Professional Skills Statistical Assignment

25/Nov/2021

Exercise 1: Histograms and normality

A) 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?

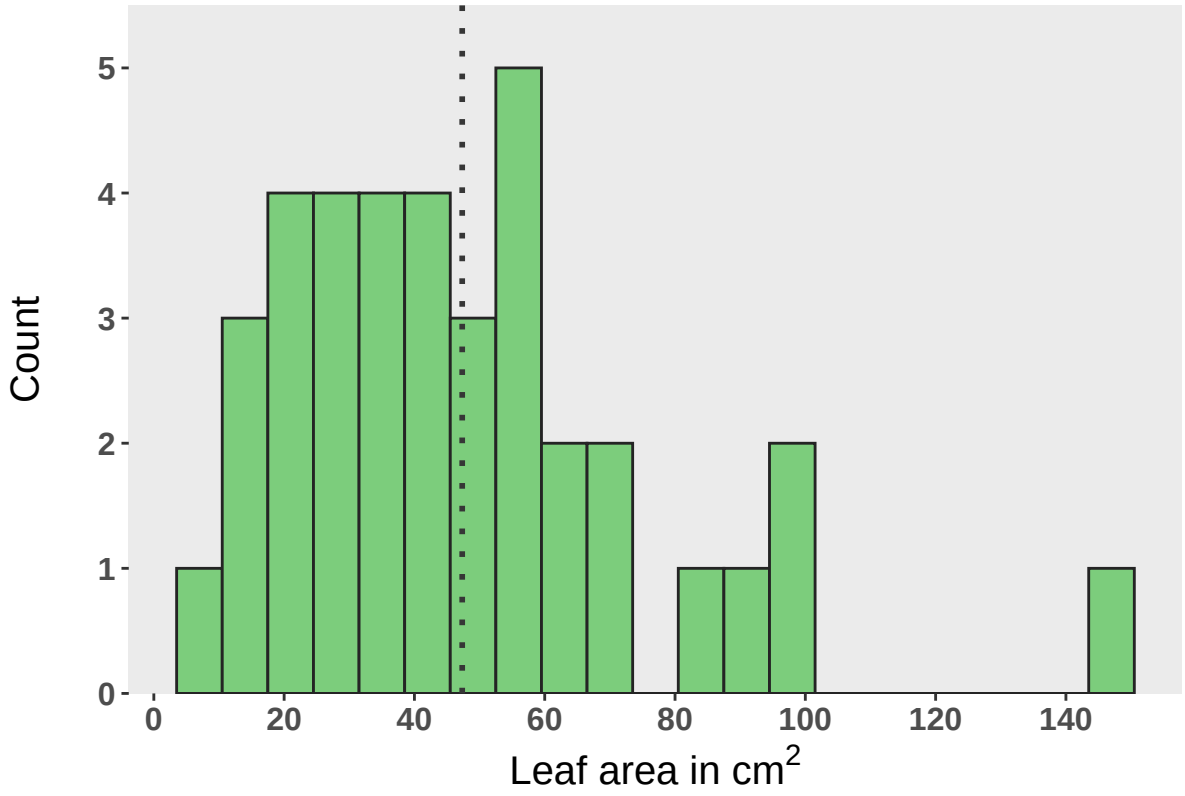

Figure 1: Histogram of leaf area (in cm^2) of tree species in the Amazon of Southeast Peru. The dotted line represents the mean leaf area.

The data appears to show a Poisson type distribution. That is the data is centered on the left and skewed towards the right as shown in Figure 1. Therefore, leaf size does not appear to be normally distributed.

B) Try log-transforming leaf area and make and present a histogram of log-transformed leaf area

