A Framework for Statistical Simulation of Physiological Responses (SSPR)

Pavitra R. Gautam *pavitragautam@dipas.drdo.in*
Biostatistics Department
DIPAS,DRDO
Delhi-110054 ,India

Dr YK Sharma *yksharma_ashu@yahoo.com*
Biostatistics Department
DIPAS,DRDO
Delhi-110054 ,India

Dr Shashi Bala Singh *director@dipas.drdo.in*
Director
DIPAS,DRDO
Delhi-110054 ,India

Abstract

The problem of variable selection from a large number of variables to predict certain important dependent variables has been of interest to both applied statisticians and other researchers in applied physiology. For this purpose, various statistical techniques have been developed. This framework embedded various statistical techniques of sampling and resampling and help in Statistical Simulation for Physiological Responses under different Environmental condition. This framework will facilitates the researchers to work on simulated population. It will also solve the problem of small sample size by providing resampling module. The generation of simulated population and other statistical calculations are based on the inputs provided by the user as mean vector and covariance matrix and the data. This framework is developed in a way that it can work for the original data as well as for simulated data generated by the software. **Approach:** The mean vector and covariance matrix are sufficient statistics when the underlying distribution is multivariate normal. This framework uses these two inputs and is able to generate simulated multivariate normal population for any number of variables. The software changes the manual operation into a computer-based system to automate the study, provide efficiency, accuracy, timelessness, and economy. **Result:** A complete framework that can statistically simulate any type and any number of responses or variables. Simulated data when analyzed using statistical techniques; results of such analysis will be the same as that using the original data. The system provides solution for missing data on some variables. **Conclusion:** The proposed system makes it possible to carry out the physiological studies and statistical calculations even if the actual data is not present and also in the case when sample size is small.

Keywords: SSPR(Statistical Simulation of Physiological Responses),Population Generation (PG), Sample Selection (SS), Simple Random Sampling With Replacement (SRSWR), Simple Random Sampling Without Replacement (SRSWOR) , Probability proportional to the Size With Replacement (PPSWR), Probability proportional to the Size Without Replacement(PPSWOR), Data Flow Diagram(DFD)

1. INTRODUCTION

The process of designing a model of a real system and conducting experiments with this model for the purpose either of understanding the behavior of the system or of evaluating various strategies (within the limits imposed by a criterion or set of criteria) for the operation of the system."

- R.E Sahnnon

The need for the statistical information is endless in the modern society for planning development and growth. One of the most important modes of data collection for satisfying such needs is sample survey, i.e. a partial investigation of finite population. Simulation is a process of designing a model of a real system & conducting experiments with this model on the computer for understating the behavior of the system or evaluating various strategies for operation of the system. The model can be defined as a presentation of a real system which can be controlled by various parameters.

The advanced computer programs can simulate weather conditions, chemical reactions, atomic reactions, even biological processes. In theory, any phenomena that can be reduced to mathematical data and equations can be simulated on a computer. In practice, however, simulation is extremely difficult because most natural phenomena are subject to an almost infinite number of influences. One of the tricks to developing useful simulations, therefore, is to determine which the most important factors are. There are various studies for building mathematical models for physiological processes. Some authors proposed the framework for modeling and simulation of physiological models to improve the modeling process. Phy-SIM is a modeling, integration and simulation environment for physiological processes from tissue level models to organ-organism levels [1].In this author proposed a layered approach modular design principles. A framework for cardiovascular system based on ontology was described by Daniel in 2006[10].But no attempts have been made to use the technique of simulation on physiological systems/ responses in the area of heat physiology and cold physiology etc. Therefore, in the present study, statistical techniques will be used to simulate some of these important physiological functions at sea level and high altitude. SSPR provide solution to overcome the problem of small sample size by embedding the resampling techniques. The Simulated data generated by SSPR can be used further for Meta analysis, statistical modeling and it can also be used for the development of physiological index. There are many organizations public and private that are working in the field of life sciences. They collect the data either by experiments or by other techniques as interview, questionnaire, and measurements by instruments or from previous studies etc. Many times in human studies problem of non-response encountered for example, when administering a survey people may answer some questions and not others due to this reason it is not possible to collect data that is complete. There are many reasons for missing values as missing by design; or not asked or not applicable. This missing data causes a problem for researchers specially when using structural equation modeling (SEM) techniques for data analysis. Because SEM and multivariate methods require complete data [2][3]. All over the world there are several organizations, working on sensitive information regarding the entities or subjects. These organizations use several different approaches to prevent disclosure of the sensitive information. These include restricting access to specific individuals, using database control techniques, masking the data prior to providing access, releasing interval data rather than individual data points, etc. [4]. In such situations use of simulated data rather than the original data is safe, easy and time saving. One can easily calculate mean and covariance matrix of all the parameters by avoiding missing values. When literature survey of previous studies is done only mean and covariance matrix of the parameters is available but not the actual data. Sometimes it is not possible for researchers to work on the entire population in that case they have to work on the subset of individuals from within a population to yield some knowledge about the whole population. The three main advantages of sampling are that the cost is lower, data collection is faster, and since the data set is smaller it is possible to ensure homogeneity and to improve the accuracy and quality of the data.[3][5]. Thus, there is a need to develop software that can help the individual to generate the simulated population of desired size by supplying mean and co-variance matrix as input and also be able to select random samples by different sampling and resampling techniques and further for using that sample for some basic statistical calculation as mean, median, mode standard deviation etc.

