

A simplified way to analyze data



# **NINS-STAT**

*An Automated Statistical Analysis Software*

**OCTOBER 2021**



NEUROIMAGING AND NEUROSPECTROSCOPY LAB,  
NATIONAL BRAIN RESEARCH CENTER,  
MANESAR, INDIA

This toolbox is being developed under the supervision of Dr. Pravat Kumar Mandal, at the Neuroimaging and Neurospectroscopy lab, National brain Research Center, Manesar.

For any further clarifications and bug reporting please report at [pravat.mandal@gmail.com](mailto:pravat.mandal@gmail.com).

*First release, OCTOBER 2021*

# Contents

<b>1</b>	<b>Background .....</b>	<b>7</b>
1.1	What is Bio-statistics ?	7
1.2	Errors in testing of Hypothesis	7
1.3	The P Value	8
<b>2</b>	<b>NINS-STAT - A Brief Introduction .....</b>	<b>11</b>
2.1	NINS-STAT - Introduction	11
2.2	NINS-STAT - Features	12
2.3	NINS-STAT - Pre-requisites	12
2.3.1	Add R and Python in Linux and Windows .....	13
2.4	NINS-STAT - Input Data	17
2.5	NINS-STAT - Study Designs and Objectives	18
<b>3</b>	<b>NINS-STAT - Objectives and Statistics .....</b>	<b>21</b>
3.1	Objective Pipelines	21
<b>4</b>	<b>Steps to run - Automation Pipeline .....</b>	<b>35</b>
<b>5</b>	<b>Steps to run - Descriptive Analysis .....</b>	<b>43</b>
<b>6</b>	<b>Steps to run - Data Visualization .....</b>	<b>45</b>



## List of Figures

2.1	Search environment variable . . . . .	14
2.2	System properties interface opens . . . . .	15
2.3	Environment variables for a system . . . . .	15
2.4	Edit system variable . . . . .	16
2.5	Add path to system variable . . . . .	16
2.6	R console opens in command prompt . . . . .	17
2.7	Input data in wide format. . . . .	17
2.8	Complete Procedural Flowchart . . . . .	18
4.1	Initial GUI Interface . . . . .	36
4.2	Browse Tab . . . . .	36
4.3	Upload Data . . . . .	37
4.4	Data details . . . . .	37
4.5	Initial GUI . . . . .	38
4.6	Objective List . . . . .	39
4.7	Analysis Results . . . . .	40
4.8	Upload Complete . . . . .	40
4.9	Analysis Parameters . . . . .	41
4.10	Results . . . . .	41
5.1	Descriptive analysis GUI . . . . .	44
5.2	Descriptive analysis results with Grouping variable (Group 0, Group 1, Group 2) . . . . .	44
6.1	Visualization analysis GUI . . . . .	46
6.2	Scatter plot . . . . .	46





# 1. Background

## 1.1 What is Bio-statistics ?

While undertaking any research an understanding of basics of statistics is necessary for scientific planning, appropriate analysis and valid interpretation of findings. Importantly statistics deals with variables which are broadly categorized as Quantitative and Qualitative variables. The parameter of interest in a research study could be either a proportion (for categorical response) or an average (for quantitative outcome). Quantitative variables are those which are measurable and are always a number which can be either Continuous or Discrete. On the other hand qualitative variables are non-numerical which can be either Ordinal or Nominal. Qualitative variables also called "categorical variables," classify or divide individuals into classes/groups, so that quantities variables may be compared among the classes. **All research hypothesis are related to claim for improved outcome and superior drug/diagnostics. This claim is required to be tested against a null hypothesis negating the claim by indicating that the effect of interest is zero.** In most of the situations the Statistical null hypothesis is the negation of the research hypothesis. Commonly Used Null Hypothesis is indicated in the diagram below:

The null hypothesis is evaluated against a competing hypothesis (alternative hypothesis) in which the effect of interest is considered as non-zero.

Hypotheses are always stated in terms of population parameter, such as the mean ( $\mu$ ) or proportion (P). An alternative hypothesis may be one-sided or two-sided. A one-sided hypothesis claims that the parameter under the alternative hypothesis is either larger or smaller than the value given by the null hypothesis. A two-sided hypothesis claims that the parameter is not equal to the value given by the null hypothesis - the direction does not matter and could be either side.

## 1.2 Errors in testing of Hypothesis

In testing of Hypothesis there are probabilities of committing two kinds of errors, false rejection (Type I error) and false acceptance (Type II error) of the null hypothesis. These are as under:

Type 1 error	False rejection	Reject $H_0$ when $H_0$ is true
Type 2 error	False acceptance	Accept $H_0$ when $H_0$ is false

In order to explain these concepts, let us consider the possible errors in Clinical Practice as under:

- Diagnosing a person wrongly as suffering from the disease, when he/she is not.
- Missing the diagnosis and person is declared as free from the disease when he/she is actually suffering from the disease.

The Null Hypothesis in the present case is 'Person is not suffering from the disease'.

Type 1 error	Diagnosing a person who is not suffering from the disease as suffering from the disease would imply Rejecting Null Hypothesis when it is true.
Type 2 error	Declaring a person free from the disease when he/she is actually suffering from the disease would imply Accepting the Null Hypothesis when it is false.

The probabilities of Type I error and Type II error are denoted by  $\alpha$  and  $\beta$  respectively. There are two terms **Confidence level** and **Power** which are complement of these errors as under:

**Confidence Level** = Probability of accepting the null hypothesis when it is true.  
 = 1 - Probability of type I error  
 = 1 -  $\alpha$

**Power** = Probability of rejecting the null hypothesis when it is false.  
 = 1 - Probability of type II error  
 = 1 -  $\beta$

Importantly,  $\alpha$  is commonly referred to as level of significance.

The implication of Type I error in the context of a clinical research is that the new regimen, although not effective, is adopted and prescribed, i.e. an ineffective drug is allowed to be marketed.

In a trial on a new regimen, Type II error would imply that the new regimen is not approved when it is actually effective. Thus the medical profession and the society is deprived of the benefits of this new regimen.

### 1.3 The P Value

The p value is calculated based on the Null distribution, which is a distribution of the test statistic when the null hypothesis is true. The p-value indicates how probable the results are due to chance. Statistical inferences indicating the strength of the evidence corresponding to different values of p are explained as under:



Values of P	Inferences
$p > 0.10$	No evidence against the Null Hypothesis.
$0.05 < p < 0.10$	Weak evidence against the Null Hypothesis.
$0.01 < p < 0.05$	Good evidence against Null Hypothesis.
$0.05 < p < 0.001$	Strong evidence against the Null Hypothesis.
$p < 0.001$	Very strong evidence against the Null Hypothesis.

Conventionally,  $p < 0.05$  is referred as statistically significant and  $p < 0.001$  as statistically highly significant. When presenting p values it is a common practice to use the asterisk rating system.





## 2. NINS-STAT - A Brief Introduction

### 2.1 NINS-STAT - Introduction

In evidence-based research, statistical analysis and data visualization are the two crucial facets. . In research-oriented disciplines where data are collected through scientifically planned studies, , a well-informed coherent pathway needs to be arrived at to test the hypothesis and to derive appropriate inference and conclusions using scientific statistical principle . Simultaneously, there should be minimum subtleties involved while performing such data exploration and in choosing the appropriate statistical method for data analysis. Traditional computer programming languages such as FORTRAN, C, C++, and Java were unremarkably used in the past however, these usually require a researcher to have a basic understanding about the underlying mathematics and statistical theories within each method. Additionally, it also expects the user to be adept at modern computer programming fundamentals which presents a significant barrier for researchers having little exposure to mathematical or computational sciences. Researchers have frequently acknowledged the difficulty in identifying the appropriate statistical test for their study as a compelling obstacle. The advent of statistical analysis software such as SAS, Stata, R, SPSS, MedCalc, R, GraphPad, Prism and OriginPro have made the process of data analysis simpler but have not yet addressed this shortfall.

The shortfall lies in the fact that they do not consider underlying assumptions in use of specific tests for statistical analysis., which is desirable to identify the appropriate statistical method to answer the specific research question. Some of the important assumptions are normality, homogeneity linearity, sample size etc. The scientific validity and the credibility of a study depend on the use of right study design to address the objectives of the study. The broad classification of study design are observational or interventional/experimental. The observation studies include cross-sectional study, cohort or longitudinal study, case-control study, survival analysis, diagnostic accuracy study. The interventional studies could be randomized or non-randomized control trials. Each of these study designs is associated with one or more study objectives like the comparison of proportion or mean, association, agreement, diagnostic accuracy, regression, and survival analysis. Each of these study objectives has its own underlying set of statistical analysis methodologies. The entire procedure requires the researcher to have a meticulous understanding of statistics to choose the appropriate statistical method to ascertain the study outcome.

A review of related literature revealed recent attempts to integrate automation into statistics. One of the studies demonstrated RBiplot an R based package, to perform automated statistics and data visualization

in the field of molecular biology and biochemistry. Another package, The Automatic Statistician, performs on artificial intelligence-based statistical analysis for data science applications. REGSTATTOOLS is another attempt towards a web-based tool intended to provide analysis for the burden of cancer, or other group of disease registry data. REGSTATTOOLS includes three software applications: SART (analysis of disease's rates and its time trends), RiskDiff (analysis of percent changes in the rates due to demographic factors and risk of developing or dying from a disease) and WAERS (relative survival analysis). Similarly, a python-based domain-specific language (DSL) 'Tea' has been developed to integrate automation into statistical analysis limited to null hypothesis statistical testing (NHST) based modules. All the packages attempted to integrate automation into statistics however, they are bound by certain restrictions. The functionality of RBioplot is restricted to molecular biology and biochemistry just as the Automatic Statistician is restricted to data science application only [4]. Similarly, REGSTATTOOLS attempts to perform automatic calculations for a specific segment of an entire methodology [5] and Tea integrates statistics limited to null NHST modules and requires a basic understanding of the python programming language.

While several statistical packages are available, very few have focused on automating the statistical method selection. To the best of our knowledge, no graphical user interface (GUI) based automation-enabled statistical package was found that took into consideration different clinical study designs as well as research objectives accounting for diverse biomedical research applications for the final statistical test selection. Taking this into account, in this manuscript, we present, a Matlab based automated statistical analysis software, 'NINS-STAT' with a user-friendly GUI that implements automated test selection and execution using limited inputs from the researcher. The manuscript provides an overview of the methodology implemented in developing this toolbox with an illustrative algorithm along with functions that accumulate the acquired information for translating it into a constraint satisfaction problem (CSP). NINS-STAT also helps to identify the appropriate statistical test for hypothesis testing formed using the CSP based on the study design, objectives, and data provided. The manuscript further discusses the functions available and performs a comparison of the obtained result with other unremarkably used statistical software packages. Furthermore, the manuscript also presents results obtained from NINS-STAT using secondary data obtained from clinical studies with dissimilar research objectives. The implementation of the automation workflows integrated into the GUI, NINS-STAT, does not require prior programming experience or a detailed understanding of the mathematical underpinnings of a statistical method.

## 2.2 NINS-STAT - Features

To offer operational efficiency to NINS-STAT, a user-friendly GUI was designed with consideration of various important unique features. The NINS-STAT front panel is designed to converge in a **single window** which helps the user to visualize all the components in a single window, making **transparent** and **automated workflow** appearance to identify the inputs demanded to attain the outcome of interest. Each panel segment is designed in a **clean** and **attractive** approach for **ease** and **convenient field selection**. The statistical output from the NINS-STAT is also generated in a **minimalist** approach so as to avoid overwhelmed information making it difficult to interpret and digest. Therefore, it presents the results and graphs in a **clean** and is **publication-ready** which further can be saved in **traditional save as options**, allowing for successive disclosure of results. All these above design considerations were manifested while developing the GUI structural architecture.

