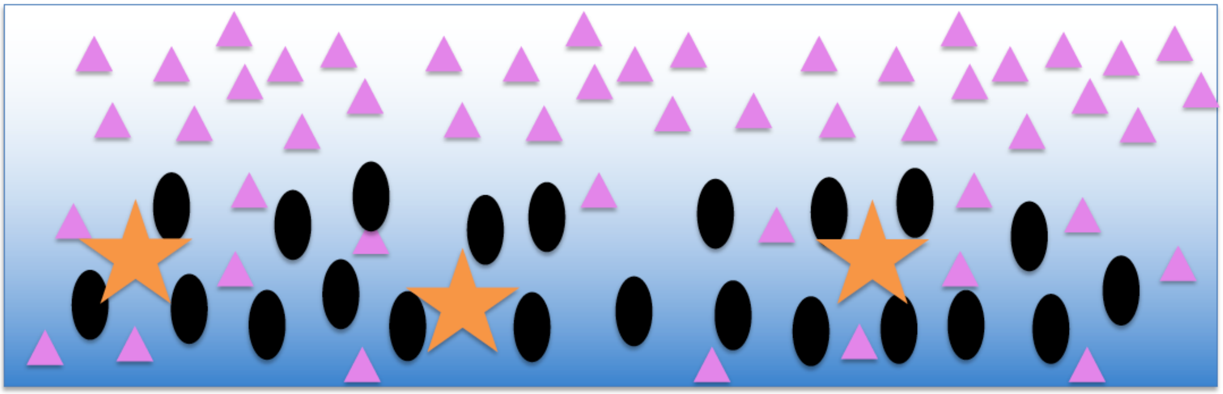
Lecture 10. Design and Inference.

Now that we've learned about linear models and generalized linear models, we're going to spend some time thinking about how we can draw scientific inferences from these models. In other words, so far we've focused on how to choose probability distributions and a deterministic model structure that is appropriate for your data. But an equally important component of data analysis is designing your study in a way that allows you to draw valid conclusions, and conducting appropriate analyses of your fitted model so that you can highlight the real signal and avoid spurious noise.

In this lecture we will focus on basic principles of study design and how those relate to the structure of statistical models. There are widely accepted principles of design for controlled experiments, and there are also common challenges and strategies for analyzing observational data where a controlled experiment was not possible (e.g., surveys). One important fact is that we use the same statistical models to analyze controlled experiments and non-experimental surveys. However, *the way we interpret those models differs depending on whether the data come from a controlled experiment or not.*

Controlled experiments are generally considered the best way to draw conclusions about causality in nature. Why is that? Let's use an imaginary case study to illustrate. We are interested in the structure of rocky intertidal communities, and we observe that barnacles (pink triangles) are an abundant component of the community, and that there are more barnacles in the high intertidal zone than the low intertidal zone:

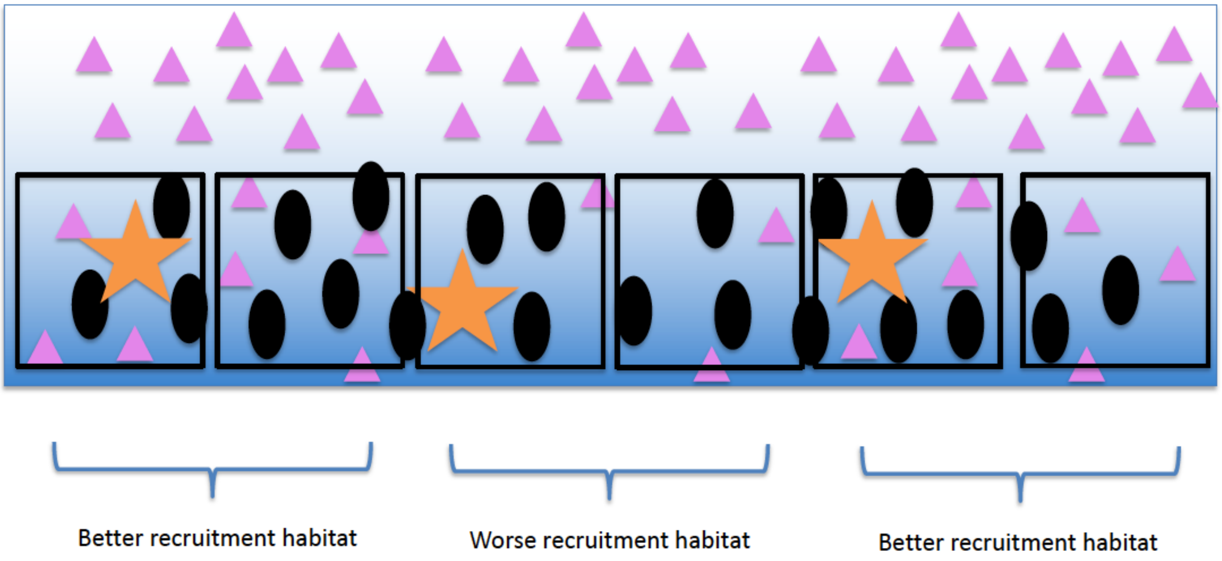


We also observe that mussels (black ovals), which may compete for living space (or food or oxygen), are more abundant in the low intertidal, and sea stars (orange), which predate upon barnacles, are also more abundant in the low intertidal. Based on these observations we hypothesize that low abundance of barnacles in the low intertidal could be due to (1) competition from mussels, (2) predation by sea stars, or (3) physiological effects of greater immersion in seawater.