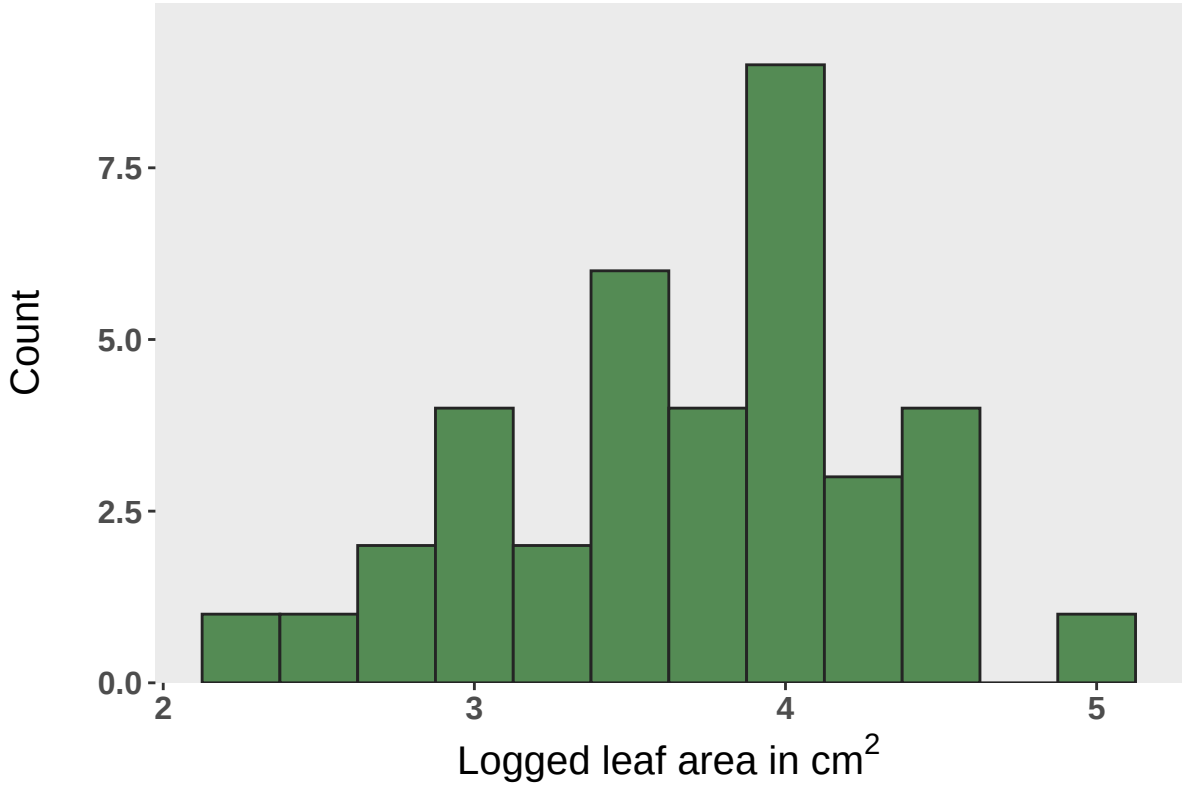

Figure 2: Histogram of logged leaf area (in cm^2) of tree species in the Amazon of Southeast Peru.

Figure 2 shows the distribution of leaf size after the data has been log-transformed. Such transformation is often usueful when the original data is not normally distributed (as seen in Figure 1). We can now see that the log-transfrom data show a more normal distribution.

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

In the Amazon of Southeast Peru, it seems like leaves have a mean area of 47 cm^2. This can be explained by most of the trees in the Amazon forest having leaves which sizes are between 10 and 60 cm^2. After the 60 cm^2 line, the proportion of leaves greater than this greatly decreases.

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

A) 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.

