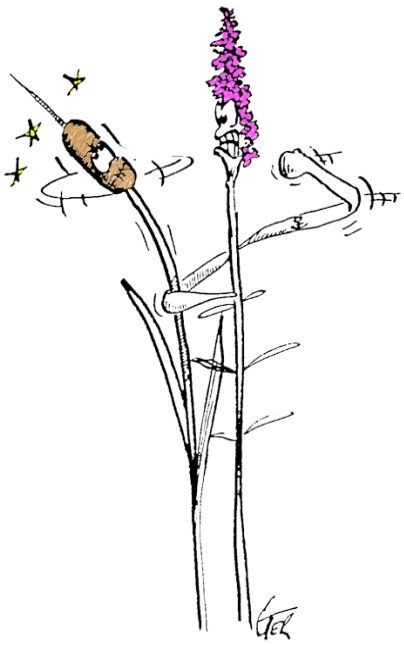
BIOL343 – Assignment #4

Entrance to Inference



Assigned Sunday 2 February
Due Saturday 8 February 1159pm

We will revisit the study investigating the efficacy of biocontrol on purple loosestrife that you delved into for Assignment #3, where you tested three hypotheses with graphical analyses and summary statistics. This time, you will calculate confidence intervals and apply some statistical tests to determine the statistical support for some of the differences you observed.

Recall the three hypotheses you tested:

(H1) *Galerucella* beetle damage of *L. salicaria* is lower in central than eastern Ontario.
(H2) *Galerucella* damage was higher at sites where the biocontrol agents had been deliberately released than sites where they had not.
(H3) Damage by beetles reduced the local abundance of *L. salicaria*.

For this assignment, you will focus on H2 and H3 only. For H2 you will contrast **mean** damage between samples of populations. For H3, I want you to focus **only** on whether local abundance varies with leaf damage (as measured by damage index). This involves regression analysis. You will use the loosestrife.csv file posted for this assignment. This data set is slightly different from the assignment #3 version of the file, so use the new one. See assignment #3 for a detailed description of the variables in the dataset. This assignment will focus only on **damage** as a response variable for H2 and a predictor for H3.

The first step of any good analysis is to make graphs that directly address the hypothesis at hand. A comparison of means (H2) would be best supported by graphs that compare group means and 95% confidence intervals. You should do this using a `geom_point()` ggplot with means and 95% confidence intervals calculated using a chain of piped **dplyr** commands. For assignment #3, you've already made a scatterplot to evaluate H3 but please include a revised version of the figure here (hint: a $\log_{10}$-transformation of stems per $50m^2$ would be highly appropriate).

In lecture, you learned how to compare group means with a 2-sample t-test (`t.test()`). Here, you will use the more general linear models approach (the `lm()` function), which we explored in lecture. This function also performs linear regression, so we can also use it to evaluate H3.

You should always verify that the assumptions of statistical tests are met. Evaluating the assumption that residuals from a linear model are normally distributed should involve plotting a histogram of residuals and overlaying a normal distribution with the same mean and standard deviation. In lecture, we discussed whether it was a good idea to use Shapiro's test to test for deviations from normality. Do that here, but then tell me why this practice might be problematic.

Linear models also assume that there is no relation between the residuals and predicted values from a model (i.e. error around the model must be purely random). How you go about verifying this assumption depends somewhat on whether the predictor variable is categorical or continuous. Draw upon what we discussed in lecture to test this assumption for both linear models you have run. Perform appropriate statistical tests.

If assumptions are not met, it's good to know alternative approaches to evaluate the null hypothesis. You can test the null hypothesis using the permutation approach we developed in lecture. To do this, compare the observed *F* value (calculated above) to distributions of *F* values generated from the data under the null hypothesis. Does this permutation result match what you obtained with the formal tests based on the theoretical *F* distribution?

So, here's a summary of what Hana and I are expecting from you:

1. A code chunk loading the required packages.
2. Code importing the dataframe and check it with the single most useful data-checking function.
3. Code that makes a graph to evaluate the hypothesis that damage is higher at release than nonrelease sites (H2).
4. Code that executes a linear model testing for a difference in mean damage between release and nonrelease sites (H2).
5. Code and a figure testing the normality of residuals assumption for H2.
6. Code and a figure testing the independence of residuals assumption for H2.
7. A permutation test for H2.
8. A section of text interpreting your results with respect to H2.
9. Code that makes a graph to evaluate H3.
10. Code that uses linear regression to evaluate H3.
11. Code and a figure testing the normality of residuals assumption for H3.
12. Code and a figure testing the independence of residuals assumption for H3.
13. A permutation test for H3.
14. A section of text interpreting your results with respect to H3.
15. All figures must have "professional" figure captions. If you are still wondering how to write a proper figure caption please look at the captions of figures in published scientific papers.
16. All figures must be super-pretty, publication-ready and finessed using the `scale…()` and `theme()` functions in **ggplot**.
17. The figures that check linear model assumptions for each hypothesis should be put together in a composite figure using functions in the **cowplot** package (i.e. 2 composite figures: one for H2 and another for H3).
18. All formatting conventions used in assignment #3 apply here.
19. Make sure all code is as concise as possible and neatly organized in code chunks.

For this pop assignment, you will submit an **PDF version** of your .html R notebook document called "StudentNumber_A4.pdf", where the file name starts with your student number.

Please make sure that you final PDF file is complete and nicely formatted before submitting it

You should upload your file to the Assignment #4 OnQ dropbox by Saturday 8 February at 1159pm.

Good luck!!