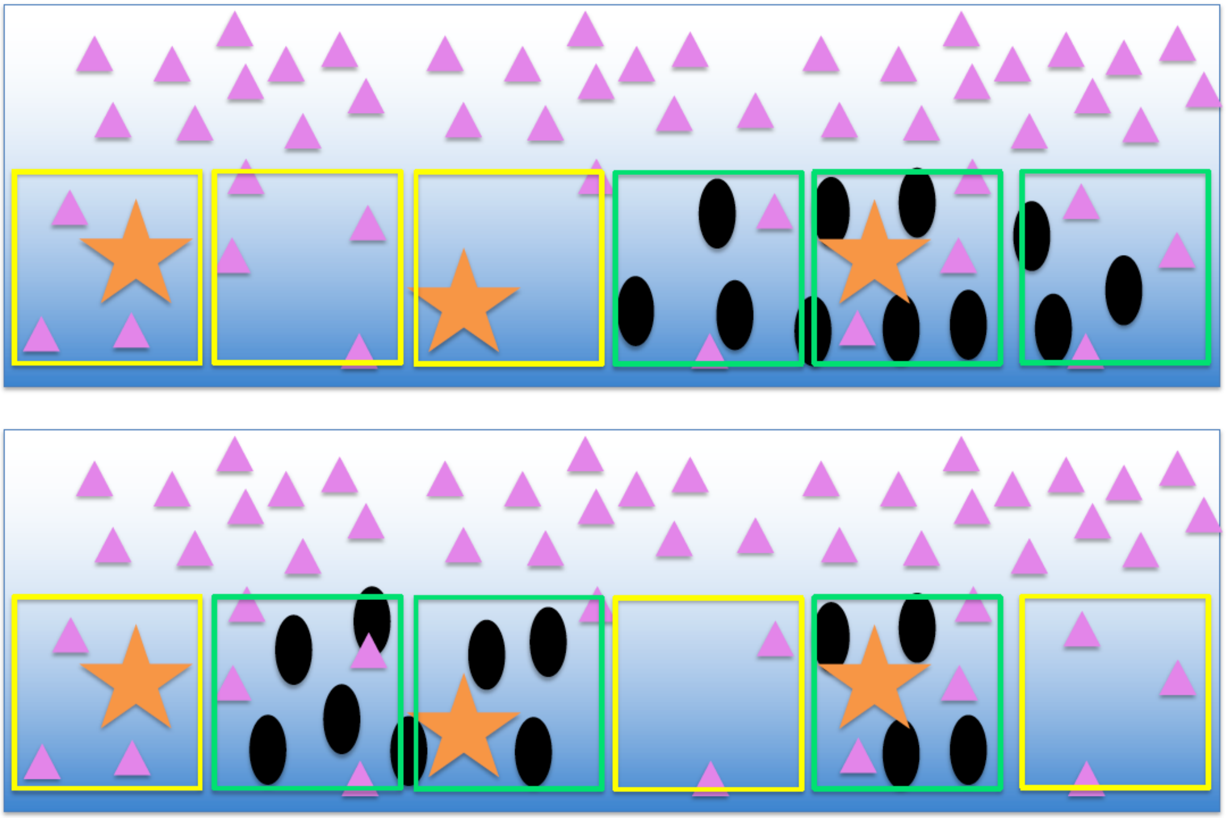
To test hypothesis #1 we can remove mussels from the low intertidal and see if the abundance of barnacles increases. We also want to establish experimental controls where we don't remove mussels, to allow us to distinguish the effect of mussels from natural variation in barnacle abundance over time. The next question is: what is the best way to arrange experimental removals and controls?

Randomization

A common and powerful approach to this problem is randomization. We can outline a certain number of experimental plots, and randomly choose which plots receive experimental mussel removals and which plots are left alone as controls. The reason randomization is a useful strategy is because it *minimizes systematic error*. Here 'systematic error' would be an apparent effect of mussels that is due to some other natural variable that covaries with our experimental treatment. For example, maybe there is small-scale variation in topography or some other factor that affects barnacle recruitment, with the result that settlement of barnacle larvae is not uniform among our experimental plots:

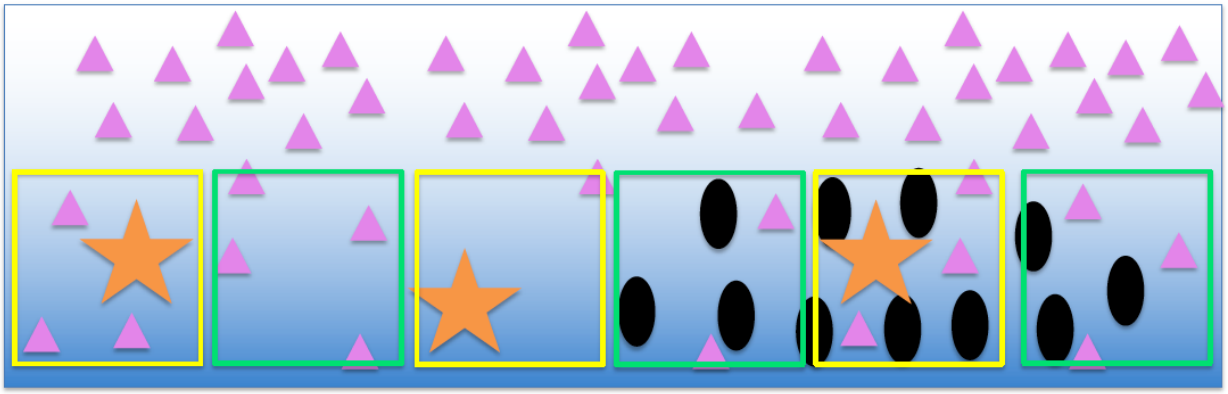

Better recruitment habitat
Worse recruitment habitat
Better recruitment habitat

We don't know in advance how these covarying factors work, so we don't know which plots will receive more barnacle settlement. By randomizing the placement of the mussel removals, we reduce the chance that the experimental treatment will inadvertently align with natural variation in barnacle settlement:



For example, in the figure above, if we placed all the mussel removals together (yellow) and all the controls together (green), this increases the chance that a horizontal gradient in barnacle settlement will be confounded with the experimental treatment (top row). In contrast, if we randomize placement of the treatments (bottom row), this will break up such an association, unless we are unlucky.

