ANOVA

One Way ANOVA

- t-test: compares the means between 2 samples
- What if? there are more than 2 samples in an experiment, then an ANOVA is required
- The one-way analysis of variance (ANOVA) is used to determine whether there are any significant differences between the means of three or more independent (unrelated) groups
- One-way because there is One Independent Variable
- Extension of One Way ANOVA is two-way ANOVA that examines the influence of two different categorical independent variables on one dependent variable

https://statistics.laerd.com/statistical-guides/one-way-anova-statistical-guide.php

One Way ANOVA… contd

- **Null Hypothesis (Ho):** All means are same
- **Alternate Hypothesis (Ha):** Atleast one of the means is different from the others

- One-way ANOVA is an ***omnibus*** test statistic and cannot tell you which specific groups were significantly different from each other, only that at least two groups were.

- To determine which specific groups differed from each other, you need to use a *post hoc* test.

ANOVA and F Test

- The ANOVA produces an F-statistic, the ratio of the variance calculated among the means to the variance within the samples.

The formula for the one-way **ANOVA** *F*-test statistic is

$$F = \frac{\text{explained variance}}{\text{unexplained variance}},$$

or

$$F = \frac{\text{between-group variability}}{\text{within-group variability}}.$$

The "explained variance", or "between-group variability" is

$$\sum_i n_i(\bar{Y}_{i\cdot} - \bar{Y})^2/(K-1)$$

where $\bar{Y}_{i\cdot}$ denotes the sample mean in the i^{th} group, n_i is the number of observations in the i^{th} group, $\bar{Y}$ denotes the overall mean of the data, and K denotes the

The "unexplained variance", or "within-group variability" is

$$\sum_{ij}(Y_{ij} - \bar{Y}_{i\cdot})^2/(N-K),$$

where Y_{ij} is the j^{th} observation in the i^{th} out of K groups and N is the overall sample size.

Let us first set the working directory path

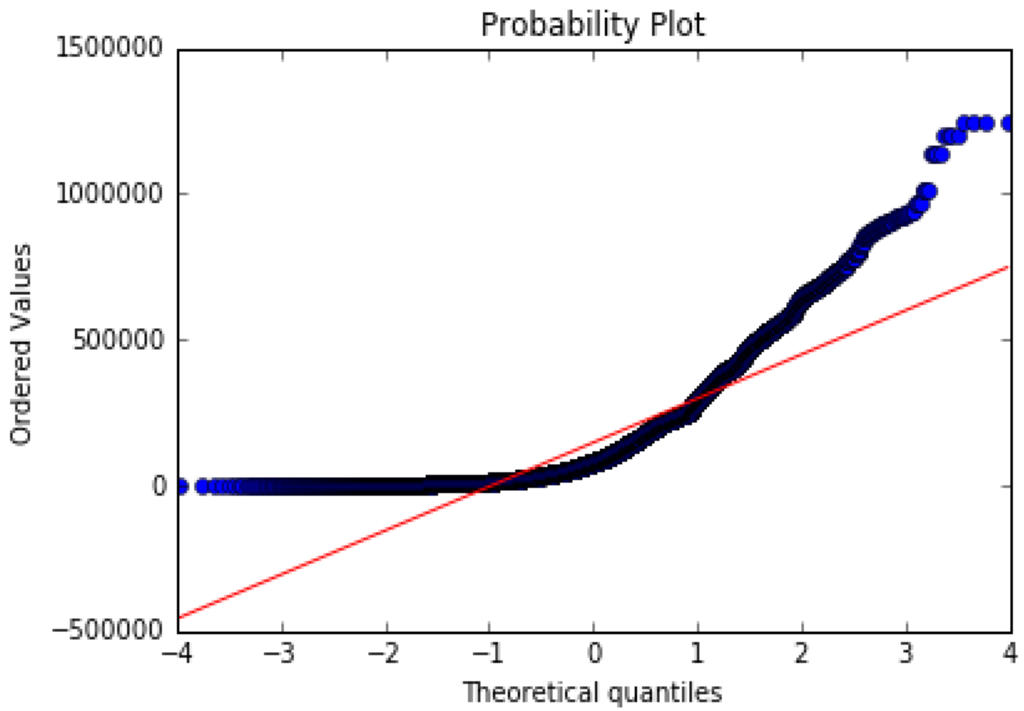
```
import os
os.chdir("D:\K2Analytics\Datafile\")
```

Import the dataset

```
dst = pd.read_csv("hypothesis_test.csv")
len(dst)
dst.dtypes
```

QQ Plot to see normality

```
st.probplot(dst.Balance, dist="norm", plot=pylab)
pylab.show()
```



Bartlett Test of Homogeneity of Variances;

Null Hypothesis : Variance is same across all groups;

