EXPERIMENTAL DESIGN IN LONG-TERM ECOLOGICAL RESEARCH

Christina Staudhammer, Associate Professor
Department of Biological Sciences
University of Alabama

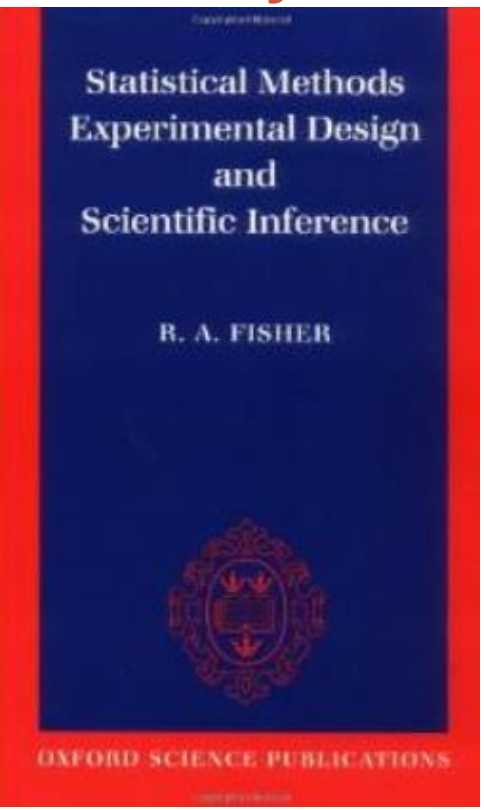
The importance of planning your study design

- 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

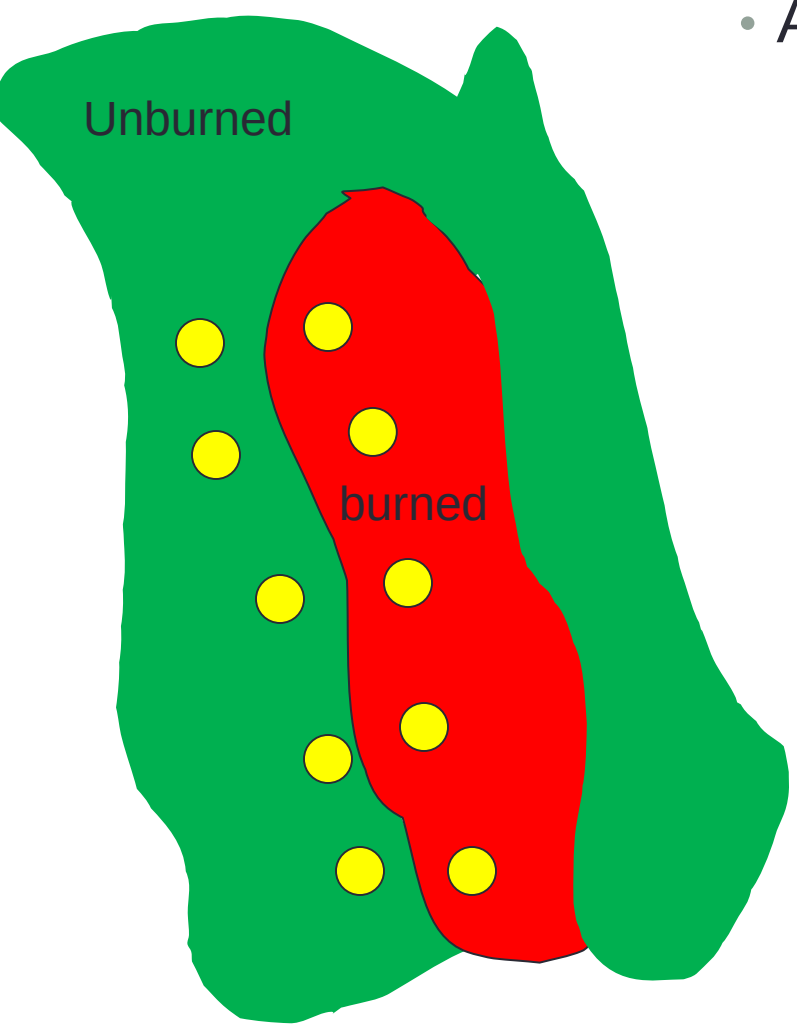
Replication

- 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 <u>one</u> true replicate (in each of unburned and burned areas) with pseudoreplicates, or <u>subsamples</u>

→ We need *multiple fires* in order to appropriately evaluate impact of fire <u>in general</u>; otherwise, our inference is only to <u>this fire</u>

What is meant by “experimental design”?

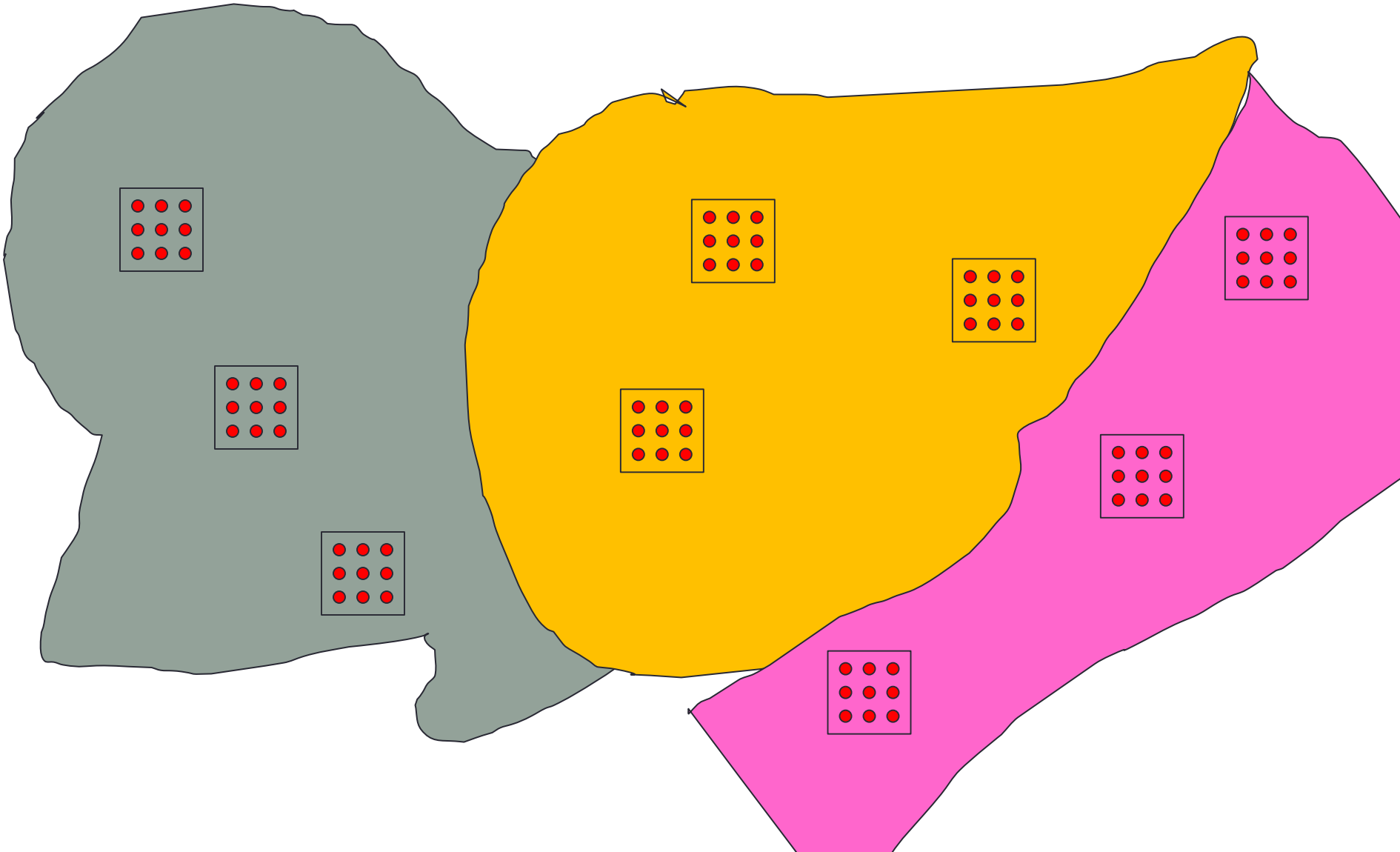
Controls how we apply *treatments* to observational units, or select data from different populations