## 2.3 NINS-STAT - Pre-requisites

These are the basic requirements that are user needs to have before running NINS-STAT on their systems. Requirement have been described below into software and hardware configurations.

Domain	Pre-requisites
Study Information	<b>Study design</b> - Cross-sectional, cohort, case-control, randomized and non-randomized. <b>Study objectives</b> - Prevalence, incidence, relative risk, odds ratio, proportions, means, associations, regression, survival analysis, agreement and diagnostic parameters.
Input Data	<b>Raw data extension</b> - Excel sheet with .xlsx extension. <b>Input data structure</b> - Wide format.
Software	<b>MATLAB</b> - R2017b and above. Add R and Python as environment variables to enable tests and procedures. Procedure has been explained below.
System	<b>Operating system</b> - Windows or Linux <b>Processor</b> - Inter or AMD x86-64 processor <b>Disk Space</b> - 4-6 GB for a typical installation (2 GB for MATLAB only) <b>RAM</b> - 2GB and above <b>Graphics</b> - No specific graphic card is required.

### 2.3.1 Add R and Python in Linux and Windows

#### 1. Windows environment

**Add Java JRE** - If JAVA Runtime Environment is not installed on your windows system.

- Download file from <https://www.oracle.com/in/java/technologies/javase-downloads.html>.
- Install it the usual way.

**Add R** - If R is not set as environment variable. Same steps can be followed for setting **python** as environment variables.



Test if R is set as environment variable or not. Open command prompt and then type 'R' in the command line. If R command line opens up then the path is set. Otherwise if the following message is displayed '**R' is not recognized as an internal or external command, operable program or batch file.**', in that case follow the below given steps.

2. Download R files from <https://cran.r-project.org/bin/windows/base/>.
3. Install it the usual way.
4. Open it's location.
5. Search environment variable as shown below.

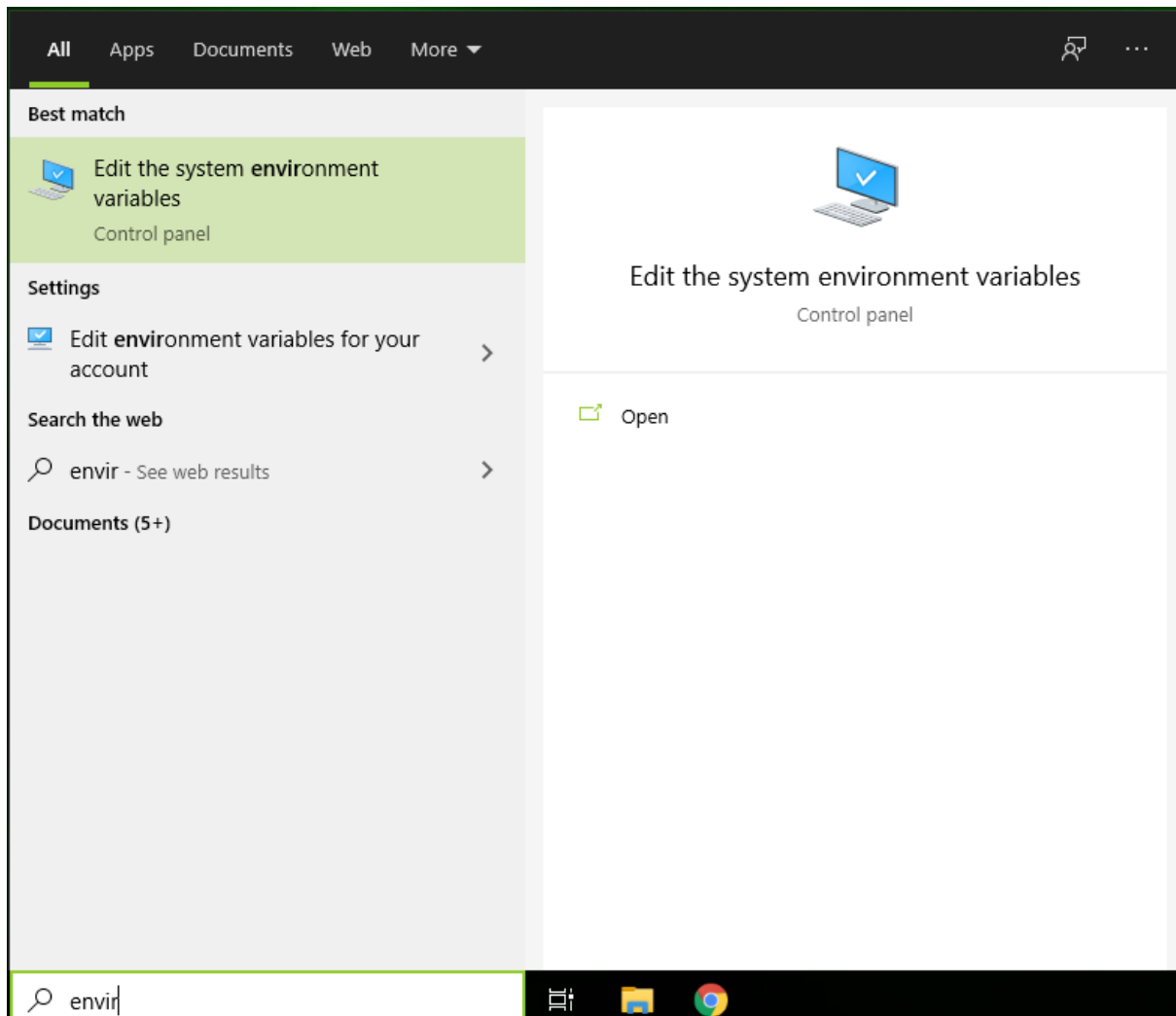


Figure 2.1: Search environment variable

6. Click on the option on start menu. An interface similar to the figure shown below opens. Click on the **Environment Variables** button to get to Fig 2.3.



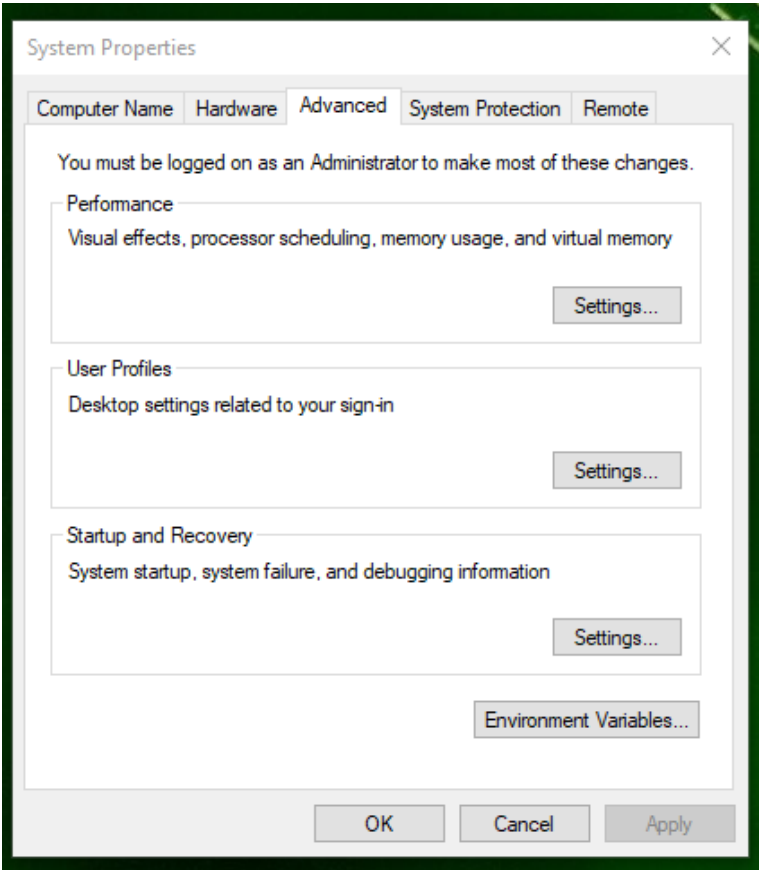


Figure 2.2: System properties interface opens

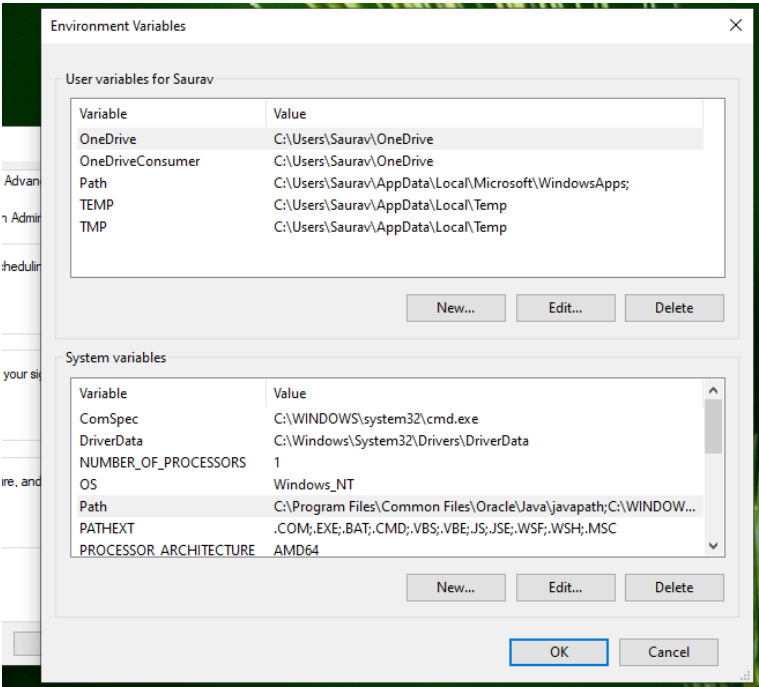


Figure 2.3: Environment variables for a system

7. double click the **Path** variable to add the path of R.

