World Bank – ICAR funded
National Agricultural Higher Education Project
Centre for Advanced Agricultural Science and Technology
(CAAST)

On

# *Genomics Assisted Crop Improvement and Management*

Training Manual

# *Analytical Techniques for Impact Assessment of Agricultural Technologies & Policies*

March 17- 27, 2021

**Division of Agricultural Economics**
**ICAR – Indian Agricultural Research Institute**
**New Delhi – 110012**
**www.nahep-caast.iari.res.in**

**NAHEP Sponsored**
**Online Workshop Programme**
**On**
# Analytical Techniques for "Impact Assessment of Agricultural Technologies and Policies

**Course Director**
**Alka Singh**
Professor and Head
Division of Agricultural Economics
ICAR-Indian Agricultural Research Institute
New Delhi

**Course Coordinators**
**Anjani Kumar**
Senior Research Fellow
International Food Policy Research Institute
South Asia Regional Office
New Delhi
**Praveen K V**
Scientist
Division of Agricultural Economics
ICAR-Indian Agricultural Research Institute
New Delhi 110 012
**Aditya K S**
Scientist
Division of Agricultural Economics
ICAR-Indian Agricultural Research Institute
New Delhi
**Nithyashree M L**
Scientist
Division of Agricultural Economics
ICAR-Indian Agricultural Research Institute
New Delhi

**Division of Agriculture Economics**
**ICAR-Indian Agricultural Research Institute**
**New Delhi- 110 012**

# About NAHEP-CAAST at IARI, New Delhi

**Centre for Advanced Agricultural Science and Technology (CAAST)** is a new initiative and student centric subcomponent of World Bank sponsored **National Agricultural Higher Education Project (NAHEP)** granted to the Indian Council of Agricultural Research, New Delhi to provide a platform for strengthening educational and research activities of post graduate and doctoral students. The ICAR-Indian Agricultural Research Institute, New Delhi was selected by the NAHEP-CAAST programme. NAHEP sanctioned Rs 19.99 crores for the project on "**Genomic assisted crop improvement and management**" under CAAST programme. The project at IARI specifically aims at inculcating genomics education and skills among the students and enhancing the expertise of the faculty of IARI in the area of genomics.

**Objectives:**

1. **To develop online teaching facility and online courses for enhancing the teaching and learning efficiency, and scientific communication skills**
2. **To develop and/or strengthen state-of-the art next-generation genomics and phenomics facilities for producing quality PG and Ph.D.students**
3. **To develop collaborative research programmes with institutes of international repute and industries in the area of genomics and phenomics**
4. **To enhance the skills of faculty and PG students of IARI and NARES**
5. **To generate and analyze big data in genomics and phenomics of crops, microbes and pests for genomics augmentation of crop improvement and management**

IARI's CAAST project is unique as it aimed at providing funding and training support to the M.Sc. and Ph.D. students from different disciplines who are working in the area of genomics. It will organize lectures and training programmes, send IARI students for training at expert laboratories and research institutions abroad, and cover students from several disciplines. It will provide opportunities to the students and faculty to gain international exposure. Further, the project envisages developing a modern lab named as **Discovery Centre** that will serve as a common facility for students' research at IARI.

**Core-Team Members**

| S.No | Name of the Faculty | Discipline | Institute |
|---|---|---|---|
| 1. | Dr. Ashok K. Singh | Genetics | ICAR-IARI |
| 2. | Dr. Vinod | Genetics | ICAR-IARI |
| 3. | Dr. Gopala Krishnan S | Genetics | ICAR-IARI |
| 4. | Dr. A. Kumar | Plant Pathology | ICAR-IARI |
| 5. | Dr. T.K. Behera | Vegetable Science | ICAR-IARI |
| 6. | Dr. R.N. Sahoo | Agricultural Physics | ICAR-IARI |
| 7. | Dr. Alka Singh | Agricultural Economics | ICAR-IARI |
| 8. | Dr. A.R. Rao | Bioinformatics | ICAR-IASRI |
| 9. | Dr. R.C. Bhattacharya | Molecular Biology & Biotechnology | ICAR-NIPB |
| 10. | Dr. K. Annapurna | Microbiology<br>**Nodal officer, Grievance Redressal, CAAST** | ICAR-IARI |
| 11. | Dr. R. Roy Burman | Agricultural Extension<br>**Nodal officer, Equity Action Plan, CAAST** | ICAR-IARI |
| 12. | Dr. K.M. Manjaiah | Soil Science & Agri. Chemistry<br>**Nodal officer, CAAST** | ICAR-IARI |
| 13. | Dr.Viswanathan Chinnusamy | Plant Physiology | ICAR-IARI |

| | | | |
|---|---|---|---|
| | | **PI, CAAST** | |

## Associate Team

| S.No. | Name of the Faculty | Discipline | Institute |
|---|---|---|---|
| 14. | Dr. Kumar Durgesh | Genetics | ICAR-IARI |
| 15. | Dr. Ranjith K. Ellur | Genetics | ICAR-IARI |
| 16. | Dr. N. Saini | Genetics | ICAR-IARI |
| 17. | Dr. D. Vijay | Seed Science & Technology | ICAR-IARI |
| 18. | Dr. Kishor Gaikwad | Molecular Biology & Biotechnology | ICAR-NIPB |
| 19. | Dr. Mahesh Rao | Genetics | ICAR-NIPB |
| 20. | Dr. Veena Gupta | Economic Botany | ICAR-NBPGR |
| 21. | Dr. Era V. Malhotra | Molecular Biology & Biotechnology | ICAR-NBPGR |
| 22. | Dr. Sudhir Kumar | Plant Physiology | ICAR-IARI |
| 23. | Dr. Dhandapani R | Plant Physiology | ICAR-IARI |
| 24. | Dr. Lekshmy S | Plant Physiology | ICAR-IARI |
| 25. | Dr. Madan Pal | Plant Physiology | ICAR-IARI |
| 26. | Dr. Shelly Praveen | Biochemistry | ICAR-IARI |
| 27. | Dr. Suresh Kumar | Biochemistry | ICAR-IARI |
| 28. | Dr. Ranjeet R. Kumar | Biochemistry | ICAR-IARI |
| 29. | Dr. S.K. Singh | Fruits & Horticultural Technology | ICAR-IARI |
| 30. | Dr. Manish Srivastava | Fruits & Horticultural Technology | ICAR-IARI |
| 31. | Dr. Amit Kumar Goswami | Fruits & Horticulture Technology | ICAR-IARI |
| 32. | Dr. Srawan Singh | Vegetable Science | ICAR-IARI |
| 33. | Dr. Gograj S Jat | Vegetable Science | ICAR-IARI |
| 34. | D. Praveen Kumar Singh | Vegetable Science | ICAR-IARI |
| 35. | Dr. V.K. Baranwal | Plant Pathology | ICAR-IARI |
| 36. | Dr. Deeba Kamil | Plant Pathology | ICAR-IARI |
| 37. | Dr. Vaibhav K. Singh | Plant Pathology | ICAR-IARI |
| 38. | Dr. Uma Rao | Nematology | ICAR-IARI |
| 39. | Dr. S. Subramanium | Entomology | ICAR-IARI |
| 40. | Dr. M.K. Dhillon | Entomology | ICAR-IARI |
| 41. | Dr. B. Ramakrishnan | Microbiology | ICAR-IARI |
| 42. | Dr. V. Govindasamy | Microbiology | ICAR-IARI |
| 43. | Dr. S.P. Datta | Soil Science & Agricultural Chemistry | ICAR-IARI |
| 44. | Dr. R.N. Padaria | Agricultural Extension | ICAR-IARI |
| 45. | Dr. Satyapriya | Agricultural Extension | ICAR-IARI |
| 46. | Dr. Sudeep Marwaha | Computer Application | ICAR-IASRI |
| 47. | Dr. Seema Jaggi | Agricultural Statistics | ICAR-IASRI |
| 48. | Dr. Anindita Datta | Agricultural Statistics | ICAR-IASRI |
| 49. | Dr. Soumen Pal | Computer Application | ICAR-IASRI |
| 50. | Dr. Sanjeev Kumar | Bioinformatics | ICAR-IASRI |
| 51. | Dr. S.K. Jha | Food Science & Post Harvest Technology | ICAR-IARI |
| 52. | Dr. Shiv Dhar Mishra | Agronomy | ICAR-IARI |
| 53. | Dr. D.K. Singh | Agricultural Engineering | ICAR-IARI |
| 54. | Dr. S. Naresh Kumar | Environmental Sciences;<br>**Nodal officer, Environmental Management Framework** | ICAR-IARI |

# Foreword

The Division of Agricultural Economics is a constituent of the School of Social Sciences of Indian Council of Agricultural Research-Indian Agricultural Research Institute, was established in 1960. The mandate of the Division is to conduct research in frontier areas and serve as a centre for academic excellence in post-graduate education. Since its inception, the Division has been making contributions in basic and applied research with significant implications for agricultural policy. The Division has achieved excellence in post-graduate education and research as an ICAR-UNDP Centre of Excellence through a faculty exchange program for human resources development and strengthening of infrastructure facilities. Since 1995 it has been functioning as an ICAR Centre of Advanced Faculty Training (CAFT) to strengthen the capacity for agricultural economics and policy research in the national agricultural research system.

The research contributions of the Division have been globally recognized and many of the alumni occupy positions of repute in national and international organizations. The Division has maintained good academic liaison with other divisions at IARI and other national and international agricultural research institutions. The research focus of the Division has been continuously reoriented to address contemporary development challenges. Current research thrust areas of the division include investment in agriculture, inclusive growth, and poverty alleviation, the impact of agricultural technologies and policies, price forecasting and market outlooks, natural resource use in agriculture and ecosystem services, climate change effects, mitigation and adaptations, and food and nutritional security.

The division has carved a niche for itself in the use of rigorous empirical methods in social science research. Considering the strength of the division in the area of research methods, the division in collaboration with the International Food Policy Research Institute, South Asia regional office New Delhi has organized a 10 days online short training program on '**Analytical techniques for Impact assessment of agricultural technologies and policies'** from March 17th to March 27th, 2021'. The objective of the training program was to train young students and research scholars on the application of econometric tools for impact assessment of agricultural technologies and programs. The advances in the impact assessment methodology covered in this training program will facilitate students to update their knowledge and skill and facilitate them in addressing their research questions with empirical rigour

.

**Rashmi Aggarwal**

Dean and Joint Director (Edn)
ICAR-IARI, New Delhi

Date: 27.03.2021

# Preface

The focus on Agricultural Economics research in the recent period has shifted considerably towards estimating the impact of policies and programs. Therefore, improving the knowledge of the students in this area will thus act as a strong base in their development as researchers in the field of Agricultural Economics. This training manual is prepared considering the target of upgrading the research skills of the post-graduate students of social sciences. The manual is based on the NAHEP-CAAST sponsored online short training course titled '**Analytical techniques for Impact assessment of agricultural technologies and policies'** from March 17th to March 27th, 2021' organized by the Division of Agricultural Economics, ICAR-Indian Agricultural Research Institute, New Delhi in collaboration with International Food Policy Research Institute, South Asia regional office, New Delhi. Centre for Advanced Agricultural Science and Technology (CAAST) is a new initiative and student-centric sub-component of the World Bank-sponsored National Agricultural Higher Education Project (NAHEP), granted to IARI to provide a platform for strengthening education and research activities of post-graduate students.

The chapters in the compendium will present an overview of the important techniques for impact evaluation. The chapters are designed in such a way that the basic idea of statistics and software tools are provided, followed by specific impact assessment methodologies. They also provide guidelines on practical issues in implementing each of these techniques and also provide codes for implementing them in software programs. The aim of this program and the compendium is to get the participants comfortable with ideas and principles behind different impact assessment methods to enable them to use these tools in their research works. We take this opportunity to sincerely acknowledge the contribution of all the authors in the preparation of this manual. Considering the diversity and comprehensive nature of the topics covered, the manual can act as a quick and effective reference source for the students in their future research endeavours.

Alka Singh
Anjani Kumar
Praveen K V
Aditya K S
Nithyashree M L

Date: 27.03.2021

# Acknowledgments

1. Secretary DARE and Director General ICAR, New Delhi
2. Deputy Director General (Education), ICAR, New Delhi
3. Assistant Director General (HRD), ICAR, New Delhi
4. National Coordinator, NAHEP, ICAR, New Delhi
5. CAAST Team, ICAR-IARI, New Delhi
6. P.G. School, ICAR-IARI, New Delhi
7. Director, ICAR-IARI, New Delhi
8. Dean & Joint Director (Education), ICAR-IARI, New Delhi
9. Joint Director (Research), ICAR-IARI, New Delhi
10. Head, Division of Agriculture Economics, ICAR-IARI, New Delhi
11. Professor, Division of Agriculture Economics, ICAR-IARI, New Delhi
12. AKMU, ICAR-IARI, New Delhi
13. Staff & Students, Division of Agriculture Economics, ICAR-IARI, New Delhi

# Content

# Overview on R Software

**Ranjit Kumar Paul**

*Senior Scientist, Division of Statistical Genetics, ICAR-IASRI, New Delhi*

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. R is a vehicle for newly developing methods of interactive data analysis. It has developed rapidly, and has been extended by a large collection of packages.

## R environment

The R environment provides an integrated suite of software facilities for data manipulation, calculation and graphical display. It has

- a data handling and storage facility,
- a suite of operators for calculations on arrays and matrices,
- a large, integrated collection of intermediate tools for data analysis,
- graphical facilities for data analysis and display, and
- a well developed, simple and effective programming language (called 'S') which includes conditionals, loops, user defined functions and input and output facilities.

## Origin

R can be regarded as an implementation of the S language which was developed at BellLaboratories by Rick Becker, John Chambers and Allan Wilks, and also forms the basis of the S-Plus systems. Robert Gentleman and Ross Ihaka of the Statistics Department of the University ofAuckland started the project on R in 1995 and hence the name software has been named as 'R'.

R was introduced as an environment within which many classical and modern statistical techniquescan be implemented. A few of these are built into the base R environment, but many are suppliedas packages. There are a number of packages supplied with R (called "standard" and "recommended" packages) and many more are available through the CRAN family of Internet sites (via http://cran.r-project.org) and elsewhere.

## Availability

Since R is an open source project, it can be obtained freely from the website www.**r**-project.org. One can download R from any CRAN mirror out of several CRAN (Comprehensive R Archive Network) mirrors. Latest available version of R is *R version 4.0.4* and it has been released on 15.02.2021.

## Installation

To install R in windows operating system, simply double click on the setup file. It willautomatically install the software in the system.

## Usage

R can work under Windows, UNIX and Mac OS. In this note, we consider usage of R in Windowsset up only.

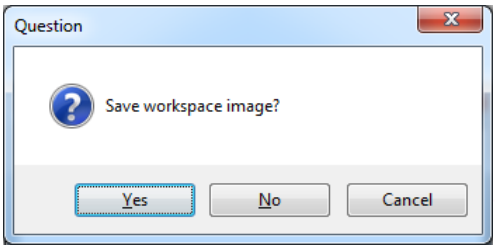
## Difference with other packages

There is an important difference between R and the other statistical packages. In R, a statistical analysis is normally done as a series of steps, with intermediate results being stored in objects. Thus whereas SAS and SPSS will give large amount of output from a given analysis, R will give minimal output and store the results in an object for subsequent interrogation by further R functions.

## Invoking R

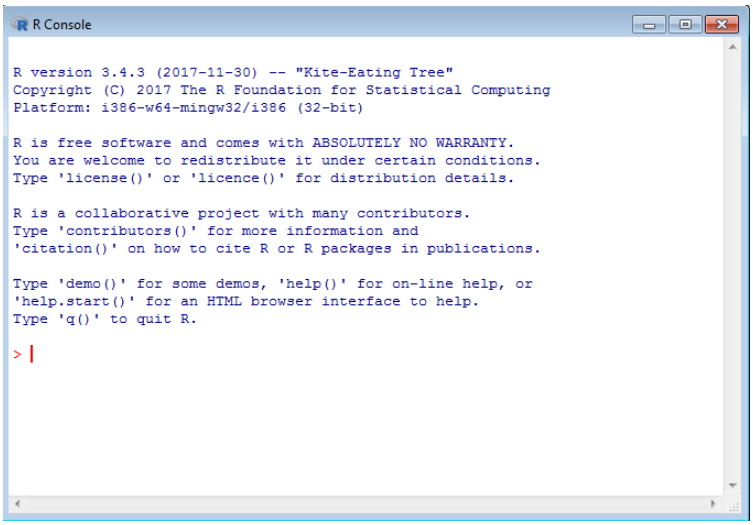
If properly installed, usually R has a shortcut icon on the desktop screen and/or we can find it under Start|All Programs|R menu.


To quit R, type *q()* at the R prompt (>) and press Enter key. A dialog box will ask whether to savethe objects we have created during the session so that they will become available next time when R will be invoked.




## Windows of R

R has only one window and when R is started it looks like




## R commands

i. R commands are case sensitive, so X and x are different symbols and would refer todifferent variables.
ii. Elementary commands consist of either expressions or assignments.
iii. If an expression is given as a command, it is evaluated, printed and the value is lost.
iv. An assignment also evaluates an expression and passes the value to a variable but the resultis not automatically printed.
v. Commands are separated either by a semi-colon (';'), or by a newline.
vi. Elementary commands can be grouped together into one compound expression by braces'{' and '}'.
vii. Comments can be put almost anywhere, starting with a hashmark ('#'). Anything writtenafter # marks to the end of the line is considered as a comment.
viii. Window can be cleared of lines by pressing Ctrl + L keys.

## Data Types

**R** has a wide variety of data types including scalars, vectors (numerical, character, logical),matrices, dataframes, and lists.

## Vectors

```
a <- c(1,2,5.3,6,-2,4) # numeric vector
b <- c("one","two","three") # character vector
c <- c(TRUE,TRUE,TRUE,FALSE,TRUE,FALSE) #logical vector
```

Refer to elements of a vector using subscripts.a[c(2,4)]
# 2nd and 4th elements of vector **Matrices**

All columns in a matrix must have the same mode(numeric, character, etc.) and the same length.
The general format is

```
mymatrix <- matrix(vector, nrow=r, ncol=c, byrow=FALSE,
  dimnames=list(char_vector_rownames, char_vector_colnames))
```

**byrow=TRUE** indicates that the matrix should be filled by rows. **byrow=FALSE** indicates that the matrix should be filled by columns (the default). **dimnames** provides optional labels for the columns and rows.

```
# generates 5 x 4 numeric matrixy<-
matrix(1:20, nrow=5,ncol=4)

# another example cells <-
c(1,26,24,68)
rnames <- c("R1", "R2")
cnames <- c("C1", "C2")
  mymatrix <- matrix(cells, nrow=2, ncol=2, byrow=TRUE,
    dimnames=list(rnames, cnames))
Identify rows, columns or elements using subscripts.x[,4] #
4th column of matrix
x[3,] # 3rd row of matrix
x[2:4,1:3] # rows 2,3,4 of columns 1,2,3
```

## Arrays

Arrays are similar to matrices but can have more than two dimensions. See **help(array)** for details.

## Dataframes

A dataframe is more general than a matrix, in that different columns can have different modes(numeric, character, factor, etc.). This is similar to SAS and SPSS datasets.

```
d <- c(1,2,3,4)
e <- c("red", "white", "red", NA)
f <- c(TRUE,TRUE,TRUE,FALSE)
mydata <- data.frame(d,e,f)
names(mydata) <- c("ID","Color","Passed") # variable names There are a
variety of ways to identify the elements of a dataframe .
myframe[3:5] # columns 3,4,5 of dataframe myframe[c("ID","Age")] #
columns ID and Age from dataframemyframe$X1 # variable x1 in the
dataframe
```

## Lists

An ordered collection of objects (components). A list allows us to gather a variety of (possiblyunrelated) objects under one name.

```
# example of a list with 4 components -
# a string, a numeric vector, a matrix, and a scaler
w <- list(name="Fred", mynumbers=a, mymatrix=y, age=5.3)

# example of a list containing two listsv <-
c(list1,list2)
Intentify elements of a list using the [[]] convention.
mylist[[2]] # 2nd component of the list
mylist[["mynumbers"]] # component named mynumbers in list
```

## Factors

Tell **R** that a variable is **nominal** by making it a factor. The factor stores the nominal values as a vector of integers in the range [ 1... k ] (where k is the number of unique values in the nominal variable), and an internal vector of character strings (the original values) mapped to these integers.

```
# variable gender with 20 "male" entries and# 30
"female" entries
gender <- c(rep("male",20), rep("female", 30))gender
<- factor(gender)
# stores gender as 20 1s and 30 2s and associates#
1=female, 2=male internally (alphabetically)
# R now treats gender as a nominal variable
summary(gender)
```

An ordered factor is used to represent an **ordinal variable**.

```
# variable rating coded as "large", "medium", "small'rating <-
ordered(rating)
# recodes rating to 1,2,3 and associates
# 1=large, 2=medium, 3=small internally# R
now treats rating as ordinal
```

**R** will treat factors as nominal variables and ordered factors as ordinal variables in statistical proceedures and graphical analyses. We can use options in the **factor( )** and **ordered( )** functionsto control the mapping of integers to strings (overiding the alphabetical ordering). We can also use factors to create value labels.