→ Controls how we analyze the data

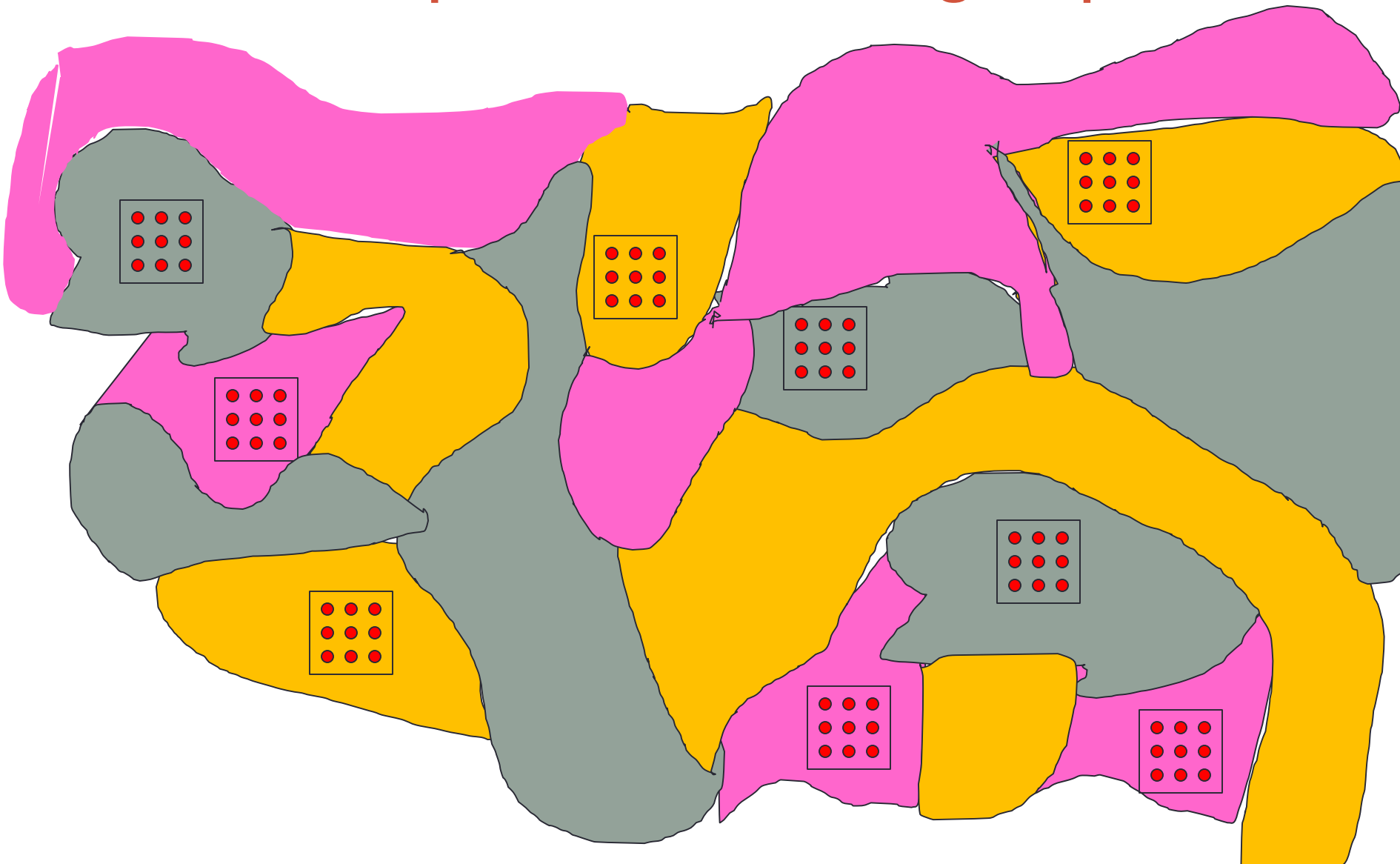
- <u>is often intimately related to the sampling design under which the data was collected</u>

- 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
  - <u>OR</u>, 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”? - 2

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

→ Controls how we analyze the data

- <u>is often intimately related to the sampling design under which the data was collected</u>

- 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
  - <u>OR</u>, 3 frogs with each of the moisture regimes in each of block #1, #2, #3

One experimental design option

A	A	A
A	A	A
A	A	A

B	B	B
B	B	B
B	B	B

C	C	C
C	C	C
C	C	C

Another experimental design option

A	B	C
A	B	C
A	B	C

C	A	B
C	A	B
C	A	B

B	C	A
B	C	A
B	C	A

Another experimental design option

A	B	C
B	C	A
C	A	B

C	A	C
A	C	B
B	B	A

C	B	A
A	A	C
B	C	B

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 <u>moderate</u> *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

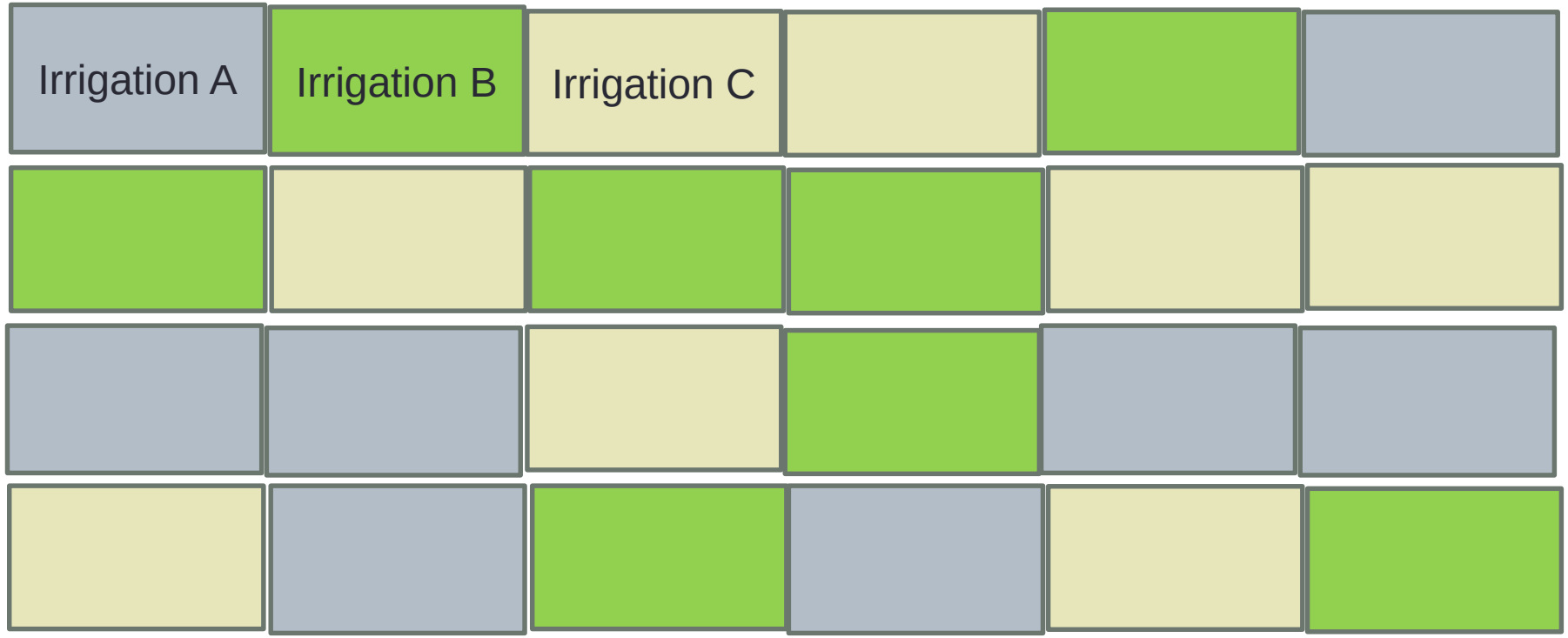
Some types of experimental designs

- 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



Completely randomized design – Analysis of Variance (ANOVA) table

- We would analyze this as a simple one-way ANOVA, or could (equivalently) use regression techniques
- Either is termed a *General Linear Model (GLM)*

Source	Degrees of freedom (DoF)	Mean Squares	F test
Irrigation	k-1=2	MS_{IRR}	$F_{(2,21)}= MS_{IRR}/MS_E$
Experimental Error	k(n-1) = 21	MS_E	
Total	kn-1 = 23		

