EXPERIMENTAL DESIGN IN LONG-TERM ECOLOGICAL RESEARCH

Christina Staudhammer, Associate Professor
Department of Biological Sciences
University of Alabama
The importance of planning your study design

- The first step in rigorous exploration is formulating testable hypotheses or posing critical research questions.
- To apply the scientific method, we must collect data that allow us to discriminate between different hypotheses.
  → we collect data to:
    - estimate values of characteristics of the parent population
    - conduct hypothesis tests
- Before we collect data, we *plan and design data collection procedures in support of those hypotheses and/or questions*.
- Data should be collected with a *purpose*:
  - Independent variables (for explanation)
  - Dependent variables (for inference)
  → *Your research hypotheses/questions define what variables need to be measured*.
Requirements for statistically defensible analysis of data

- Randomization
  - Why?
- Replication
  - Why?
- Design Control
  - What does this mean?

Assures that our own biases do not enter the data.
Necessary to meet assumption of required by most statistical tests

Permits calculation of experimental error, “Insurance” against chance events, Averages out “noise”

Use homogeneous experimental/sampling units,
OR If material is heterogeneous, then use blocking
Randomization

• Random sampling ensures that population parameter estimates are unbiased, e.g.:
  • Plants randomly selected from population of interest
  • Fixed area plot locations randomly selected from within study area
• If we do not obtain a random sample, we reduce our inferential population
• Experimental units should be randomly allocated to treatment groups
Replication

- In order to analyze data, we must have multiple observations of each factor combination we are interested in:
  - If we have one factor we are interested in (e.g. two species), we must have at least two observations per species (4 obs) in order to assess the variability within species and between species.
  - BUT NOTE: two is dangerous – what if one individual dies?

- Replication reduces the chances that we have inherent consistent differences in experimental units that receive the same treatment:
  - i.e., we can be more confident in attributing differences to treatments rather than other factors.
Replication, pseudoreplication, and independence

- Biologists in particular often find it difficult to replicate the exact same conditions, e.g.:
  - Are two pots of soil the same?
  - Are two rivers the same?
- To properly replicate conditions, “pseudo-replicates” are often chosen
- Pseudoreplication also arises when observations are not independent
  - Can arise over space, time, or can be due to genetics
- Independence is necessary for basic statistical techniques (but can be mitigated with more complex methods)
Example: sampling from burned and unburned areas

- Are these really replicates?
  1. If the scale is small (e.g., 1 ha), these are not true replicates, but they are as good as it gets in ecology!
  2. Since the fire was applied to the entire area, we really have only one true replicate (in each of unburned and burned areas) with pseudoreplicates, or subsamples

→ We need multiple fires in order to appropriately evaluate impact of fire in general; otherwise, our inference is only to this fire
What is meant by “experimental design”? 

Controls how we apply treatments to observational units, or select data from different populations

→ Controls how we analyze the data

• [underline]is often intimately related to the sampling design under which the data was collected[/underline]
• E.g., we want to describe longleaf pine regeneration in a 90 ha area with 3 understory types (20 ha in shrub oak, 30 ha in wiregrass, 40 ha in mixed grass/shrub oak)
  • each understory type covers a contiguous and non-overlapping area, so we choose 3 1-ha areas, and within each install 9 grid plots
  • OR, each understory type is patchy over our study area; we choose 3 random areas of each type, and within each install 9 grid plots
One experimental design option
Another experimental design option
What is meant by “experimental design”?

Controls how we apply treatments to observational units, or select data from different populations

→ Controls how we analyze the data

• is often intimately related to the sampling design under which the data was collected

• E.g., we want to describe disease presence in frogs under three moisture regimes (9 each of low, medium, high), and have 3 blocks of space available (in three different locations)

  • In block #1, we observe 9 frogs with low moisture, in block #2, we observe 9 frogs with medium moisture, and in block #3, we observe 9 frogs with high moisture

  • OR, 3 frogs with each of the moisture regimes in each of block #1, #2, #3
One experimental design option

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>A</th>
<th>A</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>B</th>
<th>B</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>B</td>
<td>B</td>
<td>B</td>
</tr>
<tr>
<td>B</td>
<td>B</td>
<td>B</td>
<td>B</td>
</tr>
<tr>
<td>B</td>
<td>B</td>
<td>B</td>
<td>B</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>C</th>
<th>C</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>C</td>
<td>C</td>
<td>C</td>
<td>C</td>
</tr>
</tbody>
</table>

How are these designs different? Under what circumstances is each design more appropriate/more efficient?
Assumptions of “traditional” statistical hypothesis testing

*Note: most tests are robust to *moderate* violations*

1. **Samples are from a ~Normal population**
   - If population is very skewed or multi-modal, tests not valid
   - Transformation can often fix this

2. **Samples are from homoscedastic (equal variance) populations**
   - Often, fixing #1 will fix this problem

3. **Samples are randomly selected from the population**
   - considered in the design stage of your experiment

4. **Samples are independent**
   - If samples are not independent, however, there are often ways to mitigate it in the analysis process
Some types of experimental designs

- **Common Designs**
  - Completely randomized design (CRD)
  - Randomized complete block (RCB)
  - Split-Plot Design (SPD)
  - Others (e.g., Latin Square Design)…

- **Methods of treatment application**
  - Repeated measures experiments
  - Factorial experiments

**NOTE** that experimental design concepts apply to both mensurative and manipulative experiments.
## Completely randomized designs (CRD)

- Treatments are randomly assigned to experimental units.
- Units are randomly selected for the experiment from among the set of interest.
- We assume that units are approximately homogeneous.
  - E.g., we sample understory biomass (kg) in 0.01 ha plots under three irrigation regimes.

