3. Intro to mixed models

- Understand when the use of a mixed model is necessary
- Understand the two processes that are occurring in a mixed model
- Understand the difference between fixed and random effects and their parameter estimates

Mixed models to the rescue!

- Almost all biological data is inherently grouped
  - Lakes
  - Incubators
  - Individual animals
  - Sites
- These all violate your assumptions of independence!
- Mixed models should probably be the rule, not the exception (at least according to Andrew Gelman)

Advantages of mixed models

- Allows us to account for repeated measurements on the same group
  - And allows for missing data. Does not drop entire subject if missing one observation for that subject.

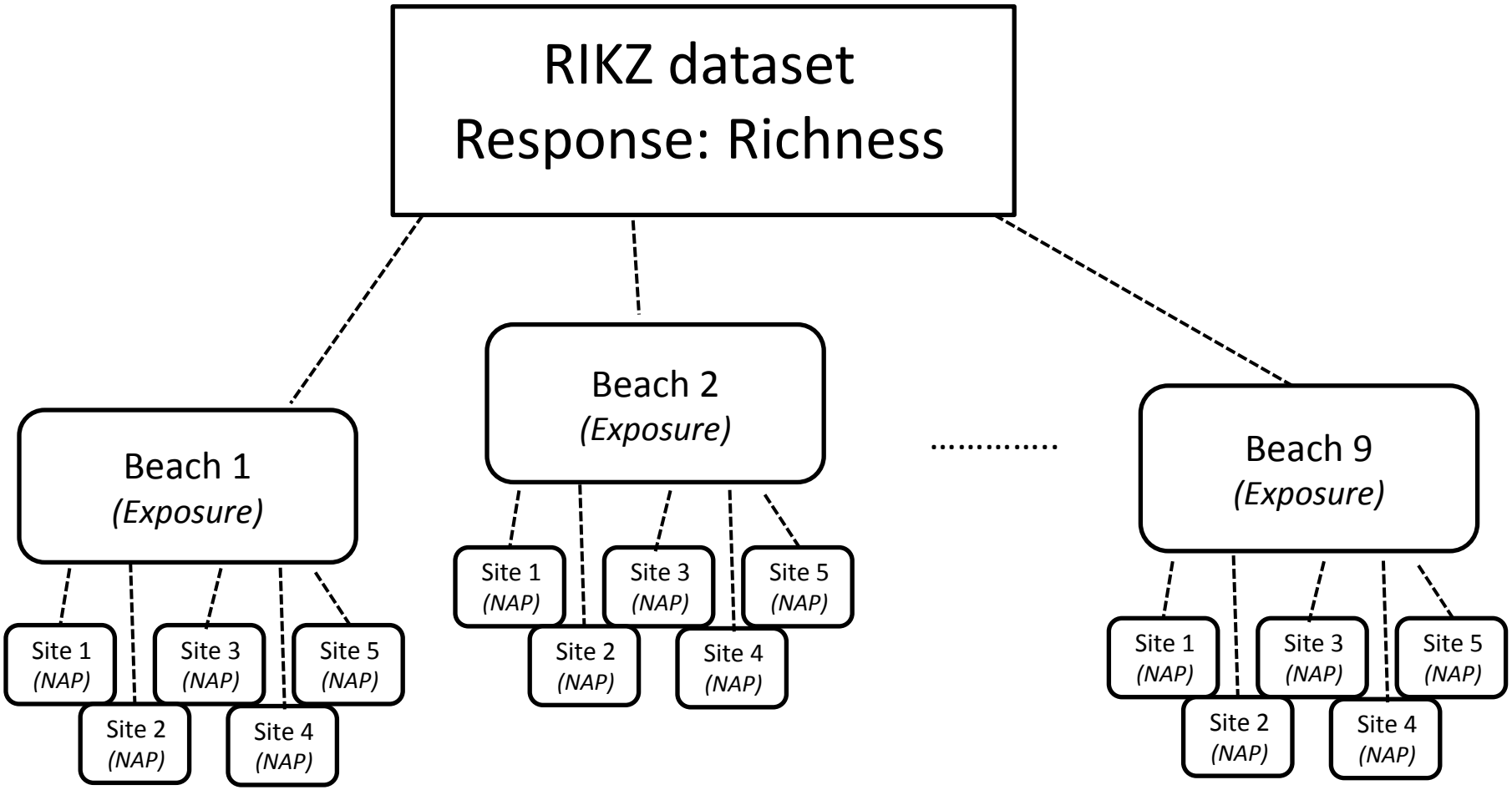
- Helps limit problem of overfitting many different parameters (fixed effect – one parameter **per level of the effect**; random effect – one parameter **per effect**)