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

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 = $\Sigma n_i - 1$

Fitting CRD models in 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

```
> lm.irr <-lm(biomass ~ irrig, data=data.irr)
> anova(lm.irr)
Analysis of Variance Table

Response: biomass
           Df  Sum Sq Mean Sq F value    Pr(>F)
Irrig       2  2021.0  1010.5  40.374    0.0003***
Residuals  21  525.26    25.0
---
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

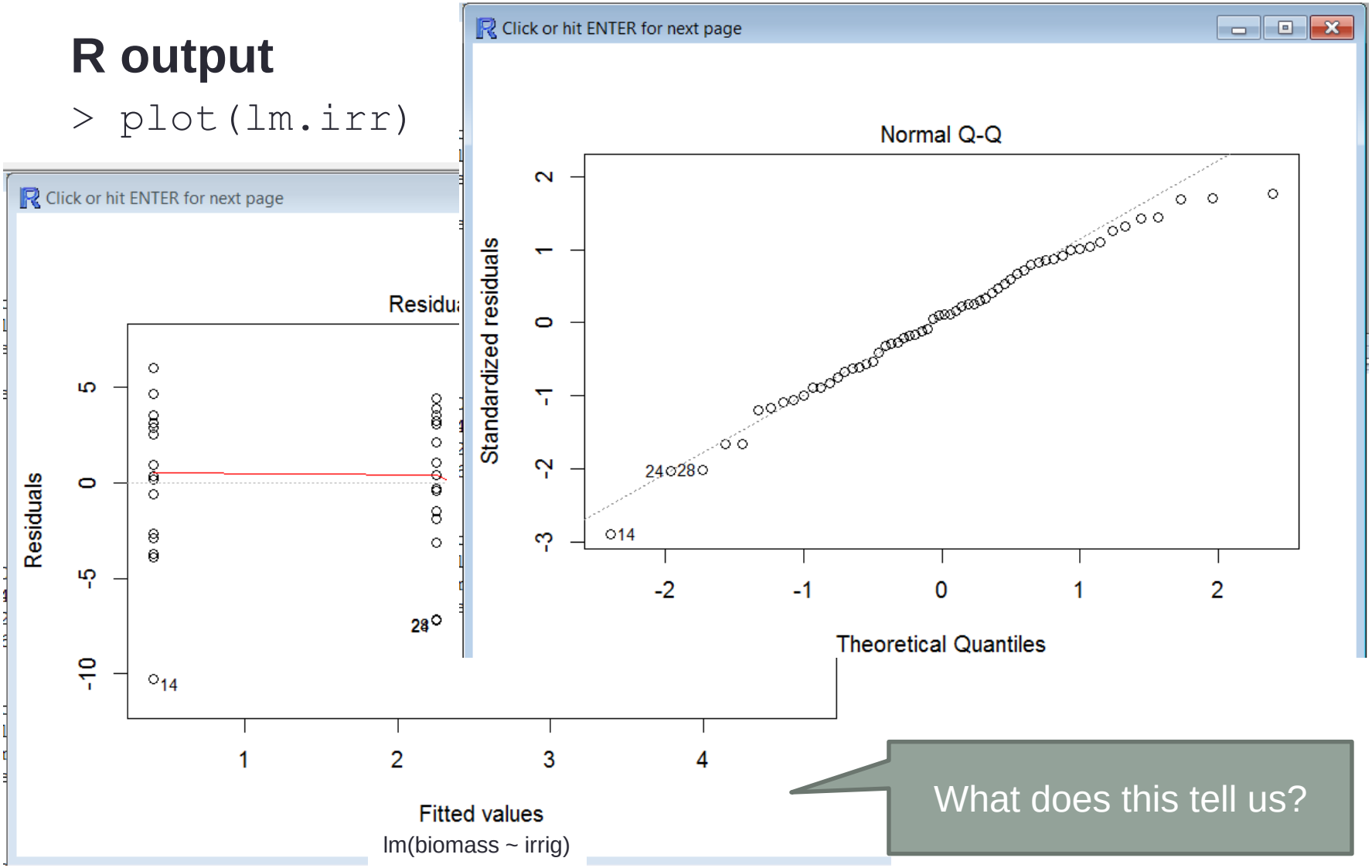
```
> summary(lm.irr)

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)   2.0210     0.4111   4.926   0.0011 **
irrig.B      12.1991     0.6022  20.257  4.1e-08 ***
Irrig.C      17.9911     0.6022  29.874  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

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



What does this tell us?

Fitting CRD models in R – marginal means

R output

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

$lsmeans
 irrig     lsmean       SE df lower.CL upper.CL
 A         2.0216 1.762224 21 -1.43252  5.47452
 B        14.2201 1.762224 21 10.76658 17.67362
 C        20.0121 1.762224 21 16.55855 23.46562

Confidence level used: 0.95

$contrasts
 contrast  estimate      SE df t.ratio  p.value
 A - B    -12.19911 2.49216 21  -4.895  7.67e-05
 A - C    -17.89117 2.49216 21  -7.219  4.10e-07
 B - C     -5.79252 2.49216 21  -2.324  0.030225

