ANOVA VERIFICA SI HAY DIFERENCIA ENTRE MEDIAS DE DIFERENTES MUESTRAS, UNA OPCION RAPIDA ES CON BOXPLOT, SE DEBE TENER CUIDADO QUE LAS MUESTRAS TENGAN IGUAL NUMERO DE CASOS Y QUE SEA SIGNIFICATIVA

1. 1 FACTOR – 4 GRUPOS

import pandas as pd

df = pd.read\_csv("https://reneshbedre.github.io/assets/posts/anova/onewayanova.txt", sep="\t")

# reshape the d dataframe suitable for statsmodels package

df\_melt = pd.melt(df.reset\_index(), id\_vars=['index'], value\_vars=['A', 'B', 'C', 'D'])

# replace column names

df\_melt.columns = ['index', 'treatments', 'value']

Graphical user interface, application

Description automatically generated

#BOXPLOT TO EXPLORE DIFFERENCES BETWEEN TREATMENTS

import matplotlib.pyplot as plt

import seaborn as sns

ax = sns.boxplot(x='treatments', y='value', data=df\_melt, color='#99c2a2')

ax = sns.swarmplot(x="treatments", y="value", data=df\_melt, color='#7d0013')

plt.show()

|  |  |
| --- | --- |
| Conclusiones  Hay diferencia visible entre medias y dispersión entre los 4 grupos.  El grupo A y C son similares, se puede explorar con mas detalle si la diferencia de medias entre ellos es significativo, por pares de medias, si fueran mas de 2 se usaría ANOVA. | Chart, box and whisker chart  Description automatically generated |

1. 2 FACTORES – 6 GRUPOS

import pandas as pd

import seaborn as sns

d = pd.read\_csv("https://reneshbedre.github.io/assets/posts/anova/twowayanova.txt", sep="\t")

# reshape the d dataframe suitable for statsmodels package

# you do not need to reshape if your data is already in stacked format.

# Compare d and d\_melt tables for detail

# understanding

d\_melt = pd.melt(d, id\_vars=['Genotype'], value\_vars=['1\_year', '2\_year', '3\_year'])

# replace column names

d\_melt.columns = ['Genotype', 'years', 'value']

d\_melt.head()

# output

Genotype years value

0 A 1\_year 1.53

1 A 1\_year 1.83

2 A 1\_year 1.38

3 B 1\_year 3.60

4 B 1\_year 2.94

As there are 6 and 3 levels for genotype and years, respectively, this is a 6 x 3 factorial design yielding 18 unique combinations for measurement of the response variable.

# boxplot to see the data distribution by genotypes and years.

# differences between different groups

sns.boxplot(x="Genotype", y="value", hue="years", data=d\_melt, palette="Set3")

|  |  |
| --- | --- |
| Conclusión  Lo mas afin es   * B y C en año 1 * A y E en año 1 * D y F en año 2 | Chart, box and whisker chart  Description automatically generated |

from statsmodels.graphics.factorplots import interaction\_plot

import matplotlib.pyplot as plt

fig = interaction\_plot(x=d\_melt['Genotype'], trace=d\_melt['years'], response=d\_melt['value'],

colors=['#4c061d','#d17a22', '#b4c292'])

plt.show()

Chart, line chart

Description automatically generated

* The interaction plot helps to visualize the means of the response of the two factors (Genotype and years) on one graph. Generally, the X-axis should have a factor with more levels.
* From the interaction plot, the interaction effect is significant between the Genotype and years because three lines are not parallel (roughly parallel factor lines indicate no interaction - additive model). This interaction is also called ordinal interaction as the lines do not cross each other.
* For a more reliable conclusion of the interaction plot, it should be verified with the *F* test for interaction