<table>
<thead>
<tr>
<th>Irrigation A</th>
<th>Irrigation B</th>
<th>Irrigation C</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
What about unbalanced designs? As long as \( n_i \) are not "too different", we can still use ANOVA techniques, but EE DoF = \( \Sigma n_i - k \) and Total DoF = \( \Sigma n_i - 1 \)

• We would analyze this as a simple one-way ANOVA, or could (equivalently) use regression techniques

• Either is termed a *General Linear Model (GLM)*

<table>
<thead>
<tr>
<th>Source</th>
<th>Degrees of freedom (DoF)</th>
<th>Mean Squares</th>
<th>F test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Irrigation</td>
<td>( k-1=2 )</td>
<td>( MS_{IRR} )</td>
<td>( F_{(2,21)} = \frac{MS_{IRR}}{MS_E} )</td>
</tr>
<tr>
<td>Experimental Error</td>
<td>( k(n-1) = 21 )</td>
<td>( MS_E )</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>( kn-1 = 23 )</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

As the number of experimental units ↑, experimental power ↑

What happens when the number of experimental units ↑?

Where: \( k=3 \) is the number of “treatments”, \( n=8 \) is the number of experimental units per treatment
Fitting CRD models in R

```r
> lm.irr <- lm(biomass ~ irrig, data=data.irr)
> anova(lm.irr)
> summary(lm.irr)
> plot(lm.irr)
> lsmeans(lm.irr, pairwise~irrig)
```

- The function `lm` estimates a linear model (Y~X) using data in the dataframe `data.irr`
- The function `anova` partitions the variation into its different sources (in this case, irrigation and error), and displays F-tests for each effect
- The function `summary` gives estimates of the model coefficients, standard errors, and t-tests, statistics on the model goodness of fit
- The function `plot` produces graphs to verify assumptions
- The function `lsmeans` produces marginal means for each effect level
- NOTE that character-valued X variable(s) are assumed to be categorical predictors, whereas numeric-valued X variables are assumed to be continuous predictors
  → If your factors are numbered (e.g., 1=blue, 2=red, 3=green), then you will have to declare the variable as a factor
Fitting CRD models in R - output

**R output**

```r
> lm.irr <- lm(biomass ~ irrig, data=data.irr)
> anova(lm.irr)
```

Analysis of Variance Table

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Irrig</td>
<td>2</td>
<td>2021.0</td>
<td>1010.5</td>
<td>40.374</td>
<td>0.0003***</td>
</tr>
<tr>
<td>Residuals</td>
<td>21</td>
<td>525.26</td>
<td>25.0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

What does this tell us?
Fitting CRD models in R - output

R output

> summary(lm.irr)

Call:
lm(formula = lm(biomass ~ irrig, data = data.irr)

Residuals:
   Min     1Q Median     3Q    Max
-2.9233 -1.2752 -0.2657  1.3976  3.0226

Coefficients:
               Estimate  Std. Error    t value  Pr(>|t|)
(Intercept) 2.021050  0.4111166   4.92600   0.0011 **
irrig.B     12.199100  0.6021865  20.25670  4.1e-08 ***
Irrig.C     17.991100  0.6021865  29.87470  1.7e-09 ***
---

Residual standard error: 1.09 on 21 degrees of freedom
Multiple R-squared:  0.9406,  Adjusted R-squared:  0.9375
F-statistic: 40.4 on 2 and 21 DF,  p-value: 0.0003

What do these fit statistics tell us?
Fitting CRD models in R - plots

R output

```r
> plot(lm.irr)
```

What does this tell us?
Fitting CRD models in R – marginal means

R output

```r
> lsmeans(lm.irr, pairwise ~ irr)
```

<table>
<thead>
<tr>
<th>irrigation</th>
<th>lsmean</th>
<th>SE</th>
<th>df</th>
<th>lower.CL</th>
<th>upper.CL</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>2.0216</td>
<td>1.7622</td>
<td>21</td>
<td>-1.43252</td>
<td>5.47452</td>
</tr>
<tr>
<td>B</td>
<td>14.2201</td>
<td>1.7622</td>
<td>21</td>
<td>10.76658</td>
<td>17.67362</td>
</tr>
<tr>
<td>C</td>
<td>20.0121</td>
<td>1.7622</td>
<td>21</td>
<td>16.55855</td>
<td>23.46562</td>
</tr>
</tbody>
</table>

