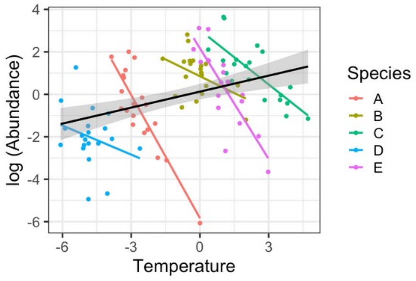
Job trends in statistics and data analysis

With so much hinging on statistics - national and international policy, funding, corporate decision making, government commitment and research - **demand for statisticians and data analysts is expected to grow around 34% between 2014 and 2024**. This vastly outstrips the average across all jobs.

source - https://www.environmentalscience.org/career/environmental-data-analyst

2

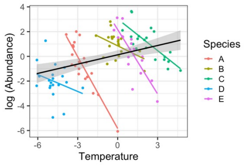
One common feature here is that species may differ in the way they are structured by environmental and/or trait variation. This can be well described by the Simpson's paradox (Simpson 1951), which is defined "as a phenomenon in probability and statistics in which a trend appears in several groups of data but disappears or reverses when the groups are combined."

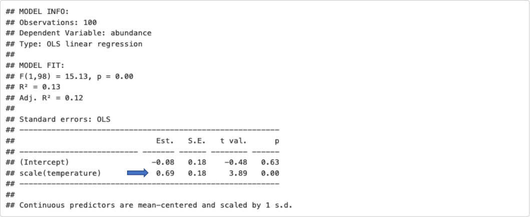


Important enough to have its own Wikipedia page:
https://en.wikipedia.org/wiki/Simpson%27s_paradox

3

Fixed effect model

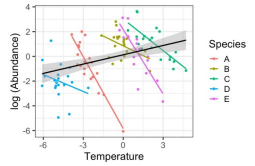


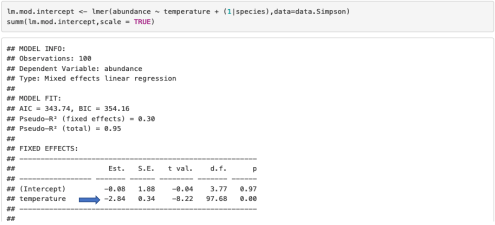


```
## MODEL INFO:
## Observations: 100
## Dependent Variable: abundance
## Type: OLS linear regression
##
## MODEL FIT:
## F(1,98) = 15.13, p = 0.00
## R² = 0.13
## Adj. R² = 0.12
##
## Standard errors: OLS
## ---------------------------------------------------------
##                              Est.   S.E.   t val.      p
## -------------------------- ------- ------ -------- ------
## (Intercept)                  -0.08   0.18    -0.48   0.63
## scale(temperature)      ➡    0.69   0.18     3.89   0.00
## ---------------------------------------------------------
##
## Continuous predictors are mean-centered and scaled by 1 s.d.
```

4

Mixed effect model





```
lm.mod.intercept <- lmer(abundance ~ temperature + (1|species),data=data.Simpson)
summ(lm.mod.intercept,scale = TRUE)
```

```
## MODEL INFO:
## Observations: 100
## Dependent Variable: abundance
## Type: Mixed effects linear regression
##
## MODEL FIT:
## AIC = 343.74, BIC = 354.16
## Pseudo-R² (fixed effects) = 0.30
## Pseudo-R² (total) = 0.95
##
## FIXED EFFECTS:
## --------------------------------------------------------
##                         Est.    S.E.   t val.    d.f.      p
## ----------------- -------- ------- -------- -------- ------
## (Intercept)             -0.08    1.88    -0.04    3.77   0.97
## temperature     ➡       -2.84    0.34    -8.22   97.68   0.00
## --------------------------------------------------------
##
```

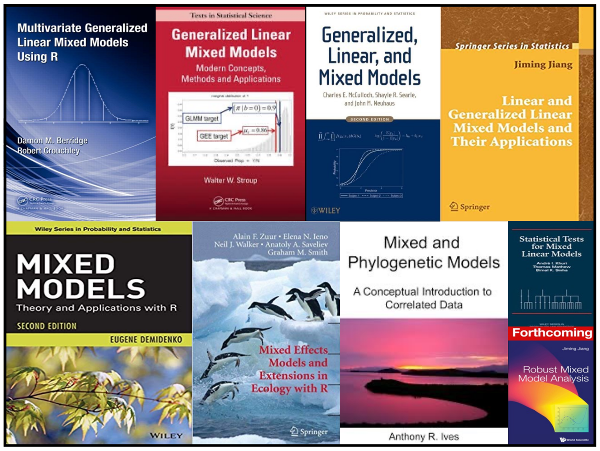
5

General linear models (not Generalized linear model)

