Differences between `ANOVA`, `MANOVA`, `ANCOVA` & `MANCOVA`

The objective of this notebook are:

- Build understanding of MANOVA, ANCOVA and MANCOVA using examples
- How we can trace back from multi-variate analysis to uni-variate analysis to post-hoc comparisons.

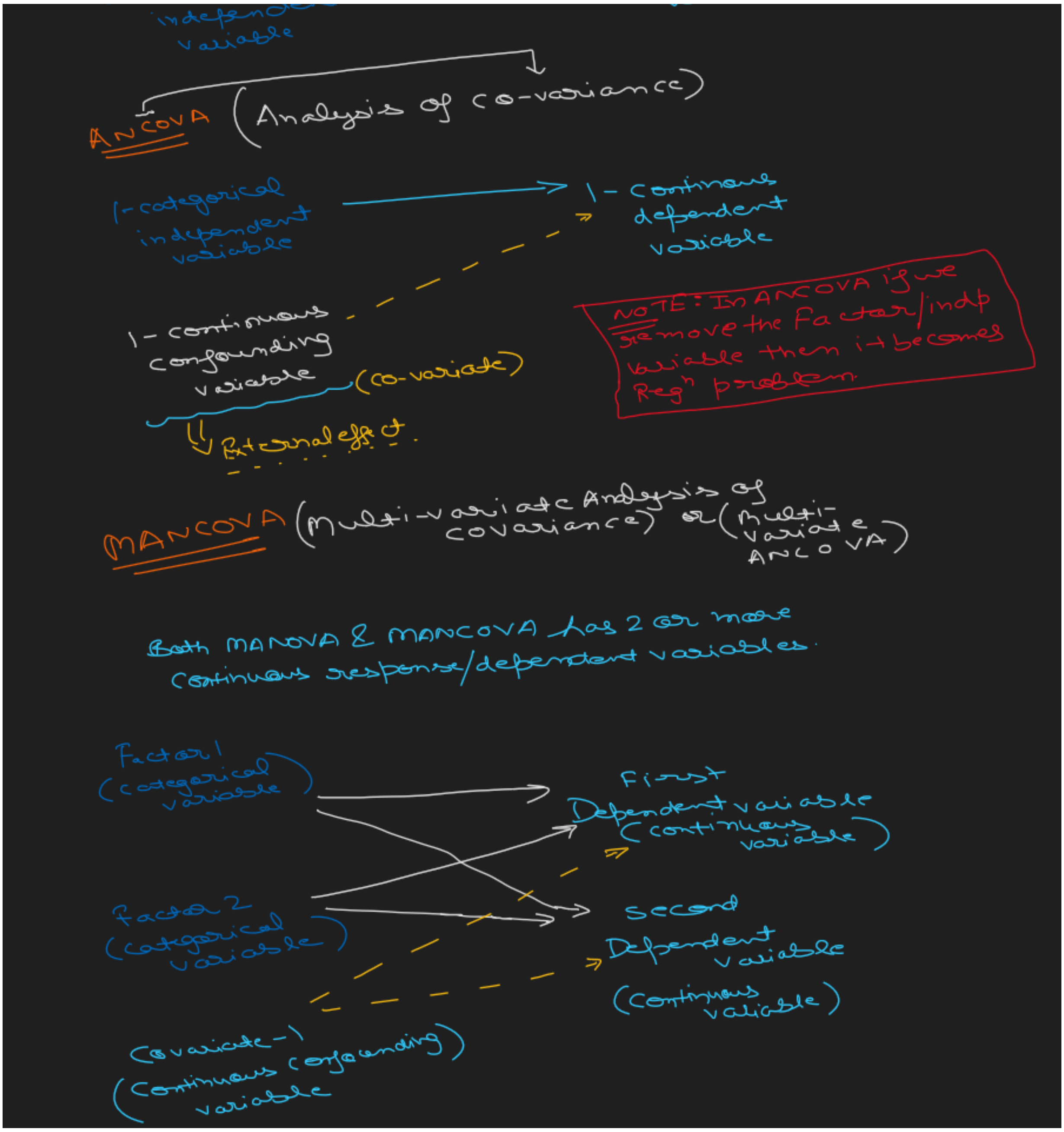
Notebook Contents

Notes_Cheatsheets

In [2]:
```
from IPython.display import Image

Image("Handwritten_Notes/Stats_Revision-4.png",width=1000,height=1000)
```

Out[2]:



Import_Packages

In [8]:

```
## Data wrangling and visualization libraries
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

## Sklearn Datasets
import sklearn.datasets as skd

## External statistics package
import pingouin

## Post-hoc tests package
import scikit_posthocs as post_hocs

## Scientific python package and statistics package
import scipy.stats as sci_st
import statsmodels.api as sm_api

## OLS and MANOVA
from statsmodels.formula.api import ols
```

```
from statsmodels.multivariate.manova import MANOVA

## Pair-wise tukey test
from statsmodels.stats.multicomp import pairwise_tukeyhsd

%matplotlib inline
```

