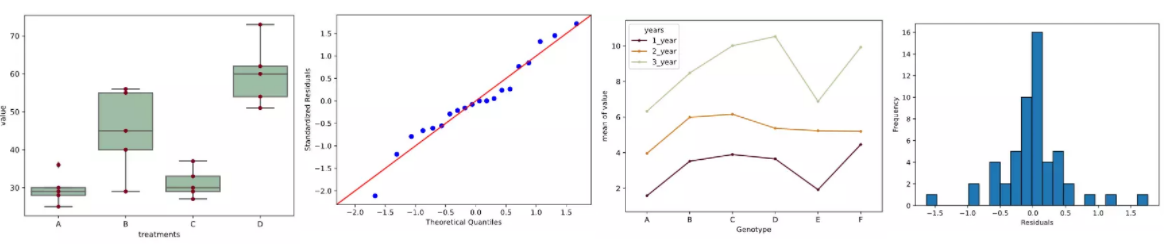
**ANOVA using Python (with examples)**

**What is ANOVA (ANalysis Of VAriance)?**

****

* ANOVA test used to compare the means of more than 2 groups (t-test can be used to compare 2 groups)
* Groups mean differences inferred by analyzing variances
* ANOVA uses variance-based *F* test to check the group mean equality. Sometimes, ANOVA *F* test is also called omnibus test as it tests non-specific null hypothesis i.e. all group means are equal
* Main types: One-way (one factor) and two-way (two factors) ANOVA (factor is an independent variable)
* It is also called univariate ANOVA as there is only one dependent variable in the model. [MANOVA](https://www.reneshbedre.com/blog/manova.html) is used when there are multiple dependent variables in the dataset. If there is an additional continuous [independent variable](https://www.reneshbedre.com/blog/manipulated-variable.html) in the model, then ANCOVA is used.
* If you have repeated measurements for treatments or time on same subjects, you should use [Repeated Measure ANOVA](https://www.reneshbedre.com/blog/repeated-measure-anova.html)

Note: In ANOVA, group, factors, and independent variables are similar terms

**ANOVA Hypotheses**

* *Null hypothesis*: Groups means are equal (no variation in means of groups)  
  H0: μ1=μ2=…=μp
* *Alternative hypothesis*: At least, one group mean is different from other groups  
  H1: All μ are not equal

*The null hypothesis is tested using the omnibus test (F test) for all groups, which is further followwd by post-hoc test to see individual group differences.*

**ANOVA Assumptions**

* [Residuals](https://www.reneshbedre.com/blog/learn-to-calculate-residuals-regression.html) (experimental error) are approximately normally distributed (Shapiro-Wilks test or histogram)
* homoscedasticity or Homogeneity of variances (variances are equal between treatment groups) (Levene’s or Bartlett’s Test)
* Observations are sampled independently from each other (no relation in observations between the groups and within the groups) i.e., each subject should have only one response
* The dependent variable should be [continuous](https://www.reneshbedre.com/blog/others.html). If the dependent variable is [ordinal or rank](https://www.reneshbedre.com/blog/others.html) (e.g. Likert item data), it is more likely to violate the assumptions of normality and homogeneity of variances. If these assumptions are violated, you should consider the non-parametric tests (e.g. [Mann-Whitney U test](https://www.reneshbedre.com/blog/mann-whitney-u-test.html), [Kruskal-Wallis test](https://www.reneshbedre.com/blog/kruskal-wallis-test.html)).

**How ANOVA works?**

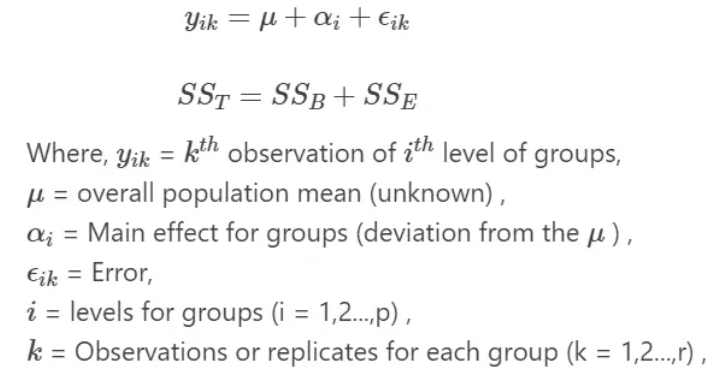
* Check sample sizes: equal number of observation in each group
* Calculate Mean Square for each group (MS) (SS of group/level-1); level-1 is a degrees of freedom (df) for a group
* Calculate Mean Square error (MSE) (SS error/df of residuals)
* Calculate *F* value (MS of group/MSE)
* Calculate *p* value based on *F* value and degrees of freedom (df)
* **One-way (one factor ANOVA) : 1 factor: treatment among 5 groups**
* **Two-way (two factor ANOVA : 2 factors: treatment – years, among 5 groups**

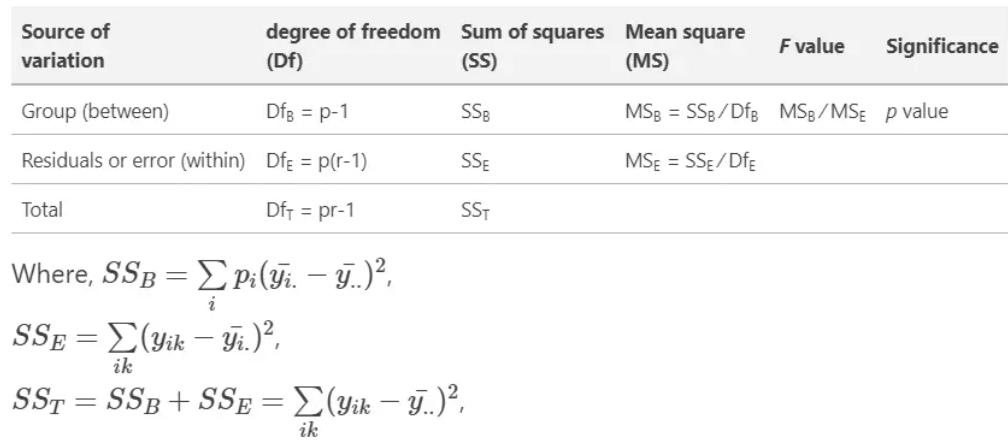
**One-way (one factor) ANOVA with Python**

**ANOVA effect model, table, and formula**

The ANOVA table represents between- and within-group sources of variation, and their associated degree of freedoms,

* the sum of squares (SS), and mean squares (MS).
* The total variation is the sum of between- and within-group variances.
* The F value is a ratio of between- and within-group mean squares (MS).
* p value is estimated from F value and degree of freedoms.





**ANOVA example**

Example data for one-way ANOVA analysis tutorial,

| **A** | **B** | **C** | **D** |
| --- | --- | --- | --- |
| 25 | 45 | 30 | 54 |
| 30 | 55 | 29 | 60 |
| 28 | 29 | 33 | 51 |
| 36 | 56 | 37 | 62 |
| 29 | 40 | 27 | 73 |

* There are four treatments (A, B, C, and D), which are groups for ANOVA analysis.
* Treatments are independent variable and termed as factor.
* As there are four types of treatments, treatment factor has four levels.
* There is only factor (treatments) or independent variable to evaluate,
* And therefore, one-way ANOVA method is suitable for analysis.

import pandas as pd

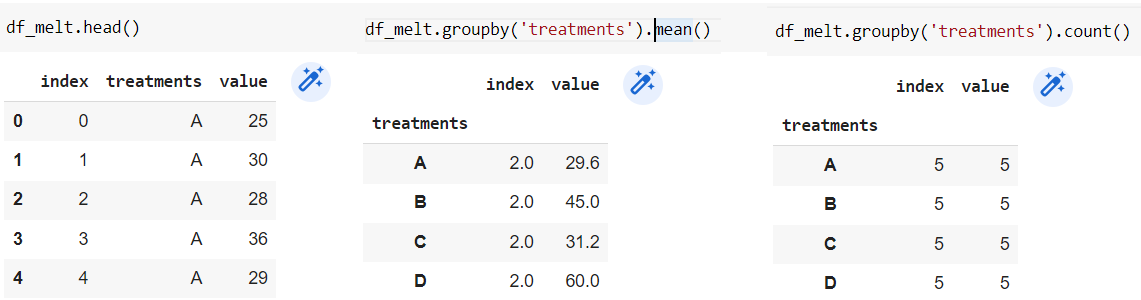
df = pd.read\_csv("https://reneshbedre.github.io/assets/posts/anova/onewayanova.txt", sep="\t")

# 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']



#BOXPLOT TO EXPLORE DIFFERENCES BETWEEN TREATMENTS

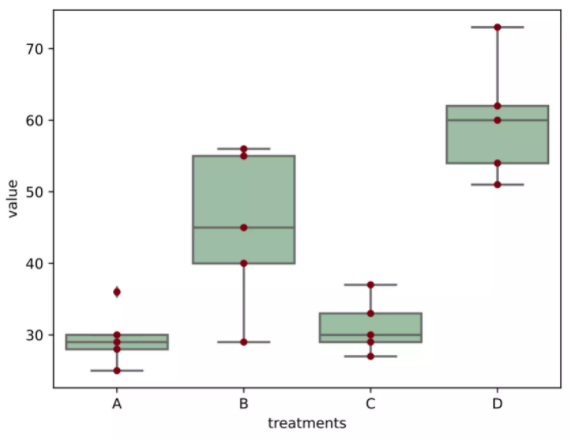
import matplotlib.pyplot as plt

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()



There are differences visible between means, however treatment A and C are similars, to explore later if there is a significance different between them.

**OPTION 1) F and p values**

#stat f\_oneway functions takes the groups as input and return ANOVA F and p value

import scipy.stats as stats

fvalue, pvalue = stats.f\_oneway(df['A'], df['B'], df['C'], df['D'])

print(fvalue, pvalue)

OUTPUT ANOVA F and p\_value

**17.492810 2.63924e-05**

**OPTION 2) ANOVA table as R like output**

import statsmodels.api as sm

from statsmodels.formula.api import ols

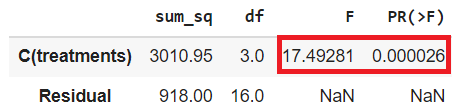
# Ordinary Least Squares (OLS) model

model = ols('value ~ C(treatments)', data=df\_melt).fit()

anova\_table = sm.stats.anova\_lm(model, typ=2)

anova\_table

OUTPUT ANOVA F AND P VALUE



**OPTION 3) ANOVA table USING bioinfokit v1.0.3**

# ANOVA table using bioinfokit v1.0.3 or later (it uses wrapper script for anova\_lm)

from bioinfokit.analys import stat

res = stat()

res.anova\_stat(df=df\_melt, res\_var='value', anova\_model='value ~ C(treatments)')

res.anova\_summary

# output (ANOVA F and p value)

df sum\_sq mean\_sq F PR(>F)

C(treatments) 3.0 3010.95 1003.650 17.49281 **0.000026**

Residual 16.0 918.00 57.375 NaN NaN

**Interpretation**

The *p* value obtained from ANOVA analysis is significant (*p* < 0.05), (0.000026 < 0.05)

Therefore, we conclude that there are significant differences among treatments.

We can confirm the interpretation with the boxplot.

Note on *F* value: *F* value is inversely related to *p* value and higher *F* value (greater than *F* critical value) indicates a significant *p* value.

**Note**: If you have unbalanced (unequal sample size for each group) data, you can perform similar steps as described for one-way ANOVA with balanced design (equal sample size for each group).

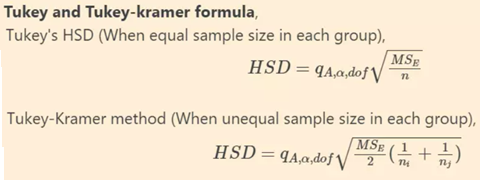
From ANOVA analysis, we know that treatment differences are statistically significant, but ANOVA does not tell which treatments are significantly different from each other.

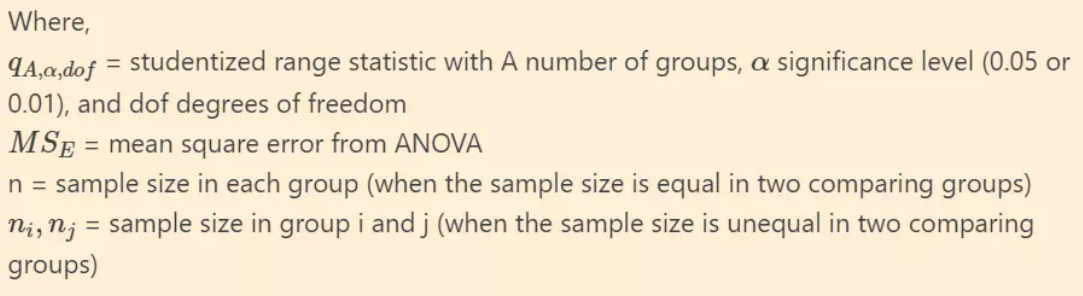
To know the pairs of significant different treatments, we will perform multiple pairwise comparison (**post hoc comparison**) analysis for all unplanned comparison using **Tukey’s honestly significantly differenced (HSD)** test.

**Note**: When the ANOVA is significant, post hoc tests are used to see differences between specific groups. post hoc tests control the family-wise error rate (inflated type I error rate) due to multiple comparisons. post hoc tests adjust the *p* values (Bonferroni correction) or critical value (Tukey's HSD test).