Although randomization is a good default strategy, it is not the only option. In this example we have relatively little replication (an issue discussed below), and randomization may, by chance, lead to a clumping of the treatment plots and control plots. Instead we could systematically alternate treatments and controls, like in the top row below:



In this case the treatment would be confounded with natural variation in barnacle settlement only if there happens to be systematic, periodic variation in settlement that aligns with our design (unlikely but not impossible).

Blocking

An additional common strategy for dealing with natural variation in an experiment is blocking. If our experiment takes place over some spatial extent, and we would like to quantify and experimental effect while controlling for natural variation over space, we can define a certain number of 'blocks' in advance, and then randomly assign treatment and control plots within those blocks. The following figure from Hurlbert 1984, "Pseudoreplication and the Design of Ecological Field Experiments", illustrates a randomized block design as option A-2:

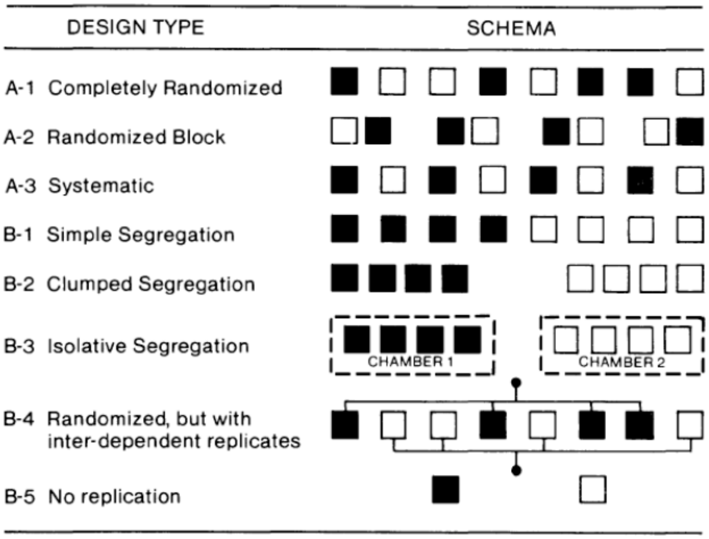
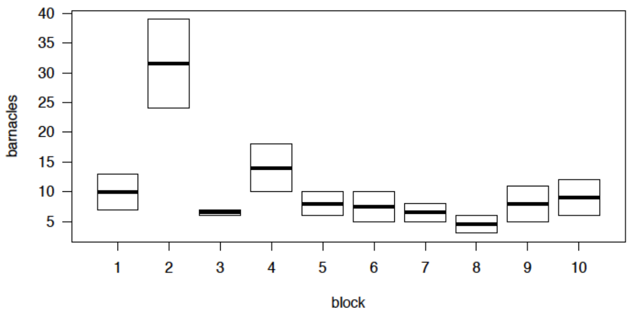


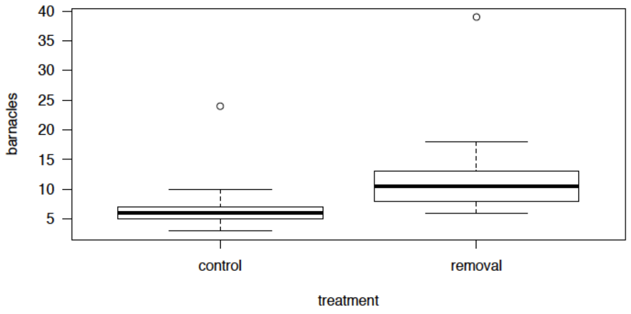
FIG. 1. Schematic representation of various acceptable modes (A) of interspersing the replicates (boxes) of two treatments (shaded, unshaded) and various ways (B) in which the principle of interspersion can be violated.

I highly recommend you read this paper – although it is decades old, the problems of design it discusses are still common, and apply to all experiments (not just ecological field experiments). Option A-3, 'systematic', is the acceptable non-randomized option I described above. Options B-1 through B-5 are different designs that are very susceptible to confounding natural variation. Note that blocks do not have to be spatial groups. They could be temporal groups, where experimental manipulations are repeated over time. They could be other kinds of groups: maybe you apply the same treatment to multiple species or clades of organisms, in which case the species or clades are like blocks in the analysis.

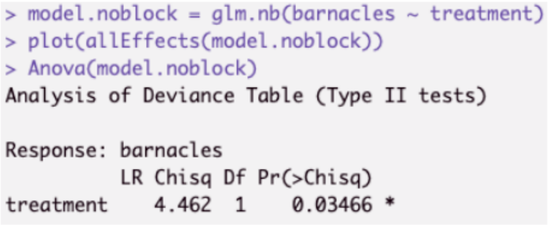
Blocking accomplishes two things at the same time: (1) it *increases precision* when estimating treatment effects, and (2) it allows for spatial/temporal/other *non-independence to be controlled for*. Blocking increases precision because we can include both the treatment effect and the block effect in our statistical model. Even though we don't know what environmental factors cause the block effect, by including block in our model we account for that variation, which reduces unexplained variation, making the treatment effect more distinguishable from unexplained 'noise'. Blocking controls for source of non-independence because that non-independence (e.g., spatial variation in barnacle settlement) is now a part of our model. That means that the remaining residual variation is more likely to be independently distributed, which is an assumption needed to make calculations of p-values (and other kinds of model inference) valid.

