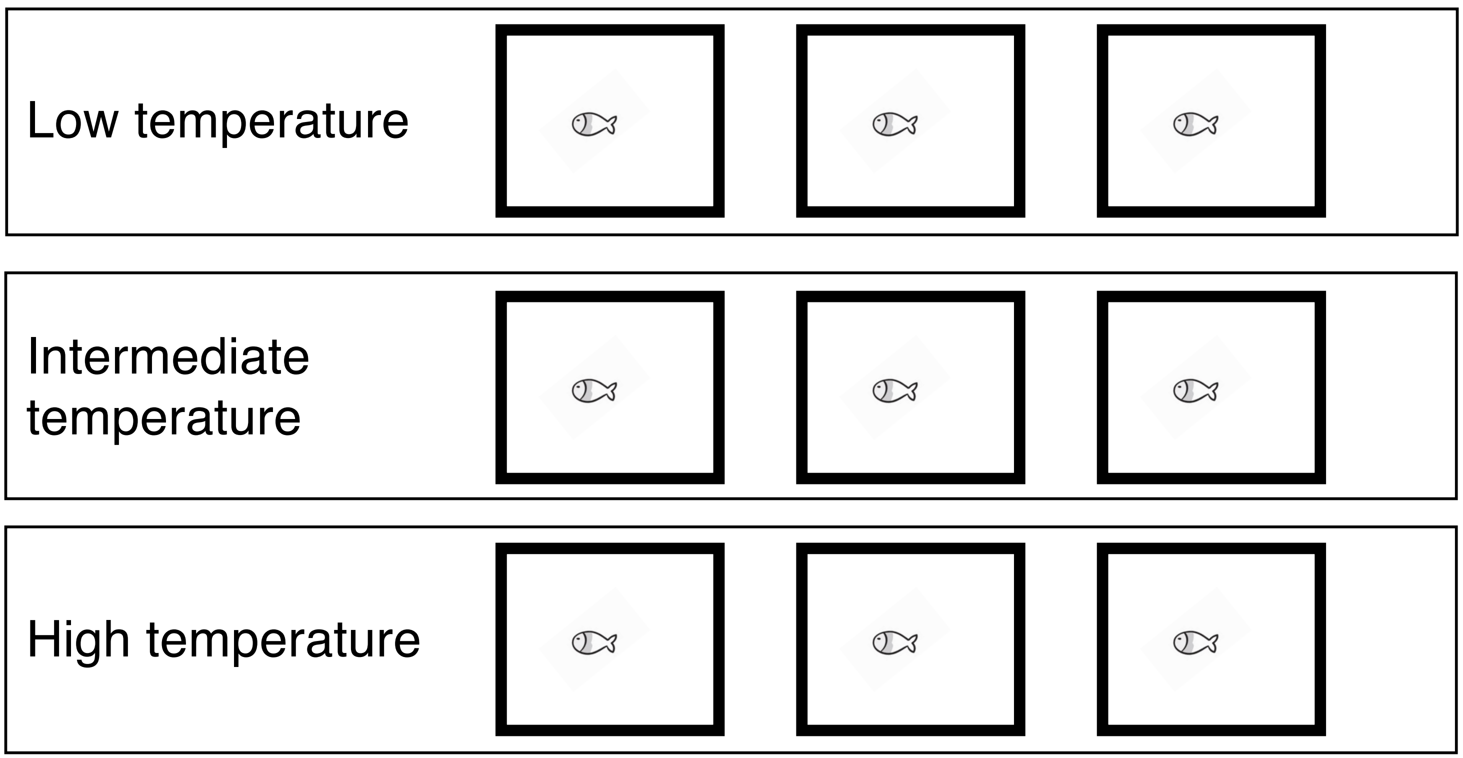
Understanding mixed models for ANOVAs (mixed model ANOVA or Linear Mixed Effects ANOVA)

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?

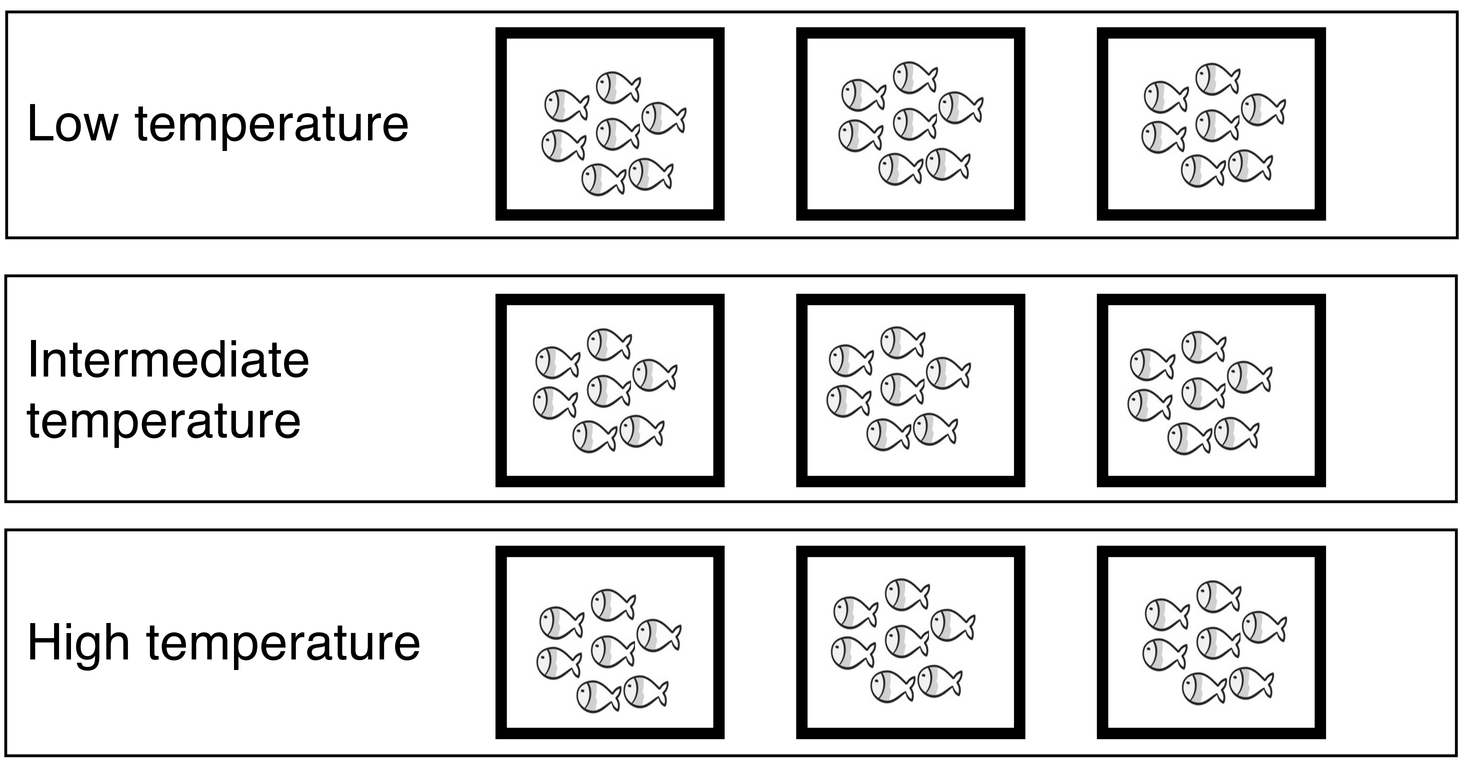
Effects of temperature on fish growth
(difference in growth begin/end of study)



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?