Tukey’s HSD test accounts for multiple comparisons and corrects for [family-wise error rate (FWER)](https://www.reneshbedre.com/blog/multiple-hypothesis-testing-corrections.html) (inflated [type I error](https://www.reneshbedre.com/blog/hypothesis-testing.html" \l "type-i-%CE%B1-type-ii-errors-%CE%B2-and-power-1-%CE%B2))

**Tukey and Tukey-kramer formula**





Alternatively, Scheffe’s method is completely coherent with ANOVA and considered as more appropriate post hoc test for significant ANOVA for all unplanned comparisons. However, it is highly conservative than other post hoc tests.

# we will use bioinfokit (v1.0.3 or later) for performing tukey HSD test

# check documentation here https://github.com/reneshbedre/bioinfokit

from bioinfokit.analys import stat

# perform multiple pairwise comparison (Tukey's HSD)

# unequal sample size data, tukey\_hsd uses Tukey-Kramer test

res = stat()

res.tukey\_hsd(df=df\_melt, res\_var='value', xfac\_var='treatments', anova\_model='value ~ C(treatments)')

res.tukey\_summary

# output

group1 group2 Diff Lower Upper q-value p-value

0 A B 15.4 1.692871 29.107129 4.546156 0.025070

1 A C 1.6 -12.107129 15.307129 0.472328 0.900000

2 A D 30.4 16.692871 44.107129 8.974231 0.001000

3 B C 13.8 0.092871 27.507129 4.073828 0.048178

4 B D 15.0 1.292871 28.707129 4.428074 0.029578

5 C D 28.8 15.092871 42.507129 8.501903 0.001000

# Note: p-value 0.001 from tukey\_hsd output should be interpreted as <=0.001

Above results from Tukey’s HSD suggests that except A-C, all other pairwise comparisons for treatments rejects null hypothesis (*p* < 0.05) and indicates statistical significant differences.

**Note**: Tukey's HSD test is conservative and increases the critical value to control the experimentwise type I error rate (or FWER). If you have a large number of comparisons (say > 10 or 20) to make using Tukey's test, there may be chances that you may not get significant results for all or expected pairs. If you are interested in only specific or few comparisons and you won't find significant differences using Tukey's test, you may split the data for specific comparisons or use the *t*-test

**Test ANOVA assumptions**

* ANOVA assumptions can be checked using test statistics (e.g. Shapiro-Wilk, Bartlett’s, Levene’s test) and the visual approaches such as residual plots (e.g. QQ-plots) and histograms.
* The visual approaches perform better than statistical tests. For example, the Shapiro-Wilk test has low power for small sample size data and deviates significantly from normality for large sample sizes (say n > 50). For large sample sizes, you should consider to use QQ-plot for normality assumption.

I will generate QQ-plot from standardized residuals ([outliers](https://www.reneshbedre.com/blog/find-outliers.html) can be easily detected from standardized residuals than normal residuals)

# QQ-plot

import statsmodels.api as sm

import matplotlib.pyplot as plt

Y1 = res.anova\_std\_residuals

Y2 = **res.anova\_model\_out.resid**

# res.anova\_std\_residuals are standardized residuals obtained from ANOVA (check above)

sm.qqplot(Y1, line='45')

plt.xlabel("Theoretical Quantiles")

plt.ylabel("Standardized Residuals")

plt.show()

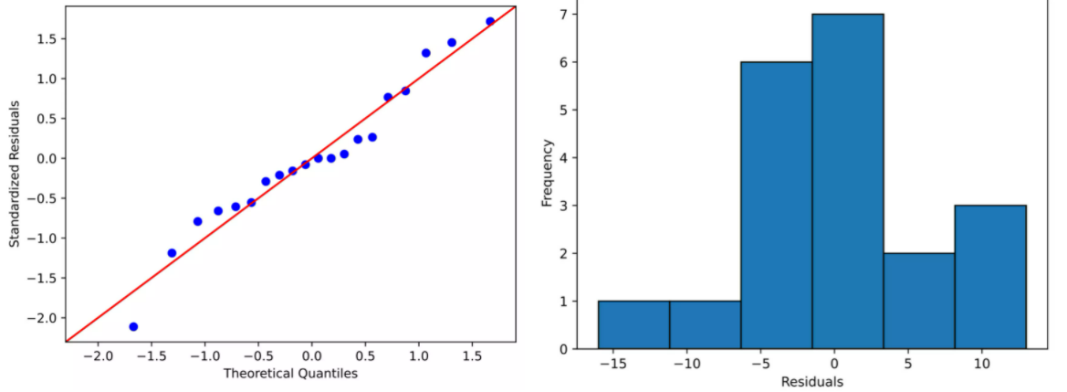
# histogram

plt.hist(**Y2**, bins='auto', histtype='bar', ec='k')

plt.xlabel("Residuals")

plt.ylabel('Frequency')

plt.show()