P value adjustment: tukey method for comparing a
estimates
```

What does this tell us?

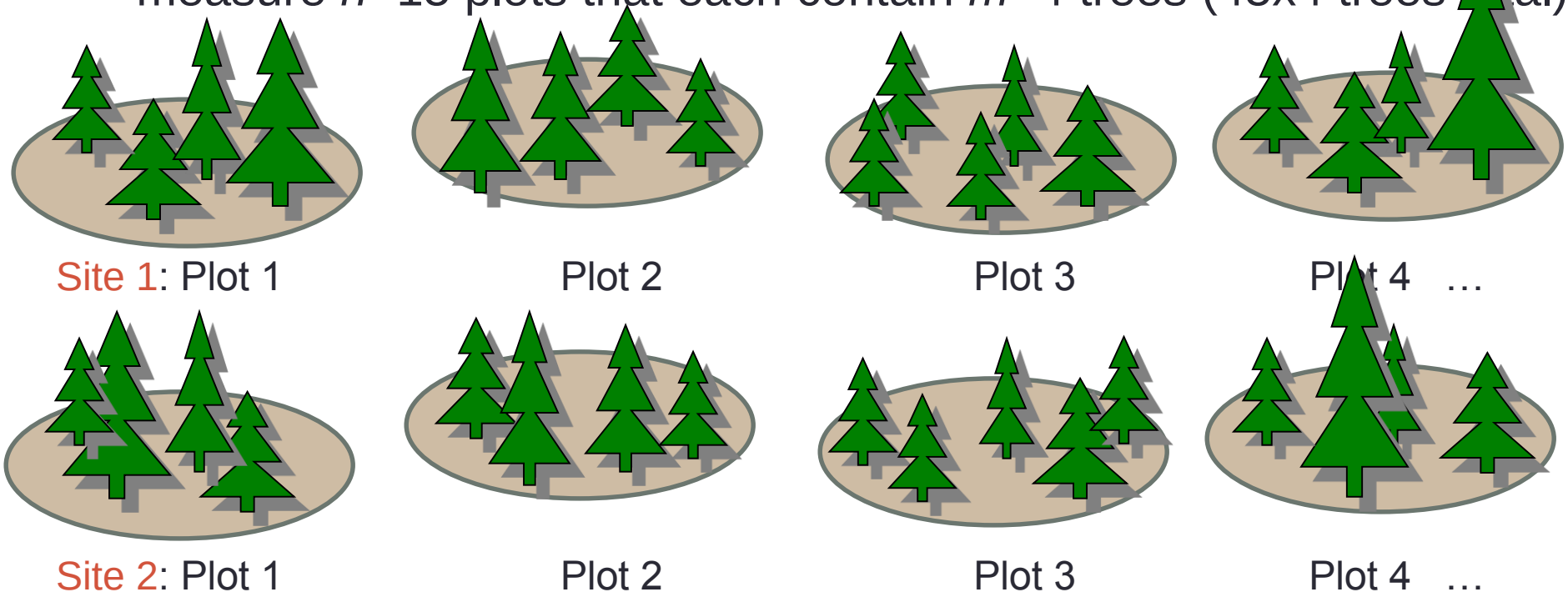
There is a 95% probability that the true mean understory biomass under irrigation C is between 16.56 and 23.47 kg

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?

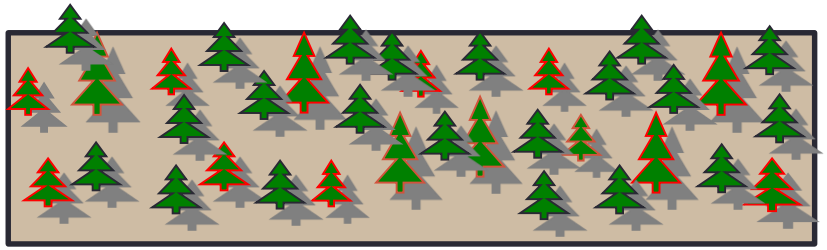
- In many situations, researchers collect data on multiple elements in the same fixed area plot
  - E.g., models of biomass as a function of *k*=3 site qualities: we measure *n*=15 plots that each contain *m*=4 trees (45x4 trees total)



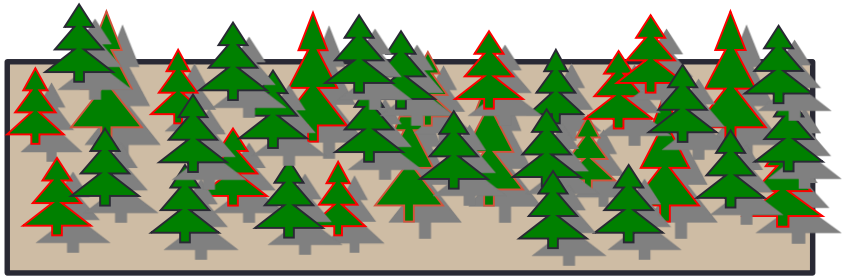
What happens if we measure the same element repeatedly over time?

- 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

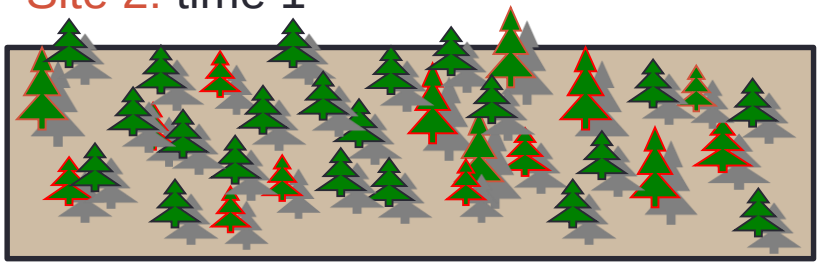


time 2

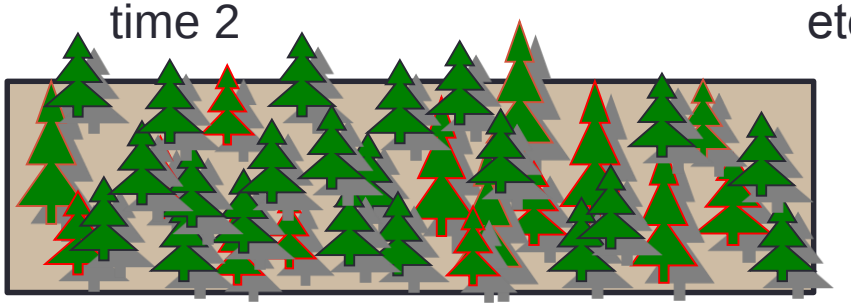


etc…

Site 2: time 1



time 2



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

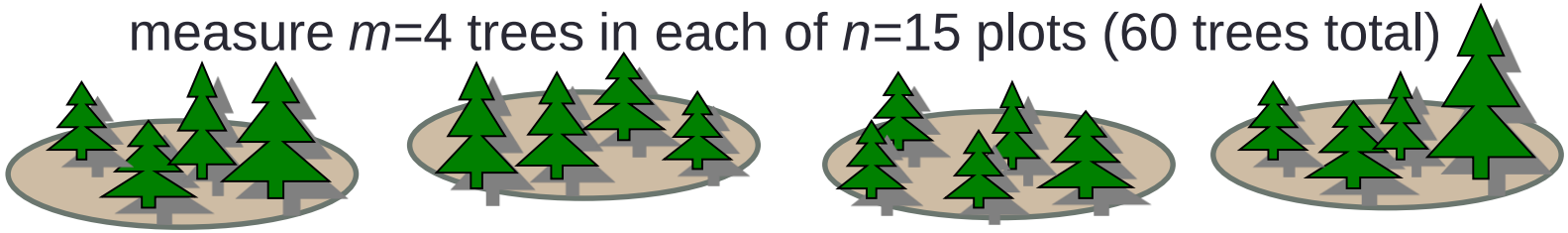
- We then want to develop models for these elements
  - For tree-level data collected in fixed area plots
    - trees within the same plot are NOT independent; they are likely more alike than those in different plots
  - For data collected on the same exact trees over time
    - Measurements on the same 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

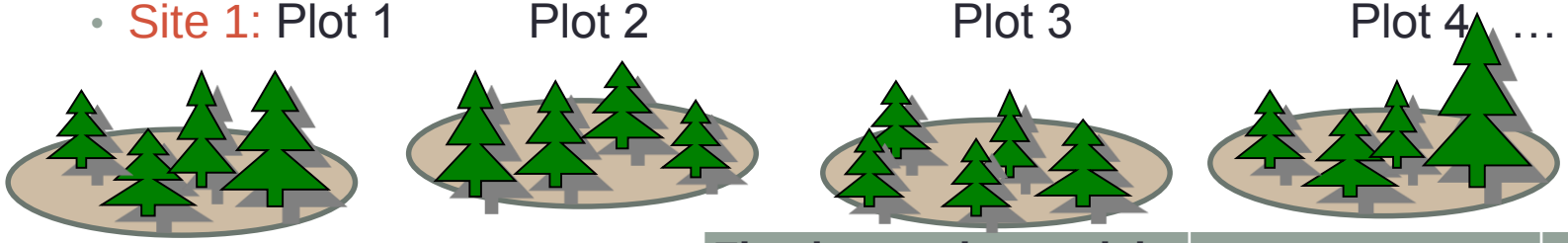
- Knowledge of these correlations can be used to formulate the correct experimental error in our models
- Moreover, this knowledge can be useful in better understanding our data!

Mixed models for multiple measurements per experimental unit (e.g., fixed area plots)

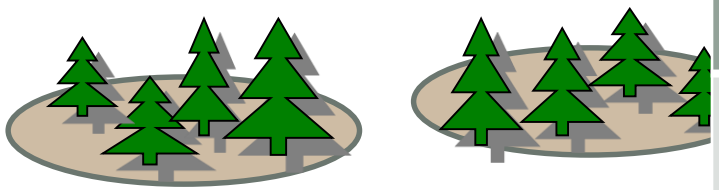
- E.g., models of biomass as a function of *k*=3 site qualities, where we measure *m*=4 trees in each of *n*=15 plots (60 trees total)



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



Site 2: Plot 1 Plot 2



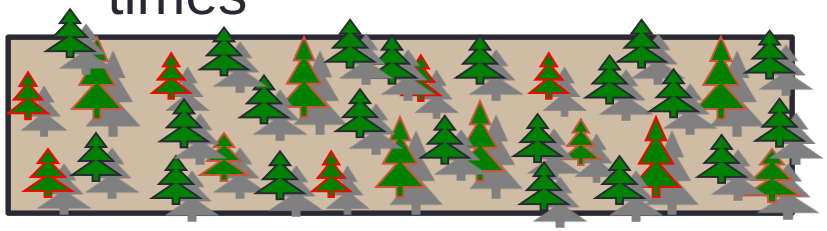
Site 3: Plot 1 Plot 2

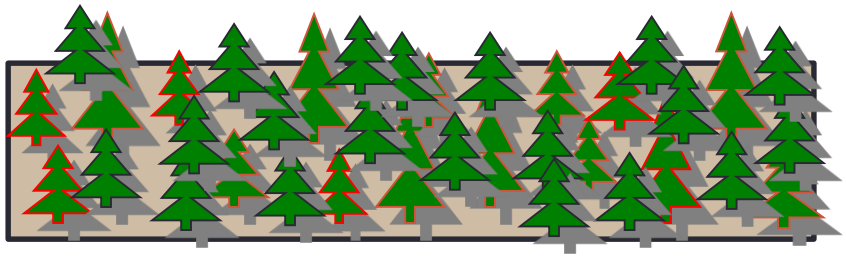
In the case of *k*=3 sites with *m*=4 trees per *n*=15 plots, each plot is a "subject"

Fixed area plot model Source	Degrees of freedom	F test
Site	k-1=2	$F_{(2,42)}= MS_S/MS_E$
Experimental Error	k(n-1) = 42	
Within plot error	nk(m-1)=135	
Total	knm-1 = 179	

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





Site 1: time 1
etc…

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

Repeated times model **Source**	**Degrees of freedom**	**F test**
Site	k-1=2	MS_S/MS_E
Experimental Error	k(n-1) = 42	
time	m-1=3	MS_t/MS_W
Site x time	(k-1)(m-1)=6	MS_{Sxt}/MS_W
Within tree error	k(n-1)(m-1)=126	
Total	knm-1 = 179	

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

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*…

Repeated times model **Source**	**Degrees of freedom**	**F test**
Site	k-1=2	MS_S/MS_E
Experimental Error	k(n-1) = 42	
time	m-1=3	MS_t/MS_W
Site x time	(k-1)(m-1)=6	MS_{Sxt}/MS_W
Within tree error	k(n-1)(m-1)=126	
Total	knm-1 = 179	

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)