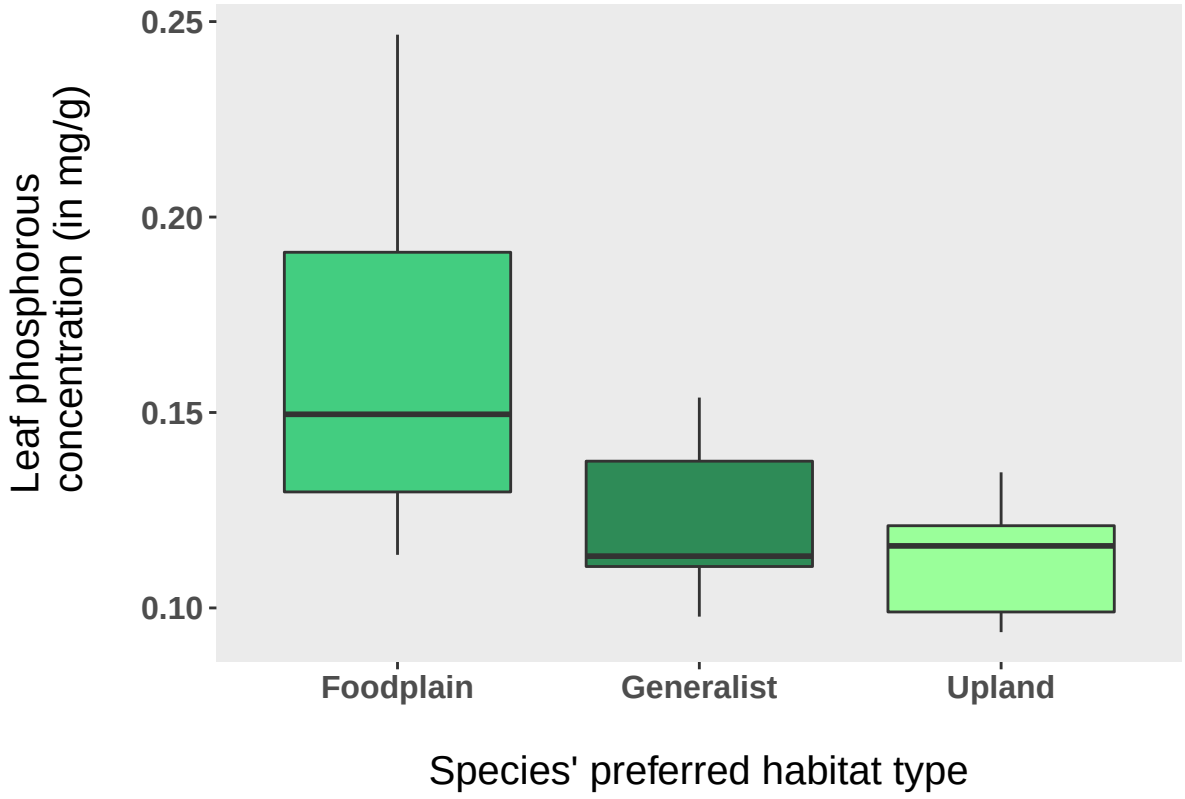

Figure 3: Boxplot representing leaf phosphorous concentration (in $mg.g^{-1}$) between species' preferred to habitat types.

Figure 3 shows that the concentration of phosphorous seems to be higher in leaves of tree species which are adapted to foodplain habitats and is lowest for tree species adapted to upland habitats. This could be due to the fact that food plain habitats get drained with rain that has ran through more land and thus that is richer in nutrients concentrations such as phosphorous.

B) 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.

We used the following model to test if species found in different habitats have significantly different phosphorous concentrations in their leaves:

```
phosph_lm <- lm(P_Leaf ~ Habitat, data = inga_traits)
```