The whole system is divided into four major modules as shown below

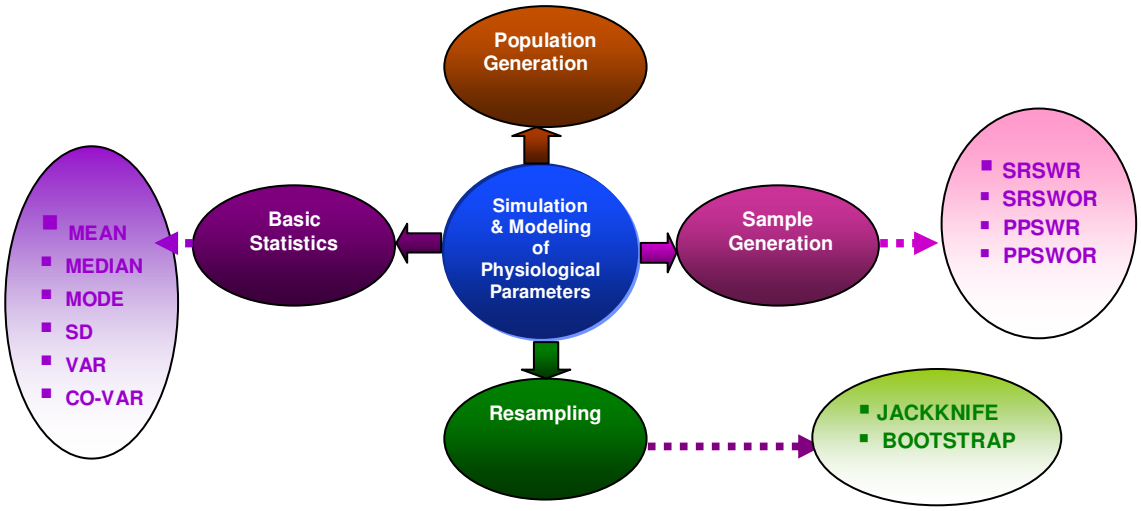

FIGURE 1: Statistical Simulation of Physiological Responses (SSPR) Modules.

The population generation module will help the researches who want to continue old studies but does not having the actual data by giving them the simulated population , This will also help the researchers who does not want to reveal sensitive information by generating the simulated population based on mean and variance covariance matrix. The Sample Selection module will help in sample selection by using various sampling techniques. The Resampling module will help the researches who are not able to continue their research due to small size of samples. The Basic statistics module will facilitates basic statistical calculations with graphical output as well.

2. MATERIALS AND METHODS

Before designing the SSPR, necessary objectives of the system were established. The objectives were created after the detailed analysis of organization work, limitations and concerns in the existing manual system. The various necessary details about the population generation, sample selection techniques and resampling techniques also gathered from the concerned authorities and users. This helped us to plan an effective SSMPP system.

System Analysis

System Analysis by definition is a process of systematic investigation for the purpose of gathering data, interpreting the facts, diagnosing the problem and using this information to either build a completely new system or to recommend the improvement to the existing system. As the new system is going to be developed, the requirement analysis a preliminary investigation, feasibility study for the required system was done. All the available resources with respect to software requirement were checked, these all step helped us in making the data flow diagram (DFD) for the SSPR as shown in the Figure: 2.