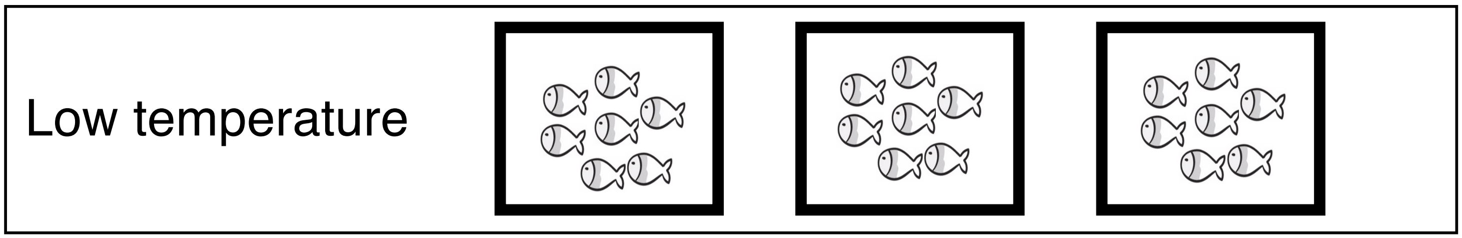
Effects of temperature on fish growth
(difference in growth begin/end of study)

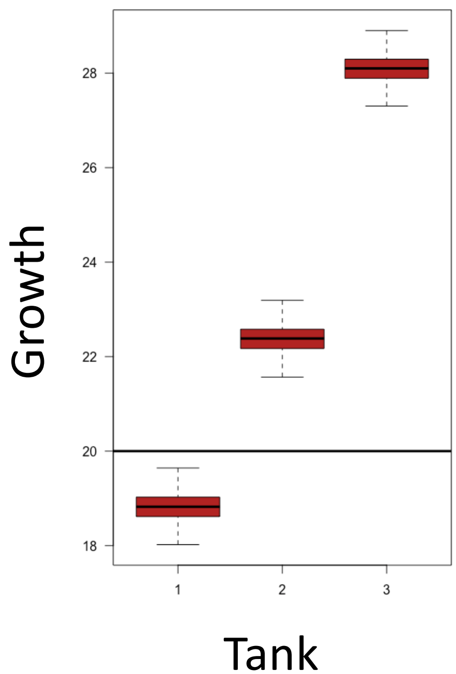


The advantages of mixed models - increase statistical power and estimation accuracy through dependent replication and design convenience (particularly in observational studies).

Do we need a random effect here?

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

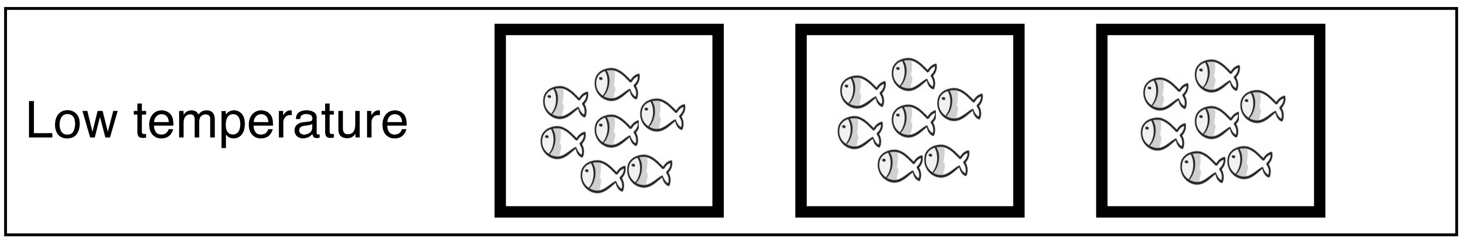


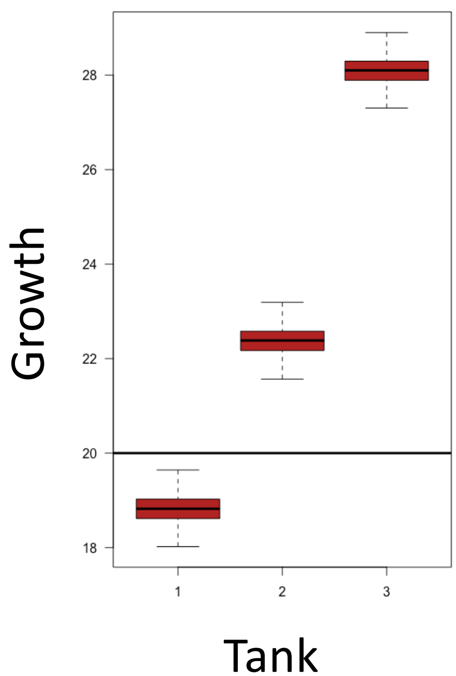


The advantages of mixed models - increase statistical power and estimation accuracy through dependent replication and design convenience (particularly in observational studies).

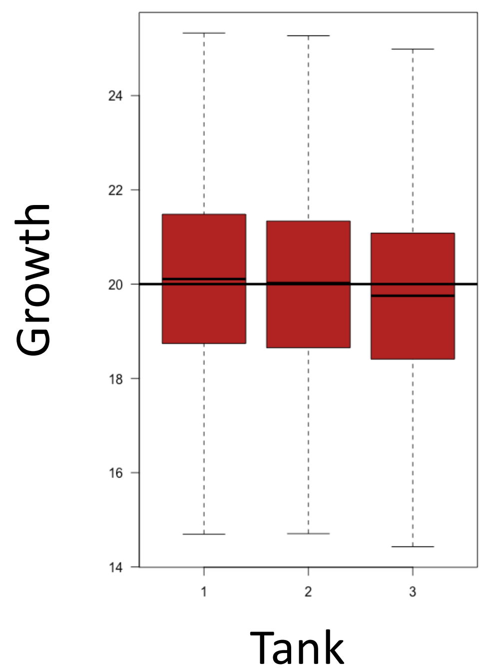
Do we need a random effect here?