Species found in different habitats have significantly different phosphorous concentrations in their leaves (ANOVA, $F_{2,27} = 8.60$, $p = 0.001$; Figure 4).

```
## Analysis of Variance Table
##
## Response: P_Leaf
##           Df   Sum Sq   Mean Sq F value   Pr(>F)
## Habitat    2 0.016347 0.0081737  8.5979 0.001291 **
## Residuals 27 0.025668 0.0009507
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 4: Results of ANOVA test between leaf phosphorous concentration and habitat type.

P-value: corresponds to the likelihood that the null-hypothesis (here being that leaf phosphorous content does not significantly differ between habitat types) is true. For p-vlaues under a pre-defined alpha-level (here we set alpha to be 0.05), we can reject the null-hypothesis and our results are therfore significant.

F-ratio: test statistic, calculated as: variance between groups (Mean Square Between) / variance within groups (Mean Squared Error). If the null hypothesis is correct, these two estimates of the variance should be close to the same and your F ratio should be near 1.0.

In this case the statistic is comparing the join effect of all habitats on leaf phosphorous concentration. As the p value is lower than 0.05, we can say that the habitat in which species are found significantly influences the leaf phosphorous content. As our F Statistic is greater than 1 we see more variation tree species in leaf phosphorous concentration between habitats than between.

C) 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?

As Figure 5 shows, the distribution of phosphorous concentration doesn't seem to be normally distributed. this is likely to create some problems in our model. We therfore perfom a Shapiro and a Bartlett test to see whether this was the case. Based on these tests, we find that the assuption of ANOVA of homoscedasticity (equal amount of variance) is not met.

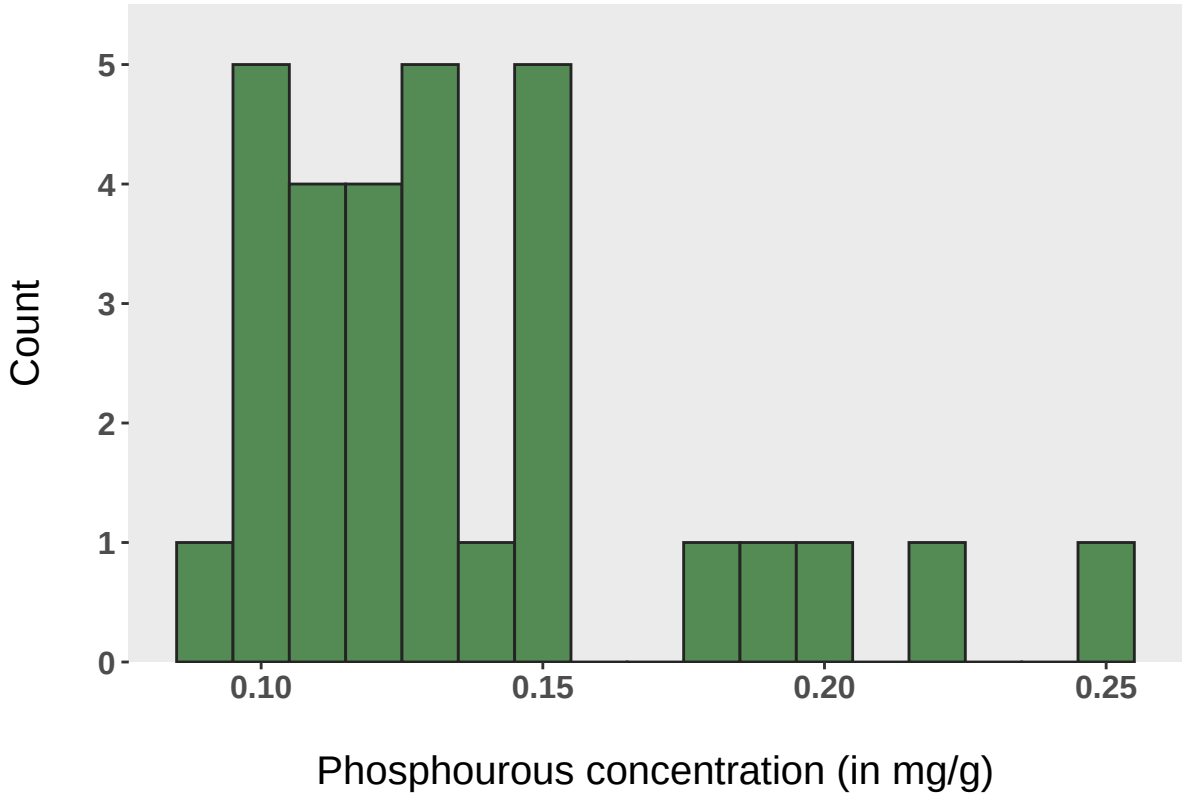

Figure 5: Histogram of phosphorous concentration (in mg/g) in leaves of trees from the Amazon of Southeast Peru.

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

As we have seen in question C, one of the reason why there was heteroscedasticity in our previous model is because the data we used in it was not normally distributed. This is something you expect when running a linear model. Mutliple transformations can be made to our data to make it normally distributed. Here, we logged transformed the data and used it to run our new model:

```
phosph_lm2 <- lm(log.phosph ~ Habitat, data = inga_traits)
```

Based on the results of our Shapiro and Bartlett test, no violation of ANOVA seem to have been met and we can be more confident of our results. Species found in different habitats therfore have significantly different phosphorous concentrations in their leaves (ANOVA, $F_{2,27} = 10.12$, $p = 0.0005$; Figure 6).

```
## Analysis of Variance Table
##
## Response: log.phosph
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## Habitat    2 0.79204 0.39602   10.12 0.0005251 ***
## Residuals 27 1.05658 0.03913
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 6: Results of ANOVA test between logged leaf phosphorous concentration and habitat type.

E) 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.

We wanted to see whether the concentration of phosphourous, a nutrient that is essential for plant growth, differed between species of trees that have different habitat preferences. Some species prefer floodplains or bottomlands, other prefer upland enviromnents and some don't mind both. To find out wether there is a difference in phosphorous conventration between these species we performed an ANOVA analysis and found that species that grow in different habitats show significant variations in their leaf phosphorous content. This could be explained by the fact that trees at the bottom of hills are fed by water that ran past a bigger proportion of land (the whole hill) when compared with upland trees which water doesn't run through much land. As water runs downhill, it picks up phosphours from leaves and soil on the ground and can therfore give more of this to trees at the bottom of hills.

