FES 524: Natural Resources Data Analysis

Reading 1.2

Contents

1 The discipline of statistics

When many people think of the discipline of statistics, they think "data analysis". However, as you can see in Figure 1, data analysis is only one aspect of statistics. Statistics is a broad field that covers many topics. There are aspects of statistics that are relevant throughout any research project. Responsible conduct of research requires the integration of all these branches of statistical practice.

While aspects of statistical practice are relevant throughout a research project, take note of what is missing from the branches within the discipline of statistics. There is no mention of the research question or the scientific theory the research is based on. Statistical practice is built on top of the science of your field. Your scientific expertise is as crucial to research as statistical practice, and the best research will combine good scientific and good statistical practice. Statistics and scientific research come together as a collaboration between disciplines.

1.1 What is a statistic?

Statistics is a discipline, but (somewhat confusingly) we also use the word "statistic" to mean a numerical summary that describes a sample. In fact, any numeric value that is calculated from a sample (the data) is a statistic, including the raw data themselves because $\mathbf{y} = I(\mathbf{y})$, where $\mathbf{y}$ is the vector of data and $I()$ is the identity function. Additional examples of statistics include a difference in means, a mean, a proportion, a median, and frequencies of events.

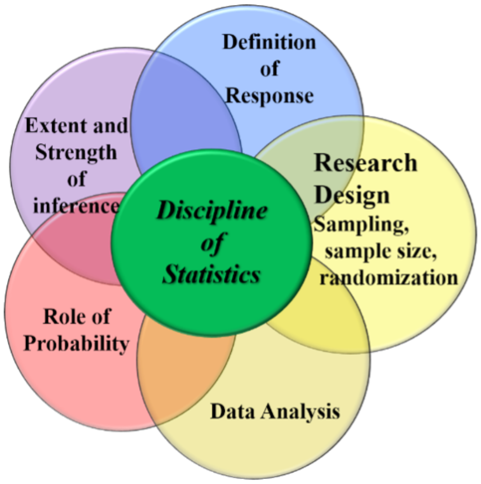


Figure 1: Venn diagram of components of research included in the discipline of statistics.

1.2 Why we use statistics

We use these statistics (i.e., numerical summaries) in a variety of ways and for a variety of reasons. Three primary reasons we use statistics is to describe, to explain, and to predict.

We use statistics to *describe* something about a sample. Description is most common for exploratory research, where we are learning about something for the first time and we have no scientific theories or past information to help inform our expectations. Exploratory research leads to new research questions. Description can certainly be a part of confirmatory research, as well.

We use statistics to *explain* when we use the sample to describe something about a population. Many statistics courses, like this one, are based on statistics-as-explanation from confirmatory studies. We design a study to address a research question that is based on current scientific theory. When we explain we estimate numerical summaries as support to help support or refute previous research or scientific hypotheses.

We use statistics to *predict* when we use the sample to describe something about future observations from similar populations. When prediction is the goal, then the statistics we calculate, how we do the analysis, and what we discuss as statistical results is different from those we use for confirmatory analyses for explanation. When the goal is to make predictions, we are most concerned with measures of how well we can make predictions on future data. This is the primary objective of *machine learning* methods, which we will not cover.

Both explanation and prediction involve using sample statistics to ascribe characteristics to members of the population we haven't seen. This means the sample must be *representative* of the population of interest. We will discuss representation more later.

A single study can use any or all of these three main uses of statistics. While some parts of the research may be confirmatory, based on prior knowledge, the same research project could have exploratory aspects to it. For example, a "surprise" result might lead to more exploration and description of the sample as a discussion point, which can be used as a springboard when planning future research paths.

2 Sampling

Statistics involves taking a sample. If we could get information on the entire population, we wouldn't need statistics; instead we could simply state exactly what the population looks like.

What does it mean to take a sample from a population? How well does any individual sample describe the population? Are statistics estimated from a sample always good estimates of what is happening in the population? Think about these questions and use the app below to explore sampling. The ultimate goal of the app is to explore the central limit theorem, which should be review for you, but along the way it touches on important aspects of sampling for you to think about.