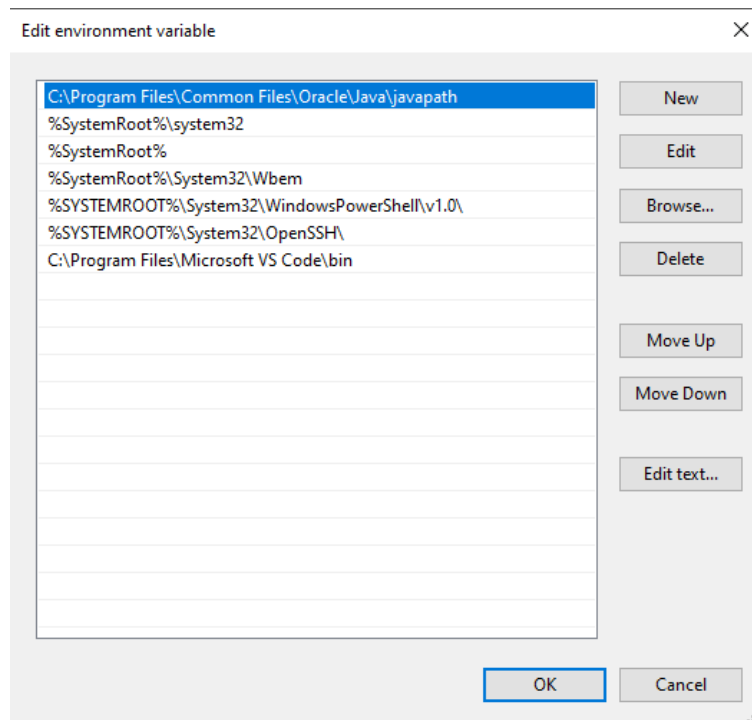


Figure 2.4: Edit system variable

8. Add path of location where R is installed on your system.

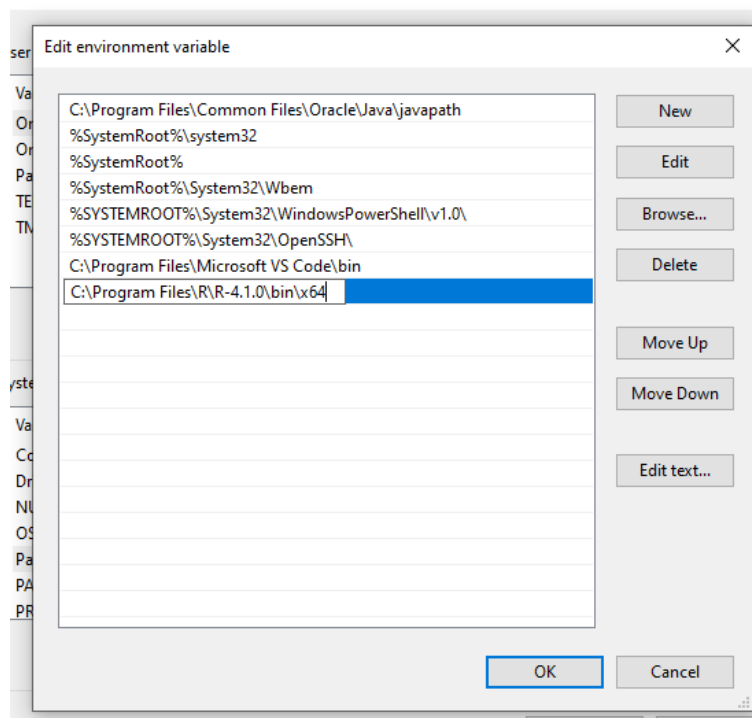


Figure 2.5: Add path to system variable

9. Once added. Click **OK** and exit.
10. See if R environment is set correctly. Open **command prompt** and write R on the console. A console similar to the one shown will open on the command prompt.

```

Rterm (64-bit)
Microsoft Windows [Version 10.0.19042.1055]
(c) Microsoft Corporation. All rights reserved.

C:\Users\Saurav>R

R version 4.1.0 (2021-05-18) -- "Camp Pontanezen"
Copyright (C) 2021 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

>

```

Figure 2.6: R console opens in command prompt

## 2.4 NINS-STAT - Input Data

The 'wide' format data (.xlsx) to be imported to NINS-STAT GUI has been represented below. it shows the wide format of data having a common row for an individual observation with multiple columns as data variables. Only the first row must be the variable names and the consecutive rows will have the individual observations. Each field must be encoded to a numerical value only, as string variables cannot be analyzed.

Variables	S.No	Participant ID	Age	Gender	Stages of Disease	Blood Pressure (Baseline)	Blood Pressure (after 1 month)
Observations	1	Participant A	25	1	1	110	115
	2	Participant B	15	0	2	90	100
	3	Participant C	28	1	3	120	130

Gender  
 -----  
 1 - Male  
 0 - Female

Stages of Disease  
 -----  
 1 - Mild  
 2 - Moderate  
 3 - Severe

Figure 2.7: Input data in wide format.

## 2.5 NINS-STAT - Study Designs and Objectives

Clinical based studies are reliant on an appropriately designed study design. Design selection is vital to a study since a poorly designed study is non-recoverable as opposed to a poorly analyzed study which is easily recoverable via reanalysis to arrive at a meaningful outcome. Clinical study designs and study objectives are the two most important factors for identifying the accurate statistical method for data analysis. Each study design has their own set of advantages and disadvantages. Therefore, it is essential to recognize and understand the type of study a researcher is planning to conduct to answer their research question. Defining a study into a specific type communicates a lot about the type of data to be collected and the data collection method, strengths and weaknesses underpin its design and conduct and the statistical analysis planning and execution and study outcomes.

Clinical based research studies are generally categorized broadly into observational studies and experimental studies or clinical trial. Observational studies are based on observations of a risk factor, diagnostic test, treatment, or other intervention without trying to change who is or is not exposed to it. They are further described as below.

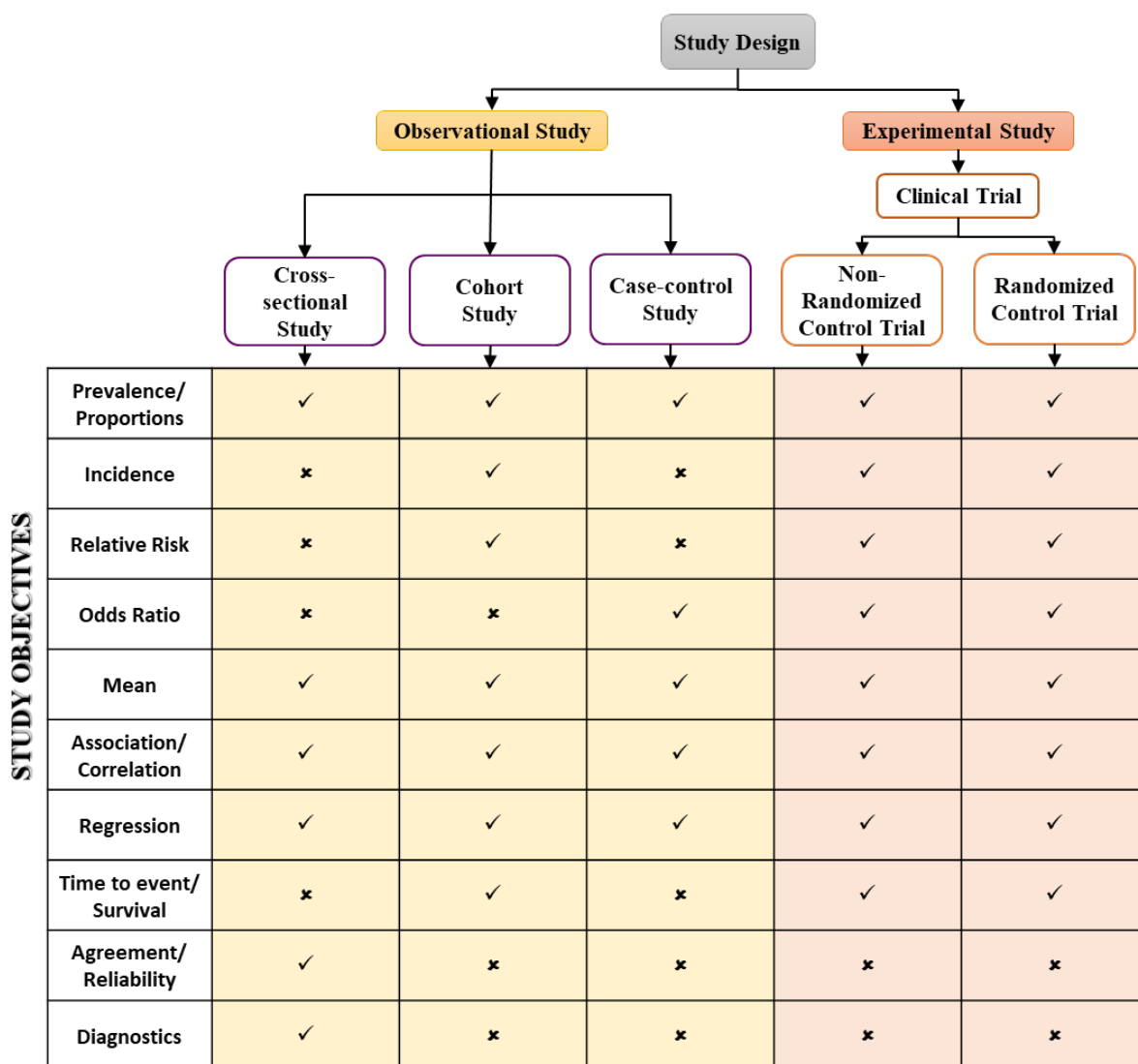


Figure 2.8: Complete Procedural Flowchart

1. **Cross-sectional / Longitudinal Study** - In a longitudinal study subjects are followed over time with continuous or repeated monitoring of risk factors or health outcomes, or both. Most longitudinal studies examine associations between exposure to known or suspected causes of disease and subsequent morbidity or mortality. In the simplest design a sample or cohort of subjects exposed to a risk factor is identified along with a sample of unexposed controls. The two groups are then followed up prospectively, and the incidence of disease in each is measured. By comparing the incidence rates, attributable and relative risks can be estimated. Allowance can be made for suspected confounding factors either by matching the controls to the exposed subjects so that they have a similar pattern of exposure to the confounder, or by measuring exposure to the confounder in each group and adjusting for any difference in the statistical analysis.
2. **Cohort Study** - Cohort studies are a type of longitudinal study—an approach that follows research participants over a period of time (often many years). Specifically, cohort studies recruit and follow participants who share a common characteristic, such as a particular occupation or demographic similarity. Cohort studies are types of observational studies in which a cohort, or a group of individuals sharing some characteristic, are followed up over time, and outcomes are measured at one or more time points.
3. **Case-control Study** - A study that compares patients who have a disease or outcome of interest (cases) with patients who do not have the disease or outcome (controls), and looks back retrospectively to compare how frequently the exposure to a risk factor is present in each group to determine the relationship between the risk factor and the disease. Case control studies are observational because no intervention is attempted and no attempt is made to alter the course of the disease. The goal is to retrospectively determine the exposure to the risk factor of interest from each of the two groups of individuals: cases and controls. These studies are designed to estimate odds.
4. **Randomised Control Trial** - The randomised control trial (RCT) is a trial in which subjects are randomly assigned to one of two groups: one (the experimental group) receiving the intervention that is being tested, and the other (the comparison group or control) receiving an alternative (conventional) treatment. The two groups are then followed up to see if there are any differences between them in outcome. The results and subsequent analysis of the trial are used to assess the effectiveness of the intervention, which is the extent to which a treatment, procedure, or service does patients more good than harm. RCTs are the most stringent way of determining whether a cause-effect relation exists between the intervention and the outcome.
5. **Non-randomised Control Trial** - A clinical trial in which the participants are not assigned by chance to different treatment groups. Participants may choose which group they want to be in, or they may be assigned to the groups by the researchers.





## 3. NINS-STAT - Objectives and Statistics

The chapter gives a brief description regarding all the statistical tests that have been included in each objective. The purpose of this is to allow the researcher to know the tests that have been included.

### 3.1 Objective Pipelines

#### 1. Test of Proportions

- **Binomial/ One sample Z-test of proportions**

Description: In case of a one sample test, the test statistic 'z' is given by -

$$Z = \frac{p - P}{\sqrt{P(1 - P)/n}}$$

Here,

$p$  = Sample proportions.

$P$  = Proportion under null hypothesis.

$n$  = Sample size.

- **Two sample Z-Test of proportions**

Description: Many times proportions (cure rate, prevalence of a condition, etc.) for 2 groups are required to be compared and in such situations two-sample z test is used, which is given by -

$$Z = \frac{p_1 - p_2}{\sqrt{(p_1(1 - p_1)/n_1) + (p_2(1 - p_2)/n_2)}}$$

Here,

$p_1$  and  $p_2$  = Proportions for sample 1 and sample 2.