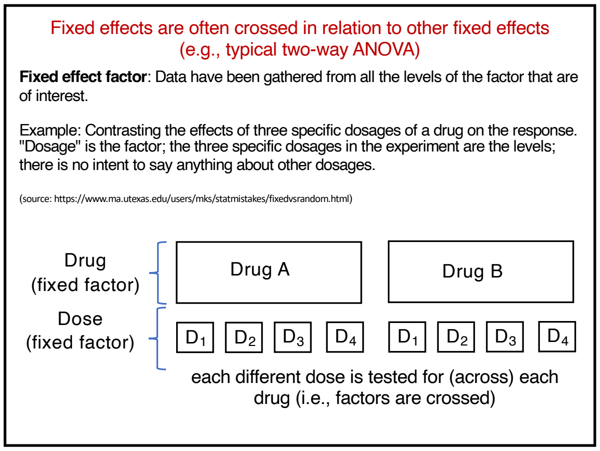
	Linear Model	Common name
✓	$Y = \mu + X$	Simple linear regression
✓	$Y = \mu + A_1$	One-factorial (one-way) ANOVA
✓	$Y = \mu + A_1 + A_2 + A_1 \times A_2$	Two-factorial (two-way) ANOVA
✓	$Y = \mu + A_1 + X\ (+A_1 \times X)$	Analysis of Covariance (ANCOVA)
✓	$Y = \mu + X_1 + X_2 + X_3$	Multiple regression
⇨	$Y = \mu + A_1 + g + A_1 \times g$	Mixed model ANOVA
	$Y_1 + Y_2 = \mu + A_1 + A_2 + A_1 \times A_2$	Multivariate ANOVA (MANOVA)

Y (response) is a continuous variable
X (predictor) is a continuous variable
A represents categorical predictors (factors)
g represents groups of data (more on this later)
$(+A_1 \times X)$ - step 1 on an ANCOVA, but not in the final analysis
Multiple factors $A_1 + A_2 +$ etc (and their interactions)

6


Multivariate Generalized Linear Mixed Models Using R
Damon M. Berridge
Robert Crouchley
Texts in Statistical Science
Generalized Linear Mixed Models
Modern Concepts, Methods and Applications
Walter W. Stroup
CRC Press
Generalized, Linear, and Mixed Models
Charles E. McCulloch, Shayle R. Searle, and John M. Neuhaus
SECOND EDITION
WILEY
Springer Series in Statistics
Jiming Jiang
Linear and Generalized Linear Mixed Models and Their Applications
Springer
Wiley Series in Probability and Statistics
MIXED MODELS
Theory and Applications with R
SECOND EDITION
EUGENE DEMIDENKO
WILEY
Alain F. Zuur • Elena N. Ieno
Neil J. Walker • Anatoly A. Saveliev
Graham M. Smith
Mixed Effects Models and Extensions in Ecology with R
Springer
Mixed and Phylogenetic Models
A Conceptual Introduction to Correlated Data
Anthony R. Ives
Statistical Tests for Mixed Linear Models
Forthcoming
Robust Mixed Model Analysis

7



Fixed effects are often crossed in relation to other fixed effects (e.g., typical two-way ANOVA)

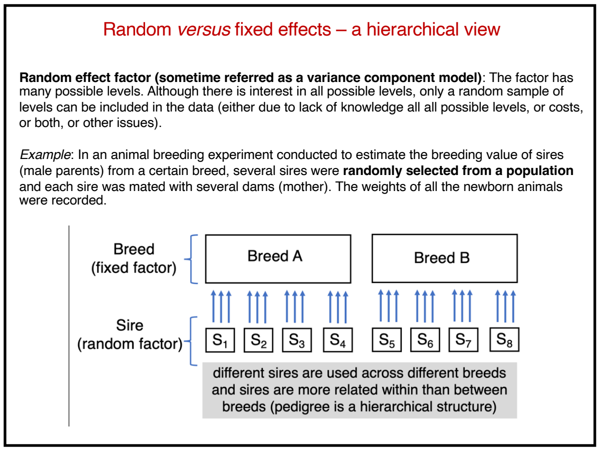
Fixed effect factor: Data have been gathered from all the levels of the factor that are of interest.

Example: Contrasting the effects of three specific dosages of a drug on the response. "Dosage" is the factor; the three specific dosages in the experiment are the levels; there is no intent to say anything about other dosages.

(source: https://www.ma.utexas.edu/users/mks/statmistakes/fixedvsrandom.html)

each different dose is tested for (across) each drug (i.e., factors are crossed)

8



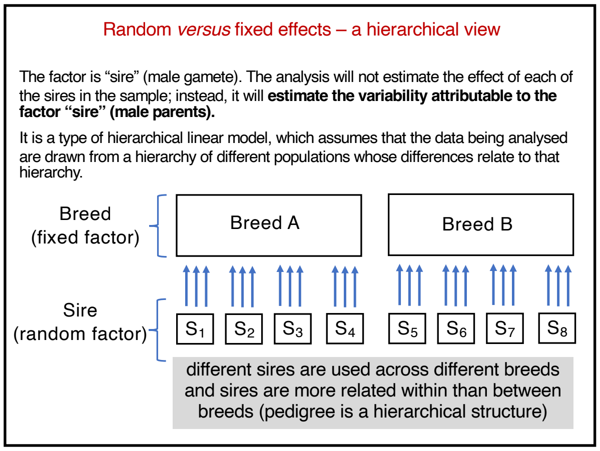
Random *versus* fixed effects – a hierarchical view

Random effect factor (sometime referred as a variance component model): The factor has many possible levels. Although there is interest in all possible levels, only a random sample of levels can be included in the data (either due to lack of knowledge all all possible levels, or costs, or both, or other issues).