In [9]:

```
dir(pingouin)
```

Out[9]:

```
['__builtins__',
 '__cached__',
 '__doc__',
 '__file__',
 '__loader__',
 '__name__',
 '__package__',
 '__path__',
 '__spec__',
 '__version__',
 '_check_dataframe',
 '_check_eftype',
 '_flatten_list',
 '_is_mpmath_installed',
 '_is_sklearn_installed',
 '_is_statsmodels_installed',
 '_perm_pval',
 '_postprocess_dataframe',
 'ancova',
 'anderson',
 'anova',
 'bayesfactor_binom',
 'bayesfactor_pearson',
 'bayesfactor_ttest',
 'bayesian',
 'chi2_independence',
 'chi2_mcnemar',
 'circ_axial',
 'circ_corrcc',
 'circ_corrcl',
 'circ_mean',
 'circ_r',
 'circ_rayleigh',
 'circ_vtest',
 'circular',
 'cochran',
 'compute_bootci',
 'compute_effsize',
 'compute_effsize_from_t',
 'compute_esci',
 'config',
 'contingency',
 'convert_angles',
 'convert_effsize',
 'corr',
 'correlation',
 'cronbach_alpha',
 'datasets',
 'dichotomous_crosstab',
 'distance_corr',
 'distribution',
 'effsize',
 'epsilon',
 'equivalence',
 'friedman',
 'gzscore',
 'harrelldavis',
 'homoscedasticity',
 'intraclass_corr',
```

```
 'kruskal',
 'linear_regression',
 'list_dataset',
 'logistic_regression',
 'mad',
 'madmedianrule',
 'mediation_analysis',
 'mixed_anova',
 'multicomp',
 'multivariate',
 'multivariate_normality',
 'multivariate_ttest',
 'mwu',
 'nonparametric',
 'normality',
 'options',
 'pairwise',
 'pairwise_corr',
 'pairwise_gameshowell',
 'pairwise_ttests',
 'pairwise_tukey',
 'parametric',
 'partial_corr',
 'pcorr',
 'plot_blandaltman',
 'plot_circmean',
 'plot_paired',
 'plot_rm_corr',
 'plot_shift',
 'plotting',
 'power',
 'power_anova',
 'power_chi2',
 'power_corr',
 'power_rm_anova',
 'power_ttest',
 'power_ttest2n',
 'print_table',
 'qqplot',
 'rcorr',
 'read_dataset',
 'regression',
 'reliability',
 'remove_na',
 'remove_rm_na',
 'rm_anova',
 'rm_corr',
 'set_default_options',
 'sphericity',
 'tost',
 'ttest',
 'utils',
 'warn_if_outdated',
 'welch_anova',
 'wilcoxon']
```

MANOVA

Multi-variate Analysis of Variance or **Multi-variate ANOVA**

- It is an extension of ANOVA and here 'M' stands for Multivariate.
- Just like ANOVA it can be 1-Way or 2-Way.
- MANOVA has 2 or more continuous dependent or response variables.

In [10]:
```
boston_data = skd.load_boston()
```

In [11]:
```
print(boston_data.DESCR)
```

```
.. _boston_dataset:

Boston house prices dataset
---------------------------

**Data Set Characteristics:**

    :Number of Instances: 506

    :Number of Attributes: 13 numeric/categorical predictive. Median Value (attribut
e 14) is usually the target.

    :Attribute Information (in order):
        - CRIM     per capita crime rate by town
        - ZN       proportion of residential land zoned for lots over 25,000 sq.ft.
        - INDUS    proportion of non-retail business acres per town
        - CHAS     Charles River dummy variable (= 1 if tract bounds river; 0 otherw
ise)
        - NOX      nitric oxides concentration (parts per 10 million)
        - RM       average number of rooms per dwelling
        - AGE      proportion of owner-occupied units built prior to 1940
        - DIS      weighted distances to five Boston employment centres
        - RAD      index of accessibility to radial highways
        - TAX      full-value property-tax rate per $10,000
        - PTRATIO  pupil-teacher ratio by town
        - B        1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
        - LSTAT    % lower status of the population
        - MEDV     Median value of owner-occupied homes in $1000's

    :Missing Attribute Values: None

    :Creator: Harrison, D. and Rubinfeld, D.L.

This is a copy of UCI ML housing dataset.
https://archive.ics.uci.edu/ml/machine-learning-databases/housing/


This dataset was taken from the StatLib library which is maintained at Carnegie Mell
on University.

The Boston house-price data of Harrison, D. and Rubinfeld, D.L. 'Hedonic
prices and the demand for clean air', J. Environ. Economics & Management,
vol.5, 81-102, 1978.   Used in Belsley, Kuh & Welsch, 'Regression diagnostics
...', Wiley, 1980.   N.B. Various transformations are used in the table on
pages 244-261 of the latter.

The Boston house-price data has been used in many machine learning papers that addre
ss regression
problems.

.. topic:: References

   - Belsley, Kuh & Welsch, 'Regression diagnostics: Identifying Influential Data an
d Sources of Collinearity', Wiley, 1980. 244-261.
   - Quinlan,R. (1993). Combining Instance-Based and Model-Based Learning. In Procee
dings on the Tenth International Conference of Machine Learning, 236-243, University
of Massachusetts, Amherst. Morgan Kaufmann.
```

In [12]:
```
boston_df = pd.concat([pd.DataFrame(boston_data.data,columns=boston_data.feature_nam
                       pd.DataFrame(boston_data.target,columns=['Label'])],
                      axis=1)

boston_df = boston_df[boston_df['RAD'] <= 4.0]
```

In [13]:
```
boston_df.shape
```

Out[13]: (192, 14)

In [15]:
```
boston_df.head()
```

Out[15]:

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	L
0	0.00632	18.0	2.31	0.0	0.538	6.575	65.2	4.0900	1.0	296.0	15.3	396.90	4.98	
1	0.02731	0.0	7.07	0.0	0.469	6.421	78.9	4.9671	2.0	242.0	17.8	396.90	9.14	
2	0.02729	0.0	7.07	0.0	0.469	7.185	61.1	4.9671	2.0	242.0	17.8	392.83	4.03	
3	0.03237	0.0	2.18	0.0	0.458	6.998	45.8	6.0622	3.0	222.0	18.7	394.63	2.94	
4	0.06905	0.0	2.18	0.0	0.458	7.147	54.2	6.0622	3.0	222.0	18.7	396.90	5.33	

1-Way:MANOVA

In [14]:
```
## Label --> House price in $1000's (Dependent/response continuous variable)
## CRIM --> Per capita crime rate by town (Dependent/response continuous variable)
## RAD --> Accessibility to radial highways (Independent categorical variable)

manova_1_way_formula = ('Label + CRIM ~ RAD')
```

In [16]:
```
manova_1_way = MANOVA.from_formula(manova_1_way_formula,data=boston_df)
```