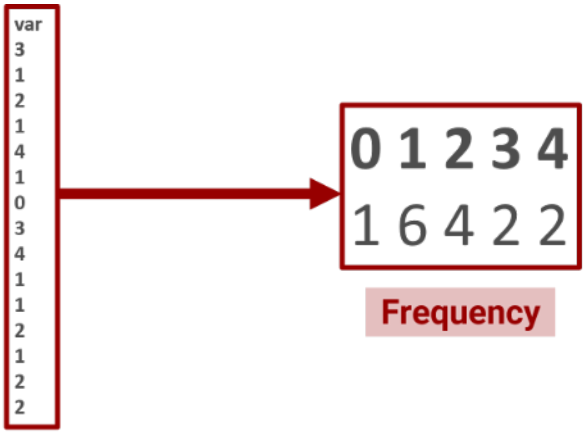
http://mfviz.com/central-limit/

3 Distributions

You will find distributions play an important role in the discipline of statistics and in this course. What is a distribution?

A distribution is a description of the relative frequency of the different values of a variable.

Figure 2 shows a simple example of the frequency of a simple integer variable as a table.



0	1	2	3	4
1	6	4	2	2

Figure 2: Simple frequency table of a sample from an integer-valued random variable.

Alternatively, we could represent the same sample distribution via a histogram (Figure 3).

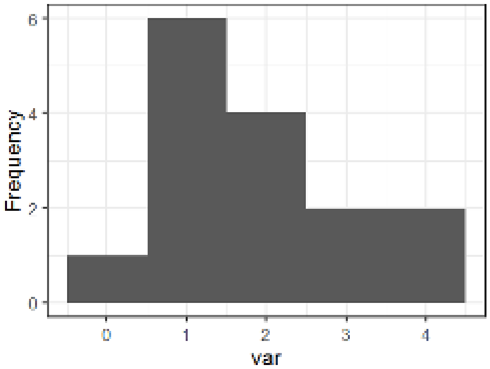


Figure 3: Simple histogram of a sample from an integer-valued random variable.

There are both continuous and categorical distributions, depending on the variable type.

Continuous variables can be *constrained* to subsets of the real number line (e.g., between 0 and 1, inclusive, denoted [0, 1]) or potentially take any real number $(-\infty, \infty)$, as long as they are continuous in the set.

Examples of categorical variables include non-negative integers (e.g., {0, 1, 2, 54, 78, 190}) and presence/absence ({0, 1}).

4 Study design

It is impossible to separate statistical analyses from the study design, since the study design directly informs what kind of analysis might be appropriate. The description of the study design needs to include a clear definition of any response variables, assignment protocols, how samples were selected, and the unit of replication for all variables of interest. All groupings must be accurately described using consistent language.

We will spend time throughout the course reading example study designs and using them to describe the scope of inference for study results and to come up with an appropriate statistical model for analysis. Students will also practice describing their own study for peers.

5 Scope of inference

The scope of inference is the set of conditions and objects to which the conclusions from the research will apply. This is based on the study design; i.e., how the sample was taken.

The conditions for the scope of inference are elements from the study design. These are specific to an individual study, but there are some general categories that are going to be common parts of the scope of inference.

- **Physical units**: The physical units that are sampled, such as, e.g., plots, stands, trees, and watersheds.
- **Geographic units**: Where the study takes place. For example, the Great Basin, the Coast Range, or western Oregon are all examples of the geographic conditions that could affect scope of inference.
- **Biological units**: Biological conditions, such as a specific species that is part of the study, an age range of the organisms sampled, or some subpopulation of a species.
- **Temporal units**: The specific time the study takes place. This could be a single year, a range of years, conditions during nights for some time frame, etc.

A research question should explicitly include the scope of inference. In the example below, notice how biological and geographic conditions are included as part of the question so the scope of inference is explicitly stated.

> Does a reduction in thinning intensity increase the 5-year growth increment in 25-30 year-old Douglas-fir trees in western Oregon?

