FES 524: Natural Resources Data Analysis

Reading 2.1

**Contents**

[Other readings to do before class 1](#_Toc76116025)

[Appropriate Language 1](#_Toc76116026)

[Research question 1](#_Toc76116027)

[Results 2](#_Toc76116028)

[Sources of variation 2](#_Toc76116029)

[Fixed vs random 3](#_Toc76116030)

[Minimize unexplained variation 3](#_Toc76116031)

[Statistical model 4](#_Toc76116032)

[Types of research goals 4](#_Toc76116033)

[Basis of statistical model 4](#_Toc76116034)

[Statistical model defines analysis 4](#_Toc76116035)

[Language about response variables 4](#_Toc76116036)

# Other readings to do before class

Read Week 2 Handout 1: “Study on thinning in western Oregon” and come prepared to use this example to talk about the topics covered in reading 2.1. Make sure you’ve had time to think about the answers to the questions in that handout.

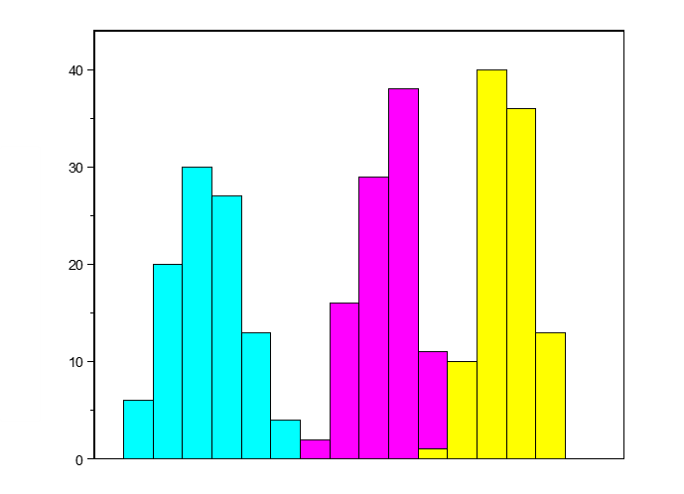
# Appropriate Language

Learning to use appropriate language in descriptions of your research question and study design is an important part of this class. When we informally discuss our study with peers we often skip doing this, but when it comes to formally writing things down in a manuscript we need to be more careful.

## Research question

When writing a research question, we should state exactly we are interested in. It is not uncommon to see the question stated something like “We want to know if the groups differ.” That’s not specific enough in any sort of formal setting, however. We need to be clear, among other things, about *which* differences we are interested in.

Below is an example of a graph showing histograms of the raw data for samples from three groups. We stated we wanted to know if the groups differ. But distributions can differ in many ways. Do we want to know if the variances differ? The ranges differ? The means differ? Or something else? Be specific about exactly what you are interested in within your research question.



An example of a research question that includes specific language:

“Of primary interest were differences in mean 5-year diameter growth between the light thinning density (325 tpa) and the other two thinning densities, moderate (225 tpa) and heavy (100 tpa) thinning, respectively.”

## Results

Similarly, the language used to talk about your results must always reflect the specific statistics you were interested in estimating or comparing. For linear models, which is the focus of this class, this will be differences in means or medians.

Examples of language that can be used to discuss results:

*On average*

*Mean* site growth increment

*Median* average diameter growth

Many studies you hear about in the news were interested in differences in means or medians (measures of *location*). When you hear a breathless report about some new study results, mentally add “on average” to the reported conclusions and consider how that changes how you think about what you heard.

# Sources of variation

Definition:

A component of a study such that *different levels* of that component result in *different values* of a given response variable.

Each study will likely have many sources of variation. Here are just a few examples:

Geographic areas

Plots

Forest stands

Species

Experimental units

Continuous covariates, e.g., rainfall, soil moisture

Protocols/treatments

Subplots within plots

Time periods

Data recorder

Refrigerator

## Fixed vs random

We will need to think about the role each source of variation plays in the study and whether or not it is relevant to the research question or not. This concerns sources of variation that are categorical. Continuous sources of variation will be fixed effects.

When a source of variation is directly related to the research question, we consider this an effect of interest. For example, you expect a protocol you applied to your study units to cause variation. You are interested in the systematic effect of that protocol; it is part of your research question. In that case, you would treat that source of variation as a *fixed* effect in any analysis.