As the standardized residuals lie around the 45-degree line, it suggests that the residuals are approximately normally distributed

In the histogram, the distribution looks approximately normal and suggests that residuals are approximately normally distributed

**Shapiro-Wilk test** can be used to check the **normal distribution of residuals**.

*Null hypothesis*: data is drawn from normal distribution.

import scipy.stats as stats

w, pvalue = stats.shapiro(model.resid)

print(w, pvalue)

# 0.968501 0.722977

As the *p* value is non significant, we fail to reject null hypothesis and conclude that data is drawn from normal distribution.

As the data is drawn from normal distribution, use Bartlett’s test to check the **Homogeneity of variances**. *Null hypothesis*: samples from populations have equal variances.

import scipy.stats as stats

w, pvalue = stats.bartlett(df['A'], df['B'], df['C'], df['D'])

print(w, pvalue)

5.687843565012841 0.1278253399753447

# if you have a stacked table, you can use bioinfokit v1.0.3 or later for the bartlett's test

from bioinfokit.analys import stat

res = stat()

res.bartlett(df=df\_melt, res\_var='value', xfac\_var='treatments')

res.bartlett\_summary

# output

Parameter Value

0 Test statistics (T) 5.6878

1 Degrees of freedom (Df) 3.0000

2 p value **0.1278**

As the *p* value (0.12) is non significant, we fail to reject null hypothesis and conclude that treatments have equal variances.

**Levene’s test** can be used to check the Homogeneity of variances when the data is not drawn from normal distribution.

# if you have a stacked table, you can use bioinfokit v1.0.3 or later for the Levene's test

from bioinfokit.analys import stat

res = stat()

res.levene(df=df\_melt, res\_var='value', xfac\_var='treatments')