Let's look at a simulated example of how blocking increases precision. Imagine we have 10 spatial blocks in the intertidal, and within each block we have a certain number of removals and controls, randomly assigned. The data look like this:

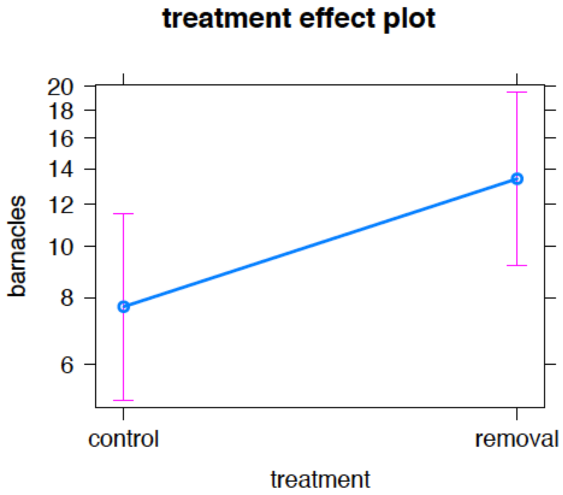


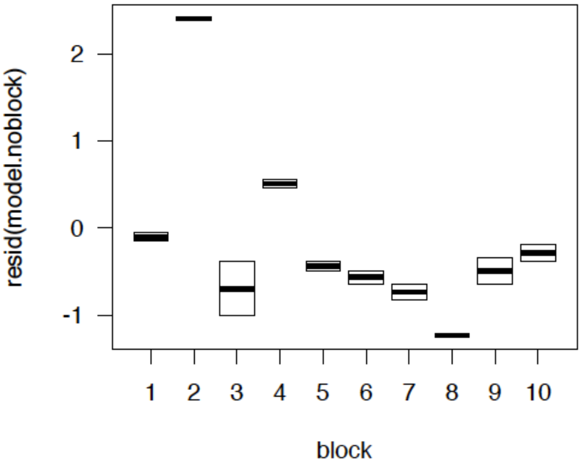


Note that there is substantial variation in barnacle abundance among blocks. There are appears to be a positive effect of mussel removal on barnacles, but our ability to clearly detect this effect will depend on whether we control for blocks in our model. If we make a simple model with no blocks (assuming counts of barnacles follow a negative binomial distribution), a likelihood ratio test yields p = 0.035:



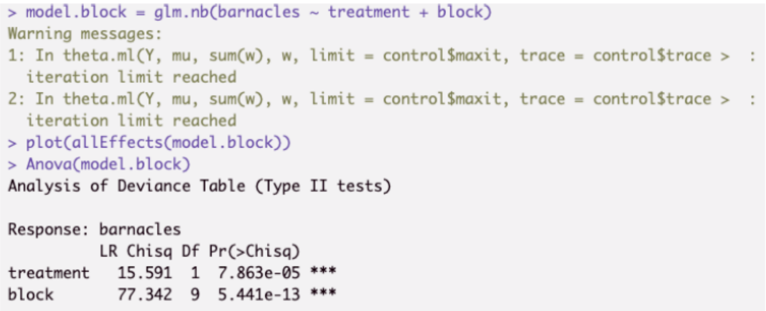
```
> model.noblock = glm.nb(barnacles ~ treatment)
> plot(allEffects(model.noblock))
> Anova(model.noblock)
Analysis of Deviance Table (Type II tests)

Response: barnacles
          LR Chisq Df Pr(>Chisq)
treatment    4.462  1    0.03466 *
```



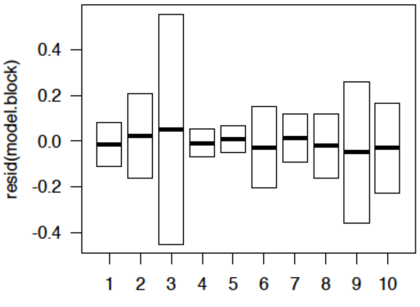


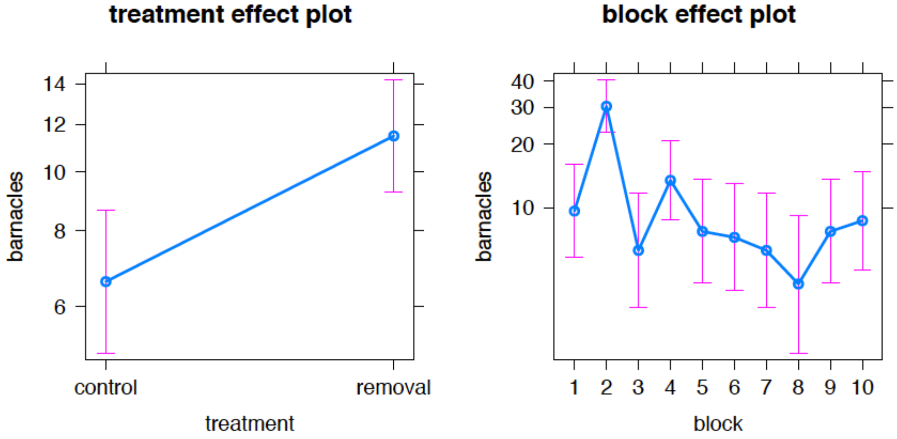
Also note that the residuals vary by block, which means our residuals are not independently distributed.