Example: In an animal breeding experiment conducted to estimate the breeding value of sires (male parents) from a certain breed, several sires were **randomly selected from a population** and each sire was mated with several dams (mother). The weights of all the newborn animals were recorded.

different sires are used across different breeds and sires are more related within than between breeds (pedigree is a hierarchical structure)

9



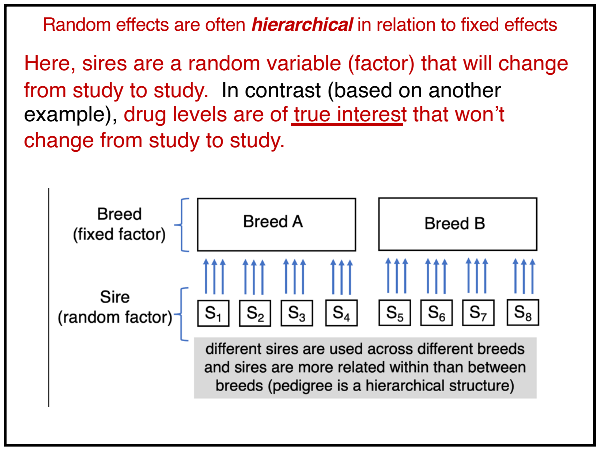
Random *versus* fixed effects – a hierarchical view

The factor is "sire" (male gamete). The analysis will not estimate the effect of each of the sires in the sample; instead, it will **estimate the variability attributable to the factor "sire" (male parents).**

It is a type of hierarchical linear model, which assumes that the data being analysed are drawn from a hierarchy of different populations whose differences relate to that hierarchy.

different sires are used across different breeds and sires are more related within than between breeds (pedigree is a hierarchical structure)

10

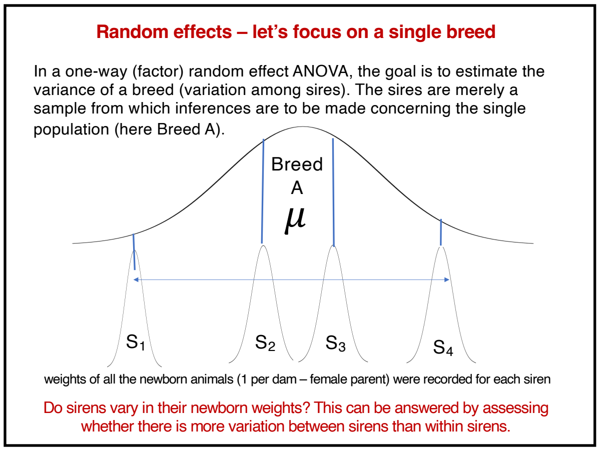


Random effects are often ***hierarchical*** in relation to fixed effects

Here, sires are a random variable (factor) that will change from study to study. In contrast (based on another example), drug levels are of true interest that won't change from study to study.

different sires are used across different breeds and sires are more related within than between breeds (pedigree is a hierarchical structure)

11



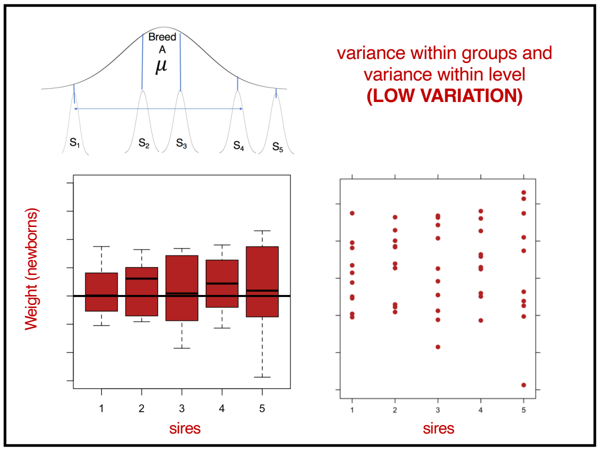
Random effects – let's focus on a single breed

In a one-way (factor) random effect ANOVA, the goal is to estimate the variance of a breed (variation among sires). The sires are merely a sample from which inferences are to be made concerning the single population (here Breed A).

weights of all the newborn animals (1 per dam – female parent) were recorded for each siren

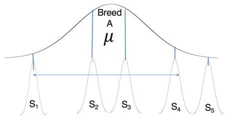
Do sirens vary in their newborn weights? This can be answered by assessing whether there is more variation between sirens than within sirens.

12

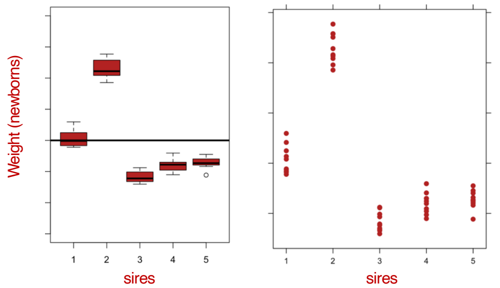


variance within groups and variance within level
(LOW VARIATION)

13



variance within groups and
variance within level
(HIGH VARIATION)



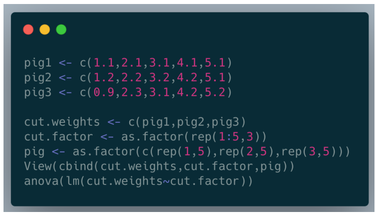
14

Fixed versus random effects may depend on the question and not always the data

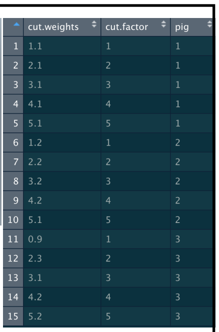
STUDY: Suppose five cuts of meat are taken from each of three pigs, all from the same breed, and the fat content is measured in each cut.

FIXED EFFECT QUESTION – Do the different cuts differ in their fat content? One-way (fixed) ANOVA with five treatment levels (cuts) and three replicates (observations) per cut (pigs).