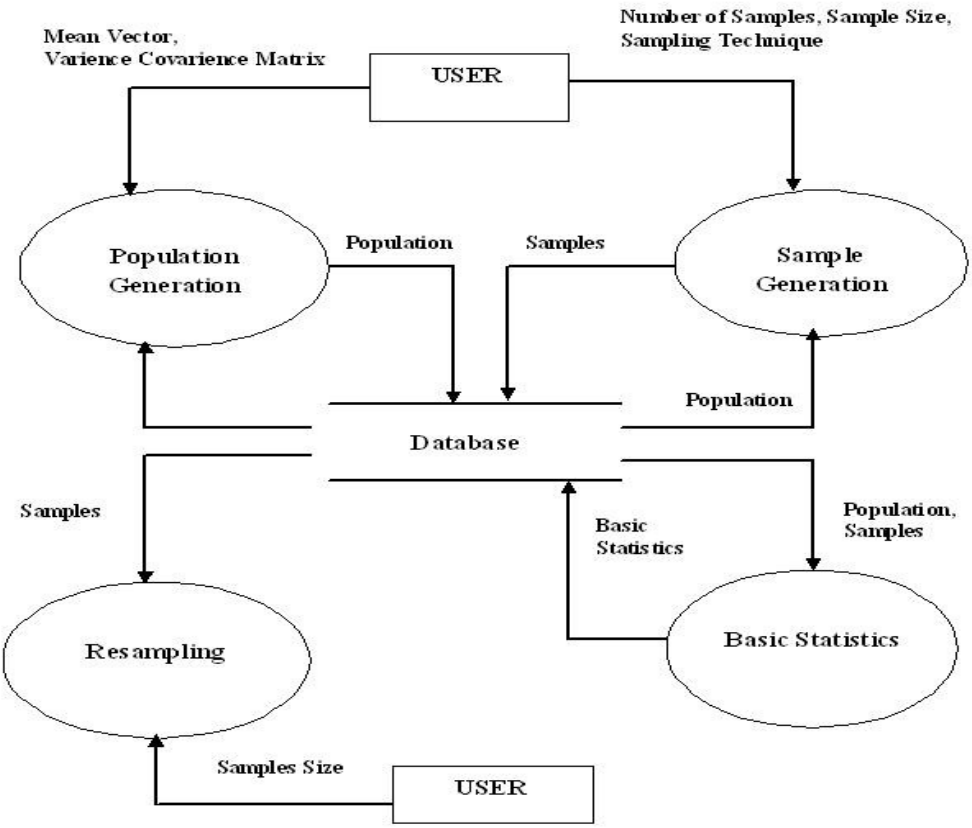

FIGURE 2: DFD for SSPR

The above diagram shows how the data flows from one process to others and from databases to the others process and vice –versa.

In the real world, most data collection schemes or designed experiments will result in multivariate data. When multivariate data are analyzed, the multivariate normal model is the most commonly used model. Many statistical techniques focus on just one or two variables but sometimes there is a need to analyze more than two variables. Multivariate analysis (MVA) techniques allow more than two variables to be analysed at once. The multivariate normal distribution is the generalized form of the one-dimensional i.e. univariate normal distribution to higher dimensions. The multivariate normal distribution of a *k*-dimensional random vector $X = [X_1, X_2, \ldots, X_k]$ can be written in the following notation[6]:

$$X \sim \mathcal{N}(\mu, \Sigma),$$

or to make it explicitly known that X is *k*-dimensional,

$$X \sim \mathcal{N}_k(\mu, \Sigma).$$

with *k*-dimensional mean vector

$$\mu = [\mathrm{E}[X_1], \mathrm{E}[X_2], \ldots, \mathrm{E}[X_k]]$$

and *k x k* covariance matrix

$$\Sigma = [\mathrm{Cov}[X_i, X_j]], i = 1, 2, \ldots, k; j = 1, 2, \ldots, k$$

Sampling refers to the process of choosing a sample of elements from a total population of elements. Simple random sampling refers to a sampling method that has the following properties: The population consists of *N* objects. The sample consists of *n* objects. All possible samples of *n* objects are equally likely to occur. The main benefit of simple random sampling is that it guarantees that the sample chosen is representative of the population. This ensures that the statistical conclusions will be valid. Sampling can be done with replacement or without replacement [7][8]. **Sampling with replacement** is accomplished by "tossing" population members back into the mix after they have been selected. In this way, all *N* members of the population have an equal chance of being selected *at each draw*. In other words -Sampling is called with replacement when a unit selected at random from the population is returned to the population and then a second element is selected at random. Whenever a unit is selected, the population contains all the same units. A unit may be selected more than once. There is no change in the size of the population at any stage. Let us assume that a sample of any size can be selected from the given population of any size. This is only a theoretical concept and in practical situations the sample is not selected by using this scheme of selection. Suppose the population size N=5 and sample size n=2, and sampling is done with replacement. Out of 5 elements, the first element can be selected in 5 ways. The selected unit is returned to the main lot and now the second unit can also be selected in 5 ways. Thus in total there are 5X5=25 samples or pairs which are possible. In contrast, **sampling without replacement** is done so that once a population member has been drawn; this person is removed from further sampling. Thus, once a population member has been drawn, their subsequent probability of selection is zero and the probability that someone else is selected goes up a little. In other words -Sampling is called without replacement when a unit is selected at random from the population and it is not returned to the main lot. First unit is selected out of a population of size N and the second unit is selected out of the remaining population of N-1 units and so on. Thus the size of the population goes on decreasing as the sample size n increases. The sample size n cannot exceed the population size N. The unit once selected for a sample cannot be repeated in the same sample. Thus all the units of the sample are distinct from one another. In simple words , when a population element can be selected more than one time, the sampling is known as sampling with replacement and when a population element can be selected only one time then the sampling is known as sampling without replacement**.** In sampling with replacement the two sample values are independent. Practically, this means that what we get on the first one doesn't affect what we get on the second. Mathematically, this means that the covariance between the two is zero. In sampling without replacement the case is just reverse of this, the two sample values aren't independent. In this sampling technique what we got on the for the first one affects what we can get for the second one and the covariance between the two isn't zero.

Probability Proportional to Size

Probability proportional to size (PPS) is a sampling technique mostly used with surveys where the probability of selection of a sampling unit (city, any village, district etc) is proportional to the size of its population. It gives a probability (i.e., random, representative) sample[5]. It is most useful when the sampling units vary considerably in size because it assures that those in larger sites have the same probability of getting into the sample as those in smaller sites, and vice verse. This method also facilitates planning for field work because a pre-determined number of respondents is interviewed in each unit selected, and staff can be allocated accordingly [8].

Sometime the situation occur when the population distribution is unknown and the sample size is small, in that case resampling plays a very important role to continue the study. **Resampling:** Resampling means that inference is based upon repeated sampling within the same sample.For

more than a century the inherent difficulty of formula-based inferential statistics has baffled scientists, induced errors in research, and caused million of students to hate the subject. In place of the formidable formulas and mysterious tables of parametric and non-parametric tests based on complicated mathematics and arcane approximations, the basic resampling tools are simulations, created especially for the task at hand by practitioners who completely understand what they are doing and why they are doing it. Resampling lets you analyze most sorts of data, even those that cannot be analyzed with formulas [9]. **Bootstrapping** is proposed by Bradly Efron in 1979. It is a statistical method for estimating the confidence interval of a parameter with or without assumption on data distribution. It also estimates the bias of an estimator. It may also be used for constructing hypothesis tests. It is often used as a robust alternative to inference based on parametric assumptions when those assumptions are in doubt, or where parametric inference is impossible or requires very complicated formulas for the calculation of standard errors[9]. **Jackknifing** is proposed by Quenouille in 1949 and in 1958 tukey named it as Jackknife .It is similar to bootstrapping, is used in statistical inference to estimate the bias and standard error (variance) of a statistic, when a random sample of observations is used to calculate it. The basic idea behind the jackknife variance estimator lies in systematically recomputing the statistic estimate leaving out one or more observations at a time from the sample set. From this new set of replicates of the statistic, an estimate for the bias and an estimate for the variance of the statistic can be calculated [10]. It is mainly proposed to reduce the bias of an estimator. It is only adaptive for estimators which are smooth function of the observation.

3. DISCUSSION

3.1 MODULE 1 : POPULATION GENERATION (PG)

PG is the module of the system SSPR, this takes mean, variance covariance matrix and population size as inputs and generates the simulated population of the size given by the user. The output of this module is the desired size simulated population matrix.

The function used to generate the Multivariate normal random numbers is mvnrnd function of MATLAB

Syntax

R = mvnrnd(MU,SIGMA)

Description

R = mvnrnd(MU,SIGMA) returns an n-by-d matrix R of random vectors chosen from the multivariate normal distribution with mean MU, and covariance SIGMA. MU is an n-by-d matrix, and mvnrnd generates each row of R using
the corresponding row of mu. SIGMA is a d-by-d symmetric positive semi-definite matrix, or a d-by-d-by-n array. If SIGMA is an array, mvnrnd generates
each row of R using the corresponding page of SIGMA, i.e., mvnrnd computes R(i,:) using MU(i,:) and SIGMA(:,:,i). If MU is a 1-by-d vector, mvnrnd replicates it to match the trailing dimension of SIGMA.[11]

Step1: Click on the button read mean Vector
Step2: Click on the button read Cov- Var Matrix
Step3: Enter the size of population you want to generate.
Step 4: Click on OK button.