$n_1$  and  $n_2$  = Sizes for sample 1 and sample 2.

#### 2. Fisher Exact Test

Description: Fisher exact test, similar to Z-test of proportions above, is used for testing the equality of proportion of individuals possessing a certain trait in two populations when the sample sizes are

small. This test is also used for finding association between two categorical variables which will be discussed later.

### 3. Test of Means

- **One Sample Z-test**

Description: In this situation, the null hypothesis is  $H_0 : \mu = \mu_0$  i.e. the population mean  $\mu$  is equal to a given value  $\mu_0$ . The alternative hypothesis could be  $\mu \neq \mu_0$  or  $\mu < 0$  or  $\mu > 0$ . For such situations the appropriate significance test is 'Z test' which is defined as -

$$Z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

Here,

$n$  = Sample size.

$\bar{x}$  = Sample mean.

The test statistic follow the standard normal distribution.

- **One Sample t-test**

Description: In most practical situations, the standard deviation for the population will not be known. In such situations, the standard deviation  $\sigma$  in the formula is replaced by the estimated standard deviation ( $s$ ). The distribution of the sample mean  $\bar{x}$  is no longer normal. Instead, the sample mean follows a  $t$  distribution. The test statistic  $t$  is defined as -

$$Z = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$$

The  $t$  distribution is described by its degrees of freedom. For a sample of size  $n$ , the  $t$  distribution will have  $(n - 1)$  degrees of freedom. As the sample size  $n$  increases, the  $t$  distribution approaches to the normal distribution.

- **Paired-t test**

Description: In many experiments, one is interested in comparing the measurements for the same set of individuals in 'before' and 'after' situation. Such studies have a matched pair design, where the difference between the two measurements in each pair is the variable of interest.

Analysis of data from a matched pairs experiment compares the two measurements by subtracting one from the other and basing the test hypothesis upon the differences. Usually, the null hypothesis  $H_0$  assumes that the mean of these differences is equal to 0, while the alternative hypothesis  $H_1$  would be that the mean of the differences is not equal to zero (the alternative hypothesis may be one- or two-sided, depending on the experiment). In this situation, using the differences between the paired measurements as single observation, the standard t-test for one sample is followed. The difference for each pair of values is denoted by  $d$ . Once, this is obtained, the situation reduces to one sample  $t$  test given by -

$$t = \frac{\bar{d}}{(S_d / \sqrt{n})}$$

where  $\bar{d}$  is the sample mean and  $s_d$  the sample standard deviation of the differences.

- **Wilcoxon Sign Rank Test**

Description: The nonparametric analog of the t-test is the Wilcoxon Signed-Rank Test and may be used when the one-sample t- test assumptions are violated. While using t-test, the key assumption relates to normality. If an assumption is not met even approximately, the significance levels and the power of the t-test are invalidated. Unfortunately, in practice it often happens that several assumptions are not met. In such situations Wilcoxon Signed- Rank Test is used. The assumptions of the Wilcoxon signed-rank test are as follows (note that the difference is between a data value and the hypothesized median or between the two data values of a pair):

- (a) The differences are continuous (not discrete).
- (b) The distribution of each difference is symmetric.
- (c) The differences are mutually independent.

The Wilcoxon signed-rank test is a popular, non-parametric substitute for the t-test. It assumes that the data follow a symmetric distribution. The test is computed using the following steps.

- (a) Subtract the hypothesized mean,  $\mu_0$ , from each data value. Rank the values according to their absolute values.
- (b) Compute the sum of the positive ranks  $S_p$  and the sum of the negative ranks  $S_n$ . The test statistic,  $W_R$ , is the minimum of  $S_p$  and  $S_n$ .
- (c) Compute the mean and standard deviation of  $W_R$  using the formulas

$$\mu_{W_R} = \frac{n(n+1)}{4}$$

$$\sigma_{W_R} = \sqrt{\frac{n(n+1)(2n+1)}{24} - \frac{\sum t^3 - \sum t}{48}}$$

where  $t$  = number of times the  $i$ th value occurs.

$$Z_W = \frac{W_R - \mu_{W_R}}{\sigma_{W_R}}$$

The significance of the test statistic is determined by computing the p-value using the standard normal distribution. If this p-value is less than a specified level (usually 0.05), the null hypothesis is rejected in favor of the alternative hypothesis. Otherwise, no conclusion can be reached.

#### • Levene Test

Description: An F-test is used to test if the variances of two populations are equal. This test can be a two-tailed test or a one-tailed test. The two-tailed version tests against the alternative that the variances are not equal. The F hypothesis test is defined as:

$$H_0 : \sigma_1^2 = \sigma_2^2$$

Similarly for a two tailed test,

$$H_a : \sigma_1^2 \neq \sigma_2^2$$

Test statistic:

$$F = \frac{s_1^2}{s_2^2}$$

where  $s_1^2$  and  $s_2^2$  are the sample variances. The more this ratio deviates from 1, the stronger the evidence for unequal population variances. The hypothesis that the two variances are equal is rejected if

$$F < F_{1-\frac{\alpha}{2}, N_1-1, N_2-1}$$

where  $F_{\alpha}, N_1 - 1, N_2 - 1$  is the critical value of the F distribution with  $N_1 - 1$  and  $N_2 - 1$  degrees of freedom and a significance level of  $\alpha$ . Levene's test (1960) is an alternative test used to test if two samples have equal variances. The formula for which is

$$W = \frac{(N - k)}{(k - 1)} \cdot \frac{\sum_{i=1}^k N_i (Z_{i.} - Z_{..})^2}{\sum_{i=1}^k \sum_{j=1}^{N_i} (Z_{ij} - Z_{i.})^2}$$

where,  $W$  is distributed as  $F$  with  $N - k$  and  $k - 1$  degrees of freedom.

Equal variances across samples is called homogeneity of variance. Some statistical tests, assume that variances are equal. The Levene test can be used to verify that assumption.

### • 2-Sample t-test

Description: In situations dealing with comparison of mean of two populations i.e. averages of one group with that of another group to test whether they are equal or not, a two sample Student's t-test is used. In this independent samples are drawn from two populations. Based on the sample values the formula for two sample  $t$  test is given by -

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{s \sqrt{(1/n_1 + 1/n_2)}}$$

Here,

$n_1$  and  $n_2$  = Sample sizes of sample 1 and sample 2.

$\bar{x}_1$  and  $\bar{x}_2$  = Mean of sample 1 and sample 2.

$s$  = Pooled sample standard deviation.

Subscript 1 and 2 refer to group 1 and group 2 respectively. here the degrees of freedom will be  $n_1 + n_2 - 2$ .

### • Welch T-test

Description: Welch t-test is very similar to 2-sample t-test, the only difference being that underlying populations are normal with unequal variances, hence the test statistic changes, otherwise everything else and interpretations remain the same.

Based on the sample values the formula for Welch  $t$ -test is given by -

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{(s_1^2/n_1 + s_2^2/n_2)}}$$

Here,

$n_1$  and  $n_2$  = Sample sizes of sample 1 and sample 2.

$\bar{x}_1$  and  $\bar{x}_2$  = Mean of sample 1 and sample 2.

$s_1$  and  $s_2$  = standard deviation of sample 1 and sample 2.

Subscript 1 and 2 refer to group 1 and group 2 respectively. here the degrees of freedom will be  $n_1 + n_2 - 2$ .

- **Mann Whitney U Test**

Description: Mann-Whitney U is similar to Wilcoxon signed-ranks test except that the samples are independent and not paired. Null hypothesis: the population means are the same for the two groups.

Rank the combined data values for the two groups. Then find the average rank in each group. Then the U value is calculated using formula -

$$U = N_1 * N_2 + \frac{N_x(N_x + 1)}{2} - R_x$$

where,  $R_x$  is larger rank total.

- **ANOVA Test**

Description: The one-way analysis of variance (ANOVA) is used to determine whether there are any statistically significant differences between the means of three or more groups. Specifically, it tests the null hypothesis:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \dots = \mu_k$$

where  $\mu$  = group mean and  $k$  = number of groups.

The alternative hypothesis ( $H_a$ ), is that there are at least two group means that are statistically significantly different from each other. The computation involves calculation of total sum of squares, sum of square between groups. Thereafter sum of squares within groups is computed as the difference total sum of squares and sum of squares between groups. The total degrees of freedom is  $N-1$  where  $N$  is total number of observations. The degrees of freedom for comparing groups is  $k-1$ . The difference of these two will give degrees of freedom for within group as  $N-k$ . The results are presented as under with usual notations:

Source	SS	df	MS	F	Sig.
Between	$SS_b$	$k - 1$	$MS_b$	$\frac{MS_b}{MS_w}$	$p$ value
Within	$SS_w$	$N - k$	$MS_w$		
Total	$SS_b + SS_w$	$N - 1$			

F statistic is computed as above which has  $k-1$  and  $N-k$  degrees of freedom. The computed value of F is compared with table value of F using F distribution with degrees of freedom. ( $k-1$ ,  $N-k$ ) If computed value of F is less than the table value, the null hypothesis is not rejected.

- **Welch ANOVA Test**

Description: Welch ANOVA is used for the same reason as of ANOVA i.e. to access the equality of means of two or more groups, but we can use it even when underlying populations have unequal variances.

- **Krushkal Wallis Test**

Description: The test is more commonly used when there are three or more groups to compare. For two groups, Mann Whitney U Test is used. The variable could be in ordinal scale, Ratio Scale or Interval scale. All groups are assumed to have the same shape distributions.

$$H = \frac{12}{n(n+1)} \sum \frac{R_i^2}{n_i} - 3(n+1)$$

where,

$n$  = sum of sample sizes for all samples.

$c$  = number of samples.

$T_j$  = sum of raks in the  $j$ th sample.

$n_j$  = size of  $j$ th sample.

The critical chi-square value, with  $c-1$  degrees of freedom is determined and  $H$  value from above is compared with the critical chi-square value. If the critical chi-square value is less than the  $H$  statistic, reject the null hypothesis that the medians are equal. If the chi-square value is not less than the  $H$  statistic, there is not enough evidence to suggest that the medians are unequal.

#### 4. Test of Correlations/ Associations

- **Kendall Tau Coefficient**

Description: The Kendal's Tau Correlation coefficient measures the strength of association between two ordinal or ranked variables taking values between -1 to 1, with value near +1 implying perfect agreement and values near -1 implying perfect disagreement. Its formula is given by:

$$\tau = \frac{2}{n(n-1)} \sum_{i < j} \text{sgn}(x_i - x_j) \text{sgn}(y_i - y_j)$$

where  $\text{sgn}$  function gives the sign of the argument.

The corresponding hypothesis test involves testing  $H_0 : \tau = 0$  vs  $H_1 : \tau \neq 0$  and the result can be concluded using the p-value of the test.

- **Karl Pearson Test**

Description: The karl pearson correlation coefficient,  $r$ , tells us about the strength and direction of the linear relationship between  $x$  and  $y$ . The sample data are used to compute  $r$ , the correlation coefficient for the sample.

The sample correlation coefficient,  $r$ , is the estimate of the population correlation coefficient.

**Null hypothesis:**  $H_0 : \rho = 0$

The population correlation coefficient is not significantly different from zero.

**Alternate null hypothesis:**  $H_a : \rho \neq 0$

The population correlation coefficient is significantly different from zero. The p-value is calculated using a t-distribution with  $n-2$  degrees of freedom. The formula for the test statistic is-

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

If the computed value of  $t$  is less than the table value of  $t$  with  $n-2$  degrees of freedom from  $t$  distribution, the null hypothesis is not rejected.

