Analytical Interpretation of Biological Data

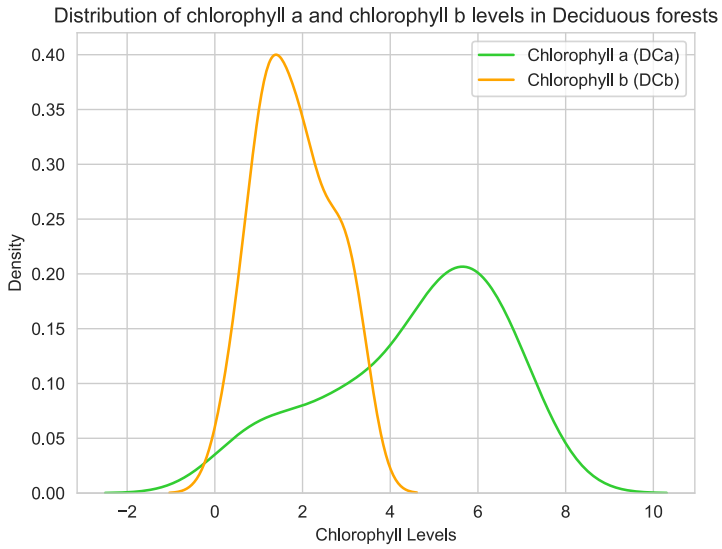
BE21B037 - Assignment 1

1 Introduction

You are provided with a dataset comprising measurements of chlorophyll levels (chlorophyll a and chlorophyll b) from 100 plant samples collected from two types of forests. Fifty samples were taken from Deciduous forests and the remaining fifty from Evergreen forests. Perform the below mentioned analysis on the given data:

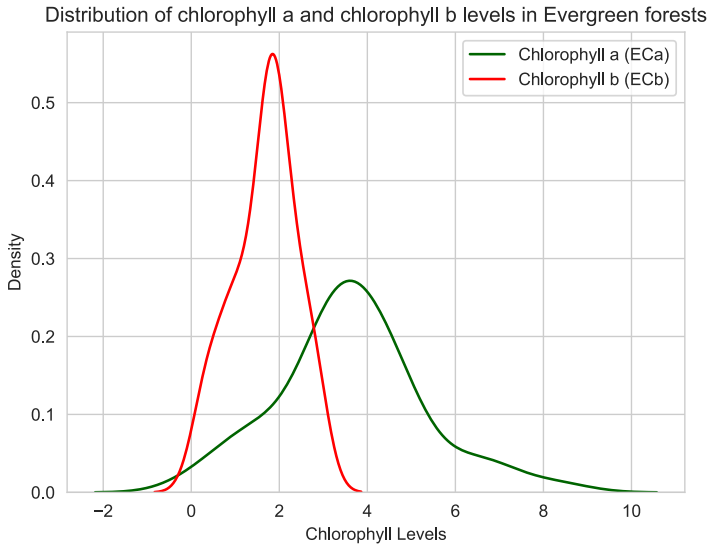
2 Questions

1. **Visualize the distribution of chlorophyll a and chlorophyll b values using histogram or density plots in Deciduous forests**



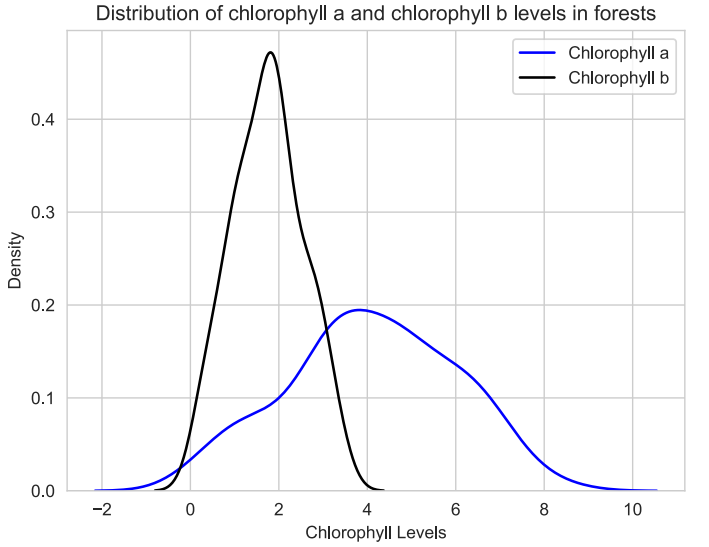
We observe see that the mean 'chlorophyll a' level is greater than the mean 'chlorophyll b' in the given sample. 'Chlorophyll a' levels are more evenly distributed than 'b' in Deciduous Forests