Population of entered size get generated and saved in excel file name pop.xls in the MyOutput Folder (automatically get created) of C drive.

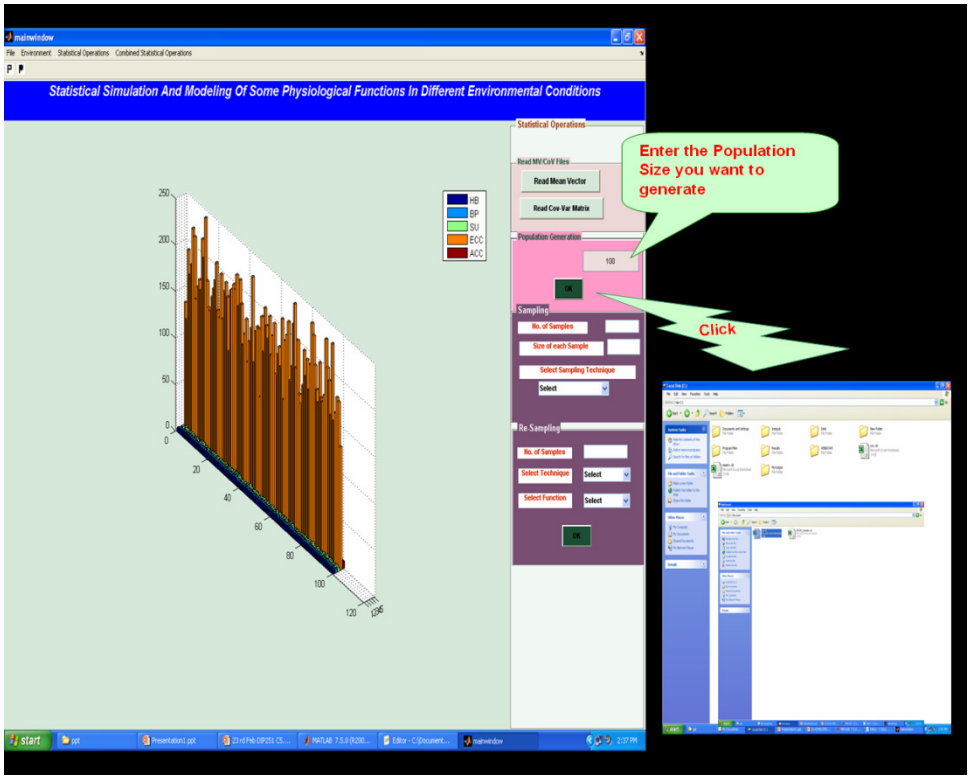

FIGURE 3: Population Generation Window

3.2 MODULE 2: SAMPLE SELECTION (SS)

Sample Selection : In this module the population generated by the module PG is sampled by the different sampling techniques selected by the user. Here the samples can be generated by Simple Random Sampling with and without replacement (SRSWR & SRSWOR) technique and also by Probability proportional to the size with and without replacement (PPSWR & PPSWOR) technique. In this module sample size, number of samples and sampling technique are the input and samples which get generated by this are the output.

Syntax

y = randsample(population,size,replace)

y = randsample(population,size,replace) returns a sample of the given size from the population . If replace is true then the samples will be taken with replacement or without replacement if replace is false [11].

Pre Requirement: Simulated population should exist in the file for example pop.xls, from which samples has to be selected.

Step 1: Enter the number of samples you want to generate (Ex 5,10,100 etc)
Step 2: Enter the size of each samples.
Step 3: Select one of the Sampling Technique given below

- SRSWR
- SRSWOR
- PPSWR
- PPSWOR

The samples get selected from the given population (SRSWOR.xls in case of SRSWOR sampling technique) and saved in the same directory as for population.

3.3 MODULE 3: BASIC SATISTICS

When performing statistical analysis on a set of data, the mean, median, mode, and standard deviation are all helpful values to calculate. The mean, median and mode are all estimates of where the "middle" of a set of data is. These values are useful when creating groups or bins to organize larger sets of data**.**In SSMPP the Basic Statistics module provide the basic statistics calculations like mean, SD, median, mode of the given population or samples. The methods which have used in software are as mean(A), median(A),mode(X),std(tsobj),var(X) etc.