res.levene\_summary

# output

Parameter Value

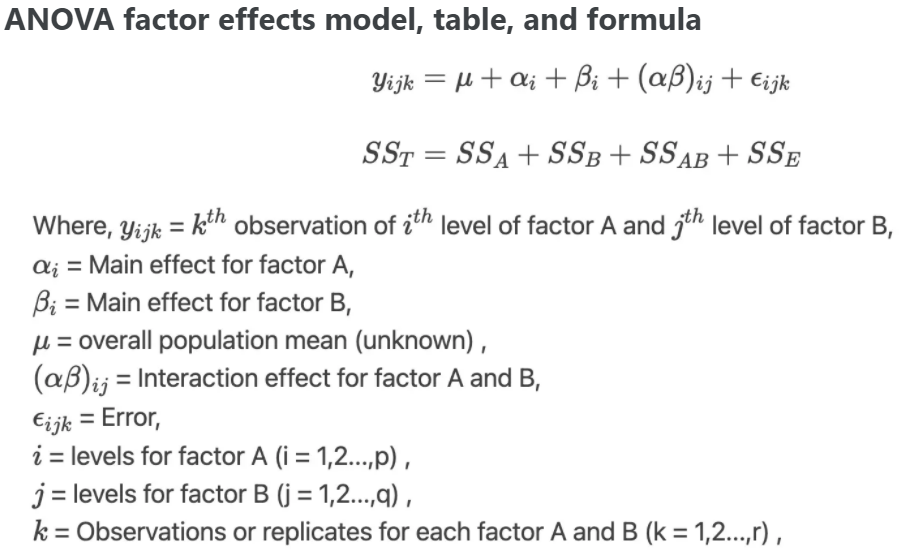
0 Test statistics (W) 1.9220

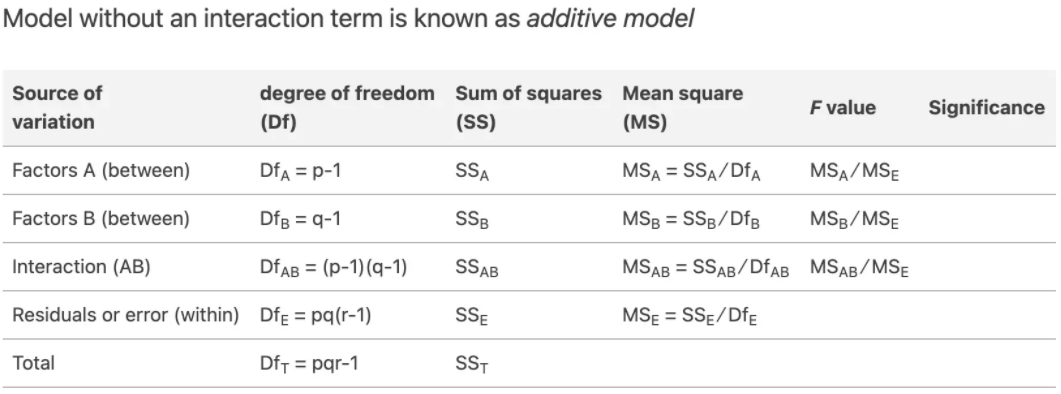
1 Degrees of freedom (Df) 3.0000

2 p value 0.1667

**Two-way (two factor) ANOVA (factorial design) with Python**

Here, I will discuss the two-way independent ANOVA, which differs from [two-way mixed ANOVA](https://www.reneshbedre.com/blog/mixed-anova.html) and [repeated measure ANOVA](https://www.reneshbedre.com/blog/repeated-measure-anova.html).





**ANOVA factor effects model, table, and formula**

Example data for two-way ANOVA analysis tutorial, [dataset](https://www.reneshbedre.com/assets/posts/anova/twowayanova.txt)

From dataset, there are two factors ([independent variables](https://www.reneshbedre.com/blog/manipulated-variable.html)) viz. genotypes and yield in years. Genotypes and years has six and three levels respectively (see one-way ANOVA to know factors and levels).

For this experimental design, there are two factors to evaluate, and therefore, two-way ANOVA method is suitable for analysis. Here, using two-way ANOVA, we can simultaneously evaluate how type of genotype and years affects the yields of plants. If you apply one-way ANOVA here, you can able to evaluate only one factor at a time.

From two-way ANOVA, we can tests three hypotheses 1) effect of genotype on yield 2) effect of time (years) on yield, and 3) effect of genotype and time (years) interactions on yield

Note: If you have your own dataset, you should import it as pandas dataframe. [Learn how to import data using pandas](https://www.reneshbedre.com/blog/import-data-pandas.html)

import pandas as pd

import seaborn as sns

# load data file

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

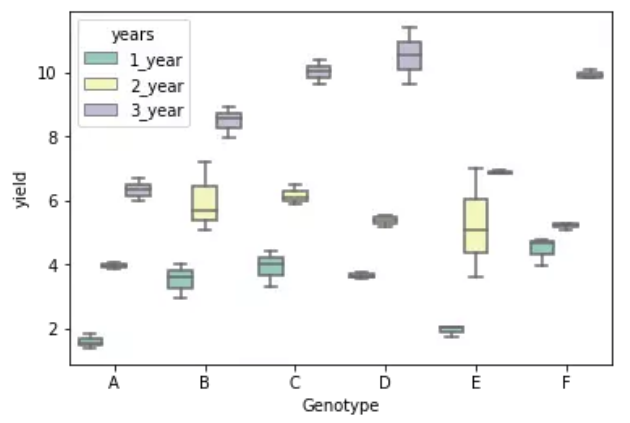
4 B 1\_year 2.94

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")



import statsmodels.api as sm

from statsmodels.formula.api import ols

model = ols('value ~ C(Genotype) + C(years) + C(Genotype):C(years)', data=d\_melt).fit()

anova\_table = sm.stats.anova\_lm(model, typ=2)

anova\_table

# output

sum\_sq df F PR(>F)

C(Genotype) 58.551733 5.0 32.748581 1.931655e-12

C(years) 278.925633 2.0 390.014868 4.006243e-25

C(Genotype):C(years) 17.122967 10.0 4.788525 2.230094e-04

Residual 12.873000 36.0 NaN NaN

# ANOVA table using bioinfokit v1.0.3 or later (it uses wrapper script for anova\_lm)

from bioinfokit.analys import stat

res = stat()

res.anova\_stat(df=d\_melt, res\_var='value', anova\_model='value~C(Genotype)+C(years)+C(Genotype):C(years)')

res.anova\_summary

# output

df sum\_sq mean\_sq F PR(>F)

C(Genotype) 5.0 58.551733 11.710347 32.748581 1.931655e-12

C(years) 2.0 278.925633 139.462817 390.014868 4.006243e-25

C(Genotype):C(years) 10.0 17.122967 1.712297 4.788525 2.230094e-04

Residual 36.0 12.873000 0.357583 NaN NaN

**Note**: If you have unbalanced (unequal sample size for each group) data, you can perform similar steps as described for two-way ANOVA with the balanced design but set `typ=3`. Type 3 sums of squares (SS) does not assume equal sample sizes among the groups and is recommended for an unbalanced design for multifactorial ANOVA.