```
> 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~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

```
> 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

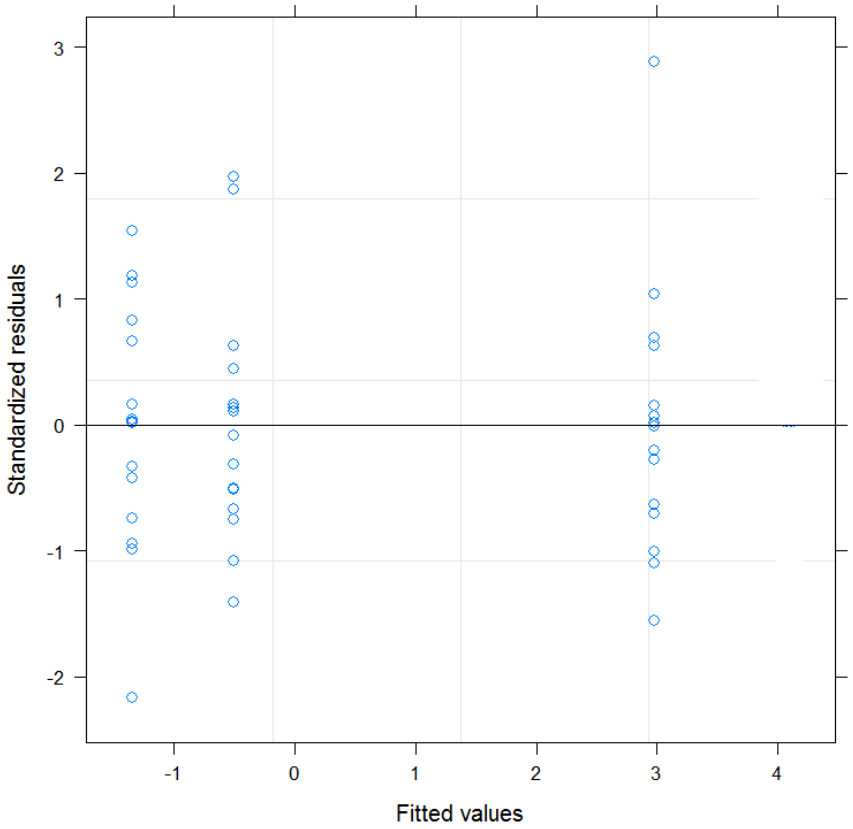
Estimates of the variance among plots versus within plots

These tests are for the effect versus the base (A)

R output: mixed model w

```
 Correlation:
           (Intr) quality.B
qualityB  -0.707
qualityC  -0.707     0.500

Standardized Within-Group Residual~
       Min          Q1         Med
-1.8651688 -0.6058632 -0.0108787

Number of Observations: 60
Number of Groups: 12

> plot(lme.sq)
```

This is <u>not</u> the correlation between the variables. It is the expected correlation of the model coefficients. This <u>might</u> 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

```
> 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~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)
            numDF denDF  F-value p-value
(Intercept)     1   126 92.46865  <.0001
site            2    42  3.59848  0.0189
time            3   126 35.55504  <.0001
site:time       6   126  0.50806  0.8673
```

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

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

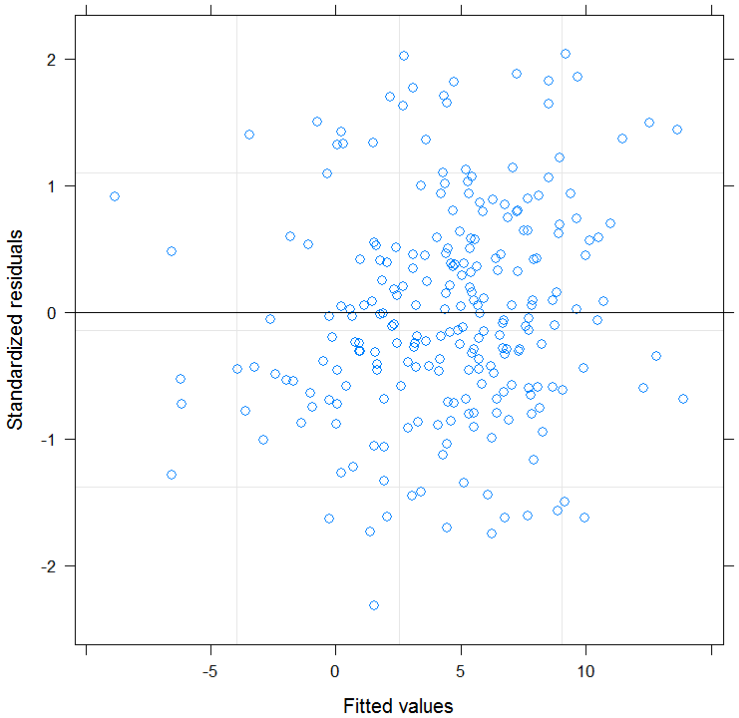
Estimates of the variance among trees versus within trees

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

R output: mixed model with repeated times - 2

```
Correlation:
                (Intr)  siteB  siteC  time2
siteB           -0.623
siteC           -0.623  0.200
time2           -0.357  0.253  0.253
time3           -0.357  0.253  0.253  0.2
time4           -0.357  0.253  0.253
siteB:time2      0.253 -0.357 -0.179 -0.623
siteC:time2      0.253 -0.179 -0.357 -0.623 -
siteB:time3      0.253 -0.357 -0.179 -0.354 -
siteC:time3      0.253 -0.179 -0.357 -0.354 -
...
Standardized Within-Group Residuals:
        Min          Q1         Med
-2.31861551 -0.58095354 -0.05834473  0.53737

Number of Observations: 180
Number of Groups: 15

> plot(lm.rm)
```

This is <u>not</u> the correlation between the variables. It is the expected correlation of the model coefficients. 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



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

```
> lme.rm <-lme(biomass ~ site*time, random =~time|tree, data=data.rm)
> anova(lme.rm)
             numDF denDF  F-value p-value
(Intercept)    1   126 42.96620  <.0001
site           2    42  3.44695  0.0226
time           3   126 44.31872  <.0001
site:time      6   126  0.63711  0.7642
> summary(lme.rm)
Linear mixed-effects model fit by REML
 Data: data.a.rm
       AIC      BIC    logLik
  1174.107 1266.222 -560.0537
...
                  Value Std.Error  DF  t-value p-value
(Intercept)    0.733993 0.9333719 126 0.786389  0.4327
siteB          1.017526 1.3199872  42 0.770860  0.4440
siteC          3.925862 1.3199872  42 2.974166  0.0043
time2          2.275080 0.8674875 126 2.622608  0.0095
time3          2.629211 0.6337064 126 4.148941  0.0001
time4          2.667666 0.6950092 126 3.838318  0.0002
```

With only random intercept:

```
             numDF denDF F-val
(Intercept)    1   126 92.46
site           2    42  3.59
time           3   126 35.55
site:time      6   126  0.50
```

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

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

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)

$$\Sigma = \sigma^2 \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix}$$

In a fixed effect model, we assume:

•variances among observations from different subjects taken at the same time (diagonal elements) are equal (<u>homoscedastic!</u>)

•co-variances between observations taken at different times (off-diagonal elements) are zero (<u>independent!</u>)

Compound Symmetric (CS) Variance-covariance matrix

$$\Sigma = \sigma^2 \begin{bmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & \dots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \dots & 1 \end{bmatrix}$$

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

→ 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 ρ

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

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)
            numDF denDF  F-value p-value
(Intercept)     1   126 95.03080  <.0001
site            2    42  3.57881  0.0194
time            3   126 40.98609  <.0001
site:time       6   126  0.54125  0.8428

> summary(lm.rm.ar1)
Linear mixed-effects model fit by REML
 Data: data.rm
       AIC      BIC    logLik
  1171.441 1236.2611 -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
```

R output: mixed model with AR(1) repeated times - 2

```
Correlation Structure: AR(1)
 Formula: ~1 | tree
 Parameter estimate(s):
       Phi
-0.1811848

Fixed effects: biomass ~ site * time
                   Value Std.Error  DF  t-value p-value
(Intercept)     0.733993 1.0318303 126 0.711351  0.4779
siteB           1.017526 1.4592283  42 0.697304  0.4885
siteC           3.925862 1.4592283  42 2.690368  0.0094
time2           2.275080 0.7709120 168 2.951154  0.0036
time3           2.629211 0.6975860 168 3.769013  0.0002
time4           2.667666 0.7114323 168 3.749712  0.0002
siteB:time2     0.375345 1.0902341 168 0.344280  0.7311

  ...
```

We now have an estimate of rho!

