ONE WAY ANOVA python.

```
import numpy as np
import pandas as pd
import statsmodels.api as sm
from statsmodels.formula.api import ols.   } for the model part.
```

```
# data part :
data = pd.read_csv(' ')
data
```

```
# model part :
model = ols('dependent variable ~ indep var', data=data).fit()
anova_result = sm.stats.anova_lm(model, type=2)
print(anova_result)
```

→ gives ANOVA table. p-value will be there reject/accept Ho.

SSB → indep-var
SSW → residual

Tukey Test :

if we accept Ha, it means atleast one pair has significant difference. To determine where that difference is, we use Tukeys Test

Tukey's Test

```
from statsmodels.stats.multicomp import pairwise_tukeyhsd

tukey = pairwise_tukeyhsd(endog={dependent variable},
                          groups={indep var ← categorical var},
                          alpha=0.05)
tukey.summary()
```

OR `tukey._results_table`

as we used a dataframe,

```
endog = data['dep var name']
groups = data[' ']
```

it gives pair wise reject Ho

- if true, (it means reject Ho i.e. means are significantly different)
- if false (accept Ho ; i.e. means are nearer)

TWO WAY ANOVA (PYTHON)

```
import pandas as pd
import numpy as np
import statsmodels.api as sm
from statsmodels.formula.api import ols
import statsmodels.stats.multicomp as mc
from statsmodels.stats.multicomp import MultiComparison
from statsmodels.stats.multicomp import pairwise_tukeyhsd
from scipy import stats

# without replication.
# create dataframe
df = pd.DataFrame ({'Lab': np.repeat(['lab1','lab2','lab3','lab4'], 4),
                    'Sample': np.tile(np.repeat(['type1','type2','type3','type4'],1),4),
                    'Recovery': [ 16 observations ] })
df
# 2 way anova w/o replication
model = ols('Recovery ~ C(Lab) + C(Sample)', data=df).fit()
anova_result = sm.stats.anova_lm(model, type=2)
# multi comparison
mc = MultiComparison(df['Recovery'], df['Lab'])
mc_result = mc.tukeyhsd(0.05)
mc_result.summary()
```

repeat each input 4 times

if we do with replic^
+ C(Lab):C(Sample)

do comparison like this for each indep. var.

```
# 2way ANOVA with replication.
model = ols('depvar ~ C(indvar1) + C(indvar2) + C(indvar1):C(indvar2)', data=df).fit
```

interaction effect

M T W T F S S
Page No.: YOUVA
Date:

MANOVA python

```
import pandas as pd
import numpy as np
from statsmodels.formula.api import ols
from statsmodels.multivariate.manova import MANOVA
from statsmodels.stats.multicomp import pairwise_tukeyhsd
from scipy import stats
```

if 2 way manova:
('depvar1 + depvar2 ~ indepvar1 + indepvar2)
for anova part:
('depvar1 ~ indvar1 + ind2)

```
# data
df =
```

```
# test manova.
maov = MANOVA.from_formula('depvar1 + depvar2 + .. ~ indep.var', data = df)
print(maov.mv_test())
```

gives 2 tables: intercept, independent variable
see Wilks lambda. if $p < \alpha$, rej Ho.
as Ha is accepted i.e there is significant differences in manova, we saw combined effect, we have to now see for each dependent variable so we do anova.

* intercept is the estimate of the dependent variable when all the indep var are 0

```
# anova.
fit1 = ols('depvar1 ~ indep var', data = df).fit()
anova1 = sm.stats.anova_lm(fit1, type=2)
anova1
```

do similarly for the dependent variables

```
# multiple comparison.
mc1 = pairwise_tukeyhsd(df['depvar1'], df['indepvar'])
mc1.__result s__table.
```

do for all.

T-Tests in python.

A pair t-test (stats.ttest_rel) is the same as an independent samples t-test on the difference in 2 groups (stats.ttest_1samp)

Eg: before = np.array([, ..])
after = np.array([,])

→ stats.ttest.rel(before, after, alternative = ' ')

OR

difference = before − after

→ stats.ttest_1samp(difference, popmean = 0, alternative = ')

it gives p-value & t-statistic.

(2 sided p-value if alternative isn't mentioned)
test

$$SST = SSW + SSB$$

$$\sum_{i=1}^{r}\sum_{j=1}^{c}(x_{ij}-\bar{x})^2 = \sum_{j=1}^{c} n_j(\bar{x}_j-\bar{x})^2 + \sum_{i=1}^{m}\sum_{j=1}^{c}(x_{ij}-\bar{x}_j)^2$$

Page No.: 1 YOUVA
Date:

ANOVA ONE WAY (solving)

given: Wheat Production of 3 wheat varities in tonnes.

fields	A	B	C
F1	6	5	5
F2	7	5	4
F3	3	3	3
F4	8	7	4

factors affecting the result = 1 which is variety of wheat

Step 1: Ho and Ha.