15

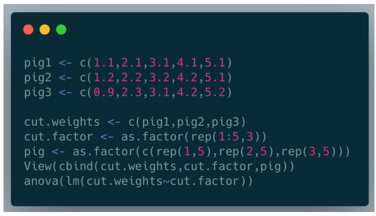


```
pig1 <- c(1.1,2.1,3.1,4.1,5.1)
pig2 <- c(1.2,2.2,3.2,4.2,5.1)
pig3 <- c(0.9,2.3,3.1,4.2,5.2)

cut.weights <- c(pig1,pig2,pig3)
cut.factor <- as.factor(rep(1:5,3))
pig <- as.factor(c(rep(1,5),rep(2,5),rep(3,5)))
View(cbind(cut.weights,cut.factor,pig))
anova(lm(cut.weights~cut.factor))
```

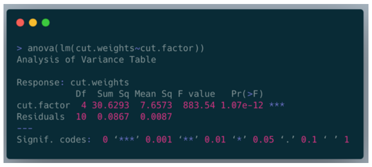


	cut.weights	cut.factor	pig
1	1.1	1	1
2	2.1	2	1
3	3.1	3	1
4	4.1	4	1
5	5.1	5	1
6	1.2	1	2
7	2.2	2	2
8	3.2	3	2
9	4.2	4	2
10	5.1	5	2
11	0.9	1	3
12	2.3	2	3
13	3.1	3	3
14	4.2	4	3
15	5.2	5	3

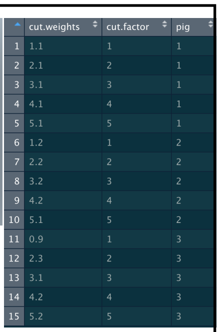
16



```
pig1 <- c(1.1,2.1,3.1,4.1,5.1)
pig2 <- c(1.2,2.2,3.2,4.2,5.1)
pig3 <- c(0.9,2.3,3.1,4.2,5.2)

cut.weights <- c(pig1,pig2,pig3)
cut.factor <- as.factor(rep(1:5,3))
pig <- as.factor(c(rep(1,5),rep(2,5),rep(3,5)))
View(cbind(cut.weights,cut.factor,pig))
anova(lm(cut.weights~cut.factor))
```



```
> anova(lm(cut.weights~cut.factor))
Analysis of Variance Table

Response: cut.weights
           Df  Sum Sq Mean Sq F value   Pr(>F)
cut.factor  4 30.6293  7.6573  883.54 1.07e-12 ***
Residuals  10  0.0867  0.0087
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



	cut.weights	cut.factor	pig
1	1.1	1	1
2	2.1	2	1
3	3.1	3	1
4	4.1	4	1
5	5.1	5	1
6	1.2	1	2
7	2.2	2	2
8	3.2	3	2
9	4.2	4	2
10	5.1	5	2
11	0.9	1	3
12	2.3	2	3
13	3.1	3	3
14	4.2	4	3
15	5.2	5	3

17

Fixed versus random effects may depend on the question and not always the data

STUDY: Suppose five cuts of meat are taken from each of three pigs, all from the same breed, and the fat content is measured in each cut.

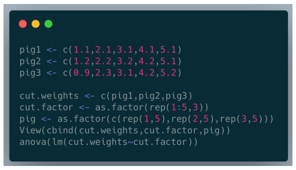
FIXED EFFECT QUESTION – Do the different cuts differ in their fat content? One-way (fixed) ANOVA with five treatment levels (cuts) and three replicates per cut (pigs).

RANDOM EFFECT QUESTION - Is there more variation in fat content among or within pigs (i.e., animal-to-animal and within-animal variation)? A fat pig could have their cuts fatter (i.e., hierarchical variation).

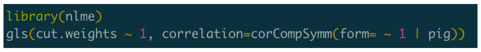
In this case, the three pigs selected are not of interest. This would be a one-way random effects ANOVA.

18

Note that there is more variation within pigs (i.e., among cuts within pigs) than between pigs



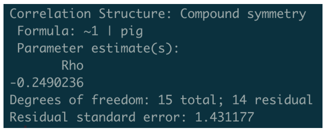
```
pig1 <- c(1.1,2.1,3.1,4.1,5.1)
pig2 <- c(1.2,2.2,3.2,4.2,5.1)
pig3 <- c(0.9,2.3,3.1,4.2,5.2)

cut.weights <- c(pig1,pig2,pig3)
cut.factor <- as.factor(rep(1:5,3))
pig <- as.factor(c(rep(1,5),rep(2,5),rep(3,5)))
View(cbind(cut.weights,cut.factor,pig))
anova(lm(cut.weights~cut.factor))
```



```
library(nlme)
gls(cut.weights ~ 1, correlation=corCompSymm(form= ~ 1 | pig))
```

Rho refers to intraclass (here pig) correlation

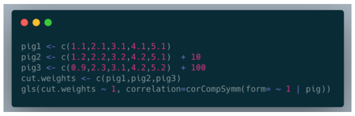
A negative or low ICC suggests that there is more variability within subjects (here pigs) than between subjects.



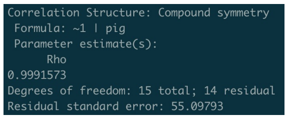
```
Correlation Structure: Compound symmetry
 Formula: ~1 | pig
 Parameter estimate(s):
       Rho
-0.2490236
Degrees of freedom: 15 total; 14 residual
Residual standard error: 1.431177
```

19

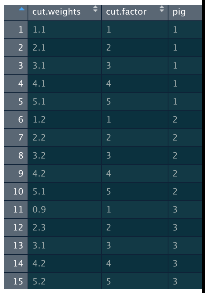
Now, these data have more variation between pigs (i.e., among cuts within pigs) than within pigs