- Avoids the need to average within a group which underestimates the true variation (gives you false confidence in your results)

Terminology

- LM, LMM, GLM, GLMMs …..?
  - LM – linear models
  - LMM – linear **mixed** models
  - GLM – **generalized** linear models
  - GLMM – **generalized** linear **mixed** models
- Fixed vs random effect
  - Fixed/predictor/independent **effect** with multiple **levels**
    - E.g. “ Light condition” is the effect; “Control” , “Medium”, “High” are the levels
  - Random **effect** with multiple **levels**
    - E.g. “Individual” is the effect, “ID 1”, “ID 2”, …, “ID 10” are the levels
- Multilevel, hierarchical, nested….

How are mixed models different?



Traditional linear model

- Traditional LM:
  - $R_{ij} = \beta_0 + \beta_1 \times NAP_{ij} + \beta_2 \times Exposure_i + \epsilon_{ij}$
  - $\epsilon_i \sim N(0, \sigma^2)$

- R_{ij} = species richness at site j on beach i
- NAP_{ij} = NAP value at site j on beach i (varies within beach)
- $Exposure_i$ = Exposure value at beach i (same across all sites in a beach; only varies between beaches)
- β_0 = overall intercept when all predictors at zero
- β_1, β_2 = relationship between predictor and response (parameter estimates – what we want to model!)

Traditional linear model

- Traditional LM:
  - $R_{ij} = \beta_0 + \beta_1 \times NAP_{ij} + \beta_2 \times Exposure_i + \epsilon_{ij}$
  - $\epsilon_i \sim N(0, \sigma^2)$

- Why is this not sufficient?

