***PROFESSIONAL SKILLS STATISTICAL ASSESSMENT***

*Have questions? Ask Kyle Dexter,* [*kyle.dexter@ed.ac.uk*](mailto:kyle.dexter@ed.ac.uk)*,* or a demonstrator

So far you have been working with data on soils from various forest plots in the Amazon of southeast Peru. In this assessment, you will work with a different dataset from the same region, which is comprised of functional trait information for various species in the genus *Inga* (Leguminosae) in the same area. This is the most abundant and diverse genus of trees in the region. Some of these species prefer floodplain, bottomland environments, others prefer upland environments, while still others are generalists that occur in both habitat types. The species also occur in different phylogenetic clades (e.g. different sub-genera). The traits that were collected relate to how tall the trees are, different aspects of leaves such as their size and hairiness, and the chemical compounds that are in the leaves. You can find the file on Learn, saved as in csv format for direct import into R and in xlsx format in case you want to check out the data in Excel or Open Office. The file name for the csv file is “Inga\_traits.csv”. Here are the units for the traits and a brief description.

Max\_Height: m, the maximum observed height of trees of this species

Leaf\_Area: cm2, the average size of leaves for the species (actually leaflets)

SiO2\_Leaf: mg/g, the concentration of silica in the leaves (used in herbivore defence)

N\_Leaf: mg/g, the concentration of nitrogen in the leaves (a major plant nutrient)

leaf C: mg/g, the concentration of carbon in the leaves (can be used in structural defences)

leaf P: mg/g, the concentration of phosphorous in the leaves (another major plant nutrient)

trichome density: number of hairs / cm2, the density of hairs on the upper leaf surface

trichome length: mm, the average length of hairs on the leaves

chlorophyll: mg/cm2, the chlorophyll content of the leaves on an area basis

expansion: percent/day, how quickly the leaves expand

Catechin: present/absent (1/0), an important chemical in herbivore defence

Gallocatechin: present/absent (1/0), an important chemical in herbivore defence

Mevalonic\_Acid: present/absent (1/0), an important chemical in herbivore defence

**Exercise 1: Histograms and normality (100 marks)**

1. Please make and present a histogram for leaf area of these species. What can you say about this distribution in statistical terms? Does leaf size appear to be normally distributed? (35 marks)

Chart, histogram

Description automatically generated

The data for leaf size appears to be skewed to the left, with a quick rise and then a tail out, suggesting a Poisson distribution.

Figure 1: Histogram of frequency of leaf area (LAI cm2) at different sizes.

1. Try log-transforming leaf area and make and present a histogram of log-transformed leaf area. (25 marks)

Chart, histogram

Description automatically generated

Now data is presented logarithmically the data now appears normally distributed with a bell shaped curve.

Figure 2: Histogram of frequency of log leaf area (LAI cm2) at different sizes.

1. Now, in simple terms, how would you describe the distribution of leaf sizes across trees in this region to a non-scientist? (40 marks)

We are investigating Leaf\_Area: cm2 in the Amazon forest plots of southeast Peru, which is the average size of leaves for the species (actual leaflets). In figure 1 we can see the leaf area results for all Species, Habitats and Clades with the highest frequency slightly skewed to the left, with a mean of 47.3. In figure 2 we have log-transformed figure 1, which reduces the effect of skewing thus showing us that the data is normally distributed (in a bell-shaped curve). This means that most Leaf\_Area readings will fall within this curve with the middle of the bell shape being the optimum size shared by the most trees leaves.

**Exercise 2: Box plots and Analysis of Variance (ANOVA) (100 marks)**

1. Now let’s see how species in different habitats might differ in leaf chemical composition. Make and present a boxplot of leaf phosphorous concentration versus habitat in which a species is found. (25 marks)

Chart, box and whisker chart

Description automatically generated

Figure 3: Boxplot of leaves phosphorous (mg/g) concentration in 3 different habitat types.