```
pig1 <- c(1.1,2.1,3.1,4.1,5.1)
pig2 <- c(1.2,2.2,3.2,4.2,5.1)  + 10
pig3 <- c(0.9,2.3,3.1,4.2,5.2)  + 100
cut.weights <- c(pig1,pig2,pig3)
gls(cut.weights ~ 1, correlation=corCompSymm(form= ~ 1 | pig))
```



```
Correlation Structure: Compound symmetry
 Formula: ~1 | pig
 Parameter estimate(s):
      Rho
0.9991573
Degrees of freedom: 15 total; 14 residual
Residual standard error: 55.09793
```



	cut.weights	cut.factor	pig
1	1.1	1	1
2	2.1	2	1
3	3.1	3	1
4	4.1	4	1
5	5.1	5	1
6	1.2	1	2
7	2.2	2	2
8	3.2	3	2
9	4.2	4	2
10	5.1	5	2
11	0.9	1	3
12	2.3	2	3
13	3.1	3	3
14	4.2	4	3
15	5.2	5	3

Rho refers to intraclass (here pig) correlation

A high positive ICC suggests that there is less variability within subjects (here pigs) than between subjects.

20

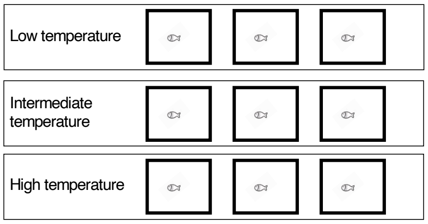
More on the hierarchical nature of data

21

The advantages of mixed models - increase statistical power and estimation accuracy through dependent replication and design convenience (particularly in observational studies).

Do we need a random effect here?

Effects of temperature on fish growth
(difference in growth begin/end of study)

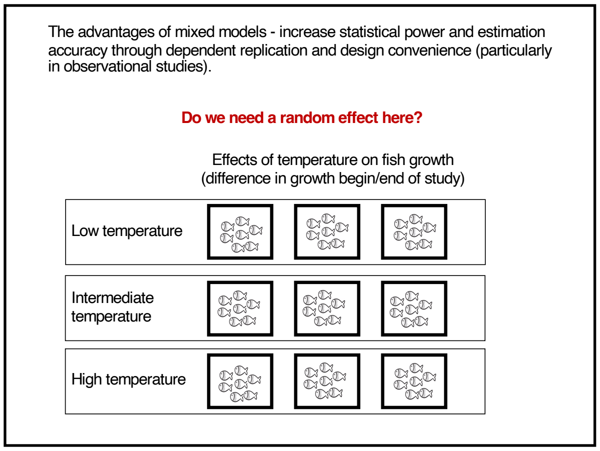


Low temperature

Intermediate temperature

High temperature

22

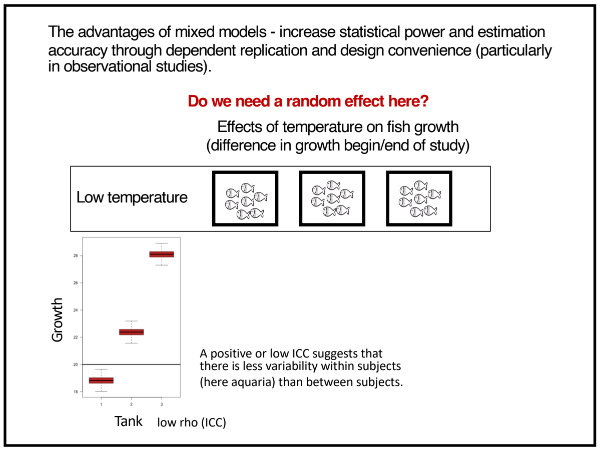


The advantages of mixed models - increase statistical power and estimation accuracy through dependent replication and design convenience (particularly in observational studies).

Do we need a random effect here?

Effects of temperature on fish growth
(difference in growth begin/end of study)

23



The advantages of mixed models - increase statistical power and estimation accuracy through dependent replication and design convenience (particularly in observational studies).

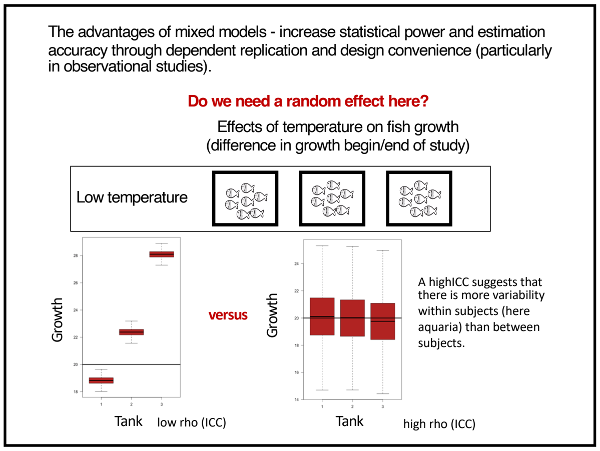
Do we need a random effect here?

Effects of temperature on fish growth
(difference in growth begin/end of study)

A positive or low ICC suggests that there is less variability within subjects (here aquaria) than between subjects.

low rho (ICC)

24



The advantages of mixed models - increase statistical power and estimation accuracy through dependent replication and design convenience (particularly in observational studies).

Do we need a random effect here?

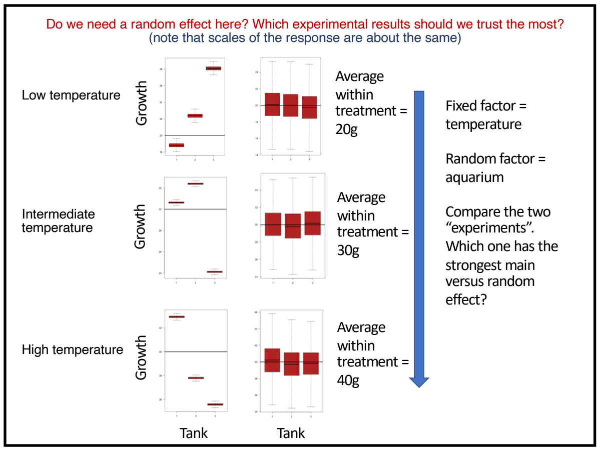
Effects of temperature on fish growth
(difference in growth begin/end of study)

low rho (ICC)

A highICC suggests that there is more variability within subjects (here aquaria) than between subjects.

high rho (ICC)

25



Do we need a random effect here? Which experimental results should we trust the most?
(note that scales of the response are about the same)

Low temperature — Growth — Average within treatment = 20g

Intermediate temperature — Growth — Average within treatment = 30g

High temperature — Growth — Average within treatment = 40g

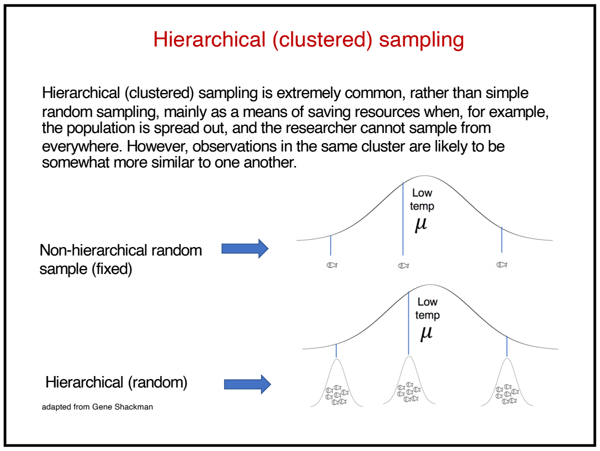
