# Statistical Analysis: Wine Dataset

We will calculate the following statistics for each attribute

* [] Central Tendancy
  + Mean
  + Median
  + Mode
* [] Dispersion
  + Range
  + Quartiles
  + InterQuartile Range
  + Variance
  + Santard Deviation
* [] Covariance Matrix
* [] Correlation Matrix

## Importing necessary Libraries

import pandas as pd

## Load datasets

iris\_df = pd.read\_csv('../iris/iris.data', header=None)  
iris\_df.columns = [  
 "sepal\_length", "sepal\_width", "petal\_length", "petal\_width", "class"   
]  
iris\_df

sepal\_length sepal\_width petal\_length petal\_width class  
0 5.1 3.5 1.4 0.2 Iris-setosa  
1 4.9 3.0 1.4 0.2 Iris-setosa  
2 4.7 3.2 1.3 0.2 Iris-setosa  
3 4.6 3.1 1.5 0.2 Iris-setosa  
4 5.0 3.6 1.4 0.2 Iris-setosa  
.. ... ... ... ... ...  
145 6.7 3.0 5.2 2.3 Iris-virginica  
146 6.3 2.5 5.0 1.9 Iris-virginica  
147 6.5 3.0 5.2 2.0 Iris-virginica  
148 6.2 3.4 5.4 2.3 Iris-virginica  
149 5.9 3.0 5.1 1.8 Iris-virginica  
  
[150 rows x 5 columns]

## Handle 'class' attribute which is a categorial data

# Handling categorical data  
  
# Iris-setosa -> 0  
iris\_df.loc[iris\_df['class'] == 'Iris-setosa', 'class'] = 0  
# Iris-versicolor -> 1  
iris\_df.loc[iris\_df['class'] == 'Iris-versicolor', 'class'] = 1  
# Iris-virginica -> 2  
iris\_df.loc[iris\_df['class'] == 'Iris-virginica', 'class'] = 2  
  
iris\_df

sepal\_length sepal\_width petal\_length petal\_width class  
0 5.1 3.5 1.4 0.2 0  
1 4.9 3.0 1.4 0.2 0  
2 4.7 3.2 1.3 0.2 0  
3 4.6 3.1 1.5 0.2 0  
4 5.0 3.6 1.4 0.2 0  
.. ... ... ... ... ...  
145 6.7 3.0 5.2 2.3 2  
146 6.3 2.5 5.0 1.9 2  
147 6.5 3.0 5.2 2.0 2  
148 6.2 3.4 5.4 2.3 2  
149 5.9 3.0 5.1 1.8 2  
  
[150 rows x 5 columns]

## Compute central tendency (mean, median, mode)

mean\_values = iris\_df.mean()  
median\_values = iris\_df.median()  
mode\_values = iris\_df.mode().iloc[0] # mode() returns a dataframe, so we take the first row  
  
# Combine them into a single DataFrame  
central\_tendency\_table = pd.DataFrame({  
 'Mean': mean\_values,  
 'Median': median\_values,  
 'Mode': mode\_values  
})  
  
# Display the table  
central\_tendency\_table

Mean Median Mode  
sepal\_length 5.843333 5.8 5.0  
sepal\_width 3.054 3.0 3.0  
petal\_length 3.758667 4.35 1.5  
petal\_width 1.198667 1.3 0.2  
class 1.0 1.0 0

## Measures of Dispersion (range, quartiles, InterQuartile Ranges, Variance, Standard Deviations)

iris\_dispersion = iris\_df.describe().T  
iris\_dispersion['IQR'] = iris\_dispersion['75%'] - iris\_dispersion['25%']  
iris\_dispersion['Variance'] = iris\_dispersion['std'] \*\* 2  
  
iris\_dispersion

count mean std min 25% 50% 75% max IQR \  
sepal\_length 150.0 5.843333 0.828066 4.3 5.1 5.80 6.4 7.9 1.3   
sepal\_width 150.0 3.054000 0.433594 2.0 2.8 3.00 3.3 4.4 0.5   
petal\_length 150.0 3.758667 1.764420 1.0 1.6 4.35 5.1 6.9 3.5   
petal\_width 150.0 1.198667 0.763161 0.1 0.3 1.30 1.8 2.5 1.5   
  
 Variance   
sepal\_length 0.685694   
sepal\_width 0.188004   
petal\_length 3.113179   
petal\_width 0.582414

# Compute covariance and correlation matrices

iris\_covariance = iris\_df.cov()  
iris\_correlation = iris\_df.corr()

### WINE Covariance Matrix

iris\_covariance

sepal\_length sepal\_width petal\_length petal\_width class  
sepal\_length 0.685694 -0.039268 1.273682 0.516904 0.530872  
sepal\_width -0.039268 0.188004 -0.321713 -0.117981 -0.148993  
petal\_length 1.273682 -0.321713 3.113179 1.296387 1.371812  
petal\_width 0.516904 -0.117981 1.296387 0.582414 0.597987  
class 0.530872 -0.148993 1.371812 0.597987 0.671141

### WINE Correlation Matrix

iris\_correlation

sepal\_length sepal\_width petal\_length petal\_width class  
sepal\_length 1.000000 -0.109369 0.871754 0.817954 0.782561  
sepal\_width -0.109369 1.000000 -0.420516 -0.356544 -0.419446  
petal\_length 0.871754 -0.420516 1.000000 0.962757 0.949043  
petal\_width 0.817954 -0.356544 0.962757 1.000000 0.956464  
class 0.782561 -0.419446 0.949043 0.956464 1.000000