→ Ho: $\mu_A = \mu_B = \mu_C$ (production of type A, B, C is same, it implies that the variety of wheat doesn't affect the production of wheat significantly)

→ Ha: atleast one of μ_A, μ_B, μ_C are significantly different. (it implies that variety of wheat affects the production of wheat significantly.

Step 2: SS Between (SSB). $SSB = n_1(\bar{x}_1-\bar{x})^2 + n_2(\bar{x}_2-\bar{x})^2 + \cdots$

A: $n_1 = 4$ $\bar{x}_1 = (6+7+3+8)/4 = 6$

B: $n_2 = 4$ $\bar{x}_2 = (5+5+3+7)/4 = 5$

C: $n_3 = 4$ $\bar{x}_3 = (5+4+3+4)/4 = 4$

Grand mean: $\bar{x} = (6+5+4)/3 = 5$

$\bar{x}_1 = 6$
$\bar{x}_2 = 5$
$\bar{x}_3 = 4$
$\bar{x} = 5$

$$SSB = n_1(\bar{x}_1-\bar{x})^2 + n_2(\bar{x}_2-\bar{x})^2 + n_3(\bar{x}_3-\bar{x})^2$$

$SSB = 4(6-5)^2 + 4(5-5)^2 + 4(4-5)^2 = 8$ • $SSB = 8$.

Step 3: SS Within (SSW).

$$SSW = \sum(x_{1i}-\bar{x}_1)^2 + \sum(x_{2i}-\bar{x}_2)^2 + \sum(x_{3i}-\bar{x}_3)^2$$

- find diff b/w each data peint & its column mean
- square that deviation & add all up.

→

A: $(6-6)^2 + (7-6)^2 + (3-6)^2 + (8-6)^2 = 14$

B: $(5-5)^2 + (5-5)^2 + (3-5)^2 + (7-5)^2 = 8$

C: $(5-4)^2 + (4-4)^2 + (3-4)^2 + (4-4)^2 = 2$

SSW = 14+8+2 = 24

Step 4: SS Total $SST = \Sigma(X_{ij} - \bar{X})^2$ → diff b/w each data point & grand mean ko square & add.

OR $SST = SSB + SSW$

SST = SSW + SSB = 24 + 8 = 32

Step 5: ANOVA TABLE. m = 3 n = 12

source of variation	SS	deg. of freedom	MS	Fratio = MSB/MSW	Fcrit
between sample	8	3-1 = 2	8/2 = 4 (SSB / df)	4/2.67	= F(2, 9)
within sample	24	12-3 = 9	24/9 = 2.67 (SSW / df within)	= 1.5	= 4.26
Total	32	12-1 = 11			

- SSB degree of freedom = m-1
- SSW degree of freedom = n-m
- SST degree of freedom = n-1
- Fcritical: F(α, df between, df within)

m = no. of samples
n = no. of elements in all samples.

Step 6: Conclusion.

Fratio > Fcrit : Reject Ho p < α, reject Ho.

Here, we accept Ho.
i.e. production of wheat variety A, B, C is same
i.e. the variety does not affect production significantly

TWO WAY ANOVA: used to estimate how the mean of a quantitive variable changes acc to the levels of 2 categorical variables. How 2 independent variables in combination affect a dependent variable.

2 way anova with replication Eg:-

	genotype AA	Aa	aa	yes interaction
Temp 10°C				
Temp 20°c				
Temp 30°c				

without replication:

	AA	Aa	aa	no interaction
Temp 10°	—	—	—	
Temp 20°	—	—	—	
Temp 30°	—	—	—	

— x —.

gender	score	age grp.
B		
B		
G	for factor 1:	
G	Ho:- no sig. diff	
B	b/w means of	
G	row factor	
⋮	Ha: sig diff b/w	
⋮	means of row fact.	
	for factor 2:	
	Column factor	

factors → age
→ gender.

Ho:- → genders will have no significant effect on score
→ age will have no significant effect on score
→ genders & age interaction will have no effect on score.

M T W T F S S

Page No.: YOUVA

Date:

* factor is an Independent variable whose values are controlled and varied by the experimenter.

* Tukey test allows us to interpret the statistical significance of our anova test & find out which specific groups mean compared w. each other are different.
it is used to check ~~if there is~~ and find out where the statistical significance is occuring.
used for pairwise comparison.

* Manova: used to examine a dataset with multiple dependent variables at a time.
we can find the effect of one or more Independent variables on two or more dependent variables.

one way : 1 indep var.
2 way : 2 indep var