- **Spearman's correlation coefficient**

Description: Spearman's correlation coefficient is a statistical measure of the strength of a monotonic relationship between paired data, which takes values between -1 to +1. Its formula is given by

$$\rho_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$



The hypothesis test involves testing  $H_0 : \rho_s = 0$  vs  $H_1 : \rho_s \neq 0$ . The test statistic is same as under Karl Pearson.

- **Point Biserial Correlation**

Description: The Point-Biserial Correlation measures the strength of association between a binary variable and a continuous variable. The point-biserial correlation is mathematically equivalent to the Pearson (product moment) correlation. The formula is given by:

$$r_{pb} = \frac{M_1 - M_0}{s_n} \sqrt{\frac{n_1 n_0}{n^2}}$$

where  $s_n$  is the standard deviation of the entire data.  $M_1$  being the mean value on the continuous variable  $X$  for all data points in group 1, and  $M_0$  the mean value on the continuous variable  $X$  for all data points in group 2. Further,  $n_1$  is the number of data points in group 1,  $n_0$  is the number of data points in group 2 and  $n$  is the total sample size.

The corresponding hypothesis test involves testing  $H_0 : r_{pb} = 0$  vs  $H_1 : r_{pb} \neq 0$  and the result can be concluded using the p-value of the test.

- **Fisher Exact Test**

Description: In medical research, a new drug/ intervention is generally compared with standard treatment and outcomes assessed in terms of cured or not cured, and occurrence and non-occurrence of side effects etc. These variables are generally dichotomous i.e. Involves only two possible values. The data is generally displayed in a 2 x 2 contingency table that shows the frequencies of occurrence of all combinations of the levels of two dichotomous variables. A research question of interest is whether the attributes in the contingency table are associated or independent of each other. The null hypothesis would be that there is no association or there is no difference in proportions. Two methods of analysis are available for this set-up.

- (a) Proportion tests based on Gaussian distribution.
- (b) Chi-square tests based on contingency table.

For large samples both tests are equivalent and give identical results.

When the row and column margins in 2 X 2 contingency table are fixed, either by design or for the analysis, independence of attributes can be tested using Fisher's exact test. This test is based on the hyper-geometric distribution. The 2 X 2 contingency table can be presented as under:

label	Outcome		
	Cured	Not-cured	Total
New Drug	$n_{11}$	$n_{12}$	$n_{1\cdot}$
Standard drug	$n_{21}$	$n_{22}$	$n_{2\cdot}$
Total	$n_{1\cdot}$	$n_{2\cdot}$	$n$

The computation of probability for the contingency table with given cell frequencies using hyper-geometric distribution is as under,

$$P_{n_{ij}} = \frac{n_{1\cdot}! n_{2\cdot}! n_{\cdot 1}! n_{\cdot 2}!}{n! n_{11}! n_{12}! n_{21}! n_{22}!}$$

**Computation of p values using Fisher's Exact Test** - To start with for given marginal totals all possible two way tables are generated which has desired marginal totals. Thereafter,

using the smallest marginal total the table for each ordered pair of integers with that sum is created. The 2 X 2 contingency table for each of these ordered pairs are then completed. The probability for each contingency table is computed using hyper-geometric distribution. For a one tailed test  $n_{11}$  is compared with its expected value computed as -

(Corresponding row total x Corresponding column total) / Grand total

If  $n_{11}$  exceeds its expected value then p value is the sum of the probabilities of all  $n_{11}$  more than equal to the expected value. Alternatively, if  $n_{11}$  is less than its expected value then p value is the sum of the probabilities of all  $n_{11}$  less than equal to the expected value. The hyper-geometric probability distribution is used to compute the probability of the observed results.

- **Chi-Square Test**

Description: Karl Pearson (1900) suggested use of test statistic as under:

$$\chi^2 = \frac{n(ad - bc)^2}{(a + b)(c + d)(a + c)(b + d)}$$

This test statistic under the null hypothesis follows a Chi-square  $\chi^2$  distribution with one degree of freedom.

## 5. Test of Agreement/ Concordance

- **Cohen Kappa Analysis**

Description: Cohen's kappa is a robust statistic useful for either interrater or intrarater reliability testing. Similar to correlation coefficients, it can range from -1 to +1, where 0 represents the amount of agreement that can be expected from random chance, and 1 represents perfect agreement between the raters. Calculation of Cohen's kappa may be performed according to the following formula:

$$\kappa = \frac{P_r(a) - P_r(e)}{1 - P_r(e)}$$

where,

$P_r(a)$  = Actual observed agreement.

$P_r(e)$  = represents chance agreement.

- **Weighted Cohen's Kappa Test**

Description: Cohen's kappa takes into account disagreement between the two raters, but not the degree of disagreement. This is especially relevant when the ratings are ordered. To address this issue, there is a modification to Cohen's kappa called weighted Cohen's kappa. The weighted kappa is calculated using a predefined table of weights which measure the degree of disagreement between the two raters, the higher the disagreement the higher the weight. The table of weights should be a symmetric matrix with zeros in the main diagonal (i.e. where there is agreement between the two judges) and positive values off the main diagonal. The farther apart are the judgments the higher the weights assigned. Using the notation from Cohen's Kappa where  $p_{ij}$  are the observed probabilities,  $e_{ij} = p_i q_j$  are the expected probabilities and  $w_{ij}$  are the weights (with  $w_{ji} = w_{ij}$ ), then

$$\kappa_w = 1 - \frac{\sum_{ij} w_{ij} p_{ij}}{\sum_{ij} w_{ij} e_{ij}}$$

The weighted kappa can be expressed as

$$\kappa_w = \frac{Pa(w) - Ps(w)}{1 - Ps(w)}$$

where

$$Pa(w) = \sum_i \sum_j v_{ij} p_{ij}$$

From these formulas, hypothesis testing can be done and confidence intervals calculated, as described in Cohen's Kappa.

- **Lin's Concordance Correlation Coefficient**

Description: Lin's concordance correlation coefficient (CCC) is the concordance between a new test or measurement (Y) and a gold standard test or measurement (X). This statistic quantifies the agreement between these two measures of the same variable (e.g. chemical concentration). Like a correlation, CCC ranges from -1 to 1, with perfect agreement at 1. It is assumed that  $n$  observations  $(Y_k, X_k)$ , are selected from a bivariate population with means  $\mu_Y$  and  $\mu_X$ , variances  $\sigma_Y^2$  and  $\sigma_X^2$ , and correlation  $\rho$  (the Pearson correlation coefficient). Here, Y represents a measure from a candidate test or method and X represents the corresponding measure from the gold standard test or method. The value of CCC is estimated from a sample by  $\hat{CCC}$  where the usual sample counterparts are substituted into the above formula to obtain

$$\hat{CCC} = \frac{2S_{YX}}{(\bar{Y} - \bar{X})^2 + S_Y^2 + S_X^2}$$

- **Intra Class Correlation Coefficient**

Description: Denote response  $j$  of subject  $i$  by  $Y_{ij}$ , where  $i = 1, 2, \dots, N$  and  $j = 1, 2, \dots, K$ . The model for this situation is,

$$Y_{ij} = \mu + a_i + e_{ij}$$

The intra-class correlation is then defined as

$$\rho = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2}$$

The hypothesis test is stated formally as,

$$H_0 : \rho = \rho_0$$

$$H_1 : \rho = \rho_1 > \rho_0$$

This hypothesis is tested from the data of a one-way analysis of variance table using F Test. The critical value for the test statistic is

$$C(F_{1-\frac{\alpha}{2}, df_1, df_2})$$

where,

$$C = 1 + \frac{\kappa \rho_0}{1 - \rho_0}$$

$$df_1 = N - 1$$

$$df_2 = N(K - 1)$$

- **Bland Altman Analysis**

Description: Bland and Altman introduced the Bland-Altman (B&A) plot to describe agreement between two quantitative measurements by constructing limits of agreement. These statistical limits are calculated by using the mean and the standard deviation (s) of the differences between two measurements. The resulting graph is a scatter plot XY, in which the Y axis shows the difference between the two paired measurements (A-B) and the X axis represents the average of these measures ((A+B)/2). Thus, the difference of the two paired measurements is plotted against the mean of the two measurements. The B&A graph plot simply represents every difference between two paired methods against the average of the measurement. Plotting difference against mean also allows studying any possible relationship between measurement error and the true value. The visual examination of the plot allows us to evaluate the global agreement between the two measurements. One can summarize the lack of agreement by estimating the mean difference and the standard deviation of the differences.

## 6. Test of Diagnostics

- Diagnostic Tests

- (a) **Sensitivity**

Description: Sensitivity measures the proportion of all patients with the condition (true positives + false negatives) who indeed have a positive test result (true positives). In other words, it is the ability of a test to yield a positive result for a subject that has that disease.

$$\text{Sensitivity} = TP/(TP+FN)$$

- (b) **Specificity**

Description: Specificity is the proportion of all patients without the disease and a negative test result (true negatives) of all those without the disease (true negatives + false positives). Specificity is how the test performs in people who are known to not have disease.

$$\text{Specificity} = TN/(TN + FP)$$

- (c) **Positive Likelihood Ratio**

Description: Probability that a positive test would be expected in a patient divided by the probability that a positive test would be expected in a patient without a disease.

$$\text{Positive Likelihood Ratio} = \text{Sensitivity}/(1 - \text{Specificity})$$

- (d) **Negative Likelihood Ratio**

Description: Probability of a patient testing negative who has a disease divided by the probability of a patient testing negative who does not have a disease.

$$\text{Negative Likelihood Ratio} = (1 - \text{Sensitivity})/\text{Specificity}$$

- (e) **Positive Predictive Value and Negative Predictive Value**

Description: PPV indicates the probability of having the disease after a positive test result, whereas the NPV is the probability of not having the disease after a negative test result.

Positive Predictive Value =  $TP/(TP + FP)$

Negative Predictive Value =  $TN/(TN + FN)$

(f) **Disease Prevalence**

Description: Prevalence is a measure of the burden of disease in a population in a given location and at a particular time, as represented in a count of the number of people affected.

(g) **Accuracy**

Description: Diagnostic accuracy expressed as the proportion of correctly classified study participants  $(TP + TN)$  among all study participants  $(TP + TN + FP + FN)$ .

Accuracy =  $(TP + TN)/(TP + FP + TN + FN)$

- **ROC Analysis**

Description: The Receiver operative curve (ROC) analysis is undertaken to evaluate the discriminatory power of a variable for an event (disease condition) in comparison to a control (normal). In ROC sensitivity is plotted against 1-Specificity corresponding to each value of study variables. The area under curve (AUC) is calculated to assess the prediction power of the variable under study between disease condition and control. The optimum cut-off value of the variable is identified and the sensitivity and specificity at that threshold is calculated. The odds ratio are calculated for variable associated with disease condition.

## 7. Test of Survival

- **Kaplan Meier Analysis**

Description: The Kaplan-Meier estimate involves estimation of probabilities of occurrence of event at a certain point of time. For each time point or interval, survival probability is calculated as the number of subjects surviving divided by the number of patients at risk. Total probability of survival till that time interval is calculated by multiplying all the probabilities of survival at all time intervals preceding that time. The graph of the survival probability with time  $t$  is defined as survival curve. Kaplan-Meier methods is used to compute the survival functions of the two alternative situations and to test the difference in survival patterns,

## 8. Test of Regression

- **Linear Regression**

Description: The Regression relationship is a relationship of a dependent variable  $Y$  with an independent variable  $x$ . In most cases a linear regression relationship is generally studied. The statistical relation between  $x$  and  $Y$  may be expressed as follows:

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

The above equation is the linear regression model that can be used to explain the relation between  $x$  and  $Y$ . The estimates of  $\beta_0$  and  $\beta_1$  are obtained using least square technique as under:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n y_i x_i - \frac{(\sum_{i=1}^n y_i)(\sum_{i=1}^n x_i)}{n}}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

The t test is used to conduct hypothesis tests on the regression coefficients obtained in simple linear regression. A statistic based on the t distribution is used to test the two-sided hypothesis that the true slope,  $\beta_1$ , equals some constant value,  $\beta_{1,0}$ . The statements for the hypothesis test are expressed as:

$$H_0 : \beta_1 = \beta_{1,0}$$

$$H_1 : \beta_1 \neq \beta_{1,0}$$

The test statistic used for this test is:

$$T_0 = \frac{\hat{\beta}_0 - \beta_{1,0}}{se(\hat{\beta}_1)}$$

where  $\hat{\beta}_1$  is the least square estimate of  $\beta_1$  and  $se(\hat{\beta}_1)$  can be calculated as follows:

$$se(\hat{\beta}_1) = \sqrt{\frac{\frac{\sum_{i=1}^n e_i^2}{n_2}}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

The test statistic, follows a t distribution with  $(n-2)$  degrees of freedom, where  $n$  is the total number of observations. The null hypothesis,  $H_0$ , is accepted if the calculated value of the test statistic is less than the table value from t distribution. Failure to reject  $H_0 : \beta_1 = 0$  implies that no linear relationship exists between  $x$  and  $Y$ .

- **Non-Linear Regression**

Description: Nonlinear regression is another regression technique in which a nonlinear mathematical model is used to describe the relationship between two variables. For example:

(a) Log linear

$$y = ax^b$$

(b) Quadratic

$$y = a + bx + cx^2$$

- **Logistic Regression** Description: In the usual regression analysis the dependent variable is continuous but Logistic model is used in situation when the dependent variable is dichotomous that is binary. Logistic regression analysis studies the association between a binary dependent variable and a set of independent (explanatory) variables using a logit model.

For a multiple variable situation, the model is -

$$P = \frac{e^{(\beta + \alpha_1 X_1 + \dots + \alpha_n X_n)}}{1 + e^{(\beta + \alpha_1 X_1 + \dots + \alpha_n X_n)}}$$

where,

$P$  = probability of an event.

$\beta_i$  = regression coefficients associated with the  $i$ -th explanatory variables.

In this analysis the reference group is constituted separately for each and every variable. The exponential of regression coefficient is referred to as odds ratio. The useful application of



logistic regression is in case control studies.

- **Conditional Logistic Regression**

Description: Conditional logistic regression is a specialized type of logistic regression usually employed when case subjects with a particular condition or attribute are each matched with control subjects. The most common design is 1:1 matching, however there could 1:k matching. The Conditional Logistic Regression Model in case of S matched sets and  $p$  independent variables, is given by -

$$\text{logit}(p) = \alpha_0 + \alpha_1 z_1 + \alpha_2 z_2 + \dots + \alpha_s z_s + \beta_1 x_1 + \dots + \beta_p x_p$$

where,

$\alpha$ 's = regression coefficients associated with the matched indicator variables.

$\beta$ 's = regression coefficients to be estimated each covariate adjusted for the others.

This analysis provides odds ratio of each covariates adjusted for matched variables.

- **Cox Proportional Hazard Model**

Description: Cox regression is a form of regression model where the outcome is hazard of developing the event outcome taking into account both time and several other factors,

$$h_i(t) = h_0(t) \exp(\beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_n x_{in})$$

where,

$h_i(t)$  = Hazard function for individual i.

$h_0(t)$  = Baseline hazard function and can take any form.

$x_{i1}, \dots, x_{in}$  = Covariates.

$\beta_1, \dots, \beta_n$  = regression coefficients estimated from the data.



One important assumption = the effect of co-variables does not change with time (proportional hazards).





## 4. Steps to run - Automation Pipeline

This section comprehensively describes steps that are required to operate the automated test selection workflow of NINS-STAT. Descriptions of each step has been provided pictorially for the researcher's ease.

### 1. Start the GUI.

- **NINS-STAT Scripts** - Browse to the folder directory into NINS-STAT main folder and type **start** on the command line. This step is applicable only if the user has chosen to download the .m files.
- **NINS-STAT executable files** - Double click on the icon to start the interface.

### 2. Once successfully initialized, the GUI, as depicted below, will show up. The GUI is primarily divided into 3 segments.

- **Initialize** : This segment is further divided into sub-sections which are as follows -
  - (a) **Data Import** - This section is meant for importing of excel sheet only.
  - (b) **Automation Pipeline** - This section is meant for initializing the operation of automation test selection workflow, which we will be dealing with in the subsequent steps.
  - (c) **Data Visualization** - This section is meant to initialize the operation of the data visualization and plotting tools of NINS-STAT. The steps of this section will be dealt in the following chapters.
  - (d) **Descriptive Analysis** - This section is meant to initialize the operation of the data descriptive tools of NINS-STAT. The steps of tis section will also be dealt in the following chapters.
- **Objectives** : This segment displays the objectives for the aforementioned sections for the user to select.
- **Analysis** : This segment displays the additional limited inputs congruent to the previous steps that are required for the final steps required before the completion of the workflow procedure.

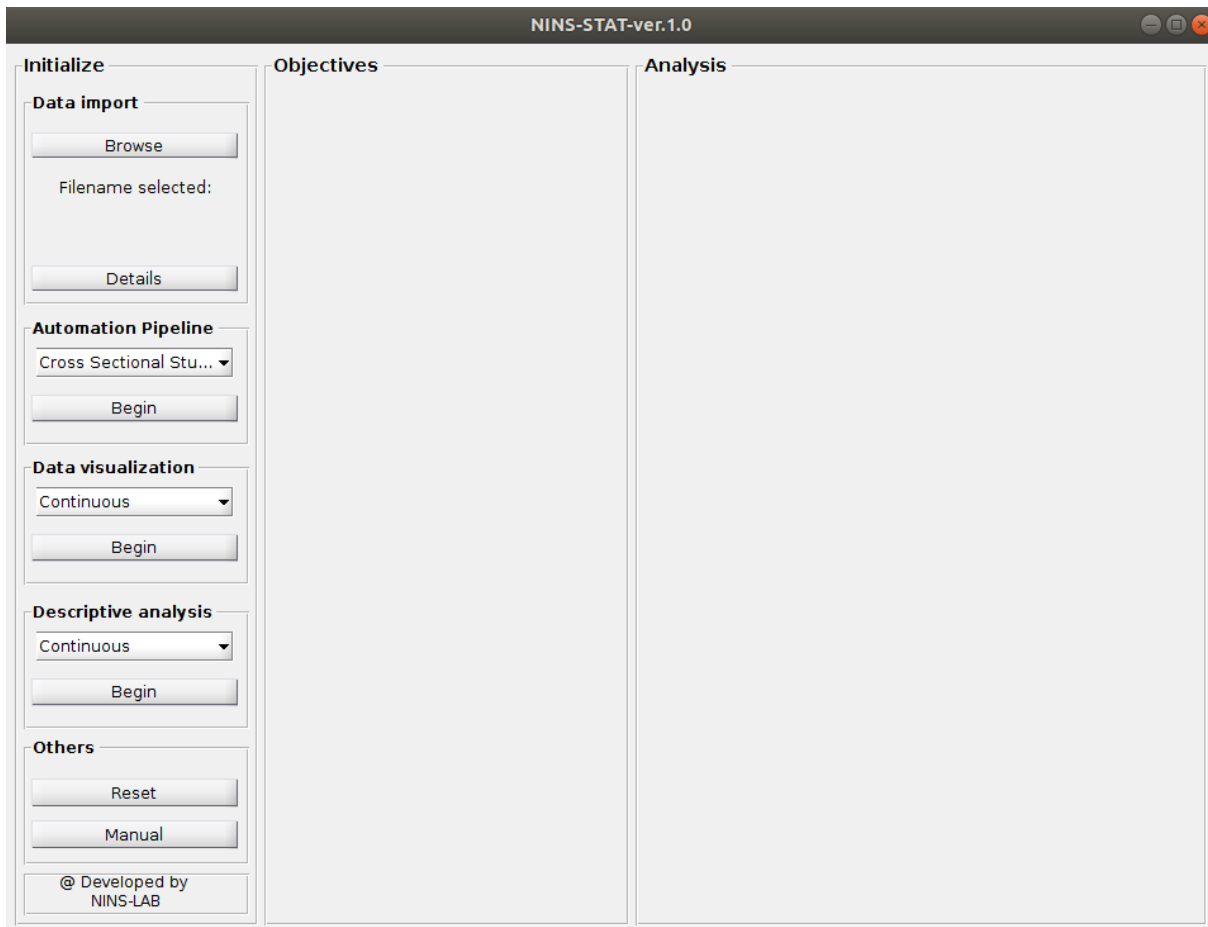


Figure 4.1: Initial GUI Interface

3. Browse for the Excel Sheet containing the data. The data should be in the prescribed format.

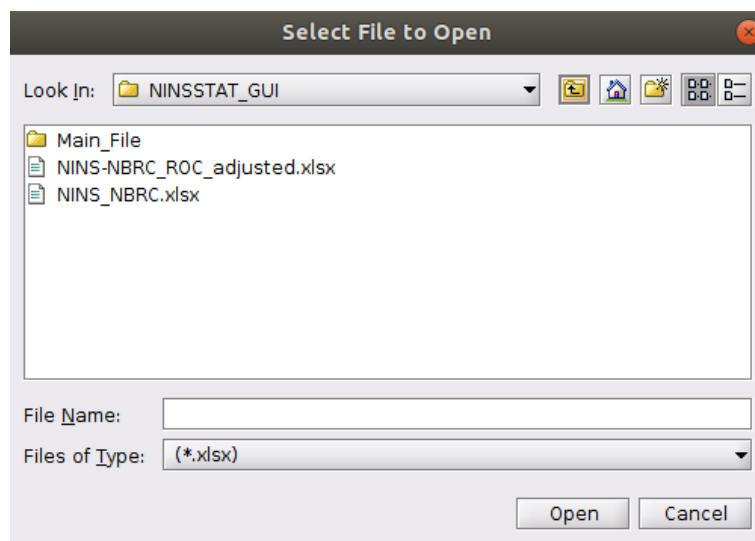


Figure 4.2: Browse Tab

4. Upload the data into the interface.

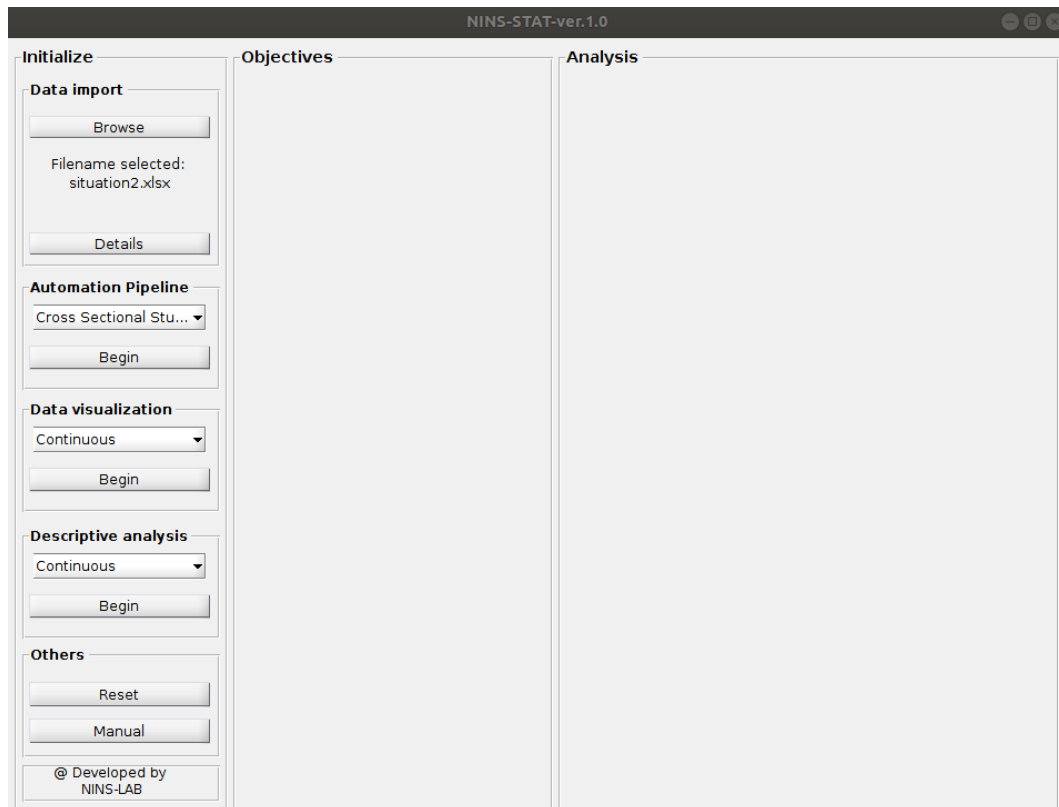


Figure 4.3: Upload Data

5. The user upon requirement can click the **Details** to see the details of the excel sheet. Information such as the filename, sample size as well as the variables will be shown on a separate window.

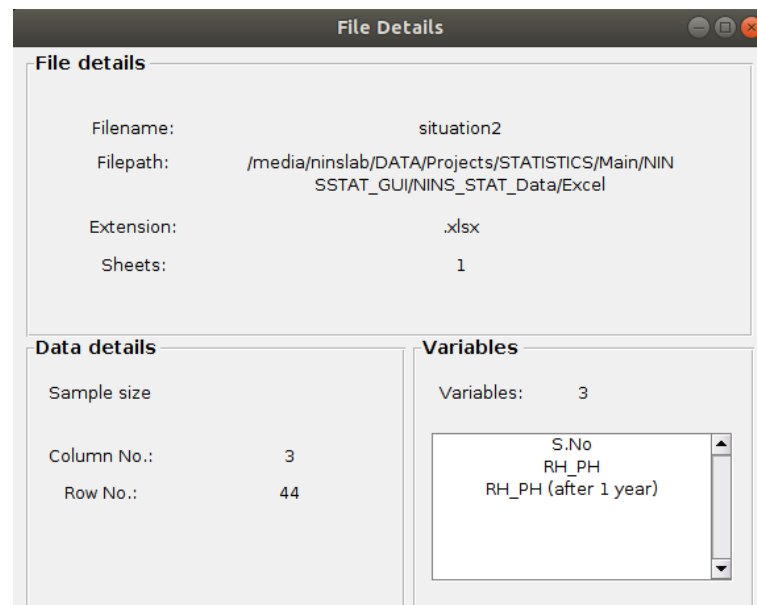


Figure 4.4: Data details

6. Once the data has been imported into the interface. The user needs to select the study design from the given list as shown. This will load a list of objectives that are appropriate for the aforementioned

selected study design.

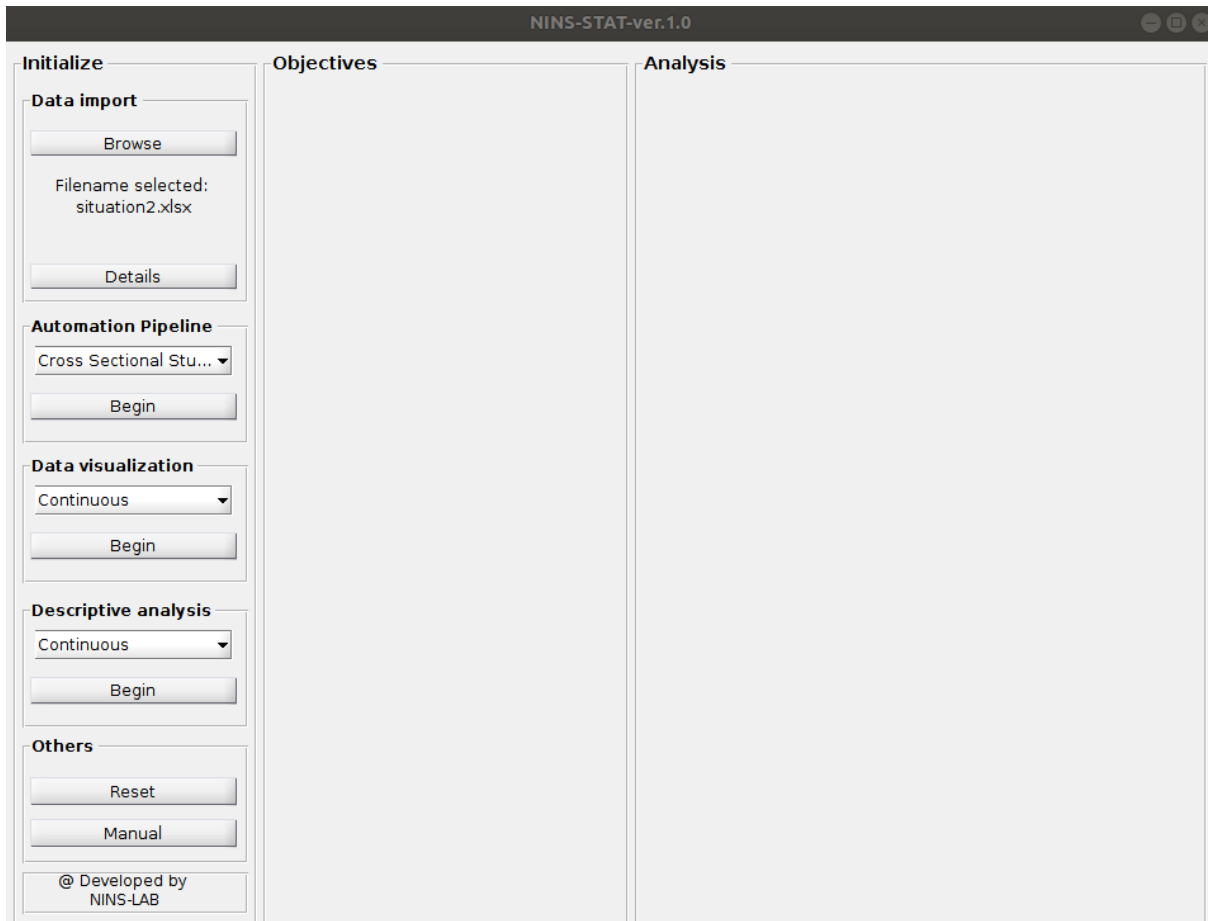


Figure 4.5: Initial GUI

7. Once the parameters are loaded the user has to select **Begin**.

The list of objectives will be divided into **Recommended list** and **Optional list**.

- **Recommended List** - This list of objectives are suggested by NINS-STAT as the most appropriate objectives for a particular study design.
- **Optional List** - This list of objectives are not recommended by NINS-STAT but are available for use should the user want to use them apart from the recommended list.

NINS-STAT-ver.1.0

**Initialize**

**Data import**

Browse

Filename selected:  
situation2.xlsx

Details

**Automation Pipeline**

Cross Sectional Stu... ▼

Begin

**Data visualization**

Continuous ▼

Begin

**Descriptive analysis**

Continuous ▼

Begin

**Others**

Reset

Manual

@ Developed by  
NINS-LAB

**Objectives**

Recommended

☒ Proportions

☐ Mean

☐ Association

☐ Agreement

☐ Diagnostic

☐ Regression

Optional

☐ Incidence

☐ Relative Risk

☐ Odds Ratio

☐ Survival

Proceed

**Analysis**

Figure 4.6: Objective List

8. Select the appropriate objective for the relevant study as per requirement.
9. Select **Proceed**. This will enable limited inputs fields in the **Analysis** segment which are required by NINS-STAT to perform automated test selection.
10. Select the relevant options in accordance to the user study objectives.
11. Click **Proceed** after making the appropriate selection at each step till the data selection phase which has been described in the next step.

**NINS-STAT-ver.1.0**

**Initialize**

**Data import**

Browse

Filename selected: situation2.xlsx

Details

**Automation Pipeline**

Cross Sectional Stu...

Begin

**Data visualization**

Continuous

Begin

**Descriptive analysis**

Continuous

Begin

**Others**

Reset

Manual

@ Developed by NINS-LAB

**Objectives**

Recommended

☐ Proportions

☒ Mean

☐ Association

☐ Agreement

☐ Diagnostic

☐ Regression

Optional

☐ Incidence

☐ Relative Risk

☐ Odds Ratio

☐ Survival

Proceed

**Analysis**

**A**

☒ Population variance is unknown (default)

☐ Population variance is known

Proceed

**B**

☐ One Sample Analysis

☒ Two Sample Analysis

☐ > 2 Sample Analysis

Proceed

**C**

☒ Paired Sample

☐ Independent Sample

Proceed

Continuous Data 1 Upload Data

Continuous Data 2 Upload Data

Run

Figure 4.7: Analysis Results

12. Upload data as indicated on the left side of the buttons.
13. Click **Upload Data** to select the columns for data analysis from the excel sheet.
14. The Uploaded Excel Sheet will appear on the screen.

	1	2	3
1	S.No	RH_PH	RH_PH (a...
2	1	6.9381	7.0381
3	2	6.9712	7.9712
4	3	6.9774	5.9774
5	4	6.9791	6.9991
6	5	6.9963	7.3856
7	6	7.0107	7.5107
8	7	7.0183	7.9183
9	8	7.0208	7.1208
10	9	7.0215	6.0205
11	10	7.0254	6.0254
12	11	7.0351	7.9351
13	12	7.0362	7.8362
14	13	7.0439	7.0499

Figure 4.8: Upload Complete



15. Select the column required for the analysis.

**R** Selecting one cell of a column will lead to the selection of the entire column.

16. After selection click **DONE** to close the data selection window. The label of the column selected will show up on the edit box next to the label as shown below.

The screenshot shows the NINS-STAT-ver.1.0 software interface with three main panels: Initialize, Objectives, and Analysis.

- Initialize Panel:**
  - Data import:** A 'Browse' button and 'Filename selected: situation2.xlsx'.
  - Automation Pipeline:** A dropdown menu set to 'Cross Sectional Stu...' and a 'Begin' button.
  - Data visualization:** A dropdown menu set to 'Continuous' and a 'Begin' button.
  - Descriptive analysis:** A dropdown menu set to 'Continuous' and a 'Begin' button.
  - Others:** 'Reset' and 'Manual' buttons.
  - Footer: '@ Developed by NINS-LAB'.
- Objectives Panel:**
  - Recommended:** Radio buttons for 'Proportions', 'Mean' (selected), 'Association', 'Agreement', 'Diagnostic', and 'Regression'.
  - Optional:** Radio buttons for 'Incidence', 'Relative Risk', 'Odds Ratio', and 'Survival'.
  - A 'Proceed' button at the bottom.
- Analysis Panel:**
  - A:** Radio buttons for 'Population variance is unknown (default)' (selected) and 'Population variance is known'. A 'Proceed' button.
  - B:** Radio buttons for 'One Sample Analysis', 'Two Sample Analysis' (selected), and '> 2 Sample Analysis'. A 'Proceed' button.
  - C:** Radio buttons for 'Paired Sample' (selected) and 'Independent Sample'. A 'Proceed' button.
  - Data Upload:** Two rows for 'Continuous Data 1' and 'Continuous Data 2'. Each row has an 'Upload Data' button and a text box containing 'RH\_PH'.
  - A 'Run' button at the bottom.

Figure 4.9: Analysis Parameters

**R** For statistical analysis that require multiple inputs, the user can press **Ctrl** and then select multiple number of columns.

17. To begin the automated test selection, press **Run**. The final results will be shown on the command line as shown below.

```
Data = Not Normalized
Test -----> Wilcoxon Signed Rank Test
-----
Results :
```

P_Value	H_Value	Signed_Rank
0.023558	true	270.5

Figure 4.10: Results



## 5. Steps to run - Descriptive Analysis

This section deals with the operation of the descriptive data analysis tools. It helps the user to perform very basic descriptive analysis for data with and without a grouping variables. The following steps, with the help of pictorial description, acts as a guide to perform the analysis using NINS-STAT.

**R** This sub-section is independent of the automated test selection workflow.

1. Start the GUI.
2. Then the user has to browse for the excel sheet containing the raw data.
3. The user has to also select the type of variable that is to be analyzed. It is divided into **Continuous** and **Categorical**. The subsequent steps will show options according to the selection made in this step.

**Continuous** - indicates the variable type to be analyzed as continuous in nature. The tools available are :

- Number of Observations
- Mean
- Median
- Mode
- Standard Deviation
- Coefficient of Variation
- Range
- Inter-Quartile Range (IQR)

**Categorical** - indicates the variable type to be analyzed as categorical in nature. The tools available are :

- Number of Observations
  - Frequency
  - Percentage
4. The user now has to select the appropriate options (single or multiple) as shown below.

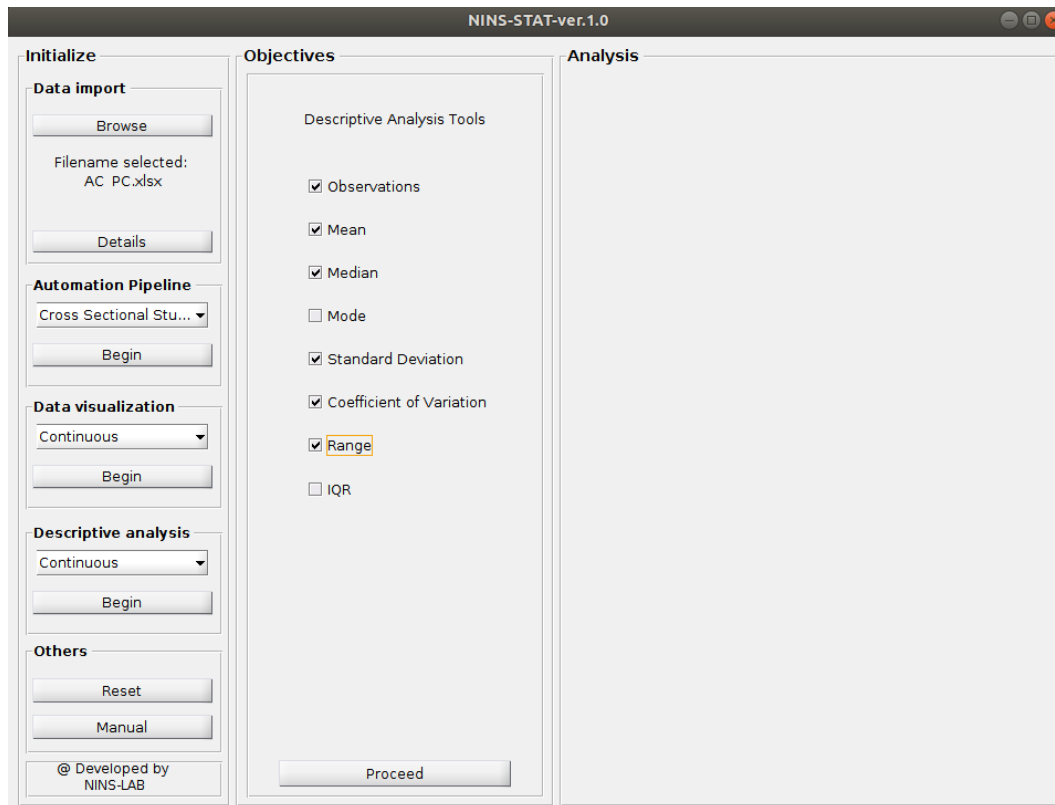


Figure 5.1: Descriptive analysis GUI

5. Select the appropriate column required from the imported dataset for analysis.

**R** Please note number of observations selected can be multiple. Also the **Grouping variable** is optional. It is to be added only if the user has a grouping variable.


6. Select **Proceed**.

Group - 0						
LABEL	OBSERVATIONS	MEAN	MEDIAN	STDDEV	CV	RANGE
'ACC GSH Conc.'	27	1.9502	1.9152	0.42239	0.21658	1.4748
'PCC GSH Conc.'	27	2.3675	2.3146	0.40498	0.17106	1.9049
'CINGULATE GSH Conc.'	27	4.3222	4.2247	0.70526	0.16317	3.1237
Group - 1						
LABEL	OBSERVATIONS	MEAN	MEDIAN	STDDEV	CV	RANGE
'ACC GSH Conc.'	19	1.6225	1.5542	0.38653	0.23824	1.6591
'PCC GSH Conc.'	19	1.7708	1.7616	0.34447	0.19452	1.2904
'CINGULATE GSH Conc.'	19	3.3019	3.18	0.56583	0.17136	2.0227
Group - 2						
LABEL	OBSERVATIONS	MEAN	MEDIAN	STDDEV	CV	RANGE
'ACC GSH Conc.'	18	1.5625	1.631	0.34075	0.21808	1.3007
'PCC GSH Conc.'	18	1.7795	1.7385	0.43616	0.2451	1.3877
'CINGULATE GSH Conc.'	18	3.3849	3.3797	0.66096	0.19527	2.033

Figure 5.2: Descriptive analysis results with Grouping variable (Group 0, Group 1, Group 2)

## 6. Steps to run - Data Visualization

This section deals with the operation of the data visualization and plotting tools. It helps the user to perform very basic plotting and visualization for data. The following steps, with the help of pictorial description, acts as a guide to perform the visualization using NINS-STAT.

 This sub-section is independent of the automated test selection workflow.

1. Start the GUI.
2. Then the user has to browse for the excel sheet containing the raw data.
3. The user has to also select the type of variable that is to be analyzed. It is divided into **Continuous** and **Categorical**. The subsequent steps will show options according to the selection made in this step.

**Continuous** - indicates the variable type to be analyzed as continuous in nature. The tools available are :

- Histogram
- Standard plot
- Scatter plot

**Categorical** - indicates the variable type to be analyzed as categorical in nature. The tools available are :

- Bar plot
- Line plot
- Stack Chart
- Pie Chart
- Box and Whiskers plot

4. The user now has to select the appropriate options (single or multiple) as shown below.

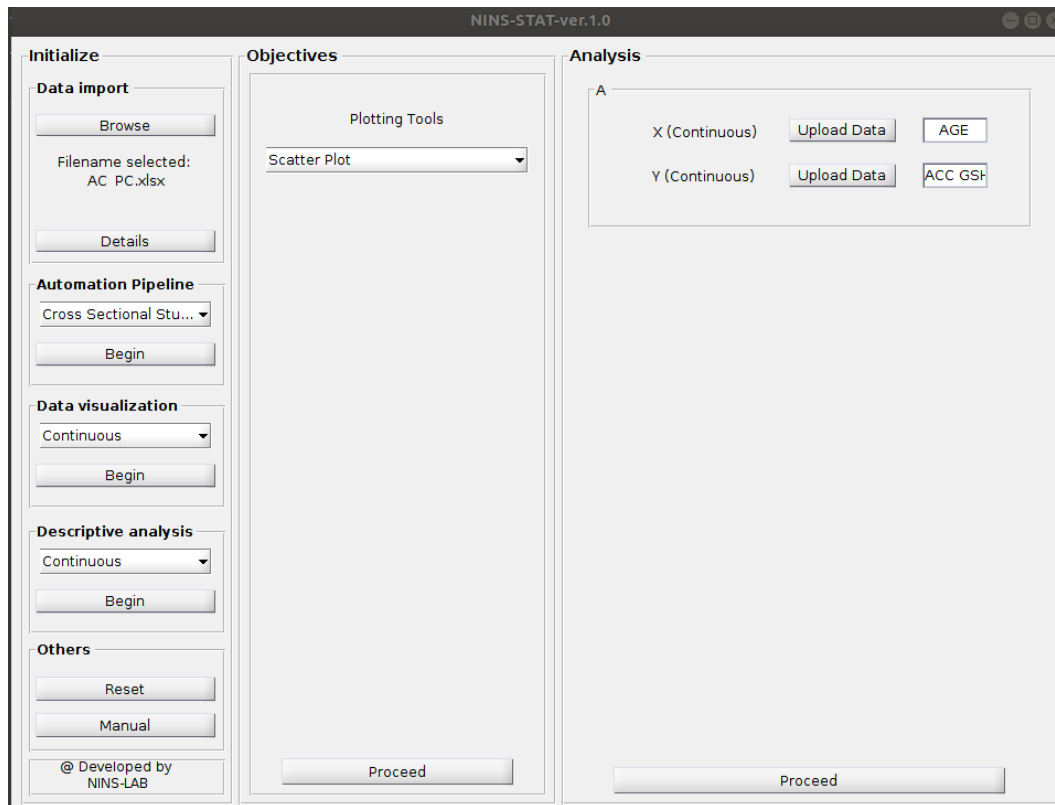


Figure 6.1: Visualization analysis GUI

5. Select the appropriate column required from the imported dataset for analysis.

**R** Please note number of observations selected can be multiple.

6. Select **Proceed**.

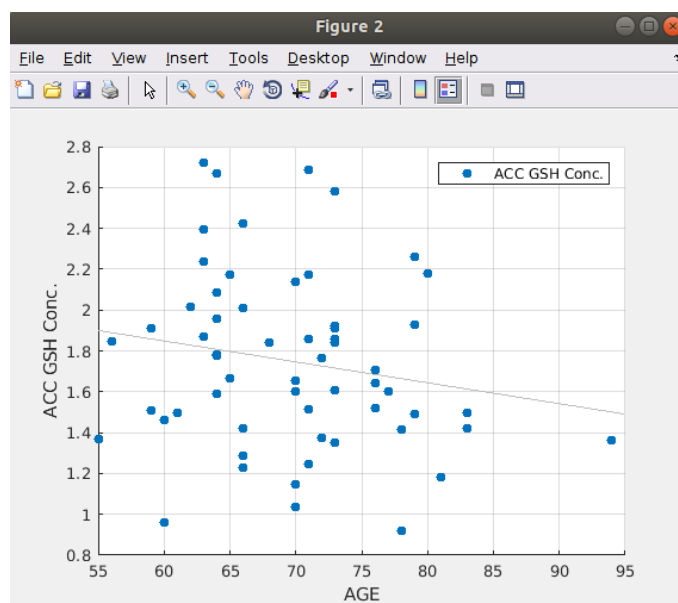


Figure 6.2: Scatter plot