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

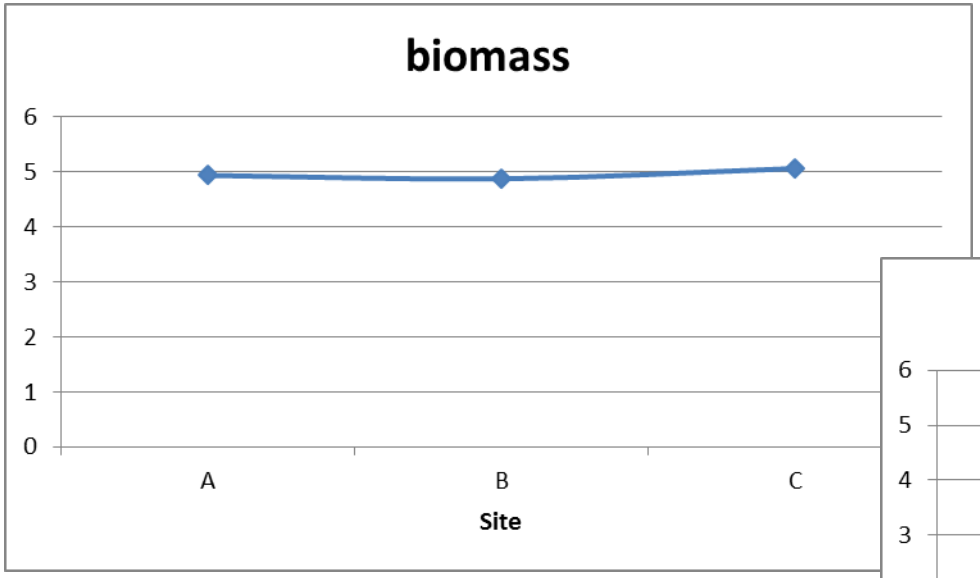
We could try other kinds of correlation matrices and find the one with lowest AIC

What is an Interaction?

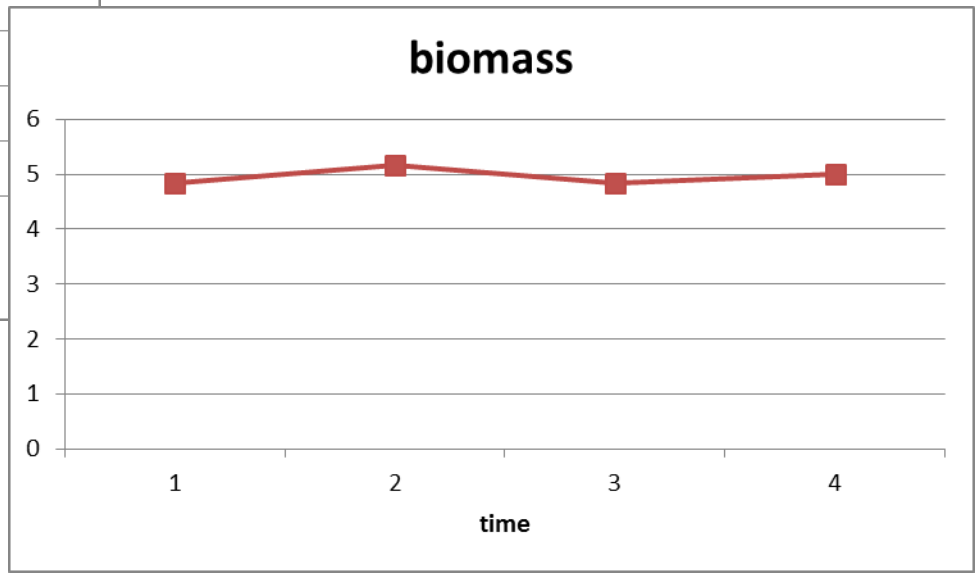
- When there is a significant interaction, the effect of Factor A depends on the level of Factor B, and
- the effect of Factor B depends on the level of Factor A
- For example:
  - We are studying the effects of 3 levels of Site *and* 4 levels of Time.
  - Neither Site nor Time is significant on its 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



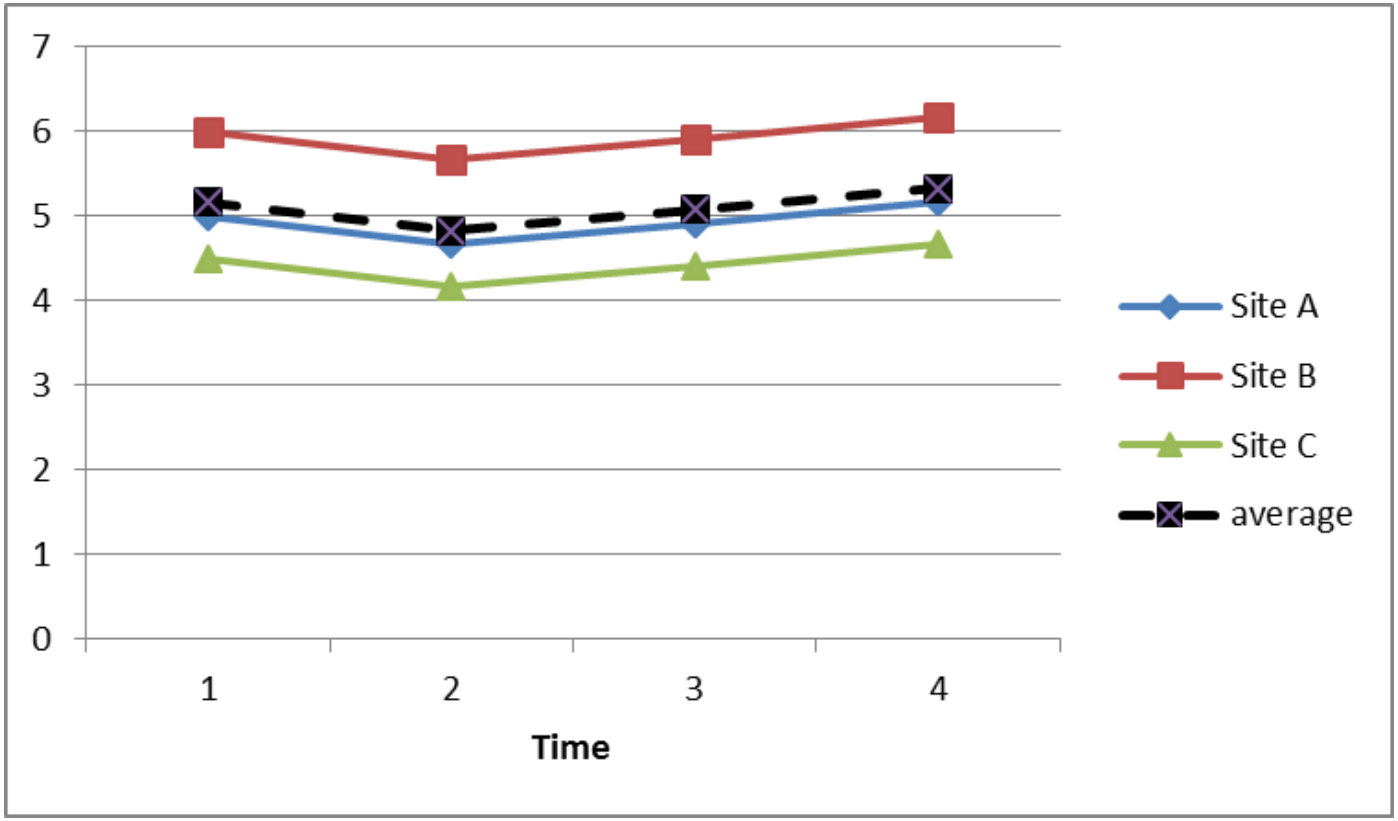
and/or Time?