- **Mode** - The mode of a distribution is simply defined as the most frequent or common score in the distribution. The mode is the point or value of *X* that corresponds to the highest point on the distribution. For this mode function of matlab is used.
- **Median**- The median is the score that divides the distribution into halves; half of the scores are above the median and half are below it when the data are arranged in numerical order. M = median(A) to calculate the median of the given array or matrix.
- **Mean** - The mean is the most common measure of central tendency and the one that can be mathematically manipulated. It is defined as the average of a distribution is equal to the $\sum X / N$. tsmean = mean(tsobj) computes the arithmetic mean of all data in all series in tsobj and returns it in tsmean.
- **Variance/Co-varience** - The variance is a measure based on the deviations of individual scores from the mean. The two functions used to calculate the varience and co-varience of the variables are var and cov.
- **Standard deviation** (SD)- The standard deviation (*s* or σ) is defined as the positive square root of the variance. The variance is a measure in squared units and has little meaning with respect to the data. Thus, the standard deviation is a measure of variability expressed in the same units as the data. The method used for it is std method.

Step 1: Go to the menu bar click on Statistical operations.
Step2: Select the menu item basic statistical operations. A new window will come out as shown below.
Step 3: Click on the operation you want to perform as Mean, Median Mode, SD, Covariance. A pop up window for file selection will come; select the file on which you want to perform basic statistical calculation.

This will provide the calculated values on the left side and the respective graphical output in the space provided for the graph. User can save the generated graph and the result in the excel sheet.

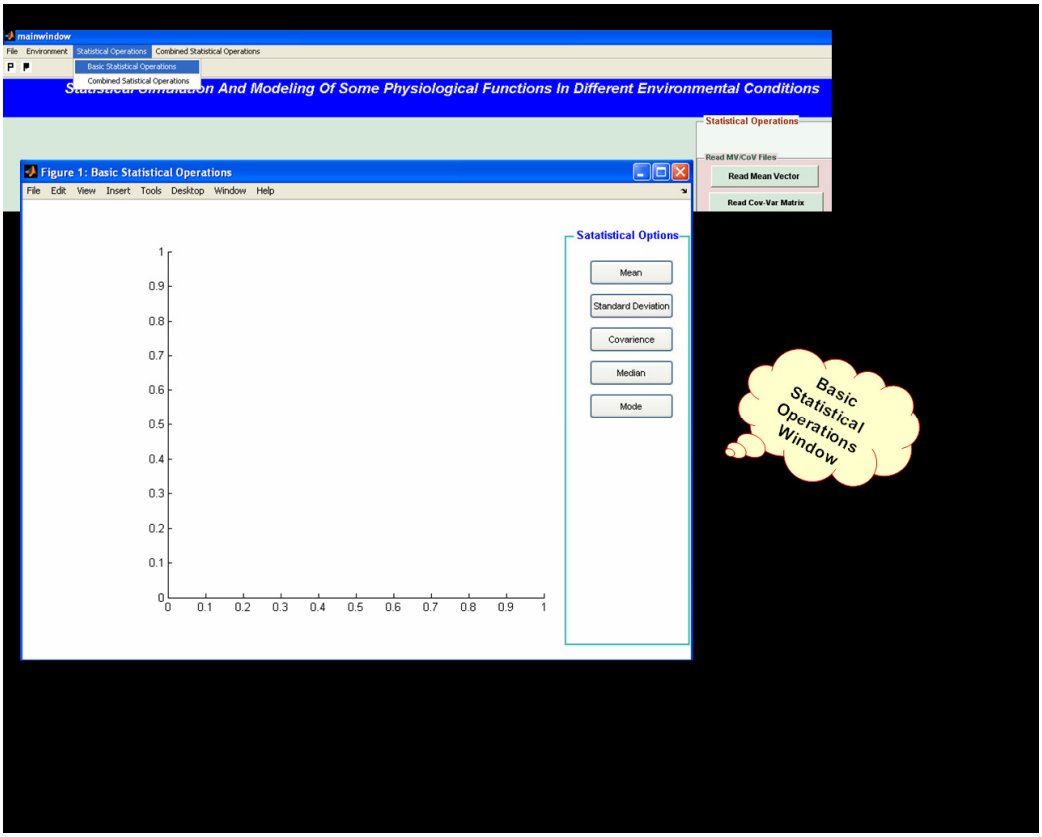

FIGURE 4 : Window for Basic Statistics

3.4 MODULE 4: RESAMPLING

Many researchers face the situation when they are not having the sample of sufficient size to do statistical analysis, at that point of time they has to close their research just because of lack of records or data. Resampling, the advance statistical technique is very useful in this situation. Resampling is also the type of sampling but in this the repeated sampling is done within the same sample, this is the reason it is called resampling. There are many techniques for resampling, common resampling techniques include bootstrapping, jackknifing and permutation tests. The Resampling module perform resampling using Jackknife Technique and Bootstrap Technique on samples generated by the module SS. Under this module user has to select the resampling technique from the drop down list and also the resample function , then he has to click on OK button. The output will save in the same folder MyOutput in C drive.