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



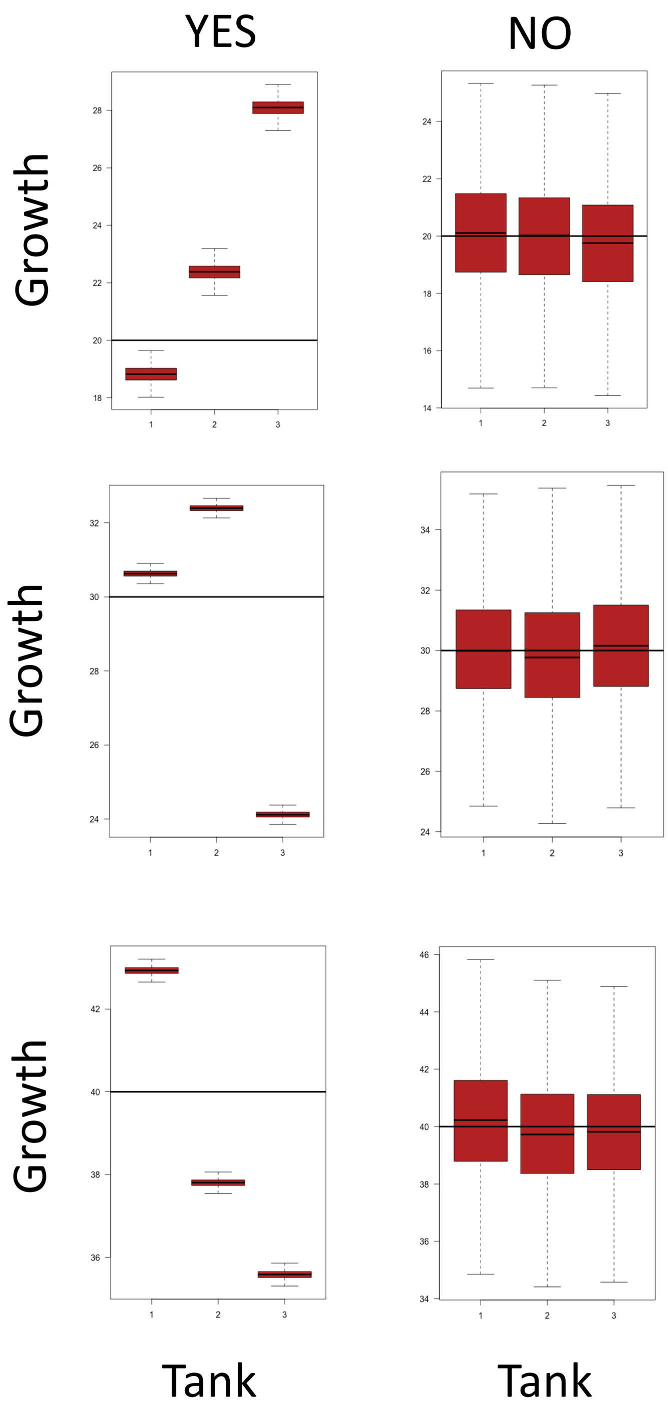


versus



Note that treatment (low temp.) means considering all individuals
are the same in both situation

Do we need a random effect here? Which experimental results you should trust the most?



Low temperature

Average within treatment = 20g

Intermediate temperature

Average within treatment = 30g

High temperature

Average within treatment = 40g

Fixed factor = temperature

Random factor = TANKS

Data structure (fixed effect) - in a regular fixed factor ANOVA individual fish would be treated as an individual replicate regardless of tank, i.e., 21 individual fish per temperature treatment (potential reason: put fish in tanks just to reduce logistics). Tank variation is NOT considered.

```
anova(lm(fish ~ treatments))
```

	fish	treatments	tanks
1	7.68879385558452	low	1
2	6.47888930110018	low	1
3	6.47892888030225	low	1
4	8.55969484798301	low	1
5	5.79259787456118	low	1
6	6.83157449123729	low	1
7	7.85100756943637	low	1
8	7.41725296210829	low	2
9	6.29519824043759	low	2
10	6.06298802779087	low	2
11	5.40829647470436	low	2
12	7.87241370948466	low	2
13	8.22833788242819	low	2
14	8.10957429138447	low	2
15	7.93614167499662	low	3
16	7.29008162923408	low	3
17	5.13694849834886	low	3
18	6.40461476423012	low	3
19	4.89813795483537	low	3
20	9.5462312170886	low	3
21	8.16317982538637	low	3
22	9.62690169935034	int	4
23	9.20257785497386	int	4
24	9.01893202941115	int	4
25	7.58914617378194	int	4
26	8.834403824626	int	4
27	7.45255303344116	int	4
28	8.75606739425442	int	4
29	10.4376338916638	int	5
30	7.57755965449669	int	5
31	8.03430191540074	int	5
32	7.44741063075402	int	5
33	9.21946502797646	int	5
34	9.88207816139077	int	5
35	10.6775358229243	int	5
36	8.93048959369115	int	6
37	11.5693982144121	int	6
38	8.89016900672762	int	6
39	8.14128191918386	int	6
40	9.91474011715798	int	6
41	7.84251427001656	int	6
42	8.62427903729362	int	6
43	10.4066254445766	high	7
44	12.4232739378679	high	7
45	9.96451908736249	high	7
46	10.6908937723269	high	7
47	10.3409644958928	high	7
48	11.9085404331858	high	7
49	9.91568701144683	high	7
50	10.7480065848387	high	8
51	11.0831245999676	high	8
52	11.2625466216413	high	8
53	11.8275933700664	high	8
54	10.985788480584	high	8
55	11.3298206853252	high	8
56	10.5663915372226	high	8
57	11.0895925175086	high	9
58	10.6686619304702	high	9
59	12.050356325582	high	9
60	11.352528167564	high	9
61	12.774538937538	high	9
62	8.71320417410089	high	9
63	10.9823492737741	high	9

Data structure (mixed effect) – here individual fish are treated as replicates within tanks and tank variation within treatments is also considered; hence we need to use a one-factorial mixed-effects ANOVA:

```
lme(fish ~ treatments, random=~1|tanks)
```