But these graphs do NOT tell the whole story… they are hiding something… THE INTERACTION!

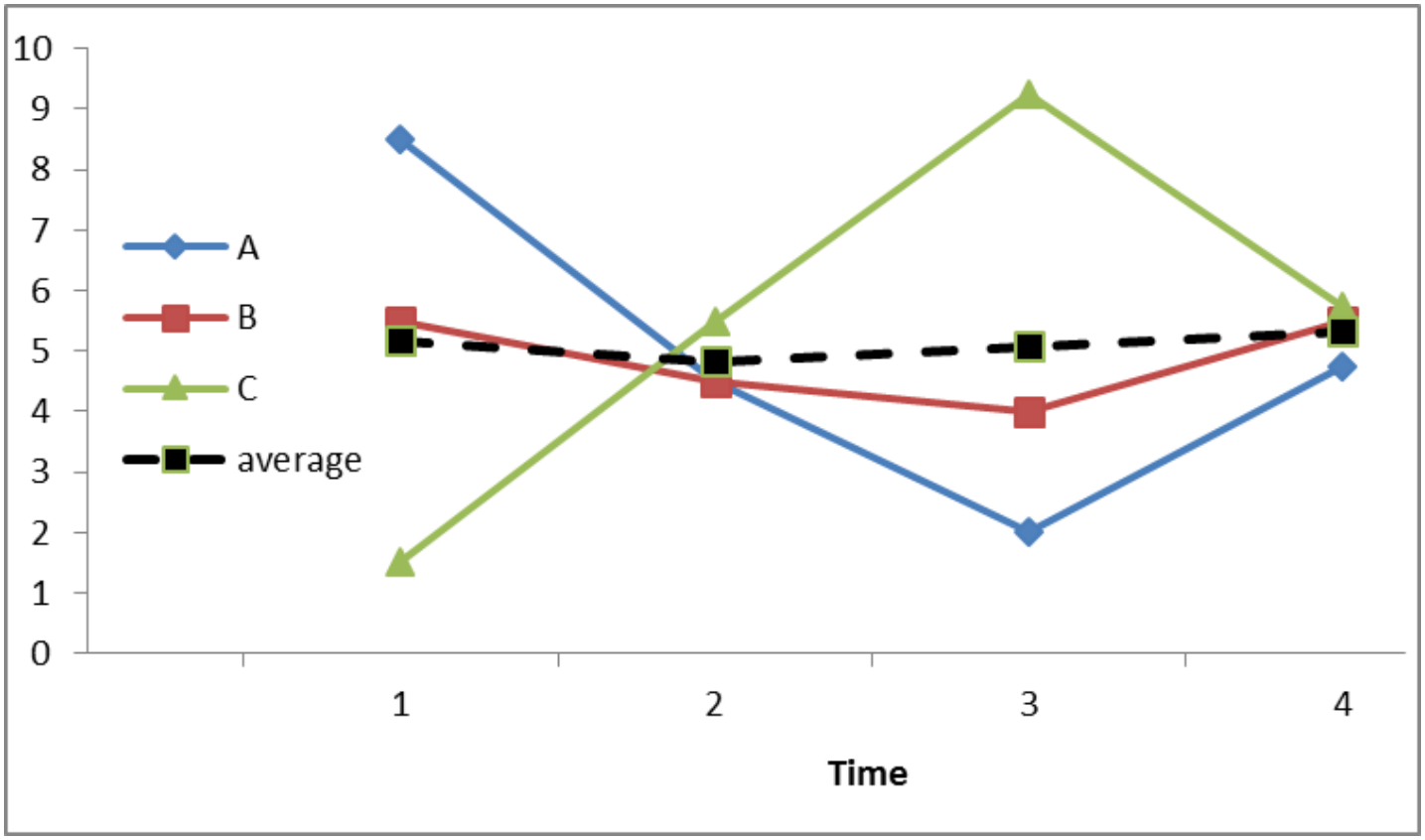
Example: no interaction

- When we have no significant interaction, the effect of factor A does not depend on the level of factor B, and vice-versa



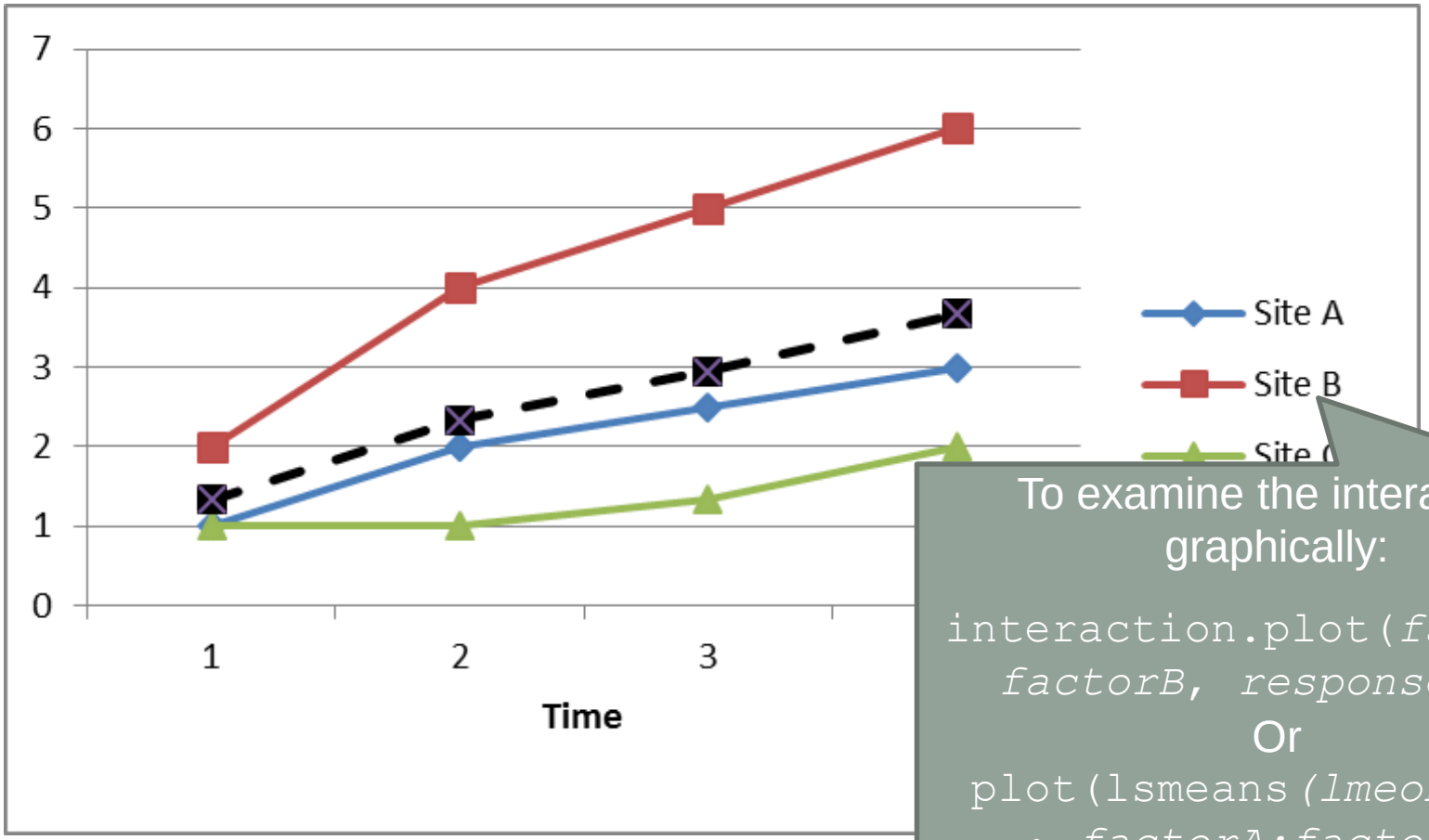
Example: significant interaction

- 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:

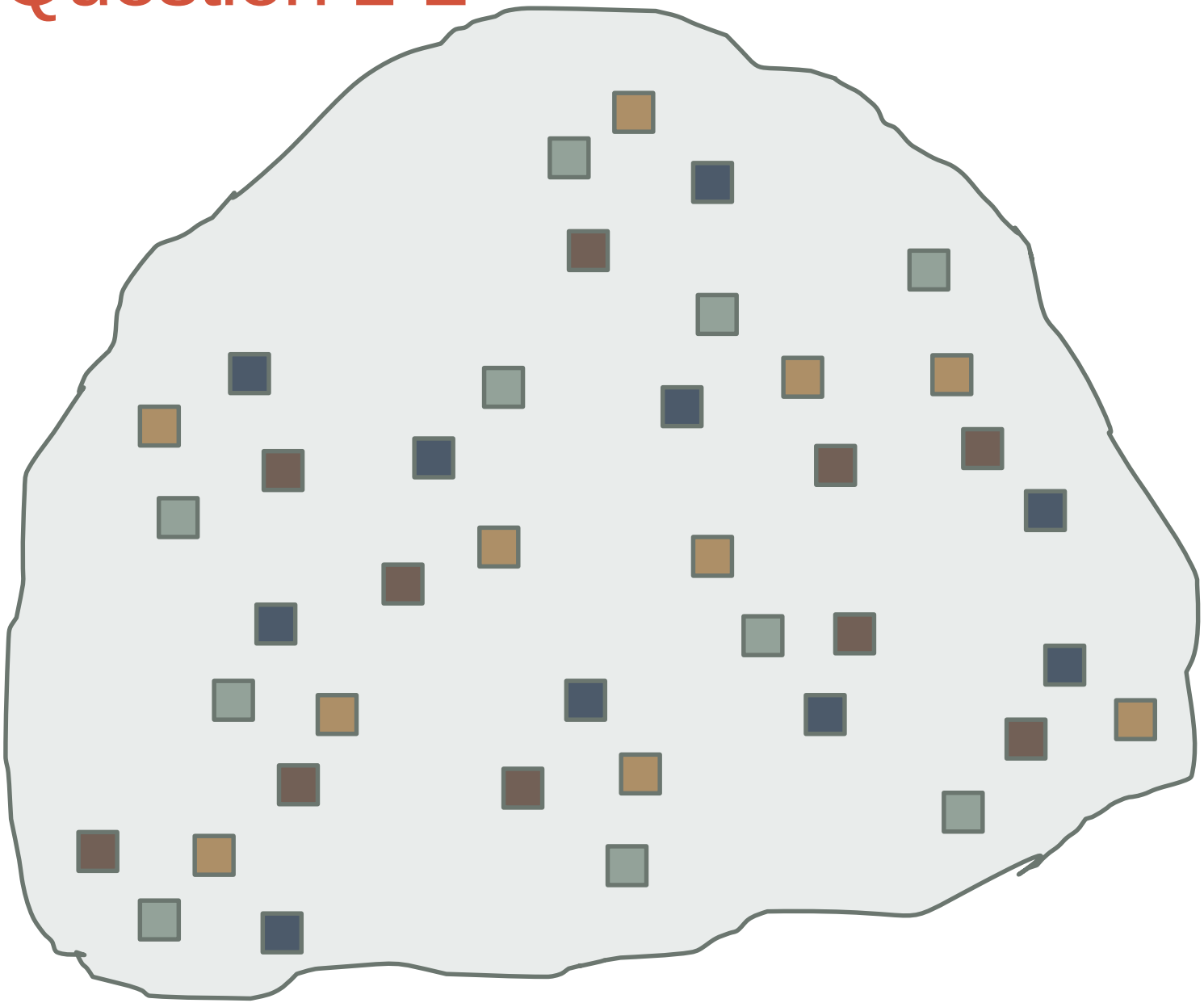
```
interaction.plot(factorA, factorB, responsevar)
```

Or

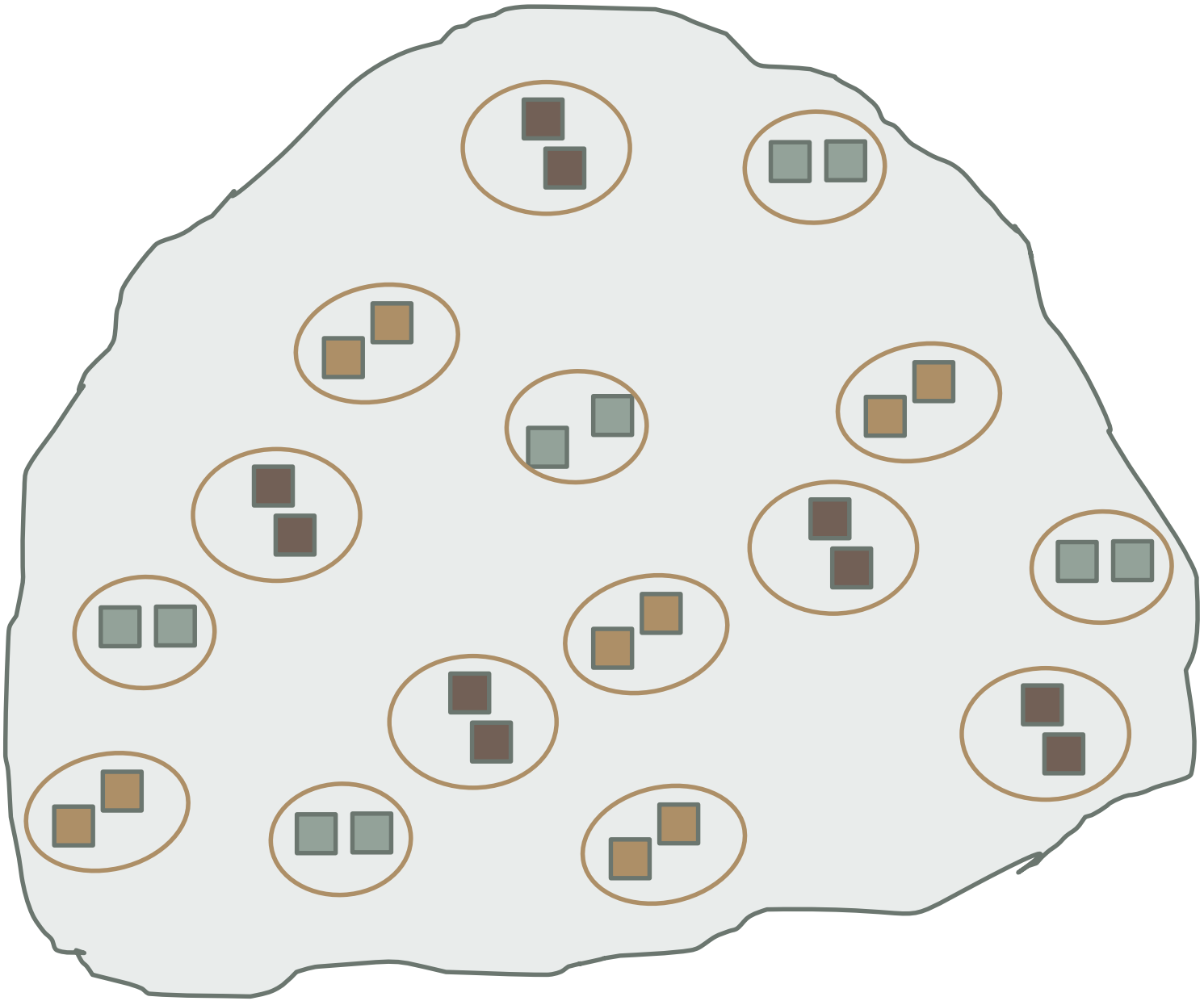
```
plot(lsmeans(lmeobject, ~ factorA:factorB))
```

HANDS-ON EXERCISE

Question 1-2



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)

BLOCK 1	BLOCK 2	BLOCK 3
Irrigation I 1	Irrigation III 2	Irrigation II 3
Irrigation II 1	Irrigation II 2	Irrigation III 3
Irrigation I 1	Irrigation III 2	Irrigation I 3

Randomized complete block designs – ANOVA table

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

Source	Degrees of freedom	Mean Squares	F test
Block	n-1=2	MS_B	
Irrigation	k-1=2	MS_{IRR}	$F_{(2,4)}= MS_{IRR}/MS_E$
Experimental Error	(k-1)(n-1) = 4	MS_E	
Total	kn-1 = 8		

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?

```
> 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~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