3.5 RESULT

The simulated population generation takes mean, variance- covariance matrix and the population size as input and produces the simulated population as output. The generated simulated population get stored in the excel worksheet a saved in the MyOutput folder under C drive. The second module Sample selection (SS) takes the population generated by the PG module or some other existing population, Number of Samples and Sample size as input and gives Samples as output based on the sampling techniques selected by the user. The third module is the basic statistics (BS) that can do all the basic statistical calculation as mean, median, mode, standard deviation on the selected file. This input file can be the population or samples of any size. The last module of SSMPP is the resampling. This module is based on two resampling techniques as jackknife and bootstrap.

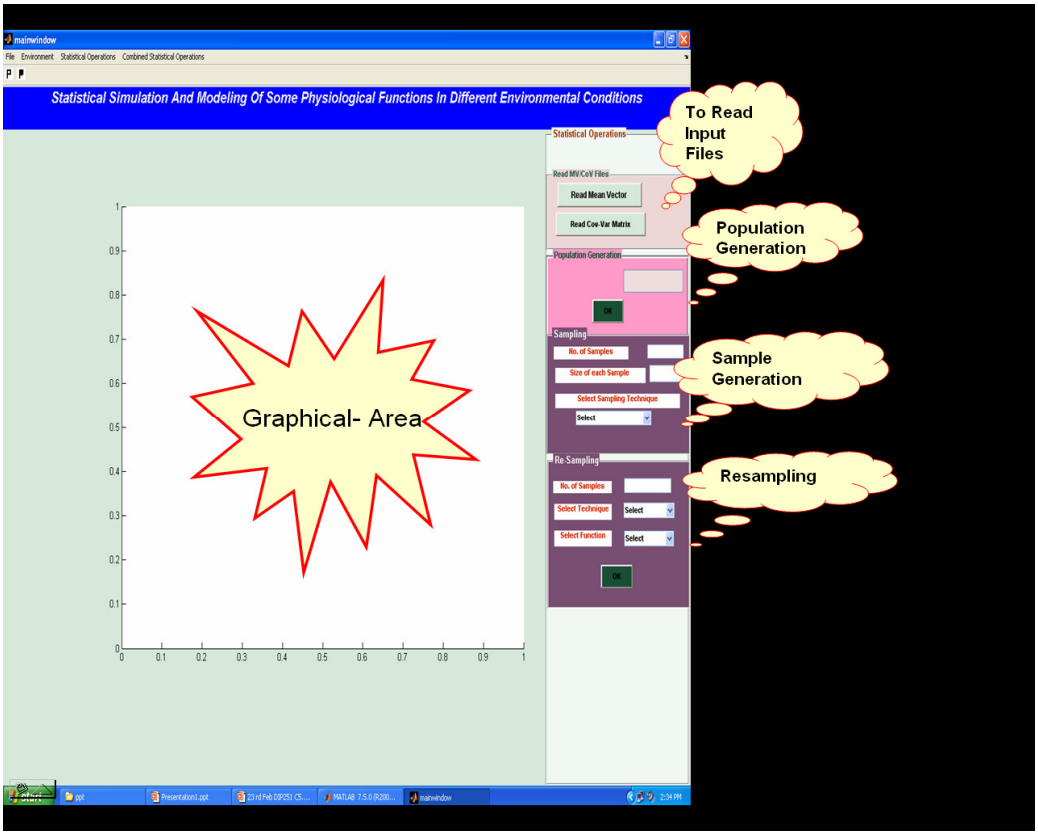

FIGURE 5: Main Window of SSPR

The results generated by this can further be used for statistical modeling, Multivariate analysis and Meta analysis. This framework is designed and developed using MATLAB an interactive numerical computing environment which is very reliable in terms of mathematical computing.