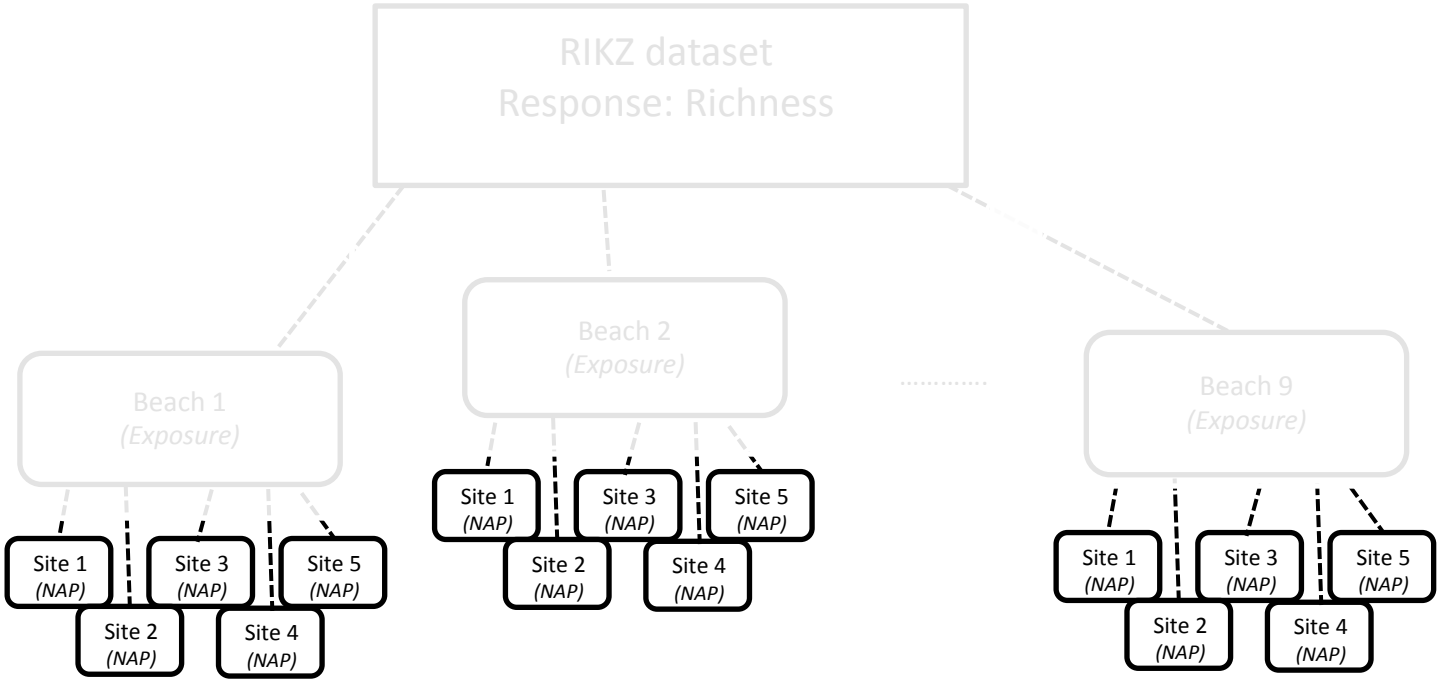
Modeling two separate processes:

- We need to account for the fact that some of our predictors are occurring at different grouping levels:
  - Within a beach, **NAP** levels vary
  - Across beaches, **Exposure** levels vary

Modeling two separate processes

- Within a single beach – what do these terms mean?
  - $R_{ij} = \beta_{0[i]} + \beta_{1[i]} \times NAP_{ij} + \epsilon_{ij} \qquad \epsilon_{ij} \sim \mathrm{N}(0, \sigma^2)$