**Interpretation**

The *p* value obtained from ANOVA analysis for genotype, years, and interaction are statistically significant (*p*<0.05). We conclude that type of genotype significantly affects the yield outcome, time (years) significantly affects the yield outcome, and interaction of both genotype and time (years) significantly affects the yield outcome.

As the interaction is significant, let’s visualize the interaction plot (also called profile plot) for interaction effects,

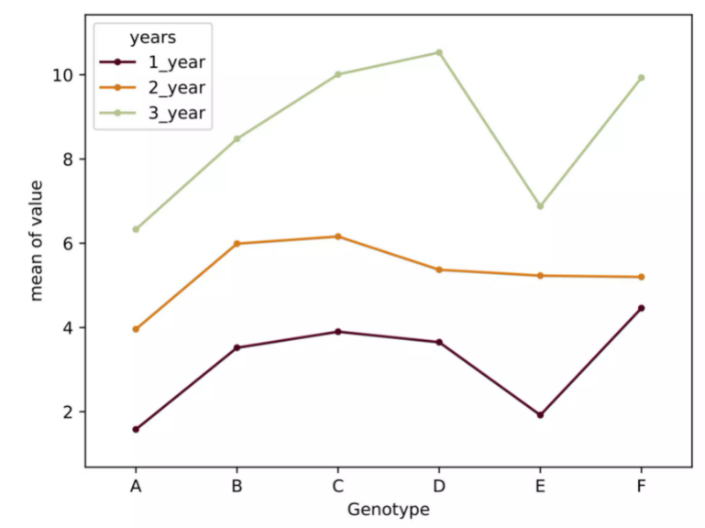
from statsmodels.graphics.factorplots import interaction\_plot

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()



* 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

**Multiple pairwise comparisons (Post-hoc test)**

Now, we know that genotype and time (years) differences are statistically significant, but ANOVA does not tell which genotype and time (years) are significantly different from each other. To know the pairs of significant different genotype and time (years), perform multiple pairwise comparison (**Post-hoc comparison**) analysis using **Tukey’s HSD** test.

# we will use bioinfokit (v1.0.3 or later) for performing tukey HSD test

# check documentation here https://github.com/reneshbedre/bioinfokit

from bioinfokit.analys import stat

# perform multiple pairwise comparison (Tukey HSD)

# unequal sample size data, tukey\_hsd uses Tukey-Kramer test

res = stat()

# for main effect Genotype

res.tukey\_hsd(df=d\_melt, res\_var='value', xfac\_var='Genotype', anova\_model='value~C(Genotype)+C(years)+C(Genotype):C(years)')

res.tukey\_summary

# output

group1 group2 Diff Lower Upper q-value p-value

0 A B 2.040000 1.191912 2.888088 10.234409 0.001000

1 A C 2.733333 1.885245 3.581421 13.712771 0.001000

2 A D 2.560000 1.711912 3.408088 12.843180 0.001000

3 A E 0.720000 -0.128088 1.568088 3.612145 0.135306

4 A F 2.573333 1.725245 3.421421 12.910072 0.001000

5 B C 0.693333 -0.154755 1.541421 3.478361 0.163609

6 B D 0.520000 -0.328088 1.368088 2.608771 0.453066

7 B E 1.320000 0.471912 2.168088 6.622265 0.001000

8 B F 0.533333 -0.314755 1.381421 2.675663 0.425189

9 C D 0.173333 -0.674755 1.021421 0.869590 0.900000

10 C E 2.013333 1.165245 2.861421 10.100626 0.001000

11 C F 0.160000 -0.688088 1.008088 0.802699 0.900000

12 D E 1.840000 0.991912 2.688088 9.231036 0.001000

13 D F 0.013333 -0.834755 0.861421 0.066892 0.900000

14 E F 1.853333 1.005245 2.701421 9.297928 0.001000

# Note: p-value 0.001 from tukey\_hsd output should be interpreted as <=0.001

# for main effect years

res.tukey\_hsd(df=d\_melt, res\_var='value', xfac\_var='years', anova\_model='value ~ C(Genotype) + C(years) + C(Genotype):C(years)')

res.tukey\_summary

# output

group1 group2 Diff Lower Upper q-value p-value

0 1\_year 2\_year 2.146667 1.659513 2.633821 15.230432 0.001

1 1\_year 3\_year 5.521667 5.034513 6.008821 39.175794 0.001

2 2\_year 3\_year 3.375000 2.887846 3.862154 23.945361 0.001

# for interaction effect between genotype and years

res.tukey\_hsd(df=d\_melt, res\_var='value', xfac\_var=['Genotype','years'], anova\_model='value ~ C(Genotype) + C(years) + C(Genotype):C(years)')

res.tukey\_summary.head()