4. CONCLUSION

Kelton and Maria (1997) describe method to design the run for simulation models and interpreting their output. Statistical methods are described for several different purposes and related problem like comparisons , variance reduction etc. as has also been done in the present paper[3][12].Our work is similar to the work done by Thomson(1996) in estimating the uncertainty in physiological based pharmacokinetics model output by using Monte Carlo simulation to furnish random sample values for model parameters for further statistical analysis[13].Rubin(2006) created an ontologically guided methodology for representing a physiological model of circulation. Ontology provided a framework to construct a graphical representation of the model. Providing a simpler visualization than the large set of mathematical equations [10].In this study software SSPR that can do statistical simulation of some physiological functions in different environmental conditions for DIPAS is introduced. The study elaborates the system analysis, software design and system development. With the given inputs this system is able to do basic statistical calculations, generate simulated population, sampling and resampling by using different techniques as SRSWR, SRSWOR, PPSWR, PPSWOR, jackknife etc. In addition, it is a GUI based system so it is very user friendly and controls the data effectively. It uses the MS Excel as a database and also produces the results in the excel format, which can be easily readable in other statistical software, where further advance statistical analysis can be done. There is a less scope of manual error when this software is used for statistical operations. It is very flexible in terms of development. This framework will help the researchers and scientists who are working on sensitive information and wants statistical analysis without revealing the information. Researcher can generate the simulated population with respect to their data using PG module and can

perform statistical analysis. SSPR also help the researchers who want to continue the research based on old studies and are having the some descriptive statistics only. SSPR also help in the situation when sample of sufficient size is not available by providing the Resampling module. The result generated by SS and Resampling Module can further be used for statistical modeling and meta analysis by exporting the output file into other statistical software. The output provided by the software is the graphical output as well as in excel sheet which is compatible with all statistical software. In future SSPR can be upgraded by including other statistical estimators, sampling and resampling techniques as like two stage sampling techniques stratified and unstratified etc.

REFERENCES

[1] E. Z. Erson, M. C. Cavusoglu. "Design of a Framework for Modeling, Integration and Simulation of Physiological Models" .In Proceedings of the International Conference of the IEEE Engineering in Medicine and Biology Society(EMBC '10),August 31 - September 4, 2010, Buenos Aires, Argentina.

[2] R.L. Carter. "Solutions for Missing Data in Structural Equation Modeling". Research & Practice in Assessment Volume 1, Issue 1 March 2006.

[3] A. Maria "Introduction to Modeling and Simulation" . *Proceedingg of the 1997 Winter simulation conference ed.S.Andradottir, K.J.Healy, D.H. Withers, and B.L.nelson,* Atlanta, GA, USA ,1997.

[4] Krishnamurty Muralidhar and R. Sarathy. "Generating Sufficiency based Non Synthetic Perturbed Data". *Transations on Data Privacy* 117-23, 2008.

[5] Therese McGinn ,2004. "Instructions for Probability Proportional to Size Sampling Technique". RHRC Consortium monitoring and evaluation toolkit , October 2004. www.rhrc.org/.../55b%20PPS%20sampling%20technique.doc

[6] P. Bratley ,Bennett L.Fox, Linus E. Schrage."A Guide to Simulation" ,Second Edition .*Springer, Eds. Publisher, New York*, 164-165, 1987.

[7] P.V. Sukhatme, B.V.Sukhatme ,S.Sukhatme ,C. Asok . "Sampling Theory of surveys with application.Indian society of agricultural statistics", new delhi India , and IOWA State University Press Ames , USA, 21-25, 1984.

[8] William G.Cochran. "*Sampling Techniques",* Third Edition. *Wiley Eastern Limited Publication New Delhi*, 1985, Pp-18-30, 250-259

[9] Wu, C.F.J. "Jackknife, Bootstrap and other resampling methods in regression analysis". *The Annals of Statistics.* Vol. 14, 4, pp. 1261–1295, 1986.

[10] D.L. Rubin, D.Grossman, M. Neal, D.L. Cook. "Ontology-Based Representation of Simulation Models of Physiology". *AMIA 2006 Symposium Proceedings* Page – 664-668.

[11] Matlab help Demo

[12] W. D.Kelton ,"Statistical Analysis of Simulation Output". *Proceedings of the 1997 Winter Simulation Conference, Atlanta, GA, USA ,* 1997,Page-23-30.

[13] R. S. Thomas, W. E. Lytle, T. J. Keefe, Alexander A. Constan, And R. S. H. Yang "Incorporating Monte Carlo Simulation into Physiologically Based Pharmacokinetic Models Using Advanced Continuous Simulation Language (ACSL): A Computational Method". *Fundamental and applied Toxicology 31*, 1996.Page-19-28.