1. Now statistically test if species found in different habitats have significantly different phosphorous concentrations in their leaves. Report the F Statistic, p-value and degrees of freedom for your test. Then, tell me what these two measures mean in general and what the specific values mean in the context of this analysis. (20 marks)

ANOVA F2,27 = 8.598, p = 0.00129

The F statistic is the ration between variation between sample means : variation within the samples. We have a large F value, which means that the variation among group means is more than you'd expect to see by chance. In our example this means that the variation between phosphorous concentrations in different habitats is not by chance.

The degrees of freedom show the maximum number of logically independent values, which are values that have the freedom to vary, in the data sample. In this example 2 is the total number of values minus 1 (28-1), so 27, and the number of habitat types (3-1) 2.

The p value shows us that if under 0.05 then there is statistical significance and in the case of this example it means the phosphorous concentrations in different habitats are significantly different.

1. Try and conduct an evaluation of your model. I do not need to see any model validation figures, but I do want some written explanation of why you think your model is good (or not). Have you likely violated any of the assumptions of ANOVA? If so, which ones? (15 marks)

AIC = -118.7751

nullAIC = -107.9913

The AIC is better in the original model compared to the null. This means that the model fits better than the null

We have violated the assumption that data is normally distributed

1. How might you improve your model? Try doing so and report the revised F Statistic and p-value. (15 marks)

With log data with have a higher F statistic and a lower p-value,

ANOVA F2,27 = 10.12, p = 0.00052

1. Now, provide an explanation of your analysis, the results and what they mean, in non-technical terms that would be accessible to a relative or someone you meet in a pub (or elevator if you don’t frequent pubs). Your explanation should cover why species in different habitats might (or might not) have different amounts of P in their leaves. (25 marks)

It was found that species in floodplains have a higher concentration of phosphorous than in upland and generalist habitats. Phosphorous is a main leaf nutrient and is absorbed by the roots and transported to the xylem of leaves. The reason in the difference in phosphorous concentration may be due to the soil quality of these habitats. Floodplains are subject to erosion from flooding meaning that phosphorous in the soil can re-enter the top soil. Whereas in upland habitats and generalist this type of event does not occur so the cycling of the soil and phosphorous within is at a much slower rate.

NB: There are some missing values (NA) for some species for these variables. R will automatically ignore these missing data points when conducting analyses, but it is good to be aware of this. Some of the subsequent variables examined may also have missing data points.

**Exercise 3: Multiple explanatory variables (100 marks)**

1. Make a plot of leaf phosphorous concentrations versus leaf carbon concentrations (with leaf phosphorous on the y-axis). Use different symbols for species in each habitat category (floodplain, upland and generalist), and place a best fit trendline (linear) on the plot for each group of species. Let me know in the figure legend (the text at the bottom of the figure) which symbols and lines belong to each group. (25 marks)

Chart

Description automatically generated

Figure 4 : Scatter plot of the interaction between carbon concentrations and phosphorous concentrations in leaves in 3 different habitats.

1. Which groups of species show a similar pattern and which group of species shows a divergent pattern. Create a new categorical variable that categorises all species into just two categories in a sensible way. Tell me what those categories are. Then construct a statistical model where you have both habitat group and leaf carbon concentration as predictors of leaf phosphorous concentrations. You can either include an interaction term or not, but please justify this choice. Now, run an analysis of variance on this statistical model and give me the results for each term. (30 marks)

The groups generalist and upland show a similar pattern (carbon increases so does phosphorous) whereas floodplain shows a divergent patter with a decreasing phosphorous trend as carbon rises.

Model 1 <- lm(P\_Leaf ~ C\_Leaf + New\_Habitat, data = inga\_new)

LM: F3,26 = 5.563, p = 0.004371

AIC = -116.87

Model 2 <- lm(P\_Leaf ~ C\_Leaf \* New\_Habitat, data = inga\_new)

LM: F5,24 = 4.505, p = 0.00489

AIC = -117.8491

Second is better model by stating habitat as an interaction term.