When we include block in the model, our test yields p = 7.9e-5, which is much smaller:



```
> model.block = glm.nb(barnacles ~ treatment + block)
Warning messages:
1: In theta.ml(Y, mu, sum(w), w, limit = control$maxit, trace = control$trace >  :
  iteration limit reached
2: In theta.ml(Y, mu, sum(w), w, limit = control$maxit, trace = control$trace >  :
  iteration limit reached
> plot(allEffects(model.block))
> Anova(model.block)
Analysis of Deviance Table (Type II tests)

Response: barnacles
          LR Chisq Df Pr(>Chisq)
treatment   15.591  1  7.863e-05 ***
block       77.342  9  5.441e-13 ***
```





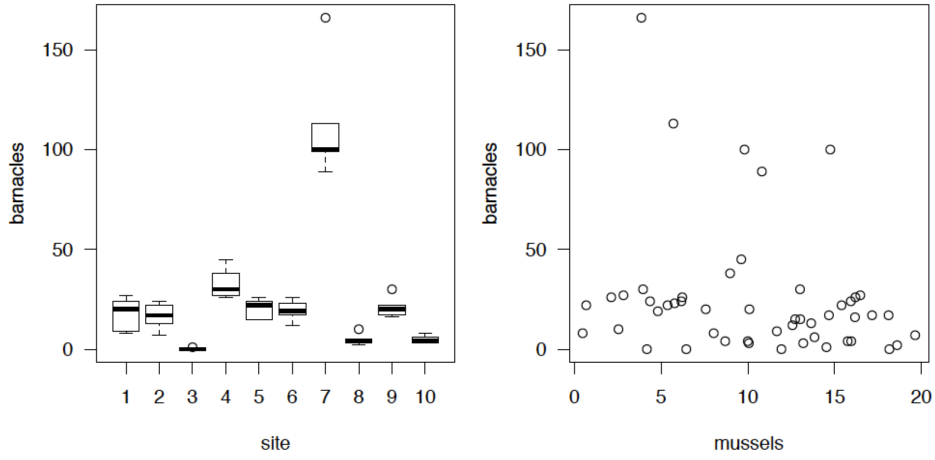
Note that the confidence intervals on our treatment and control are substantially smaller – this is a visualization of the increase in precision.

‘Blocking’ for non-experimental data

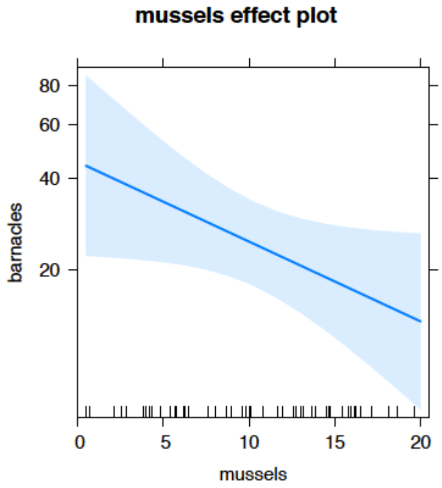
The principles that make blocking useful for experiments also apply to surveys and other observational datasets. Specifically, if we are interested in testing a relationship using observational data, we will estimate the relationship more precisely if we can control for spatial/temporal/other variation in the data. At the same time, controlling for those other sources of variation reduces non-independence of the residuals, making our statistical inferences more valid.

For example, imagine we have intertidal survey data and we want to ask whether barnacles are less abundant when mussels are more abundant (due to putative competition). We go to 10 different sites and lay down 5 quadrats per site to count barnacle and mussel abundance. If barnacle abundance varies among sites for some other reason (e.g., settlement driven by physical processes), and the effect of mussels on barnacles is mostly within sites, then including ‘site’ as a model predictor will account for site-level variation and make the correlation with mussels much clearer.

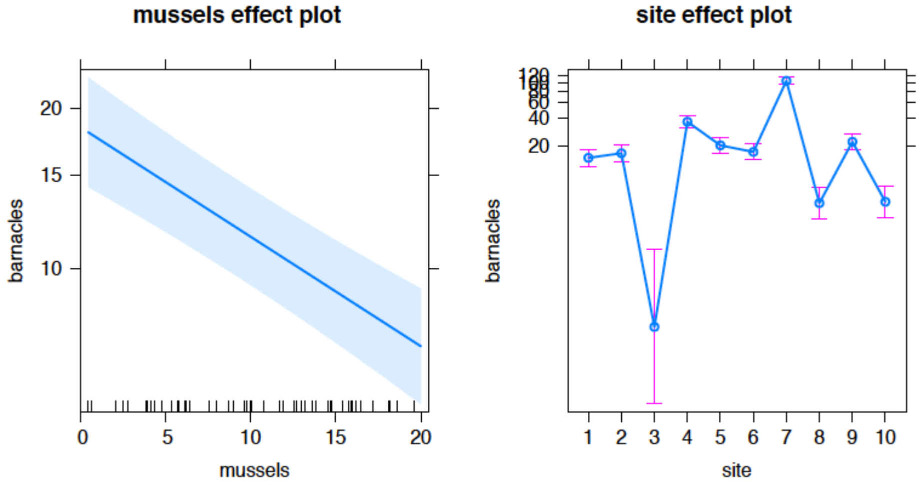
Here is simulated raw data:



Here is the effect of mussels without site included in the model:



And here is the effect of mussels with site included:



Note the increase in precision when site is included. In addition, if barnacles vary mostly between sites, rather than within sites, then it is essential to include site in the model to account for this lack of independence, which could greatly affect our apparent amount of replication.