In [17]:
```
print(manova_1_way.mv_test())
```

```
                  Multivariate linear model
==============================================================

--------------------------------------------------------------
       Intercept        Value  Num DF  Den DF  F Value  Pr > F
--------------------------------------------------------------
          Wilks' lambda 0.4118 2.0000 189.0000 134.9634 0.0000
         Pillai's trace 0.5882 2.0000 189.0000 134.9634 0.0000
 Hotelling-Lawley trace 1.4282 2.0000 189.0000 134.9634 0.0000
    Roy's greatest root 1.4282 2.0000 189.0000 134.9634 0.0000
--------------------------------------------------------------

--------------------------------------------------------------
           RAD          Value  Num DF  Den DF  F Value Pr > F
--------------------------------------------------------------
           Wilks' lambda 0.8552 2.0000 189.0000 15.9953 0.0000
          Pillai's trace 0.1448 2.0000 189.0000 15.9953 0.0000
  Hotelling-Lawley trace 0.1693 2.0000 189.0000 15.9953 0.0000
     Roy's greatest root 0.1693 2.0000 189.0000 15.9953 0.0000
==============================================================
```

Here, we have performed MANOVA for 2 response variables `House Rate` and `Crime Rate` with one factor or categorical variable i.e. `Highways Accessibility`.

- In the result, it is quite evident that there is a significant difference in means of Highways Accessibility groups for the combination of House and Crime Rates.
- I'll use `Wilk's Lambda` as a metric and it assumes that the homogenity of variances exist in the dataset (I'm assuming this assumption holds TRUE here).

- **Next, will perform the `univariate analysis to identify how much difference in means exists in both the dependent variables with respect to Highways Accessibility.`**

`Uni-variate_Analysis:1`

Dependent Variable 1 : House Rate

NOTE

- We have two dependent variables(House Rate and CRIME Rate) so the L.O.S needs to be re-calculated as 0.025(i.e. 0.05/2). If we don't perform this re-calculation of alpha then we will end up with more number of Type-1 Errors.

In [18]:
```
reg_res_val1 = ols('Label ~ RAD',data=boston_df).fit()
```

In [19]:
```
sm_api.stats.anova_lm(reg_res_val1,typ=1)
```

Out[19]:

	df	sum_sq	mean_sq	F	PR(>F)
RAD	1.0	620.716217	620.716217	10.361464	0.001513
Residual	190.0	11382.182950	59.906226	NaN	NaN

So, the univariate test for dependent variable 1 is also significant. Therefore, we will go ahead and perform the pair wise comparison test to identify which groups of Highway Accessibility have significant differences for House Rate.

In [20]:
```
reg_res_val1_post_hocs = pairwise_tukeyhsd(endog=boston_df['Label'],groups=boston_df
```

In [21]:
```
reg_res_val1_post_hocs.summary()
```

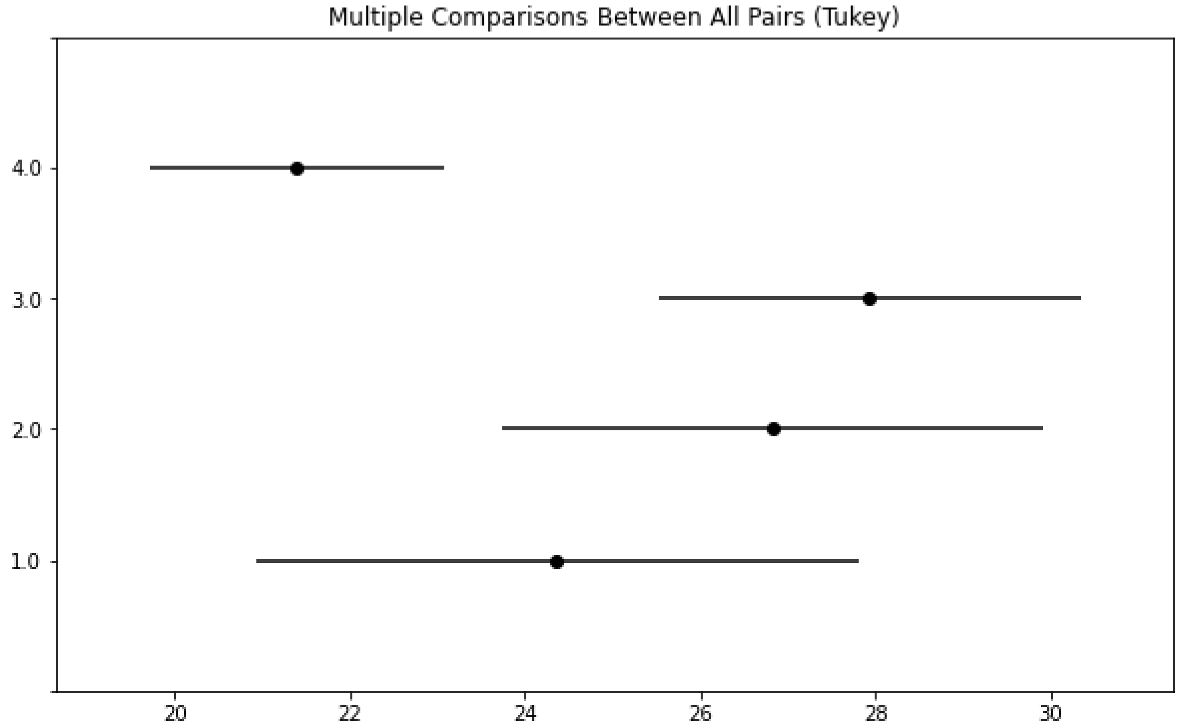
Out[21]:

Multiple Comparison of Means - Tukey HSD, FWER=0.03

group1	group2	meandiff	p-adj	lower	upper	reject
1.0	2.0	2.4683	0.6734	-3.973	8.9097	False
1.0	3.0	3.5639	0.3127	-2.3133	9.4412	False
1.0	4.0	-2.9777	0.3595	-8.1494	2.1939	False
2.0	3.0	1.0956	0.9	-4.4515	6.6427	False
2.0	4.0	-5.4461	0.0078	-10.2392	-0.6529	True
3.0	4.0	-6.5417	0.001	-10.5449	-2.5384	True

In [22]:
```
reg_res_val1_post_hocs.plot_simultaneous();
```

```
c:\users\rajsh\appdata\local\programs\python\python36\lib\site-packages\statsmodels
\sandbox\stats\multicomp.py:775: UserWarning: FixedFormatter should only be used tog
ether with FixedLocator
  ax1.set_yticklabels(np.insert(self.groupsunique.astype(str), 0, ''))
```



Here, we found out that groups (2 & 4) and (3 & 4) have significant difference in house rates.