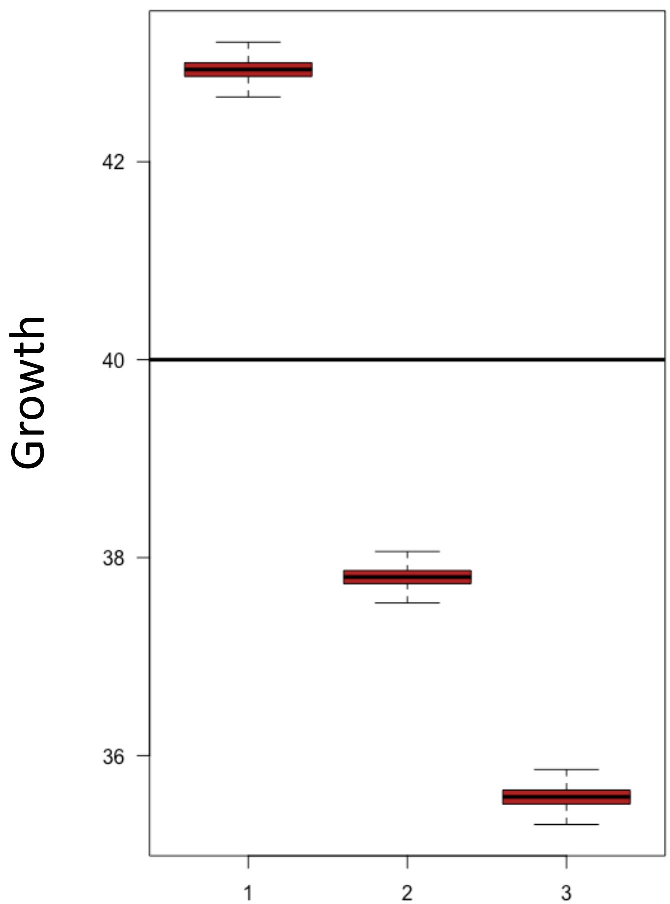
	fish	treatments	tanks
1	7.68879385558452	low	1
2	6.47888930110018	low	1
3	6.47892888030225	low	1
4	8.55969484798301	low	1
5	5.79259787456118	low	1
6	6.83157449123729	low	1
7	7.85100756943637	low	1
8	7.41725296210829	low	2
9	6.29519824043759	low	2
10	6.06298802779087	low	2
11	5.40829647470436	low	2
12	7.87241370948466	low	2
13	8.22833788242819	low	2
14	8.10957429138447	low	2
15	7.93614167499662	low	3
16	7.29008162923408	low	3
17	5.13694849834886	low	3
18	6.40461476423012	low	3
19	4.89813795483537	low	3
20	9.5462312170886	low	3
21	8.16317982538637	low	3

	fish	treatments	tanks
22	9.62690169935034	int	4
23	9.20257785497386	int	4
24	9.01893202941115	int	4
25	7.58914617378194	int	4
26	8.834403824626	int	4
27	7.45255303344116	int	4
28	8.75606739425442	int	4
29	10.4376338916638	int	5
30	7.57755965449669	int	5
31	8.03430191540074	int	5
32	7.44741063075402	int	5
33	9.21946502797646	int	5
34	9.88207816139077	int	5
35	10.6775358229243	int	5
36	8.93048959369115	int	6
37	11.5693982144121	int	6
38	8.89016900672762	int	6
39	8.14128191918386	int	6
40	9.91474011715798	int	6
41	7.84251427001656	int	6
42	8.62427903729362	int	6

	fish	treatments	tanks
43	10.4066254445766	high	7
44	12.4232739378679	high	7
45	9.96451908736249	high	7
46	10.6908937723269	high	7
47	10.3409644958928	high	7
48	11.9085404331858	high	7
49	9.91568701144683	high	7
50	10.7480065848387	high	8
51	11.0831245999676	high	8
52	11.2625466216413	high	8
53	11.8275933700664	high	8
54	10.985788480584	high	8
55	11.3298206853252	high	8
56	10.5663915372226	high	8
57	11.0895925175086	high	9
58	10.6686619304702	high	9
59	12.050356325582	high	9
60	11.352528167564	high	9
61	12.774538937538	high	9
62	8.71320417410089	high	9
63	10.9823492737741	high	9

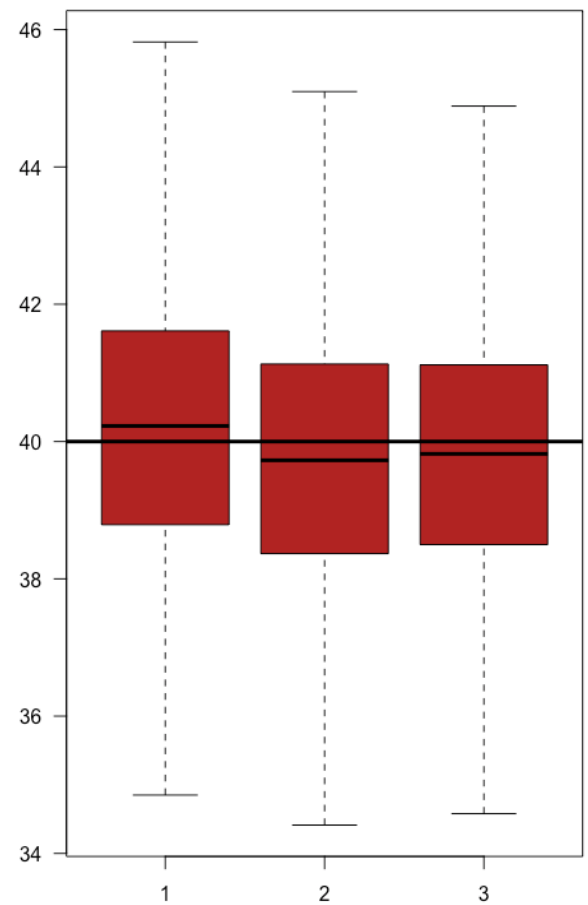
The plural of anecdote is not data (Roger Brinner)

Case 1
(random effect very strong, i.e., more uncertainty/variation among replicates (tanks))



High temperature - Tanks

Case 2
(random effect weak, i.e., small uncertainty/variation among replicates (tanks))