Exercice 3: Multiple explanatory variables

A) Make a plot of leaf phosphorous concentrations versus leaf carbon concentrations (with leaf phosphorous on the y-axis).

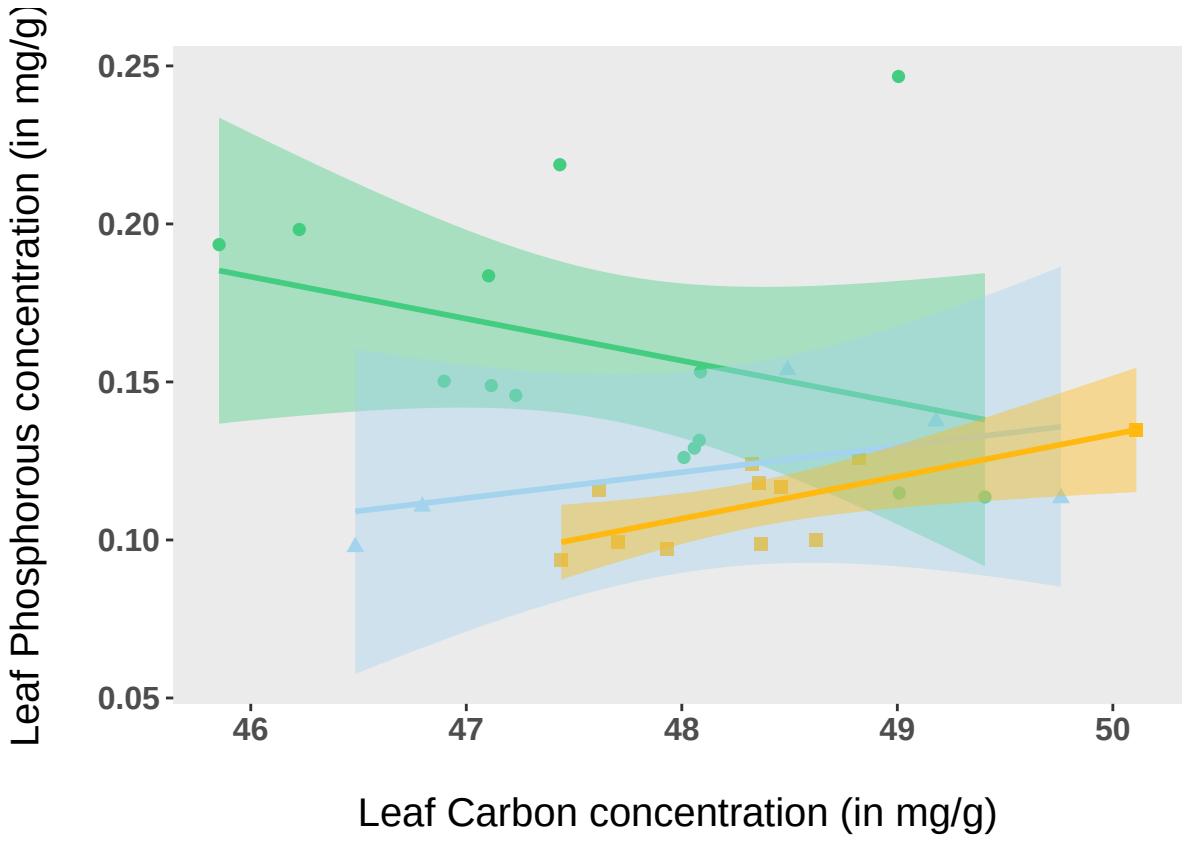

Figure 7: Scatter plot of lead carbon concentration (in mg/g) against leaf phophsphorous concentration (in mg/g) and associated best fit trend line and associated 95% confidence interval for each habitat category (foodplain in green, generalist in blue and upland in yellow).

B) 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.

Tree species that are generalist or that are found in upland habitats show a similar pattern between leaf carbon concentration and leaf phosphorous concentration as opposed to tree species found in bottomland habitats.

Based on this observation, we then created a new categorical variable with two categories: "Altitude adpated" and "Non-altitude adapted". We believe that species that are able to live in upland habitats might show similar adaptations that is reflected in their leaves' phosphorous and carbon concentration.

