Diving Deeper into the Statistics of a Scientific Study

Case Study: In this article I will discuss the statistical methods used in a paper I published from my graduate school research. For context, the paper is called *Symbiotic nitrogen fixation does not stimulate soil phosphatase activity under temperate and tropical trees* and was published in Oecologia in March 2023.

Research Summary:

* Symbiotic nitrogen fixation is a process where trees establish partnerships with nitrogen-converting bacteria in the soil and roots, providing plants with accessible nitrogen. This vital nutrient supports plant growth, and in some cases, bacteria reside in specialized root structures called nodules.
* Researchers hypothesized that fixed nitrogen might contribute to the production of phosphatase enzymes, releasing phosphorus from organic matter. However, the association between nitrogen fixation rates and phosphatase activity remains uncertain.
* In this study, soil phosphatase activity was measured under nitrogen-fixing and non-fixing trees transplanted to tropical and temperate sites in the USA. Surprisingly, no significant difference in phosphatase activity was found between the two groups, and the connection between nitrogen fixation rates and phosphatase activity was not supported.

Here is a snapshot of the data from this study:

Graphical user interface, text

Description automatically generated

First, we analyzed tree growth in response to fertilizer treatments using a linear mixed effects model (LME) for each species, using Estimated Marginal Means (emmeans) for posthoc pairwise comparisons.

Linear mixed effects model (LME): a type of statistical model used for analyzing data with clustered structures. They account for both fixed effects variables and random effects (subject specific or group specific). This was important to use because we used several species of both N-fixing and non-fixing trees (controls), so this type of model controlled for variation between species that were due to the fact that the species may vary in ways that cannot be related to our treatments.

Estimated Marginal Means (emmeans): a post hoc analysis that is typically used alongside LME models to estimate and compare the marginal means of groups or levels in the presence of both fixed and random effects. This helps to make comparisons between groups more accurate and statistically valid.

For the remaining analysis we used several linear regression models.

Question 1: How does soil phosphatase vary between N-fixing and non-fixing trees?

Question 2: How does N fertilization affect soil phosphatase activity?

* For Question 1 and Question 2 we used a linear model for each study location and year using soil phosphatase activity as the dependent variable and species and treatment as fixed effects.

Statistical concept – Fixed effects

* Fixed effects are a type of independent variable that is treated as a constant across groups in the data. This way, they are able to account for variation between groups in the dataset that are not the focus of the analysis but could introduce bias into the results of other variables of interest.
* Here, we used species and treatment as fixed effects to isolate the effect of species (a proxy for looking at N-fixing vs. non-fixing trees while also factoring in the effects of the fertilization treatments.

Question 3: How does symbiotic nitrogen fixation affect soil phosphatase activity?

* For Question 3, we also used a linear model with soil phosphatase activity as the dependent variable and N-fixation (yes or no), fertilizer treatment, and species as fixed effects. Only some sites had more than one species for each category of N-fixing and non-fixing trees, so species was only included for those sites as one of the fixed effects.
* We ran separate models for the indicators of nitrogen fixation: the rate of N fixation and %N dfa
* For each linear model, we used ANOVAs to determine the significance of each driver (independent variable or fixed effects).
* We followed the ANOVAs with Tukey HSD tests, when relevant to determine differences between pairs of means, determining which specific pairs differ significantly from each other.

Statistical concept – Power Analysis:

* To ensure the robustness of our conclusions, we conducted a power analysis using the variability and sample size from our data.
* We used the standard deviation of the residuals from our test of soil phosphatase as a function of treatment and species

Statistical concept – Residuals

* Residuals are the difference between the observed data points and the values predicted by a statistical model. In this study, we used linear regression models that best fit the data and explains the relationships between variables in our data, so the residuals represent the difference between that best fit line and the data points that it is fit to. On a scatterplot, the residuals represent the vertical difference between the observed data points and the regression line. With a model that is well fit to the data we should expect low residuals, that are close to 0 on average.
* Homoscedasticity is an assumption that the residuals are spread evenly, meaning that the data points are spread evenly along the regression line, and all have similar vertical distances from the regression line.

Statistical concept – Power analysis

* Power analysis refers to a statistical technique used in hypothesis testing and experimental design to measure the ability of a study to demonstrate an effect, if it exists within the data.
* This includes elements such as effect size, sample size, significance level (alpha), and variability.

To determine our ability to detect differences, we conducted a power analysis using the variability and sample size from our data. Specifically, we used the standard deviation of the residuals from our test of soil phosphatase as a function of treatment and species (the test we used to address Q1 and Q2), then used the same test with simulated data with different treatment effect sizes, running each simulation 1000 times. Using the standard α = 0.05 cutoff, our sample sizes and variability allowed us to detect treatment effects of ~ 10, 6, 5, 5, and 15 μmol·g dry soil–1·h–1 in New York in 2017, New York in 2018, Oregon, Waiakea, and Volcano, respectively.

Summary of results: