# **Input:**

The input for the script is path of the text file (please refer to line 528 in the python script for calling the main function)

Main function starts at the line 106

Way of approach:   
 Regular Expression based approach

# Ways to find the p-value:

## p-value along with a test statistic: (please see Regular Expressions from line 157 to 191)

Following p-values and the test statistics are extracted from the script: For Example: t (20) = 1.2, p = 0.05

a. T test statistic

𝑡 (𝑑 𝑓) = 𝑓𝑙𝑜𝑎𝑡, 𝑝 (=, >, <) 𝑓l𝑜𝑎𝑡

𝑡 = 𝑓 𝑙𝑜𝑎𝑡, 𝑝 (=, >, <) 𝑓𝑙𝑜𝑎𝑡

where 𝑡 is t-score test statistic, 𝑝 is 𝑝-value, 𝑑 𝑓 is one degree of

freedom as per the T hypothesis test

b. F test statistic

𝐹 (𝑑 𝑓 1, 𝑑 𝑓 2) = 𝑓𝑙𝑜𝑎𝑡, 𝑝 (=, >, <) 𝑓𝑙𝑜𝑎𝑡

where 𝐹 is F-score test statistic, 𝑝 is 𝑝-value, 𝑑 𝑓 1 and 𝑑 𝑓 2 are two

degrees of freedom as per the ANOVA hypothesis test

c. correlation r test statistic

𝑟 (𝑑 𝑓) = 𝑓l𝑜𝑎𝑡, 𝑝 (=, >, <) 𝑓𝑙𝑜𝑎𝑡

𝑟 = 𝑓𝑙𝑜𝑎𝑡, 𝑝 (=, >, <) 𝑓𝑙𝑜𝑎𝑡

where r is correlation r test statistic, 𝑝 is 𝑝-value, 𝑑 𝑓 is one degree

of freedom based on correlation hypothesis test

d. Chi-square test statistic

𝜒2 (𝑑 𝑓, 𝑁 = 𝑖𝑛𝑡) = 𝑓𝑙𝑜𝑎𝑡, 𝑝 (=, >, <) 𝑓𝑙𝑜𝑎𝑡

where 𝜒2 is the Chi-square test statistic, 𝑝 is the 𝑝-value, 𝑑 𝑓 is the

degrees of freedom, 𝑁 is the sample size

e. Z test statistic

𝑍 = 𝑓l𝑜𝑎𝑡, 𝑝 (=, >, <) 𝑓𝑙𝑜𝑎𝑡

where 𝑧 is the 𝑧 test statistic, 𝑝 is the 𝑝-value

f. Q test statistic

𝑄 (𝑏, 𝑏𝑒𝑡𝑤𝑒𝑒𝑛, 𝑤, 𝑤𝑖𝑡ℎ𝑖𝑛) (𝑑 𝑓) = 𝑓𝑙𝑜𝑎𝑡, 𝑝 (=, >, <) 𝑓𝑙𝑜𝑎𝑡

where 𝑄 is the 𝑄 test statistic, 𝑏, 𝑏𝑒𝑡𝑤𝑒𝑒𝑛, 𝑤, 𝑤𝑖𝑡ℎ𝑖𝑛 defines if the

statistic is between or within, 𝑝 is the 𝑝-value, 𝑑 𝑓 is degrees of

freedom

g. logistic Regression/OR test statistic

𝑂𝑅 = 𝑓𝑙𝑜𝑎𝑡, 𝑝 (=, >, <) 𝑓𝑙𝑜𝑎𝑡

where 𝑂𝑅 is the 𝑂𝑅 test statistic, 𝑝 is the 𝑝-value

h. b test statistic

𝑏 (=, <, >) 𝑓𝑙𝑜𝑎𝑡, 𝑝 (=, >, <) 𝑓𝑙𝑜𝑎𝑡

where 𝑏 is the 𝑏 test statistic, 𝑝 is the 𝑝-value

i. d test statistic

𝑑 (=, <, >) 𝑓𝑙𝑜𝑎𝑡, 𝑝 (=, >, <) 𝑓𝑙𝑜𝑎𝑡

where 𝑑 is the 𝑑 test statistic, 𝑝 is the p-value

j. Hazard Ratio test statistic

𝐻𝑅 = 𝑓𝑙𝑜𝑎𝑡, 𝑝 (=, >, <) 𝑓𝑙𝑜𝑎𝑡

where 𝐻𝑅 is the Hazard Ratio test statistic, 𝑝 is the 𝑝-value

## p-value without a test statistic: (please see line for Regular Expression 191)

A usual inequality or equality expression. For example: p=0.01, p< 0.03, p>0.005

## p-value from the claims given in the form of CSV (calling csv at the main function)

## if the paper defines a p-value range in the form of (for example: p = 01-3.4).

please find the Regular Expression on line 398

# ways to find sample size:

## directly from the text (intext sample size) (refer to line 376)

Representation: 𝑁 = 𝑖𝑛𝑡 or 𝑛 = 𝑖𝑛𝑡𝑒𝑔𝑒𝑟

## 2. The statistical test statistic and 𝑝-value expression can state the

sample size in a mathematical way.

Representation: The test statistic included p-value expression which

includes a sample size is represented as 8. Here the sample size in

the expression is defined as 𝑁 = 𝑖𝑛𝑡 but also when ’𝑁 =’ is missing

we can state that second argument inside 𝜒2 is the sample size.

3. Sample size can be formulated and computed from the test statistic

expressions mentioned in 3.7.1. A few of such test statistics

considered to find the sample size are:

a. T-statistic (3)

Expression: 𝑡 (𝑑𝑓) = 𝑓 𝑙𝑜𝑎𝑡, 𝑝 (=, >, <) 𝑓 𝑙𝑜𝑎𝑡

From the T-statistic expression the sample size (N) is calculated:

𝑁 = 𝑑 𝑓 + 1

b. F-statistic (5)

Expression: 𝐹 (𝑑𝑓1, 𝑑𝑓2) = 𝑓 𝑙𝑜𝑎𝑡, 𝑝 (=, >, <) 𝑓 𝑙𝑜𝑎𝑡

From the F-statistic expression the sample size (N) is calculated as:

𝑁 = 𝑑 𝑓 1 + 𝑑 𝑓 2 + 1

c. correlation statistic (6)

Expression: 𝑟 (𝑑𝑓) = 𝑓 𝑙𝑜𝑎𝑡, 𝑝 (=, >, <) 𝑓 𝑙𝑜𝑎𝑡

From the correlation statistical expression, the sample size (N) is

calculated as:

𝑁 = 𝑑 𝑓 + 2

d. 𝜒2 statistic (8)

Expression: 𝜒2 (𝑑𝑓, 𝑁 = 𝑖𝑛𝑡) = 𝑓𝑙𝑜𝑎𝑡, 𝑝 (=, >, <) 𝑓𝑙𝑜𝑎𝑡

From the 𝜒2 expression the sample is appeared as N = int within the expression

# **Output:**

A dictionary of p-value features is returned. The return statement is at the line 522.

Sample example: {'num\_hypo\_tested': 1, 'real\_p': 0.05, 'real\_p\_sign': -1, 'p\_val\_range': 0.0, 'num\_significant': 1, 'sample\_size': 811.0, 'extend\_p': 0}

## Output Feature Definitions:

‘num\_hypo\_tested’ = number of p-value equations along with the test statistic are found in a paper

‘real\_p’ = smallest p-value found amoung all the p-values defined in the paper

‘real\_p\_sign’ = sign respective the ‘real\_p’ feature

‘p-val\_range’ = difference of maximum p-value and minimum p-value among all the p-values found

‘num\_significant’ = the count of all the p-values which are less than or equal to 0.05

‘sample\_size’ = maximum of the list of all the sample sizes found

‘extended\_p’ = flag that is set in case the other p-value features are derived from non-statistical p-values