## Executing commands from or diverting output to a file

If commands are stored in an external file, say 'D:/commands.txt' they may be executed at anytime in an R session with the command

```
> source("d:/commands.txt")
```

For Windows Source is also available on the File menu. The function *sink()*,

```
> sink("d:/record.txt")
```

will divert all subsequent output from the console to an external file, 'record.txt' in D drive. Thecommand

```
> sink()
```

restores it to the console once again.

## Simple manipulations of numbers and vectors

R operates on named data structures. The simplest such structure is the numeric vector, which is asingle entity consisting of an ordered collection of numbers. To set up a vector named x, say, consisting of five numbers, namely 10.4, 5.6, 3.1, 6.4 and 21.7, use the R command

```
> x <- c(10.4, 5.6, 3.1, 6.4, 21.7)
```

The function *c()* assigns the five numbers to the vector x. The assignment operator (<-) 'points' tothe object receiving the value of the expression. Once can use the '=' operator as an alternative.

A single number is taken as a vector of length one.

Assignments can also be made in the other direction, using the obvious change in the assignmentoperator. So the same assignment could be made using

```
> c(10.4, 5.6, 3.1, 6.4, 21.7) -> x
```

If an expression is used as a complete command, the value is printed. So now if we were to use thecommand

```
> 1/x
```

the reciprocals of the five values would be printed at the terminal.

## Operators

**R**'s binary and logical operators will look very familiar to programmers. Note that binary operatorswork on vectors and matrices as well as scalars.

## Arithmetic Operators

| Operator | Description |
|---|---|
| + | addition |
| - | subtraction |
| * | multiplication |
| / | division |
| ^ or ** | exponentiation |
| x %% y | modulus (x mod y) 5%%2 is 1 |
| x %/% y | integer division 5%/%2 is 2 |

## Logical Operators

| Operator | Description |
|---|---|
| < | less than |
| <= | less than or equal to |
| > | greater than |
| >= | greater than or equal to |
| == | exactly equal to |
| != | not equal to |
| !x | Not x |
| x \| y | x OR y |
| x & y | x AND y |

**isTRUE(x)** test if X is TRUE

```
 # An example x <-
c(1:10)    x[(x>8)    |
(x<5)]
# yeilds 1 2 3 4 9 10

# How it worksx <-
c(1:10)
x
1 2 3 4 5 6 7 8 9 10
x > 8
F F F F F F F F T T
x < 5
T T T T F F F F F F
x > 8 | x < 5
T   T   T   T   F   F   F   F   T   T
x[c(T,T,T,T,F,F,F,F,T,T)]1  2  3
4 9 10
```

## Built-in Functions

Almost everything in **R** is done through functions. Here I'm only refering to numeric and characterfunctions that are commonly used in creating or recoding variables.

## Numeric Functions

| Function | Description |
|---|---|
| **abs(*x*)** | absolute value |
| **sqrt(*x*)** | square root |
| **ceiling(*x*)** | ceiling(3.475) is 4 |
| **floor(*x*)** | floor(3.475) is 3 |
| **trunc(*x*)** | trunc(5.99) is 5 |
| **round(*x*, digits=*n*)** | round(3.475, digits=2) is 3.48 |
| **signif(*x*, digits=*n*)** | signif(3.475, digits=2) is 3.5 |

| | |
|---|---|
| **cos(*x*), sin(*x*), tan(*x*)** | also acos(*x*), cosh(*x*), acosh(*x*), etc. |
| **log(*x*)** | natural logarithm |
| **log10(*x*)** | common logarithm |
| **exp(*x*)** | e^*x* |

## Character Functions

| Function | Description |
|---|---|
| **substr(*x*, start=*n1*, stop=*n2*)** | Extract or replace substrings in a character vector.x <- "abcdef"<br>substr(x, 2, 4) is "bcd"<br>substr(x, 2, 4) <- "22222" is "a222ef" |
| **grep(*pattern*, *x* , ignore.case=FALSE, fixed=FALSE)** | Search for *pattern* in *x*. If fixed =FALSE then *pattern* is aregular expression. If fixed=TRUE then *pattern* is a text string. Returns matching indices.<br>grep("A", c("b","A","c"), fixed=TRUE) returns 2 |
| **sub(*pattern*, *replacement*, *x*, ignore.case =FALSE, fixed=FALSE)** | Find *pattern* in *x* and replace with *replacement* text. If fixed=FALSE then *pattern* is a regular expression.<br>If fixed = T then *pattern* is a text string.<br>sub("\\s",".","Hello There") returns "Hello.There" |
| **strsplit(*x*, *split*)** | Split the elements of character vector *x* at *split*.<br>strsplit("abc", "") returns 3 element vector "a","b","c" |
| **paste(..., sep="")** | Concatenate strings after using *sep* string to seperate them.<br>paste("x",1:3,sep="") returns c("x1","x2" "x3")<br>paste("x",1:3,sep="M") returns c("xM1","xM2" "xM3")<br>paste("Today is", date()) |
| **toupper(*x*)** | Uppercase |
| **tolower(*x*)** | Lowercase |

## Statistical Probability Functions

The following table describes functions related to probaility distributions. For random number generators below, we can use set.seed(1234) or some other integer to create reproducible pseudo-random numbers.

| Function | Description |
|---|---|
| **dnorm(***x***)** | normal density function (by default m=0 sd=1)# plot standard normal curve<br>x <- pretty(c(-3,3), 30)y <- dnorm(x)<br>plot(x, y, type='l', xlab="Normal Deviate", ylab="Density", yaxs="i") |
| **pnorm(***q***)** | cumulative normal probability for q<br>(area under the normal curve to the right of q)<br>pnorm(1.96) is 0.975 |
| **qnorm(***p***)** | normal quantile.<br>value at the p percentile of normal distribution<br>qnorm(.9) is 1.28 # 90th percentile |
| **rnorm(***n***, m=0,sd=1)** | n random normal deviates with mean mand standard deviation sd.<br>#50 random normal variates with mean=50, sd=10x <- rnorm(50, m=50, sd=10) |
| **dbinom(***x***, *size*, *prob*)**<br>**pbinom(***q***, *size*, *prob*)**<br>**qbinom(***p***, *size*, *prob*)**<br>**rbinom(***n***, *size*, *prob*)** | binomial distribution where size is the sample sizeand prob is the probability of a heads (pi)<br># prob of 0 to 5 heads of fair coin out of 10 flips<br>dbinom(0:5, 10, .5)<br># prob of 5 or less heads of fair coin out of 10 flipspbinom(5, 10, .5) |
| **dpois(***x***, *lamda*)**<br>**ppois(***q***, *lamda*)**<br>**qpois(***p***, *lamda*)**<br>**rpois(***n***, *lamda*)** | poisson distribution with m=std=lamda<br>#probability of 0,1, or 2 events with lamda=4<br>dpois(0:2, 4)<br># probability of at least 3 events with lamda=41-ppois(2,4) |
| **dunif(***x***, min=0, max=1)**<br>**punif(***q***, min=0, max=1)** | uniform distribution, follows the same patternas the normal distribution above. |

| | |
|---|---|
| **qunif(***p***, min**=0**, max**=1**)** | #10 uniform random variates |
| **runif(***n***, min**=0**, max**=1**)** | x <- runif(10) |

## Other Statistical Functions

Other useful statistical functions are provided in the following table. Each has the option na.rm tostrip missing values before calculations. Otherwise the presence of missing values will lead to a missing result. Object can be a numeric vector or dataframe.

| Function | Description |
|---|---|
| **mean(***x***, trim**=0**, na.rm**=FALSE**)** | mean of object x<br># trimmed mean, removing any missing values and# 5 percent of highest and lowest scores<br>mx <- mean(x,trim=.05,na.rm=TRUE) |
| **sd(***x***)** | standard deviation of object(x). also look at var(x) for variance andmad(x) for median absolute deviation. |
| **median(***x***)** | median |
| **quantile(***x***,** *probs***)** | quantiles where x is the numeric vector whose quantiles are desiredand probs is a numeric vector with probabilities in [0,1].<br># 30th and 84th percentiles of xy <-<br>quantile(x, c(.3,.84)) |
| **range(***x***)** | range |
| **sum(***x***)** | sum |
| **diff(***x***, lag=***1***)** | lagged differences, with lag indicating which lag to use |
| **min(***x***)** | minimum |
| **max(***x***)** | maximum |
| **scale(***x***, center**=TRUE**, scale**=TRUE**)** | column center or standardize a matrix. |

## Other Useful Functions

| Function | Description |
|---|---|
| **seq(***from* **,** *to***,** *by***)** | generate a sequence indices<br><- seq(1,10,2) #indices is<br>c(1, 3, 5, 7, 9) |
| **rep(***x***,** *ntimes***)** | repeat *x n* timesy<br><- rep(1:3, 2)<br># y is c(1, 2, 3, 1, 2, 3) |
| **cut(***x***,** *n***)** | divide continuous variable in factor with *n* levelsy <-<br>cut(x, 5) |

Note that while the examples on this page apply functions to individual variables, many can beapplied to vectors and matrices as well.

## Other objects in R

Matrices or arrays - multi-dimensional generalizations of vectors.
Lists - a general form of vector in which the various elements need not be of the same type, andare often themselves vectors or lists.
Functions - objects in R which can be stored in the project's workspace. This provides a simpleand convenient way to extend R.

## Matrix facilities

A matrix is just an array with two subscripts. R provides many operators and functions those areavailable only for matrices. Some of the important R functions for matrices are
t(A) – transpose of the matrix A
nrow(A) – number of rows in the matrix A ncol(A) – number of columns in the matrix A A%*% B– Cross product of two matrices A and B
A*B – element by element product of two matrices A and B
diag (A) – gives a vector of diagonal elements of the square matrix A diag(a) – gives a matrix with diagonal elements as the elements of vector aeigen(A) – gives eigen values and eigen vectors of a symmetric matrix A

rbind (A,B) – concatenates two matrix A and B by appending B matrix below A matrix (Theyshould have same number of columns)

cbind(A, B) - concatenates two matrix A and B by appending B matrix in the right of A matrix(They should have same number of rows)

mydata

| id | time | x1 | x2 |
|---|---|---|---|
| 1 | 1 | 5 | 6 |
| 1 | 2 | 3 | 5 |
| 2 | 1 | 6 | 1 |
| 2 | 2 | 2 | 4 |

```
# example of melt function
library(reshape)
mdata <- melt(mydata, id=c("id","time"))
```

newdata

| id | time | variable | value |
|---|---|---|---|
| 1 | 1 | x1 | 5 |
| 1 | 2 | x1 | 3 |
| 2 | 1 | x1 | 6 |
| 2 | 2 | x1 | 2 |
| 1 | 1 | x2 | 6 |
| 1 | 2 | x2 | 5 |
| 2 | 1 | x2 | 1 |
| 2 | 2 | x2 | 4 |

```
# cast the melted data
# cast(data, formula, function)
subjmeans <- cast(mdata, id~variable, mean)
timemeans <- cast(mdata, time~variable, mean)
```

### subjmeans

| id | x1 | x2 |
|---|---|---|
| 1 | 4 | 5.5 |
| 2 | 4 | 2.5 |

**timemeans**

| time | x1 | x2 |
|---|---|---|
| 1 | 5.5 | 3.5 |
| 2 | 2.5 | 4.5 |

There is much more that we can do with the **melt( )** and **cast( )** functions. See the documentationfor more details.

## Subsetting Data

**R** has powerful indexing features for accessing object elements. These features can be used to select and exclude variables and observations. The following code snippets demonstrate ways tokeep or delete variables and observations and to take random samples from a dataset.

### Selecting (Keeping) Variables

```
# select variables v1, v2, v3 myvars
<- c("v1", "v2", "v3")newdata <-
mydata[myvars]

# another method
myvars <- paste("v", 1:3, sep="")
newdata <- mydata[myvars]

# select 1st and 5th thru 10th variables
newdata <- mydata[c(1,5:10)]
```

### Excluding (DROPPING) Variables

```
# exclude variables v1, v2, v3
myvars <- names(myvars) %in% c("v1", "v2", "v3")newdata
<- mydata[!myvars]
```

income (weight, income and all columns between them).

```
# using subset function (part 2)
newdata <- subset(mydata, sex=="m" & age > 25,
select=weight:income)
```

## Random Samples

Use the **sample( )** function to take a **random sample of size n** from a dataset.# take a random sample of size 50 from a dataset *mydata*

```
# sample without replacement
mysample <- mydata[sample(1:nrow(mydata),
  50replace=FALSE),]
```

## Descriptive Statistics

**R** provides a wide range of functions for obtaining summary statistics. One method of obtainingdescriptive statistics is to use the **sapply( )** function with a specified summary statistic.

```
# get means for variables in dataframe mydata #
excluding missing values
sapply(mydata, mean, na.rm=TRUE)
```

Possible functions used in sapply include **mean, sd, var, min, max, med, range, and quantile**.

There are also numerous **R** functions designed to provide a range of descriptive statistics at once. For example

```
# mean,median,25th and 75th quartiles,min,max
summary(mydata)

# Tukey min,lower-hinge, median,upper-hinge,max
fivenum(x)
```

Using the Hmisc package

```
library(Hmisc)
describe(mydata)
# n, nmiss, unique, mean, 5,10,25,50,75,90,95th percentiles # 5
lowest and 5 highest scores
```

Using the pastecs package

```
library(pastecs)
stat.desc(mydata)
# nbr.val, nbr.null, nbr.na, min max, range, sum,
# median, mean, SE.mean, CI.mean, var, std.dev, coef.varUsing the
psych package
library(psych)
describe(mydata)
# item name ,item number, nvalid, mean, sd, #
median, mad, min, max, skew, kurtosis, se
```

## Summary Statistics by Group

A simple way of generating summary statistics by grouping variable is available in the psychpackage

```
library(psych) describe.by(mydata,
group,...)
```

The **doBy** package provides much of the functionality of SAS PROC SUMMARY. It defines thedesired table using a model formula and a function. Here is a simple example.

```
library(doBy)
summaryBy(mpg + wt ~ cyl + vs, data = mtcars, FUN =
function(x) { c(m = mean(x), s = sd(x)) } ) # produces
mpg.m wt.m mpg.s wt.s for each
# combination of the levels of cyl and vs
```

**See also**: aggregating data.

## Frequencies and Crosstabs

This section describes the creation of frequency and contingency tables from categorical variables, along with tests of independence, measures of association, and methods for graphicallydisplaying results.

### Generating Frequency Tables

**R** provides many methods for creating frequency and contingency tables. Three are describedbelow. In the following examples, assume that A, B, and C represent categorical variables.

We can generate frequency tables using the **table( )** function, tables of proportions using the **prop.table( )** function, and marginal frequencies using **margin.table( )**.

```
# 2-Way Frequency Table
attach(mydata)
mytable <- table(A,B) # A will be rows, B will be columnsmytable
# print table

margin.table(mytable, 1) # A frequencies (summed over B)
margin.table(mytable, 2) # B frequencies (summed over A)

prop.table(mytable) # cell percentages
prop.table(mytable, 1) # row percentages
prop.table(mytable, 2) # column percentages
```

**table( )** can also generate multidimensional tables based on 3 or more categorical variables. Inthis case, use the **ftable( )** function to print the results more attractively.

```
# 3-Way Frequency Table
mytable <- table(A, B, C)
ftable(mytable)
```

**Table ignores missing values.** To include **NA** as a category in counts, include the table option exclude=NULL if the variable is a vector. If the variable is a factor we have to create a new factor using newfactor <- factor(oldfactor, exclude=NULL).

### xtabs

The **xtabs( )** function allows us to create crosstabulations using formula style input.

```
# 3-Way Frequency Table
mytable <- xtabs(~A+B+c, data=mydata)
ftable(mytable) # print table
summary(mytable) # chi-square test of indepedence
```

If a variable is included on the left side of the formula, it is assumed to be a vector of frequencies(useful if the data have already been tabulated).

### Crosstable

The **CrossTable( )** function in the gmodels package produces crosstabulations modeled afterPROC FREQ in **SAS** or CROSSTABS in **SPSS**. It has a wealth of options.

```
# 2-Way Cross Tabulation
library(gmodels)
CrossTable(mydata$myrowvar, mydata$mycolvar)
```

There are options to report percentages (row, column, cell), specify decimal places, produce Chi-square, Fisher, and McNemar tests of independence, report expected and residual values (pearson, standardized, adjusted standardized), include missing values as valid, annotate with row and column titles, and format as **SAS** or **SPSS** style output!

See **help(CrossTable)** for details.

## Tests of Independence

### Chi-Square Test

For 2-way tables we can use **chisq.test(***mytable***)** to test independence of the row and column variable. By default, the p-value is calculated from the asymptotic chi-squared distribution of thetest statistic. Optionally, the p-value can be derived via Monte Carlo simultation.

## Fisher Exact Test

**fisher.test(*x*)** provides an exact test of independence. *x* is a two dimensional contingency table inmatrix form.

## Mantel-Haenszel test

Use the **mantelhaen.test(*x*)** function to perform a Cochran-Mantel-Haenszel chi-squared test of the null hypothesis that two nominal variables are conditionally independent in each stratum, assuming that there is no three-way interaction. *x* is a 3 dimensional contingency table, where thelast dimension refers to the strata.

## Loglinear Models

We can use the **loglm( )** function in the **MASS** package to produce log-linear models. Forexample, let's assume we have a 3-way contingency table based on variables A, B, and C.

```
library(MASS)
mytable <- xtabs(~A+B+C, data=mydata)We can
perform the following tests:
```

**Mutual Independence**: A, B, and C are pairwise independent. loglm(~A+B+C, mytable)

**Partial Independence**: A is partially independent of B and C (i.e., A is independent of thecomposite variable BC).

```
loglin(~A+B+C+B*C, mytable)
```

**Conditional Independence:** A is independent of B, given C. loglm(~A+B+C+A*C+B*C, mytable)**No Three-Way Interaction** loglm(~A+B+C+A*B+A*C+B*C, mytable)

Martin Theus and Stephan Lauer have written an excellent article on Visualizing Loglinear Models, using mosaic plots. There is also great tutorial example by Kevin Quinn on analyzingloglinear models via glm.

## Measures of Association

The **assocstats(*mytable*)** function in the **vcd** package calculates the phi coefficient, contingencycoefficient, and Cramer's V for an rxc table. The **kappa(*mytable*)** function in the **vcd** package calculates Cohen's kappa and weighted kappa for a confusion matrix. See Richard Darlington's article on Measures of Association in Crosstab Tables for an excellent review of these statistics.

The **rcorr( )** function in the **Hmisc** package produces correlations/covariances and significancelevels for pearson and spearman correlations. However, input must be a matrix and pairwise deletion is used.

```
# Correlations with significance levels
library(Hmisc)
rcorr(x, type="pearson") # type can be pearson or spearman

#mtcars is a dataframe
rcorr(as.matrix(mtcars))
```

We can use the format **cor(X, Y)** or **rcorr(X, Y)** to generate correlations between the columns ofX and the columns of Y. This similar to the VAR and WITH commands in **SAS** PROC CORR.

```
# Correlation matrix from mtcars# with
mpg, cyl, and disp as rows# and hp,
drat, and wt as columnsx <- mtcars[1:3}
y <- mtcars[4:6]
cor(x, y)
```

## Other Types of Correlations

```
# polychoric correlation
# x is a contingency table of counts
library(polychor)
polychor(x)

# heterogeneous correlations in one matrix #
pearson (numeric-numeric),
# polyserial (numeric-ordinal),
# and polychoric (ordinal-ordinal)
# x is a dataframe with ordered factors# and
numeric variables library(polychor)
hetcor(x)


# partial correlations
library(ggm)
data(mydata)
pcor(c("a", "b", "x", "y", "z"), var(mydata))
# partial corr between a and b controlling for x, y, z
```

## Visualizing Correlations

Use **corrgram( )** to plot correlograms .

Use the pairs() or splom( ) to create scatterplot matrices.

A great example of a plotted **correlation matrix** can be found in the R Graph Gallery.

### t-tests

The **t.test( )** function produces a variety of t-tests. Unlike most statistical packages, the defaultassumes unequal variance and applies the Welsh df modification.# independent 2-group t-test t.test(y~x) # where y is numeric and x is a binary factor

```
# independent 2-group t-test
t.test(y1,y2) # where y1 and y2 are numeric

# paired t-test
t.test(y1,y2,paired=TRUE) # where y1 & y2 are numeric

# one samle t-test t.test(y,mu=3)
# Ho: mu=3
```

We can use the **var.equal = TRUE** option to specify equal variances and a pooled variance estimate. We can use the **alternative="less"** or **alternative="greater"** option to specify a onetailed test.

Nonparametric and resampling alternatives to t-tests are available.

### Visualizing Results

Use box plots or density plots to visualize group differences.

### Nonparametric Tests of Group Differences

**R** provides functions for carrying out Mann-Whitney U, Wilcoxon Signed Rank, Kruskal Wallis,and Friedman tests.# independent 2-group Mann-Whitney U Test

```
wilcox.test(y~A)
# where y is numeric and A is A binary factor

# independent 2-group Mann-Whitney U Test
wilcox.test(y,x) # where y and x are numeric

# dependent 2-group Wilcoxon Signed Rank Test
wilcox.test(y1,y2,paired=TRUE) # where y1 and y2 are numeric

# Kruskal Wallis Test One Way Anova by Ranks
kruskal.test(y~A) # where y1 is numeric and A is a factor
```

```
# Randomized Block Design - Friedman Test
friedman.test(y~A|B)
# where y are the data values, A is a grouping factor# and B
is a blocking factor
```

For the wilcox.test we can use the **alternative="less"** or **alternative="greater"** option tospecify a one tailed test.