Within beach effects



Beach 1: $R_{1j} = \beta_{0[1]} + \beta_{1[1]} \times NAP_{1j} + \epsilon_{1j}$

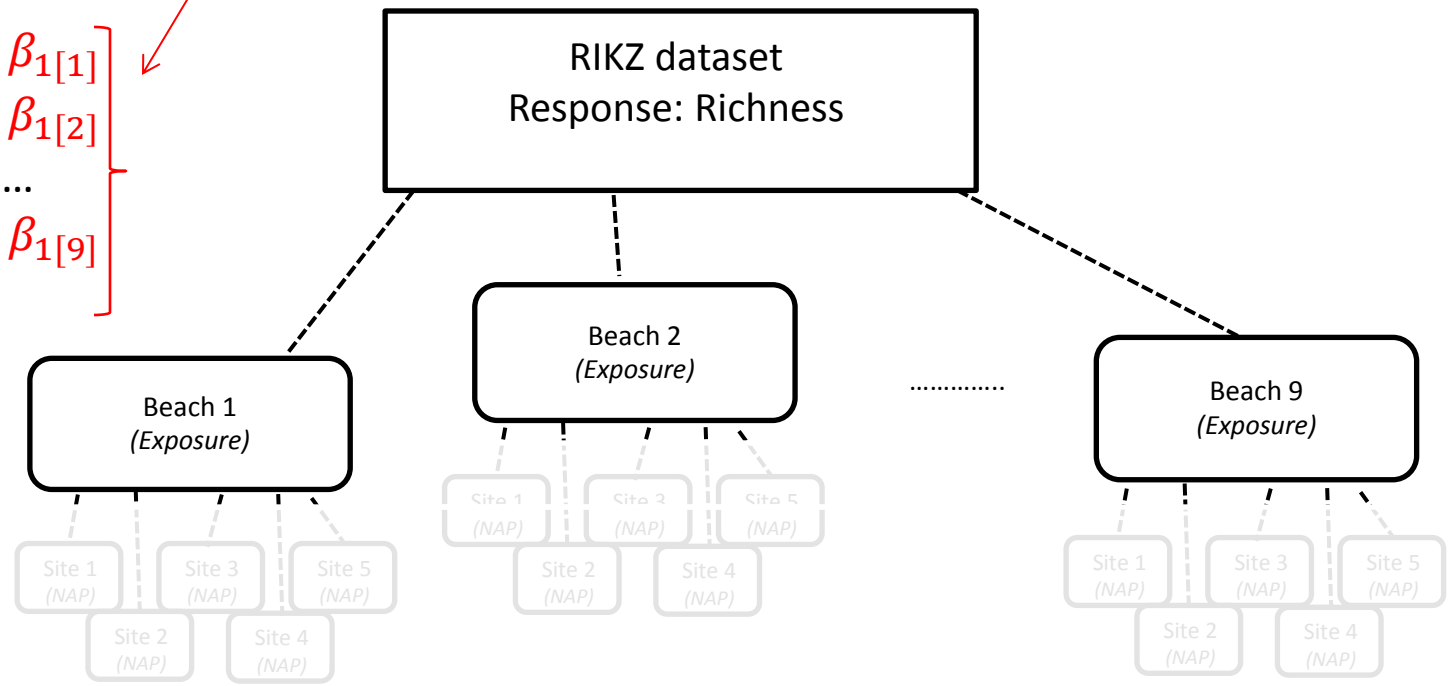
Beach 2: $R_{2j} = \beta_{0[2]} + \beta_{1[2]} \times NAP_{2j} + \epsilon_{2j}$ …

… Beach 9: $R_{9j} = \beta_{0[9]} + \beta_{1[9]} \times NAP_{9j} + \epsilon_{9j}$

Between beach effects

$$\hat{\beta}_i = \beta_0 + \beta_2 \times Exposure_j + b_j \qquad b_j \sim \mathrm{N}(0, \sigma_b^2)$$

$\beta_{1[1]}$
$\beta_{1[2]}$
…
$\beta_{1[9]}$



Beach 1: $R_{1j} = \beta_{0[1]} + \beta_{1[1]} \times NAP_{1j} + \epsilon_{1j}$

Beach 2: $R_{2j} = \beta_{0[2]} + \beta_{1[2]} \times NAP_{2j} + \epsilon_{2j}$ …

… Beach 9: $R_{9j} = \beta_{0[9]} + \beta_{1[9]} \times NAP_{9j} + \epsilon_{9j}$

Modeling two separate processes

- Between beaches – What do these terms mean?
  - $\hat{\beta}_i = \beta_0 + \beta_2 \times Exposure_j + b_j \qquad b_j \sim \mathrm{N}(0, \sigma_b^2)$

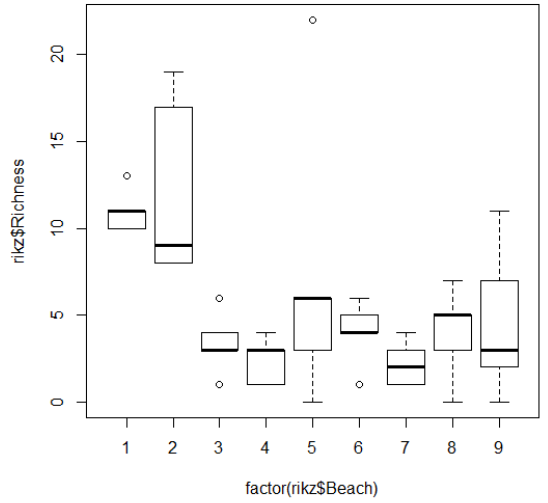
- Couldn't you just average the Richness and NAP values within each beach?