High temperature - Tanks

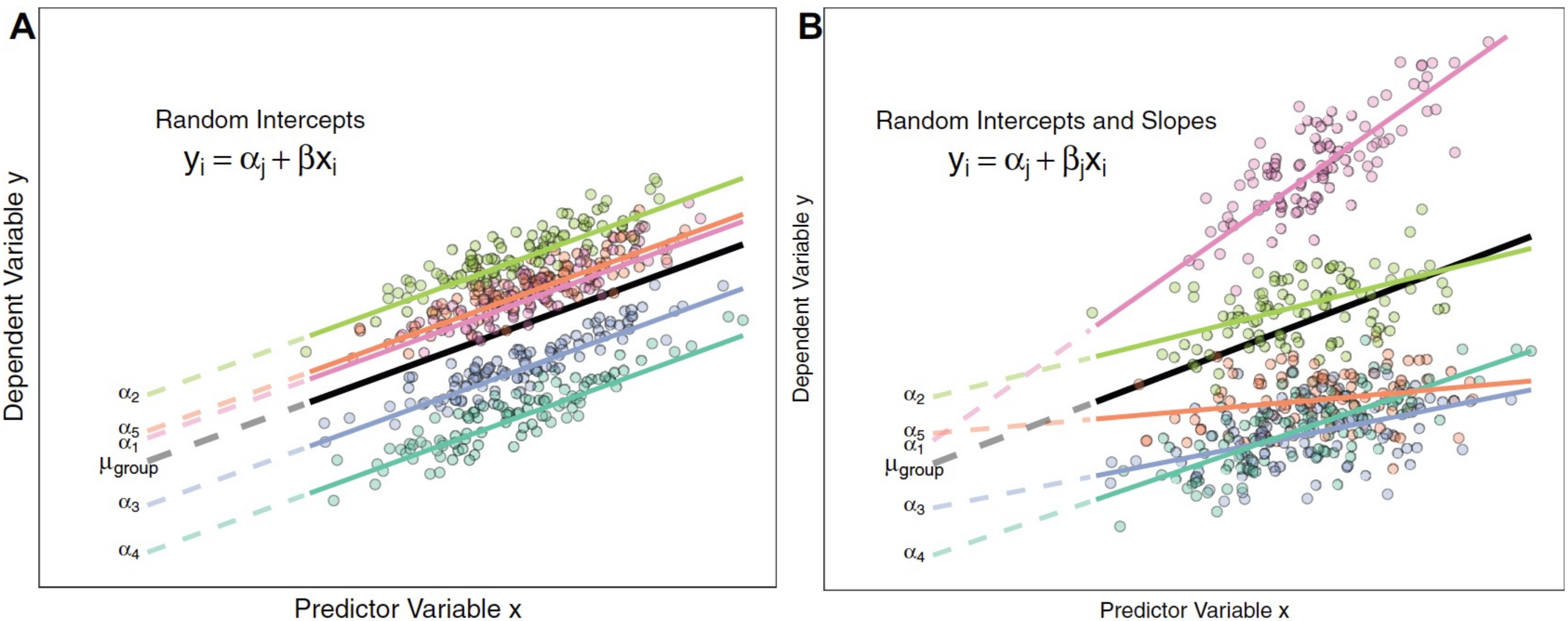
Mixed models for ANOVAs (tutorial 9)

Sources of variation:

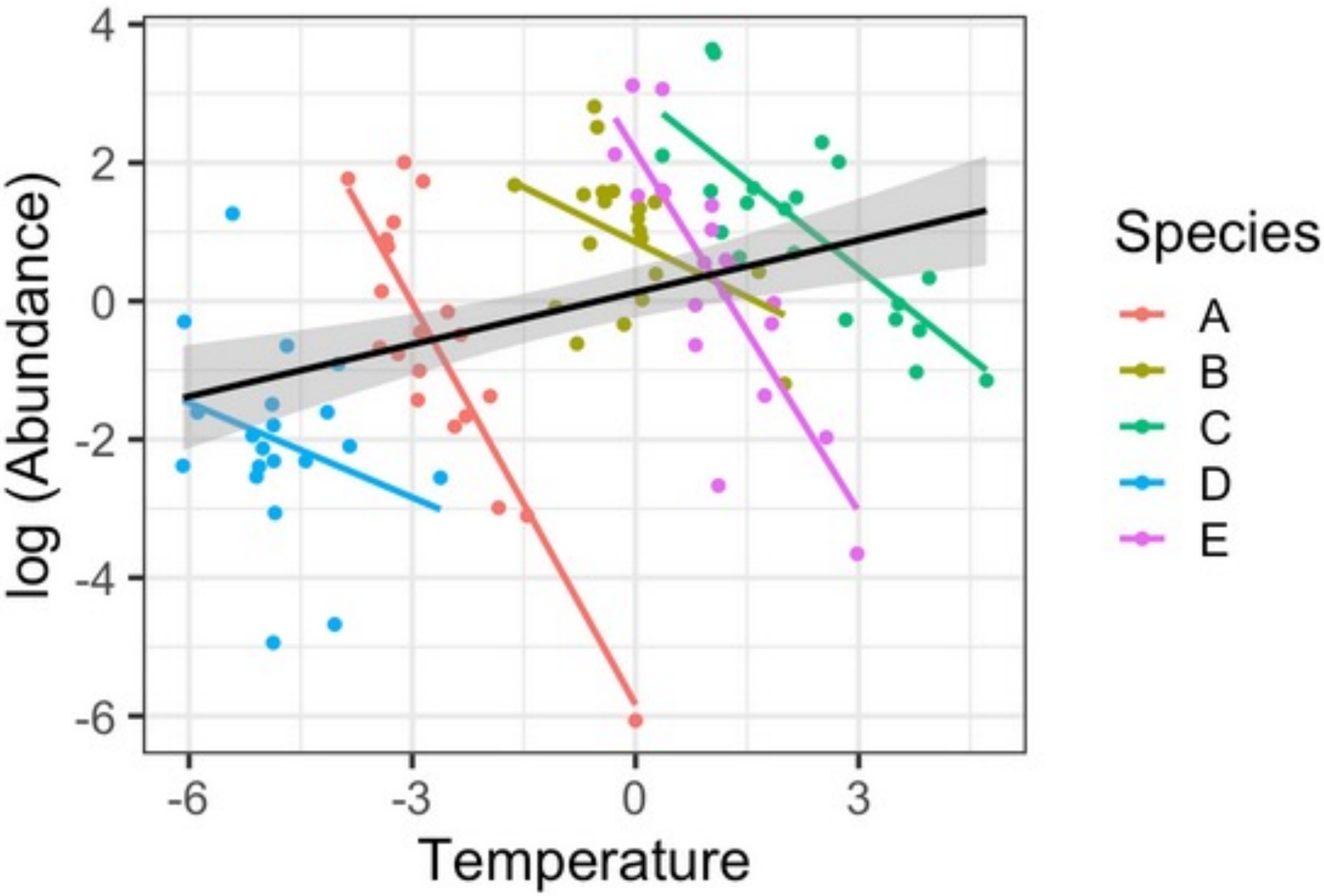
Fixed effect model -
Effects of treatments (e.g., temperature)
Residuals

Mixed effect model (fixed + random effect) -
Effects of treatments (e.g., temperature)
Residuals
Variation among replicates within fixed effect (e.g., tank)

Understanding mixed models for regressions via a two-stage method!



From Harrison et al. (2018) PeerJ 6:e4794



log (Abundance)
Temperature
4
2
0
-2
-4
-6
-6
-3
0
3
Species
A
B
C
D
E

Zuur et al. (2007) used marine benthic data from **nine inter-tidal areas** along the Dutch coast collected by the RIKZ institute (summer of 2002).

In **each intertidal zone** (zone where ocean meets land; denoted by 'beach'), five samples were taken, and the macro-fauna and abiotic variables were measured.

The goal is to model how species richness change as a function of **NAP** (Normal Amsterdam Level: the height of a sampling station compared to mean tidal level) and **Exposure** - a nominal index for the entire beach (high/low) composed of the following elements: wave action, length of the surf zone, slope, grain size, and the depth of the anaerobic layer.