Parametric and resampling alternatives are available.

The package npmc provides nonparametric multiple comparisons.

```
library(npmc)
npmc(x)
# where x is a dataframe containing variable 'var'
# (response variable) and 'class' (grouping variable)
```

## Visualizing Results

Use box plots or density plots to visual group differences.

## Multiple (Linear) Regression

**R** provides comprehensive support for multiple linear regression. The topics below are providedin order of increasing complexity.

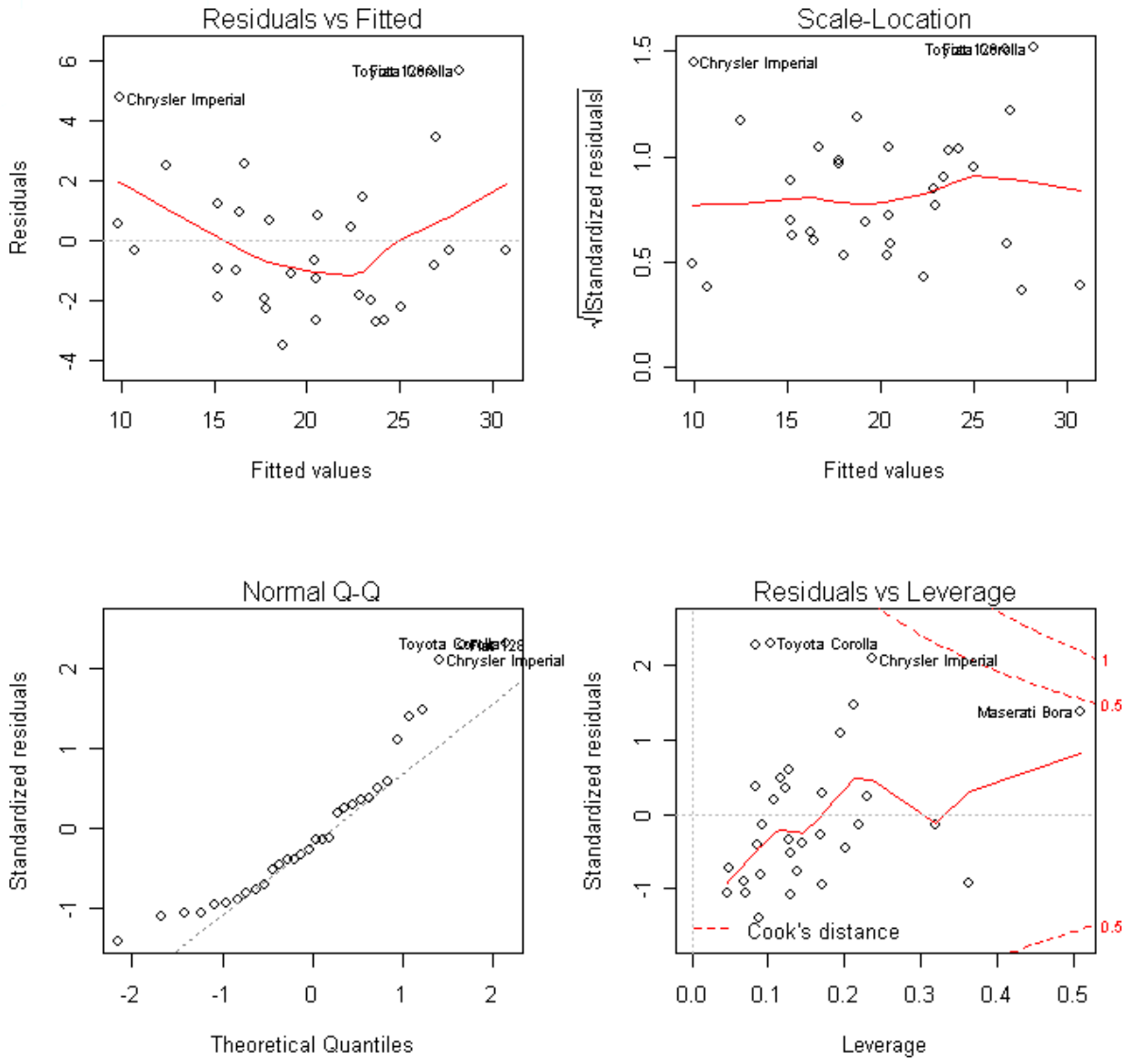
## Fitting the Model

```
# Multiple Linear Regression Example fit <-
lm(y ~ x1 + x2 + x3, data=mydata)summary(fit)
# show results

# Other useful functions coefficients(fit) #
model coefficients
confint(fit, level=0.95) # CIs for model parameters
fitted(fit) # predicted values
residuals(fit) # residuals
anova(fit) # anova table
vcov(fit) # covariance matrix for model parameters
influence(fit) # regression diagnosticsDiagnostic Plots
```

Diagnostic plots provide checks for heteroscedasticity, normality, and influential observerations

```
# diagnostic plots
layout(matrix(c(1,2,3,4),2,2)) # optional 4 graphs/pageplot(fit)
```




For a more comprehensive evaluation of model fit see regression diagnostics.

## Comparing Models

We can compare nested models with the anova( ) function. The following code provides asimultaneous test that x3 and x4 add to linear prediction above and beyond x1 and x2.

```
# compare models
fit1 <- lm(y ~ x1 + x2 + x3 + x4, data=mydata)
```

```
fit2 <- lm(y ~ x1 + x2)
anova(fit1, fit2)
```

## Cross Validation

We can do K-Fold cross-validation using the **cv.lm( )** function in the DAAG package.# K-fold cross-validation

```
library(DAAG)
cv.lm(df=mydata, fit, m=3) # 3 fold cross-validation
```

Sum the MSE for each fold, divide by the number of observations, and take the square root to getthe cross-validated standard error of estimate.

We can assess **R2 shrinkage** via K-fold cross-validation. Using the **crossval()** function from thebootstrap package, do the following:

```
# Assessing R2 shrinkage using 10-Fold Cross-Validationfit <-
lm(y~x1+x2+x3,data=mydata)
library(bootstrap)     #
define functions
theta.fit <- function(x,y){lsfit(x,y)}
theta.predict <- function(fit,x){cbind(1,x)%*%fit$coef}# matrix
of predictors
X   <-   as.matrix(mydata[c("x1","x2","x3")])   #
vector of predicted values
y <- as.matrix(mydata[c("y")])

results  <-  crossval(X,y,theta.fit,theta.predict,ngroup=10) cor(y,
fit$fitted.values)**2 # raw R2 cor(y,results$cv.fit)**2 # cross-
validated R2
```

## Variable Selection

Selecting a subset of predictor variables from a larger set (e.g., stepwise selection) is a controversial topic. We can perform stepwise selection (forward, backward, both) using the **stepAIC( )** function from the **MASS** package. **stepAIC( )** performs stepwise model selection byexact AIC.

```
#    Stepwise    Regression
library(MASS)
fit <- lm(y~x1+x2+x3,data=mydata) step <-
stepAIC(fit, direction="both")step$anova #
display results
```

Alternatively, we can perform all-subsets regression using the **leaps( )** function from the leaps package. In the following code nbest indicates the number of subsets of each size to report. Here,the ten best models will be reported for each subset size (1 predictor, 2 predictors, etc.).

```
# All Subsets Regression
library(leaps) attach(mydata)
leaps<-regsubsets(y~x1+x2+x3+x4,data=mydata,nbest=10)

# view results
summary(leaps)
# plot a table of models showing variables in each model.# models
are ordered by the selection statistic. plot(leaps,scale="r2")
# plot statistic by subset size
library(car)
subsets(leaps, statistic="rsq")
```

## Creating/editing data objects

```
> y<-c(1,2,3,4,5);y
[1] 1 2 3 4 5
```

If we want to modify the data object, use *edit()* function and assign it to an object. For example,the following command opens R Editor for editing.

```
> y<-edit(y)
```

If we prefer entering the data.frame in a spreadsheet style data editor, the following commandinvokes the built-in editor with an empty spreadsheet.

```
> data1<-edit(data.frame())
```

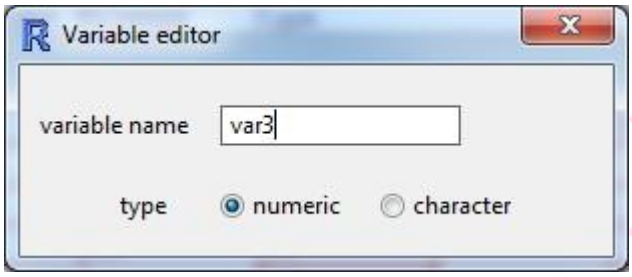
After entering a few data points, it looks like this:

RGui (64-bit) - [Data Editor]

File Windows Edit Help

| | var1 | var2 | var3 | var4 | var5 | var6 |
|---|---|---|---|---|---|---|
| 1 | 1 | aa | 100 | 0.234 | | |
| 2 | 2 | bb | 200 | 0.539 | | |
| 3 | 3 | cc | 300 | 0.625 | | |
| 4 | 4 | dd | 400 | 0.719 | | |
| 5 | | | | | | |
| 6 | | | | | | |
| 7 | | | | | | |
| 8 | | | | | | |
| 9 | | | | | | |
| 10 | | | | | | |
| 11 | | | | | | |
| 12 | | | | | | |
| 13 | | | | | | |
| 14 | | | | | | |
| 15 | | | | | | |
| 16 | | | | | | |
| 17 | | | | | | |
| 18 | | | | | | |
| 19 | | | | | | |

We can also change the variable name by clicking once on the cell containing it. Doing so opensa dialog box:




When finished, click  in the upper right corner of the dialog box to return to the Data Editor window. Close the Data Editor to return to the R command window (R Console). Check the resultby typing:

```
> data1
```

## Reading data from files

When data files are large, it is better to read data from external files rather than entering data through the keyboard. To read data from an external file directly, the external file should be arranged properly.

The first line of the file should have a name for each variable. Each additional line of the file hasthe values for each variable.

### Input file form with names and row labels:

| Price | Floor | Area | Rooms | Age | isNew |
|---|---|---|---|---|---|
| 52.00 | 111.0 | 830 | 5 | 6.2 | no |
| 54.75 | 128.0 | 710 | 5 | 7.5 | no |
| 57.50 | 101.0 | 1000 | 5 | 4.2 | yes |
| 57.50 | 131.0 | 690 | 6 | 8.8 | no |
| 59.75 | 93.0 | 900 | 5 | 1.9 | yes |

...

By default numeric items (except row labels) are read as numeric variables and non-numericvariables, such as isNew in the example, as factors. This can be changed if necessary.
The function *read.table()* can then be used to read the data frame directly

```
> HousePrice <- read.table("d:/houses.data", header = TRUE)
```

### Reading comma delimited data

The following commands can be used for reading comma delimited data into R.

*read.csv(filename)* This command reads a .CSV file into R. We need to specify the exact filename with path.

*read.csv(file.choose())* This command reads a .CSV file but the *file.choose()* part opens up an explorer type window that allows us to select a file from our computer. By default, R will take the first row as the variable names.

*read.csv(file.choose(), header=T)* This reads a .CSV file, allowing us to select the file, the header is set explicitly. If we change to header=F then the first row will be treated like the rest of the data and not as a label.

### Storing variable names

Through *read.csv()* or *read.table()* functions, data along with variable labels is read into R memory. However, to read the variables' names directly into R, one should use *attach(dataset)* function. For example,
>attach(HousePrice)
causes R to directly read all the variables' names eg. Price, Floor, Area etc. it is a good practice touse the *attach(datafile)* function immediately after reading the *datafile* into R.

### Packages

All R functions and datasets are stored in packages. The contents of a package are available only when the package is loaded. This is done to run the codes efficiently without much memory usage.To see which packages are installed at our machine, use the command

```
> library()
```

To load a particular package, use a command like

```
> library(forecast)
```

Users connected to the Internet can use the *install.packages()* and *update.packages()* functions toinstall and update packages. Use *search()* to display the list of packages that are loaded.

### Standard packages

The standard (or base) packages are considered part of the R source code. They contain the basic functions those allow R to work with the datasets and standard statistical and graphical functions.They should be automatically available in any R installation.

### Contributed packages and CRAN

There are a number of contributed packages for R, written by many authors. Various packages deal with various analyses. Most of the packages are available for download from CRAN (https://cran.r-project.org/web/packages/), and other repositories such as Bioconductor (http://www.bioconductor.org/). The collection of available packages changes frequently. As on December 11, 2018, the CRAN package repository contains 13528 available packages.

### Getting Help

Complete help files in HTML and PDF forms are available in R. To get help on a particular command/function etc., type *help (command name)*. For example, to get help on function 'mean',type *help(mean)* as shown below

```
> help(mean)
```

This will open the help file with the page containing the description of the function mean. Anotherway to get help is to use "?" followed by function name. For example,

```
>?mean
```

will open the same window again.

In this lecture note, all R commands and corresponding outputs are given in Courier New font to differentiate from the normal texts. Since R is case-sensitive, i.e. typing *Help(mean)*, would generate an error message,

```
> Help(mean)
```

```
Error in Help(mean) : could not find function "Help"
```

## Further Readings

Various documents are available in https://cran.r-project.org/manuals.html from beginners' levelto most advanced level. The following manuals are available in pdf form:

1. An Introduction to R
2. R Data Import/Export
3. R Installation and Administration
4. Writing R Extensions
5. The R language definition
6. R Internals
7. The R Reference Index

## References


1. https://www.cran.r-project.org/
2. http://www.gardenersown.co.uk/Education/Lectures/R/index.htm
3. Matloff, N. (2009). *The Art of R Programming*.
4. Quick R. http://www.statmethods.net/index.html
5. W. N. Venables, D. M. Smith and the R Development Core Team. *An Introduction to R. Notes on R: A Programming Environment for Data Analysis and Graphics Version 2.9.1 (2009-06-26)*.

# An Introduction to Stata Software for Statistical Analysis

**Nithyashree M L**

*Scientist, Division of Agricultural Economics, ICAR-IARI, New Delhi*

This chapter describes an overview and basic commands used in the Stata software which will help the beginners to get familiar and hold a grip in using the software for further statistical analysis. Those who don't have access to Stata software can avail the short-term student license service by using the link https://www.stata.com/customer-service/short-term-license/. As shown below the main interphase of the Stata has a menu system that enables the users to perform the task by interactive UI; drop-down menu. Alternatively, there is also a command window with which users can write commands directly for the statistical analysis. For example, to summarize data, by typing summarize variable/variables name, we can get the result and one can also do it by using the drop-down menu by going to current statistics (Figure 2). As a beginner one can start by exploring the possible option available by using the drop-down menu, after getting comfortable with the software, can directly type the command to perform any required analysis which is more efficient and professional, apart from this the executed commands can be copied from the history/review window and saved in do file, which can be easily shared with other researchers and this enables single-click execution. With the right side of the interphase, variables, and properties window variables and their related properties including labelling the variables and assigning variables values can be performed and visualized.

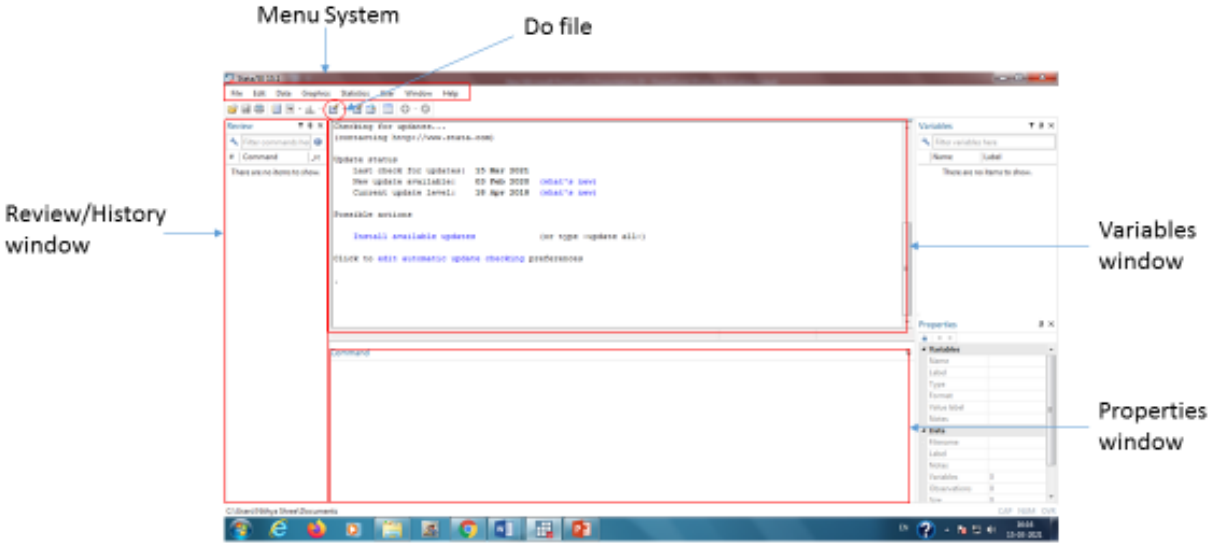



Figure 1. Main interphase of Stata

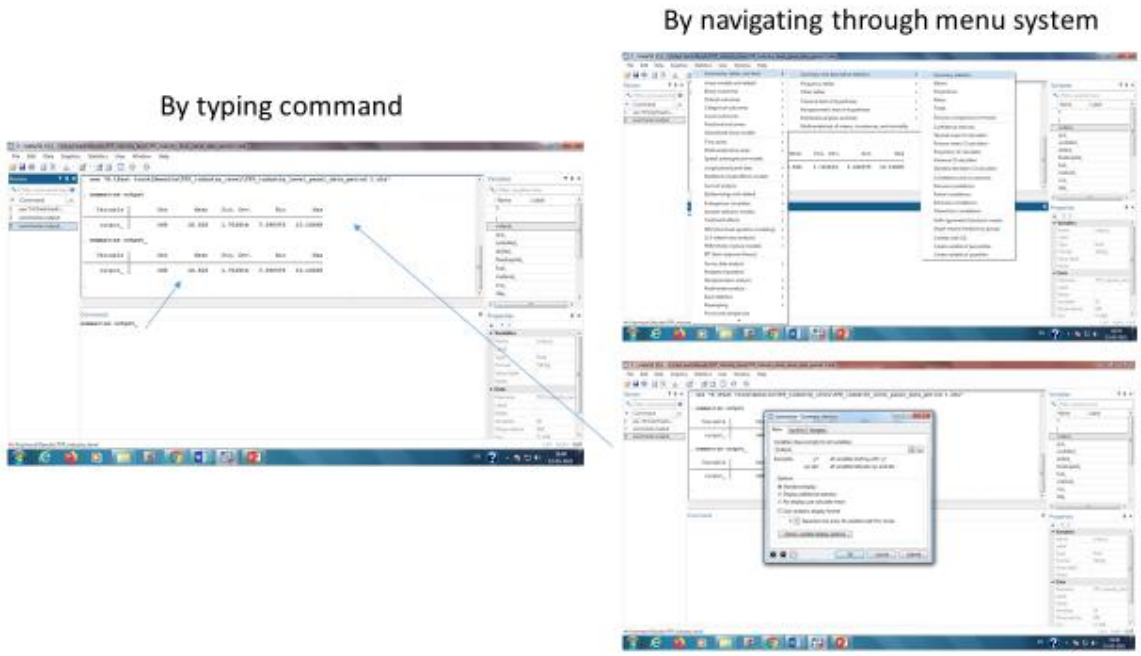
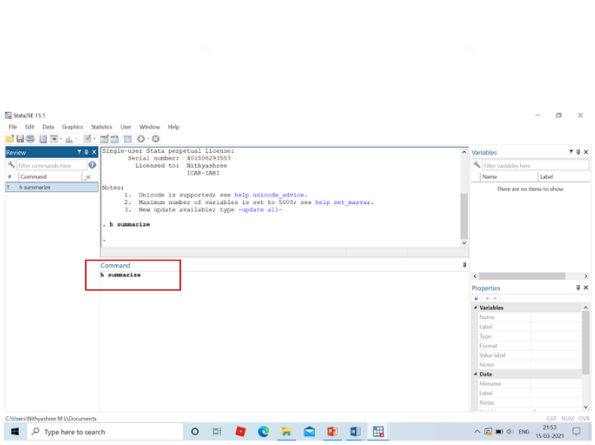


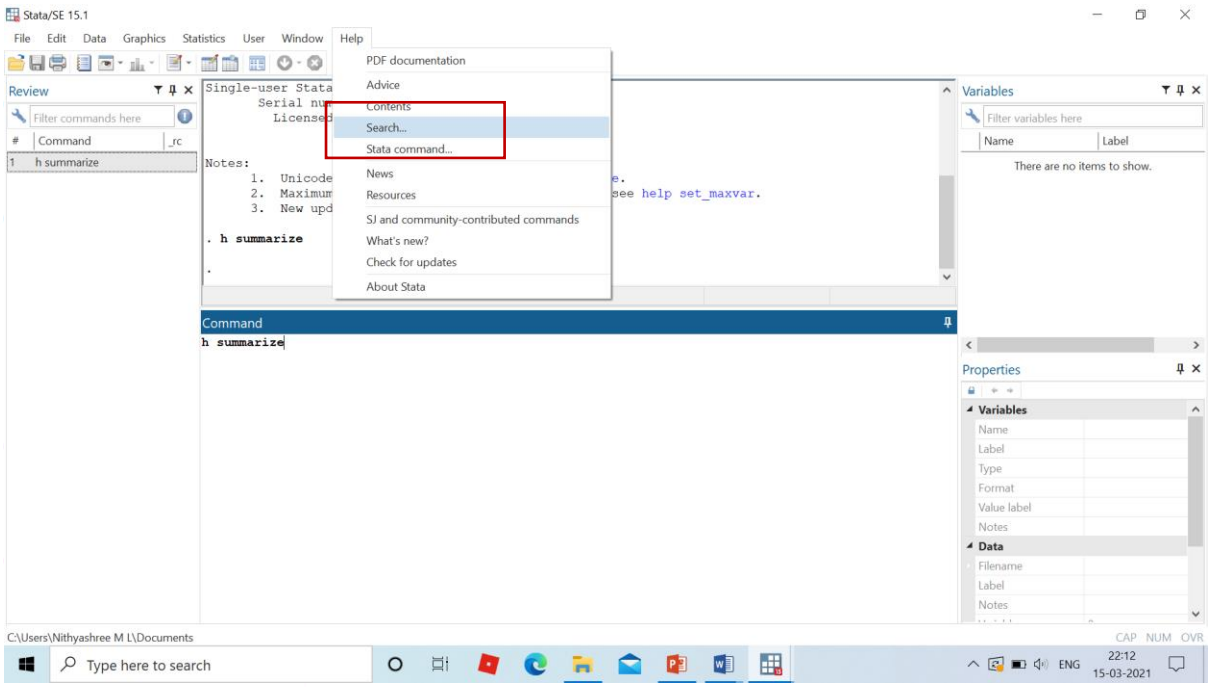

Figure 2. Two ways of performing task in Stata

Stata provides flexibility in exploring the various option to the users to learn the available functions and packages with the use of the help command. By using this command one can learn and use the applicability of different functions and packages. For example, by typing help summarize in the command window we get detailed information of the command summarize in the form of pdf also by scrolling down detailed information on using the summarize command directly as syntax as well as menu format along with some example data sets a shown below is a great source to learn Stata.