Combining the two steps in one model

- You could include Beach as **fixed** effect:

$$R_{ij} = \beta_0 + \beta_1 \times NAP_{ij} + \beta_2 \times Beach_j + \epsilon_{ij}$$

  - β_2 will estimate the intercept of each beach separately to account for average differences among beaches in richness
  - Why might this not be desirable though?



Combining the two steps in one model

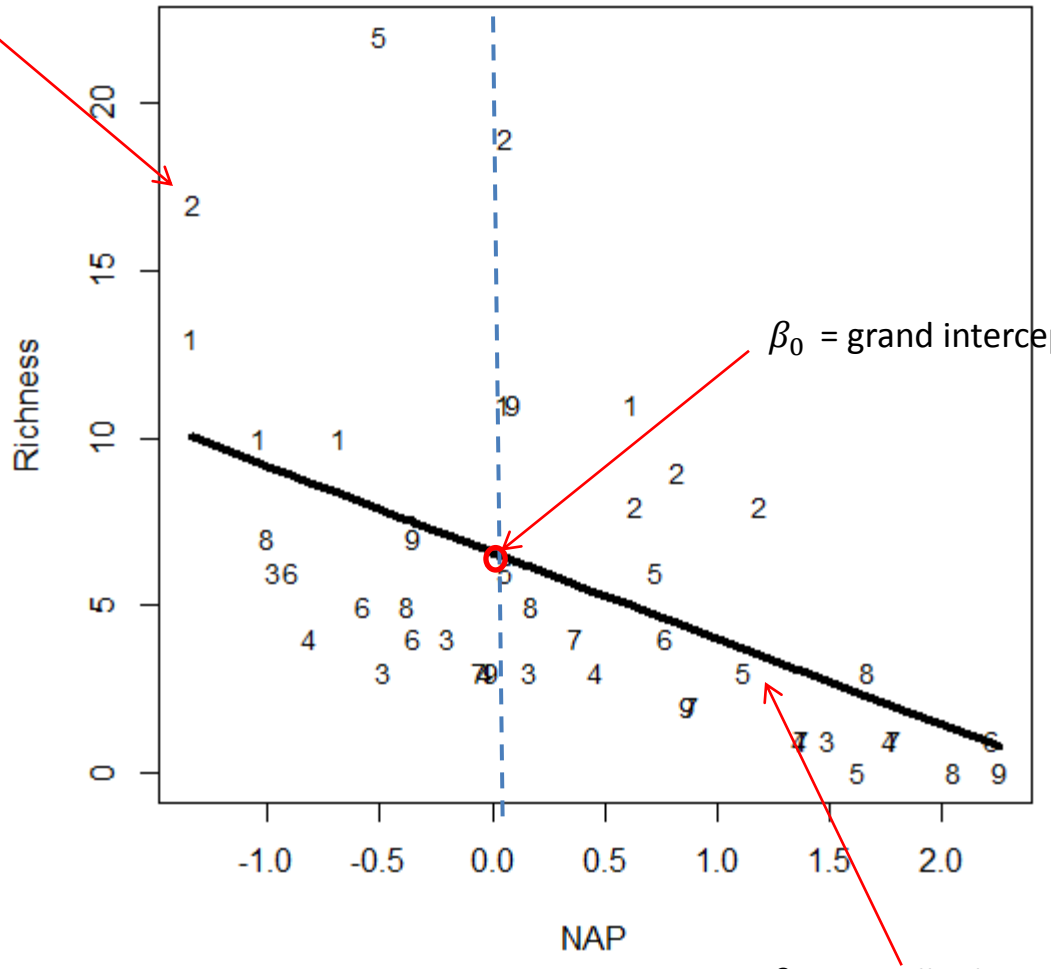
- Better to include Beach as RANDOM effect:
  - Why? What is the model doing differently?

Fixed or random?