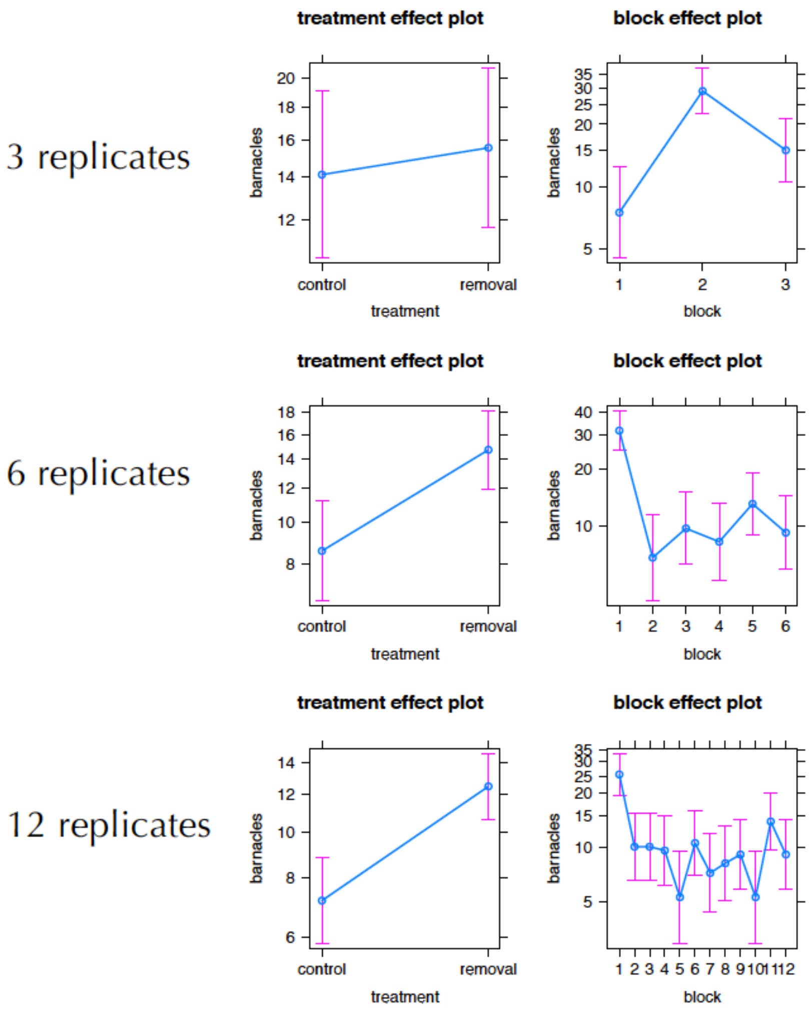
Replication

Hopefully replication of experiments seems natural based on your previous training, but it is still very common for scientists to perform unreplicated comparisons (intentionally or not), so let's try to think clearly about what replication means and what it accomplishes. Replication accomplishes three things simultaneously: (1) Verifies repeatability, (2) Enhances precision of estimated effects, (3) Allows estimation of variability in effects. These three features apply both to replication *within a study*, e.g. replicated experimental manipulations, and replication *across studies*, e.g. when different researchers perform the same experiment. We will focus here on replication within a study, because that affects how you design your study and analyze the data.

Hopefully #1 is fairly obvious – if you repeat a process and get the same result, it helps convince you that it is a robust phenomenon and not some fluke. #2 and #3 have to do with quantification and are more subtle. It can help to think of statistical inference as trying to *detect a signal*, through the noise of other processes happening simultaneously, plus the noise created by errors in our measurements. The ability to detect the signal, which I usually refer to as an 'effect', depends on:

- **Effect size**: this is the true signal that we are trying to measure. It can be defined in different ways, such as the difference between experimental treatments, or the magnitude of the slope for a continuous relationship.
- **Number of replicates or sample size**: For controlled experiments the number of replicates is clearly defined, while for survey data the effective amount of replication or sample size can be trickier and depends on the particular comparison (more on this below, and when we get to mixed models).

If we are studying a relationship that has a large effect size, it will be easier to detect this signal relative to the noise. Hopefully this is intuitive – if vaccination reduces the probability of infection by 50%, this will be a more robust signal than if if reduces infection by 5%. Let's look at an example of how number of replicates changes the ability to detect an effect:

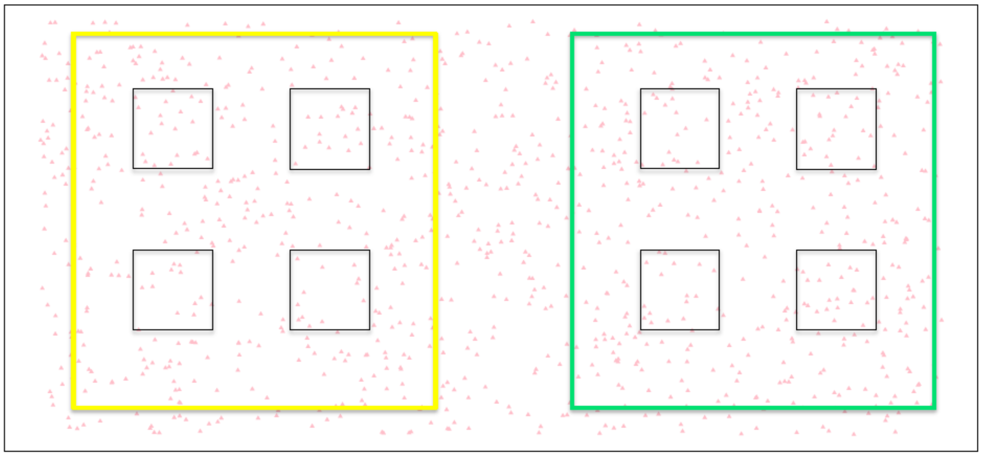


This is more simulated barnacle data, where I use the same true effect size in each case, but change the number of replicates from 3 to 6 to 12. Note that the confidence intervals around the treatment/control means decrease as replication increases – this is the increase in precision, where we become more confident what the true means are. Also note that the point estimates for the treatment and control groups change across the examples: with less replication, it is more likely that our estimated treatment effect is far from the true effect, due to other sources of variability in the data.

An important consequence of the role of sample size / replication is that *very small effects can be highly statistically significant (small p-value), with a very large sample size*. For this reason is is important to always report effect sizes (e.g. with effects plots), and to discuss the biological relevance of these effects, which could be small even if they are 'significant'.