If the user is not sure about what to type in the command window, then the search option under the help in the menu bar can be explored. Besides there are many user-written commands are available in the Stata software in the form of packages, to use them first we need to install the packages. For this one

can use findit command and after finding the suitable package install them to make use of those packages and also make sure your system connected to the internet while installing the packages.




It is always good practice to save the results of the analysis, that can be done in Stata by typing log using filename in the command window before beginning the analysis, which creates a log file by the given file name and after completing the analysis type log close and your analysis will be saved in the smcl format for example filename.smcl, which details all the activity which you carried out during the particular session or analysis. Which you may need at the later stage to communicate to the journals while submitting the research article.

Mathematical and logical operator in Stata are similar to those used in MS excel

| | |
|---|---|
| a == b | if a equals b |
| a != b | if a not equal to b |
| a > b | if a greater than b |
| a >= b | if a greater than or equal to b |
| a < b | if a less than b |
| a <= b | if a less than or equal to b |
| a & b | **&** refers to **and** |
| a \| b | \| refers to **or** |

**Handling the data set**

1. **Creating a new variable:** gen or egen command is used to create the new variable. These commands can be combined with arithmetic operators or logical operators

   *Example: gen grade=1 if marks==2*
   *egen stdev_age= std(income)*
   *gen ln_wage = ln(wage)*

2. **For labelling the variable:** to create a label to the variable, write the command lable variable and type the variable name need to be labelled then write the label with in the invited comma

   *Example: label variable ln_wage "Log of hourly wage"*

3. **The replace command:** Replace command generally helps in editing value of already existing or generated variable, for this command ***replace*** can be used.

   *Example: replace gender=0 if missing(gender)*
   *replace gender=0 if gender==.*

4. To sort the data in ascending order: use command ***sort variable name***

5. Command that can convert string to numeric variables: ***Tostring and destring***

**Few Commands for Basic Statistical analysis**

a. Regression: *reg or probit or logit*

b. Correlation: *corr or pwcorr*

c. Student T test: *ttest*

d. Factor analysis: *factor*



e. Marginal Effects after probit or logit: *mfx*

f. Chisquare test: *tab, all*

g. Principal Component Analysis: *PCA*

**Few useful user written commands**

h. *tatable2*: Calculate group wise mean value and test the significance

i. *orth_out*: Perform t-test for any number of variables at once

j. *dea*: Data envelopment analysis

k. *acfest*: Production function est. using Ackerberg-Caves-Frazer method

l. *levpet*: Production function est. using Levinsohn and Petrin approach

m. doubleb: Perform Double Bound Contingent Valuation.

n. clustersampsi: perform power calculations for RCT

**Exercise- 1**

1. Import the dataset “stata_intro”, which has been shared with you already.

2. Summarize the data so that you see the means, standard deviation, min, and max of each variable

3. generate the logarithmic transformation of the variable wage as ln_wage

4. label variable ln_wage, wage, collgrad and union as Log of hourly wage, Hourly wage, College graduate and Union member respectively.

5. define label values for the variable collgrad

6. encode racecat as race

7. obtain mean wage sd wage for the variable union

9. draw histogram for the variable ln_wage and superimpose normal curve

10. generate scatter plot for the variables wage tenure

11. generate scatter plot for the variables wage tenure by union

12. obtain a liner regression equation of ln_wage and tenure

13. obtain a liner regression equation ln_wage and tenure along with by considering any one of the categorical variable and also an interaction component

14. save the commands in do file and results in log file.

# Evidence Synthesis using Systematic Reviews and Meta-Analysis

**Praveen KV**

*Scientist, Division of Agricultural Economics, ICAR-IARI, New Delhi*

*" Evidence synthesis is a type of research method that allows researchers to bring together all relevant information on a research question. This can be useful to identify gaps in knowledge, establish an evidence base for best-practice guidance, or help inform policymakers and practitioners. There are many types of outputs that use evidence synthesis, such as policy briefs, systematic reviews, clinical practice guidelines and so on".* (Centre for Evaluation).

Systematic reviews are a sort of literature review that utilizes systematic methods to gather secondary data and blend or synthesise the research evidence qualitatively or quantitatively. With the volume of research evidence on any topic growing at an ever-expanding rate, it is very difficult for individual researchers or policymakers to survey this tremendous amount of literature and arrive at the best decision on its basis. Following a systematic approach, systematic reviews help summarize the research knowledge on an intervention. It endeavours to gather all the empirical evidence that fits pre-determined eligibility criteria to answer to a particular research question. It utilizes systematic techniques that are chosen with the end goal of minimizing bias and hence giving more dependable findings from which conclusions can be drawn and choices made (Antman et al 1992, Oxman and Guyatt 1993).

**Research questions**

As in the case of any research, the first and most significant choice in setting up a systematic review is to decide its core interest. This is best done by framing the questions that the review looks to answer. Well-formulated questions will guide the systematic review procedure, including deciding eligibility criteria, literature search, gathering data from selected publications, organizing and presenting the findings (Cooper 1984, Hedges 1994, Oliver et al 2017). The FINER standards have been proposed to make life easy for a researcher while creating research questions. As per this strategy, questions ought to be Feasible, Interesting, Novel, Ethical, and Relevant (Cummings et al 2007). These measures raise key issues to be considered at the start of the review and ought to be borne as a primary concern when questions are framed.

A systematic review can address any research question that can be answered by primary research. Studies that compare interventions utilize the outcome of the participants to arrive at the impacts of various interventions. Statistical synthesis (for example meta-analysis) centres on comparison of a new intervention with the control. The differentiation between the outcomes of two groups treated

contrastingly is known as the ‘effect’ or the 'treatment effect’. The primary objective of systematic reviews should be ideally framed in a single sentence. The objective can be structured as: 'To evaluate the impacts of [intervention or technology] for [income enhancement] in [types of individuals, region, and setting if specified]'. This may be trailed by at least one secondary targets, for instance identifying with various participant groups, varying comparison of interventions or diverse outcome measures. The detailing of review question(s) requires thought of a few key segments (Richardson et al 1995, Counsell 1997) which can be conceptualized by the 'PICO', an abbreviation for Population, Intervention, Comparison(s) and Outcome. The scope of the review should be just apt. It should not be too broad or narrow to be relevant.

**Table 1. PICO formulation**

| Sl No | Item | Example |
|---|---|---|
| 1 | Population | Farmers in developing countries |
| | | Farmers involved in farmer groups or producer companies |
| 2 | Intervention | GM crops |
| | | Integrated Pest Management |
| 3 | Comparator | Communities/famers not participating in FFS |
| | | Farmers/communities receiving alternative interventions |
| 4 | Outcome | Yield |
| | | Net revenue |

**Defining inclusion criteria**

One of the highlights that differentiate a systematic review from a narrative review is that the authors of systematic review ought to pre-indicate criteria and standards for including and barring individual studies. When building up the protocol, one of the initial steps is to decide the components of the review question (the population, intervention(s), comparator(s) and outcome, or PICO components) and how the intervention, in the identified population, creates the outcomes. Eligibility criteria depend on the PICO components in addition to a specification of the kinds of studies that have addressed these inquiries. The population, intervention, and comparators in the review question can be usually translated into the inclusion criteria, but not always directly.



**Literature search and study selection**

Systematic reviews require a careful, objective, and reproducible search of a variety of sources to extract as many studies (eligible) as possible. The quest for studies should be as broad as possible to diminish the danger of reporting bias and to identify maximum evidence as possible. Database determination ought to be guided by the survey theme. 'Grey literature' should also be considered. Authors ought to search for dissertations and conference abstracts also. They should also think about looking through the web, hand searching of journals and looking through full texts of journals electronically where accessible. They ought to inspect past reviews on a similar theme and check reference lists of included studies. Suitable search strategy should be formulated for searching in different databases. Choices about which studies to include for a review is among the most compelling choices that are made in the review procedure and they include judgment. Involvements of at least two individuals, working independently, are required to decide if each study meets the qualification standards. A PRISMA flow chart mentioning the selection of studies at each stage should be included in the report.

**Table 2. List of databases to search**

| Sl No. | Database |
|---|---|
| 1 | Web of Science (Social science citation index) |
| 2 | CeRA |
| 3 | Google scholar |
| 4 | AgEcon search |
| 5 | Econlit |
| 6 | CAB abstract |
| 7 | Medline, Pubmed |
| 8 | ERIC |

**Coding and Data collection**

Authors are urged to create layouts of tables and figures that will show up in the review to encourage the design of data collection forms. The way to effective data collection is to build simple-to-use forms and gather adequate and unambiguous information that present the source in an organized and structured way. Effort ought to be made to collect information required for meta-analysis. Data ought to be

gathered and documented in a structure that permits future access and data sharing. Coding should provide for adding data in the following components:

- Study identification
- Intervention discriptives
- Process and implementation
- Context
- Popultion characteristics
- Research methods
- Effect size data
- Outcomes
- Subgroups

**Effect measures**

The kinds of outcome data that authors are probably going to experience are dichotomous data, continuous data, ordinal data, count or rate data and time-to-event data. The nature of the collected data determines the effect measures of intervention. Effect measures are statistical constructs that compare outcome data between two intervention groups. It is mainly of two distributed into two categories: ratio measures and difference measures. Estimates of effect describe the size of the intervention effect in terms of how diverse the outcome data were between the groups. For ratio effect measures, 1 indicates no distinction between the groups, while for difference measures, 0 indicates no distinction between the groups. Larger and smaller values than these 'null' values may suggest either benefit or harm of an intervention. The true effects of interventions very difficult to arrive at, and can only be assessed by the available studies. Estimates should thus be presented with uncertainty measures like confidence interval or standard error (SE). Examples of effect measures of dichotomous outcome data: Risk ratio, Odds ratio, Risk difference. Examples of effect measures of continuous outcome data: Mean difference, Standardised mean difference, Ratio of means

**Meta-analysis**

Meta-analysis can be considered as a key step in a systematic review. Meta-analysis involves deciding on the possibility of combining the results of selected studies. This procedure results in an overall statistic with a confidence interval that summarizes the effect of an intervention compared with the counterfactual. Meta-analysis is useful since they improve precision by including more information that smaller individual studies lack. To carry out a meta-analysis, at first, a summary statistic is computed for individual studies, to present the effect of the intervention in a uniform measure. Next, the individual study's intervention effects are statistically combined using a weighted average of the intervention effects estimated in the individual studies. Undertake random-effects meta-analysis if the studies are

not all estimating the same intervention effect, but estimate intervention effects that follow a distribution across studies. On the other hand, if each study is estimating the same quantity, then a fixed-effect meta-analysis can be used. A confidence interval is derived that represents the precision of the summarized estimate. Meta-analysis can be carried out using two models:

- **Fixed effect model**
  - Under the fixed-effect model we assume that all studies in the meta-analysis share a common (true) effect size.
  - Put another way, all factors that could influence the effect size are the same in all the studies, and therefore the true effect size is the same in all the studies.
  - Since all studies share the same true effect, it follows that the observed effect size varies from one study to the next only because of the random error inherent in each study.
  - If each study had an infinite sample size the sampling error would be zero and the observed effect for each study would be the same as the true effect.
  - In practice, of course, the sample size in each study is not infinite, and so there is sampling error and the effect observed in the study is not the same as the true effect.
  - The observed effect for any study is given by the population mean plus the sampling error in that study.

- **Random effects model**
  - There is no reason to assume that studies are identical in the sense that the true effect size is exactly the same in all the studies.
  - We might not have assessed these covariates in each study.
  - If each study had an infinite sample size the sampling error would be zero and the observed effect for each study would be the same as the true effect for that study.
  - The sample size in any study is not infinite and therefore the sampling error is not zero. The observed effect for that study will be less than or greater than the true effect because of sampling error.
  - The distance between the overall mean and
  - the observed effect in any given study consists of two distinct parts: true variation in effect sizes (i) and sampling error
  - The observed effect for any study is given by the grand mean, the deviation of the study's true effect from the grand mean, and the deviation of the study's observed effect from the study's true effect.

**Meta-analysis: Demonstration (Example of meta-analysis of biofertilizer in India)**



***Setting the question***

- The effects of biofertilizer use in crop yields in India
- PICO
  - P- Experimental plots with biofertilizer application
  - I- Biofertilizer
  - C- Control plots
  - O- Yield

***Search strategy for meta-analysis***

***Search strategy for meta-analysis***

A comprehensive literature search was undertaken from February to April 2019 in the google scholar, and CeRA (Consortium for e-resources in agriculture) to identify the studies to be included in the meta-analysis. The studies published between 2000 and 2019 were searched using the following search strings: "biofertilizer", "biofertiliser", "biofertilizer OR biofertiliser" AND "response" AND "India".

***Screening, coding and data extraction***

The studies were screened independently by authors to select the ones that meet the criteria to be included for the meta-analysis. The studies based on field trials, and that provide data for pairwise comparison of the yield effect of biofertilizer treated crop to that of the control are included. Full papers were reviewed to record the data on mean yields, standard deviations and the number of replications, and also other field-specific observations that would be required for the analysis. Out of the 16700 studies that appeared during the literature search, only 236 were selected after the preliminary screening to remove studies based on biofertilizer production technology, studies that are carried out in other countries, review studies, studies dealing with regulation and policies, and other aspects of biofertilizers that are not of our interest (these are termed as ‘exclusion criteria). After removing the duplicate studies and the ones based on laboratory experiments, 86 studies were selected for full-text reviews. From this, only 18 articles were finally selected for the meta-analysis, as the others did not provide the information that we require for meta-analysis.

The flow of the search process is given in detail in the Preferred Reporting Items for Systematic Reviews and Meta-Analyses(PRISMA) flow chart given in figure . The data from all the selected studies were then extracted and classified on the basis of types of biofertilizer. Nitrogen-fixing, phosphate solubilising, VAM, Combined biofertilizers, and others were the biofertilizer categories on the basis of which data extracted from the studies were grouped. Suitable predetermined codes were prepared in advance for this purpose. Example of coded sheet is given in the figure below. Further on the basis of crop groups, data were classified into that of cereals, legumes, vegetables and oilseeds. Thus from the 18 studies selected for meta-analysis, we were able to carry out 38 pairwise comparisons between biofertilizer treatment and control.