Confidence level used: 0.95

```r
$contrasts

<table>
<thead>
<tr>
<th>contrast</th>
<th>estimate</th>
<th>SE</th>
<th>df</th>
<th>t.ratio</th>
<th>p.value</th>
</tr>
</thead>
<tbody>
<tr>
<td>A - B</td>
<td>-12.19911</td>
<td>2.49216</td>
<td>21</td>
<td>-4.895</td>
<td>7.67e-05</td>
</tr>
<tr>
<td>A - C</td>
<td>-17.89117</td>
<td>2.49216</td>
<td>21</td>
<td>-7.219</td>
<td>4.10e-07</td>
</tr>
<tr>
<td>B - C</td>
<td>-5.79252</td>
<td>2.49216</td>
<td>21</td>
<td>-2.324</td>
<td>0.030225</td>
</tr>
</tbody>
</table>
```

P value adjustment: tukey method for comparing a family of 4 estimates

What does this tell us?

There is a 95% probability that the true mean understory biomass under irrigation C is between 16.56 and 23.47 kg

What does this tell us?

There are significant differences between understory biomass values in A vs B and C (p<0.01) and B vs C (p<0.05)
What happens if we measure multiple elements in the same plot?

• In many situations, researchers collect data on multiple elements in the same fixed area plot
  
  • E.g., models of biomass as a function of $k=3$ site qualities: we measure $n=15$ plots that each contain $m=4$ trees (45x4 trees total)
What happens if we measure the same element repeatedly over time?

• In many situations, researchers collect data on the same elements over time
  
  • E.g., models of biomass at $k=3$ sites on $n=15$ trees at $m=4$ times

Site 1: time 1

<table>
<thead>
<tr>
<th>Site 1: time 1</th>
<th>Site 1: time 2</th>
<th>etc…</th>
</tr>
</thead>
<tbody>
<tr>
<td>time 1</td>
<td>time 2</td>
<td></td>
</tr>
</tbody>
</table>

Site 2: time 1

<table>
<thead>
<tr>
<th>Site 2: time 1</th>
<th>Site 2: time 2</th>
<th>etc…</th>
</tr>
</thead>
<tbody>
<tr>
<td>time 1</td>
<td>time 2</td>
<td></td>
</tr>
</tbody>
</table>

etc…
What happens if we measure repeatedly over time, or in the same plot?

? Are observations within plots or measured repeatedly by year independent? probably not!

! And if not, we violate an assumption necessary for statistical hypothesis testing

⇒ These are common occurrences in ecology and other disciplines!

⇒ Can lead to *pseudoreplication*

* To appropriately analyze, we need to consider additional non-fixed effects
Models for data correlated over space/time

- We then want to develop models for these elements
  - For tree-level data collected in fixed area plots
    - trees within the same plot are NOT independent; they are likely more alike than those in different plots
  - For data collected on the same exact trees over time
    - Measurements on the same tree over time are NOT independent; they are likely more alike than those taken on different trees
  - If we ignore these inter-relationships, estimates of the mean will still be unbiased, BUT we artificially inflate our DOF and deflate the standard errors → we are pretending to have more information than we actually have!
Mixed models for multiple measurements per experimental unit

- Knowledge of these correlations can be used to formulate the correct experimental error in our models
- Moreover, this knowledge can be useful in better understanding our data!
Mixed models for multiple measurements per experimental unit (e.g., fixed area plots)

- E.g., models of biomass as a function of $k=3$ site qualities, where we measure $m=4$ trees in each of $n=15$ plots (60 trees total)

- Site 1: Plot 1  Plot 2  Plot 3  Plot 4  ...
- Site 2: Plot 1  Plot 2  Plot 3  Plot 4  ...
- Site 3: Plot 1  Plot 2  Plot 3  Plot 4  ...

### Fixed area plot model

<table>
<thead>
<tr>
<th>Source</th>
<th>Degrees of freedom</th>
<th>F test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Site</td>
<td>$k-1=2$</td>
<td>$F_{(2,42)} = \frac{MS_S}{MS_E}$</td>
</tr>
<tr>
<td>Experimental Error</td>
<td>$k(n-1) = 42$</td>
<td></td>
</tr>
<tr>
<td>Within plot error</td>
<td>$nk(m-1)=135$</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>$knm-1 = 179$</td>
<td></td>
</tr>
</tbody>
</table>

In the case of $k=3$ sites with $m=4$ trees per $n=15$ plots, each plot is a “subject”
Mixed models for multiple measurements per experimental unit (e.g., repeated measures)

- E.g., models of biomass at $k=3$ sites on $n=15$ trees at $m=4$ times

<table>
<thead>
<tr>
<th>Source</th>
<th>Degrees of freedom</th>
<th>F test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Site</td>
<td>$k-1=2$</td>
<td>$MS_S/MS_E$</td>
</tr>
<tr>
<td>Experimental Error</td>
<td>$k(n-1) = 42$</td>
<td>$MS_t/MS_W$</td>
</tr>
<tr>
<td>time</td>
<td>$m-1=3$</td>
<td>$MS_{Sxt}/MS_W$</td>
</tr>
<tr>
<td>Site x time</td>
<td>$(k-1)(m-1)=6$</td>
<td></td>
</tr>
<tr>
<td>Within tree error</td>
<td>$k(n-1)(m-1)=126$</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>$knm-1 = 179$</td>
<td></td>
</tr>
</tbody>
</table>