- Grouping variables (e.g. beaches) should **always** be included in your model, but should it be fixed or random?
  - Are you inherently interested in its effect?
  - Do you want to 'just' exclude the noise it introduces?
  - Is it categorical or continuous (random effects **must** be categorical)
  - How much data, and how many levels do you have?

$$R_{ij} = \beta_0 + \beta_1 \times NAP_{ij} + \epsilon_{ij}$$

R_{ij} = individual observation (coded for Beach ID)

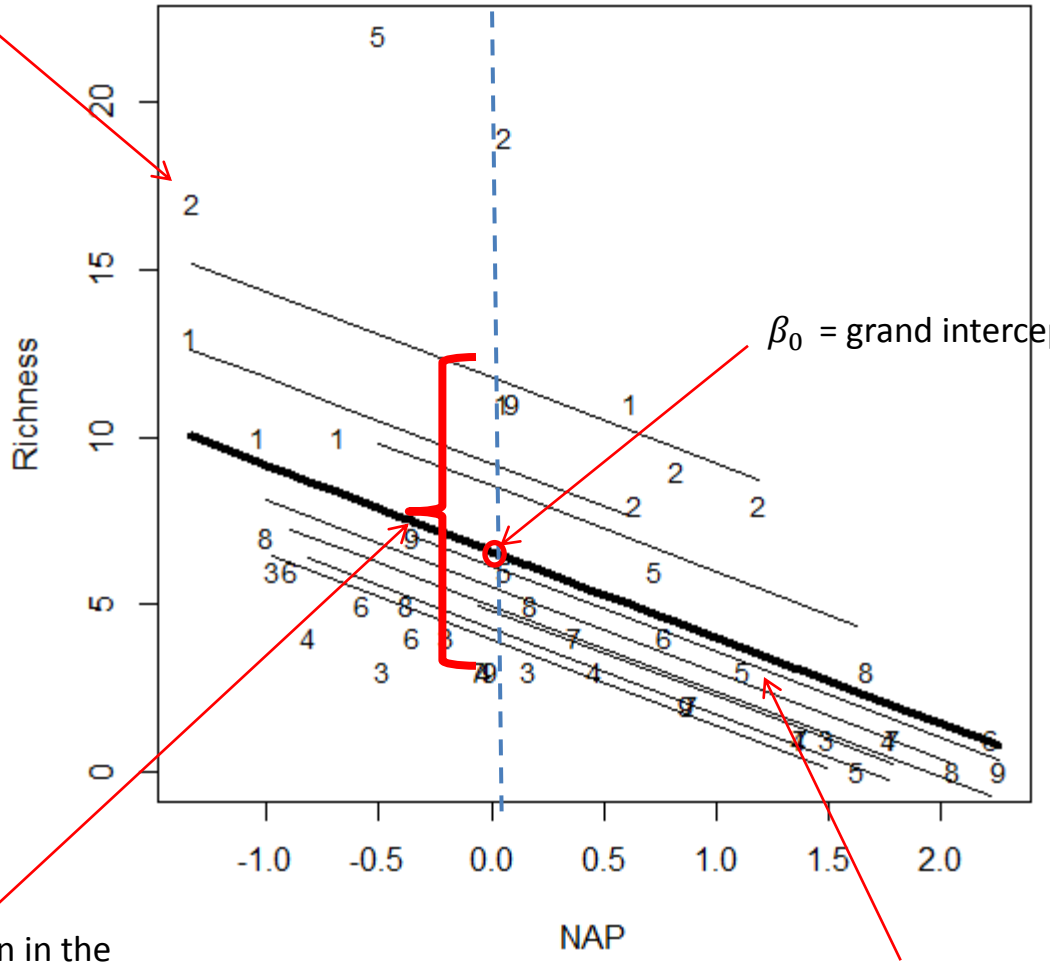


β_0 = grand intercept

β_1 = overall relationship between Richness & NAP

$$R_{ij} = (\beta_0 + \beta_{0j}) + \beta_1 \times NAP_{ij} + \epsilon_{ij}$$

R_{ij} = individual observation (coded for Beach ID)