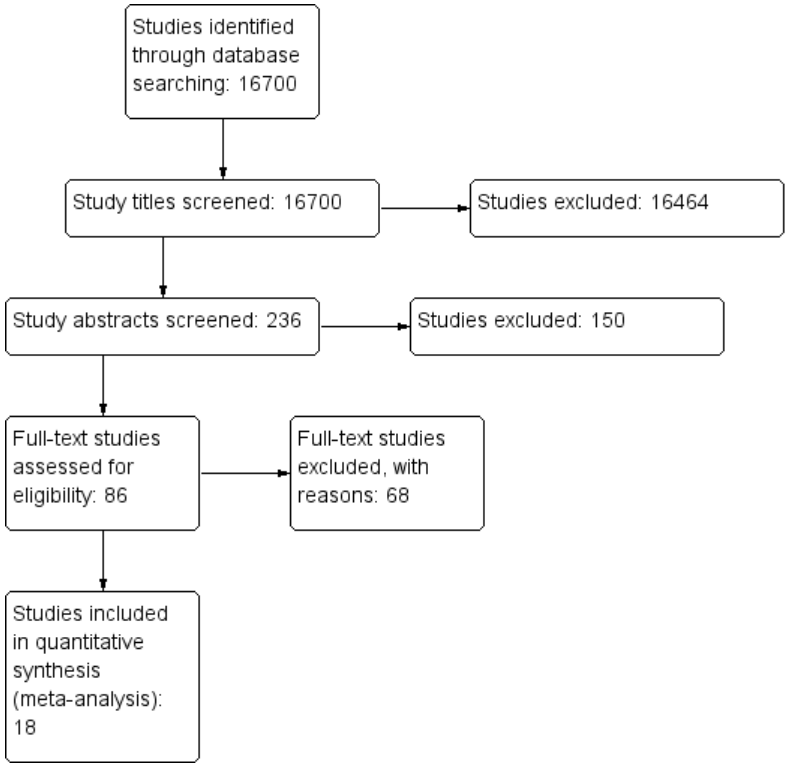



**Figure: PRISMA flow chart**

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z | AA |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Paper no | Author | Effect size | Year | Crop | Location | Agro-ecological sub region (ICAR) | No of years of experiment | Soil | pH | Organic carbon % | Available N (kg/ha) | Available P(kg/ha) | Available K(kg/ha) | Biofertilizer species | Yield treatment (tonnes/ha) | Yield Control (tonnes.ha) | SD trt | SD control | Applied N (kg/ha) | Applied P(kg/ha) | Applied K(kg/ha) | Total N (available+applied) kg/ha | Total P (available+applied) kg/ha | Total K (available+applied) kg/ha | No. of replications | se |
| 2 | 7 | Upadhyay et al | 1.62 | 2012 | Cabbage | Uttar Pradesh | Hot Sem | 2 | Sandy loa | 7.6 | 0.39 | 210.15 | 18.24 | 256.35 | Azospirillum | 41.22 | 36.63 | 2.62095 | 2.62095 | 150 | 60 | 80 | 360.15 | 78.24 | 336.35 | 6 | 1.07 |
| 3 | 7 | Upadhyay et al | 1.85 | 2012 | Cabbage | Uttar Pradesh | Hot Sem | 2 | Sandy loa | 7.6 | 0.39 | 210.15 | 18.24 | 256.35 | PSM | 41.88 | 36.63 | 2.62095 | 2.62095 | 150 | 60 | 80 | 360.15 | 78.24 | 336.35 | 6 | 1.07 |
| 4 | 7 | Upadhyay et al | 1.47 | 2012 | Cabbage | Uttar Pradesh | Hot Sem | 2 | Sandy loa | 7.6 | 0.39 | 210.15 | 18.24 | 256.35 | VAM | 40.81 | 36.63 | 2.62095 | 2.62095 | 150 | 60 | 80 | 360.15 | 78.24 | 336.35 | 6 | 1.07 |
| 5 | 8 | Yeptho | 0.32 | 2012 | Onion | Nagaland | Warm Pe | 2 | Sandy loa | 4.5 | 2 | 212.3 | 10.5 | 173.2 | Azotobacter | 17.03 | 16.74 | 0.8515 | 0.837 | 104 | 32 | 152 | 316.3 | 42.5 | 325.2 | 6 | 0.35 |
| 6 | 10 | Singh | 1.29 | 2000 | Potato | Meghalaya | Warm Pe | 3 | Sandy loa | 5.4 | 1.7 | 172 | 8.2 | 235 | Azotobacter | 17 | 15.9 | 0.85 | 0.795 | 112 | 0 | 0 | 284 | 8.2 | 235 | 12 | 0.25 |
| 7 | 10 | Singh | 2.39 | 2000 | Potato | Meghalaya | Warm Pe | 3 | Sandy loa | 5.4 | 1.7 | 172 | 8.2 | 235 | Phosphpbactrin | 18 | 15.9 | 0.9 | 0.795 | 112 | 0 | 0 | 284 | 8.2 | 235 | 12 | 0.26 |
| 8 | 11 | Gosh et al | 0.55 | 2000 | Potato | West Bengal | Hot Subh | 2 | sandy loa | 6.2 | 1.2 | 165 | 13 | 122.5 | Phosphert | 17.24 | 15.75 | 2.49848 | 2.49848 | 120 | 44.5 | 83.5 | 285 | 57.5 | 206 | 6 | 1.02 |
| 9 | 16 | Panwar | 2.97 | 2014 | Rice | Meghalaya | Warm Pe | 2 | Sandy loa | 4.9 | 2.06 | 261.2 | 5.5 | 219.7 | Azolla | 42.64 | 36.27 | 2.132 | 1.8135 | 80 | 60 | 40 | 341.2 | 65.5 | 259.7 | 6 | 0.87 |
| 10 | 16 | Panwar | 1.61 | 2014 | Rice | Meghalaya | Warm Pe | 2 | Sandy loa | 4.9 | 2.06 | 261.2 | 5.5 | 219.7 | Azospirillum | 30.38 | 27.85 | 1.519 | 1.3925 | 0 | 0 | 0 | 261.2 | 5.5 | 219.7 | 6 | 0.62 |
| 11 | 16 | Panwar | 5.42 | 2014 | Rice | Meghalaya | Warm Pe | 2 | Sandy loa | 4.9 | 2.06 | 261.2 | 5.5 | 219.7 | Azospirillum | 41.44 | 30.72 | 2.072 | 1.536 | 60 | 45 | 30 | 321.2 | 50.5 | 249.7 | 6 | 0.85 |
| 12 | 18 | Tagore et al | 1.52 | 2013 | Chickpea | Madhya Pradesh | Semi-aric | 1 | Clay loan | 7.8 | 0.45 | 204 | 9.58 | 576 | PSB | 1.7 | 1.5 | 0.10219 | 0.10219 | 0 | 0 | 0 | 204 | 9.58 | 576 | 3 | 0.06 |
| 13 | 18 | Tagore et al | 3.2 | 2013 | Chickpea | Madhya Pradesh | Semi-aric | 1 | Clay loan | 7.8 | 0.45 | 204 | 9.58 | 576 | Rhizobium | 1.9 | 1.5 | 0.10219 | 0.10219 | 0 | 0 | 0 | 204 | 9.58 | 576 | 3 | 0.06 |
| 14 | 30 | Kumar et al | 6.09 | 2009 | French b | Uttar Pradesh | Hot Sem | 2 | sandy loa | 7.2 | 0.43 | 197.02 | 23.41 | 210 | Biofertilizer | 1.83 | 1.59 | 0.0915 | 0.0795 | 0 | 0 | 0 | 197.02 | 23.41 | 210 | 6 | 0.04 |
| 15 | 31 | Kumawat et al | 1.6 | 2010 | Green gra | Rajasthan | Hot Arid | 1 | sandy loa | 8.2 | 0.3 | 78.8 | 16.3 | 180.4 | PSB | 0.64 | 0.56 | 0.03811 | 0.03811 | 0 | 0 | 0 | 78.8 | 16.3 | 180.4 | 3 | 0.02 |
| 16 | 31 | Kumawat et al | 2 | 2010 | Green gra | Rajasthan | Hot Arid | 1 | sandy loa | 8.2 | 0.3 | 78.8 | 16.3 | 180.4 | Rhizobium | 0.66 | 0.56 | 0.03811 | 0.03811 | 0 | 0 | 0 | 78.8 | 16.3 | 180.4 | 3 | 0.02 |
| 17 | 31 | Kumawat et al | 5 | 2010 | Green gra | Rajasthan | Hot Arid | 1 | sandy loa | 8.2 | 0.3 | 78.8 | 16.3 | 180.4 | Rhizobium+PSB | 0.81 | 0.56 | 0.03811 | 0.03811 | 0 | 0 | 0 | 78.8 | 16.3 | 180.4 | 3 | 0.02 |
| 18 | 33 | Singh et al | 1.76 | 2011 | Groundn | Meghalaya | Warm Pe | 2 | Sandy loa | 5 | 1.44 | 255.3 | 4.3 | 245 | PSB | 2.2 | 2 | 0.11 | 0.1 | 0 | 0 | 0 | 255.3 | 4.3 | 245 | 6 | 0.04 |
| 19 | 33 | Singh et al | 3.34 | 2011 | Groundn | Meghalaya | Warm Pe | 2 | Sandy loa | 5 | 1.44 | 255.3 | 4.3 | 245 | Rhizobium | 2.4 | 2 | 0.12 | 0.1 | 0 | 0 | 0 | 255.3 | 4.3 | 245 | 6 | 0.05 |
| 20 | 33 | Singh et al | 3.98 | 2011 | Groundn | Meghalaya | Warm Pe | 2 | Sandy loa | 5 | 1.44 | 255.3 | 4.3 | 245 | Rhizobium+PSB | 2.5 | 2 | 0.125 | 0.1 | 0 | 0 | 0 | 255.3 | 4.3 | 245 | 6 | 0.05 |
| 21 | 37 | Sharma et al | 0.11 | 2012 | Pigeon p | Karnataka | Hot arid | 3 | Clay loan | 8 | 0.5 | 180 | 25 | 350 | Biofertilizer | 0.014 | 0.013 | 0.009 | 0.009 | 25 | 50 | 0 | 205 | 75 | 350 | 9 | 0.00 |
| 22 | 39 | Majumdar et al | 2.27 | 2007 | Rice | Meghalaya | Warm Pe | 3 | Sandy loa | 4.6 | 1.85 | 222.5 | 4.5 | 180 | Azospirillum | 2.19 | 1.95 | 0.1095 | 0.0975 | 0 | 60 | 40 | 222.5 | 64.5 | 220 | 9 | 0.04 |
| 23 | 39 | Majumdar et al | 1.78 | 2007 | Rice | Meghalaya | Warm Pe | 3 | Sandy loa | 4.6 | 1.85 | 222.5 | 4.5 | 180 | Azospirillum | 3.38 | 3.08 | 0.169 | 0.154 | 60 | 60 | 40 | 282.5 | 64.5 | 220 | 9 | 0.06 |
| 24 | 39 | Majumdar et al | 3.03 | 2007 | Rice | Meghalaya | Warm Pe | 3 | Sandy loa | 4.6 | 1.85 | 222.5 | 4.5 | 180 | Azotobacter | 2.27 | 1.95 | 0.1135 | 0.0975 | 0 | 60 | 40 | 222.5 | 64.5 | 220 | 9 | 0.04 |
| 25 | 39 | Majumdar et al | 2.87 | 2007 | Rice | Meghalaya | Warm Pe | 3 | Sandy loa | 4.6 | 1.85 | 222.5 | 4.5 | 180 | Azotobacter | 3.58 | 3.08 | 0.179 | 0.154 | 60 | 60 | 40 | 282.5 | 64.5 | 220 | 9 | 0.06 |
| 26 | 43 | Mathews et al | 1.52 | 2006 | Rice | Karnataka | Hot Hum | 1 | sandy loa | 4.55 | 0.69 | 281 | 8.2 | 79 | Azospirillum+PSB | 5.71 | 4.53 | 0.62354 | 0.62354 | 0 | 0 | 0 | 281 | 8.2 | 79 | 3 | 0.36 |
| 27 | 43 | Mathews et al | 0.62 | 2006 | Rice | Karnataka | Hot Hum | 1 | sandy loa | 4.55 | 0.69 | 281 | 8.2 | 79 | Azospirillum+PSB | 8.88 | 8.4 | 0.62354 | 0.62354 | 75 | 75 | 90 | 356 | 83.2 | 169 | 3 | 0.36 |
| 28 | 45 | Ghosh and Mohiuddin | 1.05 | 2000 | Sesame | West Bengal | Hot Subh | 2 | sandy loa | 6.1 | 1.2 | 165 | 20 | 165 | Bioplin | 1.04 | 0.87 | 0.14697 | 0.14697 | 50 | 25 | 25 | 235 | 45 | 190 | 6 | 0.06 |
| 29 | 45 | Ghosh and Mohiuddin | 0.99 | 2000 | Sesame | West Bengal | Hot Subh | 2 | sandy loa | 6.1 | 1.2 | 165 | 20 | 165 | Phosfert | 1.02 | 0.87 | 0.14697 | 0.14697 | 50 | 25 | 25 | 235 | 45 | 190 | 6 | 0.06 |
| 30 | 45 | Ghosh and Mohiuddin | 0.98 | 2000 | Sesame | West Bengal | Hot Subh | 2 | sandy loa | 6.1 | 1.2 | 165 | 20 | 165 | Vitormone | 1.03 | 0.87 | 0.14697 | 0.14697 | 50 | 25 | 25 | 235 | 45 | 190 | 6 | 0.06 |
| 31 | 50 | Ehrha and Flautarag | 0.45 | 2009 | Wheat | Madhya Pradesh | semi-arid | 3 | Clay loan | 8.2 | 0.51 | 204 | 9.58 | 576 | Azospirillum | 4.78 | 4.67 | 0.239 | 0.2335 | 60 | 13.1 | 16.7 | 264 | 22.68 | 592.7 | 12 | 0.07 |

**Figure: Coding**



## Meta-analysis

Mean difference was selected as the effect size. As per the results of the meta-analysis, application of biofertilizers resulted in an average yield increase of 0.36 tonnes per ha in India. The diamond shape gives the effect of subgroup and total biofertilizers. The size of the diamond shape gives the magnitude of the effect size and the edges represent the confidence interval (95% level). Meta-regerssion results suggest that only the combined biofertilizer application has a significant effect on yield improvement. The model, indicated significant yield increase due to biofertilizers in clay loam soil (in comparison to sandy loam), and soils with low K and high P content as well as low pH and low organic carbon content (in line with the findings of Schults, 2018). The variation in the performance of biofertilizers as per the agro-ecological conditions was also confirmed in this model. Most agro-ecological variables considered were significant. Among the crop groups, significant yield effects were detected in the case of cereals, legumes and vegetables (first model).

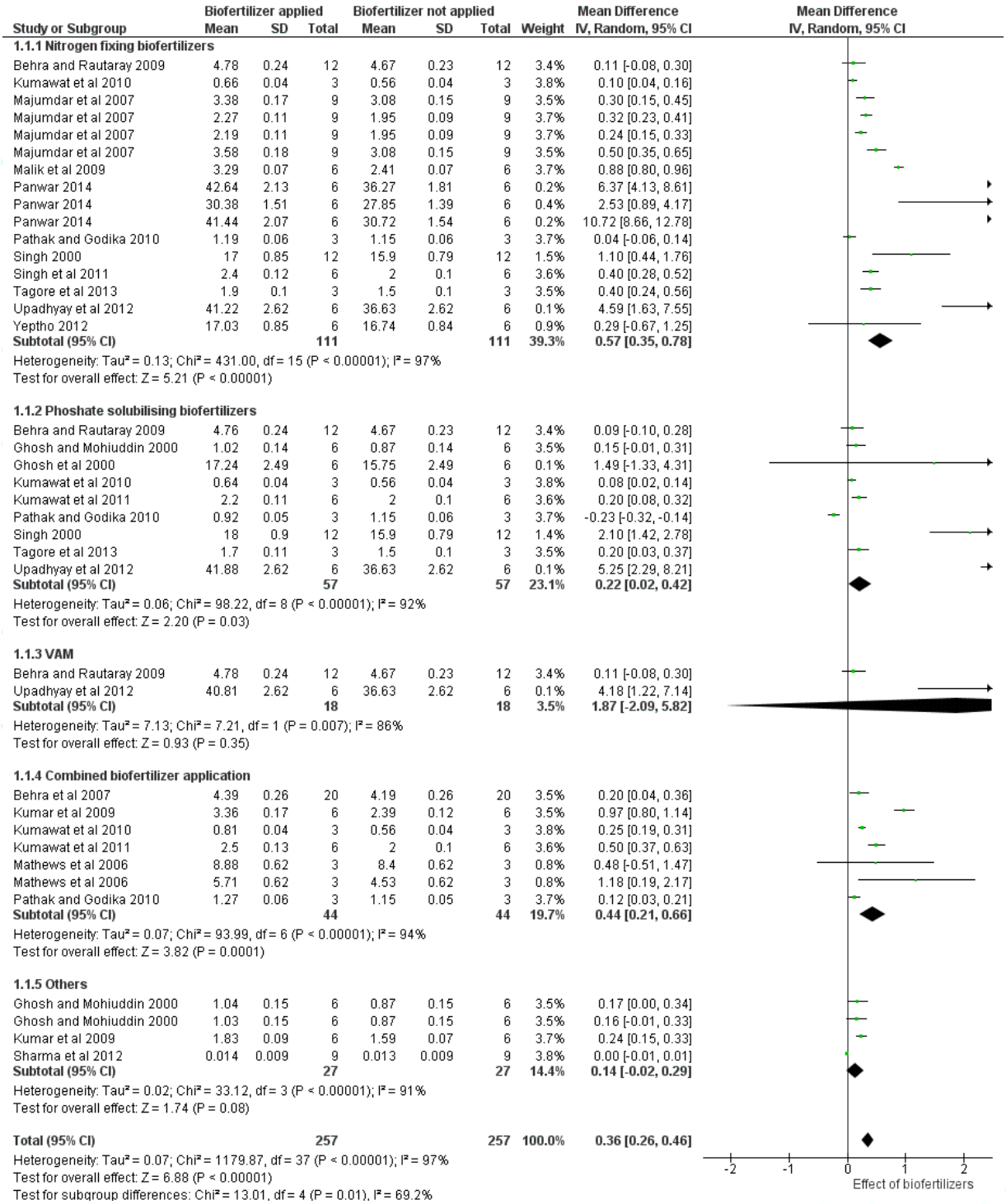



**Figure: Forest plot**

**Table: Meta-regression**

| Variables | Model with biofertilizer and agro-ecological groups | |
|---|---|---|
| | Coefficient | SE |
| Experiment duration | -6.99*** | 1.74 |
| Ph | -2.50** | 1.05 |
| organic carbon | -2.03 | 1.76 |
| Total N | 0.00 | 0.01 |
| Total P | 0.03* | 0.02 |
| Total K | -0.04*** | 0.01 |
| Replication number | 1.35*** | 0.36 |
| Clay loam | 33.72*** | 8.42 |
| VAM | -0.46 | 1.87 |
| Combined biofertilizers | 3.02** | 1.25 |
| Nitrogen fixers | 0.51 | 1.11 |
| Phosphate solubilizers | -0.38 | 1.01 |
| Hot arid eco region | 16.17** | 5.81 |
| Hot semi arid eco region | 21.56*** | 4.65 |
| Hot sub humid eco region | 14.01*** | 3.50 |
| Hot arid eco sub region | -8.67** | 3.25 |
| Northern plain | 28.35*** | 5.00 |
| Semi arid tropics | -1.18 | 1.51 |
| Warm perhumid eco region | 16.08*** | 3.66 |
| Hot arid eco sub region | 33.29*** | 6.84 |
| Constant | 16.54*** | 4.57 |
| Observations | 38 | |
| R-squared adjusted | 83.25 | |
| F statistic | 9.5 | |
| Tau-sq | 0.890 | |
| I-sq | 99.70 | |

**Improving the utility of evidence synthesis**

Policymakers are for sure a key stakeholder with regard to evidence synthesis. The best contribution of researchers in the field to the policymakers is probably an accurate, concise and unbiased synthesis of the available evidence. Four principles to improve the utility of evidence synthesis:

- Inclusivenes – Policymakers being the prime target audience (in most cases), it is ideal for them to get involved in the evidence synthesis process. If they are taking part in all the stages from the formulation of question, then through out the process and finally in the result interpretation, the quality and utility is bound to increase.

- Rigorous - No evidence shall be excluded for the convenience of the researchers. Evidence synthesis is a process that takes time, and warrants rigourous compilation of the available evidence from multiple sources.
- Transparent- Transparancy in the evidence synthesis process increases its trustworthiness. The exclusion and inclusion criteria in the entire procedure should be noted clearly in the reports so that others can understand how the researcher arrived at the final sample of evidence.
- Accessible – Improve accessibility through different forms of publications and distribution to the stakeholders. (Donnelly et al., 2018)

**Disclaimer:** This chapter of the manual is prepared referring the sources listed in the reference section. This is prepared only for distribution to the participants of the training. Kindly cite the original sources while quoting the contents of this chapter.

**References**

*Antman E, Lau J, Kupelnick B, Mosteller F, Chalmers T. A comparison of results of meta-analyses of randomized control trials and recommendations of clinical experts: treatment for myocardial infarction. JAMA 1992; 268: 240–248.*

*Centre for Evaluation. London School of Hygience and Tropical Medicine. https://www.lshtm.ac.uk/research/centres/centre-evaluation/evidence-synthesis.*

*Cooper H. The problem formulation stage. In: Cooper H, editor. Integrating Research: A Guide for Literature Reviews. Newbury Park (CA) USA: Sage Publications; 1984.*

*Counsell C. Formulating questions and locating primary studies for inclusion in systematic reviews. Annals of Internal Medicine 1997; 127: 380–387.*

*Cummings SR, Browner WS, Hulley SB. Conceiving the research question and developing the study plan. In: Hulley SB, Cummings SR, Browner WS, editors. Designing Clinical Research: An Epidemiological Approach. 4th ed. Philadelphia (PA): Lippincott Williams & Wilkins; 2007. p. 14–22.*

*Donnelly, Christl A., Ian Boyd, Philip Campbell, Claire Craig, Patrick Vallance, Mark Walport, Christopher JM Whitty, Emma Woods, and Chris Wormald. "Four principles for synthesizing evidence." 2018.*

*Hedges LV. Statistical considerations. In: Cooper H, Hedges LV, editors. The Handbook of Research Synthesis. New York (NY): USA: Russell Sage Foundation; 1994.*

*Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (editors). Cochrane Handbook for Systematic Reviews of Interventions version 6.0 (updated July 2019). Cochrane, 2019. Available from www.training.cochrane.org/handbook.*

*Oliver S, Dickson K, Bangpan M, Newman M. Getting started with a review. In: Gough D, Oliver S, Thomas J, editors. An Introduction to Systematic Reviews. London (UK): Sage Publications Ltd.; 2017.*

*Oxman A, Guyatt G. The science of reviewing research. Annals of the New York Academy of Sciences 1993; 703: 125–133.*

*Praveen KV, Singh A. Realizing the potential of a low-cost technology to enhance crop yields: evidence from a meta-analysis of biofertilizers in India. Agricultural Economics Research Review. 2019;32(conf):77-91.*

*Richardson WS, Wilson MC, Nishikawa J, Hayward RS. The well-built clinical question: a key to evidence-based decisions. ACP Journal Club 1995; 123: A12 –13*

.

# Sampling Techniques for Progrm Evaluation

**Girish Kumar Jha**

*Scientist, Division of Agricultural Economics, ICAR-IARI, New Delhi*

## INTRODUCTION

The main aim of a program evaluation study is to learn about how a population of interest is affected by the program. In experimental studies, establishing such causation is much easier as the experimenter has complete control on the experiment and one can keep all other things constant across groups except for the treatment. In such case, the counterfactual outcome is directly observed and impact (change in outcome due to treatment or cause is simply the difference in outcomes between units which received the treatment and the counterfactual, which didn't receive the treatment. But, in observational studies, the experimenter is a passive collector of the data and the counterfactual outcome is difficult to observe. The treatment and control groups vary not only with respect to treatment but also with respect to many other variables and hence it is difficult to attribute the difference in outcomes only to the treatment. This is also called as 'self-selection bias' where units with certain attributes tend to self-select themselves into either treatment or control groups. The root cause of the problem is lack of random allocation-if the units were to be allocated to treatment and control groups randomly, on an average the two groups will be similar to each other on all observable and unobservable characters except for treatment. However, very rarely policies are implemented by selecting beneficiaries randomly or technologies are given at random. The main aim of this lecture note is to provide an overview of sampling procedures which is necessary to construct a suitable sampling plan for achieving the goals of an impact evaluation study.

In sample surveys, population or an aggregate to represent the whole exists in its own way and we need to device the sampling techniques in such a way that the sample should represent the population. Here the objective is to infer about the population on the basis of a 'part' of the population which is known as a sample. Sampling is frequently used in everyday life in all kinds of investigations. Almost instinctively, before deciding to buy a lot, we examine a few articles preferably from different parts of the lot. Another example relates to a handful of grain taken from a sack to determine the quality of the grain. These are examples where inferences are drawn on the basis of the results obtained from a sample. Sampling is not always necessary. When a population is small, one may choose to collect the data for each and every unit belonging to the population and this procedure of obtaining information is termed as complete enumeration. However, the effort, money and time required for carrying out complete enumeration for large population, generally is extremely large. For example, suppose we are interested in assessing the impact of a programme on the adoption of a new technology by farmers in a region. Should we collect data about the adoption of the new technology from each and every farmer of the region or data from a sample of farmers will serve the purpose? In most of the cases, sample can provide sufficient evidence in form of data useful for the programme evaluation. A sample can also save valuable time, money and the labour of Extension professionals. Time is saved because only fewer (part of the population) farmers, households, etc. must be interviewed or surveyed, thus the complete set of data can be collected quickly. Money and labour are saved because less data must be collected. In addition, errors from handling the data (e.g. entering data into a computer file) are likely to be reduced because there are fewer opportunities to make mistakes.