Uni-variate_Analysis:2

Dependent Variable 2 : Crime Rate

In [23]:
```
reg_res_val2 = ols('CRIM ~ RAD',data=boston_df).fit()
```

In [24]:
```
sm_api.stats.anova_lm(reg_res_val2,typ=1)
```

Out[24]:

	df	sum_sq	mean_sq	F	PR(>F)
RAD	1.0	3.854328	3.854328	30.875438	9.202109e-08
Residual	190.0	23.718603	0.124835	NaN	NaN

So, the univariate test for dependent variable 2 is also significant. Therefore, we will go ahead and perform the pair wise comparison test to identify which groups of Highway Accessibility have significant differences for Crime rate.

In [25]:
```
reg_res_val2_post_hocs = pairwise_tukeyhsd(endog=boston_df['CRIM'],groups=boston_df[
```

In [26]:
```
reg_res_val2_post_hocs.summary()
```

Out[26]:

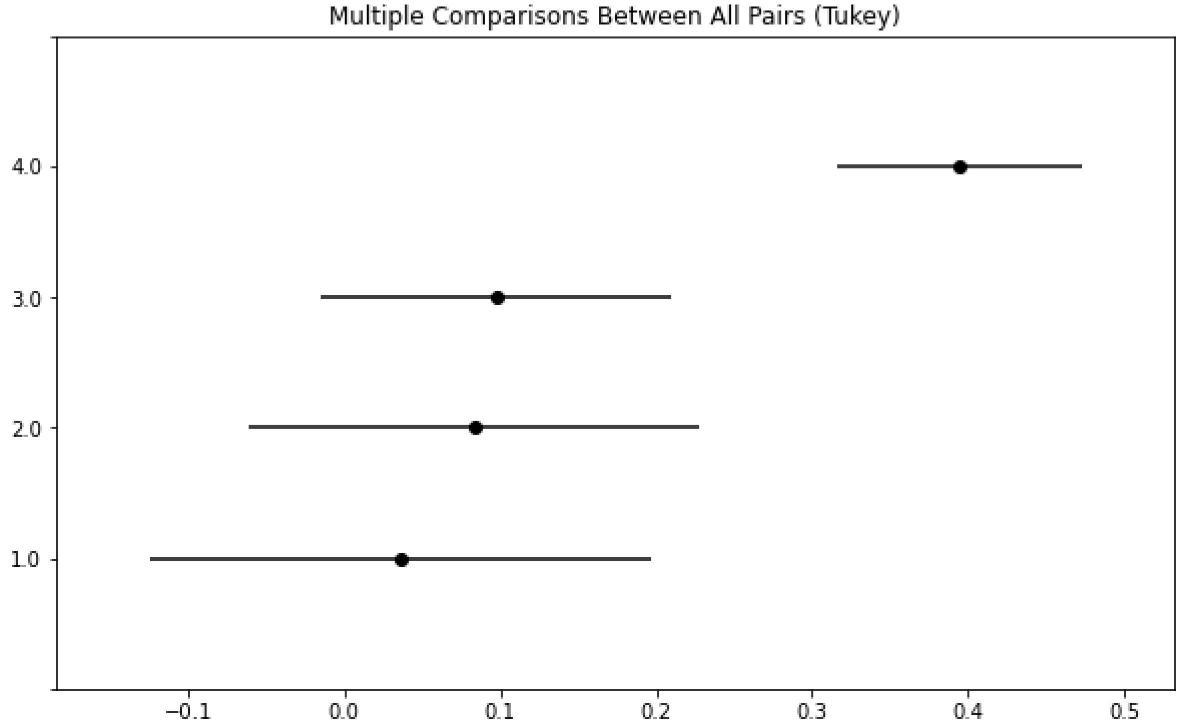
Multiple Comparison of Means - Tukey HSD, FWER=0.03

group1	group2	meandiff	p-adj	lower	upper	reject
1.0	2.0	0.0473	0.9	-0.2533	0.3478	False
1.0	3.0	0.0613	0.9	-0.2129	0.3356	False
1.0	4.0	0.3579	0.001	0.1165	0.5992	True
2.0	3.0	0.0141	0.9	-0.2448	0.2729	False

2.0	4.0	0.3106	0.001	0.087	0.5343	True
3.0	4.0	0.2965	0.001	0.1097	0.4833	True

In [27]:
```
reg_res_val2_post_hocs.plot_simultaneous();
```

```
c:\users\rajsh\appdata\local\programs\python\python36\lib\site-packages\statsmodels
\sandbox\stats\multicomp.py:775: UserWarning: FixedFormatter should only be used tog
ether with FixedLocator
  ax1.set_yticklabels(np.insert(self.groupsunique.astype(str), 0, ''))
```



Here, we found out that groups (1 & 4), (2 & 4) and (3 & 4) have significant differences in crime rates.

