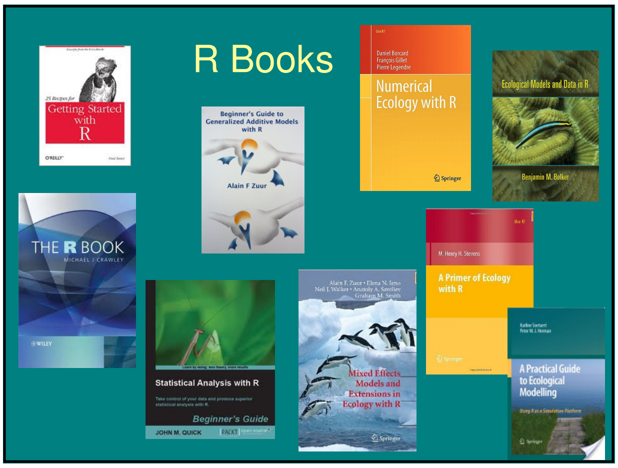
Multivariate

- When lots of predictors:
  - Multiple regression (R Book Ch 10)
  - Regression trees (R Book Ch 21)
- When lots of dependent variables:
  - Multivariate techniques
  - Very strong in ecology
  - R Book Ch 23, Day 5
- When you're not sure:
  - Structural equation modelling (SEM)

Time series

- R Book Chapter 22

R Books



Assessment

- Quiz (5%)
- Assignment Report (15%)

Day 5

- More detail on mixed effects models
- Multivariate analysis for ecologists