In the case of $k=4$ sites and $m=4$ measurements per $n=15$ trees, each tree is a “subject”
Mixed models for multiple measurements per experimental unit (e.g., repeated measures)

- The most important aspect of the mixed model is the formulation of the F tests
- The site effect in the model are tested against the Experimental Error, whereas time is tested against the within-tree error
- This ensures that we appropriately account for within subject correlations

<table>
<thead>
<tr>
<th>Repeated times model Source</th>
<th>Degrees of freedom</th>
<th>F test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Site</td>
<td>k-1=2</td>
<td>MS_S/MS_E</td>
</tr>
<tr>
<td>Experimental Error</td>
<td>k(n-1) = 42</td>
<td></td>
</tr>
<tr>
<td>time</td>
<td>m-1=3</td>
<td>MS_t/MS_W</td>
</tr>
<tr>
<td>Site x time</td>
<td>(k-1)(m-1)=6</td>
<td>MS_Sxt/MS_W</td>
</tr>
<tr>
<td>Within tree error</td>
<td>k(n-1)(m-1)=126</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>knm-1 = 179</td>
<td></td>
</tr>
</tbody>
</table>

In the case of k=4 sites and m=4 measurements per n=15 trees, each tree is a “subject”

But this assumes our times are independent. But it is likely that we have correlations among times within tree…
How to formulate the appropriate model?

• The observations are “clustered” within a “subject” (e.g., plot for fixed area example, tree for repeated measures example)

→the observations, and their residuals, are not independent, but correlated.

• There are two ways to deal with this correlation
  • A Marginal or Population Averaged approach.
  • A Mixed Model
The Marginal (Population Averaged) approach

- Instead of modeling correlation among residuals, the covariance structure of the residuals is modeled
  - While in linear models, observations are assumed independent, in marginal models, residuals from a single subject are assumed related.
  - Covariances among subjects are assumed non-zero
    → covariances among residuals from each subject are estimated
- not truly a mixed model, although you can use mixed methods to estimate them.
- (In SAS or SPSS, you use a repeated statement instead of a random statement)
The Mixed Model approach

- The model is altered by controlling for subject as a factor in the model
- Residuals are re-defined as the distance between the observed value and the mean value for that subject
- Subjects are not fixed effects in the model but instead are treated as a random effect
  - This uses less degrees of freedom
Fixed versus random effects

• FIXED effects
  • An effect is fixed if all possible levels about which inferences will be made are represented
  • A level of a fixed effect is an unknown constant, which does not vary
  • If we were to repeat the study, we would choose the same factor levels
  • Examples
    • Regression models are fixed effects models, as X is assumed fixed
    • Most effects that we purposely study are considered fixed

• RANDOM effects
  • Effects are random if the levels represent only a random sample of possible levels
  • Sub-sampling, clustering, and random selection of treatments result in random effects in models
  • If we were to repeat the study, a different set of effect levels would be obtained
How to fit a mixed model with subsamples?

**Recall:** biomass as a function of \( k=3 \) site qualities, where we measure \( m=4 \) trees in each of \( n=15 \) plots (60 trees total)

```r
> library(nlme)
> data.sq$plot <- as.factor(data.sq$plot)
> lme.sq <- lme(biomass ~ quality, random =~1|plot, data=data.sq)
> anova(lme.sq)
> summary(lme.sq)
> plot(lme.sq)
```

- The function `lme` estimates a linear mixed effects model (\( Y \sim X \)) using data in the dataframe `data.sq`
- A random effect is added to account for grouping of trees within plots
  - `~1|plot` fits a model with a random intercept for each plot
- The functions `summary`, `anova`, `plot` are used in the same manner as with the simpler model

**NOTE:** in order for this to work properly in R, you **must** have unique plot numbers, e.g., you cannot have a plot 1 in each site quality!!
R output: mixed model with subsamples

```r
> anova(lm.sq)
numDF denDF F-value p-value
(Intercept) 1 135 29.516138 <.0001
quality 2 42 4.722407 0.0152

> summary(lme.sq)
Linear mixed-effects model fit by REML
  Data: data.sq
    AIC   BIC logLik
344.7039 342.842 -166.3519

Random effects:
  Formula: ~1 | plot
      (Intercept) Residual
StdDev: 1.582772 4.060305

Fixed effects: biomass ~ quality
  Value Std.Error DF   t-value p-value
(Intercept) 1.212249 1.264953 135 0.9583354  0.3437
qualityB 3.316992 1.788913  42 1.8541942  0.0822
qualityC -0.029265 1.788913  42 -0.0163590  0.9872
```

Note the difference in denDF. DoF for EE = 42

Estimates of the variance among plots versus within plots

These tests are for the effect versus the base (A)
R output: mixed model with subsamples

Correlation:

<table>
<thead>
<tr>
<th></th>
<th>quality.B</th>
<th>qualityC</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intr)</td>
<td>-0.707</td>
<td>-0.707</td>
</tr>
<tr>
<td>qualityB</td>
<td>-0.707</td>
<td>0.500</td>
</tr>
</tbody>
</table>

Standardized Within-Group Residuals:

<table>
<thead>
<tr>
<th>Min</th>
<th>Q1</th>
<th>Med</th>
<th>Q3</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1.8651688</td>
<td>-0.6058632</td>
<td>-0.0108787</td>
<td>0.7179328</td>
<td>1.8724672</td>
</tr>
</tbody>
</table>

Number of Observations: 60
Number of Groups: 12

> plot(lme.sq)

This is **not** the correlation between the variables. It is the expected correlation of the model coefficients. This might indicate multicollinearity; it indicates that if you did the experiment again and the coefficient for A got smaller, it is likely that those of B and C would get larger.
How to fit a mixed model with repeated times?

Recall: biomass at $k=3$ sites on $n=15$ trees at $m=4$ times

```r
> library(nlme)
> data.rm$time <- as.factor(data.rm$time)
> lme.rm <- lme(biomass ~ site*time, random = ~1|tree, data=data.rm)
> anova(lme.rm)
> summary(lme.rm)
> plot(lme.rm)
```

- The function `lme` estimates a linear mixed effects model ($Y \sim X$) using data in the dataframe `data.rm`
- `site*time = site + time + site:time`
- A random effect is added to account for grouping of measurements on the same tree
- The functions `summary`, `anova`, `plot` are used in the same manner as with the simpler model
R output: mixed model with repeated times

> anova(lm.rm)

<table>
<thead>
<tr>
<th></th>
<th>numDF</th>
<th>denDF</th>
<th>F-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>1</td>
<td>126</td>
<td>92.46865</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>site</td>
<td>2</td>
<td>42</td>
<td>3.59848</td>
<td>0.0189</td>
</tr>
<tr>
<td>time</td>
<td>3</td>
<td>126</td>
<td>35.55504</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>site:time</td>
<td>6</td>
<td>126</td>
<td>0.50806</td>
<td>0.8673</td>
</tr>
</tbody>
</table>

> summary(lm.rm)
Linear mixed-effects model fit by REML
Data: data.rm
AIC      BIC      logLik
1172.093 1233.503 -428.0465

Random effects:
Formula: ~1 | tree
(Intercept) Residual
StdDev: 3.448301 2.019734

Fixed effects: biomass ~ site * time

<table>
<thead>
<tr>
<th></th>
<th>Value</th>
<th>Std.Error</th>
<th>DF</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>0.733993</td>
<td>1.0318303</td>
<td>126</td>
<td>0.711351</td>
<td>0.4779</td>
</tr>
<tr>
<td>siteB</td>
<td>1.017526</td>
<td>1.4592283</td>
<td>42</td>
<td>0.697304</td>
<td>0.4885</td>
</tr>
<tr>
<td>siteC</td>
<td>3.925862</td>
<td>1.4592283</td>
<td>42</td>
<td>2.690368</td>
<td>0.0094</td>
</tr>
<tr>
<td>time2</td>
<td>2.275080</td>
<td>0.7375026</td>
<td>126</td>
<td>3.084843</td>
<td>0.0024</td>
</tr>
<tr>
<td>time3</td>
<td>2.629211</td>
<td>0.7375026</td>
<td>126</td>
<td>3.425019</td>
<td>0.0005</td>
</tr>
<tr>
<td>time4</td>
<td>2.667666</td>
<td>0.7375026</td>
<td>126</td>
<td>3.617162</td>
<td>0.0004</td>
</tr>
<tr>
<td>siteB:time2</td>
<td>0.375345</td>
<td>1.0429862</td>
<td>126</td>
<td>0.359876</td>
<td>0.7194</td>
</tr>
</tbody>
</table>

Note the difference in denDF. DoF for EE of site = 42

Estimates of the variance among trees versus within trees

These tests are for the effect versus the base site (A) and base time (1)
Correlation:

<table>
<thead>
<tr>
<th></th>
<th>(Intr)</th>
<th>siteB</th>
<th>siteC</th>
<th>time2</th>
<th>time3</th>
<th>time4</th>
<th>siteB:time2</th>
<th>siteC:time2</th>
<th>siteB:time3</th>
<th>siteC:time3</th>
</tr>
</thead>
<tbody>
<tr>
<td>siteB</td>
<td>-0.623</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-0.357</td>
<td>0.253</td>
<td>0.253</td>
<td>0.253</td>
</tr>
<tr>
<td>siteC</td>
<td>-0.623</td>
<td>0.200</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.253</td>
<td>-0.357</td>
<td>0.253</td>
<td>0.253</td>
</tr>
<tr>
<td>time2</td>
<td>-0.357</td>
<td>0.253</td>
<td>0.253</td>
<td></td>
<td></td>
<td></td>
<td>0.253</td>
<td>-0.357</td>
<td>0.253</td>
<td>0.253</td>
</tr>
<tr>
<td>time3</td>
<td>-0.357</td>
<td>0.253</td>
<td>0.253</td>
<td>0.253</td>
<td></td>
<td></td>
<td>0.253</td>
<td>-0.357</td>
<td>0.253</td>
<td>0.253</td>
</tr>
<tr>
<td>time4</td>
<td>-0.357</td>
<td>0.253</td>
<td>0.253</td>
<td>0.253</td>
<td>0.253</td>
<td></td>
<td>0.253</td>
<td>0.253</td>
<td>0.253</td>
<td>0.253</td>
</tr>
<tr>
<td>siteB:time2</td>
<td>0.253</td>
<td>-0.357</td>
<td>-0.179</td>
<td>-0.623</td>
<td></td>
<td></td>
<td>0.253</td>
<td>-0.179</td>
<td>-0.357</td>
<td>-0.623</td>
</tr>
<tr>
<td>siteC:time2</td>
<td>0.253</td>
<td>-0.179</td>
<td>-0.357</td>
<td>-0.623</td>
<td>-0.623</td>
<td></td>
<td>0.253</td>
<td>-0.179</td>
<td>-0.357</td>
<td>-0.623</td>
</tr>
<tr>
<td>siteB:time3</td>
<td>0.253</td>
<td>-0.357</td>
<td>-0.179</td>
<td>-0.354</td>
<td>-0.354</td>
<td></td>
<td>0.253</td>
<td>-0.179</td>
<td>-0.357</td>
<td>-0.354</td>
</tr>
<tr>
<td>siteC:time3</td>
<td>0.253</td>
<td>-0.179</td>
<td>-0.357</td>
<td>-0.354</td>
<td>-0.354</td>
<td>0.200</td>
<td>0.253</td>
<td>-0.179</td>
<td>-0.357</td>
<td>-0.354</td>
</tr>
</tbody>
</table>

Standardized Within-Group Residuals:

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>Q1</th>
<th>Med</th>
<th>Q3</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-2.31861551</td>
<td>-0.58095354</td>
<td>-0.05834473</td>
<td>0.53737553</td>
<td>2.04025551</td>
</tr>
</tbody>
</table>

Number of Observations: 180
Number of Groups: 15

> plot(lm.rm)
Are random intercepts enough?

Random intercepts model
- Intercepts are allowed to vary
- biomass is predicted by an intercept that varies across subject (tree)
- assumes that slopes are fixed (the same pattern across time)
- information about intra-subject correlations help determine whether there is correlation among measurements on the same subject (tree)

Random slopes model
- Slopes are allowed to vary
- slopes are different across subject (tree)
- assumes that intercepts are fixed

Random intercepts and slopes model
- includes both random intercepts and random slopes
- most complex
- both intercepts and slopes are allowed to vary across subject (tree), meaning that they are different across times
How to fit a mixed model with random slope and intercept

**Recall:** biomass at $k=3$ sites on $n=15$ trees at $m=4$ times

```r
> lme.rm <- lme(biomass ~ site*time, random =~time|tree, data=data.rm)
> anova(lme.rm)
```

<table>
<thead>
<tr>
<th></th>
<th>numDF</th>
<th>denDF</th>
<th>F-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>1</td>
<td>126</td>
<td>42.96620</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>site</td>
<td>2</td>
<td>42</td>
<td>3.44695</td>
<td>0.0226</td>
</tr>
<tr>
<td>time</td>
<td>3</td>
<td>126</td>
<td>44.31872</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>site:time</td>
<td>6</td>
<td>126</td>
<td>0.63711</td>
<td>0.7642</td>
</tr>
</tbody>
</table>

```r
> summary(lme.rm)
```

Linear mixed-effects model fit by REML
Data: data.a.rm

<table>
<thead>
<tr>
<th></th>
<th>Value</th>
<th>Std.Error</th>
<th>DF</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>0.733993</td>
<td>0.9333719</td>
<td>126</td>
<td>0.786389</td>
<td>0.4327</td>
</tr>
<tr>
<td>siteB</td>
<td>1.017526</td>
<td>1.3199872</td>
<td>42</td>
<td>0.770860</td>
<td>0.4440</td>
</tr>
<tr>
<td>siteC</td>
<td>3.925862</td>
<td>1.3199872</td>
<td>42</td>
<td>2.974166</td>
<td>0.0043</td>
</tr>
<tr>
<td>time2</td>
<td>2.275080</td>
<td>0.8674875</td>
<td>126</td>
<td>2.622608</td>
<td>0.0095</td>
</tr>
<tr>
<td>time3</td>
<td>2.629211</td>
<td>0.6337064</td>
<td>126</td>
<td>4.148941</td>
<td>0.0001</td>
</tr>
<tr>
<td>time4</td>
<td>2.667666</td>
<td>0.6950092</td>
<td>126</td>
<td>3.838318</td>
<td>0.0002</td>
</tr>
</tbody>
</table>

With only random intercept:

<table>
<thead>
<tr>
<th></th>
<th>numDF</th>
<th>denDF</th>
<th>F-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>1</td>
<td>126</td>
<td>92.46</td>
<td></td>
</tr>
<tr>
<td>site</td>
<td>2</td>
<td>42</td>
<td>3.59</td>
<td></td>
</tr>
<tr>
<td>time</td>
<td>3</td>
<td>126</td>
<td>35.55</td>
<td></td>
</tr>
<tr>
<td>site:time</td>
<td>6</td>
<td>126</td>
<td>0.50</td>
<td></td>
</tr>
</tbody>
</table>

Does the AIC indicate a better model? (AIC=1172 in intercept only model)

The estimates are the same, but the standard errors are very different!
What correlation pattern do we expect among observations on the same subject?