Units of replication, pseudoreplication, scales of inference

When we talk about replication, what, exactly, is replicated? For experiments, it is the *experimental unit*. This is the unit at the scale where the treatment is applied (or not). In our barnacle example, imagine that we just made two very large plots, and applied a mussel removal to one of the plots:



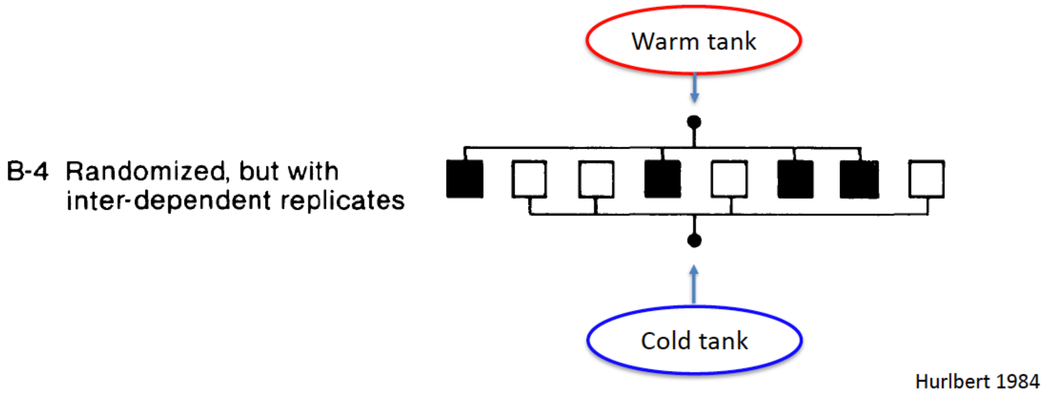
Because the plots are so big, the number of barnacles per plot (pink triangles) is very big, and we don't want to count them all. So we subsample the plots with four subplots (black squares), and count all the barnacles in those. Subsampling in this way can make sampling more efficient, but *these subsamples are not replicates of the experimental treatment*. If you treat them as replicates in a statistical model, you will not have the correct inferences – your p-values and confidence intervals will be too small, or if you are using a model selection approach like AIC, your AIC differences will be too large. This is called *pseudoreplication*. This may seem obvious, but I can assure you that pseudoreplication is a recurrent problem, and in the early days of experimental ecology there were some influential papers that did exactly what I have diagrammed here. How should you deal with this data instead? Take the average of the subsamples for each experimental unit, and use those in the analyses.

Although in this instance the pseudoreplication problem may seem clear, there are situations where it is tricky. And in fact, the validity of the numbers returned by your model will depend on what kinds of inferences you want to draw. If you only want to ask "Are the yellow and the green plots different", then you can treat the subsamples as replicates of yellow vs. green. However, you can't confidently attribute the yellow vs. green difference in that model to the mussel removal, because you only have one replicate of the mussel removal.

In an ideal world we could study all questions with controlled experiments, but the reality is that there are many important questions for which it is not tractable or not ethical to perform replicated experiments. In some cases experimental manipulation is possible, but not with a high level of replication at the scale necessary to answer the question. This has been referred to as the replication/scale tradeoff. For example, there have been a number of iron addition experiments in the ocean, where a large amount of dissolved iron was added to a large patch (many square kilometers), and a nearby control patch was followed for comparison. A large phytoplankton bloom in the iron addition patch is consistent with iron limitation of this region of the ocean. The scale of these experiments is impressive, and they are some of the few, true in situ field experiments in oceanography. At the same time, they are technically unreplicated. However, additional experiments have been peformed in other locations, allowing scientists to ask whether the evidence for iron limitation is repeatable. It should also be noted that there are many other lines of evidence supporting iron limitation of primary production in certain ocean regions, such as more easily replicated bottle incubations, patterns of dissolved nutrient concentrations, etc.

A tricky case of pseudoreplication

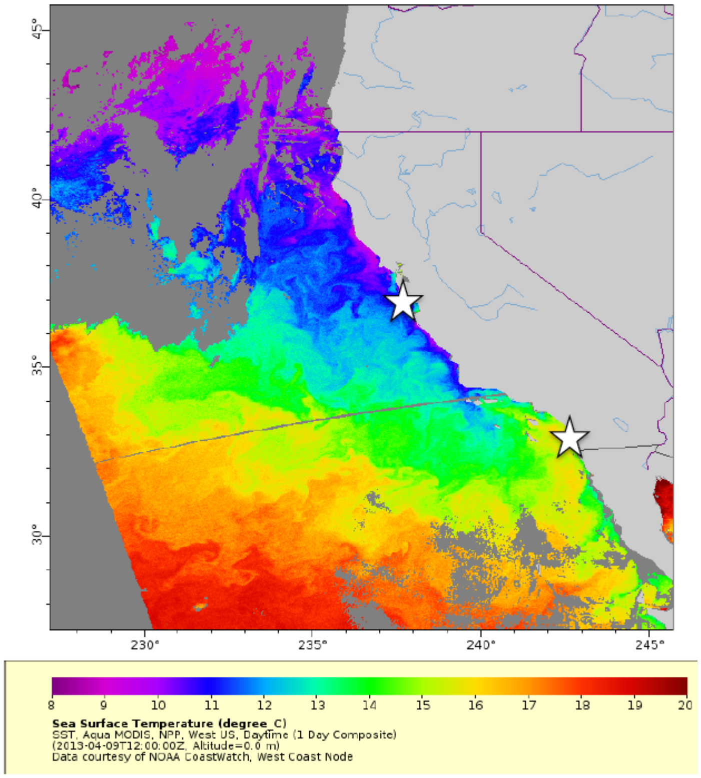
In Hurlbert's schematic of different experimental designs, case B-4 illustrates how it can be challenging to clearly define pseudoreplication:



In this case there are 8 experimental tanks randomly assigned a 'warm' or 'cold' treatment. However, all of the warm tanks are supplied with warm water from a single source, and all of the cold tanks are supplied with cold water from a single source. So the question is: does this experiment have four replicates of warm and cold, or just one replicate each? If nothing goes awry, and the only different between the warm source tank and the cold source tank is temperature, then it seems defensible to think of the 8 experimental tanks as the real experimental units. However, if there is some other cause of variation between the source tanks, such as other aspects of water quality that affect the study organisms (maybe the warm tank gets contaminated with an unpleasant microbe, and the cold tank does not), then this effect is going to propagate to all four tanks fed by the source tank. And it will cause a difference between the two treatments that appears to be due to temperature, but is not. This is why we have to think about the scale at which the experimental manipulation is applied, and define that as the scale at which the experiment can be replicated (or not). And, this is why Hurlbert considers this design an example of pseudoreplication.