2-Way:MANOVA

In [28]:
```
boston_df.head()
```

Out[28]:

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	L
0	0.00632	18.0	2.31	0.0	0.538	6.575	65.2	4.0900	1.0	296.0	15.3	396.90	4.98	
1	0.02731	0.0	7.07	0.0	0.469	6.421	78.9	4.9671	2.0	242.0	17.8	396.90	9.14	
2	0.02729	0.0	7.07	0.0	0.469	7.185	61.1	4.9671	2.0	242.0	17.8	392.83	4.03	
3	0.03237	0.0	2.18	0.0	0.458	6.998	45.8	6.0622	3.0	222.0	18.7	394.63	2.94	
4	0.06905	0.0	2.18	0.0	0.458	7.147	54.2	6.0622	3.0	222.0	18.7	396.90	5.33	

In [29]:
```
## Label --> House price in $1000's (Dependent/response continuous variable)
## CRIM --> Per capita crime rate by town (Dependent/response continuous variable)
## RAD --> Accessibility to radial highways (Independent categorical variable)
## CHAS --> Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)

manova_2_way_formula = ('Label + CRIM ~ RAD + CHAS')
```

In [30]:
```
manova_2_way = MANOVA.from_formula(manova_2_way_formula,data=boston_df)
```

In [31]:
```
print(manova_2_way.mv_test())
```

```
                 Multivariate linear model
==============================================================

--------------------------------------------------------------
       Intercept        Value  Num DF  Den DF   F Value  Pr > F
--------------------------------------------------------------
          Wilks' lambda 0.4062 2.0000 188.0000 137.4338 0.0000
         Pillai's trace 0.5938 2.0000 188.0000 137.4338 0.0000
 Hotelling-Lawley trace 1.4621 2.0000 188.0000 137.4338 0.0000
    Roy's greatest root 1.4621 2.0000 188.0000 137.4338 0.0000
--------------------------------------------------------------

--------------------------------------------------------------
           RAD           Value  Num DF  Den DF  F Value Pr > F
--------------------------------------------------------------
           Wilks' lambda 0.8487 2.0000 188.0000 16.7570 0.0000
          Pillai's trace 0.1513 2.0000 188.0000 16.7570 0.0000
  Hotelling-Lawley trace 0.1783 2.0000 188.0000 16.7570 0.0000
     Roy's greatest root 0.1783 2.0000 188.0000 16.7570 0.0000
--------------------------------------------------------------

--------------------------------------------------------------
           CHAS          Value  Num DF  Den DF  F Value Pr > F
--------------------------------------------------------------
           Wilks' lambda 0.9707 2.0000 188.0000  2.8346 0.0613
          Pillai's trace 0.0293 2.0000 188.0000  2.8346 0.0613
  Hotelling-Lawley trace 0.0302 2.0000 188.0000  2.8346 0.0613
     Roy's greatest root 0.0302 2.0000 188.0000  2.8346 0.0613
==============================================================
```

Here, I'll use the Wilk's lambda, thus, assumes that homogenity in the variables holds TRUE.

- **Some more background on Wilk's Lambda Value i.e. RAD --> 0.8487 and CHAS -->0.9707.**
  - **Here, the point to understand is that Wilk's Lambda are the coefficients between 0 and 1. The higher this value the lower is the contribution of this factor. Thus, the contribution of CHAS is less as compared to RAD in the deviation of dependent variables.**

- The results of second factor (Charles River Bound) are not significant. Therefore, for this variable we will stop here because it means that there is no significant difference in the dependent variables w.r.t to Charles River Bound.

- For the first factor (RAD : Highways accessibility) the results are significant. Therefore, we can go ahead with the univariate analysis and post-hoc tests(just like performed in 1-way MANOVA).

e.g. `(Label + CRIM ~ RAD)`

If the results are found significant then we perform the 2 uni-variate analysis

`(Label ~ RAD)`

`(CRIM ~ RAD)`

Additional_INFO

- **If in case both of the factor results were significant then we would have straight away performed the univariate analysis.**

e.g. `(Label ~ RAD + CHAS)` and `(CRIM ~ RAD + CHAS)`

If the results are found to be significant then perform the pair-wise comparisons