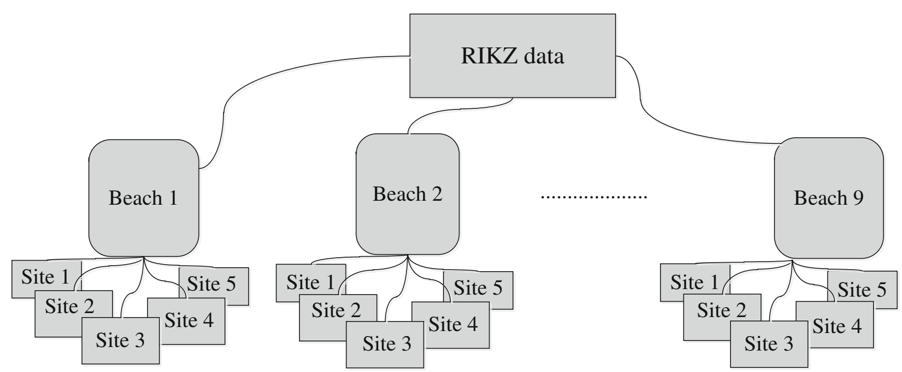
$$R_{ij} = b_0 + b_1 \times NAP_{ij} + b_2 \times Exposure_j + e_{ij}$$

Each site for each beach has a NAP value

One value per beach

$$\varepsilon_{ij} \sim N(0, \sigma^2)$$

$$i = \text{sites};$$
$$j = \text{beach}$$



Zuur AF, Ieno EN, Smith GM (2007) Analysing Ecological Data. Springer.

RIKZ data

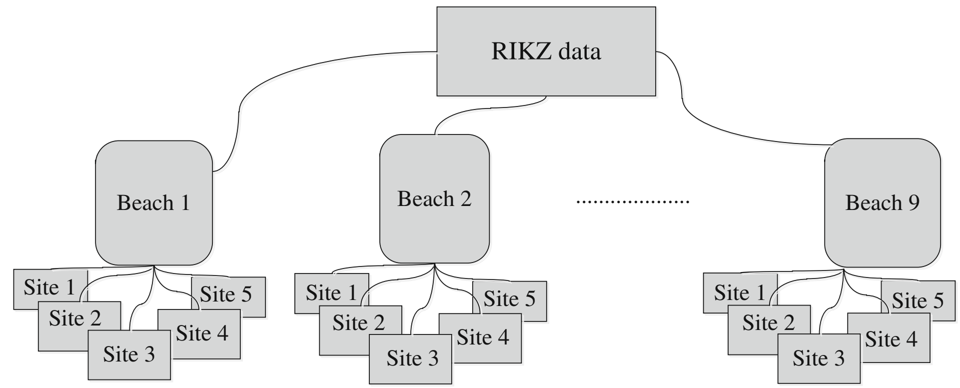
A	B	C	D	E
Sample	Richness	Exposure	NAP	Beach
1	11	1	0.045	1
2	10	1	-1.036	1
3	13	1	-1.336	1
4	11	1	0.616	1
5	10	1	-0.684	1
6	8	1	1.19	2
7	9	1	0.82	2
8	8	1	0.635	2
9	19	1	0.061	2
10	17	1	-1.334	2
11	6	2	-0.976	3
12	1	2	1.494	3
13	4	2	-0.201	3
14	3	2	-0.482	3
15	3	2	0.167	3
16	1	2	1.768	4
17	3	2	-0.03	4
18	3	2	0.46	4
19	1	2	1.367	4
20	4	2	-0.811	4
21	3	1	1.117	5
22	22	1	-0.503	5
23	6	1	0.729	5

.

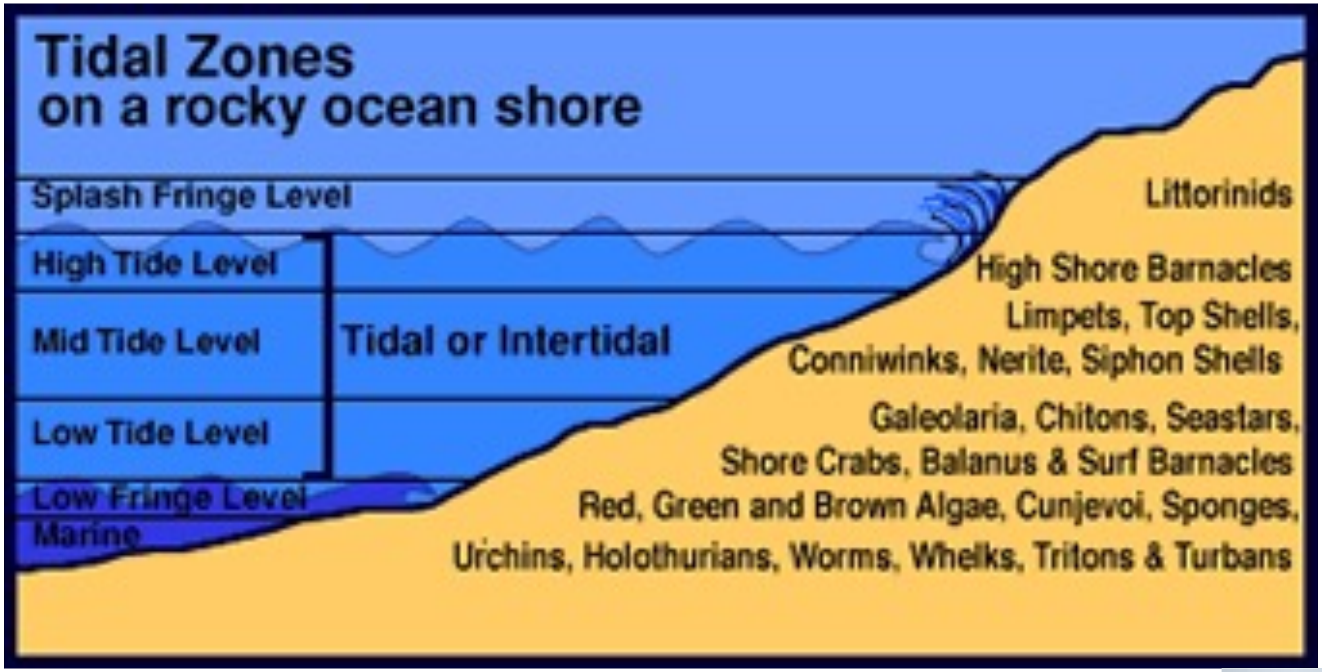
.

.

45



$$R_{ij} = b_0 + b_1 \times NAP_{ij} + e_{ij}$$



Understanding mixed models for regressions via a two-stage method!

Mixed effects models for regression are often introduced first by using an easy-to-understand framework called two-stage analysis.

We then understand better how a mixed model for regression works BUT also understand that the two-stage analysis is not optimal for the analysis.