- The models we fit assumed a compound symmetric correlation structure (CS) among measurements taken on the same subject (trees in the same plots or times on the same tree)
- What if we think measurements taken closer together in time/space might be more correlated than those taken farther apart?
General form of a variance-covariance matrix

$$
\Sigma = \begin{bmatrix}
\sigma_{11}^2 & \sigma_{21}^2 & \ldots & \sigma_{t1}^2 \\
\sigma_{21}^2 & \sigma_{22}^2 & \ldots & \sigma_{t2}^2 \\
\vdots & \vdots & \ddots & \vdots \\
\sigma_{t1}^2 & \sigma_{t2}^2 & \ldots & \sigma_{tt}^2 
\end{bmatrix}
$$

Diagonal elements are the variances among observations from different subjects taken at the same time.
Off-diagonal elements are the co-variances between observations taken at different times.
Variance components – type matrix (VC)

\[
\Sigma = \sigma^2 \begin{bmatrix}
1 & 0 & \ldots & 0 \\
0 & 1 & \ldots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \ldots & 1
\end{bmatrix}
\]

In a fixed effect model, we assume:
• variances among observations from different subjects taken at the same time (diagonal elements) are equal (homoscedastic!)
• co-variances between observations taken at different times (off-diagonal elements) are zero (independent!)
Compound Symmetric (CS) Variance-covariance matrix

\[ \Sigma = \sigma^2 \begin{bmatrix} 1 & \rho & \ldots & \rho \\ \rho & 1 & \ldots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \ldots & 1 \end{bmatrix} \]

- Variances among observations from different subjects taken at the same time (diagonal elements) are the equal (homoscedastic!)
- Co-variances between observations taken at different times (off-diagonal elements) are equal

\[ \rightarrow \] Regardless of time between measurements, observations from same subject are equally correlated
Autoregressive order 1 (AR(1)) variance-covariance structure

- Variances among obs from different subjects taken at the same time (diag. elements) are the equal (homoscedastic!)

- Covariances between obs taken at different times (off-diag. elements) are correlated, with constant decay $\rho$

$$
\Sigma = \sigma^2 \begin{bmatrix}
1 & \rho & \rho^2 & \ldots & \rho^{t-1} \\
\rho & 1 & \rho & \ldots & \rho^{t-2} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\rho^{t-1} & \rho^{t-2} & \rho^{t-3} & \ldots & 1
\end{bmatrix}
$$

Correlations decrease as time between obs. increases

Add to lme call: `corr=corAR1()`
R output: mixed model with AR(1) repeated times

> lme.rm.ar1 <- lme(biomass ~ site*time, random = ~1|tree, correlation = corAR1(), data=data.rm)
> anova(lm.rm.ar1)

<table>
<thead>
<tr>
<th>numDF</th>
<th>denDF</th>
<th>F-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>1</td>
<td>126</td>
<td>95.03080</td>
</tr>
<tr>
<td>site</td>
<td>2</td>
<td>42</td>
<td>3.57881</td>
</tr>
<tr>
<td>time</td>
<td>3</td>
<td>126</td>
<td>40.98609</td>
</tr>
<tr>
<td>site:time</td>
<td>6</td>
<td>126</td>
<td>0.54125</td>
</tr>
</tbody>
</table>

> summary(lm.rm.ar1)
Linear mixed-effects model fit by REML
Data: data.rm

AIC      BIC      logLik
1171.441 1236.261 -566.719

Random effects:
Formula: ~1 | tree
(Intercept) Residual
StdDev: 3.492924 1.942569

AIC are very close to those without AR(1): AIC was 1172.1,
Random effects:
Formula: ~1 | rep
(Intercept) Residual
StdDev: 3.448301 2.019734
Correlation Structure: AR(1)

Formula: ~1 | tree

Parameter estimate(s):

\[ \Phi = -0.1811848 \]

Fixed effects: biomass ~ site * time

<table>
<thead>
<tr>
<th></th>
<th>Value</th>
<th>Std.Error</th>
<th>DF</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>0.733993</td>
<td>1.0318303</td>
<td>126</td>
<td>0.711351</td>
<td>0.4779</td>
</tr>
<tr>
<td>siteB</td>
<td>1.017526</td>
<td>1.4592283</td>
<td>42</td>
<td>0.697304</td>
<td>0.4885</td>
</tr>
<tr>
<td>siteC</td>
<td>3.925862</td>
<td>1.4592283</td>
<td>42</td>
<td>2.690368</td>
<td>0.0094</td>
</tr>
<tr>
<td>time2</td>
<td>2.275080</td>
<td>0.7709120</td>
<td>168</td>
<td>2.951154</td>
<td>0.0036</td>
</tr>
<tr>
<td>time3</td>
<td>2.629211</td>
<td>0.6975860</td>
<td>168</td>
<td>3.769013</td>
<td>0.0002</td>
</tr>
<tr>
<td>time4</td>
<td>2.667666</td>
<td>0.7114323</td>
<td>168</td>
<td>3.749712</td>
<td>0.0002</td>
</tr>
<tr>
<td>siteB:time2</td>
<td>0.375345</td>
<td>1.0902341</td>
<td>168</td>
<td>0.344280</td>
<td>0.7311</td>
</tr>
</tbody>
</table>

Effect values are the same. Standard errors are different for times only!