2. **Visualize the distribution of chlorophyll a and chlorophyll b values using histogram or density plots in Evergreen forests**



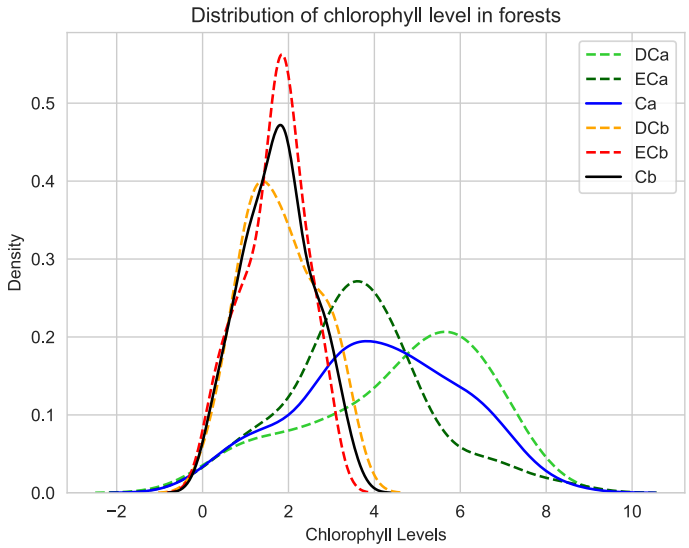
We observe see that the mean 'chlorophyll a' level is greater than the mean 'chlorophyll b' in the given sample. 'Chlorophyll a' levels are more evenly distributed than 'b' in Evergreen Forests as well

3. **Visualize the distribution of chlorophyll a and chlorophyll b values using histogram or density plots without separating the measurements from each forests**



We observe that regardless of the forest type, the mean 'chlorophyll a' level is greater than 'chlorophyll b' in the given sample data. 'Chlorophyll a' levels are more evenly spread out than that of 'b'

4. **Plot 1, 2 and 3 in same plot together and explore how it changes**



We observe that regardless of forest type, the mean 'chlorophyll a' level is greater than mean of 'chlorophyll b' level in this sample dataset. The curve resembles a bell shaped curve, but that does not necessarily imply the distribution of the population will be normal. We note that for the given sample the distribution for 'chlorophyll b' is more sharp and centered around the sample mean, while that of 'chlorophyll a' is more spread out. We can get a sense of the variance by observing the graph, since 'chlorophyll a' is more spread out than 'b', its variance will be more for this particular sample. The next logical step would be to analyze the means & variances and then perform hypothesis testing for the same

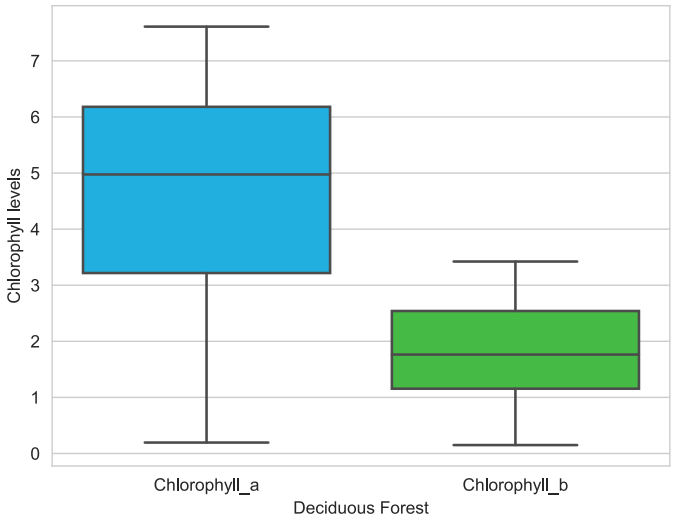
5. **Calculate summary statistics (mean, median, mode and standard deviation) of chlorophyll a and chlorophyll b measurements from Deciduous forests separately, Evergreen forests separately and both the forests together**

