***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)
2. Try log-transforming leaf area and make and present a histogram of log-transformed leaf area. (25 marks)
3. Now, in simple terms, how would you describe the distribution of leaf sizes across trees in this region to a non-scientist? (40 marks)

**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)
2. 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)

F= 8.598, P= 0,00129, DF = 2,27

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, so first is better

What are assumptions of ANOVA

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

Not sure

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)

Look for a reading maybe here but focus on what I’ve just answered

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)
2. 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)
3. 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)
4. In non-statistical terms, please describe your analysis and what the results mean for the biology of Inga species. (25 marks)

**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)
2. 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)
3. 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)
4. 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)

NB: Remember that to use information criteria (e.g. AIC) to compare models (as you should be doing here!), the response data need to be identical. This is a hint that different traits considered in this question are missing data for different species, and that you might need to restrict comparisons to a subset of data that has information for all the variables you are looking at. Thus, you will have to be very careful in your model comparison to be sure that you are always comparing AIC values for the same set of response data.