Other sources of variation are not of interest. For example, while we generally believe stands of trees will vary in many ways because they have different environmental conditions, investigators are often not interested in the systematic effect of different stands on their response variable. This type of source of variation would be treated as *random* in an analysis. Random effects are usually based on variables that cause variation but are not from the protocol(s) of interest.

We will be discussing fixed vs random effects more extensively throughout the quarter.

In some studies, there may be sources of variation caused by subsampling within the replicates of a study. You may see these referred to as *pseudoreplicates* or *subsamples*. For example, if a protocol is assigned to the stand level but we measure multiple trees within each stand, the trees are pseudoreplicates. This sort of source of variation could be treated as an additional random effect in analysis, since it’s not relevant to the research question. However, since the protocol was applied overall to the replicates, we could also average over the subsamples and work on the scale of the replicate. This simplifies the analysis and, generally speaking, does not change any results.

## Minimize unexplained variation

When we are designing a study, one important goal is to minimize unexplained variation. Unexplained variation should be small so that we can reasonably estimate any changes/differences of interest with precision. If we fail to minimize variation we may find ourselves in a scenario where our estimated differences in means are large enough to seem practically important but we have so much variation that the plausible values for the true difference encompass a huge range of values and so the results are inconclusive.

We should always spend a lot of time thinking about sources of variation prior to collecting data. Sources of variation will drive the study design and can be potentially controlled within the study design. For example, *blocked* designs, which we will discuss next week, are designed specifically to help control one source of unexplained variation. Another way to minimize unexplained variation is to collect information on continuous variables (i.e., *covariates*) that we believe will cause variation in the response variable and will need to be accounted for in any analysis. Which covariates will be collected must be defined during the design stage of a study. Finally, limiting the scope of inference, as discussed last week, can also limit unexplained variation.

Unexplained sources of variation are considered limitations to a study and should be discussed in the conclusions section of a manuscript. If, e.g., investigators failed to collect an important covariate, they should discuss this and recommend that future research be designed in a way to mitigate the effect of that covariate.

# Statistical model

In this class, when I talk about a statistical model I mean the theoretical model that defines how we believe the response variable was generated. The statistical model is written in mathematical notation and defines all assumptions of the model. We will see examples of the mathematical notation in class but I will not ask you to write the notation on assignments. Instead you will describe the statistical model in words.

If the statistical model we choose is a poor approximation of how the response variable is generated, results from that analysis based on that statistical model will not do a good job of addressing the research question or improving our understanding of the science in our field.

## Types of research goals

There are two main types of research goals that we commonly create statistical models for.

*Group means*

Estimating group means or group proportions is one common goal. This is often the goal in survey research, for example. When estimating group means, any bias in the measurement is a big problem. When measurements are biased in some way the estimated mean will not be accurate.

Note that prediction is a special case of estimating means.

*Differences in group means or estimating slopes*

The goal of estimating *differences* in group means, which is the focus of this class, is a little easier than estimating group means. As long as any bias in the measurement is the same among groups, the bias cancels out when the difference is taken. Similarly, an estimate of the slope is still unbiased when all values were shifted up or down due to a bias in the measurements. The intercept of the line is biased, but not the estimated relationship of interest.

## Basis of statistical model

The statistical model is based on the research question/goals and elements of the study design.

* The research question tells us the question of interest that the model needs to address and defines the response variable of interest.
* The known sources of variation tell us which variables need to be in model in order to minimize variation.
* The study design helps define the structure of the statistical model based on how the protocol of interest was applied and the replicates of that protocol.

## Statistical model defines analysis

The statistical model, whether we formally write it in mathematical notation or describe it, is the basis for the actual analysis. We must have an idea of what the statistical model is before we can choose an appropriate analysis and fit a model with software.

# Language about response variables

Sometimes we end up working with complicated response variables. This is particularly true when we’ve averaged over subsamples to work on the scale of the replicate of some protocol. In this case, our response variable is an average.

When describing the results of an averaged response variable we will end up talking about differences in “mean average response variable”. This is perfectly appropriate, but can feel a little awkward. If you want to avoid that sort of language you can define a term to use for the averaged response variable early in your manuscript and use it throughout the rest of the paper. Be careful to be consistent; don’t go back and forth using multiple terms to refer to the same response variable.

Example language defining a new term:

“This averaged 5-year diameter growth will be referred to as the site growth increment throughout this paper.”