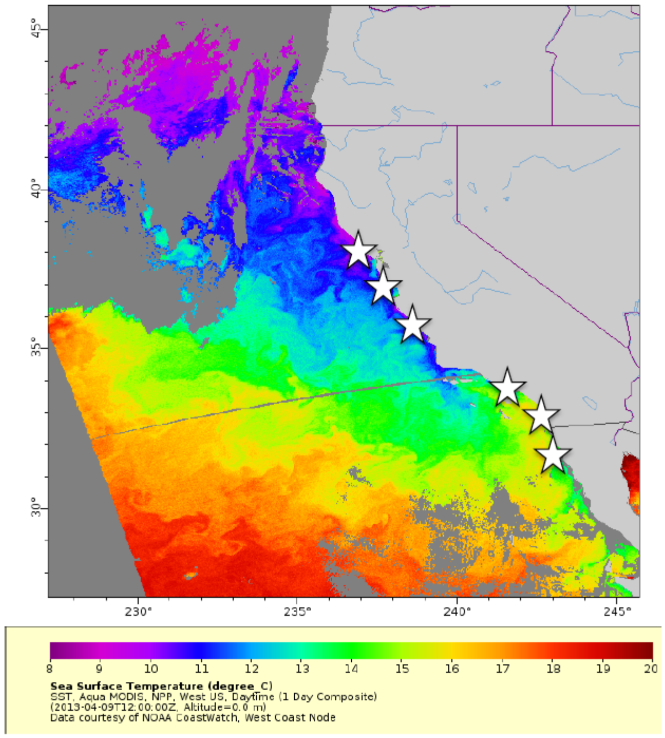
Thinking about replication in observational datasets

Although there are many situations where we are asking scientific questions without controlled experiments, we can still use the logic of experimental design to create survey designs and statistical models that test questions as rigorously as possible. For example, imagine that are measuring mussel and barnacle abundances in 20 quadrats at each of two sites (the two stars):


45°
40°
35°
30°
230°
235°
240°
245°
8
9
10
11
12
13
14
15
16
17
18
19
20
Sea Surface Temperature (degree_C)
SST, Aqua MODIS, NPP, West US, Daytime (1 Day Composite)
(2013-04-09T12:00:00Z, Altitude=0.0 m)
Data courtesy of NOAA CoastWatch, West Coast Node

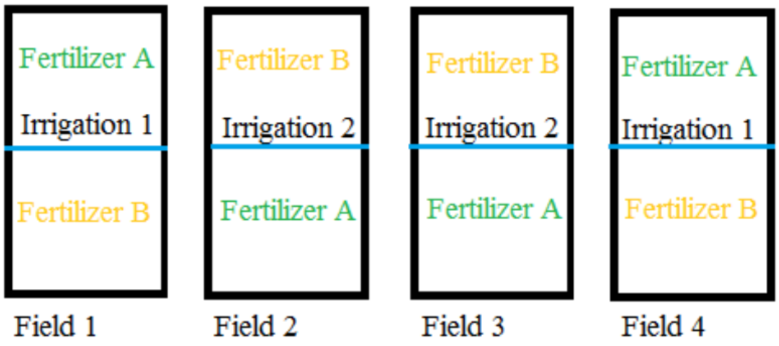
When we look at correlations between the organisms we find that barnacles steeply decline as mussels increase at the northern site, but there is no correlation at the southern site. We also know that temperature is greater at the southern site. Could temperature be changing the mussel-barnacle interaction? Maybe, but how many 'replicates' of the two temperature 'treatments' do we have here? If temperature doesn't vary much within sites, only across sites, then we essentially have an unreplicated comparison of two temperatures, and we can't really conclude anything about the role of temperature here. Any other difference between the two sites could be responsible for the difference in mussel-barnacle correlation.

In contrast, if we did the same surveys at six sites…


45°
40°
35°
30°
230°
235°
240°
245°
8
9
10
11
12
13
14
15
16
17
18
19
20
Sea Surface Temperature (degree_C)
SST, Aqua MODIS, NPP, West US, Daytime (1 Day Composite)
(2013-04-09T12:00:00Z, Altitude=0.0 m)
Data courtesy of NOAA CoastWatch, West Coast Node

can we now do a better job of testing how temperature might affect competition? We now have the option to include an interaction between mussels and temperature in our statistical model, while also include a site predictor to account for other differences between sites not due to temperature (this is best accomplished with a random effect for site, which we will cover in depth later). Because this is still survey data, we always have to keep in mind that an apparent mussel:temperature interaction could be caused by something else covarying with temperature. But at least we now have the option of testing whether there is a statistical temperature effect.

A final point to make is that the level of replication can differ for different treatments in an experiment, and this same phenomenon is equally important for survey data. A class experimental design for this situation is the split plot:



Here irrigation treatments are applied at the level of entire fields, while fertilizer treatments are applied within field. Irrigation is only replicated twice, because it is more challenging to manipulate, but fertilization is replicated four times, which increases the statistical power and precision for detecting a fertilization effect. Later we will discuss one way to analyze this kind of design with a mixed model, but there are also classical ANOVA methods (check out documentation of the function aov()).

In survey datasets the effect amount of replication will also vary for different predictors of interest. For example, in the mussel:temperature scenario we are imagining that temperature varies mostly at the site scale, while mussel abundance may vary a lot within sites, as well as across sites. When we learn about random effects we will see how to explicitly include predictors in a model that vary at different scales.