# output

group1 group2 Diff Lower Upper q-value p-value

0 (A, 1\_year) (A, 2\_year) 2.38 0.548861 4.211139 6.893646 0.002439

1 (A, 1\_year) (A, 3\_year) 4.75 2.918861 6.581139 13.758326 0.001000

2 (A, 1\_year) (B, 1\_year) 1.94 0.108861 3.771139 5.619190 0.028673

3 (A, 1\_year) (B, 2\_year) 4.41 2.578861 6.241139 12.773520 0.001000

4 (A, 1\_year) (B, 3\_year) 6.90 5.068861 8.731139 19.985779 0.001000

**Test ANOVA assumptions**

Similar to one-way ANOVA, you can use visual approaches, **Bartlett’s** or **Levene’s**, and **Shapiro-Wilk test** to validate the assumptions for homogeneity of variances and normal distribution of residuals.

# QQ-plot

import statsmodels.api as sm

import matplotlib.pyplot as plt

# res.anova\_std\_residuals are standardized residuals obtained from two-way ANOVA (check above)

sm.qqplot(res.anova\_std\_residuals, line='45')

plt.xlabel("Theoretical Quantiles")

plt.ylabel("Standardized Residuals")

plt.show()

# histogram

plt.hist(res.anova\_model\_out.resid, bins='auto', histtype='bar', ec='k')

plt.xlabel("Residuals")

plt.ylabel('Frequency')

plt.show()

# Shapiro-Wilk test

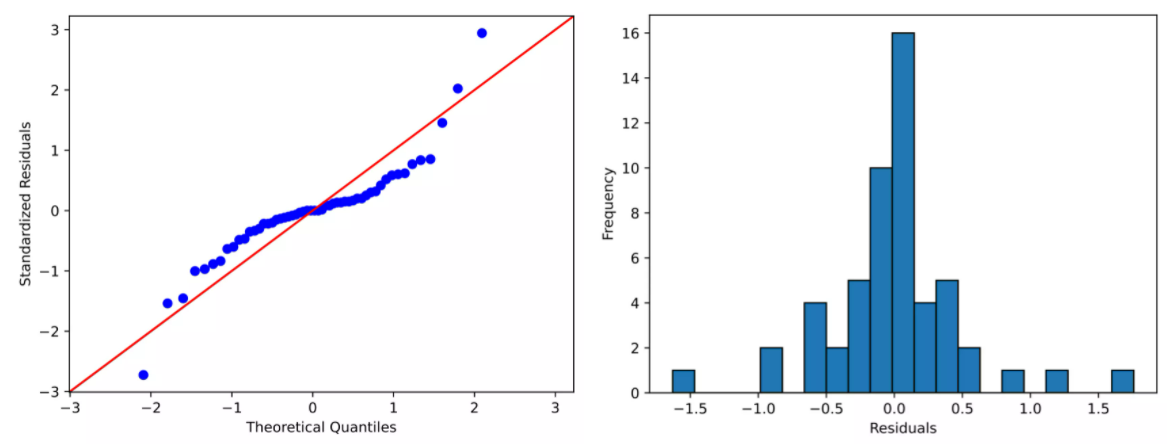
import scipy.stats as stats

w, pvalue = stats.shapiro(res.anova\_model\_out.resid)

print(w, pvalue)

0.8978844 0.0002398

Even though we rejected the Shapiro-Wilk test statistics (*p* < 0.05), we should further look for the residual plots and histograms. In the residual plot, standardized residuals lie around the 45-degree line, it suggests that the residuals are approximately normally distributed. Besides, the histogram shows the approximately normal distribution of residuals.



**Note**: The ANOVA model is remarkably robust to the violation of normality assumption, which means that it will have a non-significant effect on Type I error rate and *p* values will remain reliable as long as there are no outliers

We will use Levene’s test to check the assumption of homogeneity of variances,

# if you have a stacked table, you can use bioinfokit v1.0.3 or later for the Levene's test

from bioinfokit.analys import stat

res = stat()

res.levene(df=d\_melt, res\_var='value', xfac\_var=['Genotype', 'years'])

res.levene\_summary

# output

Parameter Value

0 Test statistics (W) 1.6849

1 Degrees of freedom (Df) 17.0000

2 p value 0.0927

As the *p* value (0.09) is non-significant, we fail to reject the null hypothesis and conclude that treatments have equal variances.