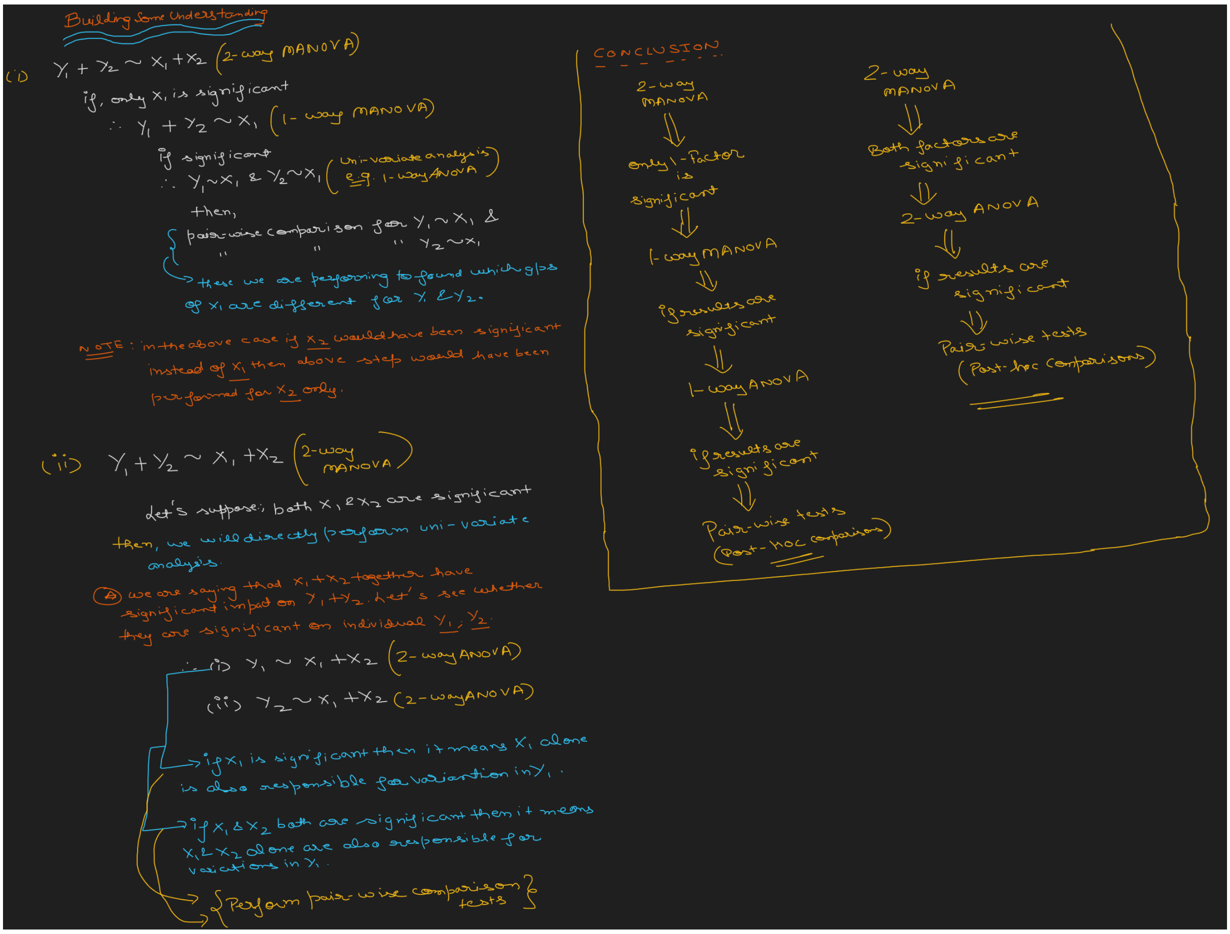
---

In [52]:

```python
from IPython.display import Image

Image("Handwritten_Notes/Stats_Revision-5.png",width=1000,height=1000)
```

Out[52]:



ANCOVA

It stands for Analysis of Covariance. It is an extension of ANOVA that is referred as ANOVA + Regression.

- It is used to determine whether or not there is a statistically significant difference between the means of three or more independent groups, after controlling for one or more covariates or confounding variable(an external variable that influences the response variable).

Example of ANCOVA

A teacher wants to know if three different studying techniques have an impact on exam scores, but she wants to account for the current grade that the student already has in the class.

- She will perform an ANCOVA using the following variables:
  - Factor variable: studying techniques
  - Covariate: current grade
  - Response variable: exam score

DATASET-1

In [32]:
```
df = pd.DataFrame({'technique': np.repeat(['A', 'B', 'C'], 5),
                   'current_grade': [67, 88, 75, 77, 85,
                                     92, 69, 77, 74, 88,
                                     96, 91, 88, 82, 80],
                   'exam_score': [77, 89, 72, 74, 69,
                                  78, 88, 93, 94, 90,
                                  85, 81, 83, 88, 79]})
df
```

Out[32]:

	technique	current_grade	exam_score
0	A	67	77
1	A	88	89
2	A	75	72
3	A	77	74
4	A	85	69
5	B	92	78
6	B	69	88
7	B	77	93
8	B	74	94
9	B	88	90
10	C	96	85
11	C	91	81
12	C	88	83
13	C	82	88
14	C	80	79

In [33]:
```
from pingouin import ancova
```

In [232...
```
ancova(data=df, dv='exam_score', covar='current_grade', between='technique')
```

Out[232...

	Source	SS	DF	F	p-unc	np2
0	technique	390.575130	2	4.809973	0.031556	0.466536
1	current_grade	4.193886	1	0.103296	0.753934	0.009303
2	Residual	446.606114	11	NaN	NaN	NaN

From above ANCOVA table, we can understand that the p-value (p-unc = "uncorrected p-value") for `study technique` is 0.03155.

- As, p-value < 0.05, therefore, we are rejecting the null hypothesis and concluding that the study techniques leads to different exam scores, even after accounting for the student's current grade in the class.

DATASET-2

In [36]:
```
boston_df.head()
```

Out[36]:

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	L
0	0.00632	18.0	2.31	0.0	0.538	6.575	65.2	4.0900	1.0	296.0	15.3	396.90	4.98	
1	0.02731	0.0	7.07	0.0	0.469	6.421	78.9	4.9671	2.0	242.0	17.8	396.90	9.14	
2	0.02729	0.0	7.07	0.0	0.469	7.185	61.1	4.9671	2.0	242.0	17.8	392.83	4.03	
3	0.03237	0.0	2.18	0.0	0.458	6.998	45.8	6.0622	3.0	222.0	18.7	394.63	2.94	
4	0.06905	0.0	2.18	0.0	0.458	7.147	54.2	6.0622	3.0	222.0	18.7	396.90	5.33	

In [37]:
```
pingouin.ancova(data=boston_df,dv='Label',between='RAD',covar='AGE')
```

Out[37]:

	Source	SS	DF	F	p-unc	np2
0	RAD	1199.526554	3	8.955124	1.423559e-05	0.125618
1	AGE	2141.154605	1	47.954681	6.794349e-11	0.204102
2	Residual	8349.464568	187	NaN	NaN	NaN

From the ANCOVA table, we can understand that there is the significant effect of Highways accessibility on the difference in house prices.

- In addition to this, AGE of the property also playing a significant role in the deviation among House rates.

`MANCOVA`

It stands for Multi-variate Analysis of Covariance. It is an extension of MANOVA that is referred as MANOVA + Regression.

- It is used to determine whether or not there is a statistically significant difference between the means of three or more independent groups, after controlling for one or more covariates or confounding variable(an external variable that influences the response variable).

Two Factor or Categorical variables with one or more response variables with a Covariate.