We then constructed a model to test for the effects of habitat group and leaf carbon concentration as predictors of leaf phosphorous concentration. As we have seen that the relationship bewteen phosphorous and carbon varies in the two habitats, we used an interaction term to allow for relationship between leaf phosphourous carbon concentration to differ between both habitats. The model code used can be seen below:

```
habitat_leaf_C_lm <-lm(P_Leaf ~ new_habitat*C_Leaf, data = inga_traits)
```

The leaf phosophorous concentration of trees in the Amazon of Southeast Peru is significantly influenced by leaf carbon concentration and habitat type together (ANCOVA, $F_{3,26} = 7.615$, $p = 0.0008$, Adj $R^2 = 0.41$).

C) 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.

After evaluating our statistical model using diagnostic plots, we can see that our residuals are not normally distributed around the model fit. We then tried to log-transform our data to see whether this helped. This gives a slightly better distribution of the residuals around the model fit.

```
habitat_leaf_C_lm2 <-lm(log.phosph ~ new_habitat*C_Leaf, data = inga_traits)
```

Leaf phosphorous concentration is significantly influenced by leaf carbon concetration and by habitat type (ANCOVA, $F_{3,26} = 10.15$, $p = 0.0001$, Adj $R^2 = 0.48$).

However, we cannot really trust the outputs of this test and we would have to use a non-parametric test since the redisuals are still not normally distributed around the model fit.

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

We first realized that the relationship between leaf phosphorous and carbon concentration was different based on whether trees are adapated to live in upland environments or not. This led us to test whether leaf phosphorous concentration for a leaf could be predicted based on whether it is adapted to live in upland habitat and its carbon concentration. We found that this was the case. This means that trees of the Inga species show specific adaptations to the habitat in which they can be found and that is reflected in the way they allocate phosphorous and carbon nutrients in their leaves. Studies have found that vegetation types, climate, and soil nutrient concentrations together can explain most of the variance in phosophorous, nitrogen and carbon in leaf tissues (Tang et al., 2017).

For trees in the Amazon of Southeast Peru, this could mean that trees that can be found in upland habitats and that invest in their structural defence (that is increased C concentration) are able to store more phosphours in their leaves as a result. On the contrary, investment in structural defences in trees solely adapted to bottomland environment might lead to reduced phosphorous concentration in their leaves.

Exercice 4: Generalised Linear Models

A) Now let's try and understand variation in the presence versus absence of one of the chemical defences in this dataset, specifically mevalonic acid. 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?

As the data we used for this model use a bionomial response type (either 1 or 0), we used a generalized linear model and specified the use of a binomial distibution wiht the formula "family = binomial".

Model 1: influence of leaf expansion rate on whether or not leaves produce the defence chemical mevalonic acic

```
MA_LER_glm <- glm(Mevalonic_Acid ~ Expansion, data = inga_traits_finite, family = binomial)
```

Model 1 seems to show a bit of overdispersion (which was calculated by dividing the residual deviance by the residual degrees of freedom). As we set the value to one when Mevalonic acid is present, our model indicates that as the rate of expansion of leaves increases, Mevalonic acid is more likely to be present (p = 0.05). However, this relationship is relatively small. Using a null model and the Akaike Information Criterion (AIC) to assess the explanatory power of our model, we found that using expansion rate as our explanotory variable increases our understanding of Mevalonic Acid but that this increase in explantory power is relatively weak (the difference in AIC between our model and the null model is between 2 and 4).

Model 2: influence of leaf trichome density on whether or not leaves produce the defence chemical mevalonic acid

```
MA_LTD_glm <- glm(Mevalonic_Acid ~ Trichome_Density, data = inga_traits_finite, family = binomial)
```

Model 2 also shows a bit of overdispersion. It seems that the density of hair on the upper leaf surface is negatively correlated to the presence of Mevalonic acid in leaves but this relationship is not significant (p = 0.20). Again, using a null model and the AIC to assess the explanatory power of our model, we found that using trichome density as our explanotory variable increases our understanding of Mevalonic Acid but that this increase in explantory power is relatively weak (the difference in AIC between our model and the null model is between 2 and 4).

Based on our evaluation of these models, the amount of overdispersion and the relativery weak explanatory power of our models means that we cannot say that either variables has a strong influence on whether or not trees produce melavonic acid.

B) 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?