β_0 = grand intercept

β_{0j} = variation in the intercepts among the beaches

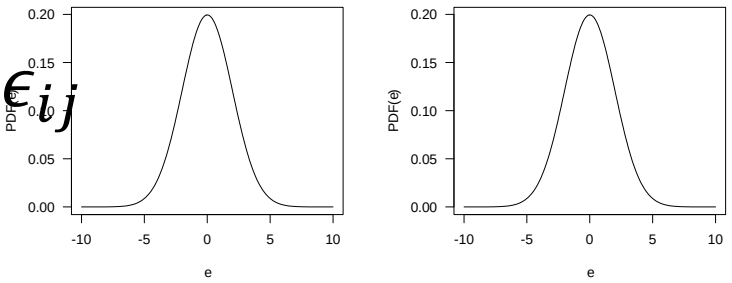
β_1 = overall relationship between Richness & NAP

Combining the two steps in one model

- $y_{ij} = (\beta_0 + \beta_{0j}) + \beta_1 \times x_{ij} + \epsilon_{ij}$

  $\beta_{0j} \sim \mathrm{N}(0, \sigma^2_{int})$

  $\epsilon_{ij} \sim \mathrm{N}(0, \sigma^2_e)$



- $R_{ij} = (\beta_0 + \beta_{0j}) + \beta_1 \times NAP_{ij} + \beta_2 \times Exposure_i + \epsilon_{ij}$

  $\beta_{0j} \sim \mathrm{N}(0, \sigma^2_{int})$

  $\epsilon_{ij} \sim \mathrm{N}(0, \sigma^2_e)$

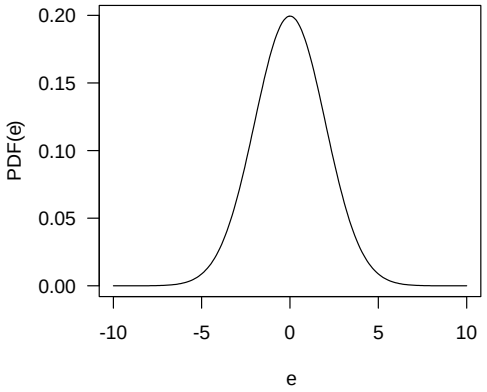
- Now you have TWO separate variance estimates, one for error and one for random intercepts

Random effects

- Random effects are estimating VARIANCE – which means you need multiple observations of these effects.
  - Must have several levels of an effect
    - Most authors recommend at least 4
    - $y_{ij} = (\beta_0 + \beta_{0j}) + \beta_1 \times x_{ij} + \epsilon_{ij}$