In [53]:
```
mancova_2_way_formula = ('Label + CRIM ~ RAD + CHAS + AGE')
```

In [54]:
```
print(MANOVA.from_formula(formula=mancova_2_way_formula, data=boston_df).mv_test())
```

```
                  Multivariate linear model
==============================================================

--------------------------------------------------------------
       Intercept        Value  Num DF  Den DF   F Value  Pr > F
--------------------------------------------------------------
          Wilks' lambda 0.3406 2.0000 187.0000 181.0278 0.0000
         Pillai's trace 0.6594 2.0000 187.0000 181.0278 0.0000
 Hotelling-Lawley trace 1.9361 2.0000 187.0000 181.0278 0.0000
    Roy's greatest root 1.9361 2.0000 187.0000 181.0278 0.0000
--------------------------------------------------------------

--------------------------------------------------------------
           RAD           Value  Num DF  Den DF   F Value Pr > F
--------------------------------------------------------------
           Wilks' lambda 0.8550 2.0000 187.0000 15.8507 0.0000
          Pillai's trace 0.1450 2.0000 187.0000 15.8507 0.0000
  Hotelling-Lawley trace 0.1695 2.0000 187.0000 15.8507 0.0000
     Roy's greatest root 0.1695 2.0000 187.0000 15.8507 0.0000
--------------------------------------------------------------

--------------------------------------------------------------
           CHAS          Value  Num DF  Den DF   F Value Pr > F
--------------------------------------------------------------
           Wilks' lambda 0.9620 2.0000 187.0000  3.6916 0.0268
          Pillai's trace 0.0380 2.0000 187.0000  3.6916 0.0268
  Hotelling-Lawley trace 0.0395 2.0000 187.0000  3.6916 0.0268
     Roy's greatest root 0.0395 2.0000 187.0000  3.6916 0.0268
--------------------------------------------------------------

--------------------------------------------------------------
           AGE           Value  Num DF  Den DF   F Value Pr > F
--------------------------------------------------------------
           Wilks' lambda 0.7289 2.0000 187.0000 34.7734 0.0000
          Pillai's trace 0.2711 2.0000 187.0000 34.7734 0.0000
  Hotelling-Lawley trace 0.3719 2.0000 187.0000 34.7734 0.0000
     Roy's greatest root 0.3719 2.0000 187.0000 34.7734 0.0000
==============================================================
```

All the results are significant, therefore, we directly perform the univariate analysis.

Univariate Analysis of Dependent Variable-1

In [55]:
```
mancova_2_way_val1 = ols('Label ~ RAD + CHAS + AGE',data=boston_df).fit()
```

In [56]:
```
sm_api.stats.anova_lm(mancova_2_way_val1)
```

Out[56]:

	df	sum_sq	mean_sq	F	PR(>F)
RAD	1.0	620.716217	620.716217	13.163748	3.676702e-04
CHAS	1.0	324.828139	324.828139	6.888745	9.387280e-03
AGE	1.0	2192.504690	2192.504690	46.497219	1.215439e-10
Residual	188.0	8864.850121	47.153458	NaN	NaN

All factors significant results therefore, will perform the post-hoc comparisons for both the factors.

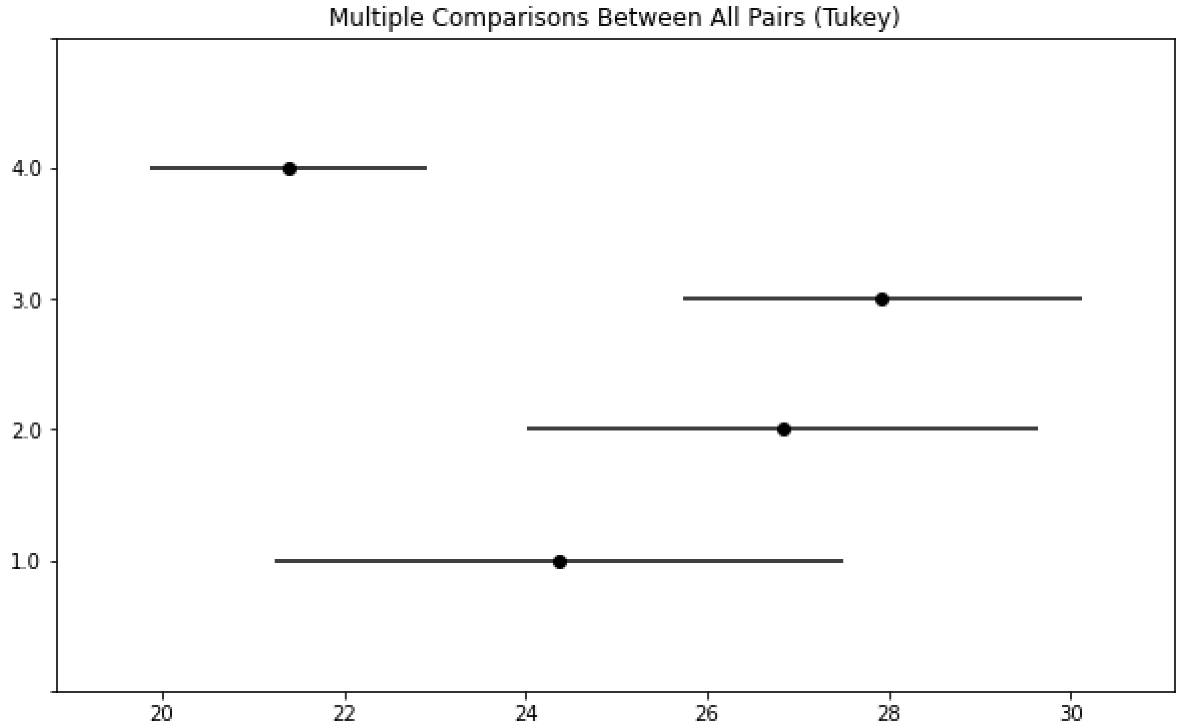
In [57]:
```
pairwise_tukeyhsd(endog=boston_df['Label'],groups=boston_df['RAD']).summary()
```