CONCEPTS AND DEFINITIONS

**Sampling Unit**: Elementary units or groups of units which are clearly defined, identifiable, observable and convenient for sampling purposes are called sampling units.

**Sampling Frame**: The list or map of sampling units is called the sampling frame and provides the basis for the selection of units in the sample. The common problem of sampling frames are noncoverage or incomplete frame, foreign units, duplicate listings and cluster of elements combined into one listings etc.

**Population**: A population consists of a group of units defined according to the objectives of the survey. The population may consist of all the households in a village/ locality, all the fields under a particular crop. We may also consider a population of persons, families, fields, animals in a region, or a population of trees, birds in a forest depending upon the nature of data required.

**Probability and Non-probability Sampling**: The two standard ways to draw a sample are probability and non-probability sampling. If the purpose of the evaluation is to generalize for the whole group on the basis of sample results or to provide a statistical basis for saying that the sample is representative, a probability sample is appropriate. If the aim of the evaluation is to learn about individuals or cases for some purpose other than generalizing to the population, or if random selection is not feasible, sometimes study is limited to those participants that agree to be included then non-probability sampling is appropriate.

Probability sampling is a method of selecting samples according to certain laws of probability in which each unit of the population has some known and positive probability of its being selected in the sample. Because the probability is known, the sample statistics can be generalized to the population at large (at least within a given level of precision).

In non-probability sampling procedures, choice of selection of sampling units depends entirely on the discretion or judgment of the sampler. This method is called purposive or judgment sampling. In this procedure, the sampler inspects the entire population and selects a sample of typical units which he

considers close to the average of population. One thing common to all these non-probability sampling methods is that the selection of the sample is confined to a part of the population. None of the methods give a sample which can be considered to represent the entire population. A particular sample may prove to be very good or very bad but unless one has the knowledge of the complete population, it is not possible to know the performance of a particular sample. Moreover, since these methods lack a proper mathematical basis, these are not amenable to the development of the sampling theory. Nonprobabilty samples are quite convenient and economical.

Non-probability samples include haphazard, convenience, quota and purposive samples. Haphazard samples are those in which no conscious planning or consistent procedures are employed to select sample units. Convenience samples are those in which a unit is self-selected (e.g., volunteers) or easily accessible. Quota samples are those in which a predetermined number of units which have certain characteristics are selected. A sample of 50 small and 50 large farmers to be interviewed in a village is an example of this type. Researchers select units (e.g., individuals) for a purposive sample on the basis of characteristics or attributes that are important to the evaluation. The units used in a purposive sample are sometimes extreme or critical units. Suppose we are evaluating the adoption rates of a technology by farmers and we want to know if large farmers differ from small farmers. A sample of extreme units, e.g., farms of 10 or more hectares and farms of 2 or less hectares, would provide information to make this comparison.

**Sampling and Non-sampling Errors**: The sampling errors arise because estimates of parameter is based on a ‘part’ from the ‘whole’ population while non-sampling errors mainly arise because of some departure from the prescribed rule of the survey such as survey design, field work, tabulation & analysis of data etc. The sampling error usually decreases with increase in sample size (number of units selected in the sample) and is non-existent in a complete enumeration survey, since the whole population is surveyed. However, non-sampling errors are common to both to complete enumeration and sample surveys.

VARIOUS SAMPLING METHODS

A sampling method is a scientific and objective procedure of selecting units from the population and provides a sample that is expected to be representative of the population. One of the vital issues in sample surveys is the choice of a proper sampling strategy, which essentially comprise of a sampling method and the estimation procedure. In the choice of a sampling method there are some methods of selection while some others are control measures which help in grouping the population before the selection process. In the methods of selection, schemes such as simple random sampling, systematic sampling and varying probability sampling are generally used. Among the control measures are procedures such as stratified sampling, cluster sampling and multi-stage sampling etc. A combination of control measures along with the method of selection is called the sampling scheme.

We shall describe in brief the different sampling methods in the following sections.

Simple Random Sampling

Simple random sampling (SRS) is a method of selecting 'n' units (sample size) out of 'N' units (population size) such that every one of the non-distinct samples has an equal chance of being chosen. In practice, a simple random sample is drawn unit by unit. The units in the population are numbered from 1 to N. A series of random numbers between 1 and N are then drawn either by means of a table of random numbers or by means of a computer program that produces such a table. Sampling where each member of a population may be chosen more than once is called sampling with replacement. Similarly a method of sampling in which each member cannot be chosen more than once is called sampling without replacement.

***Procedure of Selecting a Random Sample***: Since probability sampling theory is based on the assumption of random sampling, the technique of random sampling is of basic significance. Some of the procedures used for selecting a random sample are as follows:

(i) Lottery Method,

(ii) Use of Random Number Tables.

***Lottery Method***: Each unit in the population may be associated with a chit/ticket such that each sampling unit has its identification mark from 1 to N. All the chits/tickets are placed in a container, drum or metallic spherical device, in which a thorough mixing is possible before each draw. Chits/tickets may be drawn one by one may be continued until a sample of the required size is obtained. When the size of population is large, this procedure of numbering units on chits/tickets and selecting one after reshuffling becomes cumbersome. In practice, it may be too difficult to achieve a thorough shuffling. Human bias and prejudice may also creep in this method.

***Use of Random Number Tables***: A random number table is an arrangement of digits 0 to 9, in either a linear or rectangular pattern, where each position is filled with one of these digits. A Table of random numbers is so constructed that all numbers 0, 1,2…,9 appear independent of each other. Some random Tables in common use are:

(a) Tippett's random number Tables

(b) Fisher and Yates Tables

(c) A million random digits Table.

A practical method of selecting a random sample is to choose units one-by-one with the help of a Table of random numbers. By considering two-digits numbers, we can obtain numbers from 00 to 99, all having the same frequency. Similarly, three or more digit numbers may be obtained by combining three or more rows or columns of these Tables. The simplest way of selecting a sample of the required size is by selecting a random number from 1 to N and then taking the unit bearing that number. This procedure involves a number of rejections since all numbers greater than N appearing in the Table are not considered for selection. The used numbers is, therefore, modified as remainder approach.

***Remainder Approach***: Let N be a r-digit number and let its r-digit highest multiple be N'. A random number k is chosen from 1 to N' and the unit with serial number equal to the remainder obtained on dividing k by N is selected. If the remainder is zero, the last unit is selected. As an illustration, let N = 123, the highest three-digit multiple of 123 is 984. For selecting a unit, one random number from 001 to 984 has to be selected. Let the random number selected be 287. Dividing 287 by 123, the remainder is 41. Hence, the unit with serial number 41 is selected in the sample.

Stratified Random Sampling

In SRS the precision of a sample estimate of the population mean depends not only upon the size of the sample but also on the population variability. Selection of a simple random sample from the entire population may be desirable when we do not have any knowledge about the nature of population, such as, population variability etc. However, if it is known that the population has got differential behaviour regarding variability, in different pockets, this information can be made use of in providing a control in the selection. The approach through which such a controlled selection can be exercised is called stratified sampling.

In stratified sampling, the whole population is divided into several homogenous groups (strata), thereby, controlling variability within each group and a random sample of pre-determined size is drawn independently from each one of the groups. To obtain full benefit from stratification, the strata sizes must be known. If a simple random sample is taken in each stratum then the procedure is termed as stratified

random sampling. As the sampling variance of the estimate of mean or total depends on within strata variation, the stratification of population is done in such a way that strata are homogeneous within themselves with respect to the variable under study. However, in many practical situations it is usually difficult to stratify with respect to the variable under consideration especially because of physical and cost consideration. Generally the stratification is done according to administrative groupings, geographical regions and on the basis of auxiliary characters correlated with the character under study.

Cluster Sampling

A cluster may be defined as a group of units. When the sampling units are clusters, the method of sampling is known as cluster sampling. Cluster sampling is used when the frame of units is not available or it is expensive to construct such a frame. Thus, a list of all the farms in the districts may not be available but information on the list of villages is easily available. For carrying out any district level survey aimed at estimating the yield of a crop, it is practically feasible to select villages first and then enumerating the elements (in this case farms) in the selected village. The method is operationally convenient, less time consuming and more importantly such a method is cost-wise efficient. The main disadvantage of cluster sampling is that it is less efficient than a method of sampling in which the units are selected individually.

*Multi-Stage Sampling*

Generally, elements belong to the same cluster are more homogeneous as compared to those elements which belong to different clusters. Therefore, a comparatively representative sample can be obtained by enumerating each cluster partially and distributing the entire sample over more clusters. This will increase the cost of the survey but the proportionate increase in cost vis-à-vis cluster sampling will be less as compared to increase in the precision. This process of first electing cluster and then further sampling units within a cluster is called as two-stage sampling. The clusters in a two-stage sample are called as primary-stage units (psu) and elements within a cluster are called as second-stage units (ssu).

A two-stage sample has the advantage that after psu's are selected, the frame of the ssu's is required only for the sampled psu's. The procedure allows the flexibility of using different sampling design at the different stages of selection of sampling units. A two-stage sampling procedure can be easily generalized to multi-stage sampling designs. Such a sampling design is commonly used in large scale surveys. It is operationally convenient, provides reasonable degree of precision and is cost-wise efficient.

*Systematic Sampling*

In systematic sampling, only the first unit is selected at random and then proceeds with the selection of every k-th unit from then onwards. In this case, k = (population size/sample size). The method of systematic sampling is used on account of its low cost and simplicity in the selection of the sample. It makes control of field work easier. Since every k-th unit will be in the sample, the method is expected to provide an evenly balanced sample.

Systematic sampling can be used in situations such as selection of k-th strip in forest sampling, selection of corn fields every k-th mile apart for observation on incidence of borers, or the selection of every k-th time interval for observing the number of fishing craft landing at a centre.

For example, suppose we wish to sample people from a long street that starts in a poor district (house #1) and ends in an expensive district (house #1000). A simple random selection of addresses from this street could easily end up with too many from the high end and too few from the low end (or vice versa), leading to an unrepresentative sample. Selecting every $10^{th}$ street number along the street ensures that the sample is spread evenly along the length of the street, representing all of these districts.

However, systematic sampling is especially vulnerable to periodicities in the list. If periodicity is present and the period is a multiple or factor of the interval used, the sample is especially likely to be unrepresentative of the overall population, making the scheme less accurate then simple random sampling. Another drawback of systematic sampling is that it is not possible to get an unbiased estimation of the variance of the estimator.

*Varying Probability Sampling*

In simple random sampling without replacement (SRSWOR), the selection probabilities are equal for all the units in the population. However, if the sampling units vary in size considerably, SRSWOR may not be appropriate as it doesn't take into account the possible importance of the larger units in the population. To give possible importance to larger units, there are various sampling methods in which this can be achieved. A simple methods is of assigning unequal probabilities of selection to the different units in the population. Thus, when units vary in size and the variable under study is correlated with size, probabilities of selection may be assigned in proportion to the size of the unit e.g. villages having larger geographical areas are likely to have larger populations and larger areas under food crops. For estimating the crop production, it may be desirable to adopt a selection scheme in which villages are selected with probabilities proportional to their populations or to their geographical areas. A sampling procedure in which the units are selected with probabilities proportional to some measure of their size is known as sampling with probability proportional to size (pps). The units may be selected with or without replacement. In sampling with replacement, the probability of drawing a specified unit at a given draw is the same. In this case sample is selected either through cumulative total method or Lahiri's method.

Now let us move to the practical consideration of sample survey.

**PLANNING AND EXECUTION OF SAMPLE SURVEYS**

Sample Surveys are widely used as a cost effective instruments of data collection and for making valid inferences about population parameters. Most of the steps involved while planning a sample survey are common to those for a complete enumeration. These major stages of a survey are **planning, data collection and tabulation of data**. Some of the important aspects requiring attention of the planning stage are as follows:

- Formulation of data requirements – objectives of the survey
- Ad-hoc or repetitive survey
- Method of data collection
- Questionnaire versus schedules
- Survey, reference and reporting periods
- Problems of sampling frames
- Choice of sampling design
- Planning of pilot survey
- Field work

The different aspects listed above are inter-dependent.

***(i) Formulation of Data requirements:***

The user i.e., the person or organizations requiring the statistical information are expected to formulate the objectives of the survey. The user's formulation of data requirements is not likely to be adequately precise from the statistical point of view. It is for the survey statistician to give a clear

formulation of the objectives of the survey and to check up whether his formulation faithfully reflects the requirements of the users.

***(ii) Survey: Ad-hoc or repetitive:***

An ad-hoc survey is one which is conducted without any intention of or provision for repeating it, whereas a repetitive survey is one, in which data are collected periodically for the same, partially replaced or freshly selected sample units. If the aim is to study only the current situation, the survey can be an ad-hoc one. But when changes or trends in some characteristics overtime are of interest, it is necessary to carry out the survey repetitively.

***(iii) Method of collecting Primary Data:***

There are variety of methods that can be used to collect information. The method to be followed has to be decided keeping in view the cost involved and the precision aimed at. The methods usually adopted for collecting primary data are:

- Recorded information
- Physical observation
- Face-to-face interviewing
- Postal enquiries
- Telephone interviews
- Web based survey etc.

***(iv) Questionnaire Vs. Schedule:***

In the questionnaire approach, the informants or respondents are asked pre-specified questions and their replies to these questions are recorded by themselves or by investigators. In this case, the investigator is not supposed to influence the respondents. This approach is widely used in mail enquiries. In the schedule approach, the exact form of the questions to be asked are not given and the task of questioning and soliciting information is left to the investigator, who backed by the training and instructions has to use his ingenuity in explaining the concepts and definitions to the informant for obtaining reliable information.

However, these two terms are often used synonymously. Designing questionnaire is one of the vital aspects of survey. Few suggestions for wording questions

- Use Simple words
- Questions should be concise
- Avoid multiple meaning questions
- Avoid ambiguous questions
- Minimum amount of writing on schedule
- Check on accuracy & consistency
- Handbook of instructions

***(v) Survey, Reference and Reporting Periods:***

Another aspect requiring special attention is the determination of survey period, reference period and reporting periods.

- Survey Period: The time period during which the required data is collected.
- Reference Period: The time period to which the collective data for all the units should refer.

- Reporting Period: The time period for which the required statistical information is collected for a unit at a time (reporting period is a part or whole of the reference period).

***(vi) Choice of Sampling Design:***

The principle generally adopted in the choice of a design is either reduction of overall cost for a pre-specified permissible error or reduction of margin of error of the estimates for given fixed cost. Generally a stratified uni-stage or multi-stage design is adopted for large scale surveys. For efficient planning, various auxiliary information which are normally available are utilized at various stages e.g. the area under particular crop available for previous years is normally used for size stratification of villages. If the information is available for each and every unit of the population and there is wide variability in the information then it may be used for selecting the sample through probability proportional to size methods.

***(vii) Pilot surveys:***

Where some prior information about the nature of population under study, and the operational and cost aspects of data collection and analysis is not available from part surveys. It is desirable to design and carry out a pilot survey. It will be useful for

- Testing out provisional schedules and related instructions
- Evolving suitable procedure for field and tabulation work, and
- Training field and tabulation staff
- Potential sources of measurement error
- Likely non-response rate
- Sensitive issues or sources of ambiguity
- Difficulties of access to chosen sample members

***(viii) Field Work:***

While planning the field work of the survey, a careful consideration is needed regarding choice of the field agency. For ad-hoc surveys, one may plan for ad-hoc staff but if survey is going to be a regular activity, the field agency should also be on a regular basis. Normally for regular surveys, the available field agencies are utilized. A regular plan of work by the Enumerators along with the rationalized supervision is an important consideration for getting a good quality of data. An initial quality check should be instituted while the interviewers are in the field to supply missing entries and correct apparent inconsistencies.

*Determination of sample size*

While planning a survey, a question often arises is that of fixing the size of the sample as unduly large sample size may mean wastage of resources while a smaller sample size limits the utility of results. The sample sizes are determined by fixing the precision of the estimate. It can be seen that sample size depends on population variance, which is generally not known. An estimated value of population variance can be obtained either from a pilot survey or by previous sampling of the same or similar population or by guess work about the structure of the population. Besides population variance, sample size depends on the minimum effect size we want to detect as well as power of the test. Illustration of power and sample size calculation through STATA will be discussed in the class. A do-file and log-file for power and sample size calculation will also be shared during the class.

REFERENCES

*Cochran, W.G. (1977) Sampling Techniques, Third Edition, John Wiley & Sons, New York.*

*Murthy, M.N. (1967) Sampling Theory and Methods, First Edition, Statistical Publishing Society, Calcutta.*

*Sukhatme, P.V., Sukhatme, B.V., Sukhatme, S. and Asok, C.(1984) Sampling Theory of Surveys with Applications. Third Edition, Iowa State University Press, USA and Indian Society of Agricultural Statistics, New Delhi.*

# Regression Adjustment and Inverse Probability Weighted Regression Adjustment

**Aditya K S**

*Scientist, Division of Agricultural Economics, ICAR-IARI, New Delhi*

## Introduction

Impact assessment methods have their roots in theory of causal inference. Theory of causal inference is concerned with establishing causation and estimating the magnitude of effect of the cause. Within the framework of causal inference, impact is defined as the expected change in the outcome of interest in the absence of the treatment/ intervention. In other words, impact assessment methods aim at estimating the difference in value of the outcome variable of the units receiving the treatment/ intervention to what would have happened in absence of the treatment. Since the value of outcome variable of treatment group in the absence of treatment is never observed, the value of outcome from the control group is used as proxy. However, in observational studies, due to non-random allocation of the treatment and due to confounding variables, it is very difficult to find a suitable counterfactual. Assessing impact in absence of suitable counterfactual can lead to either over or under estimation of the impact of the treatment. Impact assessment methods are aimed at constructing the suitable counterfactual outcomes, either through experimental setting or through quasi experiments and regression-based adjustments to minimize the element of bias in estimates of Impact.

In simple terms, to assess the impact the treated units and control units must be similar in all observable and unobservable characters except for the treatment. If this is satisfied, we can assess impact by simply comparing the outcome across the groups. In real world situation this is never satisfied and we need to account for the pre-treatment differences across the treated and control groups. Using dummy variable regression without accounting for the pre-treatment differences will lead to biased estimates due to confounding. Confounders are the variables which affect both the outcome as well as the treatment allocation. If we estimate the impact without controlling for confounding, the estimate will be biased and direction of bias cannot be determined too.

One simple method to address the problem of endogeneity and confounding is known as Regression Adjustment (RA). The Regression Adjustment model fits two separate regressions – for the treated and control units – and estimate the partial regression coefficients for all the control variables included in the model (dependent variable - outcome variable; like income). In the next step, the model estimates 'Potential Mean Outcomes' (PMO). PMO is the average value of the outcome if all the units in the sample are either in treated or control. For example, PMO for treatment is the mean income in case all the units in our sample were to be a beneficiary of the programme. In the Regression Adjustment model, we calculate the expected value of dependent variable for the entire sample based on coefficients of regression estimated on 'treated units'. Mean of the expected value is termed as PMO of the treated group. Similarly, expected value of dependent variable for the entire sample based on coefficients of regression estimated on control units is used to estimate PMO for control units. The difference between PMO of treated and control groups is considered as estimate of impact.

Again, a word of caution, the Regression Adjustment method is very sensitive to functional form of the outcome equation and model specification. In many cases, the estimate of impact changes drastically with addition/deletion of a control variable indicating model dependency leading to bias. In spite of these limitations, RA can be used as a method to assess impact, particularly when the size of sample is not large enough for semi-parametric matching methods, such as Propensity Score Matching.

For the demonstration, I will be using hh_98 data provided with the Khandker et al 2010 book. The data is regarding the impact of microfinance in Bangladesh. There are two 'treatment variables'; dmmfd – households which have a male microfinance participant and dfmmd – household which has a female microfinance member. The outcome variable of interest is the household expenditure – 'exptot'. Let us examine how Regression Adjustment method can be used estimate the impact of male microfinance participation on the total household expenditure.