Tank | Tank

Fixed factor = temperature

Random factor = aquarium

Compare the two "experiments". Which one has the strongest main versus random effect?

26



Hierarchical (clustered) sampling

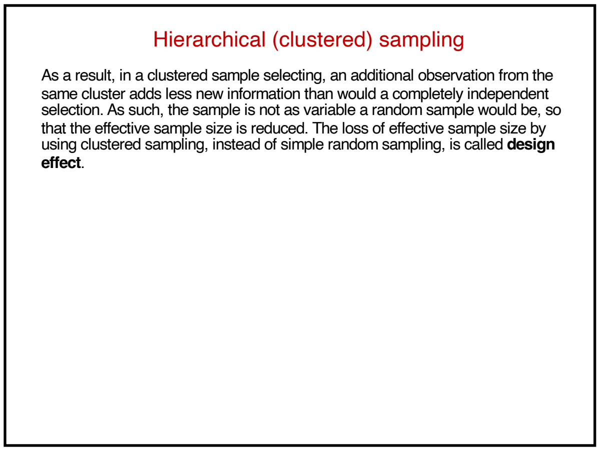
Hierarchical (clustered) sampling is extremely common, rather than simple random sampling, mainly as a means of saving resources when, for example, the population is spread out, and the researcher cannot sample from everywhere. However, observations in the same cluster are likely to be somewhat more similar to one another.

Non-hierarchical random sample (fixed)

Hierarchical (random)

adapted from Gene Shackman

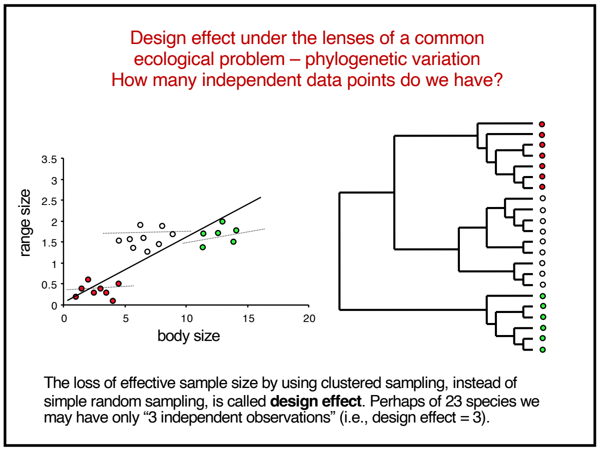
27



Hierarchical (clustered) sampling

As a result, in a clustered sample selecting, an additional observation from the same cluster adds less new information than would a completely independent selection. As such, the sample is not as variable a random sample would be, so that the effective sample size is reduced. The loss of effective sample size by using clustered sampling, instead of simple random sampling, is called **design effect**.

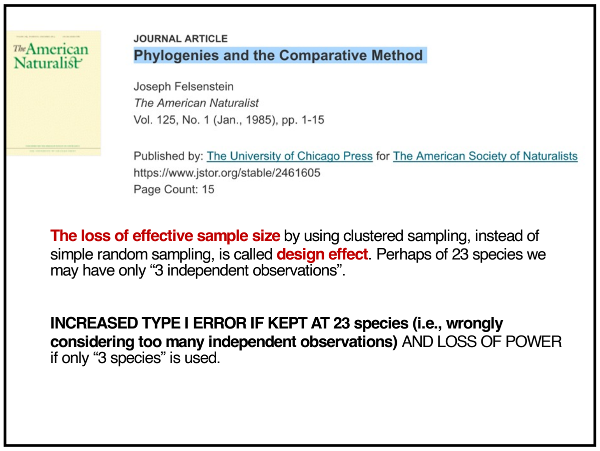
28



Design effect under the lenses of a common ecological problem – phylogenetic variation
How many independent data points do we have?

The loss of effective sample size by using clustered sampling, instead of simple random sampling, is called **design effect**. Perhaps of 23 species we may have only “3 independent observations” (i.e., design effect = 3).

29



The American Naturalist

JOURNAL ARTICLE
Phylogenies and the Comparative Method

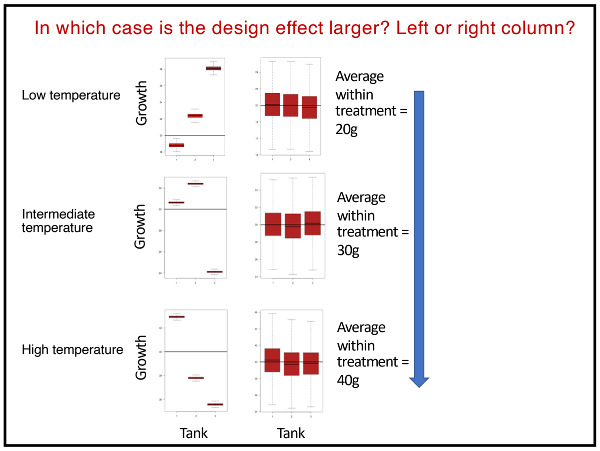
Joseph Felsenstein
The American Naturalist
Vol. 125, No. 1 (Jan., 1985), pp. 1-15

Published by: The University of Chicago Press for The American Society of Naturalists
https://www.jstor.org/stable/2461605
Page Count: 15

The loss of effective sample size by using clustered sampling, instead of simple random sampling, is called **design effect**. Perhaps of 23 species we may have only “3 independent observations”.

INCREASED TYPE I ERROR IF KEPT AT 23 species (i.e., wrongly considering too many independent observations) AND LOSS OF POWER if only “3 species” is used.

30



In which case is the design effect larger? Left or right column?

Low temperature

Average within treatment = 20g

Intermediate temperature

Average within treatment = 30g

High temperature

Average within treatment = 40g

31

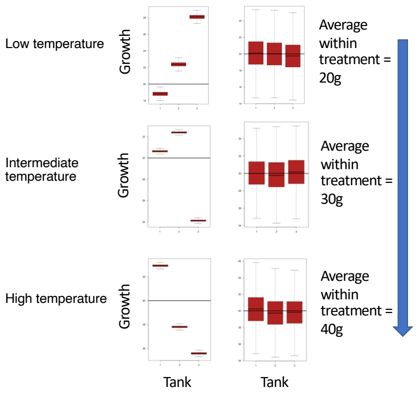
Hierarchical (clustered) sampling

The **design effect** is a correction factor that is used to adjust the sample size based on clustered sampling. This accounts for the loss of information inherent in the clustered design and is used when estimating random effects.

Once the design effect is calculated, the sample size calculated for a standard design can be adjusted accordingly (i.e., degrees of freedom are corrected). As such, the statistical power may change according to the design effect.