Out[57]: Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
1.0	2.0	2.4683	0.6734	-3.3946	8.3312	False
1.0	3.0	3.5639	0.3127	-1.7856	8.9135	False
1.0	4.0	-2.9777	0.3595	-7.685	1.7295	False
2.0	3.0	1.0956	0.9	-3.9534	6.1446	False
2.0	4.0	-5.4461	0.0078	-9.8088	-1.0833	True
3.0	4.0	-6.5417	0.001	-10.1854	-2.8979	True

In [58]:
```
pairwise_tukeyhsd(endog=boston_df['Label'],groups=boston_df['RAD']).plot_simultaneou
```

```
c:\users\rajsh\appdata\local\programs\python\python36\lib\site-packages\statsmodels
\sandbox\stats\multicomp.py:775: UserWarning: FixedFormatter should only be used tog
ether with FixedLocator
  ax1.set_yticklabels(np.insert(self.groupsunique.astype(str), 0, ''))
```



In [59]:
```
pairwise_tukeyhsd(endog=boston_df['Label'],groups=boston_df['CHAS']).summary()
```

Out[59]: Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
0.0	1.0	5.0144	0.0413	0.1989	9.8298	True

Univariate Analysis of Dependent Variable-2

In [60]:
```
mancova_2_way_val2 = ols('CRIM ~ RAD + CHAS + AGE',data=boston_df).fit()
```

In [61]:
```
sm_api.stats.anova_lm(mancova_2_way_val2)
```

Out[61]:

	df	sum_sq	mean_sq	F	PR(>F)
RAD	1.0	3.854328	3.854328	37.400353	5.452660e-09
CHAS	1.0	0.196842	0.196842	1.910047	1.685979e-01

	df	sum_sq	mean_sq	F	PR(>F)
AGE	1.0	4.147249	4.147249	40.242711	1.631436e-09
Residual	188.0	19.374512	0.103056	NaN	NaN

In [62]:
```
pairwise_tukeyhsd(endog=boston_df['CRIM'],groups=boston_df['RAD']).summary()
```

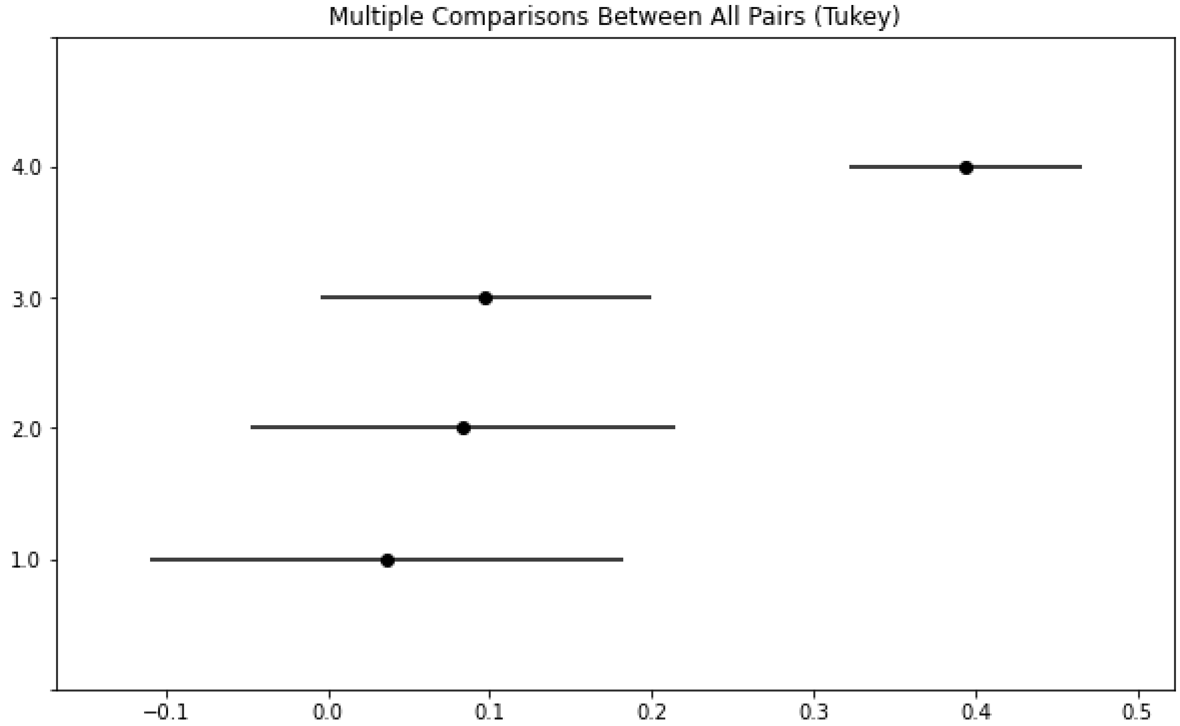
Out[62]:

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
1.0	2.0	0.0473	0.9	-0.2263	0.3208	False
1.0	3.0	0.0613	0.9	-0.1883	0.311	False
1.0	4.0	0.3579	0.001	0.1382	0.5775	True
2.0	3.0	0.0141	0.9	-0.2215	0.2497	False
2.0	4.0	0.3106	0.001	0.107	0.5142	True
3.0	4.0	0.2965	0.001	0.1265	0.4666	True

In [63]:
```
pairwise_tukeyhsd(endog=boston_df['CRIM'],groups=boston_df['RAD']).plot_simultaneous
```

```
c:\users\rajsh\appdata\local\programs\python\python36\lib\site-packages\statsmodels
\sandbox\stats\multicomp.py:775: UserWarning: FixedFormatter should only be used tog
ether with FixedLocator
  ax1.set_yticklabels(np.insert(self.groupsunique.astype(str), 0, ''))
```



In [64]:
```
pairwise_tukeyhsd(endog=boston_df['CRIM'],groups=boston_df['CHAS']).summary()
```

Out[64]:

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
0.0	1.0	-0.0927	0.4335	-0.3257	0.1403	False

Both the factors are not significant, same gets displayed in the post-hoc tests.