First, let us see the variables in the dataset and the description.

```
. describe

Contains data from C:\Users\abc\Dropbox\Impact assessnent\PSM- NIAP\PSM data stata 11.dta
  obs:         1,129
 vars:            36                          25 Sep 2019 06:47
 size:       176,124
-------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------
nh              double  %7.0f                 HH ID
year            float   %9.0g                 Year of observation
villid          double  %9.0g                 Village ID
thanaid         double  %9.0g                 Thana ID
agehead         float   %3.0f                 Age of HH head: years
sexhead         float   %2.0f                 Gender of HH head: 1=M, 0=F
educhead        float   %2.0f                 Education of HH head: years
famsize         float   %9.2f                 HH size
hhland          float   %9.0g                 HH land: decimals
hhasset         float   %9.0g                 HH total asset: Tk.
expfd           float   %9.0g                 HH per capita food expenditure: Tk/year
expnfd          float   %9.0g                 HH per capita nonfood expenditure: Tk/year
exptot          float   %9.0g                 HH per capita total expenditure: Tk/year
dmmfd           byte    %8.0g                 HH has male microcredit participant: 1=Y, 0=N
dfmfd           byte    %8.0g                 HH has female microcredit participant: 1=Y, 0=N
weight          float   %9.0g                 HH sampling weight
vaccess         float   %9.0g                 Village is accessible by road all year: 1=Y, 0=N
pcirr           float   %9.0g                 Proportion of village land irrigated
rice            float   %9.3f                 Village price of rice: Tk./kg
wheat           float   %9.3f                 Village price of wheat: Tk./kg
milk            float   %9.3f                 Village price of milk: Tk./liter
potato          float   %9.3f                 Village price of potato: Tk./kg
egg             float   %9.3f                 Village price of egg: Tk./4 counts
```

Now, let us examine the mean values of the variables across the treated and control groups. We will use a user written command 'ttable2' to do a simultaneous t test for the independent variables. (Note: if you have not installed ttable2 before, then do it by using command 'ssc install ttable2')

```
. ttable2 sexhead agehead educhead lnland vaccess pcirr rice wheat milk oil egg, by(dmmfd)
```

| Variables | G1(0) | Mean1 | G2(1) | Mean2 | MeanDiff |
|---|---|---|---|---|---|
| sexhead | 909 | 0.891 | 220 | 0.977 | -0.086*** |
| agehead | 909 | 46.491 | 220 | 44.036 | 2.454*** |
| educhead | 909 | 2.215 | 220 | 2.741 | -0.526** |
| lnland | 909 | 0.392 | 220 | 0.324 | 0.067* |
| vaccess | 909 | 0.832 | 220 | 0.850 | -0.018 |
| pcirr | 909 | 0.562 | 220 | 0.553 | 0.009 |
| rice | 909 | 10.223 | 220 | 10.532 | -0.309*** |
| wheat | 909 | 7.452 | 220 | 7.530 | -0.078 |
| milk | 909 | 10.965 | 220 | 10.609 | 0.357 |
| oil | 909 | 39.398 | 220 | 39.426 | -0.028 |
| egg | 909 | 1.978 | 220 | 1.851 | 0.127*** |

Now we will estimate the regression adjustment model using teffects stata command.

```
. teffects ra (lexptot sexhead agehead educhead vaccess pcirr rice wheat milk oil) (dmmfd), atet aequations

Iteration 0:   EE criterion =  2.147e-27
Iteration 1:   EE criterion =  1.009e-31

Treatment-effects estimation                    Number of obs     =      1,129
Estimator      : regression adjustment
Outcome model  : linear
Treatment model: none
```

| lexptot | Coef. | Robust Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| **ATET** | | | | | | |
| dmmfd (1 vs 0) | -.046551 | .0326823 | -1.42 | 0.154 | -.1106072 | .0175051 |
| **POmean** | | | | | | |
| dmmfd 0 | 8.459292 | .0230563 | 366.90 | 0.000 | 8.414103 | 8.504482 |

Here coefficient of ATET is the impact which is -0.04651, however it is statistically not significant at 5% level of significance. So we conclude that there is no empirical evidence that the microfinance participation has an effect on the total household expenditure. If we want to see the coefficients of the two regressions used to estimate the Potential Mean Outcomes, we can use the following command.

| | | | | | | |
|---|---|---|---|---|---|---|
| **OME0** | | | | | | |
| sexhead | -.071767 | .0594638 | -1.21 | 0.227 | -.188314 | .0447799 |
| agehead | .0047035 | .00127 | 3.70 | 0.000 | .0022144 | .0071927 |
| educhead | .0609195 | .0046448 | 13.12 | 0.000 | .0518158 | .0700232 |
| vaccess | -.0250455 | .0456103 | -0.55 | 0.583 | -.1144401 | .064349 |
| pcirr | .1652743 | .0530079 | 3.12 | 0.002 | .0613807 | .2691679 |
| rice | .0014569 | .0101711 | 0.14 | 0.886 | -.018478 | .0213918 |
| wheat | -.0358387 | .0183419 | -1.95 | 0.051 | -.0717882 | .0001108 |
| milk | .030303 | .005062 | 5.99 | 0.000 | .0203815 | .0402244 |
| oil | .0066701 | .0039328 | 1.70 | 0.090 | -.0010381 | .0143783 |
| _cons | 7.755241 | .258653 | 29.98 | 0.000 | 7.24829 | 8.262192 |
| **OME1** | | | | | | |
| sexhead | .1207559 | .088697 | 1.36 | 0.173 | -.053087 | .2945988 |
| agehead | .0042302 | .0024225 | 1.75 | 0.081 | -.0005179 | .0089783 |
| educhead | .0338929 | .007527 | 4.50 | 0.000 | .0191403 | .0486456 |
| vaccess | -.0305123 | .087211 | -0.35 | 0.726 | -.2014428 | .1404182 |
| pcirr | .1056915 | .1323397 | 0.80 | 0.425 | -.1536896 | .3650725 |
| rice | .0360498 | .0157368 | 2.29 | 0.022 | .0052063 | .0668934 |
| wheat | -.0137422 | .0379113 | -0.36 | 0.717 | -.0880471 | .0605627 |
| milk | .0114263 | .0132784 | 0.86 | 0.390 | -.0145989 | .0374515 |
| oil | .0176904 | .004849 | 3.65 | 0.000 | .0081866 | .0271942 |
| _cons | 6.888147 | .4641885 | 14.84 | 0.000 | 5.978354 | 7.797939 |

## Inverse Probability Weighted Regression Adjustment (IPWRA)

Under the stronger condition of **conditional exchangeability**, wherein exchangeability holds within each strata of the confounding variables (i.eY (1),Y(0)⊥T|X), then there are methods that can be used to eliminate confounding and estimate the causal effect. This is the same principle used in the Propensity Score Matching (PSM). But here instead of using the Propensity Scores for matching, they are used as weights to adjust for the difference in preatreatment covariates.

Suppose that there are measurable differences between the control and treated groups. For example, educated farmers are more probable in the adopter category (treated group). But this doesn't mean that all the educated farmers will be adopters. Some of the educated farmers can always end up in control groups. In this case, it would make sense that comparing the outcome of few educated farmers in the control group with the outcome of the many educated farmers in the treatment group. So, in regression, our purpose is to give more weightage to those educated farmers in control group and to give less weightage to large number of educated farmers in the control group. On generalization, we could say that those who are more likely to be in treated group than the control group, but still in control group are a good counterfactual for the treated group (probability of being in treated group or control group calculated based on selection equation) should get higher weightage in the regression, as they are smaller in number. Whereas the much larger group of individuals who were placed in the expected treatment group need less weight, simply because there are so many of them and we have much more information on them.

Thus, for treated units', Weight = $(1/p_x)$ - (Less weightage to those units which has high probability of being in the treated group), where the $p_x$ is the Propensity Score of the particular unit. For control units, Weight = $(1/1-p_x)$ - (More weightage to those units which has high probability of being in the treated group). This method is also known as 'doubly robust method'. There are two models involved here in estimation, in the first step the propensity score is estimated and in the second stage, the outcome equation is estimated with propensity score as weights. Studies have indicated that only one these two models need to be correctly specified to estimate the impact and hence it is known as doubly robust method.

Now we will estimate the IPWRA using teffects command of stata

```
. teffects ipwra (lexptot sexhead agehead educhead vaccess pcirr rice wheat milk oil)(dmmfd sexhead agehead educhead vaccess pcirr rice wheat milk oil, pr
> obit), atet

Iteration 0:   EE criterion =  3.041e-17
Iteration 1:   EE criterion =  1.267e-31

Treatment-effects estimation                    Number of obs     =      1,129
Estimator      : IPW regression adjustment
Outcome model  : linear
Treatment model: probit
```

| lexptot | Coef. | Robust Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| **ATET** | | | | | | |
| dmmfd (1 vs 0) | -.0482186 | .0326016 | -1.48 | 0.139 | -.1121165 | .0156793 |
| **POmean** | | | | | | |
| dmmfd 0 | 8.46096 | .0230571 | 366.96 | 0.000 | 8.415769 | 8.506151 |

Here the variables in the first parenthesis indicates the selection equations to estimate the propensity score and the variables in the second parenthesis indicates the outcome equations. Though teeffects can estimate the IPWRA in one step, it is advisable to use pscore command first, estimate pscore and test the balancing property before using the teffects model. As in case of RA, the IPWRA also indicates that the estimate of impact is not statistically significant.

# Regression Adjustment Models and Regression Discontinuity Design for estimating impact

**Aditya K S and Subash S P**

*Scientist, Division of Agricultural Economics, ICAR-IARI, New Delhi*

## Regression Discontinuity Design

### Introduction

In non-experimental studies, it is difficult to get a proper counterfactual, which makes job of assessing impact a challenging task. However, discontinuities and delays in program implementation, based on eligibility criteria or other exogenous factors, can be very useful in non-experimental program evaluation in measuring the impact of interventions. The basic idea is that people above and below the threshold, assuming they are similar in observed characteristics, form a comparable treatment and counterfactual hence can be distinguished in terms of outcomes. However, the samples across both sides of discontinuity should be close to ensure homogeneity. Discontinuity approaches are therefore similar to instrumental variable (IV) methods because they introduce an exogenous variable that is highly correlated with participation, albeit not akin to participation.

For instance, take the example of a programme where only small farmers are eligible for the programme. Here, being in the treated or control group depends on amount of land one has. Hence, land is the forcing variable in this case and cutoff is at 1ha. If the land is less than 1 ha, unit will be under treated and if it is more than 1, then in control group. Underlying assumption is that units around the cutoff point, on either side of the cutoff point, are homogeneous except for treatment. RDD uses this assumption to measure impact.

### Regression Discontinuity Design

Regression Discontinuity Design is a quasi-experimental approach of impact assessment. It cues from discontinuity approach with assignment is determined by 'threshold' or a 'cut-off' point. Potential beneficiaries (units) just above the cut-off point are very similar to potential beneficiaries just below cut-off. However, only the units above the threshold receives the treatment. Hence, impact can be assessed by comparing outcomes for units just above and below cut-off (Figure 1).

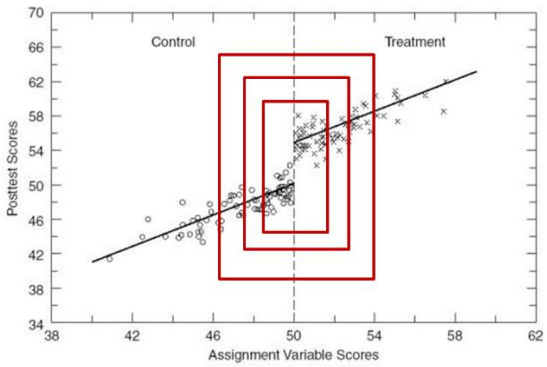


Figure 1. Regression discontinuity experiment with an effective treatment

The situation in which RDD would apply is with observations with narrower bandwidth and need many observations around the cut-off.

## Type of RDD

There are two kind RDD; Sharp RDD and Fuzzy RDD.

### *Sharp regression discontinuity setup*

This approach is used when forcing variable precisely determines the treatment status. Treatment status is a deterministic and discontinuous function of the 'forcing variable' (The variable, which determine the allocation of treatment, is called as 'forcing variable').

### *Fuzzy regression discontinuity setup*

This approach is used when forcing variable does not precisely determine the treatment status. There are two type of fuzzy. Type 1 fuzzy is a condition in which some treatment do not receive treatment. Type 2 fuzzy is a situation which is a combination of type 1 and some control revive treatment.

Standard nonparametric regression can be used to estimate the treatment effect in either the sharp or the fuzzy regression discontinuity setup. For a sharp discontinuity design, the treatment effect can be estimated by a simple comparison of the mean outcomes of individuals to the left and the right of the threshold.

## Graphical analysis

The steps involved in graphical analysis are given below

1. Divide forcing variable into bins
2. Plot forcing variable and average outcome variable in each bin
3. Superimpose a flexible regression line
4. Inspect whether there is discontinuity at the threshold

## Steps in applying RD approach

### *Step 1: Check whether the conditions given below are satisfied*

Two conditions are to be met before employing RDD approach.

1. Continuous eligibility index- assignment variable /forcing variable
2. Clearly defined cut-off score

### *Step 2: Check for internal Validity Assumption*

This includes two set of testing

**1. Covariate balance:** Test whether other covariates jump at the cut-off (Figure 2). Check: Re-estimate the RD model with covariates as the dependent variable.

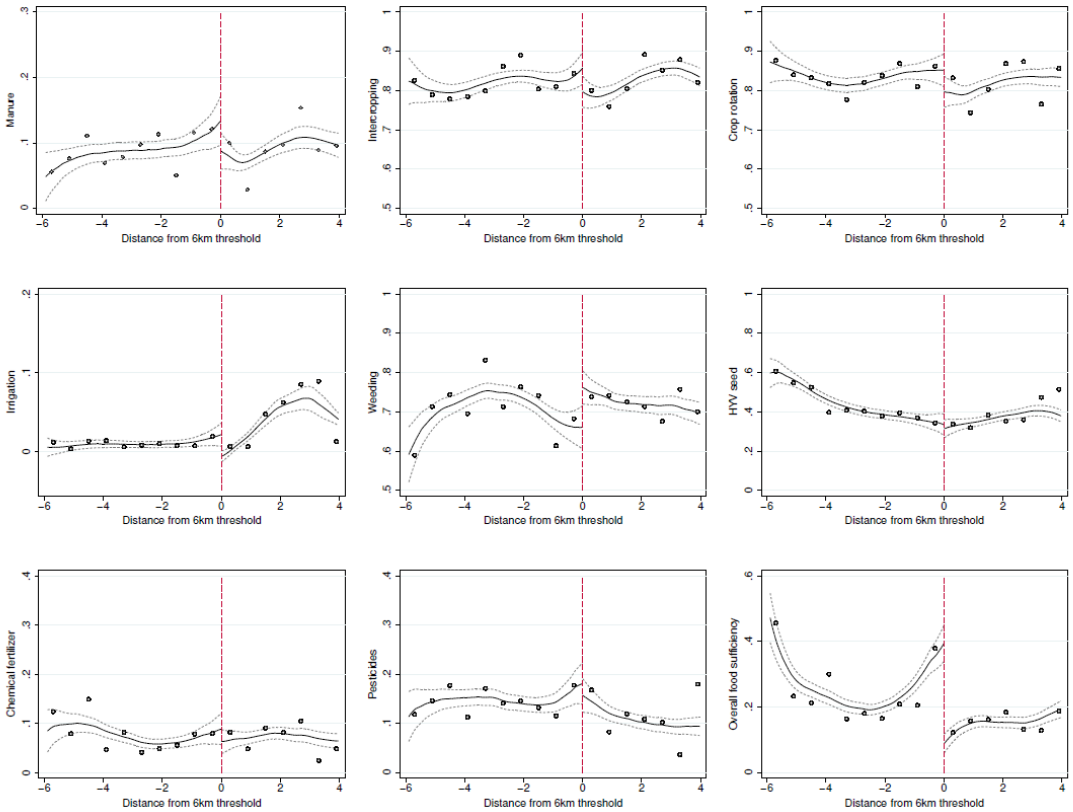
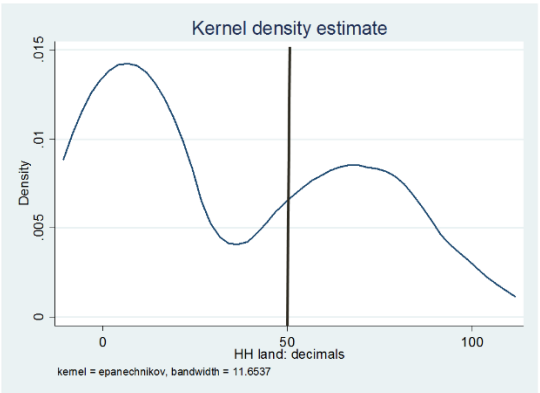



Figure 2. Checking covariates jump ate the cut-off point

**2. Continuous density of the forcing variable:** This is done to understand if there is a perceived benefit from a treatment (Self-select). Individual on one side systematically different from other side (self-selection bias). Check: Plot density of the forcing variable, no significant jump around cut-off (Figure 3).




Note: After dropping values of HH land above 100

Figure 3. Continuous density of forcing variable at cut off point

*Step 3: Bandwidth selection*

Optimal bandwidth is estimated by different methods provided in literature.

Advantages and Disadvantages

The advantages of the RD method are (a) that it yields an unbiased estimate of treatment effect at the discontinuity, (b) that it can many times take advantage of a known rule for assigning the benefit that is common in the designs of social policy, and (c) that a group of eligible households or individuals need not be excluded from treatment. However, the concerns with RD are (a) that it produces local average treatment effects that are not always generalizable; (b) that the effect is estimated at the discontinuity, so, generally, fewer observations exist than in a randomized experiment with the same sample size; and (c) that the specification can be sensitive to functional form, including nonlinear relationships and interactions.

One concern with the RD method is behavioral (Ravallion 2008). Program officials may not always know precisely the eligibility criteria; hence, behavioral responses to the program intervention may be confused with actual targeting rules. Data collected prior to program intervention, in the form of a baseline, for example, may help to clarify program design and corresponding uptake.

Another concern is that the exercise focuses only on individuals or units closely situated around the threshold. Whether this group is materially interesting for the evaluator needs to be addressed; if program officials are interested, for example, in identifying the effects of a program around a geographic border and in determining whether the program should be expanded across borders, the limited sample may not be as great a concern. A similar example can be constructed about a poverty alleviation program concerned with households whose status hovers near the poverty line.

If the eligibility rules are not adhered to or change over time, the validity of the discontinuity approach also needs to be examined more carefully. Robustness checks can be conducted to examine the validity of the discontinuity design, including sudden changes in other control variables at the cutoff point. Examining the pattern in the variable determining eligibility can also be useful—whether, for example, the average outcome exhibits jumps at values of the variable other than the eligibility cutoff, as well as any discontinuities in the conditional density of this variable. If the control data exhibit nonlinearities—for example, a steeper slope than the treatment data—then a squared term for the selection variable can be added in the outcome regression equation. Nonlinearities in the functional form can also be addressed by interacting the selection variable with the cutoff point or, perhaps, by using shorter, denser regression lines to capture narrower comparisons.

## Reference

*Khandker, Shahidur R. Gayatri B. Koolwal, Hussain A. Samad. 2010. "Handbook on impact evaluation : quantitative methods and practices". The International Bank for Reconstruction and Development / The World Bank. Washington DC.*

*Ravallion, Martin. 2008. "Evaluating Anti-poverty Programs." In Handbook of Development Economics, vol. 4, ed. T. Paul Schultz and John Strauss, 3787–846. Amsterdam: North-Holland.*

# Application of Psychometrics for Behavioural Research

**R.N.Padaria**

*Head/ Principal Scientist, Division of Agricultural Extension, ICAR-IARI, New Delhi*

Psychometrics refers to psychological measurement. One of the important applications of psychometrics has been construction and validation of scales and tests, which have been immensely used in agricultural extension research to measure psychological variables like attitude, achievement motivation, entrepreneurial orientation, risk orientation, etc.

**Approaches and methods of Scale Development**

*Arbitrary approach*: Very often when a researcher or an evaluator needs to understand whether there is difference in degree of judgment or response towards any characteristic or trait, scales are used. Sometimes a scale is developed with an arbitrary collection of statements based on heuristics and subjective judgment of the researcher and is administered to respondents for measuring the characteristics in question. Such scales are called arbitrary scales.

*Differential scales approach*: There are scales in which the items or statements are selected and rank ordered on a continuum by a group of experts. Such scales are known as differential scales. Various methods of judgments for relative ranking of statements based on Thurstone's principles are used like method of paired comparison or equal appearing intervals. Therefore, differentials scales are also referred as Thurstone's type scales. The person's response to the statement fixes his or her position on the continuum.

*Item analysis approach*: The statements are selected for a scale based upon its discriminatory power. Likert scale falls under this category. Since the total score of an individual is obtained by summation of scores of responses to all the statements of a scale, such scales are also called summated scales.

*Cumulative scale approach*: Cumulative scale or Guttman scale lay emphasis upon unidimensionality of a scale and it is checked through scalogram analysis. A scale is unidimensional if the statements of scale fall along a single dimension. There are two techniques for conducting scalogram analysis i.e. Cornell technique and Goodenough technique. A perfect scale makes it possible to reproduce the responses to the individual statements from knowledge of total scores. Scalogram analysis provides an estimate of coefficient of reproducibility, which indicates the percent accuracy with which responses to the various statements can be reproduced from the total scores.

*Factor analysis approach*: Factor analysis is another approach for scale development. Scale developed through factor analysis is called factor scale. Factor analysis helps to determine the number of latent variables underlying a set of statements. Semantic Differential (S.D.) and the multidimensional scaling are based upon factor analysis.

*Comparative and non-comparative scaling techniques*

The scaling techniques can be compared as comparative scales and non-comparative scales. The scaling technique in which there is a direct comparison of stimulus objects with each other is known as comparative scale. Paired comparisons, rank order, constant sum scales, Q-sort and other procedures are comparative scales. On the contrary, the stimulus objects are scaled independently of other objects in the stimulus set, it is known as non-comparative scale. It can be continuous rating or itemized rating

scales. The itemized rating scale can be classifies as Likert scale, semantic differential scale and staple scale.

**Modern approaches to psychological scaling**

The psychometric methods could be divided into three major classes viz., psychological scaling; factor analysis, and test theory. Psychological scaling comprises a set of techniques for assignment of quantitative values to objects or events based upon the data obtained through human judgment. Factor analysis methods aim at explaining the observed co-variation among a set of variables. Item Response Theory (IRT) assumes that one or more unobserved (latent) variables underlie the responses to test items in the sense that variation among the individuals on those latent variables explains the observed co-variation among item responses.