1. Evaluate your statistical model using diagnostic plots. Do not present the diagnostic plots themselves, but explain any issues you might have found with your statistical model. How would you manage any potential issues, i.e. how would you amend your statistical analysis to deal with these issues? Please do so and give the revised results for the analysis of variance. (20 marks)

The residuals vs fitted diagnostic plot shows no linear relationship so it does not meet the regression assumptions well. The normal Q-Q plot shows residuals lined well on the dotted line meaning that residuals are normally distributed. The scale location plot shows if residuals are spread equally along the ranges of predictors. In ours points appear to not have randomly + equally spread points. The residuals vs leverage plot shows if there are any influential cases, because no data points are out of cooks lines none of the data points need to be removed to improve model.

The main issues here are the residuals vs fitted and scale location plots. I believe this is because I’ve used a linear model when the data is not normally distributed

1. In non-statistical terms, please describe your analysis and what the results mean for the biology of Inga species. (25 marks)

Increasing carbon levels are good for the phosphorous concentration in generalist and upland but bad for floodplain species leaves. This means that floodplain species will start to lack nutrients as carbon concentration increases.

**Exercise 4: Generalised Linear Models (100 marks)**

1. Now let’s try and understand variation in the presence versus absence of one of the chemical defences in this dataset, specifically mevalonic acid. Investment in chemical defence may trade off with investment in other defences. One mechanism of defence involves having leaves that expand quickly. Once leaves are expanded, they can harden and be harder for herbivores to eat (e.g. have you ever noticed that freshly flushed leaves in the spring are a bit limp or weak?). Another mechanism of defence involves having hairs on the leaves. If there is a high density of hairs, it might be difficult for herbivores to actually crawl around on a leaf and eat it. Construct separate generalised linear models that individually test the influence of leaf expansion rate and leaf trichome density on whether or not leaves produce the defence chemical mevalonic acid (1 = yes, 0 = no). Based on your evaluation of these models, do you think either variable has a strong influence on whether or not trees produce mevalonic acid? (25 marks)

Leaf expansion has a slight positive interaction and trichome density has a negative between them and acid. Though neither have a significant interaction with the acid so you cannot say for sure there is a relationship there.

leaf expansion rate = 1 AIC: 44.09, p = 0.1070

leaf trichome density = 0 AIC: 42.77, p = 0.232

1. Now construct a model incorporating both expansion rate and trichome density to explain whether or not trees produce mevalonic acid in their leaves. Has assessing these models with multiple explanatory variables changed your understanding from the univariate analyses in part a? Why or why not? (25 marks)

Adding both interactions together produces a better model fit AIC: 44.188, p = 0.999 With an negative interaction but the P value goes up showing there is no significant result there

1. Explain in simple terms what your results mean? Was your expectation met, that there are tradeoffs between investing in different types of herbivore defence? (25 marks)

I think that there is a tradeoff for different leaf defences that maybe the effort required for a defence means other defences are unable to occur. From my results I have found that expansion rate of the leaf size and producing mevalonic acid have a positive interaction meaning that they can occur in the same leaf. Whereas trichome density (hairs on leaves) and mevalonic acid have a negative interaction meaning that prevalence of hairs on leaves means that acid is not created. Though it should be noted no statistically significant relationship was found. This may be due to there not being an interaction or lack of data due to missing values. So, to further this understanding more samples should be taken. Additionally, I believe there might be something else influencing the tradeoffs between herbivore defence.

1. Now visualise your results. Make a figure that shows how one or both of your predictor variables influence your response variable (presence vs. absence of mevalonic acid in leaves), and present that here. (25 marks)

Chart, scatter chart

Description automatically generated

Figure 5 : Presence and absence of mevalonic acid represented by red for present and blue for absent. Scatter points show higher trichome density influences the presence of acid.