32

In which case is the statistical power greater? Left or right column?



33

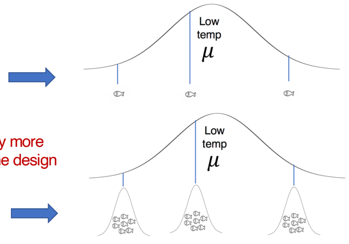
Hierarchical (clustered) sampling

Hierarchical (clustered) sampling is extremely common, rather than simple random sampling, mainly as a means of saving resources when, for example, the population is spread out, and the researcher cannot sample from everywhere. However, observations in the same cluster are likely to be somewhat more similar to one another, decreasing effect size.

Non-hierarchical random sample (fixed)

Which design is statistically more powerful? It depends on the design effect of the data.

Hierarchical (random)



adapted from Gene Shackman

34

Intraclass correlation as a way to understand why to use a random effect and how much "random" has an "effect"

35

Hierarchical (clustered) sampling – intraclass correlation & design effect (one extreme example)

- Consider a study that wants to estimate fish size between two regions in Quebec (Laurentians and Eastern Townships).
- Draw a random sample of 50 lakes in each region.
- Randomly sample 10 individuals in each lake (*n* = 10).

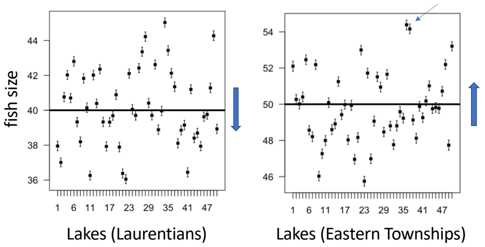
One fixed factor = region (these two regions cannot change).

One random factor = lakes (because they are more likely to be similar within regions and they are not crossed between regions, i.e., different lakes are used).

36

- Consider a study that wants to estimate fish size between two regions in Quebec (Laurentians and Eastern Townships).
- Draw a random sample of 50 lakes in each region.
- Randomly sample 10 individuals in each lake (*n* = 10).

Each confidence interval represents variation is size of 10 fish



Very high (~1) intraclass correlation here because individuals are very similar within lakes than among lakes; so, a random effect would be important.

37

- Consider a study that wishes to estimate fish size between two regions in Quebec (Laurentians and Eastern Townships).
- Draw a random sample of 50 lakes in each region.
- Randomly sample 10 individuals in each lake (*n* = 10).

- Assume for the sake of discussion that all individuals within each lake had the exact same size but size differed between each of the lakes, then

  intraclass correlation = 1 and *a design effect* = 1 + 1(50 -1) = 50.

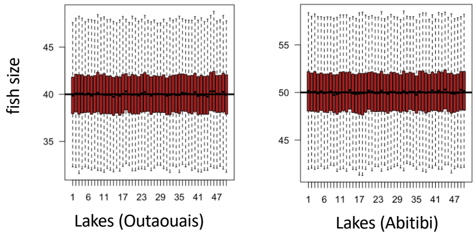
- In this case, we have started our sample with 500 individuals across lakes, but the "design" can only "use" 50 values ("observations") in the analysis because all individuals were too similar within lakes! So, to increase the degrees of freedom, we would need to sample now 500 lakes per region instead! So, this design effect reduced the statistical power of the ANOVA.

38

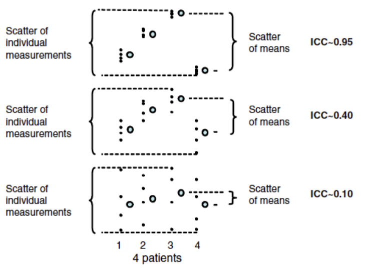
- Consider now these data (two other regions): If variation in size *within* each lake is the same for all lakes, then

  intraclass correlation = 0 and *what we call design effect* = 1 + 0(50 -1) = 1.

- In this extreme case, each additional lake adds no new information about the fish size in each region.
- Only surveying one lake would give us the same information (with the same standard error) about fish size as we get from surveying 50 lakes. So, to get the same degrees of freedom we only need to sample now 500 fish in a single lake per region instead of 10 individuals across 50 lakes!



39


Scatter of individual measurements
Scatter of means
ICC~0.95
Scatter of individual measurements
Scatter of means
ICC~0.40
Scatter of individual measurements
Scatter of means
ICC~0.10
1
2
3
4
4 patients

intraclass correlation coefficient (ICC) or rho

Schematic demonstrating intraclass correlation coefficient (ICC) as a measure of reproducibility. 4 patients each have measurements made 4 times (small dots) with each patient also summarised by an individual average (large dot). In the top panel, there is little within-patient scatter, and therefore the ratio of variance of mean (large dots) to the variance of the raw data (small dots) is almost 1, so ICC≈1. In the middle panel, the ICC is lower. In the bottom panel, within-patient scatter is large, and the means much less varied than the raw data, so ICC is low.

Extracted from Moraldo et al. (2013); International Journal of Cardiology, 166:688-695.

40

Mixed models mix random and fixed effects and allows estimating conducting statistical testing (inference) via proper estimation of design effects for hierarchical (clustered) sampling! It also affects parameter estimation (e.g., Simpson's paradox)

41

Why consider a mixed-model?
Some factors you may be able to control (fixed) and others you won't (random)

(i) Models using random effects are important for inference when analyzing data that exhibit non-independence (hierarchical structure).

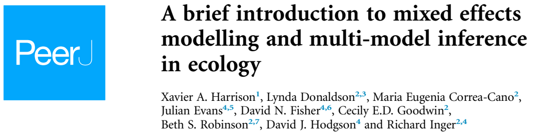
(ii) Random effects provide a unifying statistical framework for models that might otherwise seem unrelated, for example, time-series models for populations, spatial models, genetics models, and models for variation among individuals;

(iii) Models that include random effects are increasingly easy to build and customize for specific problems using publicly available modelling tools and software.

adapted from Thorson and Minto

42

Time for **reading**

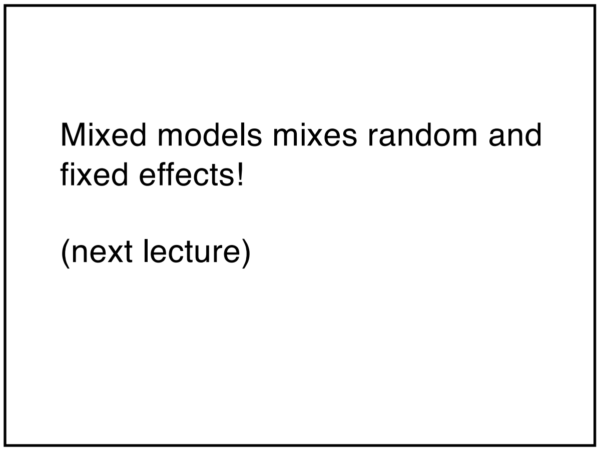


PeerJ

A brief introduction to mixed effects modelling and multi-model inference in ecology

Xavier A. Harrison[1], Lynda Donaldson[2,3], Maria Eugenia Correa-Cano[2],
Julian Evans[4,5], David N. Fisher[4,6], Cecily E.D. Goodwin[2],
Beth S. Robinson[2,7], David J. Hodgson[4] and Richard Inger[2,4]

43



Mixed models mixes random and fixed effects!

(next lecture)

44