STAT4102 Overview

Course Description: Sampling distributions; likelihood and sufficiency. Estimation; significance tests; distribution free methods; power; application to regression, analysis of variance, and analysis of count data.

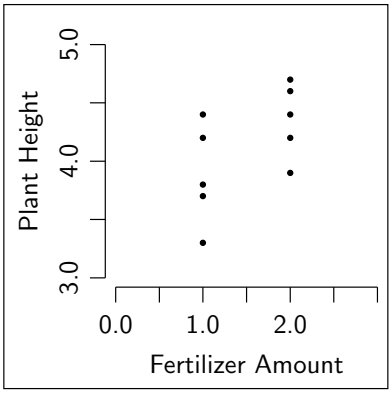
1 Introduction

To understand what this course is about, we first need to understand what variability is and the difference between a population and a sample.

1.1 Variability

Almost every measurement involves variability. This can mean several things. The measurements themselves can be variable, such as when measuring the height of a individual plant, or individual differences in a group of subjects can be variable, such as when measuring the heights of several plants that have been treated with a certain fertilizer. Or it can refer to variability within a fixed group, such as the the heights of an entire field full of plants.

Variability can make it difficult to draw conclusions about what we are actually measuring. For instance, consider an experiment where plants are grown with either one scoop or two scoops of fertilizer, five plants for each amount. If there was no variability in plant growth or between plants, then each of the five plants treated with the same amount would grow to the same height, and there would be no problem determining which amount of fertilizer was better. But because there is variability, which is better may not be clear. Perhaps the data is as the graph on the right, where overall, the plants with more fertilizer seemed to grow taller, but there were some plants with more fertilizer that were shorter than some of the plants with less.



How do we decide whether the fertilizer made the plants grow taller? Perhaps it was just random variation that made it look like the plants with more fertilizer grew taller. This variability makes a decision unclear.

Statistics consists of tools we use to understand this variability and determine the answers to questions like these. That is, if there was no variability, there would be no need for statistics. Some even say that statistics is the study of variability.

1.2 Population vs. Sample

To help to understand variability, statisticians use the concepts of *population* and *sample*. Population refers to the various possible measured values, while the sample refers to the actual measurements taken. For example, we might be interested in the heights of students at the University; the population is the collection of measurements of every student. Since it's impractical to measure every student's height, we take a sample of fifty. The fifty heights are the sample.

In the plant and fertilizer example, there isn't a fixed group of plant heights. The population in this case refers to the possible heights that the plants could have grown to. (This takes a bit more imagination.) The sample refers to the actual heights that we measured.

You may be beginning to realize that what we're actually interested in is not the sample, but the population. If we could observe the population directly in our plants and fertilizer example, we could make a definitive conclusion about whether or not more fertilizer helped. Unfortunately, we can never know for certain about the population, only the sample.

However, we may be able to infer something about the population, given the sample we have. This semester, we will study how this is done.

2 Topics covered

The main technique we will use for understanding populations is to model the population with one of the probability models we learned last semester—usually the normal distribution. Questions we'll cover, though not quite in this order, include:

- How can we decide that one population model is better than another? (likelihood)
- Do we need the entire sample to make this decision, or is some summary of it enough? (sufficiency)
- If a summary of the sample is enough, how is that summary distributed? (sampling distributions, central limit theorem)
- Often we'll be interested in a single feature, or parameter, of the population. How can we estimate this parameter? (estimation)
- How can we describe the uncertainty in our estimate? (confidence intervals)
- How can we determine if a population parameter is different than a given value? (significance tests)
- What if we need to make an either/or decision about it? (decision rules, tests)
- How can we describe the uncertainty (or certainty) in our decision? (power, Type I/II errors)
- How do we tell if the probability model we're using is applicable? (goodness of fit)
- Can we make conclusions even if we don't have a good probability model? (distribution-free methods)

When these basic questions are answered, we'll study more complicated situations, including:

- comparing two populations
- categorical responses (count data)
- continuous responses, more than one populations (ANOVA)
- continuous responses, and continuous predictors (regression)