We now have an estimate of rho!

We could try other kinds of correlation matrices and find the one with lowest AIC.
What is an Interaction?

- When there is a significant interaction, the effect of Factor A depends on the level of Factor B, and
- the effect of Factor B depends on the level of Factor A
- For example:
  - We are studying the effects of 3 levels of Site and 4 levels of Time.
  - Neither Site nor Time is significant on its own, but the interaction is significant
  - if we plot means for each factor separately, we may see…:
Example: mean values for each factor separately

- Looking at these graphs, what would you conclude about the effects of Site and/or Time?

But these graphs do NOT tell the whole story... they are hiding something... THE INTERACTION!
When we have no significant interaction, the effect of factor A does not depend on the level of factor B, and vice-versa.

Example: no interaction

![Graph showing no interaction between Site A, Site B, and Site C over time. The lines for each site are parallel, indicating no interaction.]
Example: significant interaction

- Where there is a significant interaction, we cannot make statements about A or B without the context of B or A, respectively.
Example: significant interaction - 2

- Sometimes the significant interaction is not directional; rather, it means that the direction is the same for all levels, while the magnitude is different by level.

To examine the interaction graphically:

interaction.plot(factorA, factorB, responsevar)

Or

plot(lsmeans(lmeobject, ~ factorA:factorB))
HANDS-ON EXERCISE
Question 5-6
What if our experimental/sampling area is not homogenous?

• **Blocking**
  • A block is a group of homogeneous experimental units
  • Blocks are chosen so as to maximize variation among blocks with the aim of minimizing the variation within blocks

• **Reasons for blocking**
  • To remove block-to-block variation from the experimental error (which should increase precision)
  • To allow more uniform treatment comparisons
  • To allow the researcher to sample a wider range of conditions
Randomized complete block designs (RCB)

- Blocks are chosen so that the experimental material within block is homogeneous – and generally we do NOT care to make inferences about blocks (it is a ‘nuisance’ variable)
- Treatments are randomly assigned *within block* (restricted randomization)
Randomized complete block designs – ANOVA table

- We would analyze as a two-way ANOVA – also a GLM

<table>
<thead>
<tr>
<th>Source</th>
<th>Degrees of freedom</th>
<th>Mean Squares</th>
<th>F test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Block</td>
<td>n-1=2</td>
<td>MS_B</td>
<td></td>
</tr>
<tr>
<td>Irrigation</td>
<td>k-1=2</td>
<td>MS_IRR</td>
<td>F_{(2,4)}= MS_IRR/MS_E</td>
</tr>
<tr>
<td>Experimental Error</td>
<td>(k-1)(n-1) = 4</td>
<td>MS_E</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>kn-1 = 8</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Experimental error is partitioned so that we separate out block-to-block variation ➔ lose DOF but (hopefully) decrease Exp.Error
How to fit a mixed model with blocking?

```r
> data.rcb$block <- as.factor(data.rcb$block)
> lme.rcb <- lme(biomass ~ irr, random =~1|block/irr, data=data.rcb)
> anova(lme.rcb)
> summary(lme.rcb)
> plot(lme.rcb)
```

- The function `lme` estimates a linear mixed effects model ($Y \sim X$) using data in the dataframe `data.rcb`.
- Block is not a fixed effect.
- Irrigation types are nested inside each block in the random effect.
- The functions `summary`, `anova`, `plot` are used in the same manner as with the other analyses.
More complex designs

• What if you have more time points than experimental units?
  • E.g. eddy covariance data
    → Time series models, wavelet analyses

• What if you have multiple simultaneous experimental treatments?
  • No restriction on randomization
    → Factorial experiment (can be used with CRD or RCB)
  • Restriction on randomization
    → Split-plot experiment (can be used with CRD or RCB)

• What if you have additional explanatory variables?
  • E.g. soil moisture measured at each site and time
    → Analysis of covariance
Conclusions: Why does design matter?

- Experimental designs have HUGE impacts on how we collect and analyze the data
  - How we set up the experiment controls:
    - What effects are testable
    - What error terms are appropriate
    - The number of ‘true replicates’
Take home messages

• In the design stage, be sure to be very clear about how you intend to collect the data!
  • Draw a picture
  • Make a table
  • Consider ‘confounding factors’, such as aquaria or greenhouse space or other things that might introduce bias
• Using well-studies designs enables us to easily analyze data and construct uncertainty estimates

Now on to hands-on exercises!
Question 5