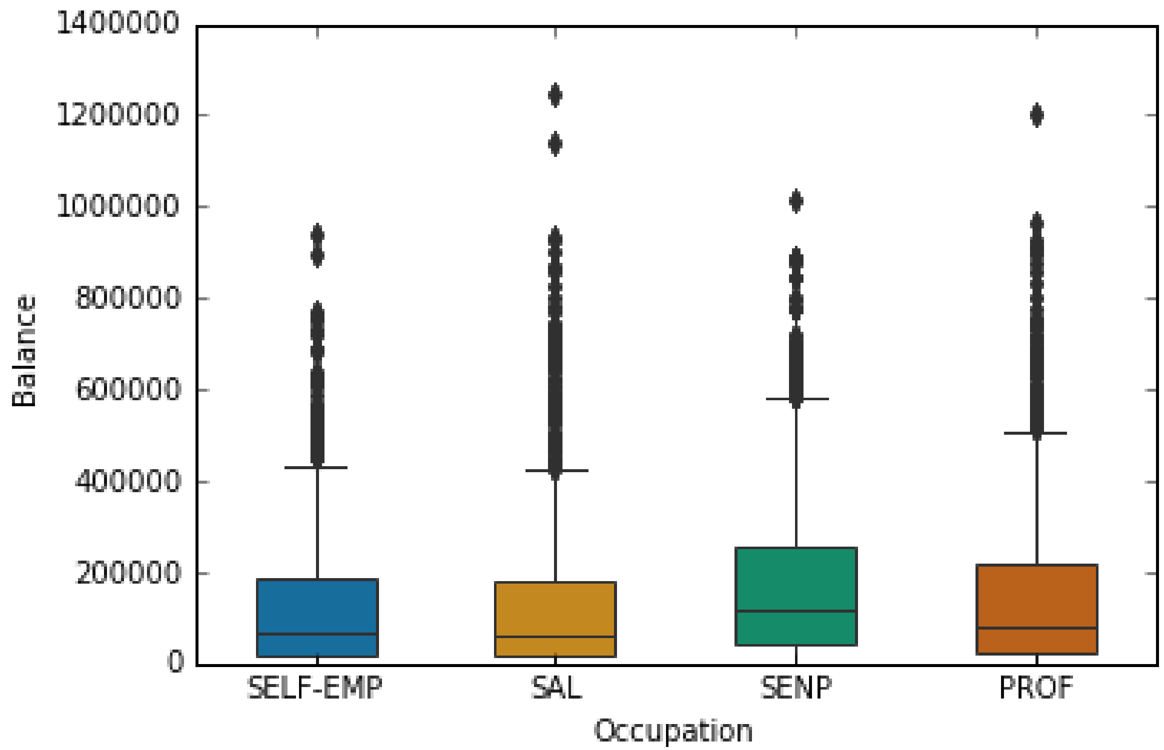
We require p-value to be greater than 0.05 to satisfy the ANOVA criteria

```
grp = dst['Occupation'].unique().tolist()

for i in range(0,len(grp)):
globals()['occ_%s' % int(i+1)] = dst['Balance'][dst['Occupation']==grp[i]].values

st.bartlett(occ_1, occ_2, occ_3, occ_4)
```

```
BartlettResult(statistic=96.4005309887855, pvalue=9.232247879183098e-21)
```

...contd

```
## create box plot to see the variable distribution
bplot = sns.boxplot(y='Balance', x='Occupation',
                    data=dst,
                    width=0.5,
                    palette="colorblind")

#### create an ANOVA table
mod = ols('Balance ~ Occupation’,
                    data=dst).fit();
aov_table = sm.stats.anova_lm(mod, typ=2)
print(aov_table)
```




```
                  sum_sq       df           F        PR(>F)
Occupation  1.051773e+13      3.0  123.819694  1.831565e-79
Residual    5.661793e+14  19996.0         NaN           NaN
```

...contd

Let us see the Tukey's Honest Significant Difference; It helps see the Factor Level which are statistically different

mod = ols('Balance ~ Occupation', data=dst).fit();

aov_table = sm.stats.anova_lm(mod, typ=2)

print(aov_table)

```
    Multiple Comparison of Means - Tukey HSD,FWER=0.05
=============================================================
 group1   group2    meandiff     lower        upper     reject
-------------------------------------------------------------
  PROF     SAL     -23151.2296 -31288.9331 -15013.5262  True
  PROF   SELF-EMP -18592.1778 -28065.2784  -9119.0772  True
  PROF     SENP    34199.9152  25877.3022  42522.5282  True
  SAL    SELF-EMP  4559.0518   -4797.0438  13915.1474 False
  SAL      SENP    57351.1448  49161.9582  65540.3314  True
SELF-EMP   SENP    52792.093   43274.7302  62309.4557  True
-------------------------------------------------------------
```

```
In [27]: dst.groupby(by = ['Occupation'])['Balance'].mean()
    ...:
Out[27]:
Occupation
PROF        146951.673258
SAL         123800.443624
SELF-EMP    128359.495443
SENP        181151.588434
```

Kruskal–Wallis test (Non-Parametric Test)

- **Kruskal–Wallis test** is used in place of ANOVA if the distribution is not normal

```
## Rank all data from all groups together; i.e.,
## rank the data from 1 to N ignoring group membership.
## Assign any tied values the average of the ranks
## they would have received had they not been tied.
occupation = {}
for grp in dst['Occupation'].unique():
    occupation[grp] = dst['Balance'][dst['Occupation']==grp].values
args = occupation.values()
args = [occupation[grp] for grp in sorted(dst['Occupation'].unique())]
st.kruskal(*args)
```

```
KruskalResult(statistic=582.542507805765, pvalue=6.135488837094009e-126)
```