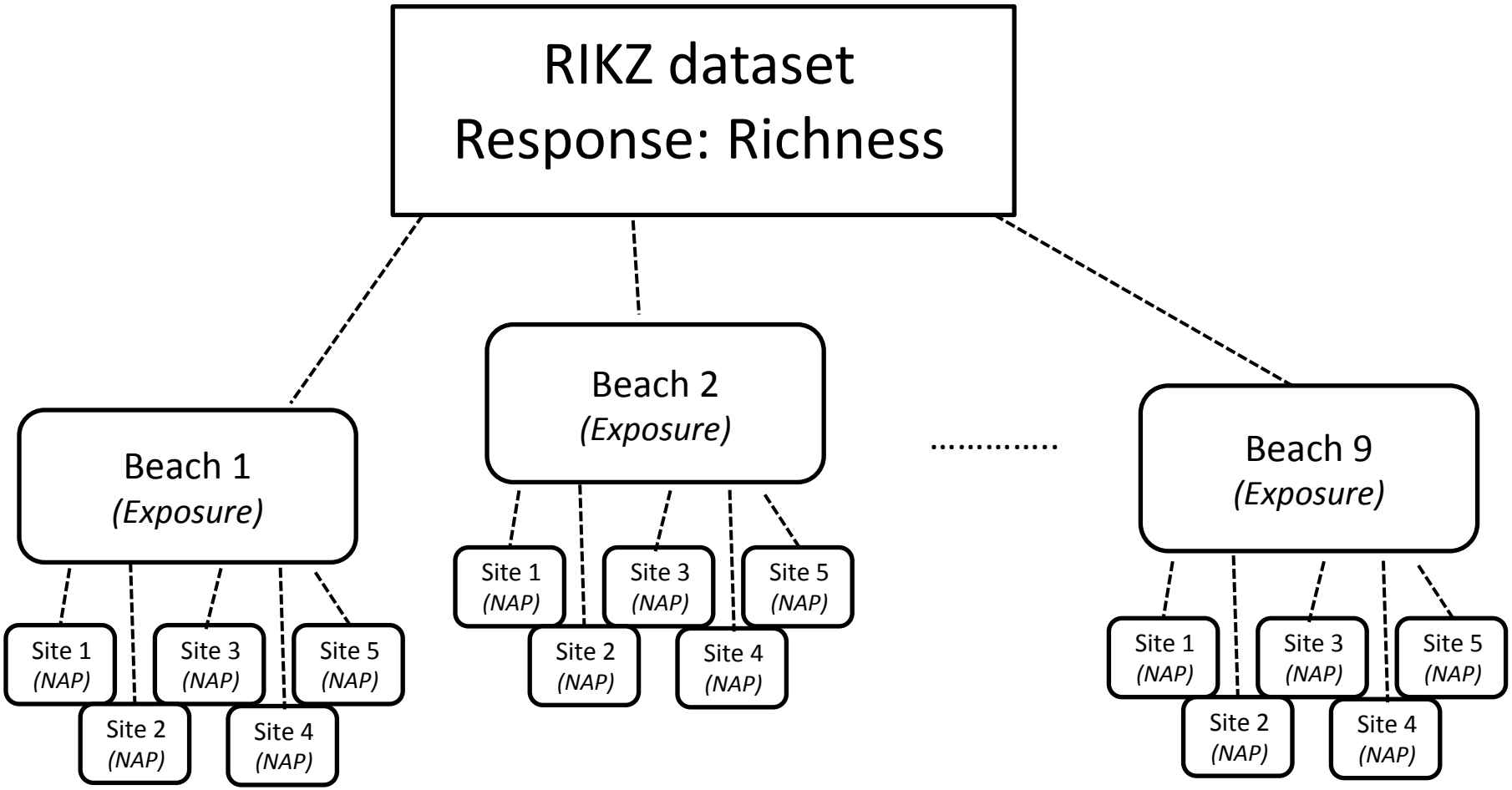
      $\beta_{0j} \sim \mathrm{N}(0, \sigma^2_{int})$

      $\epsilon_{ij} \sim \mathrm{N}(0, \sigma^2_e)$



  - Know where your error is coming from
    - Must have multiple observations at each level
    - Bare minimum is 3; more is better.

Where is your error coming from?



Random effects

- You want to see how fish abundances are influenced by climate factors like latitude, average temperature and average rainfall. You find a database that has measures of abundance across different lakes in Germany. You find 349 lakes that each have one measure of abundance.
- Can you include “lake” as a random effect?

Random effects

- You are interested in the effects of artificial light on insect abundances. You have 8 experimental fields. Half the fields are artificially lit at night, half are not. You sample insects every week for 2 months in the summer. You repeat the experiment over two years.
- Can you include field as a random effect?
- Can you include year as a random effect?

- FURTHER READING on mixed model processes:
  - - Zuur Chapter 5
  - For more explicit statistical theory:
    - Sullivan et al. 1999. Tutorial in biostatistics: an introduction to hierarchical linear modeling. *Statistics in Medicine 18.*
  - Should it be fixed or random? https://dynamicecology.wordpress.com/2015/11/04/is-it-a-fixed-or-random-effect/