FES 524: Natural resources data analysis

Reading 2.1

2024-01-11

Contents

1 Appropriate language **1**
- 1.1 Research question . . . 1
- 1.2 Results . . . 2

2 Sources of variation **2**
- 2.1 Fixed versus random (the classic conundrum) . . . 3

1 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.

1.1 Research question

When writing a research question, we should state exactly what 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.

Figure 1 shows 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.

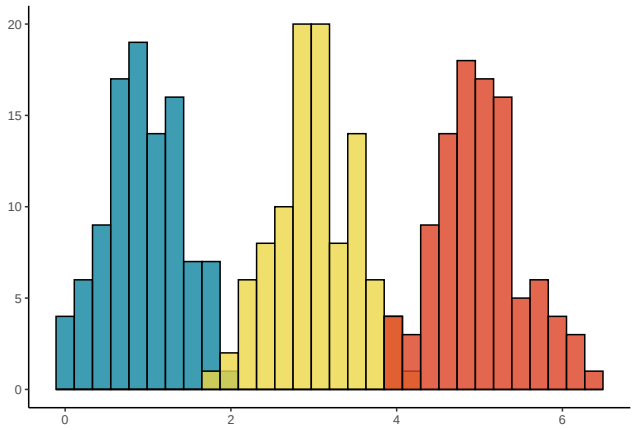

Figure 1: Example of histograms of samples from three groups in which the mean differs but the variance is constant.

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."

1.2 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 generally 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.

2 Sources of variation

> A source of variation is 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

2.1 Fixed versus random (the classic conundrum)

In this class, we will only discuss models in which random effects are included for categorical variables. Continuous explanatory variables will always be treated as fixed for our purposes (though there exist models and approaches that bend that rule).

From a philosophical standpoint, so-called *random effects* (also known as varying effects) are those variables for which the *levels* that are present in your study are a random sample of the population of possible levels. Someone aiming to reproduce your experiment or study could easily select a different subset of levels because those specific levels in your study were not the target of the research question. *Fixed effects*, on the other hand, are the variables for which you as the experimenter or observer *chose* the levels because you were interested in those particular levels of the variable as they pertain to the research question.

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 during 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. I like to ask myself:

> If someone were to test the reproducibility of my results (in terms addressing the research question), would they need the same levels of this variable, or could they choose another subset of levels of

this variable and still test my conclusions regarding the research question?

We will be discussing fixed versus 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.