Using AIC we first compared the explanatory power of using an interaction term to see which model performed better. Based on our results we selected the following model (not incoporating an interaction terms yields a lower AIC):

```
MA_both_glm <- glm(Mevalonic_Acid ~ Trichome_Density + Expansion, data = inga_traits_finite, family = bi
```

This model shows less overdispersion than the previous ones. We then used AIC to assess the explanatory power of this new model when compared to a null model. We found that using trichome density and expansion rate as our explanotory variables together increases our understanding of whether or not trees produce Mevalonic Acid as the difference in AIC between our new model and the null model is between 7 and 10 meaning that the models are really different.

It seems that the density of hair and the expansion rate of leaves together explain whether or not trees produce melavonic adic best. Based on our final model, as the rate of expansion of leaves increases, Mevalonic acid is more likely to be present ($p = 0.03$) but trichome density doesn't seem to influence whether or not trees produce mevalonic acid ($p = 0.26$).Assessing the model with multiple explanatory variables has therefore not changed our understanding from the univariate analyses in question a but has made us more confident in interpreting our results.

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

Plants harbour a wide range of defences to ensure their survival and can take the form of chemical and structural defences. however, llants often have to trade-off between different forms of investments. An example of plant chemical defence found in trees of the Amazon of Southeast Peru is melavonic acid. Here, we wanted to assess whether plants have to make tradeoff with other defences when producing melavonic acid. To do so, we built models to separatly test for the influence of leaf expansion rate (a physical defense) and leaf hair density (another physical defense) on whether or not trees produce melavonic acid. Based on our analysis, we found that when assessed as independent varibles in a model expansion rate and hair density best explain whether or not trees produce melavonic acid. However, it does not seem that there is a trade-off between producing melavonic acid and investing in physical defences. As expansion rate increases, melavonic is significantly more likely to be present and the relationship between hair density and melavonic acid is not significant.

D) 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.

Figure 8 represents the relationship between leaf expansion rate and the presence or absence of Mevalonic Acid in leaves.