Then the two-stages (or multiple stages) of the model are combined into a single mixed effect model.

Understanding mixed models via a two-stage method!

The first stage is to fit a linear regression model to each category of the random factor (here beach). Separate intercepts and slopes are calculated for each beach.

$$R_{i1} = b_0 + b_1 \times NAP_{i1} + e_i \qquad j = 1$$

$$R_{i2} = b_0 + b_1 \times NAP_{i2} + e_i \qquad j = 2$$

... .

$$R_{i9} = b_0 + b_1 \times NAP_{i9} + e_i \qquad j = 9$$

Each beach would have a different slope and intercept

Understanding mixed models via a two-stage method!

The first stage is to fit a linear regression model to each category of the random factor (here beach). Separate intercepts and slopes are calculated for each beach. HERE BEACH 1 WAS MODELLED

$$R_{i1} = b_0 + b_1 \times NAP_{i1} + e_i$$

$$\begin{pmatrix} R_{11} \\ R_{21} \\ R_{31} \\ R_{41} \\ R_{51} \end{pmatrix} = \begin{pmatrix} 1 & NAP_{11} \\ 1 & NAP_{21} \\ 1 & NAP_{31} \\ 1 & NAP_{41} \\ 1 & NAP_{51} \end{pmatrix} \times \begin{pmatrix} b_{0_1} \\ b_{11} \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \end{pmatrix}$$

i = sites;

j = beach

Ri is a vector of length 5 containing the species richness values of the 5 sites on beach 1

Understanding mixed models via a two-stage method!

The first stage is to fit a linear regression model to each category of the random factor (here beach). Separate intercepts and slopes are calculated for each beach.

$$R_{i1} = b_0 + b_1 \times NAP_{i1} + e_i$$

Let's say beach 1 had 4 observations instead of 5, then:

$$\begin{pmatrix} R_{11} \\ R_{21} \\ R_{31} \\ R_{41} \end{pmatrix} = \begin{pmatrix} 1 & NAP_{11} \\ 1 & NAP_{21} \\ 1 & NAP_{31} \\ 1 & NAP_{41} \end{pmatrix} \times \begin{pmatrix} b_{0_1} \\ b_{1_1} \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \end{pmatrix}$$

RIKZ data

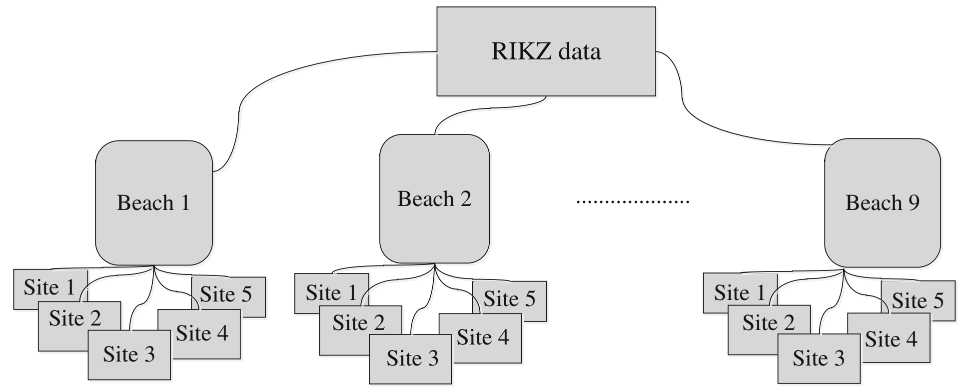
A	B	C	D	E
Sample	Richness	Exposure	NAP	Beach
1	11	1	0.045	1
2	10	1	-1.036	1
3	13	1	-1.336	1
4	11	1	0.616	1
5	10	1	-0.684	1
6	8	1	1.19	2
7	9	1	0.82	2
8	8	1	0.635	2
9	19	1	0.061	2
10	17	1	-1.334	2
11	6	2	-0.976	3
12	1	2	1.494	3
13	4	2	-0.201	3
14	3	2	-0.482	3
15	3	2	0.167	3
16	1	2	1.768	4
17	3	2	-0.03	4
18	3	2	0.46	4
19	1	2	1.367	4
20	4	2	-0.811	4
21	3	1	1.117	5
22	22	1	-0.503	5
23	6	1	0.729	5

.

.

.

45



Understanding mixed models via a two-stage method!

The first stage is to fit a linear regression model to each category of the random factor (here beach). Separate intercepts and slopes are calculated for each beach.

$$R_{ij} = b_0 + b_1 \times NAP_{ij} + e_{ij} \qquad j = 1, \ldots, 4$$

```r

RIKZ <- read.table("RIKZ.txt",header=TRUE)
Beta <- vector()
for (i in 1:9){
  result <- summary(lm(Richness ~ NAP,subset = (Beach==i), data=RIKZ))
  Beta[i] <- result$coefficients[2, 1]
}

```

Understanding mixed models via a two-stage method!

The first stage is to fit a linear regression model to each category of the random factor (here beach). Separate intercepts and slopes are calculated for each beach.