*Item Response Theory (IRT):* Though Classic test theory (CTT) has been the basis for developing psychological scales and test scoring for many decades, Item Response Theory (IRT), is a new approach being applied by psychometricians for explaining and analyzing the relationship between the characteristics of an individual and his/her response to the individual items. Item Response Theory (IRT) emphasizes that besides the item properties like item difficulty and item discrimination, a respondent's response to an item of a psychological scale or test also very much depends upon the standing of the respondent on the psychological characteristic being measured by the item.

IRT provides information about the quality of a scale's item and of the entire scale. There are several important uses of IRT. With item information and test information items having good discriminative ability could be identified. The second important use of IRT is examination of differential item function. It occurs when an item functions differently in different groups of respondents. The third key use is examination of person-fit. Analysis of person-fit identifies people whose response pattern does not fit the expected pattern of responses to a set of items. IRT also facilitates Computerized Adaptive Testing (CAT) intended to produce accurate and efficient assessment of individual's psychological characteristics.

*Factor analysis for evaluating dimensionality, internal structure and validity of scale*

Factor analysis (FA) is the most important statistical tool for validating the structure of our instruments. There are other components of construct validity that are not addressed by factor analysis. FA is usually a two-stage process. The first stage of FA offers a systematic means of examining inter-relationships among items on a scale. This stage of FA is exploratory factors analysis. Exploratory factor analysis (EFA) is the most common method of evaluating the dimensionality of psychological scales. If all scale items are well correlated with each other at about equal levels, the scale is unidimensional. Exploratory factor analysis (EFA) is useful when a researcher has a few hypotheses about a scale's internal structure. On the contrary, when a researcher has a clear hypothesis about a scale i.e. the number of factors or dimensions underlying its items, links between items and factors, and the association between factors, Confirmatory factor analysis (CFA) is useful. Confirmatory factor analysis (CFA) is a statistical method appropriate for testing whether a theoretical model of relationships is consistent with a given set of data. CFA allows researchers to evaluate the degree to which their measurement hypotheses are consistent with actual data produced. CFA facilitates theory testing, theory comparison, and theory development in a measurement context.

Application of Structural Equation Modeling (SEM) has gained attention for scale development. While EFA is used to study single relationships individually, SEM deals with multiple dependence

relationship. Structural modeling refers to the systematic identification of possible relationship among concepts. The structure of relationship is represented with mathematical equations.

*Multidimensional scaling (MDS):* MDS represents a set of stimuli as points in a multidimensional space in such a way that those points corresponding to similar stimuli are located close together, while those corresponding to dissimilar stimuli are located far apart. The basic idea behind MDS is similarity/dis-similarity data or proximity data obtained by various spatial distance models. Most frequently used model is Euclidean model.. For social science, Non-metric MDS is the most suited one since data are mostly at ordinal level. MDS can be used rigorously in the field of extension in measuring multi-dimensional variables, attitude, perception, semantic differential, positioning of innovation, audience segmentation, targeting, discovering underlying behavioral and personality factors, understanding audience preferences *etc.* Since it is based on respondents' subjective perception and subjective evaluation it gives us greater understanding of our target clientele and individual differences prevailing among them.

**Testing the reliability and validity of the scale**

***Reliability****:* It refers to the accuracy or precision of the measuring instruments. Reliability can be defined in terms of relative absence of errors of measurement in a measuring instrument.

**Methods to measure reliability:** There are different methods of testing reliability of any psychological measurement tools such as Test-retest method, Parallel forms method, and Split- half method.

*Test-retest:* As the name of the method signifies, the scale to be tested for reliability, is administered to a group of individuals at two points of time (usually at a gap of 15 to 30 days) and the scores obtained are correlated. Value of correlation r gives us the reliability coefficient. Higher the value r, higher is the reliability of the test.

*Parallel form or equivalent form:* Two separate scales comprising similar items on the psychological object are used simultaneously. Both the tests are equivalent in terms of their items. The two tests on psychological objects are administered one after the other to a group of subjects and the scores on the two scales are then correlated. The value of r obtained is the reliability coefficient and is also known as coefficient of equivalence.

*Split- half method*- In this method, the test scores on one half of the scale are correlated with the scores on other half. The split- half method is based on the assumption that if the test items are homogeneous and there is internal consistency the scores on any item or set of items on the scale would yield high correlation value with any other item or set of items (the number of items being the same in the two subsets). The r-value worked out in this case is for half of the scale; hence, to have the reliability coefficient of the entire test scale we need to apply 'Spearman-Brown formula'

$$r_{tt} = \frac{nr_1}{1+(n-1)r_1}$$

rtt = reliability of the original test

r = reliability coefficient of the subsets

n =number of times the length of the original test-is shortened

***Validity*:** It is the degree to which a measuring instrument measures what it is supposed to measure. Validity refers to the appropriateness of the instruments/test. A test is said to be valid when it measures what it is supposed to measure. Statistically, validity is the problem of common factor variance to total variance. According to Guttman, there are two broad types of validity, i.e., internal and external.

*Internal validity* expresses a logical relationship between the theoretical and operational definition of the concept under study.

*External validity* expresses an empirical relationship between the theoretical definition and the operational definition.

Validity has different levels viz, content, criterion, and construct validity. Content validity is evaluated by determining the degree to which the items of a scale/test represent the universe of content of the object phenomenon being measured by it and their adequacy. It is related with the representativeness and adequacy of the context of the test/scale. Criterion validity has two forms i.e. Predictive validity and Concurrent validity. *Predictive validity* is estimated by showing how accurately we can guess some future performance, on the basis of the measure on external or other criteria, e.g., on the basis of the score on leadership scale of an individual, one can predict accurately his behavior as a manager. While predictive validity is used for forecasting the presence or absence of the trait in future, based on the scores on the criteria obtained today. The concurrent validity is based on simultaneous comparison of scores on one test/scale with other established criteria. Construct validity is the strongest and the most meaningful validity compared to the other two. Construct validity is useful in validating the construct, the theory behind the test. It is evaluated by a determination of the relationship between the test attitude score and other aspects of the individual personality

# Qualitative Techniques useful for impact studies (thematic analysis/ Content analysis)

**P. Sethuraman Sivakumar**

*Principal Scientist, ICAR- Central Tuber Crops Research Institute, Kerala*

Qualitative approaches are increasingly being used to asses he impact of field intervention programme mainly due to their ability "to explain causality"" in realistic way. These methods are integrated with quantitative approaches to provide detailed account of "cause-effect relationship" in these field programmes.

Generally, the 'qualitative' and 'quantitative' approaches differ in the "type of data generated" during the research work, while quantitative approaches produce data in the form of numbers while qualitative research produce "prose or text" data.

Hentschel's (1999) proposed a method-data framework (Fig. 1) to help the researchers to choose "right method" based on contextual and qualitative/ quantitative approaches.

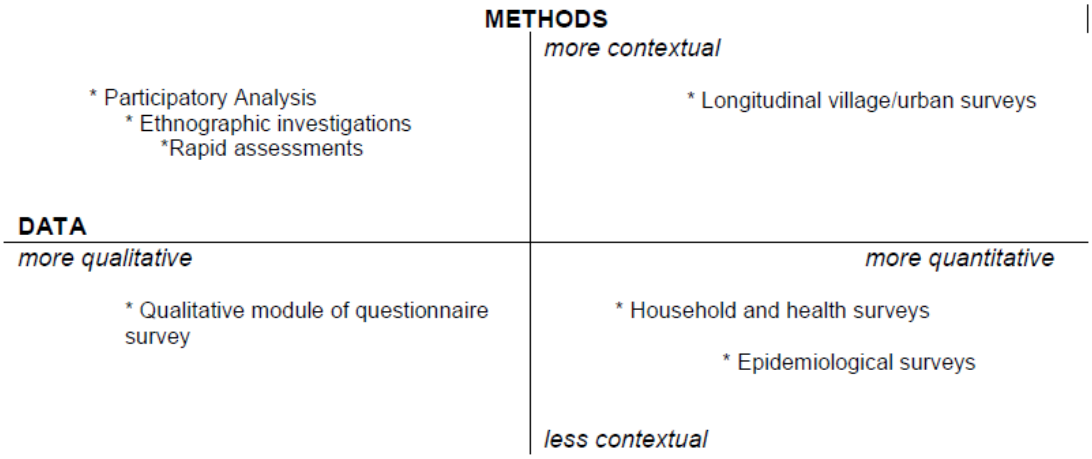



Fig 1. The method data framework for choosing the data collection method for impact analysis (Hentschel, 1999)

**Thematic analysis**

Thematic analysis is one of the popular qualitative approaches, used for identifying, analysing, and reporting patterns within data concerning a social phenomenon (Barun and Clarke, 2006). Though thematic analysis is considered as a foundational method for qualitative analysis, many social researchers used it to assist other forms of qualitative analysis (Holloway and Todres, 2003), primarily due to inadequate methodological rigor and lack of clarity in implementing this methodology (Nowell *et al*., 2017). The purpose of this paper is to introduce thematic analysis as a research method for assessing large volumes of narrative data to derive meaningful interpretation about the phenomenon under study.

**Content Analysis Vs Thematic Analysis**

In general, the qualitative approaches seek to understand a phenomenon from the people who are experiencing it, through a structured research framework. Along this framework, there is a considerable overlap among these approaches in terms of methods, procedures, and techniques. Both qualitative content

analysis and thematic analysis share same goal of assessing the phenomenon by breaking the narrative text into relatively small units of content and submitting them to descriptive treatment (Sparker, 2005). However, there are few functional differences among these methods (Vaismoradi *et al.*, 2013) (Table 1).

Table 1. Differences between content analysis and thematic analysis

| | **ontent analysis** | **hematic analysis** |
|---|---|---|
| urpose | o describe the characteristics of th rrative of a specific phenomenon b amining who says what, to whom, and wit hat effect. | o identify common threads or theme om that extend across an entir rrative |
| esearch approach | ixed methods – Both qualitative an antitative | ualitative |
| ocus of research desig | escription and more interpretation tha ematic analysis | inimal description and interpretation |
| onsideration of contex data | anger of missing context- only th equency of codes is counted to fin gnificant meanings in the text | ombines analysis of the meanin erived from data within particula ntext |
| ontent type | onsiders either manifest (developin tegories) or latent contents (developin emes). | onsiders both manifest and laten ntents |
| ature of theme | erived based on frequency of occurrence o ntent; represents only surface meaning | bstract, mostly latent and derive rough an intense qualitative process |
| apping of themes | o | es |
| ssessment of reliabilit coders | ssessed through inter-coder reliability | ode book, Audit Trails |

(Adapted from Vaismoradi *et al.*, 2013)

**When to use Thematic Analysis**

Thematic analysis is suitable for understanding a phenomenon through stakeholder's views, opinions, knowledge, experiences or values, derived from a set of qualitative data. The common sources of qualitative data are audio/video recorded personal interviews; audio-recorded telephonic interviews; blogs on a specific topic; social media posts on a specific topic; case studies/success stories; newspaper reports; other audio or video recordings of events, views, etc.; reports; policy documents; and feedback forms.

**Different approaches to thematic analysis**

Themes or patterns within data can be identified in one of two primary ways in thematic analysis: in an inductive or 'bottom up' way or in a theoretical or deductive or 'top down' way (Braun and Clarke, 2006).

✓ Inductive approach – Data- driven analysis which involves coding the data without trying to fit it into a pre-existing coding frame, or the researcher's analytic preconceptions

✓ Deductive approach – Analyst-driven approach which follows researcher's theoretical or analytic interest in the area.

Besides, thematic analysis also looks into nature of themes – semantic and latent ( Braun & Clarke, 2006).

✓ Semantic themes – Focus on the surface meanings of the data confined to what a participant has said or what has been written.

✓ Latent themes – Analysis extend beyond respondents views to identify or examine the underlying ideas, assumptions, and conceptualisations and ideologies, which shape semantic content of the data

**Advantages of thematic analysis**

Thematic analysis has several advantages over other qualitative methods (Braun and Clarke, 2006; King, 2004; Nowell *et al.*, 2017).

- Highly flexible approach customised to the needs of researchers to assess complex qualitative data.
- Useful method for examining the phenomenon from perspectives of different research participants, which help in highlighting similarities and differences, and generating unanticipated insights.
- Useful for summarizing key features of a large data set.

**Disadvantages**

Though thematic analysis has several advantages, it has few limitations too (Braun and Clarke, 2006; Holloway and Todres, 2003; Nowell *et al.*, 2017).

- The lack of substantial literature on thematic analysis compared to other qualitative methods like grounded theory, ethnography, and phenomenology.
- It does not allow researcher to make claims about language use
- Inconsistency and lack of coherence when developing themes derived from the research data

**Steps in conducting Thematic Analysis**

Considering the strengths and limitation of thematic analysis as a qualitative approach, a systematic approach and trustworthy approach is proposed following the guidelines suggested by Braun and Clarke (2006) and Nowell et al., (2017) on the trustworthiness criteria suggested by Lincoln and Guba (1985).

**Step 1: Define the research problem and question**

The researcher develop the research problem in a systematic way by specifying appropriate research goals, along with clear, concise and sound research questions.

**Step 2: Sampling strategy**

The sampling strategy is the plan devised by the researcher to ensure that the sample chosen for the research work represents the selected population. Robinson (2014) proposed a four-point sampling process for systematically selecting adequate samples for obtaining quality results. It involves (i) Defining a sample universe – inclusion/ exclusion criteria for respondents; (ii) Selecting adequate samples; (iii) Choosing relevant sampling method and (iv) ways of sourcing samples.

**Step 3: Collect data in a systematic way**

The researcher collects the data in a systematic way following the sampling strategy in an unbiased and error free manner. After collecting the data, a data corpus containing all the information gathered for the thematic analysis is prepared.

**Step 4: Transcription and translation of verbal data**

The collected data is often in the form of audio or video forms, which need to be converted into textual form for analysis. It involves two processes – Transcription and translation.

Transcription is the process of converting the verbal data i.e., interviews, audio/video clips and speeches into written form of the same language (Barun and Clarke, 2006) while translation is the process of translating

the transcribed verbatim into English or any other language in written form. A systematic process of objective ways of doing transcription and translation is described by Chen and Boore (2009).

**5. Familiarising with data**

At this stage, the researcher makes a quick glance at the verbatim/ transcripts as a whole and takes notes from first impressions. Then, the verbatim/transcripts are thoroughly read line by line by looking for meanings and patterns from the data. The trustworthiness of data can be ensured through (i) Triangulating different data collection modes; (ii) documenting theoretical and reflective thoughts, (iii) documenting thoughts about potential codes/themes and (iv) Storing data properly along with all records including field notes, transcripts, and reflexive journals (Nowell et al., 2017)

**6. Generating initial codes or labels**

After initial reading, identify the data extracts or specific word(s) which represent a dimension of the research problem. The coding process is performed systematically across the entire data set by collating data relevant to each code.

During the coding process, the researcher highlights or underlines the specific data extracts while reading the transcript. The researcher can also write them down in a separate notebook for grouping in the later stages.After identifying the data extract, assign a code to it based on your perception of what it signifies.Group all the codes along with relevant data extracts and prepare a long list of codes and data extracts.

At this phase, trustworthiness can be ensured through (i) Peer debriefing; (ii) Researcher triangulation; (iii) Reflexive journaling; (iv) Use of a coding framework; (iv) Audit trail of code generation; and (v) Documentation of all team meeting and peer debriefings (Nowell et al., 2017).

**7. Searching for themes**

This phase involves sorting the identified codes into preliminary levels of themes based on their perceived closeness, and pooling all the relevant coded data extracts into the identified themes. A theme is essentially a coherent and meaningful pattern in the verbatim/ transcript relevant to the research question. It is technically a construct or a dimension of the construct. Visual techniques like mind maps, tables and cards may be used to pool the relevant codes into preliminary themes.

At this phase, trustworthiness can be ensured through (i) Researcher triangulation and (ii) Storing detailed notes about development and hierarchies of concepts and themes (Nowell et al., 2017)

At the end of this phase, the preliminary themes, and sub-themes, and all extracts of data that have been coded in relation to them are identified and plotted.

**8. Reviewing themes**

At this stage, the researcher checks if the themes work in relation to the coded extracts and the entire data set, generating a thematic map. The reviewing and refining are performed following Patton's (1990) dual criteria for judging categories - internal homogeneity and external heterogeneity

**Level 1. Internal Homogenity** - Reviewing at the level of the coded data (Reviewing the codes and data extracts)

✓ Read all the collated codes and respective data extracts for each theme and sub-theme to check if the data forms a coherent pattern.

✓ If the main and sub- themes do not fit, you would rework your theme, creating a new theme, finding a home for those extracts that do not.

**Level 2. External Homogenity** - Over all reviewing of themes with the data set Consider each theme in relation to your data corpus.

✓ Generating and checking Thematic map - Do the relationships between the themes reflect the meaning of your data as a whole?

At this phase, trustworthiness can be ensured through (i) Researcher triangulation; (ii) vetting of themes and subthemes by team members; (iii) conducting test for referential adequacy by returning to raw data (Nowell et al., 2017)

**Step 9: Defining and Naming Themes**

This phase captures the essence of what each theme is about and what aspect of the data each theme captures. In this phase, the themes, sub-themes are examined carefully to see that they are coherent and internally consistent. Each theme get a name - concise, punchy and immediately give the reader a sense of what the theme is about. The final thematic map is drawn with description.

At this phase, trustworthiness can be ensured through (i) researcher triangulation; (ii) peer debriefing; (iii) team consensus on themes; (iv) documentation of team meetings regarding themes; and (v) Documentation of theme naming (Nowell et al., 2017).

**Step 10: Producing the Report**

The final phase begins once the researcher has fully established the themes and is ready to begin the final analysis and write-up of the report (Braun and Clarke, 2006). The write-up of a thematic analysis should provide a concise, coherent, logical, non-repetitive, and interesting account of the data within and across themes (Braun and Clarke, 2006).

At this phase, trustworthiness can be ensured through (i) member checking, (ii) peer debriefing; (iii) describing process of coding and analysis in sufficient details; (iv) thick descriptions of context; (v) description of the audit trail; (vi) report on reasons for theoretical, methodological, and analytical choices throughout the entire study (Nowell et al., 2017).

**Software for thematic analysis**

Thematic analysis is often performed manually since it involves identification of semantic and latent themes. With the recent advances in Natural Language Processing, Verbatim analysis or text analytics and Word embedding applications, many qualitative analysis software have inbuilt capacities to do thematic analysis. Popular software used for thematic analysis are Nvivo (https://www.qsrinternational.com/nvivo-qualitative-data-analysis-software/home), QDS Miner (https://provalisresearch.com/products/qualitative-data-analysis-software/freeware/) and ATLAS.ti (https://atlasti.com/). The QDA Miner Lite- free version of QDA Miner is popular freeware for conducting thematic analysis.

**References**

*Braun, V., and Clarke, V. (2006). Using thematic analysis in psychology. Qualitative Research in Psychology, **3**: 77–101.*

*Chen, H-Y. and Boore, J.R.P. (2009.) Translation and back-translation in qualitative nursing research: Methodological review. Journal of Clinical Nursing **19**:234–239*

*Hentschel, J. (1999) Contextuality and data collection methods: A ramework and application to health service utilisation. In: The Journal of Development Studies 35, pp: 64-94.*

*Holloway, I., and Todres, L. (2003). The status of method: Flexibility, consistency and coherence. Qualitative Research, **3:** 345–357.*

*King, N. (2004). Using templates in the thematic analysis of text. In C. Cassell & G. Symon (Eds.), Essential guide to qualitative methods in organizational research (pp. 257–270). London, UK: Sage.*

*Lincoln, Y., and Guba, E. G. (1985). Naturalistic inquiry. Newbury Park, CA: Sage.*

*Nowell, L. S., Norris, J. M., White, D. E., and Moules, N. J. (2017). Thematic Analysis: Striving to meet the trustworthiness criteria. International Journal of Qualitative Methods,* ***16 (1),*** *1-13.*
*Robinson, O.C. (2014.) Sampling in interview-based qualitative research: A theoretical and practical guide. Qualitative Research in Psychology,* ***11****:25–41.*
*Vaismoradi, M., Turunen, H., and Bondas, T. (2013). Content analysis and thematic analysis: Implications for conducting a qualitative descriptive study. Nursing and Health Sciences,* ***15(3),*** *398–405.*

# Practical Manual

## Advances in Research Methodology for Social Sciences

**March 17- 27, 2021**

**Division of Agriculture Economics, ICAR-IARI, New Delhi**


**Course Director**

**Alka Singh**
Professor and Head
Division of Agricultural Economics
ICAR-Indian Agricultural Research Institute
New Delhi

**Course Coordinators**

**Anjani Kumar**
Senior Research Fellow
International Food Policy Research Institute
New Delhi

**Praveen K V**
Scientist
Division of Agricultural Economics
ICAR-Indian Agricultural Research Institute
New Delhi 110 012

**Aditya K S**
Scientist
Division of Agricultural Economics
ICAR-Indian Agricultural Research Institute
New Delhi

**Nithyashree M L**
Scientist
Division of Agricultural Economics
ICAR-Indian Agricultural Research Institute
New Delhi



**Division of Agricultural Economics**
**ICAR – Indian Agricultural Research Institute**
**New Delhi – 110012**
**www.nahep-caast.iari.res.in**