The following table includes the summary statistics for the given sample

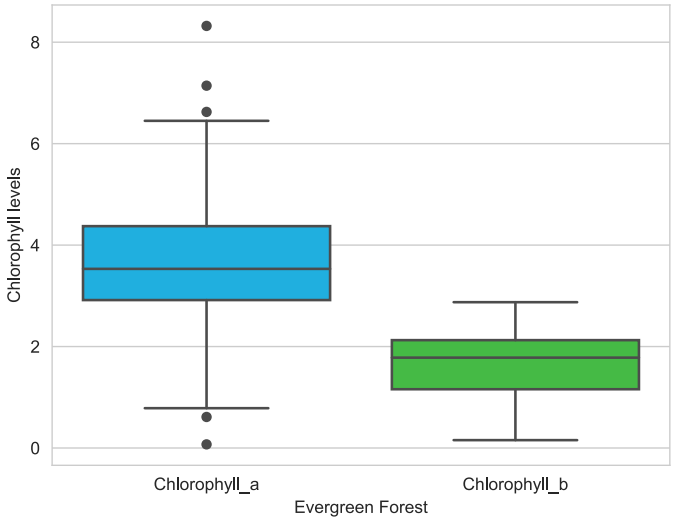
	Deciduous		**Evergreen**		**Both**	
	chlorophyll a	chlorophyll b	chlorophyll a	chlorophyll b	chlorophyll a	chlorophyll b
Mean	4.564366	1.817811	3.625706	1.66338	4.095036	1.742074
Median	4.974721	1.763595	3.531464	1.781322	4.098685	1.769393
Mode	–	–	–	–	–	–
Std. dev	1.964399	0.863713	1.642608	0.718783	1.862227	0.794192

Table 1: Summary statistics

6. **In a same box plot, compare how the distribution of chlorophyll a and chlorophyll b values compare in Deciduous forests and Evergreen forests**



In **Deciduous forests** we observe that more than 75% of the 'chlorophyll a' values are greater than 'chlorophyll b'. Comparing the IQR of both the values, we can conclude that 'a' is more spread out than 'b'. There is a slight negative skewness in 'a', while there is a slight positive skewness in 'b'.



In **Evergreen forests** too we observe that more than 75% of the 'chlorophyll a' values are greater than 'chlorophyll b'. The IQR of both the values are similar but there is a slight positive skewness in 'a', while there is a slight negative skewness in 'b'. The sample of 'chlorophyll a' include a few outliers as well.

7. **Are the variances between chlorophyll a and chlorophyll b measurements differ significantly? Perform appropriate statistical tests to support your claim. Compare variances of chlorophyll content from Deciduous forests separately, Evergreen forests separately and both the forests together**

	Deciduous		**Evergreen**		**Both**	
	chlorophyll a	chlorophyll b	chlorophyll a	chlorophyll b	chlorophyll a	chlorophyll b
Variance(σ^2)	3.858863	0.746	2.698161	0.516649	3.467891	0.630741

Table 2: Variance values

The variance differs significantly between 'chlorophyll a' and 'b' in all three cases. To check this we use the *F-test* for sample variances. We denote the variance of 'chlorophyll a' as σ_a^2 and 'chlorophyll b' as σ_b^2

Deciduous:

$H_0 : \sigma_a^2 = \sigma_b^2, H_A : \sigma_a^2 \neq \sigma_b^2$

$\sigma_a^2 = 3.858863, v_1 = 50 - 1 = 49$

$\sigma_b^2 = 0.746, v_2 = 50 - 1 = 49$

Significance value $\alpha = 5\%$

$F^{'} = \frac{\sigma_a^2}{\sigma_b^2} = \frac{3.858863}{0.746} = 5.172739$

$F_{\alpha v1 v2} = 1.607289$

Since $F^{'} > F_{\alpha v1 v2}$, **reject** H_0 at 0.05 level of significance and we conclude that the variances are not equal in Deciduous forests 'chlorophyll a' and 'b' values.

Evergreen:

$H_0 : \sigma_a^2 = \sigma_b^2, H_A : \sigma_a^2 \neq \sigma_b^2$

$\sigma_a^2 = 2.698161, v_1 = 50 - 1 = 49$

$\sigma_b^2 = 0.516649, v_2 = 50 - 1 = 49$

Significance value $\alpha = 5\%$

$F^{'} = \frac{\sigma_a^2}{\sigma_b^2} = \frac{2.698161}{0.516649} = 5.222425$

$F_{\alpha v1 v2} = 1.607289$

Since $F^{'} > F_{\alpha v1 v2}$, **reject** H_0 at 0.05 level of significance and we conclude that the variances are not equal in Evergreen forests 'chlorophyll a' and 'b' values.

Both:

$H_0 : \sigma_a^2 = \sigma_b^2, H_A : \sigma_a^2 \neq \sigma_b^2$

$\sigma_a^2 = 3.467891, v_1 = 100 - 1 = 99$

$\sigma_b^2 = 0.630741, v_2 = 100 - 1 = 99$

Significance value $\alpha = 5\%$

$F^{'} = \frac{\sigma_a^2}{\sigma_b^2} = \frac{3.467891}{0.630741} = 5.498122$

$F_{\alpha v1 v2} = 1.394061$

Since $F^{'} > F_{\alpha v1 v2}$, **reject** H_0 at 0.05 level of significance and we conclude that the variances are not equal in both forests 'chlorophyll a' and 'b' values.

8. **Finally, test whether the mean of chlorophyll a is greater than mean of chlorophyll b using appropriate statistical test in all three combinations: Deciduous forests separately, Evergreen forests separately and both the forests together. Clearly state your null hypothesis, chosen significance criteria and the result of hypothesis testing.**

	Deciduous		**Evergreen**		**Both**	
	chlorophyll a	chlorophyll b	chlorophyll a	chlorophyll b	chlorophyll a	chlorophyll b
Mean(μ)	4.564366	1.817811	3.625706	1.666338	4.095036	1.742074

Table 3: Mean values

We see that the mean of 'chlorophyll a' is greater than 'chlorophyll b' in all three cases. Since the 'chlorophyll a' and 'chlorophyll b' values are taken from a particular tree from a particular forest, their levels could vary based on the other's. So the right statistical to use will be the paired t-test. We denote the mean of 'chlorophyll a' as μ_a and 'chlorophyll b' as μ_b. The degree of freedom for individual forest types will be 49 and when analyzed together, it will be 99. A significance value of 5% will be used. The following is the null and alternate hypothesis for all the three cases

$H_0 : \mu_a \leq \mu_b, H_A : \mu_a > \mu_b$

The following are the p-values for each of the cases:

	p - value	**Conclusion**
Deciduous	$2.056196e^{-12}$	$p < 0.05$, hence reject H_0
Evergreen	$6.614613e^{-10}$	$p < 0.05$, hence reject H_0
Both	$1.904868e^{-20}$	$p < 0.05$, hence reject H_0

Table 4: Hypothesis Testing

We can conclude that the mean 'chlorophyll a' level of the population will be higher than the mean 'chlorophyll b' level

3 Appendix

The code to recreate all the above plots and testing can be found here.