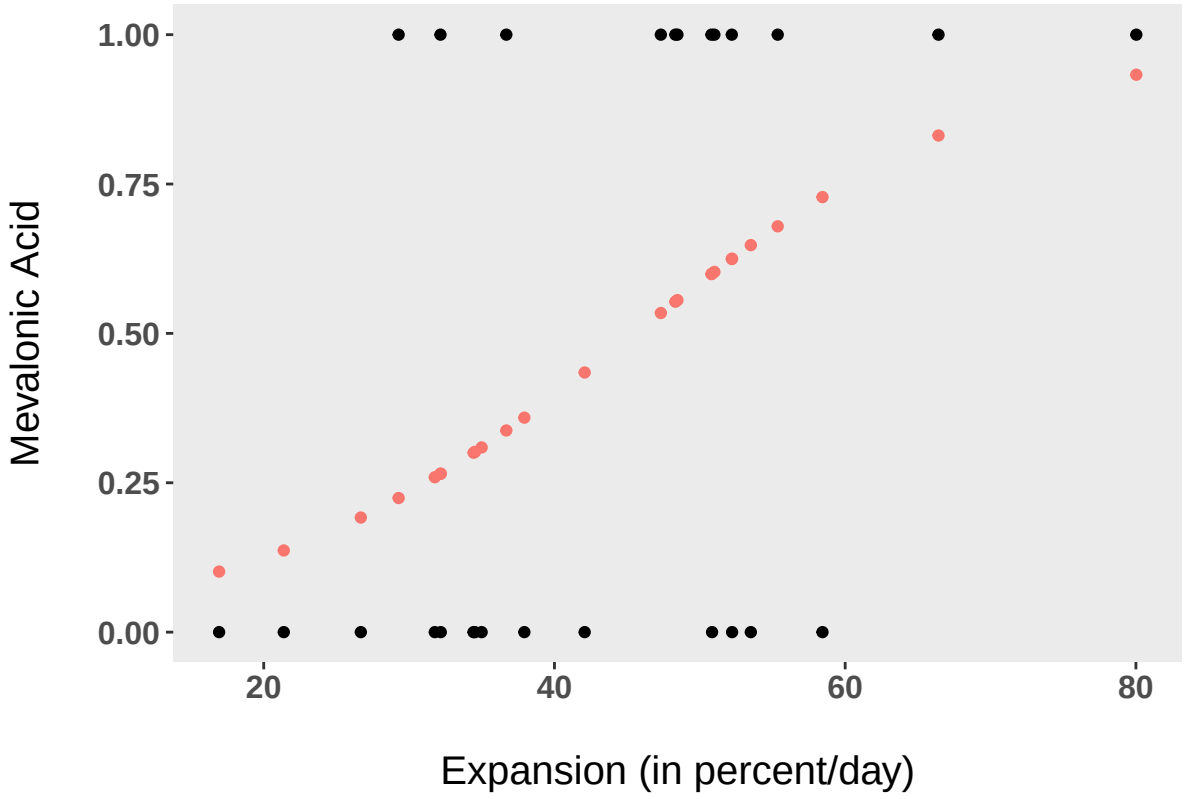

Figure 8: Visualization of the relationship between leaf expansion rate (in percent/day) and the presence (=1) or absence (=0) of Mevalonic Acid. The red line represents the model fit of this relationship.

EXTRA

The figure below also shows the variation in the rate of leaf expansion when compated between presence and absence of Mevalonic acid. The figure incorporates a violon plot showing where the data is mostly concentrated, a boxplot and the actual data as points.

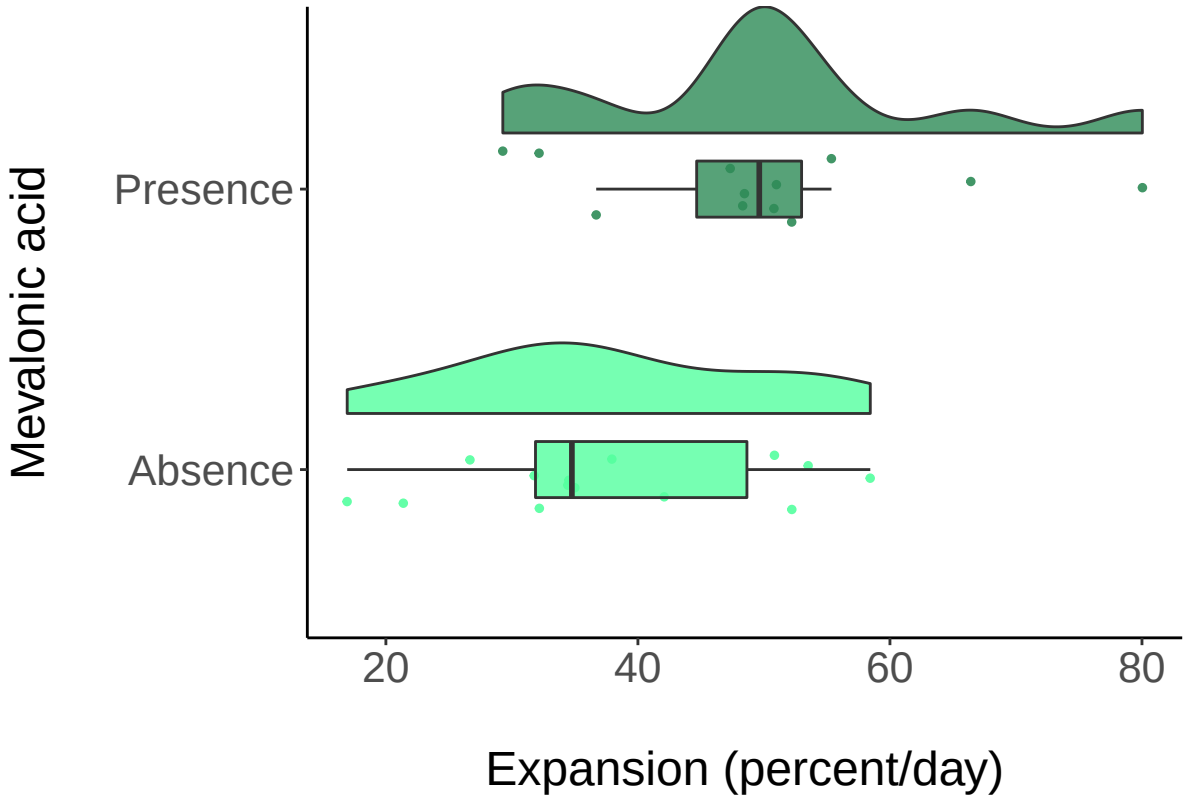

Figure 9: Cloud plot showing how the absence or presence of Mevalonic Acid influences on leaf expansion rate (in percent/day).