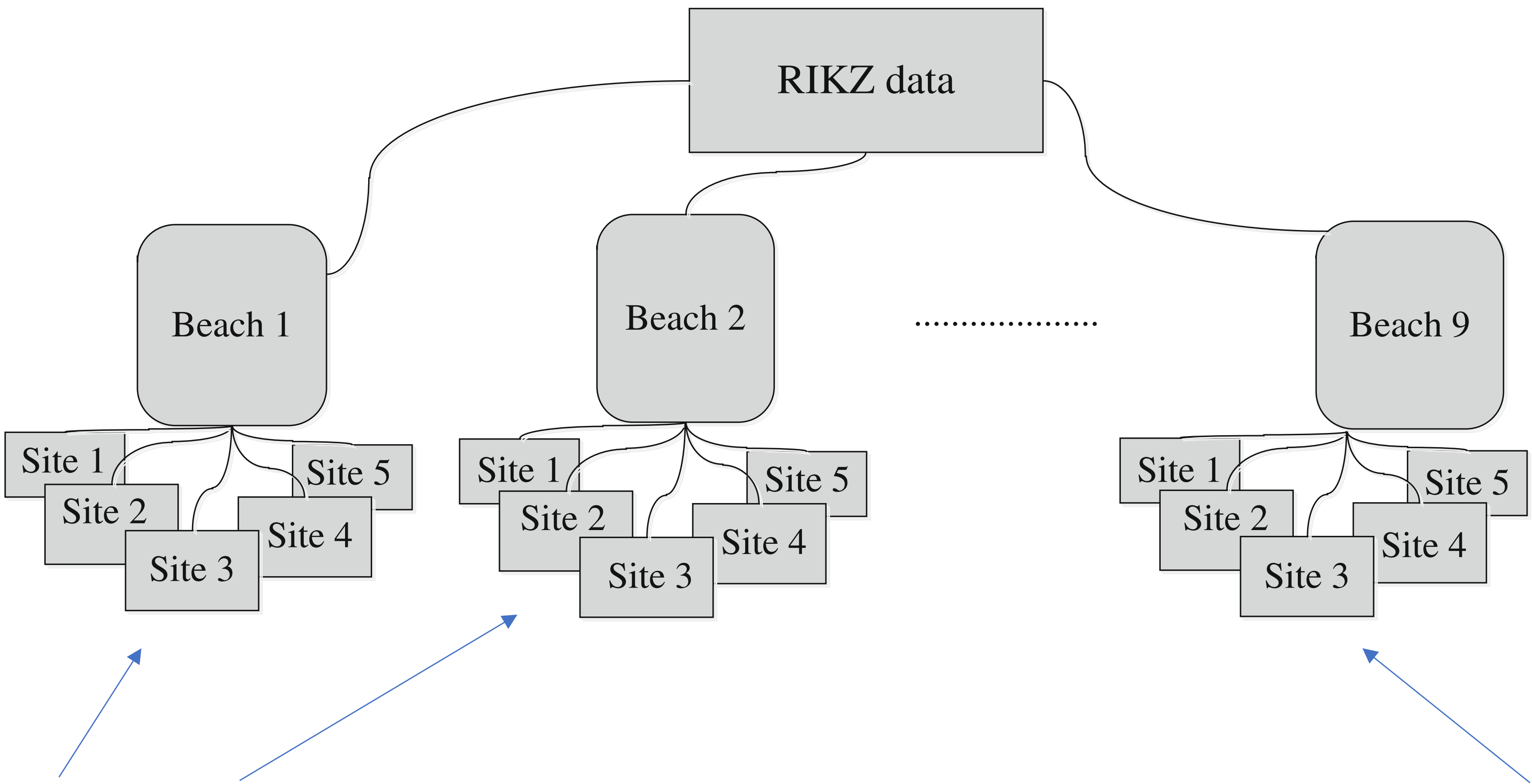
$$R_{ij} = b_0 + b_1 \times NAP_{ij} + e_{ij} \qquad j = 1, \ldots, 4$$

```r
RIKZ <- read.table("RIKZ.txt",header=TRUE)
Beta <- vector()
for (i in 1:9){
  result <- summary(lm(Richness ~ NAP,subset = (Beach==i), data=RIKZ))
  Beta[i] <- result$coefficients[2, 1]
}
```

```
> Beta
[1] -0.3718279 -4.1752712 -1.7553529 -1.2485766 -8.9001779 -1.3885120 -1.5176126 -1.8930665 -2.9675304
```

Lots of differences in slopes among beaches!

RIKZ data




```
> Beta
[1] -0.3718279 -4.1752712 -1.7553529 -1.2485766 -8.9001779 -1.3885120 -1.5176126 -1.8930665 -2.9675304
```

Understanding mixed models via a two-stage method!

The first stage is to fit a linear regression model to each category of the random factor (here beach). Separate intercepts and slopes are calculated for each beach.

$$R_{i1} = b_0 + b_1 \times NAP_{i1} + e_i \qquad j = 1$$

$$R_{i2} = b_0 + b_1 \times NAP_{i2} + e_i \qquad j = 2$$

... .

$$R_{i9} = b_0 + b_1 \times NAP_{i9} + e_i \qquad j = 9$$

Each beach would have a different slope and intercept

Remember that *i* represents the sites within each beach

The second step fits the estimated regression slopes as a function of exposure. Given that expose is a nominal variable, this would just a simple one-way ANOVA:

slope of Exposure for the slopes of R on NAP

Residuals for the slopes

$$\hat{\beta}_j = \eta + \tau \times Exposure_j + e_{b_j} \qquad j = 1, \ldots, 9$$

Intercept

```
> Beta
[1] -0.3718279 -4.1752712 -1.7553529 -1.2485766 -8.9001779 -1.3885120 -1.5176126 -1.8930665 -2.9675304
```

$$j = \text{beach}$$

How does the influence of NAP on richness (slopes of R on NAP) change as a function of exposure?

The second step fits the estimated regression slopes as a function of exposure. Given that expose is a nominal variable, this would just a simple one-way ANOVA:

slope of Exposure for the slopes of R on NAP

Residuals for the slopes

$$\hat{\beta}_{e_{b_i}} = \eta + \tau \times Exposure_{e_{b_i}} + e_{b_{e_{b_i}}} \qquad e_{b_i} = 1, \ldots, 9$$

Intercept

```
> Expose <- factor(c(0, 0, 1, 1, 0, 1, 1, 0, 0))
> anova(lm(Beta ~ Expose))
Analysis of Variance Table

Response: Beta
          Df Sum Sq Mean Sq F value Pr(>F)
Expose     1 10.600 10.6003  1.7551 0.2268
Residuals  7 42.278  6.0397
```

No significant effect of exposure on the individual beach slopes

The second step fits the estimated regression slopes as a function of exposure. Given that expose is a nominal variable, this would just a simple one-way ANOVA:

$$\widehat{\boldsymbol{\beta}_j} = \mathbf{K}_i \times \gamma + e_{b_j} \qquad e_{b_j} \sim N(0, D)$$

$$\begin{pmatrix} -0.37 \\ -4.17 \\ -1.75 \\ -1.24 \\ -8.90 \\ -1.38 \\ -1.51 \\ -1.89 \\ -2.96 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 1 & 0 \\ 1 & 0 \end{pmatrix} \times \begin{pmatrix} \eta \\ \tau \end{pmatrix} + \begin{pmatrix} e_{b_1} \\ e_{b_2} \\ e_{b_3} \\ e_{b_4} \\ e_{b_5} \\ e_{b_6} \\ e_{b_7} \\ e_{b_8} \\ e_{b_9} \end{pmatrix}$$

Understanding mixed models via a two-stage method!

The two formulae of the two-stage approach (more predictors, more stages) and some issues:

$$\mathbf{R}_i = \mathbf{Z}_i \times b_i + e_i \qquad e_i \sim N(0, \sigma^2)$$

$$\widehat{\boldsymbol{\beta}_j} = \mathbf{K}_j \times \gamma + e_{b_j} \qquad e_{b_j} \sim N(0, D)$$

hyperparameter (assumed independent)

1) all the data from a beach is summarized by one parameter (intercept and slope per beach).

2) We analyzed regression parameters, not the observed data; i.e., the variable of interest is not modelled directly but rather the slopes or intercepts or both.

3) The number of observations used to calculate the summary statistic (slopes) is not used in the second step. In this case, we had five observations for each beach. But if you have 5, 50, or 50,000 observations, you still end up with only one summary statistic.

Zuur AF, Ieno EN, Smith GM (2007) Analysing Ecological Data. Springer.

The more appropriate procedure:
Mixed models in one-single step

(next lecture)