5.1 Example of extending the scope of inference

Some studies will have a narrow scope of inference based on the study design and the sample taken. If you have a narrow scope of inference, you need to acknowledge it. However, it is possible in some cases to extend a narrow scope of inference based on scientific expertise. This does not involve simply stating that the scope is wide instead of narrow. Instead, justification based on current scientific knowledge needs to be used to say why the scope is wider and what population your sample can reasonably represent.

Below is an example of how to extend a narrow scope of inference from the Hayes et al. paper “Response of birds to thinning young Douglas-fir forests”, published in Ecological Applications in 2003.

Note that they first acknowledge the narrowest scope of inference and talk about populations that inference should not be extended to because it can’t be scientifically justified. They then discuss why the scope of inference is likely wider than it appears, using scientific justification based on how their results fit into the science of their specific field. This is the sort of thing you need to do if you think your study has a wider scope even though the sample is extremely limited.

> Our study was restricted to 35–45-yr-old Douglas fir stands in the northern Oregon Coast Range. Our results may not be applicable to substantially different thinning intensities, and should not be applied to much more intensive treatments, such as shelterwood treatments.

> Strongest inference can be applied to forests of similar structure in the same region. However, as our findings are generally consistent with what is known about the natural history for most of the species that we examined, we believe it is likely that our results can be applied to the same species in other geographic regions with coniferous forests of similar structure.

5.2 Representation

The goal of a study is most often to make inference to something more extensive than the sample taken. When we talk about representation we are talking about which population the sample represents. You will hear the term representative sample to refer to a sample that represents some population of interest.

Which population the sample can be used to make inference to is an important part of study planning. Researchers must balance having a wide scope of inference with controlling known sources of variability.

If they have limited resources, researchers may decide to focus on a very specific portion of the population. While this will narrow the scope of inference, it also decreases the variability that needs to be sampled across and they will need fewer samples to obtain a representative sample. If the sample needs to represent an extensive population that encompasses a lot of different conditions, more samples will be needed in order to get a sample that is representative of the population because of the variability that must be sampled across.

6 Randomization

Randomization is an important element of study design. How randomization is used depends on what kind of study you are doing.

6.1 Random assignment

When doing a designed experiment, units of interest can be randomly *assigned* to the protocol of interest by the investigator. Random assignment helps ensure that observed differences are due to the protocol under investigation and not to an unidentified cause. Random assignment of a treatment or protocol of interest is what allows us *causal* inferences. In other words, with a sound experimental design, we can interpret differences among treatment groups as being *caused* by the treatments.

6.2 Random selection

Unlike experiments, observational studies rarely allow for random assignment. Units often already exist in groups or with specific characteristics of interest prior to selection for measurement. The investigator can only select which units to measure. Random selection ensures that the effects we observe are reasonably believed to be true for the whole set of units we have selected from (the population), not just for the subset we observed.

Random selection is used to justify the scope of inference that is wider than only the sampled units. However, any single sample based on random selection can be poorly representative of the population, as you saw in the simulation app you looked at earlier.

6.2.1 Other sampling methods

6.2.1.1 Haphazard sampling is when the investigator chooses units to sample by some method that doesn't involve taking a statistically random sample. This is often incorrectly referred to as "random", using the word random as it is defined outside of science as "chosen without method or conscious decision". It is difficult to justify a haphazard sample has a scope of inference wider than the sample, and there have been examples that show that someone choosing samples haphazardly can unconsciously bias the result.

6.2.1.2 Professional judgment sampling involves working with an expert to explicitly choose samples that represent that population of interest. This involves more scientific expertise about the study area or study organisms than usually exists, but such a sample can work very well for getting a representative sample of the larger population.

7 Replication

Replication is the repetition of independent applications of a treatment or protocol. Some studies will have multiple protocols and multiple levels of replication, where one size of unit is a replicate of one protocol while another size of unit is the replicate of another protocol.

We will practice identifying the replicate throughout the quarter.