FES 524: Natural resources data analysis

Reading 2.1

2024-01-11

Contents

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.

An example of a research question that includes specific language:

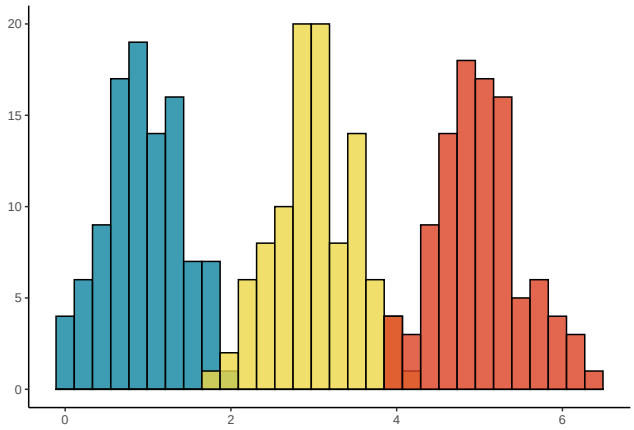

Figure 1: Example of histograms of samples from three groups in which the mean differs but the variance is constant.

> "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.

2.2 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, for example, investigators failed to collect data on an important covariate, they should discuss this and recommend that future research be designed in a way to mitigate the effect of that covariate. They should also acknowledge that including the missing covariate in the model could drastically change their results, so future research that comes to a different conclusion may be due to accounting for the effect of the missing covariate in the model.

3 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 (i.e., the data-generating process), 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.

3.1 Basis of the 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, in a regression context, 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.

3.2 Types of research goals

The same statistical methods may be used for different research goals. Research goals generally fall into three categories:

1. Exploratory
2. Confirmatory
3. Predictive

Regression is a perfect example of how the same overall methodology and framework can be used for any of these research goals. For example, I can use multiple regression to explore which of a set of variable might be related to my response. I can also use regression to test for differences in means among groups from an experiment, testing *a priori* hypotheses about how the treatments drive differences in the response, on average (confirmatory research). Finally, I can use a linear model to predict unseen outcomes based on measured or hypothetical values of the explanatory variable. However, we usually cannot accomplish all these goals with one dataset and analysis (despite common statistical frameworks) without violating some statistical assumptions or dealing with suboptimal approaches for one or two of the goals. In other words, there is a tradeoff, and clearly defining your objective before you start an analysis is necessary to choose a good approach. We will get into this more when we discuss variable selection later in the course.

3.3 The 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.

4 Language about response variables

Sometimes we end up working with complicated response variables. This is particularly true when we average 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."